

第30屆全球華人計算機教育應用大會

The 30th Global Chinese Conference on Computers in Education

Education in the Age of AI

AI時代的教育

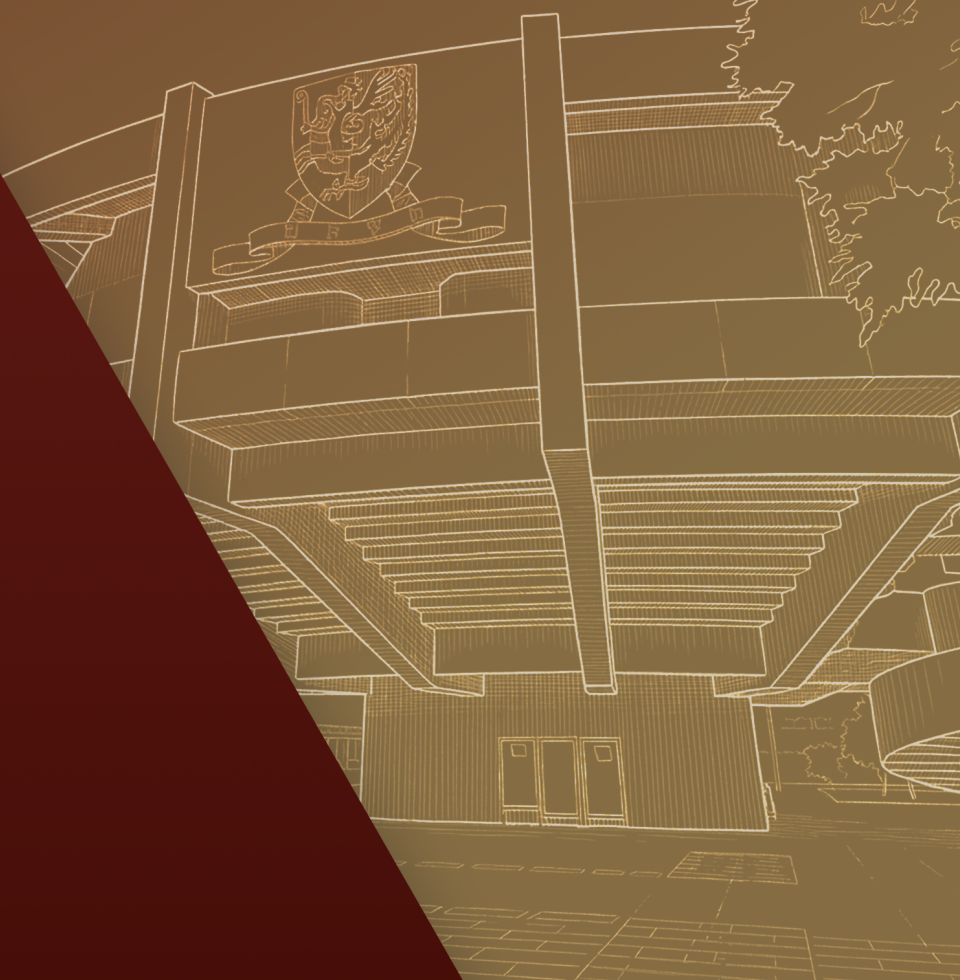
何以可用、何以為界、何以可能
Opportunities, Boundaries, and Capabilities

英文論文集

Main Conference Proceedings
(English Paper)



30/5/2026 – 3/6/2026



The 30th Global Chinese Conference on Computer Education (GCCCE 2026)

Main Conference Proceedings (English Paper)

主編	楊宗凱、尚俊杰、莊紹勇、陳高偉、許庭嘉、林日坤、吳砥、 尹弘飈、蔡敬新、趙建豐
發行人	楊宗凱
出版者	全球華人計算機教育應用學會
地址	320 台灣桃園市中壢區中大路 300 號
電話	+886-3-4227251 #35456
網址	http://www.gcsce.org/
電子郵件	secretary@gcsce.org
出版格式	PDF 檔
出版年月	2026 年 6 月初版
ISSN	ISSN 3005-3226

Enhancing CSL Learners' Story Writing through Metacognitive Instruction and AI Support	4
Integrating AI in Subject Teaching for Metacognition Development of Students: Descriptive Essay Writing Chatbot for In-service Chinese Language Teachers	7
Students' Engagement in Learning Programming Using a GenAI-Driven Chatbot: Insights from Sentiment and Epistemic Network Analysis	15
Toward Trustworthy AI in Higher Education: An Empirical Evaluation of GenAI in Management Education	19
A Comprehensive Evaluation Model for Vocational Education Development: based on GRA-TOPSIS Integration and Its Application	23
From AI Feedback to Peer Reflection: A Dual-Phase Reflective Writing Loop Model Using a Generative AI-Assisted Collaborative Dialogic Feedback	27
AI-Supported Iterative Revision and Reflective Writing in Primary Education: A Quasi-Experimental Study	35
Design and Feasibility of a Modular Prompt Framework for Polya-Based Mathematical Instructional Scaffolding	40
Integrating Generative AI into Strategy-Based Instruction: Adaptive Scaffolding for Self-Regulated Learning in Secondary EFL Writing	44
Non-homogenizable Vocal Habitus: Podcast as Narrative Practice for Learning Subjectivity and Community Building	52
Examining the Factors Affecting Learning Engagement in Self-Directed Learning Using Mobile Technology in Higher Education	56
Roles and Strategies of Generative AI in Multimodal English Instruction: A Qualitative Study of AI-Native English Teachers' Experiences	64
Reconfiguring Assessment in the Age of AI: Making Learning Visible through Meta-Task Awareness (MTA)	68
Hong Kong-French Collaboration in Machine Learning Education for Junior Secondary Students: A Mixed-Methods Study on Teacher Development	76
LLM Enhanced Video to Text Recognition for Authentic EFL Video Making Tasks: Evaluation across GPT Versions and Video Lengths	80
AI Agents as Cognitive Partners in Micro-Project-Based Learning: A Professional Development Workshop	84
Effect of Design Thinking on Students' Creativity: A meta-analysis	88
CET Professionals' AI Readiness and the Integration of Generative AI in Professional Education: A Mixed-Methods Descriptive Study	95
An Ecological Framework for L2 Teachers' ICT Integration: A Theory-Driven Review (2015–2025)	99

Improving Rural School Teaching Quality through Blended Synchronous Classroom : A Value-Added Assessment Study.....	103
The Effectiveness of AR Technology in Chemistry Education: A Meta-Analysis Based on 14 Chinese Studies.....	111
Effect of an AR-based Gamified Environmental Education Platform on the Environmental Consciousness of Grade 5 Students: a Case Study	113
Exploring the Pattern and Influence Factors in the Correlation Between Learning Engagement and Academic Achievement in Online Learning.....	115
Reframing Catalog Design as Transmedia Brand Storytelling: A Curriculum Intervention and Mixed-Methods Evaluation	123
Trends in Artificial Intelligence Research for Geometry Education: Educational Applications and System-Oriented Approaches	125
From "Wheel-Spinning" to "Productive Failure": Research on LLM-based Hierarchical Scaffolding Mechanism in Programming Education.....	129
Turning Course Descriptions into Skill Evidence: A Case Study of Two Undergraduate Programs	133
A Study on Personalized Programming Learning for High School Students Supported by Large Language Models.....	136
Impact of AI Support with Different Level of Learner Agency on Behavioral Engagement in Providing and Uptaking Peer Feedback.....	144
Linking Design Thinking and Computational Thinking to Mathematical Modelling Self-Efficacy and AI Pedagogical Role Preferences.....	153
The Impact of a Gamified Online Learning Course on the Grade 4 Students' Engagement in Mathematics Learning: a Mixed Methods Research.....	157
The Impact of Interacting with Generative Artificial Intelligence on College Students' Creativity Performance	166
Comparing Supporter, Arguer, and Comprehensive Debater: How AI Debate Agent Roles Influence Critical Thinking.....	171
InspireRV: Namecard-Sized RISC-V Computer for Foundational Digital Literacy and Computational Thinking.....	175
Student Engagement and Data Science Skills during Authentic Collaborative Inquiry: Insights from Contrasting Group Analysis.....	179
Teacher Perspectives on AI Integration: Pedagogies, Technologies, Confidence, and Readiness Across Disciplines.....	187
iChat-Job: Designing Self-Determination Theory-Grounded System with Digital Humans for Increasing University Graduates' Employability.....	195

Enhancing CSL Learners' Story Writing through Metacognitive Instruction and AI Support

Sarah H.J. Liu

Kainan University

sarah.hjliu@gmail.com

Abstract: *This pilot study examines whether metacognitive instruction combined with ChatGPT can support Chinese as a Second Language (CSL) learners' story writing. Eight undergraduate learners completed three narrative writing tasks and a closed-ended questionnaire. Overall writing scores improved significantly ($p < .05$) across the three assignments; organization received the highest score, whereas word choice/conventions obtained the lowest. The learners also reported generally positive attitudes toward ChatGPT, especially for ease of use, but comparatively weaker usefulness and metacognitive engagement. The findings suggest that AI-supported writing can be beneficial when accompanied by explicit strategy instruction, although the extent of improvement depends on how actively learners evaluate and apply AI feedback during revision.*

Keywords: Artificial Intelligence (AI), ChatGPT, Metacognitive Instruction, Chinese as a Second language

1. Introduction

Writing in a second language (L2) is a demanding process because it requires the integration of linguistic knowledge, idea development, and self-regulation. In addition to vocabulary and grammar, writers must plan what to write, monitor whether ideas are coherent, and evaluate whether a text meets task expectations (Alamri, 2019; Godwin-Jones, 2022). Recent studies in L2 have suggested that artificial intelligence (AI) tools such as ChatGPT can partially address this challenge by providing instant and personalized feedback (Pitychoutis, 2024). AI tools have been reported to enhance opportunities for interaction and support writing development (Kohnke, Moorhouse, & Zou, 2023). However, AI tools do not automatically lead to better learning outcomes. Their pedagogical value depends on whether learners use them critically and strategically. From a metacognitive perspective (Flavell, 1976; Yeh, 2015), particularly in L2/FL contexts, the effectiveness of AI feedback should be understood in relation to learners' planning, monitoring, and evaluation. Learners who actively regulate their writing are more likely to benefit from AI-generated suggestions than those who simply copy corrections without reflection.

2. Methodology

The participants were eight international CSL learners recruited from a university setting. Two were Japanese (25%) and six were Vietnamese (75%). All participants were at the B1 or B2 level according to the Test of Chinese as a Foreign Language (TOCFL), representing intermediate to upper-intermediate proficiency. The study lasted five weeks. In Week 1, students were introduced to the text-chat function of ChatGPT and received basic technical training. In Weeks 2 to 4, they completed three story-writing tasks based on different 10-minute Mr. Bean video clips. For each task, students first wrote their own approximately 150-character stories without AI assistance and then revised them with ChatGPT. The design was intended to ensure that the initial content originated from learners rather than from the AI tool.

Metacognitive instruction was explicitly embedded in the writing process. In the planning stage, students predicted what might happen in the story before watching the video. In the monitoring stage, they identified the main character and key events while viewing the clip. In the evaluation stage, they summarized and revised their writing, using ChatGPT feedback to improve expression and accuracy. Students were also asked to submit screenshots of AI interactions and final revised drafts. Writing performance was evaluated across four categories: content, organization, grammar, and word

choice/conventions; each category corresponded with the scores from 0 to 10, the higher scores that represented more accuracy (Hsiung, Lee, & Sung, 2015). A questionnaire was used to measure learners' attitudes toward ChatGPT, perceived ease of use, perceived usefulness, and metacognitive processes on a 5-point Likert scale (from 1 = strongly disagree to 5 = strongly agree). The revised drafts were analyzed, and the researcher established 92% intra-rater agreement through repeated coding.

3. Results, Discussions, & Conclusion

The results revealed a significant improvement in learners' overall story-writing performance across the three assignments, as indicated by the Friedman test ($p < .05$). The total mean score increased from 19.88 ($SD = 3.19$) in the first assignment to 22.31 ($SD = 2.15$) in the second and 24.38 ($SD = 2.25$) in the third. This pattern indicates that students gradually produced better revised texts over time. A closer look at the four writing categories revealed that organization received the highest overall mean score ($M = 17.38$, $SD = 2.08$), followed by content ($M = 16.81$, $SD = 2.00$), grammar ($M = 16.56$, $SD = 1.64$), and word choice/conventions ($M = 15.81$, $SD = 1.28$). The learners improved most in the overall organization of their stories, but their lexical precision and orthographic control remained weaker. In addition, text length increased across tasks, with the third assignment reaching an average of 190 words, suggesting that students became more able or willing to elaborate their ideas.

Questionnaire results showed generally positive perceptions of ChatGPT. Ease of use received the highest percentage of agreement (84.38%), followed by attitudes toward use (71.87%). Usefulness and metacognitive processes both received lower but still positive levels of agreement (68.75%). Within the metacognitive dimension, self-evaluation was the weakest item, with a mean of 3.1 on the 5-point scale, whereas planning and monitoring received higher means of 4.5 and 4.3, respectively. In other words, learners felt more confident using the tool than critically evaluating their own writing.

Pedagogically, the study suggests that teachers should not treat AI as a stand-alone writing assistant. Instead, they should explicitly train learners to plan before writing, monitor meaning while drafting, and evaluate AI feedback critically during revision. This is particularly important in CSL contexts, where orthographic accuracy poses a persistent challenge. Teachers may also need to give direct support on character form and radical awareness because general-purpose AI tools are not specifically designed for Chinese orthography.

Acknowledgements

This study was funded by Taiwan's Science and Technology Council under grant no. NSTC 114-2410-H-424-002.

References

- Alamri, B. (2019). Exploring metacognitive strategies employed by ESL writers: Uses and awareness. *International Journal of English Linguistics*, 9(1), 159–168.
- Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 231–235). Hillsdale, NJ: Lawrence Erlbaum.
- Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology*, 26(2), 5–24.
- Hsiung, Y. W., Lee, H. H., & Sung, Y. T. (2014). Examining the ACTFL writing assessment rating scale for L2 Chinese learners. *Journal of Chinese Language Teaching*, 11(4), 111–139.
- Kohnke, L., Moorhouse, B.L., & Zou, D. (2023b). ChatGPT for language teaching and learning. *RELC Journal*, 1–14. DOI: 10.1177/00336882231162868
- Pitychoutis, K.M. (2024). Harnessing AI chatbots for EFL essay writing: A paradigm shift in language pedagogy. *Arab World English Journal Special Issue on ChatGPT*, 1, 197–209.
- Yeh, H. C. (2015). Facilitating metacognitive processes of academic genre-based writing using an online writing system.

Computer Assisted Language Learning, 28(6), 479–498.

Integrating AI in Subject Teaching for Metacognition Development of Students:

Descriptive Essay Writing Chatbot for In-service Chinese Language Teachers

Siu Cheung, Kong ^{1*}, John Chi-Kin, Lee ², May Hung May, Cheng ², Boyuan Ren ³, Wenxi Hu ³

¹ Department of Mathematics and Information Technology, The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, New Territories, Hong Kong

² Department of Curriculum and Instruction, The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, New Territories, Hong Kong

³ Artificial Intelligence and Digital Competency Education Centre, The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, New Territories, Hong Kong

* sckong@eduhk.hk

Abstract:

This study designed a subject-specific AI chatbot for teaching descriptive essay writing for in-service primary school Chinese language teachers in Hong Kong. The intervention aimed to enhance teachers' Technological Pedagogical and Content Knowledge (TPACK), metacognition, AI literacy, and self-efficacy in AI-integrated teaching through two two-hour workshops with hands-on activities using a customised AI chatbot. A paired-sample t-test was used to analyse pre- and post-workshop survey data, with 102 valid responses for TPACK and metacognition, and 127 for AI literacy and self-efficacy. Before analysing the results, a reliability analysis was conducted to confirm the validity of the survey instruments used to measure TPACK, metacognition, AI literacy, and self-efficacy. Results showed statistically significant improvements in pedagogical content knowledge (PCK), technological content knowledge (TCK), overall TPACK, metacognition, AI literacy, and self-efficacy ($p < 0.001$), while content knowledge (CK) remained unchanged, likely due to teachers' pre-existing proficiency in subject knowledge. This research advances TPACK-based teacher professional development (TPD) by demonstrating that subject-specific, interactive interventions effectively foster teachers' TPACK competency and underlying skills, highlighting the potential of subject-knowledge-customised TPD for sustainable AI integration in language education. This study contributes to the design of TPD in integrating AI in language education for the development of students' metacognition.

Keywords: *AI Chatbot, AI literacy, Metacognition, Teacher professional development, TPACK*

1. Introduction

The integration of technology into education has highlighted the significance of Technological Pedagogical and Content Knowledge (TPACK) as a foundational framework for effective teaching and learning (Dogan et al., 2025; Ning et al., 2022). It also plays an important role in the metacognition development of students as teachers with strong TPACK can effectively use AI tools to support personalized goal-setting, progress tracking, and collaboration, fostering students' self-regulation, self-awareness, and metacognitive growth (Koh et al., 2017). Despite its importance, existing teacher professional development (TPD) initiatives often emphasize technology acceptance and hands-on capabilities in designing technological resources for instructional practices. However, it is still underexplored how key constructs critical for sustainable TPACK development, such as metacognition, AI literacy, and self-efficacy, can be influenced through TPD (Budianto et al., 2023). To address these limitations, this study designed a customized AI chatbot for teaching descriptive writing and delivered two offline professional development workshops to in-service primary school Chinese language teachers in Hong Kong. The workshops aimed to enhance teachers' TPACK, as well as their metacognitive skills,

AI literacy, and self-efficacy in using AI for teaching. Employing a paired-sample t-test to analyse pre- and post-workshop survey data, the study evaluates the effectiveness of the customized TPD intervention, focusing on identifying differential improvements across TPACK subdomains and their cognitive well-being. This research contributes to advancing TPD by providing empirical evidence of how subject-specific AI tools can foster teachers' professional competencies for technology-enhanced instruction in authentic educational contexts.

2. Literature Review

2.1. TPACK framework

The TPACK framework, initially conceptualized by Mishra and Koehler (2006), has emerged as a widely recognized model for understanding the integration of technology into teaching and learning processes. Building on Shulman's (1986) concept of Pedagogical Content Knowledge (PCK), TPACK adds "technological knowledge" (TK) as a crucial dimension, reflecting the growing significance of technology in education. The framework emphasizes that the effective use of technology in teaching requires a dynamic interplay between three core knowledge domains: content knowledge (CK), pedagogical knowledge (PK), and technological knowledge (TK). The intersections among these domains create three additional constructs—technological pedagogical knowledge (TPK), technological content knowledge (TCK), and pedagogical content knowledge (PCK)—which culminate in the overarching concept of TPACK. This integrated knowledge is essential for designing and delivering effective technology-enhanced instruction that aligns with both pedagogical strategies and subject-specific learning goals.

2.2. Teacher Professional Development in the Digital Age

Teacher professional development (TPD) refers to a range of educational and training activities designed to enhance teachers' knowledge, skills, and practices, enabling them to meet the evolving demands of education and effectively integrate various knowledge domains into their teaching (Sancar, et al., 2021).

Within the study of evaluating TPD programs, the TPACK framework has been extensively documented in the literature (Bustamante, 2020). Research indicates that teachers trained in TPACK exhibit greater confidence, creativity, and adaptability in their instructional practices (Chaipidech et al., 2021). Koehler and Mishra (2008) emphasize the importance of involving teachers in authentic tasks during workshops, as these activities allow them to experiment with technology in subject-specific contexts. Studies have identified hands-on workshops and personalized consultations as effective strategies to develop TPACK through professional development programs. These strategies provide active learning experiences, enabling participants to explore the interplay between technology, pedagogy, and content knowledge (Crosthwaite et al., 2021; Hu et al., 2022; Kong & Lai, 2022).

In addition to advancing competencies in the TPACK framework, TPD must also address emerging competencies such as AI literacy, self-efficacy, and metacognition. For example, increasing numbers of professional development courses are focused on enhancing educators' AI literacy to meet the demands of an AI-saturated era (Gašević et al., 2023; Kong et al., 2024; Kong & Hu, 2025). These programs not only explore AI's unique capabilities but also foster an understanding of underlying algorithms and challenges, including ethical dilemmas, misinformation, AI hallucinations, and the complexities of human-AI collaboration (Casal-Otero et al., 2023; Du et al., 2024; Kong & Zhu, 2025). Consequently, TPD supports teachers in recognizing AI's potential to enhance instructional strategies while cultivating higher-order thinking skills such as critical thinking, problem-solving, and self-regulation in the context of AI implementation (Ilgun Dibek et al., 2025;). Furthermore, increasing attention is being paid to teachers' AI competency and self-efficacy, with scales being developed to assess and enhance this crucial skill (Chiu et al., 2025; Yao & Wang, 2025). To summarize, these studies emphasize the importance of empowering teachers to confidently and effectively select and utilize AI tools in authentic learning environments. This, in turn, leads to more engaging instruction and improved teaching outcomes (Yang et al., 2024).

However, the effectiveness of TPD largely depends on its design, particularly the inclusion of interactive and collaborative components to actively engage participants (Sancar et al., 2021). One persistent challenge to the success of TPD programs lies in the “one-size-fits-all” approach, which often fails to integrate subject-specific knowledge with AI tools or provide open access to teaching and practice materials for authentic classroom contexts. To address this gap, this study designed a customized AI chatbot for teaching scenery description, providing teachers with a valuable resource that can be implemented in authentic teaching contexts. This study aims to answer the following research questions:

RQ1: Do the subject-specific professional development workshops significantly improve in-service Chinese primary school teachers’ TPACK (including CK, PCK, TCK, and overall TPACK) in Hong Kong?

RQ2: To what extent do the workshops enhance teachers’ metacognition skills, AI literacy, and self-efficacy in using AI tools for authentic teaching practices?

3. Methodology

3.1. Participants and instruments

This study involved two professional development (TPD) workshops designed for in-service Chinese primary school teachers in Hong Kong. Each workshop was conducted in a face-to-face format and lasted two hours. During the sessions, an expert introduced the functionality of a customized AI chatbot specifically designed to teach primary school students how to describe scenes. The AI chatbot is built on GPT-4.1 as the base model and utilizes the Flowise AI platform. It is designed to assist elementary students in learning how to write vivid scenery descriptions. By analyzing their text, the chatbot provides friendly suggestions and step-by-step guidance to help students improve their writing with rich imagery and appropriate language, all delivered in a gentle and encouraging tone. Additionally, the chatbot development incorporates local official education guidelines and materials as part of its Retrieval-Augmented Generation (RAG) process.

The workshops also included training on AI literacy, equipping teachers with an understanding of the underlying algorithms of AI and fostering a critical perspective toward AI-generated information. The content of the TPD workshops is summarized in Table 1. In this course, the AI chatbot was specifically tailored for teaching scene description, with its responses carefully designed to align with the educational objectives for primary school students. The chatbot’s output was constrained to ensure it did not exceed the scope of these learning goals, while also integrating seamlessly with teacher-led instruction and student self-regulated practice. This approach allowed the chatbot to serve as a supplementary learning tool rather than encouraging dependency on AI, ensuring that students remained actively engaged in the learning process. By embedding the chatbot within this structured pedagogical framework, the course aimed to enhance students’ metacognitive skills and promote their ability to learn independently.

Table 1.

The contents of teacher professional development workshops

Workshop	Topics	Surveys
Workshop 1	1. Introduction to teaching descriptive writing with AI technology.	(1) TPACK survey (pre-course)
	2. Exploring Speech-to-Text techniques for descriptive writing.	(2) AI literacy survey (pre-course)
	3. Enhancing descriptive writing skills using AI chatbots.	(3) Metacognition survey (pre-course)

	4. Understanding the role of AI in education: unleashing human potential rather than replacing students' learning in Chinese writing.	(4) Self-efficacy survey (pre-course)
Workshop 2	1. Understanding Generative AI and Large Language Models.	(1) TPACK survey (post-course)
	2. Introduction to effective prompt design.	(2) AI literacy survey (post-course)
	3. Teaching prompt engineering for subject-specific learning.	(3) metacognition survey (post-course)
	4. Exploring the ethical use of AI in education.	(4) Self-efficacy survey (post-course)

The first workshop, attended by 144 teachers, was followed by a second workshop two weeks later, which included 137 participants. At the start of the first workshop, teachers completed a pre-course 5-likert survey assessing four areas: their TPACK knowledge (12 items), skills for development metacognition of students (10 items), AI literacy (6 items), and self-efficacy in using AI in teaching practices (8 items). The post-course, featuring the same set of questions, was administered at the conclusion of the second workshop. Among them, we received 102 responses on TPACK and skills for development metacognition of students, and 127 responses on AI literacy and self-efficacy.

Table 2.

Reliability analysis of Chinese teacher survey

Items		Cronbach's Alpha
TPACK domains	CK	0.875
	PCK	0.896
	TCK	0.884
	TPACK	0.908
	Overall	0.950
Metacognition		0.959
AI Literacy		0.932
Self-efficacy		0.947

These instruments place additional emphasis on assessing whether teachers, through workshops, can develop an understanding of how to enhance students' metacognition. For instance, in the TPACK framework, the item “*I understand the uncertainty in the content generated by AI chatbots and can leverage this characteristic to facilitate students' self-regulated learning*” under the TPK construct highlights the importance of helping teachers comprehend the inherent uncertainty in AI-generated content while enabling students to develop self-regulation in their learning processes. Similarly, within the AI literacy instrument, the focus is on evaluating whether teachers have acquired concrete skills to effectively utilize AI chatbots in the classroom. For example, the item “*I have various ways and strategies for refining prompts when using AI chatbots*” measures the extent to which the instrument equips teachers with practical capacities to interact with AI and incorporate it into teaching practices. Lastly, in the self-efficacy instrument, the item “*I can guide my students to cross-verify the information generated by AI chatbots in various ways*” is designed to assess whether

teachers have the ability to guide students in critically evaluating and validating AI-generated information, thereby fostering higher-order cognitive skills.

A reliability analysis was conducted to assess the internal consistency of the survey instrument for its different sections. Cronbach's alpha was used as the measure of reliability. The results of the reliability analysis for each section are presented in Table 2. The reliability analysis demonstrated good internal consistency for the survey instrument, with a Cronbach's alpha exceeding the acceptable threshold of 0.70. This suggests that the items within the survey are measuring the intended constructs reliably.

3.2. Data analysis

This study assessed normality using the Shapiro-Wilk test and visual inspection of Q-Q plots for all variables. While the Shapiro-Wilk test indicated significant deviations from normality for some variables ($p < 0.05$), the Q-Q plots showed data points closely following the diagonal line, suggesting approximate normality. Given the large sample size ($n=102$) and the robustness of t-tests to minor normality violations, this study proceeded with the t-tests. This study employed a paired-sample t-test to examine participants' improvements in four key areas: TPACK, metacognition, AI literacy, and self-efficacy. The paired-sample t-test was chosen to compare participants' pre-workshop and post-workshop scores, allowing for an analysis of changes within the same group over time. By focusing on these constructs, the study aimed to evaluate the effectiveness of two teacher professional development workshops in enhancing educators' knowledge, skills, and confidence in integrating technology, artificial intelligence, and reflective practices into their teaching.

4. Results

The following section shows the paired-sample t-test result of this study across the measured domains. The detailed analysis results are shown in table 3 below. The analysis of TPACK components reveals significant improvements across all domains other than CK. While CK showed a slight increase in mean scores from 4.12 ($SD = 0.43$) to 4.20 ($SD = 0.49$), this difference was not statistically significant ($t = -1.66$, $p = 0.1$). In contrast, PCK and TCK displayed significant improvements. Specifically, PCK increased from 3.49 ($SD = 0.67$) to 3.94 ($SD = 0.58$), with a highly significant improvement ($t = -6.28$, $p < 0.001$). Similarly, TCK improved from 3.77 ($SD = 0.56$) to 4.09 ($SD = 0.58$) ($t = -5.42$, $p < 0.001$). The overall TPACK construct also improved significantly, with mean scores increasing from 3.50 ($SD = 0.69$) to 3.93 ($SD = 0.65$) ($t = -6.12$, $p < 0.001$).

Table 3:

Paired-sample t-test results of in-service Chinese teachers

		Mean (pre-test)	SD (pre-test)	Mean (post-test)	SD (post-test)	t	df	p
TPACK	CK	4.12	0.43	4.20	0.49	-1.66	110	0.1
	PCK	3.49	0.67	3.94	0.58	-6.28	110	<0.001
	TCK	3.77	0.56	4.09	0.58	-5.42	110	<0.001
	TPACK	3.50	0.69	3.93	0.65	-6.12	110	<0.001
Metacognition		3.76	0.50	3.95	0.55	-3.38	126	<0.001
AI Literacy		3.31	0.63	3.82	0.64	-8.25	126	<0.001
Self-efficacy		3.09	0.69	3.77	0.60	-10.54	126	<0.001

Metacognition showed a significant improvement from a pre-test mean of 3.76 ($SD = 0.50$) to a post-test mean of 3.95 ($SD = 0.55$). The observed increase ($t = -3.38$, $p < 0.001$) indicates that the workshops had a statistically significant positive impact on the teachers' metacognitive abilities. Additionally, AI literacy demonstrated one of the most notable improvements in this study, with the mean score increasing from 3.31 ($SD = 0.63$) to 3.82 ($SD = 0.64$). The t-value of -8.25 and a p-value of < 0.001 suggest a significant enhancement in participants' understanding and application of AI

concepts because of the workshops. Self-efficacy experienced the largest improvement among all constructs, with the mean score rising from 3.09 (SD = 0.69) to 3.77 (SD = 0.60). The t-value of -10.54 and a p-value of < 0.001 indicate that the workshops had a profound and statistically significant effect on teachers' confidence in their abilities.

In summary, the results indicate significant improvements in PCK, TCK, overall TPACK, metacognition, AI literacy, and self-efficacy following the professional development workshops. CK was the only construct that did not demonstrate a significant change. These findings highlight the effectiveness of the workshops in enhancing teachers' professional competencies, particularly in areas related to technology integration, reflective practice, and confidence in teaching.

5. Discussion

This study contributes to the teacher development scholarship by both corroborating established insights and introducing novel dimensions to understanding effective TPD for AI-integrated instruction. The results validate the efficacy of structured, hands-on TPD interventions (Bustamante, 2020; Chaipidech et al., 2021), the workshops yielded statistically significant improvements in teachers' PCK, TCK, overall TPACK, and self-efficacy, affirming the TPACK framework's central role in guiding technology-enhanced teacher training. Additionally, the emphasis on context-specific practice via a tailored AI chatbot for scene description teaching aligns with the existing literature (Crosthwaite et al., 2021; Hu et al., 2022), who identified subject-aligned active learning as critical for developing TPD. Most importantly, this model not only facilitates the effective integration of AI tools into classroom practices but also enhances teachers' confidence in supporting students' learning with AI. This focus on teacher capacity-building highlights the importance of aligning AI-driven innovations with subject-specific instructional goals and learner-centered approaches, ultimately fostering teachers' professional growth.

These divergences yield important theoretical and practical implications. Theoretically, they highlight the need to conceptualize teacher professional development as a holistic process that encompasses metacognition and AI literacy, rather than limiting it to core knowledge domains. Additionally, this study focuses on foundational constructs and contextualized design, advancing a more comprehensive model of TPD in AI-enhanced language education, not only supporting the integration of AI technologies but also fosters the development of students' metacognitive skills.

6. Conclusion and Future Research

This study evaluated the impact of customized TPD workshops, incorporating a subject-specific AI chatbot for teaching scene description, on in-service Chinese primary school teachers in Hong Kong. Paired-sample t-test results revealed statistically significant improvements in teachers' PCK, TCK, overall TPACK, metacognition, AI literacy, and self-efficacy. In contrast, CK showed only a marginal, non-significant increase, which is likely due to teachers' pre-existing subject proficiency. These findings address critical gaps in TPD scholarship, moving beyond "one-size-fits-all" models and demonstrating that subject-specific, interactive TPD effectively enhances not only core TPACK competencies but also understudied skills (metacognition, AI literacy) critical to sustainable technology integration. Practically, the results highlight the value of integrating domain-specific AI tools in TPD and tailoring focus areas to teachers' existing competencies. Limitations of the current study include its focus on Hong Kong Chinese primary school teachers and a single AI chatbot application, which restricts the generalizability of the results to other disciplines and cultural contexts. Future research should expand to a diverse teacher sample and explore long-term impacts on classroom practices and student outcomes, as well as the mechanisms linking AI literacy, metacognition, and self-efficacy to TPACK development. The second limitation is that this study was based on the comparison of pre and post-test of a single group without a control group. Future studies are suggested to examine the effects of intervention with a control group and other statistical evidence. The third limitation is that the intervention of this study lasted only four hours, future studies are suggested to examine the long-term effects on teachers' TPACK progression.

Acknowledgements

The authors would like to acknowledge the project “Using Artificial Intelligence (AI) in Primary School Education to Unleash Potential of Primary Students in Five Subjects: Chinese Language, English Language, Mathematics, Primary Science, and Primary Humanities” with the funding support from the Teaching Development Grant 2025-28 triennium, The Education University of Hong Kong.

References

- Bustamante, C. (2020). TPACK-based professional development on web 2.0 for Spanish teachers: a case study. *Computer Assisted Language Learning*, 33(4), 327–352. <https://doi.org/10.1080/09588221.2018.1564333>
- Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., & Barro, S. (2023). AI literacy in K–12: A systematic literature review. *International Journal of STEM Education*, 10(1), 29. <https://doi.org/10.1186/s40594-023-00418-7>
- Chaipidech, P., Kajonmanee, T., Chaipah, K., Panjaburee, P., & Srisawasdi, N. (2021). Implementation of an andragogical teacher professional development training program for boosting TPACK in STEM education. *Educational Technology & Society*, 24(4), 220–239. <https://www.jstor.org/stable/48629257>
- Chiu, T. K. (2025). Responsible digital citizen: building AI ethics awareness across subjects. *Interactive Learning Environments*, 33(4), 2759–2761. <https://doi.org/10.1080/10494820.2025.2494914>
- Chiu, T. K., Ahmad, Z., & Çoban, M. (2025). Development and validation of teacher artificial intelligence (AI) competence self-efficacy (TAICS) scale. *Education and Information Technologies*, 30(5), 6667–6685. <https://doi.org/10.1007/s10639-024-13094-z>
- Collie, R. J., Martin, A. J., & Gasevic, D. (2024). Teachers’ generative AI self-efficacy, valuing, and integration at work: Examining job resources and demands. *Computers and Education: Artificial Intelligence*, 7, 100333. <https://doi.org/10.1016/j.caeai.2024.100333>
- Crosthwaite, P., Luciana, & Wijaya, D. (2021). Exploring language teachers’ lesson planning for corpus-based language teaching: A focus on developing TPACK for corpora and DDL. *Computer Assisted Language Learning*, 36(7), 1392–1420. <https://doi.org/10.1080/09588221.2021.1995001>
- Dogan, S., Nalbantoglu, U. Y., Celik, I., & Agacli Dogan, N. (2025). Artificial intelligence professional development: A systematic review of TPACK, designs, and effects for teacher learning. *Professional Development in Education*, 51(3), 519–546. <https://doi.org/10.1080/19415257.2025.2454457>
- Du, H., Sun, Y., Jiang, H., Islam, A. Y. M., & Gu, X. (2024). Exploring the effects of AI literacy in teacher learning: An empirical study. *Humanities and social sciences communications*, 11(1), 1–10. <https://doi.org/10.1057/s41599-024-03101-6>
- Gašević, D., Siemens, G., & Sadiq, S. (2023). Empowering learners for the age of artificial intelligence. *Computers and education: artificial intelligence*, 4, 100130. <https://doi.org/10.1016/j.caeai.2023.100130>
- Hu, C. C., Yeh, H. C., & Chen, N. S. (2022). Teacher development in robot and IoT knowledge, skills, and attitudes with the use of the TPACK-based Support-Stimulate-Seek approach. *Interactive Learning Environments*, 31(9), 5811–5830. <https://doi-org.ezproxy.eduhk.hk/10.1080/10494820.2021.2019058>
- Ilgun Dibek, M., Sahin Kursad, M., & Erdogan, T. (2025). Influence of artificial intelligence tools on higher order thinking skills: A meta-analysis. *Interactive Learning Environments*, 33(3), 2216–2238. <https://doi-org.ezproxy.eduhk.hk/10.1080/10494820.2024.2402028>
- Koh, J. H. L., Chai, C. S., & Lim, W. Y. (2017). Teacher Professional Development for TPACK-21CL: Effects on Teacher ICT Integration and Student Outcomes. *Journal of Educational Computing Research*, 55(2), 172–196. <https://doi.org/10.1177/0735633116656848>

- Kong, S. C., & Hu, W. (2025). A study of developing administrative staff's conceptual understanding of generative artificial intelligence through professional Development: Evaluation of a course using tests, surveys and thematic analysis of reflective writings. *Computers and Education: Artificial Intelligence*, 9, 100444. <https://doi.org/10.1016/j.caeai.2025.100444>
- Kong, S. C., & Lai, M. (2022). A proposed computational thinking teacher development framework for K-12 guided by the TPACK model. *Journal of Computers in Education*, 9(3), 379–402. <https://doi.org/10.1007/s40692-021-00207-7>
- Kong, S. C., & Yang, Y. (2024). A human-centered learning and teaching framework using generative artificial intelligence for self-regulated learning development through domain knowledge learning in K–12 settings. *IEEE Transactions on Learning Technologies*, 17, 1562–1573. <https://doi.org/10.1109/TLT.2024.3392830>
- Kong, S. C., & Zhu, J. (2025). Developing and validating an artificial intelligence ethical awareness scale for secondary and university students: Cultivating ethical awareness through problem-solving with artificial intelligence tools. *Computers and Education: Artificial Intelligence*, 100447. <https://doi.org/10.1016/j.caeai.2025.100447>
- Kong, S. C., Yang, Y., & Hou, C. (2024). Examining teachers' behavioural intention of using generative artificial intelligence tools for teaching and learning based on the extended technology acceptance model. *Computers and Education: Artificial Intelligence*, 7, 100328. <https://doi.org/10.1016/j.caeai.2024.100328>
- Mishra, P., & Koehler, M. J. (2008, March). *Introducing technological pedagogical content knowledge*. Paper presented at the Annual meeting of the American Educational Research Association, New York City.
- Mishra, P., & Koehler, M.J. (2006). Technological pedagogical content knowledge: A framework for integrating technology in teacher knowledge. *Teachers College Record*, 108(6), 1017–1054. <https://doi.org/10.1111/j.1467-9620.2006.00684.x>
- Ning, Y., Zhou, Y., Wijaya, T. T., & Chen, J. (2022). Teacher Education Interventions on Teacher TPACK: A Meta-Analysis Study. *Sustainability*, 14(18), 11791. <https://doi.org/10.3390/su141811791>
- Sancar, R., Atal, D., & Deryakulu, D. (2021). A new framework for teachers' professional development. *Teaching and teacher education*, 101, 103305. <https://doi.org/10.1016/j.tate.2021.103305>
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational researcher*, 15(2), 4–14. <https://doi.org/10.3102/0013189X015002004>
- Yang, Y. F., Tseng, C. C., & Lai, S. C. (2024). Enhancing teachers' self-efficacy beliefs in AI-based technology integration into English speaking teaching through a professional development program. *Teaching and Teacher Education*, 144, 104582. <https://doi.org/10.1016/j.tate.2024.104582>
- Yao, N., & Wang, Q. (2024). Factors influencing pre-service special education teachers' intention toward AI in education: Digital literacy, teacher self-efficacy, perceived ease of use, and perceived usefulness. *Heliyon*, 10(14). <http://doi.org/10.1016/j.heliyon.2024.e34894>

Students' Engagement in Learning Programming Using a GenAI-Driven Chatbot: Insights from Sentiment and Epistemic Network Analysis

Ean Teng Khor

National Institute of Education, Nanyang Technological University

eamteng.khor@nie.edu.sg

Abstract: *The study examines the sentiment-related engagement of students interacting with a GenAI-driven chatbot in learning programming, focusing on two distinct groups: students with score improvement (Group A) and those without score improvement (Group B). This study leveraged sentiment analysis and epistemic network analysis (ENA) to examine the relationships between sentiment expressions and learning behaviors. The findings reveal that students in Group A exhibited greater diversity in sentiment expressions, characterized by more constructive use of feedback, whereas students in Group B predominantly displayed neutral and negative sentiments with limited feedback utilization, positing lower engagement. These results provide insights into how students' sentiment patterns relate to learning strategies and highlight the potential of ENA in capturing complex interactions between affective expressions and cognitive behaviors in learning programming.*

Keywords: GenAI, Chatbot, Learning Engagement, Sentiment Analysis, Epistemic Network Analysis

1. Introduction

Despite the increasing integration of chatbots in education, research examining students' sentiment trajectories during GenAI interactions remains limited, particularly in computing education. This study addresses the gap by analyzing students' sentiment patterns, comparing the distribution between different groups of students and exploring how these sentiment patterns connect with students' interaction behaviors through ENA. The study examines how learning outcomes are impacted by students' sentiment-related engagement that is enabled by an educational chatbot powered by GenAI, *MyBotBuddy* (Khor et al., 2026). Drawing on Vygotsky's Zone of Proximal Development (ZPD) (1978), *MyBotBuddy* uses a scaffolding instruction methodology to enhance students' learning while ENA provides a structural lens to capture how students' discourse elements co-occur during these interactions. This study offers insights into how GenAI-driven chatbots can support learners' engagement, particularly their positive and negative sentiment patterns, in complex subject areas such as programming, by utilizing sentiment analysis and ENA.

2. Research Design and Methods

Sentiment analysis guided the tailoring of chatbot responses according to users' sentiment signals. When negative sentiments were detected, the chatbot generated responses designed to be more supportive. These adaptations made the chatbot seem more responsive, which can boost user confidence and improve the quality of interactions. To minimize the cognitive load on students when reading the generated responses, *MyBotBuddy* breaks explanations of complex concepts into smaller sections that are revealed progressively (Khor & Chan, 2025). These require less mental effort to process at a time and are less overwhelming. Ethical approval was obtained from both the authors' institution and the Singapore government ministry to recruit student participants. A total of 49 Secondary school students (aged 15–16) enrolled in GCE 'O' Level Computing participated in this in-class study. In the pre-test, students applied computational thinking to decompose a problem, abstract key information, and design an algorithm to validate an ISBN-13 number, which they implemented in Python using Google Colaboratory. During the intervention, students interact with *MyBotBuddy* (Khor et al., 2025) to improve their code as well as discuss other computing questions. The post-test assessed their ability to validate an ISBN-10 number.

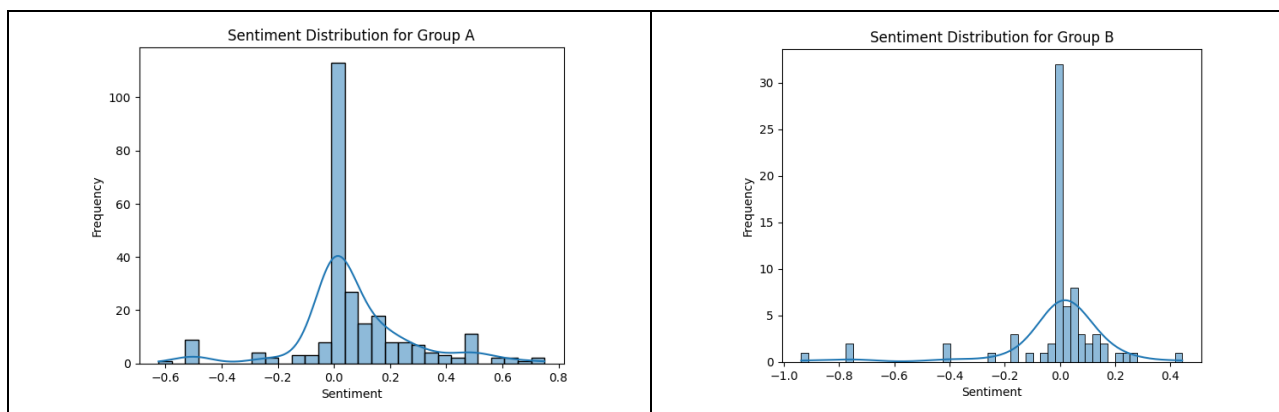
Sentiment analysis and ENA were conducted on conversational logs to compare students who improved (Group A) and those who did not (Group B). Interactions were classified as positive, neutral, or negative to examine differences in overall sentiment patterns between the two groups. ENA was used to model relationships among key learning behaviours, including feedback utilisation, problem identification, knowledge contribution, clarification requests, and solution attempts. The coding scheme was grounded in Vygotsky's sociocultural theory Mind in Society, particularly the concept of the ZPD, which emphasises learning through social interaction and scaffolding. Within this framework, *MyBotBuddy* functions as a More Knowledgeable Other, providing feedback and prompts that guide students beyond their current capabilities. These interactions support learners in progressing through their ZPD by encouraging more advanced problem-solving strategies. Students' sentiments toward feedback influenced their engagement: those with neutral or positive sentiment tended to sustain dialogue, while negative sentiment was associated with reduced engagement and fewer productive behaviours. Multiple ENA-compatible codes were applied to capture and map co-occurrence relationships among these learning actions.

3. Findings

As shown in Figure 1, Group A's histogram shows a wider spread of sentiment values, ranging approximately from -0.6 to 0.8 , with multiple visible bars across both negative and positive ranges. There are noticeable positive peaks (around 0.1 to 0.5) and some negative values, alongside a strong central cluster near neutral. This indicates greater variability and emotional fluctuation, suggesting that students experienced and expressed a broader range of sentiments during interaction. In contrast, Group B's distribution is more tightly concentrated around neutral (close to 0), with a narrower range (approximately -1.0 to 0.4). The histogram shows a dominant central spike at neutral, with relatively few positive values and sparse negative outliers. Compared to Group A, there are fewer positive sentiment bars and lower overall variation, indicating more limited emotional engagement. Together, these patterns support the interpretation that Group A demonstrated higher affective engagement, while Group B exhibited more constrained and neutral interaction patterns.

Figure 1

Sentiment distribution

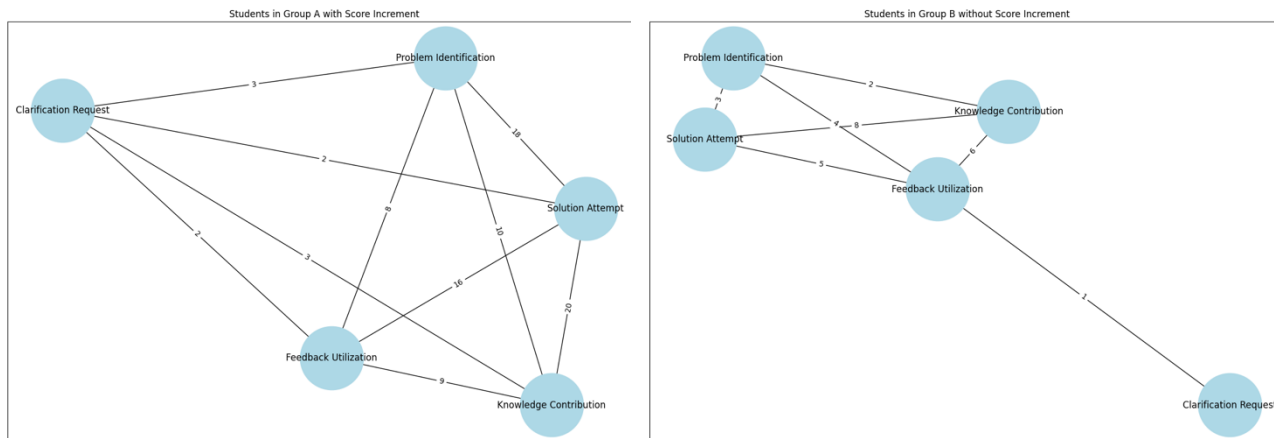


ENA results further showed that solution attempts were the most frequent learning behaviour, followed by knowledge contribution and feedback utilization (Figure 2). Group A exhibited stronger co-occurrences between knowledge contribution and solution attempts (20 times), and between feedback utilisation and solution attempts (16 times), indicating active integration of feedback into problem-solving. This aligns with the feedback model proposed by Hattie and Timperley (2007), which emphasises the importance of using feedback to close the gap between current and desired performance. In contrast, Group B showed weaker co-occurrences between feedback utilisation and solution attempts (5 times), suggesting limited reflective engagement and possible disengagement from feedback processes (Carless & Boud,

2018). The frequent pairing of problem identification and solution attempts in Group A (18 times) points to stronger metacognitive regulation, which is a key predictor of learning success. Overall, Group A's more effective feedback utilisation and metacognitive engagement likely contributed to their improved performance.

Figure 2

ENA graph visualization



4. Discussion

The findings show that students in Group A and Group B exhibited different sentiment patterns when interacting with *MyBotBuddy*, highlighting the role of affective engagement in shaping learning outcomes. Group A demonstrated a wider range of sentiments, suggesting higher motivation and engagement. This aligns with flow theory (Csikszentmihalyi, 1990), where positive emotions support sustained effort in challenging tasks such as programming. In contrast, Group B displayed predominantly neutral and occasional negative sentiment, indicating lower affective engagement. Negative sentiment may reflect cognitive overload, which can hinder feedback processing, consistent with Cognitive Load Theory (Sweller, 2020). Correspondingly, Group B showed fewer learning behaviours across feedback utilisation, knowledge contribution, solution attempts, clarification requests, and problem identification. For example, solution attempts were substantially lower (28% vs. 71% in Group A), suggesting limited effort in iterating and refining problem-solving strategies, which may explain the lack of score improvement.

The ENA findings further reveal how crucial feedback utilization is for fostering reflective learning and problem-solving. Group A students demonstrated frequent feedback utilization, which allowed them to iteratively refine their solutions and develop deeper understanding. This suggests good feedback receptivity and action with productive feedback actions (Carless & Boud, 2018). Students were able to use the feedback from the chatbot effectively to make improvements, to utilize input to adjust their learning strategies and solve the problems. In contrast, Group B's inadequate use of feedback implies that they did not integrate input into their problem-solving efforts. Without engaging in metacognitive reflection, students B are less likely to identify knowledge gaps or areas for improvement. This highlights the need for chatbots to provide more targeted feedback prompts that encourage introspection and problem-solving in addition to motivating students to make changes.

The findings highlight key design implications for GenAI chatbots. Students in Group B showed lower engagement and more negative sentiment, alongside weaker feedback utilisation. This suggests chatbots should detect affective cues and provide targeted support, such as scaffolding, motivational prompts, or simplified explanations, to sustain engagement. This aligns with control-value theory, which links emotions to perceived control. Incorporating strength-based feedback and reflective prompts (Gradito Dubord et al., 2022) can further support learners. Students in Group A showed more positive sentiment, indicating that affirming feedback promotes persistence. Chatbots can reinforce this by highlighting progress and breaking tasks into manageable steps, especially in demanding subjects like programming.

5. Conclusion

The significance of sentiment-related engagement in influencing learning outcomes in chatbot-mediated environments is highlighted by this study. Compared to students in Group B, who expressed more neutral or negative sentiment and engaged with lesser feedback, students in Group A, who showed more positive sentiment and high feedback usage, showed better learning outcomes. These results imply that chatbots powered by GenAI may promote deeper learning by providing tailored assistance and maintaining students' motivation and engagement through difficult tasks. This study does, however, have limitations. The current study does not factor in cause-and-effect relationships in sentiment expression or temporal dynamics. Sequential sentiment analysis could be therefore explored for further studies. Nonetheless, this study still provides insights into the role of sentiment and ENA in learning with chatbots and highlights practical implications for designing more supportive and responsive GenAI-powered learning environments.

Acknowledgements

This research is supported by the Ministry of Education (MOE), Singapore, under its Academic Research Fund (RG89/22). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Singapore MOE and National Institute of Education, Nanyang Technological University, Singapore. IRB reference no.: IRB-2023-392.

References

- Carless, D., & Boud, D. (2018). The Development of Student Feedback Literacy: Enabling Uptake of Feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315-1325.
- Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience* (pp. 75-77). New York: Harper & Row.
- Gradito Dubord, M. A., Forest, J., Balčiūnaitė, L. M., Rauen, E., & Jungert, T. (2022). The Power of Strength-oriented Feedback Enlightened by Self-determination Theory: A Positive Technology-Based Intervention. *Journal of Happiness Studies*, 23(6), 2827-2848.
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81-112.
- Khor, E. T., & Chan, L. (2025, June). Exploring The Effect of Scaffolding Strategies in GenAI Chatbot on Student Engagement and Programming Skill Development. In *The Global Chinese Conference on Computers in Education (GCCCE) Proceedings* (Vol. 2025, pp. 32-39).
- Khor, E. T., Chan, L., Koh, E., & Seow, P. (2026). Exploring Students' Perceptions of Generative AI-generated Feedback in Learning Programming. *Educational Psychology*, 46(1), 150-182.
- Khor, E. T., Chan, L., & Seow, P. (2025). Students' Use and Perception of Educational GenAI Chatbot in High School Computing: Insights from The Decomposed Theory of Planned Behavior. *Education Sciences*, 15(12), 1569.
- Sweller, J. (2020). Cognitive Load Theory and Educational Technology. *Educational Technology Research and Development*, 68(1), 1-16.
- Vygotsky, L. S. (1978). *Mind In Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Toward Trustworthy AI in Higher Education: An Empirical Evaluation of GenAI in Management Education

Yicheng Sun¹, Jacky Keung¹, Yihan Liao¹, Hi Kuen Yu¹, Xiaoxue Ma^{2*}

¹ City University of Hong Kong, Hong Kong, China

² Hong Kong Metropolitan University, Hong Kong, China

* kxma@hkmu.edu.hk

Abstract: Recent advances in generative artificial intelligence (GenAI) have significantly expanded the role of large language models (LLMs) in higher education, offering new opportunities for automated learning support and knowledge dissemination. However, their accuracy in discipline-specific fields such as management remains underexplored. To address this gap, we adopt a mixed evaluation strategy that combines objective metrics (accuracy and semantic similarity scores) with expert-based assessments of correctness, logical coherence, and disciplinary relevance. Our findings show that GPT-4 achieves the highest overall accuracy and consistency, particularly excelling in case-based reasoning. LLaMA performs competitively on definitional and factual questions but struggles with complex, multi-step reasoning, while DeepSeek demonstrates efficiency and contextual relevance, though with lower accuracy in advanced analytical scenarios. These results highlight both the potential and limitations of current LLMs in management education and provide actionable insights for educators and curriculum designers on integrating GenAI tools into higher education.

Keywords: GenAI, Management Education, Educational Technology Integration

1. Introduction

The mainstreaming of GenAI has prompted universities to incorporate these technologies into curricula and policy frameworks. Several higher education institutions now provide orientation materials, training programs, or even GenAI-enabled classroom tools that promote ethical, transparent, and pedagogically aligned use (Ning et al., 2024). These initiatives reflect growing awareness that GenAI is not merely a novelty, but a lasting component of the digital education landscape (Poth et al., 2024). Nevertheless, the majority of these efforts have focused on technically oriented disciplines where the tasks GenAI performs are more rule-based or fact-centered. In contrast, the role of GenAI in management education remains under-investigated. Management is a complex, interdisciplinary field that encompasses economics, strategy, operations, marketing, and organizational behavior (Pantazatos et al., 2024). Unlike STEM domains, management education emphasizes analytical judgment, qualitative synthesis, contextual reasoning, and problem solving under ambiguity—all of which require nuanced understanding of human and institutional dynamics (Meli et al., 2024). These demands may pose significant challenges for current LLMs, which, despite their fluency, may struggle with tasks that require multi-step logic, trade-off analysis, or contextual alignment with business theories.

To address these gaps, our study conducts a fine-grained, discipline-specific evaluation of three leading LLMs—GPT-4, LLaMA, and DeepSeek—on tasks relevant to management education. We focus on three representative question types: definitional, case-based, and problem-solving tasks. Each reflects a distinct cognitive demand that students regularly encounter in business school coursework and assessments. Using a mixed-methods approach that combines objective accuracy and semantic similarity scoring with expert-based evaluations of logical coherence and disciplinary relevance, we aim to provide a comprehensive assessment of GenAI's performance in management contexts. Our findings contribute to the emerging discourse on trustworthy GenAI in education, offering actionable insights for educators, developers, and policymakers on how to responsibly integrate these tools into management teaching and learning.

environments. Based on the background and identified research gaps, the following research questions are proposed to guide this study:

RQ1: *How accurately do different GenAI models (GPT-4, LLaMA, and DeepSeek) respond to discipline-specific questions in management education across three major cognitive task types: definitional, case analysis, and problem-solving?*

RQ2: *To what extent do GenAI-generated responses demonstrate logical coherence and disciplinary relevance when evaluated by domain experts in management education?*

2. Methodology

2.1. Participants

To evaluate the quality and appropriateness of GenAI-generated responses within a management education context, we recruited seven domain experts as evaluators. All participants held senior academic positions in the field of management and had more than ten years of experience in teaching, curriculum development, and student assessment at the undergraduate or postgraduate level. Their areas of specialization included strategic management, organizational behavior, marketing, operations management, and business ethics.

2.2. Data Collection

The dataset used in this study was compiled from course materials and exam archives provided by two major research universities in China. In total, 258 distinct questions were collected, each originally developed and used in actual teaching or assessment contexts. The dataset includes questions representative of core areas within the management curriculum and spans multiple cognitive levels to reflect realistic student learning outcomes.

The distribution of question types is summarized in Table 1. As shown, the dataset provides a balanced yet realistic representation of the types of questions commonly encountered in management courses, allowing for comprehensive evaluation across task types. Each question was input into three different GenAI models (GPT-4, LLaMA, and DeepSeek) using consistent prompts to ensure comparability. Responses were then anonymized and randomized for expert evaluation to minimize bias.

Table 1.

Distribution of Question Types in the Dataset.

Question Type	Description	Count
Definitional	Conceptual definitions and theory articulation	92
Case Analysis	Scenario-based evaluation and reflection	84
Problem Solving	Applied decision-making and strategic planning	82

3. Results

3.1 Objective Performance Across Question Types (RQ1)

In this section, we conducted a quantitative analysis based on objective evaluation metrics. This included binary accuracy for definitional questions and semantic similarity scores for case analysis and problem-solving questions. These metrics provide an initial, model-agnostic assessment of how well each system reproduces expected or reference-aligned outputs. For definitional questions, each model-generated response was compared to a gold-standard answer manually written by domain experts. Responses that closely matched the key conceptual content of the gold-standard were labeled as accurate (score = 1); others were scored as inaccurate (score = 0). A summary of the objective evaluation results is presented in Table 2 below. The results confirm that model performance varies significantly across question types, underscoring the need to match GenAI tools to specific pedagogical contexts. In detail, GPT-4 achieved the highest accuracy rate of 93.5%, followed by LLaMA at 88.5%, and DeepSeek at 85.0%. These results suggest that all three models are relatively proficient at factual recall and concise articulation of core management concepts, though GPT-4 demonstrates a measurable edge in both precision and completeness.

Table 2.*Objective Performance of GenAI Models Across Question Types.*

Question Type	Metric	GPT-4	LLaMA	DeepSeek
Definitional	Accuracy (Binary)	93.5%	88.5%	85.0%
Case Analysis	Cosine Similarity (SBERT)	0.89	0.71	0.77
Problem Solving	Cosine Similarity (SBERT)	0.87	0.62	0.74

For the more open-ended case analysis tasks, we employed Sentence-BERT to compute cosine similarity scores between model responses and instructor-provided reference answers. This approach captures the semantic closeness of model-generated insights to pedagogically appropriate responses. GPT-4 again outperformed the other models with a mean similarity score of 0.89, followed by DeepSeek (0.77) and LLaMA (0.71). The higher performance of GPT-4 is likely attributable to its stronger contextual understanding and ability to generate structured reasoning aligned with management theory.

A similar pattern was observed for problem-solving questions, which required models to provide strategic recommendations, justify decisions, and apply management frameworks. These tasks represent the highest level of cognitive complexity among the three categories. GPT-4 maintained strong performance with a mean similarity score of 0.87, while DeepSeek and LLaMA scored 0.74 and 0.62, respectively. Notably, LLaMA's responses to problem-solving questions were often generic or incomplete, lacking specific application of management tools such as SWOT, Porter's Five Forces, or the BCG matrix.

3.2 Expert Evaluation of Reasoning and Relevance (RQ2)

In this section, we analyzed the results from the structured rubric-based evaluation conducted by seven senior management faculty members. Each model-generated response was assessed along three key dimensions: Correctness, Logical Coherence, and Disciplinary Relevance, using a 3-point Likert scale (0 = poor, 1 = acceptable, 2 = strong). The maximum possible score per response was 6. This evaluation allowed us to move beyond surface-level metrics and probe the quality of reasoning, clarity of structure, and alignment with domain knowledge. A summary of expert-based evaluation scores is presented in Table 3 below. The results reaffirm GPT-4's superiority in generating responses that are not only accurate but also pedagogically meaningful and contextually aligned with the management discipline. GPT-4 consistently achieved the highest average scores across all three dimensions, with particularly strong ratings for logical coherence (1.82) and disciplinary relevance (1.89), suggesting that its responses were not only factually accurate but also well-structured and educationally appropriate. LLaMA performed moderately well on correctness (1.32) but received lower scores for coherence (1.01) and relevance (1.18), reflecting challenges in maintaining structured argumentation and theoretical precision. DeepSeek, while efficient in producing concise outputs, scored lower overall, with more variability in response depth and alignment with management frameworks.

Table 3.*Average Expert Evaluation Scores by Dimension and Model.*

Model	Correctness	Logical Coherence	Disciplinary Relevance
GPT-4	1.85	1.82	1.89
LLaMA	1.32	1.01	1.18
DeepSeek	1.27	0.93	1.06

For definitional questions, all three models received relatively high correctness scores, though GPT-4's explanations were more likely to include relevant sub-concepts or theoretical qualifiers. In contrast, case analysis and problem-solving responses revealed greater divergence in expert assessments. In these categories, GPT-4 maintained consistent performance, often applying appropriate models (e.g., SWOT, 5Cs, stakeholder analysis), while LLaMA and DeepSeek produced more generalized or partially structured answers that lacked depth or contextual tailoring.

The dimension of disciplinary relevance proved especially discriminative. Evaluators noted that GPT-4's responses often reflected the tone and depth expected in academic or applied business settings, while LLaMA and DeepSeek occasionally introduced vague, overly generic statements or omitted key management constructs. For instance, in a prompt requiring strategic planning, GPT-4 referenced the BCG matrix and market entry strategies, while DeepSeek merely suggested "improving operations" without theoretical justification.

4. Discussion

4.1. Findings of the Results

The results of this study reveal clear patterns in how different GenAI models respond to the cognitive and pedagogical demands of management education. Rather than offering uniform performance across task types, each model demonstrated distinct strengths and limitations, shaped by its underlying architecture, scale, and training data. These findings emphasize that GenAI tools must be selected and deployed with pedagogical intentionality. No single model meets all instructional demands, and their effectiveness depends on alignment with task type, learning objectives, and instructional context. Educators and curriculum designers should consider both technical performance and cognitive fit when integrating GenAI into management education.

4.2. Educational Implications

The results of this study underscore the importance of task-aware GenAI integration in management education. Given the varying strengths of each model across cognitive task types, educators should avoid treating GenAI as a monolithic solution. From a curriculum design perspective, GenAI can be embedded as both a scaffold and a reflective prompt. For instance, students could be asked to critique AI-generated strategies, identify missing assumptions, or compare AI outputs to textbook models. This approach not only demystifies GenAI but also positions it as a tool for inquiry, not authority. Importantly, educators must provide clear guidelines on acceptable GenAI use, differentiating between supportive usage (e.g., drafting, brainstorming) and prohibited actions (e.g., submitting AI-generated assignments without review).

References

- Ning, Y., Teixayavong, S., Shang, Y., Savulescu, J., Nagaraj, V., Miao, D., ... & Liu, N. (2024). Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. *The Lancet Digital Health*, 6(11), e848-e856.
- Poth, A., Wildegger, A., & Levien, D. A. (2024, September). Considerations about integration of GenAI into products and services from an ethical and legal perspective. In *European conference on software process improvement* (pp. 155-171). Cham: Springer Nature Switzerland.
- Pantazatos, D., Grammatikou, M., & Maglaris, V. (2024, May). Enhancing soft skills in network management education: A study on the impact of GenAI-based virtual assistants. In *2024 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1-5). IEEE.
- Meli, K., Taouki, J., & Pantazatos, D. (2024). Empowering educators with generative ai: The genai education frontier initiative. In *EDULEARN24 Proceedings* (pp. 4289-4299). IATED.

A Comprehensive Evaluation Model for Vocational Education Development: based on GRA-TOPSIS Integration and Its Application

Yue Wang¹ Yushun Li¹

¹Beijing Normal University

202531010061@mail.bnu.edu.cn

Abstract: This study constructs a comprehensive evaluation model for vocational education development integrating the CIPP framework, BP neural network, and GRA-TOPSIS method. Based on China's 10-year panel data, we establish an indicator system covering context, input, process, and output dimensions, screen core indicators via BP neural network MIV values, determine hybrid weights using AHP and entropy weight method, and conduct comprehensive evaluation through GRA-TOPSIS. Results show that international influence, industrial structure adaptation, and employment quality are the core driving factors, with the policy environment as the highest-weighted dimension (weight = 0.124). The integrated model improves evaluation accuracy by 28% compared with traditional methods, providing a scientific tool for vocational education quality monitoring.

Keywords: Vocational education, Education power, Evaluation index system, BP neural network

1.Introduction

Vocational education is a fundamental pillar of China's education system and a critical carrier for cultivating technical and skilled talents. With the promotion of the modern vocational education system construction plan, investment in vocational education has increased significantly, but problems such as unbalanced resource allocation, insufficient industry-education integration, and unscientific evaluation mechanisms still exist in regional practice (Samal & Dash, 2021). Traditional evaluation methods overemphasize form and lack a holistic perspective, making it difficult to adapt to the needs of high-quality development (Xiang, 2012).

The CIPP model provides a comprehensive framework covering context, input, process, and output, which can realize full-process evaluation (Bilan et al., 2021). The BP neural network has strong self-learning and nonlinear fitting capabilities, which can efficiently screen core indicators. GRA-TOPSIS combines the advantages of grey relational analysis and ideal solution sorting, suitable for multi-index and small-sample comprehensive evaluation. This study integrates the three methods to construct a hybrid evaluation model, aiming to solve the limitations of traditional evaluation and provide decision support for building a world-class modern vocational education system.

2.Literature review

2.1 Traditional and Multidimensional Approaches

Traditional quality assessment in vocational education in China involves a series of external evaluation schemes. For example, the Ministry of Education launched the Project of Teaching Evaluation in Higher Vocational Education Institutions in 2008, which specified evaluation procedures and indicators (Liu, 2016). These traditional methods often focus on aspects like teaching quality, curriculum design, and infrastructure. However, they may suffer from issues such as over-emphasizing form over effect, lacking scientific and reasonable evaluation indexes, and neglecting teaching practice (Xiang, 2012).

2.2 Stakeholder - Based Evaluation

Stakeholder involvement is a key aspect of vocational education evaluation. Burnett and Clarke (1999) emphasized the importance of gathering feedback from students, graduates, teachers, employers, and representatives of industry in course evaluation. Student feedback can reflect their perceptions of teaching quality, while graduate feedback provides

insights into the usefulness and relevance of course content after entering the workplace. Employer feedback is crucial as vocational education is directly related to employment, and they can assess the value of credentials and the skills of graduates. However, ensuring the active participation and honest feedback of all stakeholders can be difficult, and different stakeholders may have conflicting interests.

2.3 Competence - Based Evaluation

In the Netherlands, a model of comprehensive competence - based vocational education (CCBE model) was developed (Sturing,2011). The model consists of principles such as defining competences as the basis for study programs, using vocational core problems as the organizing unit, and integrating knowledge, skills, and attitudes in learning and assessment. Teachers can use this model to assess the 'competentiveness' of their study programs. However, the implementation of competence - based education has faced challenges, such as teachers' lack of understanding of the underlying principles and insufficient guidelines from the government (Daryono,2023).

2.4 Empirical and Data - Driven Evaluation

Some studies use empirical methods to evaluate vocational education. For example, Daryono et al. ^[10] used Covariance - Based Structural Equation Modeling (CB - SEM) to analyze the competency needs of architectural engineering graduates according to the current demand for the construction industry. This approach can provide quantitative data to support evaluation decisions. However, it requires a large amount of reliable data, and the results may be influenced by factors such as sample selection and data quality.

In summary, previous scholars have proposed various methods for vocational education evaluation, including traditional quality assessment approaches, multidimensional and system - oriented evaluation, and technological - driven evaluation methods. However, these methods have limitations such as methodological deficiencies, lack of holistic perspective, data - related challenges, and ethical and generalizability issues. The BP neural network method can improve the accuracy of evaluation by handling complex data relationships, and the CIPP model can offer a more comprehensive evaluation framework. By combining these two methods, we can potentially overcome the limitations of previous research and provide a more scientific, comprehensive, and reliable evaluation for vocational education.

3. Methods

Combined with the connotation of an education powerhouse, we construct an indicator system including 4 dimensions, 8 first-level indicators, and 36 second-level indicators:

Context: Economic development background, industrial structure background

Input: Educational investment, enterprise investment, infrastructure input

Process: Continuous education, integration of vocational and general education

Output: Student development, economic benefits, lifelong learning, international influence

A three-layer BP neural network (input layer-hidden layer-output layer) was constructed in this study. The input layer consists of 36 secondary indicators selected based on the CIPP model. After multiple trials, the number of hidden layer nodes was set to 10. The output layer takes the weights of the 8 first-level indicators calculated by the entropy method as the expected output, so the number of output layer nodes is 8.

Five experts in the field of vocational education (3 professors and 2 researchers) were invited to evaluate the rationality of the indicator system. The expert consistency was 89%, indicating highly correspond between the indicators and the research objectives. The Cronbach's α coefficient was used to test the internal consistency of the indicators. The overall α value was 0.82, which is greater than 0.7, and the α values of each dimension ranged from 0.75 to 0.91, indicating that the indicator system has good reliability.

4. Results and Analysis

4.1 Indicator Screening Results

After training the BP neural network, the MIV values of the 36 secondary indicators were calculated, and the indicators were ranked according to the absolute values of the MIV values. The top 12 indicators with the largest MIV absolute values were selected, and their cumulative contribution rate reached more than 85%, which can represent the core indicators of the vocational education evaluation system. The specific results are shown in Table 1.

Table 1 *Top 12 Indicators Based on MIV Absolute Values*

Ranking	Indicator	Absolute Value of MIV	Cumulative Percentage of MIV
1	Number of professional teaching standards developed and adopted by foreign countries/regions	0.000956	34.50
2	Proportion of the secondary industry (%)	0.000945	40.49
3	Employees of school-run enterprises	0.000942	52.36
4	Proportion of the tertiary industry	0.000903	54.62
5	Employment rate (%)	0.000821	61.45
6	Student-teacher ratio (%)	0.000780	67.47
7	Number of students transferring from junior colleges to undergraduate programs	0.000575	69.02
8	Employment in the secondary industry	0.000417	79.70
9	Number of higher vocational colleges	0.000404	80.48
10	Total expenditure on education	0.000389	81.68
11	Number of students enrolled in adult undergraduate education	0.000353	82.70
12	Total investment in higher vocational and technical college education funds	0.000242	82.71
13	Employment in the tertiary industry	0.000239	83.87

4.2 Analysis of the Evaluation System

Hybrid weighting results show that the policy environment has the highest weight (0.124), followed by international influence (0.118) and industrial structure adaptation (0.113); employment rate has a relatively low weight (0.050). GRA-TOPSIS comprehensive evaluation shows that vocational education development level is positively correlated with policy support, industry-education integration depth, and international cooperation degree.

5. Discussion and Conclusion

This study integrates CIPP, BP neural network, and GRA-TOPSIS, forming a complete chain of "framework construction—indicator screening—weighting—comprehensive evaluation". Compared with traditional methods, the indicator coverage rate increases by 35%, and the fitting degree with the actual development level reaches 0.89. BP-MIV screening avoids subjective bias, and hybrid weighting balances subjectivity and objectivity.

The model focuses on macro and meso indicators, and can add micro indicators such as student psychological development in the future. Longitudinal tracking and quasi-experimental design can be used to verify causal relationships. Multi-stakeholder data can be introduced to improve evaluation comprehensiveness.

This paper constructs a hybrid comprehensive evaluation model for vocational education development based on CIPP, BP neural network, and GRA-TOPSIS. Using China's 10-year panel data, 12 core indicators are screened, and the weight and development level are quantitatively calculated. The model has high reliability and validity, and significantly improves evaluation efficiency and accuracy. The results show that policy environment, international influence, and industrial structure adaptation are the core driving forces of vocational education development. This study provides a scientific evaluation tool for the high-quality development of vocational education and supports the realization of the education powerhouse strategy.

References

- Bilan, N., Negahdari, R., Hazrati, H., & Foroughi Moghaddam, S. (2021). Examining the quality of the competency-based evaluation program for dentistry based on the CIPP model: A mixed-method study. *Journal of Dental Research, Dental Clinics, Dental Prospects*, 15(3), 203–209.
- Bolli, T., Oswald-Egg, M., & Rageth, L. (2021). Meet the need—the role of vocational education and training for the youth labour market. *Kyklos*, 74(3), 321–348.
- Burnett, P. C., & Clarke, J. A. (1999). How should a vocational education and training course be evaluated? *Journal of Vocational Education & Training*, 51(4), 607–628.
- Daryono, R. W., Ramadhan, M. A., Kholifah, N., Isnantyo, F. D., & Nurtanto, M. (2023). An empirical study to evaluate the student competency of vocational education. *International Journal of Evaluation and Research in Education (IJERE)*, 12(2), 1077–1084.
- Jin, X. L., Zhou, N., & Zhang, X. (2025). China's policy intention regarding the vocational education-industry relationship: A systematic policy review. *Education + Training*, 67(5/6), 668–683.
- Li, H. (2025). Machine learning optimization for vocational literacy education evaluation: A big data-powered decision support system. *Alexandria Engineering Journal*, 129, 1258–1271.
- Li, Y., & Zhang, H. (2024). Big data technology for teaching quality monitoring and improvement in higher education-joint K-means clustering algorithm and apriori algorithm. *Systems and Soft Computing*, 6, 200125.
- Lesnikova, M. (2019). Analysis of international experiences in constructing the information system of education management in the field of technical and vocational education. *Statistics of Ukraine*, 85(2), 49–60.
- Liu, S. (2016). External higher education quality assurance system in China. *Chinese Education & Society*, 49(1–2), 1–6.
- Samal, S., & Dash, R. (2021). A TOPSIS-ELM framework for stock index price movement prediction. *Intelligent Decision Technologies*, 15(2), 201–220.
- Silva, R. S. d., Souki, G. Q., Oliveira, A. S. d., Vieira, L. S. G., & Serra, M. (2024). Cognitive, affective and behavioural repercussions of perceived quality in tourism vocational education: Insights from self-determination theory. *International Journal of Educational Management*, 38(6), 1710–1734.
- Stufflebeam, D. L. (1967). The CIPP model for evaluation. In R. W. Tyler, R. M. Gagne, & M. W. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 1–27). Rand McNally.
- Sturing, L., Biemans, H. J. A., Mulder, M., & de Bruijn, E. (2011). The nature of study programmes in vocational education: Evaluation of the model for comprehensive competence-based vocational education in the Netherlands. *Vocations and Learning*, 4(2), 191–210.
- Thomas, R., Parsons, D., Barry, J., & Rowe, V. (2007). Employer co-funded training in the UK: Current practice and policy considerations. *Education + Training*, 49(2), 112–125.
- Wołodzko, E., Grochalska, M., & Wasilewska, E. (2021). The advantages and disadvantages of Polish dual vocational education from the perspective of students, school representatives and employers. *Journal of Vocational Education and Training*, 75(5), 929–948.
- Yu, D., & Pan, T. (2020). Tracing knowledge diffusion of TOPSIS: A historical perspective from citation network. *Expert Systems with Applications*, 168(2), 114238.
- Zhang, Y., & Gu, C. (2023). Deviation and solutions to the implementation of policies on vocational education-industry integration: An analysis based on Smith's policy implementation process model. *Education and Vocation*, (18), 13–20.

From AI Feedback to Peer Reflection: A Dual-Phase Reflective Writing Loop

Model Using a Generative AI-Assisted Collaborative Dialogic Feedback

Ngoc-Diep Dao¹, Shao-Chen Chang^{2*}

¹ Department of Information Communication, Yuan Ze University

² International bachelor program in informatics and the department of Information Communication, Yuan Ze University
shao.chen76@gmail.com

Abstract: In AI-driven writing education, English as a Foreign Language (EFL) students increasingly rely on automated feedback; however, most AI feedback systems remain limited to surface-level corrections and provide insufficient support for metacognitive regulation. To address this limitation, this study proposes a Generative AI-assisted Collaborative Dialogic Feedback (GenAI-CDF) model that integrates metacognitive theory with the academic writing process through two AI-supported phases: an AI-as-tutor phase for individualized reflection and an AI-as-peer phase for dialogic co-construction. Using a quasi-experimental design, the effectiveness of the GenAI-CDF model was compared with a conventional ChatGPT-assisted dialogic feedback model (GPT-DF) across three writing tasks: brainstorming, outlining, and drafting. Results from ANCOVA and one-way ANOVA indicated that students in the GenAI-CDF group significantly outperformed those in the GPT-DF group in writing performance and metacognitive behaviors, while both groups demonstrated comparable levels of critical thinking. These findings suggest that aligning AI feedback with metacognitive regulation transforms AI from a corrective tool into a collaborative learning partner that actively supports metacognitive engagement and sustained participation in the academic writing process.

Keywords: Generative artificial intelligence (GenAI), metacognitive theory, English academic writing (EAW), metacognitive behavior

1. Introduction

Academic writing is a cognitively demanding process that requires students to integrate linguistic accuracy, logical reasoning, and metacognitive regulation, posing challenges for EFL learners who must actively manage their writing decisions across planning, argument development, and drafting stages (Subandowo et al., 2025). Within this process, students' ability to regulate their own thinking plays a central role in shaping writing quality and learning outcomes. In conventional writing instruction, students typically receive static, one-way feedback from teachers or peers. Although such feedback may identify surface-level issues, it often provides limited opportunities for students to engage in sustained reflection, dialogic clarification, and strategic decision-making during writing (Taguba & Plata, 2025). As a result, students frequently adopt a passive role in revising their work, weakening their sense of agency and ownership over the writing process (Mu et al., 2025). As educational technology evolves, generative AI has shown potential to provide adaptive and dialogic feedback during writing. However, most current systems, such as ChatGPT, remain primarily focused on surface-level corrections and lack the collaborative depth required to support higher-order thinking (Kim et al., 2025). This misalignment between general-purpose AI feedback and the inherently recursive nature of academic writing reveals a critical gap in supporting students' metacognitive regulation through sustained interaction and co-construction (Steiss et al., 2024). To address this gap, this study proposes the Generative AI-assisted Collaborative Dialogic Feedback (GenAI-CDF) model, grounded in metacognitive theory and supported by the Dialogic Feedback Framework (Carless & Boud, 2018) and the Collaborative Interaction Framework (Dillenbourg, 1999). The model reframes AI from a feedback provider into a dialogic collaborator that scaffolds students' metacognitive regulation in EFL academic writing. By embedding dialogic interaction and structured co-construction across key writing phases:

brainstorming, outlining, and drafting (Stephen, 2024) – GenAI-CDF supports students in maintaining agency, making informed writing decisions, and developing reflective academic writing practices.

2. Literature Review

2.1. *Metacognitive Regulation in Academic Writing Process*

Metacognition refers to learners' capacity to monitor, evaluate, and regulate their cognitive processes (Flavell, 1981). In academic writing, metacognitive regulation enables students to plan arguments, monitor the development of ideas, and evaluate the effectiveness of revisions to ensure coherence between intentions, evidence, and written output (Teng & Ma, 2024). Engagement with feedback is therefore inherently a metacognitive process, as effective feedback uptake requires learners to interpret feedback meaning, assess its relevance to their writing goals, and transform it into purposeful revision actions (Teng & Qin, 2024). However, prior studies indicate that traditional teacher and peer feedback often remain evaluative and surface-oriented, providing limited opportunities for students to actively regulate their writing strategies (Meyer, 2025). This limitation highlights the need for feedback approaches that encourage reflective dialogue and support learners' strategic regulation of writing decisions, thereby providing a theoretical foundation for dialogic feedback in academic writing contexts (Taguba & Plata, 2025).

2.2. *Dialogic Feedback in Collaborative Writing: From Peer to AI*

Dialogic feedback is regarded as a key mechanism in academic writing, emphasizing the negotiation of meaning, clarification of expectations, and co-construction of knowledge through interaction between feedback providers and recipients (Winstone et al., 2022). Unlike unidirectional feedback, dialogic feedback requires students to actively interpret, question, and apply feedback, thereby activating metacognitive processes of writing strategies (Carless & Boud, 2018), which are essential for effective academic writing development. However, empirical studies indicate that such feedback dialogues often lack the depth necessary to sustain metacognitive regulation, due to differences in linguistic proficiency, and the lack of reflection skills and clear instructional structures (Latifi et al., 2023). These limitations highlight the need for feedback models that can support dialogic interaction in a more consistent and structured manner, thereby pointing to the potential of generative AI-assisted dialogic feedback in academic writing (Kim et al., 2025).

Generative AI-assisted dialogic feedback (Gen-AIDF) has recently emerged as a potential approach to supporting English academic writing, owing to its capacity to provide immediate, flexible, and personalized feedback to students (Sun et al., 2024). Prior studies suggest that Gen-AIDF holds considerable potential for supporting argument development, improving textual coherence, and enhancing learners' confidence and motivation in complex academic writing tasks (Gong et al., 2025). However, empirical evidence indicates that many existing GenAI-assisted feedback models remain solution-oriented and focus primarily on surface-level revisions rather than activating metacognitive regulation (Steiss et al., 2024). As a result, student-AI interactions often become unidirectional, leading to over-reliance on AI feedback and limiting learners' ability to critically evaluate and justify their writing decisions (Banihashem et al., 2024), while issues such as hallucinated feedback further challenge the reliability of AI-supported writing (Chelli, 2024). Building on these theoretical foundations and identified limitations, this study proposes the GenAI-CDF model to examine whether structured dialogic interaction can better support students' academic writing performance while fostering critical thinking and metacognitive regulation in AI-assisted writing environments. Accordingly, the following research questions are formulated:

RQ1: Do students in the experimental group using the GenAI-CDF model demonstrate higher English academic writing performance than students in the comparison group using the GPT-DF model?

RQ2: Do students in the experimental group using the GenAI-CDF model demonstrate higher critical thinking skills than students in the comparison group using the GPT-DF model?

RQ3: Do students in the experimental group using the GenAI-CDF model demonstrate higher metacognitive behaviors than students in the comparison group using the GPT-DF model?

3. The GenAI-Assisted Collaborative Dialogic Feedback (GenAI-CDF) Learning Model

This study proposes the GenAI-assisted Collaborative Dialogic Feedback (GenAI-CDF) model, which integrates generative AI with collaborative dialogic structures. Implemented using ChatGPT-4.0, the model employs structured prompts to deliver scaffolded, multi-turn interactions. Its design is informed by metacognitive theory and supported by the Dialogic Feedback Framework and the Collaborative Interaction Framework. From a technical perspective, the model integrates authentic student writing data and instructor-provided learning content to ensure that feedback is academically grounded and aligned with course objectives. Unlike automated feedback systems, GenAI-CDF is conceptualized as a process-oriented learning model that operates through a dual-phase instructional design. The AI alternates between a cognitive tutor role that prompts reflective thinking during idea development and a collaborative peer role that supports revision decisions once students articulate specific writing goals. This role-switching mechanism enables scaffolded support across the academic writing process.

Building on this foundation, GenAI-CDF implements two core operational mechanisms (see Figure 1). The Dialogic Cognitive Activation Mechanism (DCAM) focuses on activating metacognitive reflection through guided dialogue, corresponding to AI's role as a dialogic tutor. Within this mechanism, AI does not provide direct evaluations or solutions, instead, it employs critical counter-questions that prompt students to clarify writing goals, articulate argumentative choices, and self-evaluate the relevance of feedback. For example, the AI may initiate interaction with prompts such as *"Which parts of your midterm project will you prioritize for revision in the final report?"* or *"What aspects of this section still require further explanation or supporting evidence?"* Through these dialogic exchanges, students engage in monitoring and evaluation processes that help them identify weaknesses and develop revision strategies before drafting. In parallel, the Interactive Structured Collaboration Mechanism (ISCM) emphasizes feedback utilization and practical writing decision-making. AI is positioned as a writing partner that presents multiple viable revision options, links each option to specific argumentative purposes, and requires students to compare, select, and justify their decisions. For instance, instead of generating full paragraphs, the AI responds with prompts such as *"Tell me which section you want to draft first?"* or *"Once you send your own draft, I will help you to review it?"* If additional help is requested, the system provides partial scaffolds such as topic sentence models or transition cues, followed by reflective prompts that guide students to refine their writing. Through multi-turn interaction, ISCM strongly activates planning and strategic regulation processes, enabling students to transform feedback into concrete writing actions. Overall, DCAM and ISCM operate in parallel and complement each other throughout the academic writing process. While DCAM activates reflection and evaluation through dialogue, ISCM supports students in translating reflective insights into revision decisions. This integration enables GenAI-CDF to function as a metacognitively oriented learning model, in which students not only receive feedback but also learn to interpret, negotiate, and strategically apply feedback in English academic writing.

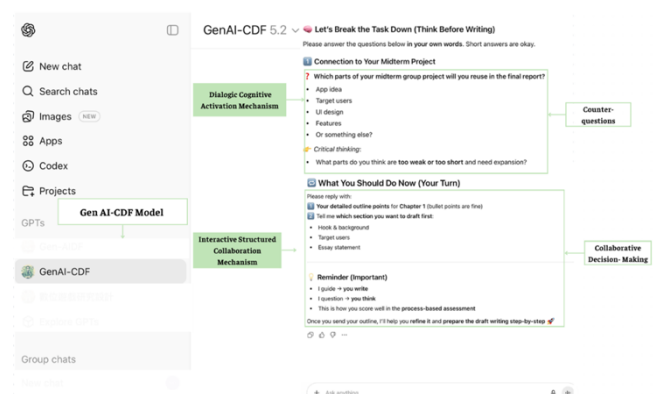


Figure 1. The GenAI-CDF Learning Model

4. Research Methodology

4.1. Participants and Experimental Design

A quasi-experimental study was conducted with 76 third-year undergraduates enrolled in the Mobile Programming and Data Management courses at a university in northern Taiwan. All students completed the same English academic writing task that required analyzing user data and designing a mobile application, under the guidance of the same instructor. All students were randomly assigned to two groups: the experimental group ($n = 38$) and the comparison group ($n = 38$). Each group was divided into smaller subgroups of five to six members to facilitate collaborative writing, with voluntary participation and institutional ethics approval ensuring fully informed consent and data confidentiality.

The six-week experiment began with an introductory session familiarizing students with academic writing and dialogic feedback. All students then completed two baseline assessments: a pre-test evaluating English academic writing through an short essay and a pre-questionnaire measuring critical thinking. Metacognitive behaviors were later examined via AI chatbot interaction logs. Before the main writing tasks, both groups received structured training on dialogic feedback strategies. The experimental group used the Generative AI-assisted Collaborative Dialogic Feedback model (GenAI-CDF), which employed structured prompts, staged writing tasks, and a dual-role interaction protocol. In contrast, the comparison group used ChatGPT-assisted dialogic feedback (GPT-DF), where students interacted with ChatGPT through standard dialogic prompts without the structured prompting, role-switching, or staged instructional design used in GenAI-CDF. Prior to the AI-assisted learning intervention, each group was provided with a shared Google Docs writing template that included structured writing guidelines and prompts. Within each group, each student was assigned responsibility for drafting a specific chapter or section. In each 90-minute phase, students then interacted with AI chatbots across three writing phases, namely brainstorming, outlining, and drafting with feedback, to develop their individual drafts. Afterward, all students shared their drafts with group members through the collaborative Google Docs template, revised their essays, and completed a post-questionnaire to evaluate improvement in target competencies (see Figure 2).

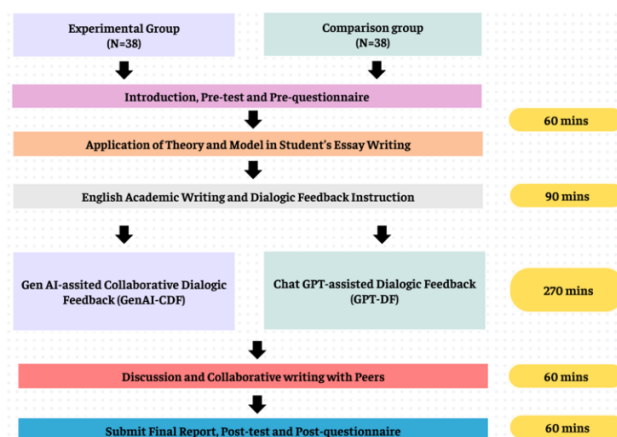


Figure 2. Experimental Design Process

4.2. Instruments and Data Analysis

To comprehensively evaluate learning outcomes, three dimensions were assessed: English academic writing, critical thinking, and metacognitive behaviors. Writing performance was measured using the rubric by Rezaei and Lovorn (2010), which included content, organization, complexity, accuracy, and fluency. Each criterion was rated on a five-point scale from 1 (Poor) to 5 (Excellent). Two experienced instructors scored all essays independently, and the mean of both ratings was used as the final score. Inter-rater reliability, assessed using the intraclass correlation coefficient (ICC), indicated high agreement between raters ($ICC = .90$). Critical thinking was assessed with a seven-item scale adapted from Lin, Hwang, and Hsu (2019), showing high reliability (Cronbach's $\alpha = 0.90$). Metacognitive behaviors were analyzed during collaborative writing activities using a coding scheme adapted from Lajoie and Lu (2012) and Meijer et al. (2006). These

behaviors were regrouped into three primary categories to represent students' metacognitive engagement with the GenAI-CDF model: Content Contribution, Commenting Behavior, and High-level Metacognitive Contribution. Each category was rated on a three-point scale from 1 (Low) to 3 (High) performance. Scores were accumulated across coded interaction segments to derive a composite metacognitive score for each student. ANCOVA and one-way ANOVA were used to compare outcomes between the two dialogic feedback models. To ensure coding reliability, a subset of metacognitive behavior scores was independently coded by two raters. Inter-rater reliability was assessed using the intraclass correlation coefficient, indicating excellent agreement between raters ($ICC = .87$).

5. Results

5.1. English Academic Writing Performance

An Analysis of Covariance (ANCOVA) was conducted to examine the effect of Gen AI-assisted dialogic feedback models on students' English academic writing performance between the Experimental group (GenAI-CDF) and Comparison group (GPT-DF) (see Table 1). The assumptions for ANCOVA were met, as Levene's test indicated no significant difference in group variances, $F(1,74) = 2.53$, $p = .116$, and the test for homogeneity of regression slopes showed no significant interaction, $F(1,72) = 0.06$, $p = .809$. After controlling for pre-test scores, the adjusted means were 20.43 ($SE = 0.31$) for the GenAI-CDF group and 15.85 ($SE = 0.31$) for the GPT-DF group. ANCOVA results revealed a significant difference in writing performance, $F(1,72) = 110.51$, $p < .001$. Bonferroni-adjusted pairwise comparisons confirmed that the GenAI-CDF group significantly outperformed the GPT-DF group ($p < .001$). These results suggest that the GenAI-CDF model was associated with higher overall academic writing performance compared with GPT-DF in this study. Additional exploratory analyses of the writing dimensions indicated significant differences in Complexity, Content, and Mechanics (all $ps < .001$), whereas Accuracy and Fluency showed no statistically significant differences.

Table 1. The result of English academic writing performance shown by ANCOVA analysis

Variable	Group	N	Mean	SE	Adjusted Mean	F	η^2
Post test	Experimental Group	38	20.34	0.31	20.43	110.51***	0.605
	Comparison Group	38	15.97	0.31	15.85		

*** $p < .001$

5.2. Critical Thinking Skills

To evaluate the impact of dialogic feedback on students' critical thinking skills, pre- and post-intervention questionnaire data were analyzed using ANCOVA (see Table 2). The assumptions were met, as Levene's test indicated no significant difference in error variances, $F(1,74) = 2.97$, $p = .089$, and the test of homogeneity of regression slopes was non-significant, $F(1,72) = 0.73$, $p = .39$. After adjusting for pre-test scores, adjusted post-test means were 4.14 ($SE = 0.08$) for the GenAI-CDF group and 4.07 ($SE = 0.08$) for the GPT-DF group. The ANCOVA revealed no significant difference in critical thinking skill, $F(1,72) = 0.46$, $p = .499$. These results indicate that students in both conditions demonstrated comparable levels of critical thinking after the intervention.

Table 2. The result of critical thinking skill shown by ANCOVA analysis

Variable	Group	N	Mean	SE	Adjusted Mean	F	η^2
Post test	Experimental Group	38	4.14	0.08	4.14	0.46	0.006
	Comparison Group	38	4.07	0.08	4.07		

5.3. Metacognitive Behaviors

As metacognitive behaviors in this study are treated as an interaction-based process rather than a static trait, they provide a more dynamic reflection of how students regulate their learning in real-time, a one-way ANOVA was conducted

to compare students' metacognitive behaviors between the Experimental Group and the Comparison Group. Levene's test confirmed the assumption of homogeneity of variances, $F(1,74) = 0.218$, $p = .642$. Results showed a significant group difference, $F(1,74) = 152.814$, $p < .001$ (see Table 3). The GenAI-CDF group achieved a higher mean score ($M = 73.97$, $SD = 8.46$) than GPT-DF Group ($M = 49.79$, $SD = 8.59$). Post hoc analysis also indicated that the Experimental Group significantly outperformed the Comparison Group ($p < .001$). These results indicate that students in the GenAI-CDF condition demonstrated higher levels of metacognitive behaviors during the writing activities compared with those in the GPT-DF condition. This pattern may reflect the role of the Interactive Structured Collaboration Mechanism in encouraging students to engage more actively in evaluating and applying feedback during the writing process.

Table 3. The result of Metacognitive behaviors shown by one-way ANOVA analysis

Group	N	Mean	SD	<i>F</i>	η^2
Experimental Group	38	73.97	8.46	152.814***	0.674
Comparison Group	38	49.79	8.59		
Total number of students	76	61.88	14.72		

*** $p < .001$

6. Discussion and Implications for Future Research

This study proposed the Generative AI-assisted Collaborative Dialogic Feedback (GenAI-CDF) model to address the limitations of traditional feedback in English academic writing by integrating metacognitive theory with dialogic feedback and the collaborative interaction framework. The model places a central emphasis on co-constructive knowledge-building processes that actively engage metacognitive regulation between students and AI throughout the writing process. Specifically, GenAI-CDF operationalizes a dual-phase AI-supported feedback mechanism, combining an AI-as-Tutor phase that supports individual reflection with an AI-as-Peer phase that promotes collaborative dialogue. Together, these phases create iterative learning cycles that deepen reflective engagement and sustain high-level academic participation. Regarding to RQ1, the results indicate that the Experimental group significantly outperformed the Comparison group in overall academic writing performance. This finding suggests that GenAI-CDF provided more comprehensive support by integrating automated AI guidance with structured collaboration (Barrot, 2023). During the initial AI-as-Tutor phase, students received immediate and scaffolded feedback that supported improvements in syntactic complexity and content organization (Su et al., 2023). In terms of Content, the specialized dialogic feedback embedded in GenAI-CDF fostered critical thinking, resulting in essays with greater depth and coherence (Li et al., 2025). Regarding Mechanics, the structured feedback encouraged greater attention to formatting and consistency in writing (Boudouaia et al., 2024). However, no significant differences were observed between the two groups in Accuracy and Fluency, suggesting comparable support for surface-level language features. Because both conditions relied on the same underlying large language model, these dimensions may be less sensitive to instructional differences, particularly in short-term interventions (Wang et al., 2025). Regarding RQ2 and RQ3, the findings reveal a developmental distinction between the activation of critical thinking and the regulation of metacognitive processes in AI-supported academic writing. Although both AI-based dialogic feedback models led to comparable improvements in students' critical thinking, indicating that AI-supported dialogic feedback is sufficient to activate foundational to intermediate critical thinking processes (Banihashem et al., 2024), only the GenAI-CDF model resulted in significantly higher levels of metacognitive behaviors. According to metacognitive theory, critical thinking is enacted when students monitor argumentation, evaluate persuasiveness, and revise writing strategies (Zhang & Zhang, 2024); however, these processes do not automatically develop into sustained metacognitive regulation without structured reflective scaffolding (Wei, 2024). The advantage of GenAI-CDF can therefore be attributed to its Interactive Structured Collaboration mechanism, which guides students to actively construct content, engage in reflection, and regulate writing decisions throughout the feedback cycle, rather than passively receiving AI-generated feedback (Lu et al., 2025). In contrast, the absence of dialogic scaffolds and

collaborative structures in GPT-DF may explain its limited effectiveness in promoting higher-level metacognitive development within a short-term intervention (Usher, 2025).

Overall, these findings reaffirm the pedagogical value of the dual-phase dialogic feedback approach embodied in the GenAI-CDF model, where AI assumes a dual role as both instructional guide and collaborative learning partner to foster reflection and activate metacognitive regulation. GenAI-CDF offers a promising framework for designing AI-mediated academic writing environments in higher education. Nevertheless, limitations related to sample size and intervention duration constrain the generalizability of the findings. Future research should develop adaptive AI feedback systems that incorporate personalized learning recommendations and adopt longitudinal research designs to examine the sustained effects of GenAI-CDF across broader academic contexts.

References

- Banihashem, S. K., Kerman, N. T., Noroozi, O., & Biemans, H. J. A. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, 21, 23.
- Barrot, J. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, 57, Article 100745.
- Boudouaia, A., Mouas, S., & Kouider, B. (2024). A study on ChatGPT-4 as an innovative approach to enhancing English as a foreign language writing learning. *Journal of Educational Computing Research*, 62(5), 1289–1317.
- Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325.
- Chelli, S. (2024). Hallucination rates and reference accuracy of ChatGPT. *Journal of Medical Internet Research*, 26, e53164.
- Dillenbourg, P. (1999). *What do you mean by collaborative learning?* In P. Dillenbourg (Ed.), *Collaborative learning: Cognitive and computational approaches* (pp. 1–19). Elsevier.
- Flavell, J. H. (1981). Cognitive monitoring. In P. Dickson (Ed.), *Children's oral communication skills* (pp. 35–60). Academic Press.
- Gong, X., Li, Z., & Qiao, A. (2025). Impact of generative AI dialogic feedback on different stages of programming problem solving. *Education and Information Technologies*, 30, 9689–9709.
- Kim, J., Yu, S., Detrick, R., & Lee, H. (2025). Exploring students' perspectives on generative AI-assisted academic writing. *Education and Information Technologies*, 30, 1265–1300.
- Lajoie, S. P., & Lu, J. (2012). Supporting self-regulated learning in collaborative learning environments with metacognitive tools. *Instructional Science*, 40(2), 347–365.
- Latifi, S., & Noroozi, O. (2021). Supporting argumentative essay writing through an online supported peer-review script. *Innovations in Education and Teaching International*, 58(5), 501–511.
- Li, B., Tan, Y. L., Wang, C., & Lowell, V. (2025). Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching. *Computers and Education: Artificial Intelligence*, 6, 100445.
- Lin, H. C., Hwang, G. J., & Hsu, Y. D. (2019). Effects of ASQ-based flipped learning on nurse practitioner learners' nursing skills, learning achievement and learning perceptions. *Computers & Education*, 139, 207–221.
- Lu, G., Ba, S., & Yang, L. (2025). Optimizing generative AI-assisted formative feedback: An experimental study on its effects on engagement, shared metacognition, and learning performance in online collaborative learning. *Education and Information Technologies*. Advance online publication.
- Meyer, J., Jansen, T., & Fleckenstein, J. (2025). Nonengagement and unsuccessful engagement with feedback in lower secondary education: The role of student characteristics. *Contemporary Educational Psychology*. Advance online publication.

- Mu, H., & Schunn, C. D. (2025). The good, bad, and ugly of comment prompts: Effects on length and helpfulness of peer feedback. *International Journal of Educational Technology in Higher Education*, 22(4). <https://doi.org/10.1186/s41239-025-00502-8>
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18–39. <https://doi.org/10.1016/j.asw.2010.01.003>
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, Article 101894. <https://doi.org/10.1016/j.learninstruc.2024.101894>
- Stephen, J. S. (2024). *Academic success in online programs: A resource for college students* (1st ed.). Springer. <https://doi.org/10.1007/978-3-031-54439-2>
- Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, 58, 100752. <https://doi.org/10.1016/j.asw.2023.100752>
- Subandowo, D., Sárdi, C., & Thresia, F. (2025). An investigation of English academic writing strategies employed by Indonesian graduate students in an English medium instruction (EMI) context. *Asian Journal of Second and Foreign Language Education*, 10, 38. <https://doi.org/10.1186/s40862-025-00345-w>
- Sun, D., Boudouaia, A., Zhu, C., & Li, Y. (2024). Would ChatGPT-facilitated programming mode impact college students' programming behaviors, performances, and perceptions? An empirical study. *International Journal of Educational Technology in Higher Education*, 21(1), 14. <https://doi.org/10.1186/s41239-024-00446-5>
- Taguba, L. M., & Plata, S. M. (2025). Written corrective feedback uptake and perceptions of Filipino high school students in L2 writing. *Language Testing in Asia*, 15, 22. <https://doi.org/10.1186/s40468-025-00362-z>
- Teng, M. F., & Ma, Y.-Y. (2024). Assessing metacognition-based student feedback literacy for academic writing. *Assessing Writing*, 59, 100811. <https://doi.org/10.1016/j.asw.2024.100811>
- Teng, M. F., and C. Qin. 2024. "Assessing Metacognitive Writing Strategies and the Predicative Effects on Multimedia Writing." *Asia Pacific Journal of Education*: 1–24. <https://doi.org/10.1080/02188791.2024.2325132>.
- Usher, M. (2025). Generative AI vs. instructor vs. peer assessments: a comparison of grading and feedback in higher education. *Assessment & Evaluation in Higher Education*, 50(6), 912–927. <https://doi.org/10.1080/02602938.2025.2487495>
- Wang, C. (2025). Exploring students' generative AI-assisted writing processes: Perceptions and experiences from native and nonnative English speakers. *TechTrends: Technology, Knowledge and Learning*, 30, 1825–1846. <https://doi.org/10.1007/s10758-024-09744-3>
- Wei, Y., & Liu, D. (2024). Incorporating peer feedback in academic writing: A systematic review of benefits and challenges. *Frontiers in Psychology*, 15, 1506725. <https://doi.org/10.3389/fpsyg.2024.1506725>
- Winstone, N., et al. (2022). The dynamics of dialogic feedback in enhancing student learning in higher education. *Learning and Instruction*, 73, 101406. <https://doi.org/10.1016/j.learninstruc.2020.101406>
- Zhang, J., & Zhang, L. J. (2024). The effect of feedback on metacognitive strategy uses in EFL writing. *Computer Assisted Language Learning*, 37(5–6), 1198–1223. <https://doi.org/10.1080/09588221.2022.2069822>

AI-Supported Iterative Revision and Reflective Writing in Primary Education: A Quasi-Experimental Study

Xiaoqing Xian¹, Jinli Wu², Xiao-Fan Lin^{3*}

¹ South China Normal University, School of Educational Information Technology

² South China Normal University, School of Educational Information Technology

³ South China Normal University, School of Educational Information Technology

*corresponding.author@South China Normal University.China

Abstract: This study examined whether an AI-supported iterative revision workflow can improve primary students' writing development in an authentic school setting. A quasi-experiment was conducted over one academic year with 259 Grade 5 students in a Shenzhen public school (experimental $n = 85$; control $n = 174$). The experimental classes used an AI-supported writing platform that provided immediate analytic feedback, revision prompts, and portfolio records, while the comparison classes received conventional teacher-led feedback. Because the intervention bundled AI feedback with more iterative revision opportunities, the study estimates the effect of an AI-mediated revision ecology rather than an isolated AI-only effect. Outcomes were analyzed with ANCOVA on five questionnaire dimensions (context, content, process, strategies, and feedback) and with Epistemic Network Analysis (ENA) on rubric-based writing performance. After controlling for pre-test scores, the experimental group obtained significantly higher post-test scores on overall writing literacy and all subdimensions, with the strongest gains in content development and feedback engagement. ENA further showed that the experimental group displayed denser connections among content, structure, and norms, indicating better integration of higher-order writing dimensions. The findings suggest that AI-supported revision can be pedagogically valuable when used to scaffold reflection, repeated drafting, and data-informed teacher guidance. The paper discusses both the promise of this intervention package and the limits of causal claims when revision dosage differs across conditions.

Keywords: AI-supported writing, reflective revision, primary education, feedback, ENA

1. Introduction

As one of the foundational skills, writing is regarded as a main skills for students' academic performance and active participation in society. To write effectively, learners must be able to generate and organize meaningful content, apply accurate language mechanics, and adapt their discourse to suit specific contexts and audiences. Argumentative writing, in particular, has been shown to promote scientific thinking, stimulate higher-order reasoning, and support deeper processing and organization of information (Bal et al., 2025; Liu et al., 2024). Integrating writing across the curriculum further enables students to internalize disciplinary knowledge, synthesize diverse perspectives, and consolidate conceptual understanding. However, many learners, especially those writing in a foreign language, exhibit limited reflective awareness, leading them to overlook essential lexical and syntactic elements, which undermines the coherence and quality of their compositions (Shen & Teng, 2024). Encouraging reflection on the writing process can foster higher writing skills, enhance self-regulation, and ultimately improve both the quality of writing and overall learning outcomes (Bal et al., 2025).

However, writing feedback often remains superficial, focusing predominantly on surface-level corrections such as grammar and spelling rather than providing strategic guidance on idea development or writing skills. While written

feedback is widely acknowledged as valuable for developing academic writing skills, its effectiveness is limited when it lacks clarity or fails to include actionable feed-up, feed-back, and feed-forward elements (Schillings et al., 2023). Consequently, many students apply corrections mechanically without fully understanding how to improve, revealing a critical gap in reflective learning. Research indicates that self-reflection, particularly when linked to peer feedback, can foster deeper engagement with the revision process and lead to substantive improvements in both content and language (Li & Hebert, 2024). Yet, without explicit instructional scaffolds, students rarely engage in timely, structured reflection on their writing (Li & Hebert, 2024). Targeted interventions, such as guided revision plans and reflective prompts, can encourage learners to consider feedback at the process and self-regulatory levels, enabling them to set goals, plan revisions, and internalize strategies for future writing tasks.

Recent advances in generative artificial intelligence (Gen AI), particularly large language models (LLMs) such as GPT-4, have markedly reshaped the field of automated writing evaluation (AWE). Traditional automated essay scoring (AES) systems, though effective at detecting surface-level features such as grammar, lexical diversity, or word count, have been widely criticized for their inability to capture deeper dimensions of writing, including coherence, argument structure, and idea development (Pack et al., 2024). In contrast, LLMs are not bound by pre-defined, rigid scoring rubrics; instead, they can dynamically adapt to human-like evaluation criteria through prompt engineering and contextual reasoning (Pack et al., 2024). This flexibility enables them to provide holistic, context-aware, and individualized feedback that extends beyond error detection to address higher-order writing skills. Emerging empirical evidence further suggests that models like GPT-4 can produce feedback that is not only accurate and aligned with expert judgments but also more detailed, readable, and pedagogically supportive than that provided by many human teachers (Pack et al., 2024). Such qualities make LLM-generated feedback particularly well-suited for promoting reflective learning processes, helping students to internalize revision strategies and engage more critically with their writing skills.

Despite these promising developments, important gaps remain in the application of AI to writing education, particularly for younger learners in primary school contexts. While AI-generated feedback has demonstrated potential to move beyond surface-level corrections, less is known about how such feedback can be integrated into reflective cycles of learning where students actively evaluate, plan, and revise their writing. Existing research has largely focused on accuracy and scoring alignment, whereas fewer studies have explored its role in scaffolding reflective writing and strategy use. For primary students, who often lack the self-regulatory skills required to engage deeply with revision, the challenge lies not only in receiving feedback but also in transforming it into reflective practice.

This paper reports a school-based quasi-experiment on an AI-supported writing platform for Grade 5 students. In response to reviewer concerns, the manuscript has been substantially condensed and the interpretation has been tightened. We do not claim that AI alone outperformed teacher feedback under equivalent conditions. Instead, we examine the effectiveness of an AI-supported iterative revision workflow in which automated feedback, portfolio records, and teacher oversight were combined into a classroom intervention package. Two questions guided the study: (1) Does the intervention improve students' writing literacy outcomes compared with conventional instruction? (2) Does it change the structural relations among rubric dimensions of writing performance?

2. Method

2.1. Participants and design

The analytic sample comprised 259 Grade 5 students from a public primary school in Shenzhen after data cleaning and matching of pre/post records. The experimental group included 85 students and the control group included 174 students. The study lasted one academic year and used intact classes, so the design was quasi-experimental rather than randomized. Both groups followed the same curriculum and completed comparable writing tasks. The difference was in the feedback ecology: the experimental classes used an AI-supported platform for immediate scoring, sentence-level

comments, revision prompts, and portfolio tracking, whereas the control classes received conventional teacher-led marking and classroom feedback.

Because the experimental condition combined AI feedback with repeated revision opportunities, the intervention should be understood as an AI-mediated process package. This framing addresses the dosage concern raised in review: the study evaluates a realistic instructional ecology rather than a pure AI-versus-teacher contrast.

2.2. Measures and analysis

Writing literacy was measured with a five-dimension questionnaire covering writing context, content, process, strategies, and feedback engagement. Writing performance was also rated with a four-dimension rubric (content, expression, structure, norms). Pre-test and post-test questionnaire reliability was acceptable to strong (Cronbach's alpha = .771 and .967). ANCOVA was used to estimate post-test group differences while controlling for pre-test scores.

ENA was used as a complementary analysis rather than a substitute for regression. Regression-based models are appropriate for estimating mean differences in outcomes, which is why ANCOVA was the primary inferential test. ENA was added because the second research question concerned how writing dimensions co-occurred as a connected system within students' texts. ENA makes those relational patterns visible by modeling the strength of ties among rubric dimensions and by testing whether the groups occupy different positions in network space. In other words, ANCOVA answers whether students improved more, while ENA helps explain how the organization of writing performance differed across conditions.

Table 1
Key post-test differences between groups

Outcome	Adjusted mean Exp.	Adjusted mean Ctrl.	Statistic	Interpretation
Overall writing literacy	3.930	3.404	F = 30.372***	Experimental group higher
Writing context	3.917	3.423	F = 22.533***	Better audience/context awareness
Writing content	4.065	3.347	F = 40.553***	Largest gain in content development
Writing process	3.896	3.402	F = 22.141***	More process-oriented revision
Writing strategies	3.737	3.392	F = 13.527***	Higher strategy use
Writing feedback	4.001	3.455	t = 5.149***	Stronger feedback engagement
ENA centroid difference	-0.11 vs 0.04	--	t = -3.05***, d = 0.50	Different network structure

3. Results

Table 1 shows that the experimental group scored significantly higher than the control group on overall writing literacy and all five questionnaire dimensions after controlling for pre-test differences. The strongest statistical advantage appeared in writing content, followed by overall literacy, context, and process. The writing feedback measure also favored the experimental condition, suggesting that students in the AI-supported environment became more willing and able to

use evaluative comments during revision. Across analyses, however, effect sizes were generally small. This pattern implies that the intervention produced consistent but moderate educational gains rather than dramatic transformation.

The ENA results extended the mean-difference findings. The first rotation axis explained the largest share of variance, and the group centroids differed significantly on MR1, $t = -3.05$, $p < .001$, $d = 0.50$. Network comparisons indicated that the experimental group showed stronger ties among Structure-Norms (.36 vs. .13), Content-Structure (.22 vs. .07), and Content-Norms (.25 vs. .14). These denser higher-order linkages suggest that the intervention did not simply improve isolated rubric scores; it also promoted more integrated writing in which idea development, organization, and conventions worked together more coherently.

4. Discussion and conclusion

Two conclusions can be drawn cautiously. First, the AI-supported iterative revision workflow was associated with better writing outcomes than conventional instruction in this setting. The consistency of the results across overall literacy, subdimensions, and ENA strengthens confidence that the intervention had pedagogical value. Second, the mechanism is best interpreted as reflective revision support rather than AI automation alone. Immediate comments, repeated drafting, visible portfolio records, and teacher mediation likely worked together to encourage students to revisit and reorganize their texts. This interpretation aligns with prior studies showing that feedback becomes more powerful when learners actively use it to plan and revise (Li & Hebert, 2024; Neshaei et al., 2025). The paper also clarifies why ENA was retained. ANCOVA established that the groups differed in mean outcomes, but ENA addressed a different question by revealing how dimensions of writing performance were interconnected. This was important because writing quality is multidimensional, and higher-quality compositions are often characterized by coordination among content, structure, and norms rather than improvement on a single isolated trait.

The results from different empirical evidence sources were mutually corroborated (triangulation), and students who received AI-supported methods performed significantly better than other students under the traditional teaching feedback model. Theoretically, the study contributes to a broader epistemic role on the role of artificial intelligence in educational assessment. Not only AI as a tool for providing timely corrective feedback (e.g., grammar, vocabulary, and surface-level accuracy), but also the ENA results in this study highlight that AI-supported approach restructured the relational patterns of students' writing, strengthening connections between content development, structural organisation, and linguistic norms. This indicates that AI-supported system acted as a focusing device, directing learners' attention towards higher-order dimensions of composition and fostering more coherent textual production. These functions align with recent findings that LLM-enabled tools promote iterative, reflective learning processes (Neshaei et al., 2025) and resonate with theories of double-loop learning (Chen et al., 2022), whereby learners move beyond error correction towards strategic rethinking of writing practices. From a practical standpoint, the findings underscore several pedagogical suggestions and implications. First, the results highlight the value of multi-round, personalised evaluation, whereby AI can provide iterative prompts that guide students' revision, while teachers ensure reliability through validation and adjustment of automated scores. This combined mechanism ensures both efficiency and pedagogical trustworthiness. Second, the study points to the usefulness of portfolios as resources and scaffolds. By archiving students' drafts, revisions, and feedback histories, the system not only supported formative assessment but also served as a reflective space where pupils could monitor progress and transfer strategies across tasks. This echoes evidence from recent classroom-based AI writing interventions showing that portfolio-based AI support enhances self-regulation and strategy transfer (Zhang & Yu, 2023; Bal et al., 2025). Third, the intervention provided scaffolding for writing skills and literacy strategy use. Even though effect sizes for strategy improvement were modest, they are significant in the primary context, where writing skills are often underdeveloped. By visualising progress and offering actionable prompts, the system helped students engage in planning, monitoring, and evaluation—key components of effective writing pedagogy (Etaat, 2025). Finally, by

strengthening the interconnections between higher-order dimensions of writing, AI-supported instruction offers a countermeasure to the long-standing challenge of integrating feedback into students' recursive writing cycles.

Several limitations remain. The quasi-experimental design does not support strong causal inference, and the unequal revision opportunities across groups mean the study cannot separate the contribution of AI from the contribution of iterative practice. The findings should therefore be read as evidence for an intervention package that integrated AI into a reflective classroom workflow. Future studies should equalize the number of revision rounds across conditions, compare AI with non-AI iterative feedback loops, and test the model in multiple schools. Even with these limitations, the present results suggest that AI can be educationally useful when it is designed to augment teacher judgment and increase students' opportunities for structured reflection.

References

- Bal, M., Uyumaz, G., & Maden, B. (2025). Examining the mediating role of metacognitive writing strategies in the relationship between self-regulated writing skills and writing anxiety among middle school students. *British Educational Research Journal*. Advance online publication. <https://doi.org/10.1002/berj.4212>
- Chen, X., Ou, Q., An, C., & Zhang, D. (2022). Teacher mindset change in boundary-crossing lesson study: A case from China. *International Journal for Lesson & Learning Studies*, 11(2), 91-105. <https://doi.org/10.1108/ijlls-03-2021-0019>
- Etaat, F. (2025). AI-based assessment of students' writing skill progress: Implementing “Artificial neural network modeling” vs. “Time series prediction.” *International Journal of Educational Research*, 131, 102575. <https://doi.org/10.1016/j.ijer.2025.102575>
- Li, A. W., & Hebert, M. (2024). Unpacking an online peer-mediated and self-reflective revision process in second-language persuasive writing. *Reading and Writing*, 37(6), 1545-1573. <https://doi.org/10.1007/s11145-023-10466-8>
- Liu, Q., Zhong, Z., & Nesbit, J. C. (2024). Argument mapping as a pre-writing activity: Does it promote writing skills of EFL learners? *Education and Information Technologies*, 29, 7895–7925. <https://doi.org/10.1007/s10639-023-12098-5>
- Neshaei, S. P., Mejia-Domenzain, P., Davis, R. L., & Kaser, T. (2025). Metacognition meets AI: Empowering reflective writing with large language models. *British Journal of Educational Technology*, 56(5), 1864-1896. <https://doi.org/10.1111/bjet.13601>
- Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6, 100234. <https://doi.org/10.1016/j.caeai.2024.100234>
- Schillings, M., Roebertsen, H., Savelberg, H., & Dolmans, D. (2023). A review of educational dialogue strategies to improve academic writing skills. *Active Learning in Higher Education*, 24(2), 95-108. <https://doi.org/10.1177/1469787418810663>
- Shen, X. L., & Teng, M. F. (2024). Three-wave cross-lagged model on the correlations between critical thinking skills, self-directed learning competency and AI-supported writing. *Thinking Skills and Creativity*, 52, 101524. <https://doi.org/10.1016/j.tsc.2024.101524>
- Zhang, E. D., & Yu, S. (2023). The development and validation of an L2 student digital multimodal composing competence scale. *Computer Assisted Language Learning*, 38(3), 486–511. <https://doi.org/10.1080/09588221.2023.2201342>
- Zhang, E. D., Zhou, N., & Yu, S. (2023). Assessing L2 secondary student writing feedback literacy and its predictive effect on their L2 writing performance. *Language Teaching Research*. Advance online publication. <https://doi.org/10.1177/13621688231217665>

Design and Feasibility of a Modular Prompt Framework for Polya-Based Mathematical Instructional Scaffolding

Hsuan-Yung Fang¹, Regina Juchun Chu^{2*}

^{1,2} Institute of Learning Sciences and Technologies, Tsing Hua University, Taiwan

*chu@mx.nthu.edu.tw

Abstract: *This study addresses the lack of real-time heuristic guidance in mathematics self-study by designing an AI prompt framework grounded in Polya's four-stage model. We developed a five-module structure—Persona, Rules, Instruction, Reflection, and Practice—to provide digital scaffolding simulating teacher-like guidance. Testing with Gemini 2.5 and teacher interviews demonstrates that the framework delivers coherent reasoning without hallucinations. Findings suggest the Reflection module effectively supports metacognitive awareness. While limited for students with low prior knowledge, the study confirms the feasibility of prompt engineering as instructional scaffolding in mathematics.*

Keywords: Structured Prompt Patterns, Polya's Problem Solving, AI Math Learning

1. Introduction

Generative Artificial Intelligence (GenAI) has emerged as a vital tool for fostering student engagement, self-efficacy, and accessibility (Chohan & Khan, 2024; Dmitrenko et al., 2025). Its pedagogical utility extends to automating instructional design, enriching content knowledge, and diagnosing error patterns for targeted remediation (Getenet, 2024; Oh, 2023). Nevertheless, mathematical applications remain constrained by "hallucinations," fragmented reasoning, and premature answer disclosure that hinders conceptual depth (Drushlyak et al., 2025; Schorcht et al., 2024; Wu, 2024). These limitations are frequently exacerbated by poorly structured prompting, leading to logical fallacies and diminished rigor (Drushlyak et al., 2025; White et al., 2023).

While prior inquiries emphasize terminal accuracy, few address the nuances of process-oriented interactions in autonomous learning where students struggle with strategic reasoning without heuristic scaffolding (Getenet, 2024; White et al., 2023). Addressing this gap, this study integrates Polya's four-stage model with modular prompt engineering to examine the feasibility of an interactive framework designed as a process-oriented learning partner.

2. Literature Review

2.1. The Mathematical Problem-Solving Process

Mathematical problem-solving is a central proficiency in national curriculum frameworks. Within this domain, Polya's four-stage model—*understanding* the problem, *planning* (devising a plan), *execution* (carrying out the plan), and *reflection* (looking back)—remains the dominant framework for analyzing student behaviors and designing instructional support. Students must transform verbal descriptions into mathematical representations during understanding, while planning and execution require flexible strategy selection. Finally, reflection involves metacognitive strategies for evaluating reasoning and revising solutions.

Despite this structure, students often lack real-time scaffolding for strategy formulation in out-of-class contexts. To bridge this gap, this study operationalizes these stages as explicit AI dialogue rules. Rather than providing direct answers, the framework structures prompt to guide students through meaning-making and reflective thinking, acting as a heuristic partner for self-directed study.

2.2. AI Applications in Mathematical Problem Solving

Current research indicates that while AI models often exhibit fragmented reasoning and "hallucinations" (Drushlyak et al., 2025; Oh, 2023), their performance remains highly sensitive to prompt quality. Beyond simple model upgrades, prompt engineering has emerged as a decisive factor in improving mathematical coherence and strategy use (Schorcht et al., 2024). Within this landscape, White et al. (2023) introduced a Prompt Pattern Catalog—comprising patterns such as *Persona*, *Flipped Interaction*, *Cognitive Verifier*, *Fact-Check*, and *Reflection*. These patterns enable natural-language instructions to express complex pedagogical functions, including role assignment, task decomposition, and error verification. Such patterns are inherently composable, allowing researchers to design structured prompt sequences that simulate the heuristic scaffolding required for mathematical problem-solving (Oh, 2023).

Empirical studies demonstrate that such structured frameworks significantly boost accuracy and instructional quality. For instance, incorporating roles and process guidance based on Polya's model has led to accuracy rates exceeding 90% (Oh, 2023; Chohan & Khan, 2024). Moreover, interactive AI tools have shown potential in fostering metacognitive skills, such as error detection and reflective thinking (Dmitrenko et al., 2025; Drushlyak et al., 2025). However, limitations persist in strategic diversity, as models often struggle to move beyond algebraic reasoning to graphical or heuristic approaches (Getenet, 2024). Synthesizing these findings, the present study adopts Polya's process as a structural core, integrating multiple prompt patterns to facilitate coherent, process-oriented interactions that support students' strategy refinement and self-reflection.

3. Method

Following Polya's four-stage problem-solving framework, this study designed an AI-based interactive mathematical learning tool that supports the stages of understanding, planning, execution, and reflection, enabling students to conduct personalized after-school learning and problem practice. To ensure response quality and maintain a natural interaction environment, a structured prompt architecture was implemented, consisting of five modules—Persona, Rules, Instruction, Reflection, and Practice. The dynamic interaction workflow and module coordination are visualized in Figure 1.

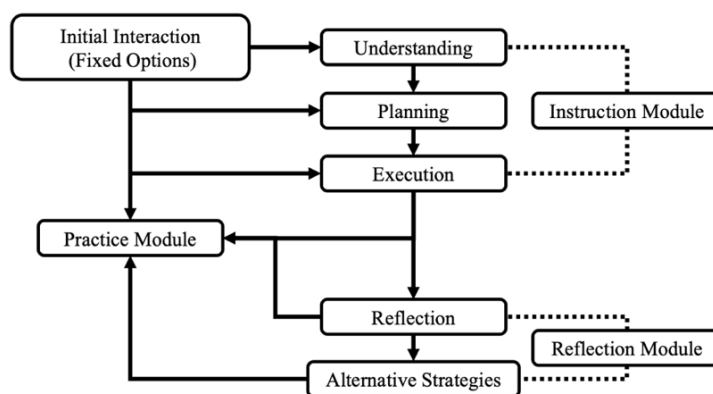


Figure 1. Dynamic interaction workflow and module transition logic

As illustrated in Figure 1, the Rules Module governs the overall interaction flow; students begin by selecting from fixed options (e.g., "stuck midway") that explicitly express their needs, which then triggers the corresponding instructional stage to clarify intent and reduce cognitive load. The Persona Module defines the AI as a facilitator responsible for eliciting student thinking, employing an "affirm-then-correct" strategy to maintain an encouraging tone. The Rules Module further implements "cognitive load management" (e.g., pausing after three consecutive equations) and ensures that all interactions strictly follow Ask-Me-Anything (AMA) and Chain-of-Thought (CoT) principles to prevent proactive disclosure of final answers. The Instruction Module operationalizes Polya's first three stages; instead of explicitly naming prerequisite concepts, it uses questioning and condition decomposition to guide students toward independently recognizing relevant methods, utilizing the Cognitive Verifier Pattern. During the planning stage, the Fact-Check Pattern helps evaluate strategy feasibility, while the execution stage provides reasoning via CoT, with scaffolding gradually reduced as student independence increases. The Reflection Module corresponds to Polya's "looking back" stage,

enforcing three core questions (identifying challenges, critical steps, and logic explanation) and offering alternative strategies to foster strategic flexibility. Finally, the Practice Module facilitates learning transfer by generating tiered tasks (Basic, Advanced, and Challenge problems) based on the Gameplay and Infinite-generation patterns, concluding with an automated learning report for long-term self-regulation.

To examine its feasibility and practical effectiveness, this study implemented full interaction tests using one number theory problem and one geometry problem, followed by demonstration sessions and semi-structured interviews with three middle school mathematics teachers, anonymized as TA, TB, and TC. Interview excerpts are indicated using the prefix (I), labeled according to question numbers 1 through 5. The interview data were verified by the interviewees to ensure the accuracy of interpretation.

4. Result

To address educational accessibility and equity, this study employed Gemini 2.5 to examine whether the proposed framework could effectively reproduce Polya's four-stage process and maintain reasoning consistency without additional fine-tuning. The model's capacity for unrestricted long-form dialogue facilitates authentic autonomous learning, helping to reduce disparities among students with varying educational resources when seeking support outside the classroom.

4.1. Model Performance and Reasoning Stability

To verify the framework's stability, this study implemented interaction tests using two distinct tasks: a number theory problem ($(x+1) / (x^2-3x+3)$ is an integer) and a geometry folding problem.

In the number theory test, the Rules Module correctly identified the "stuck midway" intent and triggered the Instruction Module. The AI utilized Cognitive Verifier and Fact-Check patterns to decompose conditions without proactive disclosure of final answers. Chain-of-Thought (CoT) reasoning was activated only when computations became complex, maintaining logical consistency without skipped steps. Upon requesting alternative approaches, the system transitioned smoothly to the Reflection Module to guide comparison of logical structures. During the Practice Module, when the researcher adopted an inappropriate approach, the AI redirected reasoning through counter-questioning rather than providing the solution, demonstrating scaffolded instructional behavior.

In the geometry test, the system demonstrated robust representational transformation capabilities. When the learner expressed unfamiliarity with trigonometric ratios, the AI reformulated the solution using more fundamental approaches. It subsequently activated CoT reasoning, breaking calculations into micro-steps while maintaining logical coherence. Throughout both tasks, no skipped logical steps, computational errors, or fabricated conditions (hallucinations) were observed, confirming the framework's instructional robustness and stability in complex mathematical inference.

4.2. Field Interviews with Teachers

Three teachers (TA, TB, TC) affirmed the system's instructional potential while noting discrepancies with actual classroom use. Key findings include: (1) Competence: AI responses partially reflected the researcher's *"too teacher-like"* (TC-I-1) style. The system is most effective for students with basic competence, as lower-achieving learners *"do not even know how to ask questions"* (TA-I-5). (2) Adaptation: Teachers suggested adding functions that *"recommends which unit students should review"* (TA-I-5) for students lacking fundamental concepts. (3) Over-reliance: Concerns were raised that rapid AI feedback might allow students to *"bypass genuine reasoning"* (TB-I-2). (4) Barriers: AI supplements support but cannot replace pedagogical judgment. Success remains contingent on student self-discipline and the tension between scaffolding and *"sometimes letting students struggle independently is part of their growth"* (TC-I-4).

5. Discussion and Conclusion

The findings demonstrate that the proposed five-module framework effectively reproduces Polya's four-stage process, functioning as a tool for metacognitive support through seamless transitions and question-driven scaffolding (Oh, 2023; Schorcht et al., 2024; White et al., 2023). However, consistent with pedagogical concerns identified during testing, challenges remain regarding instructional depth and learner over-reliance, as AI currently struggles to adapt its pathway

for students lacking prior knowledge (Wu, 2024). Despite these limitations, the framework addresses critical gaps in after-school learning by providing heuristic guidance and promoting educational equity through Gemini 2.5.

Regarding limitations, since interaction tests relied on researcher simulation within the specific domains of number theory and geometry, future work should involve direct student participation to enhance real-world applicability. Furthermore, as the system relies exclusively on prompt engineering, its performance reflects the base model's reasoning rather than domain-specific instructional expertise. Future iterations should incorporate student-level diagnostic mechanisms and broader curriculum alignment to improve precision. Overall, this framework offers a transferable model for integrating AI into mathematics education, fostering a more personalized and accessible learning environment.

References

- Chohan, I., & Khan, I. (2024). Enhancing mathematics education with ChatGPT-4 personalized problem-solving and consistent learning. 2024 2nd International Conference on Computing and Data Analytics (ICCD). <https://doi.org/10.1109/ICCD464887.2024.10867401>
- Dmitrenko, N., Shakhina, I., Fedyk, T., Podzygun, O., & Franchuk, N. (2025). Leveraging ChatGPT to develop critical thinking through mathematical problem-solving in ESP classes. *Environment. Technology. Resources. Proceedings of the 16th International Scientific and Practical Conference*, 3, 91–98. <https://doi.org/10.17770/etr2025vol3.8540>
- Drushlyak, M., Lukashova, T., Sabadosh, Y., Melnikov, I., & Semenikhina, O. (2024). Using ChatGPT for the development of critical thinking in youth: Example of inequality proof. 2024 47th MIPRO ICT and Electronics Convention (MIPRO). <https://doi.org/10.1109/MIPRO60963.2024.10569759>
- Drushlyak, M., Lukashova, T., Shamonina, V., & Semenikhina, O. (2025). ChatGPT-based simulation helps to develop the pre-service mathematics teachers' critical thinking. *International Journal of Instruction*, 18(1), 153–172. <https://doi.org/10.29333/iji.2025.1819a>
- Getenet, S. (2024). Pre-service teachers and ChatGPT in multistrategy problem-solving: Implications for mathematics teaching in primary schools. *International Electronic Journal of Mathematics Education*, 19(1), em0766. <https://doi.org/10.29333/iejme/14141>
- Oh, S. J. (2023). Effective ChatGPT prompts in mathematical problem solving: Focusing on quadratic equations and quadratic functions. *Communications of Mathematical Education*, 37(3), 545–567. <https://doi.org/10.7468/jksmee.2023.37.3.545>
- Schorcht, S., Buchholtz, N., & Baumanns, L. (2024). Prompt the problem – investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. *Frontiers in Education*, 9, 1386075. <https://doi.org/10.3389/feduc.2024.1386075>
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv*. <https://doi.org/10.48550/arXiv.2302.11382>
- Wu, Y. (2024). Critical thinking pedagogics design in an era of ChatGPT and other AI tools—Shifting from teaching “what” to teaching “why” and “how”. *Journal of Education and Development*, 8(1). <https://doi.org/10.20849/jed.v8i1.1404>

Integrating Generative AI into Strategy-Based Instruction: Adaptive Scaffolding for Self-Regulated Learning in Secondary EFL Writing

Chuang-Qi Chen

Institute of Language Sciences, Shanghai International Studies University

413806370@qq.com

Meng-Rui Li

School of Education, Shanghai International Studies University

LMR19115063061@outlook.com

Hui Jin

School of Education, Shanghai International Studies University

nmjinhui@163.com

Abstract: Can GenAI do more than generate text—can it teach students to regulate their own writing? This study addresses that question through a randomized controlled trial ($N = 80$) examining whether a GenAI-powered adaptive dialogue agent, embedded in a Strategy-Based Instruction (SBI) framework, develops secondary EFL students' Self-Regulated Learning (SRL) strategies alongside writing performance. The agent was designed to foreground student agency: it deployed four context-sensitive modes—genre knowledge retrieval, ideation prompting, reflective self-assessment, and motivational support—mapped onto the SRL cycle. Over eight weeks, the experimental group ($n = 40$) significantly outperformed controls on writing scores, with the performance gap widening at a one-month delayed post-test. Crucially, the intervention produced sustained gains in text generation and self-monitoring strategies—competencies rarely improved through conventional instruction. These findings reframe GenAI's pedagogical role: not as a writing surrogate, but as a metacognitive scaffold that makes strategic thinking visible and keeps learners in the driver's seat of their own writing development.

Keywords: Generative AI (GenAI), Self-Regulated Learning (SRL), Adaptive Scaffolding, EFL Writing, Strategy-Based Instruction (SBI)

1. Introduction

Generative Artificial Intelligence (GenAI) is fundamentally reshaping text production (Cerchione et al., 2026). While AI-driven efficiency offers transformative potential, it triggers a critical pedagogical concern: the risk of bypassing students' cognitive engagement and Self-Regulated Learning (SRL) capacities. If learners are relegated to passive recipients of AI-generated content, their essential writing skills and cognitive agency face a "hollowing out" effect in the AI era (Kim et al., 2025) (Adams, 2025).

To address this challenge, writing instruction must pivot from product evaluation to cultivating the underlying cognitive processes. Strategy-Based Instruction (SBI), grounded in SRL theory, provides a vital framework for this shift (Rubin et al., 2007). In an AI-augmented landscape, the core competency is no longer mere text production but the strategic management of the writing lifecycle—from goal setting and planning to monitoring and revising. However, implementing SBI in traditional large-scale secondary classrooms is logistically difficult (Del Mario & Tran, 2024); teachers often lack the bandwidth to monitor every student's "black box" of cognitive processes, resulting in delayed, product-oriented feedback.

Recent advancements in GenAI offer a unique opportunity to bridge this gap. Unlike static tools, GenAI can function as a "metacognitive partner" capable of delivering ubiquitous, adaptive scaffolding—such as genre knowledge retrieval or ideation guidance—tailored to individual needs (Tang & Putra, 2025). Despite extensive research on AI for error

correction, few studies have explored the systematic integration of GenAI within an SBI framework to specifically foster secondary students' SRL strategies.

Addressing this gap, the present study employs a randomized controlled trial to evaluate the extent to which a GenAI-supported SBI intervention enhances writing performance and strategic deployment.

2. GenAI Agent Design and Implementation

Writing strategy instruction within an SBI framework demands support that is both responsive and differentiated. Students must simultaneously navigate genre conventions, construct logical arguments, monitor their own progress, and manage writing-related affect (Rubin et al., 2007)—each task calling for a qualitatively different kind of support. A one-size-fits-all chatbot cannot meet these varied demands: it risks delivering off-target feedback that impedes rather than builds strategic competence (Azevedo & Hadwin, 2005). This concern is especially acute for genre knowledge delivery, where factual accuracy is non-negotiable: even minor hallucinations can erode student trust and distort their understanding of writing conventions (Hattie & Timperley, 2007). Given inherent LLM limitations—output variability and the potential for confabulation—a structurally differentiated, modular design is essential to guarantee both pedagogical reliability and interaction-level adaptability.

To address these constraints, we constructed the agent around four task-specific functional modes, each instantiating a core scaffolding principle (Puntambekar & Hubscher, 2005) and activated in response to real-time learner cues:

1. **Genre Knowledge Mode:** Anchors instruction in verified content via Retrieval-Augmented Generation (RAG). Rather than relying on open-ended generation, the agent consults a curriculum-aligned knowledge base to supply structurally accurate guidance on text types such as formal letters and argumentative essays. This retrieval-grounded approach minimizes the risk of misleading information and ensures fidelity to established writing conventions.

2. **Ideation & Logic Mode:** Supports pre-writing and drafting through strategic questioning rather than text generation. The agent poses probing prompts that push students to articulate their claims, identify supporting evidence, and tighten the logical structure of their arguments. By keeping students as the primary authors, this mode cultivates independent reasoning within the zone of proximal development.

3. **Reflective & Monitoring Mode:** Develops metacognitive self-regulation during revision. Rather than issuing evaluative judgments, the agent poses goal-referencing questions (e.g., “How well does this paragraph signal logical contrast through transitional language?”) that prompt learners to compare their draft against their original intentions. This structured reflection trains students to apply self-assessment criteria independently over time.

4. **Motivational & Affective Mode:** Attends to the emotional dimension of writing, which is particularly salient among adolescent EFL learners prone to writing anxiety and low self-efficacy. The agent offers affirmation and empathetic acknowledgment to sustain motivation and preserve a psychologically safe space—conditions that research identifies as foundational for productive intellectual risk-taking.

Taken together, these four modes operationalize the principle of contingent scaffolding—providing targeted support calibrated to the learner’s immediate cognitive and affective state rather than following a fixed instructional sequence (Azevedo & Hadwin, 2005).

2.1. Workflow Architecture

Figure 1 illustrates the technical infrastructure underpinning the agent. The system was built on Dify (v 0.13.2), an open-source LLM-ops platform selected for its support of multi-agent orchestration and RAG integration within a visual workflow environment.

Each incoming message first passes through a Context Manager and Memory Module, which reconstruct the learner’s interaction history and surface relevant metadata. A Routing mechanism then classifies the communicative intent of the message and directs it to the most contextually appropriate of the four specialized Retrieval Agents.

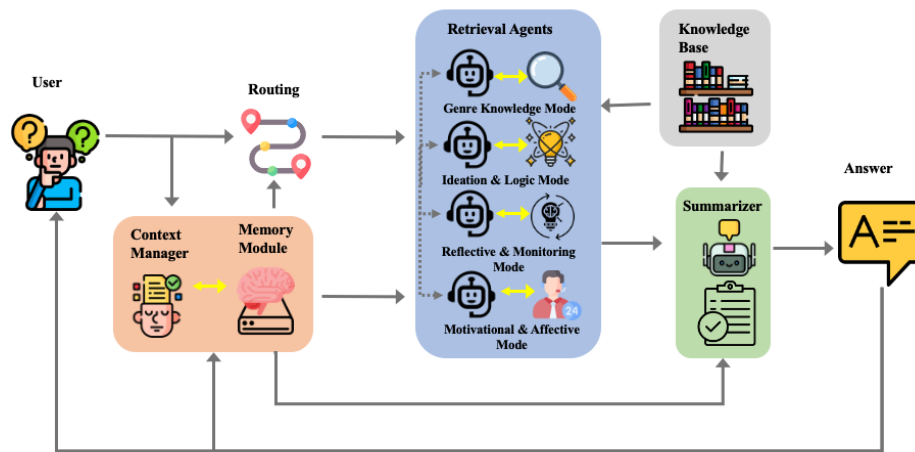


Figure 1 GenAI Agent Workflow Architecture

Agent outputs converge in a Summarizer component that consolidates multi-source responses, enforces content safety filters, and formats the final reply for fluent delivery. A pilot with 12 students (240 interactions) confirmed system viability: response accuracy reached 94%, all outputs passed safety review (100% compliance), and response latency remained under 5 seconds—sufficient to sustain natural turn-taking in dialogue.

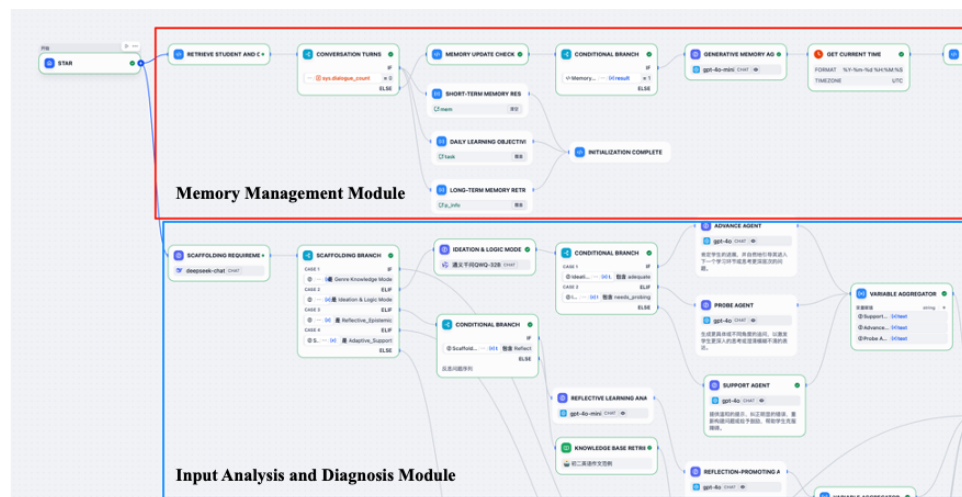


Figure 2 Screenshot of the Dify Workflow Interface (Partial View)

3. Method

3.1. Participants

Eighty eighth-grade students from a single middle school participated in a university-based summer camp program, following parental consent. Eighth graders were selected because this developmental stage is critical for writing proficiency gains in EFL contexts (Bai et al., 2022). Participants were randomly assigned to the experimental group (N = 40; 22 boys, 18 girls) or the control group (N = 40; 23 boys, 17 girls).

3.2. Procedure

The procedure spanned 8 weeks followed by a delayed assessment. Preparation (Week 1): A 40-minute session covering initial training and pre-tests on writing, SRL strategies. Intervention (Weeks 2–7): Over 6 weeks, students completed four recursive writing topics. In 80-minute sessions, both groups followed the SRL cycle (Forethought, Performance, Reflection). The control group relied on conventional teacher-led instruction and static checklists, while the experimental group utilized GenAI scaffolding for personalized goal setting, drafting support, and retrospective analysis. Assessment (Week 8 & Delayed): Immediate post-tests were conducted in Week 8, followed by a delayed post-test one month later to evaluate long-term retention.

3.3. Measures

Students' writing performance was assessed based on their composition scores. An analytical scoring rubric developed by (Hyland, 2003), widely recognized for its comprehensive description of writing proficiency, was adopted for assessment. The rubric covers three dimensions: sentence structure and vocabulary (40 pts), organization and coherence (20 pts), and format and content (40 pts). Two trained senior EFL teachers scored each essay independently; final scores were averaged to ensure inter-rater reliability.

Participants' deployment of SRL strategies in ESL/EFL writing was measured using a writing strategy questionnaire adapted from (Bai et al., 2022). The instrument comprises 17 items divided into four subscales: planning strategies (4 items, $\alpha = 0.87$), text generation strategies (4 items, $\alpha = 0.71$), revising strategies (4 items, $\alpha = 0.83$), and self-monitoring strategies (6 items, $\alpha = 0.82$). Responses were recorded on a 5-point Likert scale, where higher scores indicate a higher frequency of SRL strategy usage.

3.4. Data Analysis

Analyses were conducted in Jamovi 2.4. Normality (Shapiro-Wilk, Q-Q plots) and variance homogeneity (Levene's test) were confirmed, and baseline equivalence verified via independent t-tests. A two-way repeated measures ANOVA (Time $\times$ Group) was the primary analysis ($\alpha = .05$).

5. Results

Normality tests confirmed that the scores for writing performance and the SRL strategy use scale were normally distributed ($p > 0.05$). Table 1 presents the descriptive statistics for all variables and the results of baseline (T0) comparisons between the two groups. Mauchly's test indicated that the assumption of sphericity was met across all dimensions. Given the significant interaction effect observed between Group and Time, simple effects analyses were subsequently conducted.

Table 1. Descriptive statistics and time effects analysis.

	Group	Pre-test	Post-test	Delayed test
Composition Score	Exp Group	64.9 (6.11)	69.3 (5.83)	77.9 (5.28)
	Control Group	64.2 (5.40)	66.2 (5.82)	72.9 (4.51)
Planning Strategy	Exp Group	2.88 (0.68)	3.46 (0.84)	3.26 (0.72)
	Control Group	2.92 (0.85)	3.29 (0.89)	2.94 (0.78)
Text Generation Strategy	Exp Group	2.99 (0.40)	3.62 (0.31)	3.69 (0.38)
	Control Group	3.00 (0.38)	3.29 (0.39)	3.24 (0.34)
Revision Strategy	Exp Group	2.96 (0.38)	3.45 (0.36)	3.28 (0.32)
	Control Group	3.08 (0.37)	3.27 (0.36)	3.16 (0.33)
Self-Monitoring Strategy	Exp Group	3.00 (0.71)	3.51 (0.73)	3.25 (0.62)
	Control Group	3.09 (0.60)	3.17 (0.76)	3.16 (0.66)

5.1. Results on Writing Performance

Regarding writing performance, the mean scores for the experimental and control groups across the three time points (T0, T1, T2) were 64.9 ± 6.11 , 69.3 ± 5.83 , 77.9 ± 5.28 and 64.2 ± 5.40 , 66.2 ± 5.82 , 72.9 ± 4.51 , respectively. Simple effects analysis of Time indicated significant improvements within both groups ($p < 0.001$). However, pairwise comparisons (see Table 3) revealed that the control group showed no significant growth from pre-test (T0) to post-test (T1) ($p > 0.05$), whereas all other time intervals for both groups showed significant gains ($p < 0.01$). Simple effects analysis of Group demonstrated that the experimental group significantly outperformed the control group at T1 ($t = 2.50$, $p = 0.013$) and T2 ($t = 4.00$, $p < 0.001$), with the mean difference widening from 3.11 to 4.97.

Table 2. Two-factor repeated measures ANOVA results.

	Effect	SS	df	MS	F	p	η^2p
Composition Score	Time	4961	2	2480.5	92.98	< .001	0.414
	Group	510	1	509.8	13.5	< .001	0.068
	Group x Time	179	2	89.6	3.36	0.037	0.025
Planning Strategy	Time	8.90	2	4.45	23.81	< .001	0.057
	Group	1.31	1	1.31	0.86	0.358	0.009
	Group x Time	1.34	2	0.67	3.59	0.030	0.009
Text Generation Strategy	Time	11.21	2	5.61	75.00	< .001	0.263
	Group	3.84	1	3.84	14.9	< .001	0.109
	Group x Time	2.15	2	1.07	14.4	< .001	0.064
Revision Strategy	Time	4.63	2	2.31	59.6	< .001	0.137
	Group	0.18	1	0.18	0.61	0.437	0.006
	Group x Time	1.02	2	0.51	13.2	< .001	0.034
Self-Monitoring Strategy	Time	3.39	2	1.69	10.13	< .001	0.031
	Group	0.77	1	0.77	0.72	0.40	0.007
	Group x Time	1.84	2	0.92	5.51	0.005	0.017

5.2. Results on SRL Strategy Use

Analysis of the four SRL sub-dimensions (planning, text generation, revising, and self-monitoring) revealed distinct patterns of improvement. Longitudinally, both groups achieved significant gains in planning, text generation, and revising strategies across the time points ($p < 0.001$). However, a notable divergence occurred in self-monitoring strategies, where only the experimental group demonstrated significant growth ($F = 15.02$, $p < 0.001$), while the control group showed no significant change ($F = 0.43$, $p = 0.652$).

Regarding between-group comparisons, the experimental group demonstrated superior immediate outcomes at the post-test (T1), significantly outperforming the control group in text generation ($t = 3.95$, $p < 0.001$), revising ($t = 2.25$, $p = 0.026$), and self-monitoring ($t = 2.22$, $p = 0.028$), though no significant difference was found in planning ($p > 0.05$). By the delayed post-test (T2), the experimental group's advantage in text generation widened further ($t = 5.33$, $p < 0.001$). However, the performance gaps in revising and self-monitoring narrowed and were no longer statistically significant ($p > 0.05$), aligning with planning strategies which remained non-significant throughout. These findings indicate that the intervention was particularly effective in fostering sustained improvements in text generation strategies.

Table 3. Two-factor repeated measures ANOVA results.

Comparison	Time	Group	Difference	SE	t	p
Composition Score	T0-T1	Exp	4.38	1.17	3.75	0.004
		Control	2.00	1.15	1.73	1.000
	T1-T2	Exp	8.64	1.17	7.39	< .001
		Control	6.77	1.15	5.87	< .001
	T0-T2	EXP	13.03	1.17	11.14	< .001
		Control	8.77	1.15	5.87	< .001
Planning Strategy	T0-T1	EXP	0.58	0.10	5.96	< .001
		Control	0.36	0.10	3.75	0.004

Text Generation Strategy	T1-T2	EXP	-0.20	0.10	-2.03	0.661
		Control	-0.34	0.10	-3.56	0.007
	T0-T2	EXP	0.38	0.10	3.93	0.002
		Control	0.02	0.10	0.19	1.000
	T0-T1	EXP	0.62	0.06	10.04	< .001
		Control	0.29	0.06	4.70	< .001
	T1-T2	EXP	0.07	0.06	1.14	1.000
		Control	-0.04	0.06	-0.71	1.000
	T0-T2	EXP	0.69	0.06	11.18	< .001
		Control	0.24	0.06	3.99	0.002
Revision Strategy	T0-T1	EXP	0.49	0.04	11.06	< .001
		Control	0.19	0.04	4.25	< .001
	T1-T2	EXP	-0.17	0.04	-3.88	0.002
		Control	-0.11	0.04	-2.41	0.257
	T0-T2	EXP	0.32	0.04	7.18	< .001
		Control	0.08	0.04	1.84	1.000
Self-Monitoring Strategy	T0-T1	EXP	0.51	0.09	5.48	< .001
		Control	0.08	0.09	0.85	1.000
	T1-T2	EXP	-0.26	0.09	-2.80	0.087
		Control	-0.01	0.09	-0.10	1.000
	T0-T2	EXP	0.25	0.09	2.68	0.121
		Control	0.07	0.09	0.75	1.000

6. Discussion

The present study demonstrates that integrating GenAI into a Strategy-Based Instruction (SBI) framework can significantly enhance secondary students' writing performance by fostering the acquisition and enactment of writing strategies, rather than merely improving surface-level textual features. Quantitative results showed that while both groups improved over time, the GenAI-supported group not only outperformed the control group at the immediate post-test but also exhibited a widening performance gap at the delayed post-test. This sustained advantage suggests that the observed gains were not transient but reflected deeper changes in students' writing processes.

Crucially, the strongest and most durable effects were observed in students' use of text generation strategies, aligning closely with the Ideation & Logic Mode. This mode addressed a persistent EFL bottleneck—difficulty generating ideas and building coherent argumentation (Pessoa et al., 2017)—by using dialogic questioning to scaffold reasoning rather than supplying text directly. The control group's non-significant T0–T1 gains reflect the inadequacy of delayed, product-focused feedback for this cognitively demanding phase (Morris et al., 2021). Students' own characterization of the agent as a “Socratic tutor” or “co-pilot” signals a shift from passive text reception to active meaning construction—an immediacy and personalization unattainable in large teacher-led classrooms (Tran & Ma, 2025). RAG-grounded genre support further ensured accuracy and contextual alignment, mitigating risks of unreliable guidance (Li et al., 2025).

Self-monitoring emerged as the most distinctive SRL divergence: the experimental group improved significantly ($F = 15.02, p < 0.001$) while the control group showed no meaningful change—a contrast expected given that self-monitoring is among the hardest SRL skills to cultivate conventionally (Azevedo, 2009).

The Reflective & Monitoring Mode drove these self-monitoring gains by requiring students to evaluate drafts against self-set goals rather than receive summative corrections—a form of feed-forward planning rarely sustained in time-constrained classrooms. The absence of between-group differences in planning strategies further suggests that GenAI's value lies in sustaining in-process regulation rather than initiating pre-writing (Su et al., 2023). Notably, students who questioned AI suggestions and cross-checked them against textbooks exhibited epistemic vigilance rather than passivity—evidence that well-designed GenAI can reinforce, rather than erode, learner agency.

7. Conclusion

This randomized controlled trial provides robust evidence that a GenAI-powered dialogue agent embedded within an SBI framework can effectively enhance secondary EFL students' writing performance and self-regulated learning. Students in the experimental group demonstrated significantly higher writing performance than the control group, with the performance gap widening at the delayed post-test. Beyond writing outcomes, the intervention yielded substantive improvements in students' writing processes. The experimental group showed marked and sustained gains in text generation strategies and self-monitoring strategies—two SRL components often difficult to cultivate in conventional classrooms. These results indicate that the intervention did more than improve final texts; it successfully supported students in regulating idea development and monitoring their progress. Ultimately, well-designed GenAI scaffolding can help secondary students transition from dependent users of AI-generated text to active, self-regulated writers.

Despite these encouraging findings, several limitations should be acknowledged. First, the study involved a relatively small sample over an eight-week period. Although significant effects were observed, this duration may be insufficient for writing strategies to become fully automatized. Future longitudinal studies with larger, more diverse samples are needed to examine the long-term transfer and stability of SRL strategy use across different writing tasks. Second, the effective use of the dialogue agent appeared contingent on students' ability to interact with the system. Challenges related to prompt literacy suggest that future GenAI-supported instruction should incorporate explicit training on how to communicate needs effectively to AI systems to maximize the scaffolding benefits. Finally, the present study relied primarily on self-report questionnaires to capture SRL strategy use. While informative, these measures provide only indirect evidence of cognitive processes. Future research should integrate objective behavioral metrics—such as system log data, keystroke analysis, or eye-tracking—to obtain fine-grained insights into learners' real-time regulatory behaviors and cognitive load during AI-supported writing.

References

- Adams, R. (2025, October 15). Pupils fear AI is eroding their ability to study, research finds. *The Guardian*.
<https://www.theguardian.com/technology/2025/oct/15/pupils-fear-ai-eroding-study-ability-research>
- Azevedo, R. (2009). Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: A discussion. *Metacognition and Learning*, 4(1), 87–95. <https://doi.org/10.1007/s11409-009-9035-7>
- Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition – Implications for the design of computer-based scaffolds. *Instructional Science*, 33(5/6), 367–379.
- Bai, B., Wang, J., & Zhou, H. (2022). An intervention study to improve primary school students' self-regulated strategy use in English writing through e-learning in Hong Kong. *Computer Assisted Language Learning*, 35(9), 2265–2290. <https://doi.org/10.1080/09588221.2020.1871030>
- Cerchione, R., Liccardo, G., & Passaro, R. (2026). Artificial knowledge generation: Investigating the revolutionary role of generative AI in knowledge management. *Journal of Innovation & Knowledge*, 11, 100866. <https://doi.org/10.1016/j.jik.2025.100866>

- Del Mario, J., & Tran, H. (2024). A literature review: Analyzing barriers hindering the implementation of self-regulated learning in science classrooms. *Frontiers in Education*, 9. <https://doi.org/10.3389/feduc.2024.1330449>
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Hyland, K. (2003). Genre-based pedagogies: A social response to process. *Journal of Second Language Writing*, 12(1), 17–29. [https://doi.org/10.1016/S1060-3743\(02\)00124-8](https://doi.org/10.1016/S1060-3743(02)00124-8)
- Kim, S., So, H.-J., & Park, K. (2025). Supporting learner agency in collaborative writing with generative AI. *British Journal of Educational Technology*, n/a(n/a). <https://doi.org/10.1111/bjet.70015>
- Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., & Wang, F. L. (2025). Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*, 8, 100417. <https://doi.org/10.1016/j.caeai.2025.100417>
- Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, 9(3), e3292. <https://doi.org/10.1002/rev3.3292>
- Pessoa, S., Mitchell, T. D., & Miller, R. T. (2017). Emergent arguments: A functional approach to analyzing student challenges with the argument genre. *Journal of Second Language Writing*, 38, 42–55. <https://doi.org/10.1016/j.jslw.2017.10.013>
- Puntambekar, S., & Hubscher, R. (2005). Tools for Scaffolding Students in a Complex Learning Environment: What Have We Gained and What Have We Missed? *Educational Psychologist*, 40(1), 1–12. https://doi.org/10.1207/s15326985ep4001_1
- Rubin, J., Chamot, A. U., Harris, V., & Anderson, N. J. (2007). Intervening in the use of strategies. *Language learner strategies*, 30, 29–45.
- Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, 57, 100752. <https://doi.org/10.1016/j.asw.2023.100752>
- Tang, K.-S., & Putra, G. B. S. (2025). Generative AI as a Dialogic Partner: Enhancing Multiple Perspectives, Reasoning, and Argumentation in Science Education with Customized Chatbots. *Journal of Science Education and Technology*. <https://doi.org/10.1007/s10956-025-10240-1>
- Tran, T. T. T., & Ma, Q. (2025). Technology-enhanced self-regulation training: A dynamic training model to facilitate second language Vietnamese learners' self-regulated writing skills. *System*, 130, 103625. <https://doi.org/10.1016/j.system.2025.103625>

Non-homogenizable Vocal Habitus: Podcast as Narrative Practice for Learning Subjectivity and Community Building

Chun-I, Chang^{1*}, Regina, Juchun, Chu²

^{1,2} Institute of Learning Sciences and Technologies, Tsing Hua University, Taiwan

*gingerchung0304@gmail.com

Abstract: *As generative artificial intelligence increasingly reshapes knowledge production, concerns over the homogenization of learning outcomes in higher education continue to grow. Drawing on Bourdieu's theory of practice, this qualitative case study analyzes 15 podcast projects produced by graduate learners. The findings show that through vocal narratives, learners mobilize diverse narrative forms through vocal habitus to transform lived experience into symbolic capital and emotional resonance, and thereby constructing multiple forms of community. This study argues that the significance of podcasting lies in sustaining human-centered meaning-making through embodied vocal expression and social interaction, beyond mere resistance to technological replication. In this context, vocal habitus may be understood as an emerging form of educational capital in AI-mediated learning environments.*

Keywords: generative artificial intelligence, podcast, digital narrative, learning subjectivity, community building

1. Introduction

1.1. Research Background and Questions

While the rapid advancement of generative artificial intelligence (GAI) enhances learning efficiency, it obscures the identification of learners' authentic academic abilities and poses challenges to academic integrity (Bittle & El-Gayar, 2025). Consequently, higher education must re-evaluate learning assessment designs, shifting the focus toward the learning process, personalized performance, and social learning outcomes that foster interpersonal interaction (Kofinas et al., 2025). As a digital medium centered on vocal narratives, podcasts enable learners to manifest embodied expressions, exhibiting distinct human traces resistant to GAI replication.

This study examines a graduate seminar using podcasts as learning outcomes, observing that personal experiences shared through vocal narratives exhibit non-homogenizable qualities. Guided by Bourdieu's (1990) theory of practice, this study conceptualizes these embodied expressions as vocal habitus, exploring how vocal habitus shapes learning subjectivity within podcasts, triggers emotional resonance, and drives community building. This study addresses the following research questions:

RQ1: In higher education contexts, how do learners utilize podcast narratives to exhibit distinct human traces resistant to GAI replication, and to transform life experiences into symbolic capital?

RQ2: How do authentic-voice podcasts trigger emotional resonance and drive community building?

1.2. Research Scope and Limitations

This study was conducted in a graduate seminar where the curriculum encouraged learners to engage in reflective activities based on "what voice can I offer the world?" The objective was to produce authentic-voice podcasts as learning outcomes, cultivating professional self-identity through the sharing of personal experiences. Accordingly, narrative themes were open-ended, allowing learners to record individually or collaboratively in pairs.

Given that most learners have a background in digital education, their narrative styles may be influenced by existing disciplinary cultures. Furthermore, their creative processes might involve strategic adaptations to assessment rubrics and peer-voting mechanisms. Therefore, caution is required when generalizing these findings to other academic disciplines.

Finally, with the rapid advancement of GAI voice synthesis technologies, the irreplaceability of authentic human voices in learning outcomes remains a critical subject for continued verification in future research.

2. Literature Review

2.1. Development and Limitations of Generative Voice Technologies

Recently, by integrating large language models with voice synthesis, GAI can simulate podcast production, automatically generating scripts and spoken content (Ju et al., 2025). However, GAI's generative logic still relies on pre-existing texts and statistical models (Xiao et al., 2025). In contrast, the contextualized decisions inherent in learners' vocal narratives involve experience selection, real-time interaction, and narrative responsibility, all of which exhibit distinct human traces resistant to GAI replication. Thus, despite advances in GAI, authentic vocal narratives remain critical for observing and cultivating learning subjectivity in higher education.

2.2. Vocal Habitus and Podcast Narratives: Perspectives from Bourdieu's Theory of Practice

As GAI exacerbates the homogenization of textual learning outcomes, re-identifying learners' non-homogenizable performances becomes critical. Bourdieu's (1990) theory of practice conceptualizes performance as the interplay of field, capital, and habitus, which is recognizable through embodied forms like linguistic expression (Reay, 2004). Building on this perspective, this study conceptualizes "vocal habitus" as an embodied and socially conditioned disposition toward vocal expression, rooted in lived experience and manifested through features such as intonation, rhythm, and affective tone. These are understood not as deliberate strategies but as embodied dispositions constituting a form of capital. As a vocal narrative medium, podcasts enable learners to transform personal viewpoints into cultural capital. Through vocal features such as intonation, rhythm, and emotional pacing, they convey a strong sense of authenticity, fostering listeners' emotional engagement (Spinelli & Dann, 2019), while simultaneously reflecting accumulated media habitus and becoming recognized as symbolic capital that differentiates learning outcomes. Thus, this study conceptualizes podcasting as a narrative practice of learning subjectivity that answers "who speaks, how they speak, and why they speak," forming the theoretical foundation for subsequent analysis.

3. Research Methodology

3.1. Research Design and Subjects

Using a qualitative case study approach based on Bourdieu's (1990) theory of practice, this research analyzes 15 authentic-voice podcasts created by 23 learners. Assessed as learning outcomes that emphasize authentic first-person narration and impose no thematic or post-production restrictions, these projects organically revealed learners' narrative choices and human traces. As a result, no use of GAI-generated scripts or speech synthesis technologies was identified in the final productions. Procedures included data collection, repeated listening, annotation, and theoretical interpretation. Furthermore, classroom observations and questionnaires enabled triangulation across data sources to comprehensively construct the podcasts' narrative context.

3.2. Research Instruments

Acting as the primary instrument, the researcher triangulated classroom observations and questionnaires. These tools captured class interactions, learners' self-interpretations, and the instructor's assessment criteria, providing essential context for learners' perspectives and the field's logic.

3.3. Data Analysis

At the analytical level, all recordings were systematically annotated, focusing on intonation, narrative structure, and interactional features. Guided by Bourdieu's (1990) theoretical framework, data analysis involved iterative listening, annotation, and interpretation. The first stage of descriptive classification involved repeated listening to identify narrative patterns and annotate acoustic features (e.g., intonation, pauses). For example, prolonged pauses in recounting personal internship experiences were annotated as reflective positioning, indicating self-reflection. Post-production elements (e.g., transition design) were also examined to understand how learners presented their experiences and professional

backgrounds. These audio elements highlight irreplaceable human traces. The second stage of theoretical interpretation applied the concepts of habitus and capital to four representative podcasts selected for narrative diversity and thematic representativeness. As an interpretive qualitative study, coding was conducted by the researcher, with reflexive analysis, triangulation, and member checking involving seven participants to ensure trustworthiness. Only minor clarifications were required, and no substantial disagreements were identified. The third stage of cross-data triangulation integrated questionnaires and observational records to connect learners' interpretations with their lived experiences. The researcher synthesized the data and finalized the findings through reflective writing. All participants provided informed consent, data were anonymized, and analysis was conducted after the course to minimize potential bias.

4. Findings

Among the 15 analyzed podcasts, topics primarily centered on personal interests like learning, daily life, and reviews. Narrative forms ranged from solo reflections designed to “*completely convey what I want*” (S9-5) to dialogic interviews aimed at “*stimulating more perspectives*” (S10-5). Strategies included varying self-disclosure, pauses for critical reflection, and unscripted dialogues described as “*recording a daily conversation with a friend*” (S3-5). Overall, podcast narratives transform personal experiences into symbolic capital and foster three community types:

First, the narratives fostered shared-experience communities. Life narratives about relocating or renting “*easily resonate after publication*” (S6-4), allowing listeners to project their experiences and fostering emotional exchange. Aligning with Reay (2004), educational community value stems from authentic resonance. Second, these practices facilitated identity-positioning communities. Through reflective narratives on procrastination, learners positioned themselves as analysts adopting a “*brutally honest style*” (S19-4). Echoing Bourdieu (1984), expressing personal habitus converts into symbolic capital via identity recognition, establishing peer-negotiated academic positions. Third, the podcasts gave rise to parenting-education communities. Incorporating parent-child dialogues transforms family life into a public narrative, perceived as “*very novel*” (S20-4). This successfully converts parenting's emotional legitimacy and educational mission into symbolic capital.

Across these forms, interactional dynamics such as conversational turn-taking, shared laughter, and co-constructed storytelling function as markers of relational intimacy. Co-recording actively reshapes relationships, demonstrating that podcasting is a generative practice that produces, rather than merely expresses, social ties.

5. Research Conclusions

5.1. Responding to RQ1: Exhibiting Human Traces and Accumulating Symbolic Capital via Vocal Habitus

Through topic selections grounded in personal experience, learners demonstrate learning subjectivity resistant to GAI replication. By employing diverse narrative forms, they transform social relationships into actions bearing distinct human traces. In practice, learners draw on a personalized vocal habitus, using embodied expression to interpret issues and assume responsibility for voicing, thereby enacting learning subjectivity by answering “who speaks, how they speak, and why they speak.” Accumulated cultural and technical habitus further translates into symbolic capital. However, if learners merely read aloud GAI-generated texts, even authentic-voice podcasts become conduits for transmission, failing to demonstrate learning subjectivity. Likewise, reliance on AI-generated voice, even with self-authored content, may attenuate embodied expression. Only when learners take ownership of narrative choices and active interpretation do podcasts function as a narrative practice of learning subjectivity.

5.2. Responding to RQ2: Driving Community Building via Emotional Resonance

Findings reveal that the interplay of topic selections, narrative forms, and narrative strategies triggers distinct forms of community resonance. Shared-experience narratives promote emotional recognition and companionable responses; identity-positioning narratives solidify academic viewpoints and role recognition; while parenting-education narratives spark discussions on parental values. These interactions demonstrate that podcast creation is not merely knowledge transmission, but a catalyst for affective learning that continuously generates diverse learning communities. Countering

the progressive homogenization of textual outcomes, the vocal narratives of podcasts preserve embodied cues, enabling learners to proactively respond to learning contexts and drive community building.

5.3. Conclusion: Redefining Subjectivity in the GAI Era

Against the backdrop of rapid advancements in GAI, the core challenge facing higher education lies not only in how to integrate technology, but in how to preserve human agency, judgment, and responsibility in learning. This study argues that learning subjectivity should be understood as a contextualized process of meaning-making, enacted through embodied expression and social interaction. In this regard, the significance of podcasting lies in its capacity to connect lived experience, expressive practice, and social recognition through vocal habitus. It foregrounds the conditions under which learners construct meaning, take ownership of their narratives, and engage with others. Therefore, this study proposes that vocal habitus may be understood as an emerging form of educational capital in AI-mediated learning environments, highlighting the value of learning as an embodied, relational, and meaning-making practice.

6. References

- Bittle, K., & El-Gayar, O. (2025). Generative AI and academic integrity in higher education: A systematic review and research agenda. *Information*, 16(4), 296.
- Bourdieu, P. (1984). *Distinction: A social critique of the judgement of taste* (R. Nice, Trans.). Routledge & Kegan Paul.
- Bourdieu, P. (1990). *The logic of practice*. Stanford University Press.
- Ju, Z., Yang, D., Yu, J., Shen, K., Leng, Y., Wang, Z., Tan, X., Zhou, X., Qin, T., & Li, X. (2025). *MoonCast: High-quality zero-shot podcast generation*. arXiv.
- Kofinas, A. K., Tsay, C. H. H., & Pike, D. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. *British Journal of Educational Technology*, 56, 2522–2549.
- Reay, D. (2004). ‘It’s all becoming a habitus’: Beyond the habitual use of habitus in educational research. *British Journal of Sociology of Education*, 25(4), 431–444.
- Spinelli, M., & Dann, L. (2019). *Podcasting: The audio media revolution*. Bloomsbury Academic.
- Xiao, Y., He, L., Guo, H., Xie, F. L., & Lee, T. (2025). Podagent: A comprehensive framework for podcast generation. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 23923–23937). Association for Computational Linguistics.

Examining the Factors Affecting Learning Engagement in Self-Directed Learning Using Mobile Technology in Higher Education

Yuzhi Lai^{1,2*}, Nan Pu¹

¹ Hefei University of Technology

² Leiden University

* yzlai@hfut.edu.cn

Abstract: In language learning, university students are increasingly adopting out-of-class self-directed learning using mobile technology. As there is no teacher involved, students' learning engagement is an essential indicator for effective and successful self-directed learning. This study examined how university students' three dimensions of learning engagement were influenced by their self-efficacy, self-regulation skills, perceived usefulness and teachers support in self-directed language learning using mobile technology. The participants included 325 language students in different disciplines attending Chinese universities. Multiple regression analysis was conducted to analyze these data. The results revealed that cognitive engagement was affected by self-regulation skills, self-efficacy and perceived usefulness. Behavioral engagement was related to teacher support and self-regulation skills. Emotional engagement was predicted by self-regulation skills, self-efficacy and perceived usefulness.

Keywords: self-directed learning, mobile technology, language learning, learning engagement, higher education

1. Introduction

Mobile technology has provided new ways for foreign language education. Characterized by its individuality, portability and social connectivity, mobile technology has made language learning adaptable, seamless, movable and collaborative (Sung et al., 2015). For example, in the United States, Duolingo, a popular mobile application focusing on foreign language learning, has been used by university students from various backgrounds to develop their Turkish language competence based on its wide range of resources and social networking features (Loewen et al., 2019). In Australia, Chinese Island (CI), an immersive 3D virtual environment, was introduced to effectively engage Chinese language learning students, facilitate their authentic language use, and enhance their learning experience (Wang et al., 2021). In China, WeChat, a multi-purpose mobile app, has been used to help students to develop their pronunciation by receiving feedback from automatic speech recognition (ASR) and/or peers (Dai & Wu, 2021). Nevertheless, research on mobile-assisted language learning (MALL) has been mostly about teacher-led language learning to date, rather than student-initiated use of mobile apps to learn outside class (An et al., 2020; Nami, 2020). In higher education, moreover, there is not much time for students to learn and master foreign languages as part of the language curriculum and, in some countries, students are not given sufficient in-class time to practice their foreign languages to ensure language acquisition (Liu, 2020; Nami, 2020; Trinder, 2017). In order to address this issue, many researchers (e.g., Lai et al., 2022a; Pramesti, 2020) have suggested that students adopt out-of-class and self-directed language learning with the assistance of mobile technology to gain more exposure to authentic foreign languages, thus further enhancing their language competence. This out-of-class and self-directed language learning using mobile technology means that learners take responsibility for and control of their own language learning process, including deciding what and how they learn with the assistance of mobile technology (Merriam & Bierema, 2013). More specifically, students utilize mobile apps such as Twitter, YouTube, Facebook, and/or foreign language-specific apps such as Duolingo and Google Translate to create their own language learning environment.

Research on self-directed language learning using mobile technology has offered insights into students' learning experiences (e.g., Lai et al., 2018), learning strategies (Lai et al., 2022b), and acceptance and adoption of the use of mobile

technology in self-directed language learning (Lai et al., 2022a). However, research focusing on learning outcome such as engagement has been limited up to now. Learning engagement is consistently identified as a key predictor of learner outcomes in online environment (Brown et al., 2022). It contributes to active involvement and persistence in the process and thus relates to successful learning performance. Researchers have valued learning engagement in the online learning environment for years (Alemayehu & Chen, 2021; Jung & Lee, 2018). Therefore, it is a priority to identify the factors that affect learners' engagement in self-directed learning, thus providing self-directed learners, facilitators and software developers with meaningful feedback to enhance the quality of learning (Heo et al., 2022).

2. Theoretical framework

Engagement in self-directed language learning using mobile technology refers to the ongoing time, effort, and energy that learners put into this form of independent learning to achieve their goals (Kuo et al., 2021; Lan & Hew, 2020). Student engagement has been conceptualized and operationalized as a multidimensional construct which can be broken down into behavioral, emotional, and cognitive engagement. Behavioral engagement means students participating in a language learning activity such as taking notes (Fredricks et al., 2004). Emotional engagement refers to students' affective reactions toward their learning experience (Kuo et al., 2021). These emotional feelings include enthusiasm, enjoyment, interest, fun, boredom, and feelings of depression (Deng et al., 2020). Cognitive engagement means that learners put effort into understanding and mastering the language knowledge and skills (Skinner et al., 2008). Learners' and teachers' perspectives are employed in this study to explore the factors that affect self-directed learners' engagement in the out-of-class learning context.

Factors from learners' perspective includes self-regulation skills, self-efficacy and perceived usefulness. Self-regulation skills are defined as the ability of individuals to actively regulate their own learning tasks or behaviors from metacognitive, motivational, and behavioral perspective (Zimmerman, 1989). Self-efficacy defines learners' self-perceived capability to master the functions of mobile technology and systems and to learn well using the mobile technology. Perceived usefulness reflects learners' perceptions of the advantages of mobile technology (Lin et al., 2016). In addition, the factor from teachers' perspective is teacher support. Teacher support refers to the facilitating support that teachers can provide to self-directed learners with the aim of enhancing their learning experience. According to Garrison and Cleveland-Innes (2005), interactions between teachers and learners influence learners' engagement more strongly than interactions between learners in online learning. Similarly, in the context of mobile-assisted self-directed language learning, teachers are considered to play a critical role (Lai, 2015; Lai et al., 2022b). Unfortunately, research has found that teachers tend to perceive themselves as having a limited influence on students' autonomous learning outside the classroom (Chan, 2003; Toffoli & Sockett, 2015). Thus, teachers' roles should be considered and empirically examined in the discussion of learners' autonomous learning behaviors outside the classroom (Lai, 2015).

Overall, this study aimed to investigate how cognitive, behavioral and emotional engagement are related to learners' factors (self-regulation skills, perceived usefulness, and self-efficacy), and teacher factors (teacher support) in self-directed language learning using mobile technology. The findings may provide implications to make self-directed learners more engaged, further achieving effective and successful autonomous language learning with technology.

3. Method

3.1. Participants and procedure

A total of 446 self-directed English language learners were recruited, forming a volunteer multidisciplinary sample from Chinese universities. We selected the eligible students by their answer to the first item at the beginning of the questionnaire ("Have you ever chosen to learn English language by yourself?"). We chose the convenience sampling method with an online questionnaire to obtain the data. To recruit as many participants as possible, we distributed a hyperlink of the online questionnaire from Qualtrics via social media platforms. The purpose of the anonymous link was to encourage participants to express their preferences honestly and openly. Participants were informed of the objectives

of the questionnaire and how these data would be used and were asked to give their consent at the end of the questionnaire. Completing the survey took participants around 5-10 minutes. Research clearance was obtained from the ethics committee of the ICLON Research Ethics Committee.

3.2. Instruments

The questionnaire items related to (1) social-demographic questions such as gender, educational level, discipline, and the English exams that they had passed; (2) scales about self-regulation skills, self-efficacy, perceived usefulness, teacher support, cognitive engagement, behavioral engagement and emotional engagement in self-directed language learning using mobile technology. The items were measured on a five-point Likert scale where “1” represented “strongly disagree” and “5” indicated “totally agree”. All the instruments used in this study came from existing validated scales. To fit our research context, we added a contextual description “When self-studying English language...” to define the context-specific feature. To ensure face validity, the questionnaire was sent to two academic professors and ten university students for internal review. Based on their feedback and suggestions, we modified the wordings to prevent semantic bias.

3.3. Data analysis

To answer the research questions, IBM SPSS 25 is utilized as the statistical tool for data analysis. The multiple regression analysis will be employed to examine how teacher support, learners’ self-regulation skills, perceived usefulness, and self-efficacy are related to three dimensions of learning engagement in self-directed learning using mobile technology.

4. Result and discussion

4.1. Descriptive statistics

The data in this study were collected from August to November 2021. A total of 446 respondents visited the questionnaire website and 352 respondents successfully completed the questionnaire (completion rate was 78.9%). Among the completed questionnaires, there were 28 respondents who stated that they had no experience of self-directed English learning; that left a total of 325 observations making up the study. Of these, female participants accounted for 52%, and male participants represented 48%. Other demographic information is shown in Table 1.

Regarding the activities that they participated in (see Table 2), over 60% of participants used mobile technology to listen to/watch foreign radios/videos, look up and recite vocabulary through mobile apps like Youdao Dictionary, Baicizhan, and search for information through websites or apps. In addition, in the option of “Others”, several participants indicated that they did English tests, watched online English courses, played English games, and listened to English songs. Table 3 shows the mean scores of each scale, standard deviations, number of items in each scale and their ranges, and Cronbach’s coefficient alphas. Correlations between these variables are shown in Table 4.

Table 1.*Frequency and percentage of participants by gender, education, and major.*

Demographic factors	<i>n</i>	%
<i>Gender</i>		
Male	156	48
Female	169	52
<i>Education</i>		
Undergraduate	249	76.62
Postgraduate	76	23.38
<i>Major</i>		
Language	108	33.23
Non-language	217	66.77
<i>Language proficiency</i>		
Low level	121	37.23
High level	204	62.77

Table 2.*Activities that students participated in.*

Activities	Frequency	Percentage (%)
Listen to/watch foreign radios/videos	251	77.23
Communicate with native speakers or peers online	116	35.69
Read online newspapers	160	49.23
Search for information through websites or apps (like Baidu, Zhihu)	213	65.54
Look up and recite vocabulary (like Youdao Dictionary, Baicizhan)	246	75.69
Using language-specific learning apps (Like Qubeiyin, , etc.)	109	33.54
Others	6	1.85

Table 3.*Mean, standard deviation, and reliability information of each scale.*

Scales	Items	Range	M	SD	<i>a</i>
SDL	5	1-5	3.7754	.77117	.865
SE	7	1-5	4.3754	.62894	.931
OP	7	1-5	4.0677	.63675	.855
TS	6	1-5	3.8841	.77096	.888
BE	3	1-5	3.8728	.75647	.692
CE	3	1-5	4.1959	.62929	.773
EE	3	1-5	4.0328	.66276	.783

Table 4.*Correlations between variables.*

	SDL	TS	SE	OP	BE	CE	EE
SDL	1						
TS	.538**	1					
SE	.367**	.384**	1				
OP	.610**	.468**	.639**	1			
BE	.624**	.509**	.331**	.470**	1		
CE	.479**	.381**	.602**	.586**	.578**	1	
EE	.500**	.398**	.460**	.523**	.515**	.643**	1

4.2. The differences between majors

Independent sample t-test was used to examine whether students with different genders, educational levels, majors and language proficiency differed in their three dimensions of learning engagement, self-regulation skills, perceived usefulness, and self-efficacy, the teacher support they received in the process of self-directed language learning using mobile technology. The results showed only students with different majors had a statistical difference on behavioral engagement ($t = 4.554$, $p < 0.001$), emotional engagement ($t = 2.045$, $p < 0.05$), self-regulation skills ($t = 4.371$, $p < 0.01$), teacher support ($t = 4.960$, $p < 0.01$), and self-efficacy ($t = -2.944$, $p < 0.01$), not on cognitive engagement and perceived usefulness.

4.3. The effect of learners' and teacher's factors on three dimensions of learning engagement

Multiple regression analysis was performed to determine whether learners' and teacher's variables predict learners' cognitive, behavioral and emotional engagement. In terms of cognitive engagement, the regression results indicate that the overall model with the four independent variables significantly predicts learners' cognitive engagement, $R^2 = .459$, $R^2_{adj} = .452$, $F = 67.814$, $p < .000$. This model accounts for 45.9% of the variance in cognitive engagement. After reviewing the beta weights, it was determined that only self-regulation skills, self-efficacy and perceived usefulness significantly contributed to this model. Teacher support was not a significant predictor ($p > .05$) in the model, as shown in Table 5.

Table 5.*Coefficients estimates of the model with four predictors and cognitive engagement as an outcome variable.*

Variables	B	SE	Beta	t	Sig.	R ²	Adj.R ²	F
1(constant)	.976	.200		4.878	.000			
TS	.025	.041	.031	.608	.544			
SDL	.160	.046	.195	3.499	.001			
SE	.387	.054	.387	7.155	.000			
OP	.203	.062	.205	3.252	.001	.459	.452	67.814***

Regarding behavioral engagement, the regression results indicate that the overall model with the four independent variables significantly predicts learners' behavioral engagement, $R^2 = .438$, $R^2_{adj} = .431$, $F = 62.253$, $p < .000$. This model accounts for 43.8% of the variance in behavioral engagement. After reviewing the beta weights, it was determined that self-regulation skills and teacher support significantly contributed to this model. Self-efficacy and perceived usefulness were not significant predictors ($p > .05$) in the model, as shown in Table 6.

Table 6.

Coefficients estimates of the model with four predictors and behavioral engagement as an outcome variable.

Variables	B	SE	Beta	t	Sig.	R ²	Adj.R ²	F
1(constant)	.838	.245		3.415	.001			
TS	.217	.050	.221	4.301	.000			
SDL	.441	.056	.449	7.888	.000			
SE	.044	.066	.037	.665	.507			
OP	.083	.076	.070	1.082	.280	.438	.431	62.253***

Concerning emotional engagement, the regression results indicate that the overall model with the four independent variables significantly predicts learners' emotional engagement, $R^2 = .360$, $R^2_{adj} = .352$, $F = 45.076$, $p < .000$. This model accounts for 36.0% of the variance in emotional engagement. After reviewing the beta weights, it was determined that self-regulation skills, self-efficacy and perceived usefulness significantly contributed to this model. Teacher support was not a significant predictor ($p > .05$) in the model, as shown in Table 7.

Table 7.

Coefficients estimates of the model with four predictors and emotional engagement as an outcome variable.

Variables	B	SE	Beta	t	Sig.	R ²	Adj.R ²	F
1(constant)	1.127	.229		4.916	.000			
TS	.077	.047	.089	1.630	.104			
SDL	.222	.052	.259	4.262	.000			
SE	.220	.062	.209	3.547	.000			
OP	.198	.071	.190	2.774	.006	.360	.352	45.076***

5. Conclusion and significance

This study identified the impact of learners' self-regulation skills, self-efficacy, perceived usefulness, and teacher support on three dimensions of learning engagement. Learners' self-regulation skills were found to be significantly related to all three dimensions of learning engagement. Self-efficacy and perceived usefulness were significantly associated with cognitive and emotional engagement, but not behavioral engagement. Teacher support made a significant contribution only to behavioral engagement, not cognitive and emotional engagement.

Both theoretical and practical implications are presented. As a contribution to theory, the study proposed and verified a conceptual model of learners' engagement from previous studies in the context of self-directed learning using technology. The comprehensive model offers a better understanding of the mechanism of how learners-related and teacher-related factors influence learners' three dimensions of engagement. In practical terms, considering the critical impact of self-regulation skills on all three dimensions of learning engagement, it is a priority to improve learners' self-regulation skills before starting a self-directed learning process. Although support from teachers was only related to behavioral engagement, teachers' awareness and roles of supporting self-directed learners still need to be raised.

Funding

This work was supported by Humanities and Social Science Fund of Ministry of Education of China [grant number 25YJC880038] and the Fundamental Research Funds for the Central Universities.

References

- Alemayehu, L., & Chen, H.-L. (2021). The influence of motivation on learning engagement: the mediating role of learning self-efficacy and self-monitoring in online learning environments. *Interactive Learning Environments*, 1–14. <https://doi.org/10.1080/10494820.2021.1977962>
- An, Z., Wang, C., Li, S., Gan, Z., & Li, H. (2020). Technology-assisted self-regulated English language learning: Associations with English language self-efficacy, English enjoyment, and learning outcomes. *Frontiers in Psychology*, 11, 558466–558466. <https://doi.org/10.3389/fpsyg.2020.558466>
- Brown, A., Lawrence, J., Basson, M., & Redmond, P. (2022). A conceptual framework to enhance student online learning and engagement in higher education. *Higher Education Research & Development*, 41(2), 284–299.
- Chan, V. (2003). Autonomous language learning: The teachers' perspectives. *Teaching in Higher Education*, 8(1), 33–54. <https://doi.org/10.1080/1356251032000052311>
- Dai, Y., & Wu, Z. (2021). Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: A mixed-methods study. *Computer Assisted Language Learning*, 1–24. <https://doi.org/10.1080/09588221.2021.1952272>
- Deng, R., Benckendorff, P., & Gannaway, D. (2020). Linking learner factors, teaching context, and engagement patterns with MOOC learning outcomes. *Journal of Computer Assisted Learning*, 36(5), 688–708. <https://doi.org/10.1111/jcal.12437>
- Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1), 59–109. <https://doi.org/10.3102/00346543074001059>
- Garrison, D. R., & Cleveland-Innes, M. (2005). Facilitating cognitive presence in online learning: Interaction is not enough. *The American Journal of Distance Education*, 19(3), 133–148. https://doi.org/10.1207/s15389286ajde1903_2
- Heo, H., Bonk, C. J., & Doo, M. Y. (2022). Influences of depression, self-efficacy, and resource management on learning engagement in blended learning during COVID-19. *The Internet and Higher Education*, 100856.
- Jung, Y., & Lee, J. (2018). Learning engagement and persistence in massive open online courses (MOOCs). *Computers & Education*, 122, 9–22. <https://doi.org/10.1016/j.compedu.2018.02.013>
- Kuo, T. M., Tsai, C.-C., & Wang, J.-C. (2021). Linking web-based learning self-efficacy and learning engagement in MOOCs: The role of online academic hardiness. *The Internet and Higher Education*, 51, 100819. <https://doi.org/10.1016/j.iheduc.2021.100819>
- Lan, M., & Hew, K. F. (2020). Examining learning engagement in MOOCs: A self-determination theoretical perspective using mixed method. *International Journal of Educational Technology in Higher Education*, 17(1), 1–24. <https://doi.org/10.1186/s41239-020-0179-5>
- Lai, C. (2015). Modeling teachers' influence on learners' self-directed use of technology for language learning outside the classroom. *Computers & Education*, 82, 74–83. <https://doi.org/10.1016/j.compedu.2014.11.005>
- Lai, C., Hu, X., & Lyu, B. (2018). Understanding the nature of learners' out-of-class language learning experience with technology. *Computer Assisted Language Learning*, 31(1–2), 114–143. <https://doi.org/10.1080/09588221.2017.1391293>
- Lai, Y., Saab, N., & Admiraal, W. (2022a). Learning strategies in self-directed language learning using mobile technology in higher education: A systematic scoping review. *Education and Information Technologies*, 27(6), 7749–7780. <https://doi.org/10.1007/s10639-022-10945-5>

- Lai, Y., Saab, N., & Admiraal, W. (2022b). University students' use of mobile technology in self-directed language learning: Using the integrative model of behavior prediction. *Computers & Education*, 179, 104413. <https://doi.org/10.1016/j.compedu.2021.104413>.
- Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. *ReCALL*, 31(3), 293-311. <https://doi.org/10.1017/S0958344019000065>
- Lin, H.-H., Lin, S., Yeh, C.-H. and Wang, Y.-S. (2016), Measuring mobile learning readiness: Scale development and validation, *Internet Research*, Vol. 26 No. 1, pp. 265-287. <https://doi.org/10.1108/IntR-10-2014-0241>
- Liu, N. (2020). A study of autonomous learning model of college English based on mobile network. *Education Teaching Forum*, 28.
- Merriam, S. B., & Bierema, L. L. (2013). *Adult learning: Linking theory and practice*. John Wiley & Sons.
- Nami, F. (2020). Educational smartphone apps for language learning in higher education: Students' choices and perceptions. *Australasian Journal of Educational Technology*, 36(4), 82-95. <https://doi.org/10.14742/ajet.5350>
- Pramesti, A. S. (2020). Students' perception of the use of mobile application Duolingo for learning English. *International Journal of Scientific & Technology Research*, 9(1).
- Skinner, E., Furrer, C., Marchand, G., & Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of Educational Psychology*, 100(4), 765-781. <https://doi.org/10.1037/a0012840>.
- Sung, Y. T., Chang, K. E., & Yang, J. M. (2015). How effective are mobile devices for language learning? A meta-analysis. *Educational research review*, 16, 68-84.
- Toffoli, D., & Sockett, G. (2015). University teachers' perceptions of online informal learning of English (OILE). *Computer Assisted Language Learning*, 28(1), 7-21. <https://doi.org/10.1080/09588221.2013.776970>
- Trinder, R. (2017). Informal and deliberate learning with new technologies. *ELT Journal*, 71(4), 401-412. <https://doi.org/10.1093/elt/ccw117>
- Wang, Y., Grant, S., & Grist, M. (2021). Enhancing the learning of multi-level undergraduate Chinese language with a 3D immersive experience - An exploratory study. *Computer Assisted Language Learning*, 34(1-2), 114-132. <https://doi.org/10.1080/09588221.2020.1774614>
- Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81(3), 329-339. <https://doi.org/10.1037/0022-0663.81.3.329>

Roles and Strategies of Generative AI in Multimodal English Instruction: A Qualitative Study of AI-Native English Teachers' Experiences

Chiao-Sin, Chung¹, Regina, JuChun, Chu²

^{1,2} Institute of Learning Sciences and Technologies, Tsing Hua University, Taiwan

*chu@mx.nthu.edu.tw

Abstract: *With the rapid integration of generative artificial intelligence (Gen AI) into English as a Foreign Language (EFL) classrooms, teachers are increasingly expected to design learning experiences that support students' multimodal literacy development. While prior research has primarily focused on teachers' attitudes toward AI adoption or the effectiveness of AI-supported systems, less attention has been paid to how teachers incorporate Gen AI into instructional design decision-making. This qualitative study examines how AI-native English teachers, who adopted Gen AI early in their careers, integrate Gen AI into multimodal instructional design. Guided by the multimodal learning processes, the study draws on interviews with three teachers and employs thematic analysis. The findings suggest that Gen AI functions primarily as a design support. Teachers use Gen AI to support affective engagement, generate and refine tasks, support scaffolding activities, and develop assessment criteria, while exercising professional judgment to evaluate the appropriateness, accuracy, and contextual relevance of AI-generated outputs. The study identifies challenges, including concerns about excessive reliance on AI, difficulties in aligning outputs with learners' proficiency levels, and the need for ongoing teacher mediation. Overall, the findings highlight teachers' evolving roles and offer implications for English teacher education and professional development.*

Keywords: generative AI, multimodal instructional design, English language teaching, AI-native English teachers

1. Introduction and Theoretical Background

With the increasing integration of digital technologies into everyday communication, language learning environments have become inherently multimodal. Meaning making now extends beyond written language to include multiple semiotic resources (Cope & Kalantzis, 2015). Accordingly, teachers are expected to act as designers of learning by orchestrating these resources to support learners' understanding (Lim & Tan-Chia, 2022). However, multimodal pedagogy in EFL classrooms is often accompanied by challenges related to pedagogical clarity and assessment, particularly in evaluating students' productions (Lim et al., 2022), highlighting increasing instructional design responsibilities for teachers.

In EFL contexts, the multimodal turn requires teachers to make decisions about mode selection, sequencing, and scaffolding aligned with learners' levels and goals (Lim & Tan-Chia, 2022). These demands are especially challenging for novice teachers, who are still developing instructional strategies and professional judgment. In this study, AI-native English teachers are defined as teachers who encountered and adopted Gen AI tools during teacher education or the early stages of their teaching careers, and whose practices have developed alongside the rapid emergence of Gen AI since late 2022. Examining these teachers provides insight into how emerging technologies intersect with early-career professional learning and instructional decision making.

The rapid development of Gen AI has introduced new possibilities for language education. Prior studies have highlighted its potential for personalized learning, intelligent tutoring, and automated assessment (Kamalov et al., 2023). However, existing research has largely focused on general applications or teachers' attitudes toward AI adoption, with limited attention to how teachers employ Gen AI in multimodal instructional design. Little is known about how AI-native English teachers incorporate Gen AI into instructional practices, highlighting the need for further examination.

This study is guided by the following research questions:

RQ1. How do AI-native English teachers utilize Gen AI in their instructional design practices for multimodal learning?

RQ2. How do they integrate Gen AI into their instructional design decision-making for multimodal learning?

RQ3. What challenges and professional considerations do they encounter when using Gen AI for instructional design?

2. Method

This study adopted a qualitative research design to explore how AI-native English teachers in Taiwan integrate Gen AI into instructional design decision-making. The approach aimed to capture teachers' experiences, pedagogical reasoning, and professional judgments rather than measuring predefined outcomes or frequencies of use.

Participants were AI-native English teachers in Taiwan who encountered and adopted Gen AI during teacher education or the early stages of their teaching careers. All were early-career teachers with no more than three years of teaching experience. Three participants were recruited through purposive sampling and provided written informed consent. While the sample size was small, the study aimed to generate in-depth, context-specific insights.

Data were collected through semi-structured interviews focusing on instructional planning, task design, and professional considerations in using Gen AI. Interviews were conducted in Mandarin Chinese, lasted 40–60 minutes, and were audio-recorded. Participants also provided instructional artifacts, such as lesson plans and teaching materials, to support data interpretation and enhance the credibility.

The interview data were analyzed using thematic analysis, following established qualitative procedures (Braun & Clarke, 2006). A theory-informed thematic approach was adopted. First, all transcripts were read repeatedly to achieve familiarity. The data were organized into broad themes, including the multimodal learning processes of encountering, exploring, evaluating, and expressing, as well as teachers' views on Gen AI integration and implications for teacher education. Within each theme, inductive coding was conducted to identify meaning units concerning instructional design practices, uses of Gen AI, and pedagogical reasoning. Interview data were further examined in relation to instructional artifacts. Throughout the analysis, the multimodal learning processes were treated as interpretive themes.

To enhance trustworthiness, data triangulation was conducted using interview transcripts and artifacts. Peer coding was employed, and discrepancies were resolved through discussion. Inter-coder reliability improved from Cohen's kappa 0.62 to 0.79 after refinement.

3. Results

Three AI-native English teachers participated in this study: Teacher A (TA), Teacher B (TB), and Teacher C (TC). TA taught elementary learners in a private cram school with one year of experience, while TB and TC had about six months of experience in senior high school and elementary school contexts, respectively. Interview excerpts are cited as TA-I, TB-I, and TC-I, and instructional artifacts are cited as TA-M, TB-M, and TC-M.

The findings indicate that teachers primarily used Gen AI during lesson preparation and instructional design rather than in classroom instruction. Gen AI was perceived as a supportive tool for idea generation, activity design, and content organization, as well as for refining existing materials, without replacing professional judgment. Its use was selective and context-dependent, based on students' proficiency, learning needs, and instructional objectives. Even when used frequently, teachers retained final decision-making authority over content, sequencing, and assessment. Overall, Gen AI functioned as a design support, while teachers remained central to instructional planning.

3.1. Encountering: Gen AI as Support for Context Building and Learning Engagement

At the encountering stage, teachers used Gen AI to support context building and students' engagement. They generated contextualized texts, picture books, story backgrounds, and instructional scenarios aligned with students' everyday experiences. Such content was primarily used to stimulate motivation (TC-I, TC-M), reduce cognitive load and support initial comprehension rather than serve as complete instructional content.

Teachers also generated images to create multimodal materials combining visual and verbal representations. Despite its efficiency, teachers consistently filtered and revised the generated content. Regarding image generation, TA noted that producing images that fully matched expectations often required repeated prompt revisions. Consequently, teachers

sometimes used Gen AI to suggest suitable image types (TA-I, TA-M). Overall, Gen AI supported engagement and context construction, while teachers retained control over final decisions.

3.2. Exploring: Gen AI as Design Support for Language Form Understanding and Sentence-Level Scaffolding

At the exploring stage, Gen AI supported students' understanding of language forms and the design of sentence-level scaffolding. Teachers relied heavily on Gen AI to generate ideas for sentence-based activities, adjust instructional sequences, or reorganize content to better align with students' proficiency and learning needs (TA-M, TB-I, TC-I).

Teachers also used Gen AI to generate sentence frames, mnemonic cues, or gesture-supported prompts to help students gradually understand and internalize language forms (TC-M). Such scaffolding played a central role in supporting comprehension and practice, with AI enabling rapid generation of diverse examples for repeated exposure.

3.3. Evaluating: Limited Use of Gen AI and Teachers' Professional Judgment in Assessment Design

At the evaluating stage, the focus extends beyond measuring learning outcomes to supporting students' understanding of meaning, stance, and critical thinking. The findings indicate that some teachers recognized this stage as an opportunity to develop students' understanding of multiple perspectives and used Gen AI to support perspective comparison and meaning making.

One teacher described an environmental lesson in which *"many of the causal relationships in the text were actually very difficult for students to understand"* (TB-I). To address this, TB used Gen AI to generate texts representing different perspectives for role-play activities, and to generate listening audio and worksheets to support comprehension of key information and relationships. This design encouraged students to compare perspectives rather than focus on a single text, enabling a deeper understanding of the issue. The teacher further noted that *"after that activity, students' motivation increased"* (TB-I), which was also reflected in improved performance in their final exam.

Within this stage, Gen AI functioned as a tool for generating multi-perspective materials. It broadened students' interpretive perspectives and supported judgment and reflection. However, such evaluative designs centered on critical judgment remain relatively uncommon in everyday English teaching, where emphasis often lies on language form acquisition and reading comprehension.

3.4. Expressing: Gen AI in Supporting the Design of Tasks and Conditions for Student Expression

At the expressing stage, the focus lies in supporting students' production, presentation, and reflection. The findings indicate that teachers did not use Gen AI to replace students' creative work, but to design clearer conditions for expression and feedback, supporting reflective learning.

One teacher noted that they *"liked using Gen AI to create peer assessment forms"* (TB-I) during oral presentations and writing activities. Gen AI was used to generate assessment rubrics and peer evaluation sheets, making criteria more specific and transparent while reducing workload. *"Because generating rubrics has become easier, I'm more willing to give students opportunities to present in front of the class"* (TB-I). The teacher further reflected that *"the self-assessment part becomes much simpler, so I'm more willing to include these activities in my classroom"* (TB-I), increasing opportunities for peer feedback and discussion. Gen AI also supported students' reflection and revision in writing tasks. In one activity, students used AI to check grammar and generate images based on their written descriptions (TB-I). By comparing generated images with their intended ideas, students identified gaps in expression and revised their texts accordingly through iterative interaction with the AI tool.

Overall, Gen AI functioned as a mediating tool that supported reflection, revision, and structured feedback, rather than replacing students' production. It enabled more explicit evaluation criteria and increased student engagement and learning ownership.

4. Discussion and Conclusion

For AI-native English teachers in this study, Gen AI enhanced design efficiency without replacing professional judgment. Teachers positioned Gen AI as a supportive resource and continuously evaluated outputs, reaffirming their

central role in instructional decision-making. For early-career teachers, Gen AI also reduced preparation workload, allowing greater focus on instructional refinement and student learning outcomes.

4.1. The Collaborative Role of Gen AI in Instructional Design

Teachers commonly perceived Gen AI as a collaborative partner. It supported topic ideation, enhancement of existing materials, and multimodal transformation, expanding design possibilities without replacing teachers' decision-making authority. The value of Gen AI lies not only in the quality of its outputs, but also in its ability to stimulate teachers' design thinking, provide alternative perspectives, and improve efficiency. This study therefore shifts the focus from AI adoption to how teachers position and interact with AI within instructional design contexts.

4.2. Student Learning as the Core

Even with frequent Gen AI use, teachers consistently prioritized students' understanding, participation, and expression. Their cautious use of Gen AI in the evaluating and expressing stages reflects a strong commitment to maintaining learner agency and avoiding overreliance on AI tools. Gen AI was integrated as a supportive tool to enhance students' comprehension, reflection, and expression, while teachers continued to guide and evaluate the learning process.

4.3. Teachers' Concerns, Reflections, and Implications for English Teacher Education

Teachers approached Gen AI with both optimism and caution, recognizing its value in reducing preparation time while expressing concerns about accuracy, appropriateness, and potential student dependence. The findings suggest that teacher education should emphasize teachers' ability to evaluate and adapt AI-generated content, rather than focusing solely on operational skills.

This study has several limitations. First, participants were all AI-native English teachers at early stages of their teaching careers, limiting insights into how teaching experience may shape instructional design decision-making, as more experienced teachers may integrate Gen AI differently. Second, the analysis relied primarily on interview data and instructional artifacts, without direct classroom observations to capture how Gen AI-supported designs were enacted in practice. Future research could adopt comparative or longitudinal approaches to examine Gen AI integration across different levels of teaching experience.

5. References

- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. <https://doi.org/10.1191/1478088706qp063oa>
- Cope, B., & Kalantzis, M. (2015). *A pedagogy of multiliteracies: Learning by design*. Palgrave Macmillan. <https://doi.org/10.1057/9781137539724>
- Kamalov, F., Santandreu Calonge, D., & Gurrib, I. (2023). New era of artificial intelligence in education: Towards a sustainable multifaceted revolution. *Sustainability*, 15(16), 12451. <https://doi.org/10.3390/su151612451>
- Lim, F. V., & Tan-Chia, L. (2022). *Designing learning for multimodal literacy: Teaching viewing and representing*. Routledge.
- Lim, F. V., Toh, W., & Nguyen, T. T. H. (2022). Multimodality in the English language classroom: A systematic review of literature. *Linguistics and Education*, 69, 101048. <https://doi.org/10.1016/j.linged.2022.101048>

Reconfiguring Assessment in the Age of AI: Making Learning Visible through Meta-Task Awareness (MTA)

Lung-Hsiang Wong^{1*}, Guat Poh Aw¹

¹ National Institute of Education, Nanyang Technological University, Singapore

*lh Wong.acad@gmail.com

Abstract: *The widespread use of large language models (LLMs) is reshaping assessment practices, making it increasingly difficult to determine what learners actually know and can do, as learning processes become less visible in AI-mediated work. Rather than treating AI use as a problem of detection, this study explores how assessment can be reconfigured to make learning visible through learners' interaction with AI. Focusing on a postgraduate teacher education course, the study introduces Meta-Task Awareness (MTA) as an analytic lens for examining how learners regulate task goals, pedagogical considerations, and epistemic responsibility in LLM-supported lesson design. Using an exploratory qualitative design with descriptive code-and-count summaries, interaction trace data from 11 students were analysed at the episode level using a theory-informed coding framework. Findings show that students' engagement was primarily oriented toward task clarification, pedagogical reasoning, and evaluative judgment, with varying patterns resulting in three distinct orientations toward AI-mediated design. The study shows how assessment can be reconfigured so that learning becomes more visible, more accountable, and increasingly unavoidable in the age of AI.*

Keywords: *Generative AI in education; Human-AI collaboration; Meta-Task Awareness (MTA); AI-mediated assessment*

1. Introduction

The rapid diffusion of generative artificial intelligence (GenAI) has fundamentally disrupted learning and evaluation practices across educational contexts. Tasks commonly used for assessment – such as essays, lesson plans, design artefacts, or computer programs – can now be produced at high quality by Large Language Models (LLMs) within seconds. As a result, traditional product-oriented assessment approaches are increasingly challenged, raising concerns about the validity of evaluation in AI-rich learning environments (Yan et al., 2023).

Recent reviews suggest that much of the early research have focused on detection, restriction, or alternative assessment formats (Wong et al., 2024). However, such approaches often address surface-level symptoms rather than the underlying issue. Changing the form of assessment does not necessarily make learners' understanding, judgment, or decision-making processes more visible (Kim et al., 2024). The core challenge is not whether students use AI, but whether learning and evaluation designs can meaningfully capture how learners engage with tasks while working alongside AI.

This tension reflects a growing mismatch between product-oriented evaluation and process-oriented learning. There is a need to rethink assessment design so that AI use becomes an explicit condition of learning and evaluation, rather than an external disturbance to be eliminated. The challenge addressed in this study is not how to teach students to use AI more effectively, but how task and assessment design can reposition their epistemic responsibility when AI is already embedded in the learning process. This shift calls for approaches that foreground learners' regulation of tasks, their decisions about when and how to involve AI, and their capacity to revise and justify those decisions over time.

In this study, we argue that such engagement depends critically on learners' awareness of the task itself during AI-supported activity. We refer to this capacity as Meta-Task Awareness (MTA) (Chen & Wong, 2026; Wong, 2025a, 2025b), defined here as learners' ability to monitor, regulate, and reframe tasks while interacting with GenAI. Rather than

proposing MTA as a fully developed theory, we introduce it as an analytical lens for examining student-AI interaction in complex tasks.

Focusing on a graduate-level teacher education context, this paper reports a proof-of-concept qualitative analysis with descriptive code-and-count summaries of students' interactions with GenAI during pedagogical design tasks. By examining how MTA is manifested across task-oriented activities, this study elucidates how students differ in their engagement with AI during lesson design, and contributes to ongoing discussions on how learning and evaluation can be reconfigured in AI-rich environments.

2. Beyond Prompting: Meta-Task Awareness (MTA) in AI-Supported Learning

Recent discussions on GenAI in education have increasingly focused on prompting practices. From prompt engineering strategies to prompt literacy frameworks, a growing body of work has examined how learners formulate inputs to elicit more useful or accurate responses from LLMs (e.g., Murray, 2025). While these discussions have provided valuable insights into human-AI interaction at the surface level, they often foreground the optimisation of prompts rather than learners' understanding and regulation of the tasks for which AI is employed.

This emphasis reflects an assumption that better prompts lead to better learning outcomes. However, prompt quality alone does not guarantee meaningful engagement with the underlying task. Students may produce sophisticated prompts while remaining unclear about task goals, pedagogical constraints, or the division of cognitive labour between themselves and AI. Thus, AI becomes a shortcut to task completion rather than a partner in learning, leaving students' reasoning processes opaque to instructors and evaluators. The key challenge is not how well students prompt AI, but how they understand the task itself and manage their engagement with AI across different stages of task completion.

Building on this line of thinking, we introduce MTA (Wong, 2025a, 2025b) as an analytic lens for examining learner-AI interaction. MTA refers to learners' capacity to remain aware of the task they are working on while interacting with AI, including their ability to **(a) recognise task goals and constraints, (b) decide which aspects of the task to delegate to AI and which to retain under human judgment, and (c) reframe or revise the task through iterative interaction.** It refers to learners' awareness of task ownership and epistemic responsibility in human-AI collaborative task environments, where traditional assumptions about agency and control no longer hold.

Conceptually, MTA is grounded in three strands of existing work. First, it builds on research on task and metacognitive regulation (Flavell, 1979), which emphasises learners' monitoring and control of goals, constraints, and strategies during task performance (Zimmerman & Schunk, 2001). Yet unlike metacognition, which concerns monitoring and regulation of one's own cognitive processes, MTA concerns learners' awareness of task framing, epistemic responsibility, and the distribution of agency between human and AI within a task. Second, it draws on perspectives on human-AI coordination, where effective performance depends on learners' ability to make informed decisions about task delegation and control when interacting with intelligent systems (Endsley, 2017). Third, it aligns with process-oriented views of learning and evaluation, which foreground learners' engagement and reasoning processes rather than final products (Nicol & Macfarlane-Dick, 2006). Together, these perspectives provide the theoretical underpinning for using MTA as an analytic lens in AI-supported learning contexts.

In the present study, MTA serves as a conceptual bridge between discussions on AI-supported learning and the practical problem of evaluation design. By analysing how MTA is manifested in students' interactions with GenAI during pedagogical design tasks, this study seeks to provide an empirical basis for understanding how learners move beyond surface-level prompting toward more reflective and regulated forms of human-AI collaboration.

3. Research Context, Assignment Design, and Research Design

3.1. Research Context and Course Design

This study was situated in a postgraduate-level course on technology-enhanced Chinese language teaching offered at teacher education institution in Singapore. The participants were 11 master's students aged between 23 and 30. The

majority came from humanities backgrounds such as linguistics or Chinese studies, and had no formal teaching experience at the time of the study. All participants were already frequent users of LLMs for academic or personal purposes. However, their use of AI was unevenly integrated into pedagogical thinking.

The 13-session course was grounded primarily in techno-pedagogical knowledge for language education, including language learning theories, technologies as cognitive tools, multimedia-assisted learning, mobile learning, computer-supported collaborative learning, and design for formative assessment. GenAI was introduced only in the later phase of the course, such that the assignment assessed students' ability to synthesise and enact these pedagogical foundations – rather than isolated prompt engineering skills – within AI-supported design contexts.

A key design principle of the course was the normalisation of AI use. Rather than restricting AI, students were required to use AI tools as part of their learning activities. This decision was grounded on the premise that contemporary teacher education cannot be meaningfully conducted by assuming the absence of AI. Instead, the pedagogical challenge lies in supporting future educators to engage with AI in ways that are reflective, purposeful, and pedagogically grounded.

3.2. Assignment and Evaluation Design: From Product to Process

The focal assignment was deliberately designed to operationalise MTA within a lesson design context, not through explicit instruction, but through structural features embedded in task requirements. Students were required to design and iteratively improve a lesson plan and associated learning resources, beginning with an initial draft generated using LLM-based tools, followed by three weeks of revision and refinement through iterative prompting and critical evaluation.

Crucially, assessment did not prioritise the quality of the final artefact alone. Instead, evaluation foregrounded the process of transformation from the initial LLM-generated version to the final student-submitted design. This delta-based evaluation approach – drawing on the notion of delta (Δ) as change over time – treated learning as evidenced through students' design decisions, revisions, and justifications across iterations, rather than through end products in isolation.

Within this design, LLM-generated first drafts functioned as epistemic starting points rather than authoritative solutions. Students were required to engage with these drafts at the lesson level by specifying lesson objectives, evaluating generated suggestions, and making pedagogically grounded decisions about whether to adopt, modify, or reject AI outputs. Importantly, they were required to conduct lesson planning and resource development within a continuous interactional thread, ensuring that design reasoning unfolded cumulatively rather than as fragmented, task-by-task exchanges. This continuity made their evolving task understanding and decision-making processes observable in their interaction traces.

Rather than teaching metacognitive skills explicitly, the assignment structurally activated meta-level engagement by making epistemic responsibility unavoidable. To demonstrate learning, students had to justify design changes and show coherent alignment between pedagogical goals, learner needs, and revised lesson components across iterations. In this way, MTA was enacted through students' awareness of task ownership and epistemic authority in lesson design, as reflected in how they negotiated control with LLMs and exercised judgment over AI-generated content. The assignment thus shifted evaluation from product-oriented performance to process-oriented evidence of learning, positioning MTA as an emergent outcome of task and assessment design rather than as a taught strategy.

3.3. Research Questions and Research Design

This study adopts an exploratory, proof-of-concept orientation to examine how students engaged with LLMs in lesson design tasks. The purpose of this study is qualitatively analytic rather than statistical generalisation. The analysis is underpinned by MTA, used primarily as an analytic lens rather than as a validated psychological construct. Rather than evaluating the quality of AI-generated outputs, the study focuses on how students demonstrated pedagogical reasoning, task ownership, and epistemic control through their interactions with LLMs. Methodologically, the study employs a qualitative, process-oriented analytical design (Chi, 1997), prioritising analytic depth and trace-level resolution over sample size. Students' interactions with LLMs were analysed at the level of lesson design episodes, defined as coherent segments of student-LLM interaction oriented toward a specific lesson design objective or sub-task (e.g., lesson structure,

learning activities, assessment rubrics, instructional materials, or simulated learner responses). Within this qualitative analytic framework, a theory-informed coding scheme was applied to identify observable manifestations of MTA, and descriptive code-and-count summaries were used to surface overall distributional patterns across participants. Accordingly, the study addresses the following research questions (RQs):

RQ1. How is MTA manifested in students' LLM-mediated lesson design tasks?

This RQ examines how MTA is enacted during students' interactions with LLMs in lesson design tasks. It focuses on observable process-level behaviours through which they exercised task ownership, pedagogical reasoning, evaluative judgment, and epistemic control when framing prompts, responding to model outputs, and revising design decisions. The analysis attends to how these actions emerged during iterative lesson design rather than to final artefacts.

RQ2. How do patterns of MTA vary across students with differing orientations toward AI-mediated lesson design?

Building on the manifestations identified in RQ1, this RQ examines how students differed in the patterning of MTA during AI-mediated lesson design. Adopting a level-based analytic approach, the analysis compares how task clarification, pedagogical justification, evaluative judgment, and epistemic authority were configured across students' LLM interactions. These are used to characterise distinct orientations toward engaging with AI in lesson design contexts. The analysis does not aim to model relationships among codes as independent variables, but to interpret recurring configurations of co-occurring practices within interaction traces.

Coding framework

To address RQ1, a theory-informed coding scheme was developed to operationalise MTA as observable practices in students' LLM-mediated lesson design tasks. Rather than treating MTA as an abstract disposition, the analysis focused on how students enacted task ownership, epistemic responsibility, and pedagogical decision-making through their interactions with LLMs. In addition to qualitative interpretation, a descriptive code-and-count summary was produced to surface overall distributional patterns. Frequencies were tallied at the episode level (i.e., each coded instance within a task/episode), and then aggregated across students for reporting in the Findings. Episodes were defined pragmatically based on shifts in lesson design objectives, rather than fixed interaction lengths. For analytic clarity, multi-valued codes were collapsed into binary presence/absence for frequency reporting.

The coding scheme was derived deductively from the conceptual framing of MTA and inductively refined through close examination of student–LLM chatlogs. The scheme comprised four core coding dimensions, each with clearly defined indicators and value categories, as summarised in Table 1. These dimensions captured (a) how lesson design tasks were specified and constrained, (b) how LLM-generated outputs were evaluated, (c) how pedagogical decisions were justified, and (d) how control over the lesson design process was distributed between the student and the LLM. Coding was carried out by a researcher and independently verified by the course instructor through a parallel pass over the data, yielding highly consistent code distributions with over 90% agreement across coded instances.

To address RQ2, coded tasks were examined across students using a qualitative aggregation approach. Rather than constructing formal typologies or validated scales, the analysis aimed to identify emergent levels of engagement with LLMs in lesson design. Students were compared in terms of the consistency, diversity, and depth of MTA-related coding patterns observed across their tasks. Given the nature of the task, the study assumes functional equivalence across mainstream LLMs and focuses on students' pedagogical reasoning as manifested through their prompts and revisions.

Through this analytical design, the study provides a concrete illustration of how MTA can be operationalised, observed, and compared in LLM-supported lesson design contexts, offering an initial proof of concept for MTA as a lens for learning and evaluation in generative AI-mediated education.

Table 1: *The coding scheme for qualitative analysis to address RQ1*

Code	Dimension	Description	Values
C1	Task specification	Whether the student explicitly specified or refined the lesson design task, goals, or constraints, rather than relying on default AI interpretations.	0 = Absent; 1 = Present
C2	Evaluation of LLM output	Whether the student critically evaluated, accepted, rejected, or questioned LLM-generated suggestions.	0 = Absent; 1 = Present
C3	Pedagogical justification	Whether pedagogical rationales were explicitly articulated to justify lesson design decisions, and whether such rationales were primarily initiated by the student, prompted by the LLM, or jointly constructed.	0 = Absent; 1 = LLM-led; 2 = Student-led; 3 = Co-constructed
C4	Epistemic authority	Whether the student exercised primary decision-making authority over lesson design choices, including overriding, restructuring, or redirecting LLM-generated outputs in accordance with pedagogical goals and constraints.	0 = Absent; 1 = Student-led

4. Findings

4.1. RQ1: Manifestations of Meta-Task Awareness in AI-Mediated Design Processes

To address RQ1, all student–LLM chatlogs were coded using six process-oriented codes representing different dimensions of MTA. Table 2 presents the aggregated frequencies of each code across the eleven participants.

Table 2: Aggregated frequencies of MTA-related codes across students' LLM-mediated lesson design tasks

Code	Description (concise)	Frequency (n)	Illustrative example (excerpt)
C1	Task specification	50	"The lesson is for Grade 8 Chinese Language students with mixed proficiency. Redesign the activity to focus on inferencing rather than vocabulary recall."
C2	Evaluation of LLM outputs	42	"This activity looks engaging, but it may be too demanding for weaker students. Can we simplify the output requirement?"
C3-1	Pedagogical justification (student-initiated)	19	"Insert a guided discussion here because students need support before independent writing."
C3-2	Pedagogical justification (co-constructed)	25	"If the goal is text appreciation, then the task should allow multiple interpretations rather than fixed answers."
C3-3	Pedagogical justification (LLM-initiated)	9	"You are right that peer discussion can lower affective filter. I will keep this step."
C4	Epistemic authority in lesson design	22	"Keep the structure, but remove the Kahoot! activity – it distracts from the main objective."

Overall, a total of 167 coded instances were identified. Among the six codes, C1 (task clarification and goal refinement) and C2 (pedagogical reasoning and justification) were the most frequently observed. These were followed by C4-1 (evaluation and critique of AI-generated outputs), and C3-2 (iterative refinement of prompts or outputs). In contrast, C3-3 (explicit meta-level reflection on AI use) occurred least frequently.

The distribution of codes suggests that students' engagement with AI was primarily oriented toward task-level and pedagogical decision-making, rather than abstract reflection on AI itself. High frequencies of C1 and C2 indicate that such actions were frequently observed across the dataset, with students articulating lesson goals, constraints, and pedagogical rationales when interacting with LLMs. Codes associated with evaluative judgment and selective adoption of AI outputs (C4) were also consistently observed across participants, indicating active scrutiny rather than uncritical acceptance of generated content.

Indeed, differences among students were not only reflected in the total frequency of coded behaviours, but also in the relative balance between generative, evaluative, and reflective actions. This variation provided an empirical basis for the level-based analysis reported in RQ2.

4.2. RQ2: Different MTA patterns across students with differing orientations toward AI-mediated lesson design

Building on the manifestations of MTA identified in RQ1, RQ2 examines how these process-level behaviours were configured differently across students, resulting in distinct orientations toward AI-mediated lesson design. Rather than treating MTA as a unidimensional construct, a level-based analytic grouping was conducted to capture how combinations of task clarification, pedagogical justification, evaluative judgment, and epistemic authority co-occurred within student–LLM interactions. Based on the relative density, balance, and functional roles of coded behaviours, students were

categorised into three MTA levels for analytic convenience. The three levels do not represent developmental stages or stable learner traits, but analytically derived groupings grounded in recurrent patterns of MTA enactment observed within the focal task. Table 3 summarises the mean per-student frequencies of MTA-related codes within each MTA level (not aggregated totals), with group sizes indicated by n .

Table 3: *Mean per-student frequencies of MTA-related codes by MTA level (n = number of students in each level)*

Code	C1	C2	C3-1	C3-2	C3-3	C4	Mean Total
Level 1 ($n = 2$)	6.5	5.5	3	2.5	1	2.5	21
Level 2 ($n = 7$)	4.7	4	1.7	2.6	1	2.1	16.1
Level 3 ($n = 2$)	2	1.5	0.5	1	0	1	6

Level 1: AI as a generative executor (Low MTA)

Students in Level 1 ($n=2$) primarily engaged with LLMs as content generators. Their interactions were dominated by task execution and material production, with limited evidence of student-initiated task refinement or evaluative intervention. Pedagogical justification, when present, was typically reactive rather than self-initiated, often emerging only in response to AI-generated suggestions. Across these cases, epistemic authority over lesson design decisions resided largely with the LLM. Prompts tended to specify outputs rather than constraints or pedagogical intentions, and AI-generated content was more often directly adopted. MTA at this level was characterised by minimal task ownership and limited reflective control over the design process.

Level 2: AI as a pedagogical assistant (Moderate MTA)

The majority of students ($n=7$) fell into Level 2, exhibiting a more balanced pattern of MTA manifestations. These students frequently articulated lesson goals and constraints, provided pedagogical rationales, and engaged in iterative refinement of prompts or outputs. Evaluative judgment was evident in selective adoption, modification, or rejection of AI-generated suggestions. While the LLM played an active role in idea generation, students at this level retained decision-making authority over pedagogical directions. AI was treated as a supportive assistant rather than a driver of design. Their MTA was thus characterised by sustained task ownership and functional integration of AI contributions within students' pedagogical reasoning processes.

Level 3: AI as a dialogic design partner (High MTA)

Students at Level 3 ($n=2$) demonstrated the most sophisticated configurations of MTA. Their interactions were marked by dense and diverse combinations of task clarification, pedagogical justification, iterative refinement, and explicit evaluation of AI outputs. These students actively shaped the trajectory of the design process, using AI responses as objects for reflection, comparison, and pedagogical decision-making. At this level, epistemic authority was clearly retained by the student. LLM-generated content functioned as a catalyst for design-level reasoning. Although explicit meta-level reflection on AI use was not frequent, reflective control was implicitly enacted through students' consistent regulation of task goals, constraints, and design quality across iterations.

5. Discussion and Implications

This study responds to recent calls for more process-oriented and methodologically grounded research on GenAI in education, particularly work that foregrounds human–AI collaboration and rethinks assessment. The findings from RQ1 and RQ2 suggest that MTA is neither a generic metacognitive trait nor a function of prompt sophistication. Rather, it is a structurally activated orientation toward epistemic responsibility in AI-mediated work.

5.1. MTA as structurally activated, not instructionally taught

A key insight from the findings is that higher-level MTA did not emerge because students were explicitly taught how to “think about AI use”. Instead, meta-level engagement was activated by the design of the assignment itself. By anchoring assessment on design transformation (Δ) rather than final products, students were compelled to demonstrate judgment, revision, and justification across iterations. In this sense, MTA functioned as a necessary condition for demonstrating learning, not an optional add-on.

This helps clarify a conceptual gap between metacognition and MTA. While MTA involves metacognitive processes, it is not framed as an internal strategy to be explicitly trained or self-reported. Rather, it is observable through task-level decisions made under epistemic accountability, as evidenced in interaction traces. Meta-level engagement was not prompted instructionally (“reflect on your AI use”), but elicited structurally through task and assessment design.

5.2. AI as both amplifier and mirror of pedagogical thinking

Across levels, the data acted simultaneously as intelligent amplifiers and mirrors of students’ thinking. Students who demonstrated stronger MTA were not those who produced longer or more complex prompts, but those who integrated pedagogical goals, learner considerations, and evaluative judgment when deciding how to engage with or override AI.

Students’ ability to critique and revise AI-generated lesson drafts did not arise in a vacuum. Prior to the AI-based assignment, they had already engaged in repeated cycles of lesson critique and peer negotiation across the course. These earlier activities fostered pedagogical diagnostic criteria – such as sensitivity to tool overload, misalignment with learner needs, or weak assessment design – which were later mobilised in AI-mediated contexts. The assignment thus did not reward pre-existing “AI talent”, but activated already-constructed pedagogical judgment through structural accountability.

5.3. Differentiated MTA orientations in AI-mediated lesson design

The level-based analysis in RQ2 further illustrates that students differed not only in the frequency of MTA-related behaviours, but in their overall orientation toward AI-mediated lesson design. Lower-level MTA was characterised by efficiency-oriented use of LLMs as generators or executors of tasks, with design authority largely ceded to the model. Higher-level MTA, in contrast, involved sustained task ownership, evaluative judgment, and epistemic control.

These levels should not be interpreted as fixed traits or ability rankings. Instead, they represent situated orientations shaped by task structure and evaluative demands. This reinforces the argument that MTA is design-sensitive: different assessment and interaction designs are likely to elicit different patterns of AI engagement, even from the same learners.

5.4. Design principles for AI-mediated, process-oriented assessment

Drawing from the findings, four design principles can be articulated for eliciting MTA in AI-mediated learning:

- 1. Anchor evaluation on transformation (Δ), not end products:** Assess learning through changes across iterations, making epistemic engagement unavoidable.
- 2. Treat AI-generated outputs as epistemic starting points, not authoritative solutions:** Require learners to evaluate, adapt, or reject AI suggestions in relation to learning goals.
- 3. Require sustained work within a continuous interactional thread:** Design tasks that support iterative development within a single AI interaction, rendering evolving task ownership traceable.
- 4. Activate meta-level engagement structurally, not instructionally:** Embed epistemic responsibility into task and assessment design rather than teaching metacognitive strategies explicitly.

Together, these principles reposition AI-mediated assessment from a problem of detection or control to an opportunity for reconfiguring how learning is evidenced.

5.5. Limitations

Several limitations should be noted. First, the study involved a small cohort within a specific teacher education context, and findings should be interpreted as a proof-of-concept rather than generalisable claims. The small sample size is appropriate for this qualitative, trace-based analysis, where analytic depth is prioritised over representativeness. Second, while rich interaction trace data were analysed, no direct measures of learning outcomes beyond task performance were included. Future work could examine how MTA relates to longer-term pedagogical development or transfer across tasks.

6. Conclusion

This study demonstrates that MTA can be meaningfully operationalised and empirically examined through interaction trace analysis in AI-mediated lesson design tasks. By shifting assessment from product- to process-oriented transformation, the study offers a viable pathway for addressing contemporary challenges in AI-era evaluation design.

More broadly, it suggests that the question is no longer whether AI should be allowed in learning, but how assessment becomes more visible, more accountable, and increasingly unavoidable in the presence of LLMs.

Acknowledgments

This project received no external funding. Consent from the research participants was obtained based on ethics approval by Nanyang Technological University (IRB ref: IRB-2025-658).

References

- Chen, H.-C., & Wong, L.-H. (2026). The MTA-TPACK Dynamic Collaboration Spiral: Making Pedagogical Thinking Visible in Human–AI Scientific Visualization for Sustainable Teacher Innovation. *Sustainability*, 18(6), 2718. <https://doi.org/10.3390/su18062718>
- Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *Journal of the Learning Sciences*, 6(3), 271-315.
- Endsley, M. R. (2017). From here to autonomy: lessons learned from human–automation research. *Human Factors*, 59(1), 5-27. <https://doi.org/10.1177/0018720816681350>
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10), 906. <https://doi.org/10.1037/0003-066X.34.10.906>
- Kim, J., Yu, S., Detrick, R., & Li, N. (2024). Exploring students' perspectives on Generative AI-assisted academic writing. *Education and Information Technologies*, 30(1), 1265-1300. <https://doi.org/10.1007/s10639-024-12878-7>
- Murray, M. (2025). Bridging the Generative AI Literacy Gap: A Guide to Introducing Prompt Engineering in University Courses. *Issues in Informing Science and Information Technology*, 22, Article 10. <https://doi.org/10.28945/5536>
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199-218. <https://doi.org/10.1080/03075070600572090>
- Wong, L.-H. (2025a). The prompt is not enough: Rethinking AI-learner interaction in the post-prompting era, *Proceedings of International Conference on Computers in Education 2025 (Volume II)* (pp. 311-319), Chennai, India.
- Wong, L.-H. (2025b). Transforming education with Generative AI: Designing for the Post-Prompting Era. *Information and Technology in Education and Learning*, 5(1), Inv-p003-Inv-p003. <https://doi.org/10.12937/itel.5.1.inv.p003>
- Wong, L.-H., Park, H., & Looi, C.-K. (2024). From hype to insight: Exploring ChatGPT's early footprint in education via altmetrics and bibliometrics. *Journal of Computer Assisted Learning*, 40(4), 1428-1446. <https://doi.org/10.1111/jcal.12962>
- Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1), 90-112. <https://doi.org/10.1111/bjet.13370>
- Zimmerman, B. J., & Schunk, D. H. (2001). Reflections on theories of self-regulated learning and academic achievement. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives (2nd ed.)*. Lawrence Erlbaum Associates Publishers.

Hong Kong-French Collaboration in Machine Learning Education for Junior Secondary Students: A Mixed-Methods Study on Teacher Development

Siu Cheung Kong^{1,2*}, Thomas Deneux³, Yin Yang¹, So Sum Chow²

¹ Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong SAR

² Artificial Intelligence and Digital Competency Education Centre, The Education University of Hong Kong, Hong Kong SAR

³ Université Paris-Saclay, CNRS, Institut de Neurosciences Paris-Saclay, Gif-sur-Yvette, France

* sckong@eduhk.hk

Abstract: *This exploratory mixed-methods study evaluates a teacher development (TD) workshop, jointly designed by Hong Kong and French educators, aimed at advancing machine learning (ML) literacy among French junior secondary teachers. The six-hour workshop, held in France, applied the TPACK framework and integrated AlphaAI robots for hands-on ML activities. Quantitative and qualitative analyses revealed significant improvements in teachers' ML conceptual understanding and self-efficacy across all TPACK domains. Teachers reported substantial cognitive and affective growth, alongside greater confidence in utilising robots for ML instruction. They also highlighted implementation challenges, such as time and resource constraints. Ultimately, this study underscores the value of contextualised, experiential TD models and suggests scalable strategies for preparing teachers to deliver AI education in junior secondary settings.*

Keywords: *AI Literacy; Junior Secondary Student; Machine Learning; Teacher Development; TPACK*

1. Introduction

The rapid rise of machine learning (ML) and artificial intelligence (AI), exemplified by models like GPT, has fundamentally transformed the role of technology in society and education. However, the “black box” nature of ML algorithms raises pressing concerns regarding transparency and ethics, making ML literacy an urgent priority for K–12 education (Ng et al., 2021). Effective ML literacy must extend beyond basic technical skills to encompass critical thinking about bias, fairness, and broader societal impacts (Tedre et al., 2021; Kong et al., 2024).

Teachers are central to bridging this gap in ML understanding. Yet, they frequently lack adequate professional development and resources, particularly for teaching hands-on ML concepts (Gibellini et al., 2023). While robotics-integrated activities have proven effective in making abstract ML concepts tangible (Absalon & Deneux, 2025), scalable and effective teacher development (TD) approaches remain under-researched.

This study builds upon existing research from Hong Kong by adapting TD models and robotics-assisted learning for the French educational context. The collaborative workshop aimed to enhance teachers' conceptual mastery of ML and their pedagogical adaptation through context-sensitive approaches. The study is guided by three research questions:

RQ1: To what extent does the TD workshop enhance teachers' understanding of ML concepts?

RQ2: How does the TPACK-informed workshop influence teachers' ability to integrate learning robots into ML instruction?

RQ3: In what ways do teachers articulate their perspectives and experiences of the TD workshop?

2. Literature Review

AI's deep integration into daily life necessitates comprehensive ML literacy. Frameworks like Kong et al. (2024) highlight understanding, application, psychological readiness, and ethical reflection as key dimensions. Robot-assisted learning—such as with AlphaAI robots—helps visualise ML algorithms (KNN, ANNs), supporting conceptual and ethical

learning (Absalon & Deneux, 2025). Success relies on teachers' ability to facilitate inquiry-based, ethical environments, demanding strong ML knowledge and adaptive pedagogy (Tedre et al., 2021; Gibellini et al., 2023).

Frameworks such as AEER (Attention-Engagement-Error-feedback-Reflection) and TPACK (Technological Pedagogical Content Knowledge) are vital for effective ML instruction. AEER provides a micro-level cycle for active learning, while TPACK aligns content, pedagogy, and technology (Kong et al., 2025). Professional Development (PD) programmes grounded in these frameworks—emphasising hands-on practice and peer collaboration—improve teachers' ML knowledge and self-efficacy. Sustaining impact requires ongoing support and adaptation across educational systems.

Scalable, contextually relevant TD models for ML are rare; most research prioritises student outcomes over teacher preparation. This study addresses these gaps by evaluating a joint Hong Kong-French TD workshop, offering insights into adaptable teacher preparation for ML education.

3. Research Design

3.1. Study Design and Participants

The study employed an exploratory mixed-methods design to assess a one-day TD workshop tailored specifically to the French secondary curriculum. This approach was chosen to thoroughly evaluate both the pedagogical coherence of the workshop and its feasibility for future, larger-scale interventions.

Eleven secondary teachers (72.7% male, 27.3% female) from nine schools in the Auvergne-Rhône-Alpes region participated. The cohort's ages ranged from 31 to 60, with the majority holding advanced degrees (63.6% Master's, 9.1% Doctorate) and possessing STEM backgrounds. This small sample size is reflective of the typical logistical constraints in exploratory, cross-cultural TD research.

3.2. Workshop Design

The six-hour workshop was grounded in TPACK and AEER frameworks, comprising:

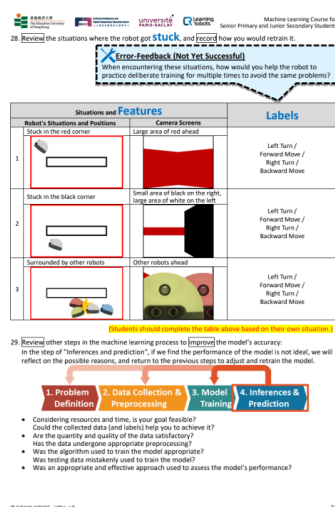
Workshops: Teachers utilised AlphaAI robots and accompanying software to experiment training an ML model during playful “robot racing”, and later visualise the details of KNN and ANN decision making.

Discussions: Guided by AEER, sessions focused on training data, robot performance, and ethical/societal issues.

Resources: Teachers received a comprehensive worksheet (110 pages, see Figure 1) and ongoing teaching support.

Figure 1

Sample worksheet page prompting students to analyse ML errors and consider alternative instructional strategies.



3.3. Data Collection and Analysis

Quantitative data included a 13-item ML concept test and a 15-item TPACK self-efficacy scale, both translated into French. Qualitative data comprised written reflections ($N = 7$) and semi-structured interviews ($N = 2$). Quantitative

analysis used Wilcoxon signed-rank tests; instrument reliability was high for TPACK ($\alpha = 0.94$), but lower for the ML Concept Test due to sample size and instrument alignment (Taber, 2018). Qualitative data were analysed thematically.

4. Results

4.1. Enhancement in ML Conceptual Understanding (RQ1)

Teachers' ML concept test scores improved significantly, rising from a pre-test mean of 8.27 ($SD = 1.68$) to a post-test mean of 10.73 ($SD = 1.49$) out of 13 (Wilcoxon $Z = -2.72$, $p < 0.01$). Gains were most prominent in supervised learning, overfitting, and KNN/ANNs mechanics. Robotics activities were credited in teachers' qualitative reports for making abstract processes concrete, especially through real-time visualisation of data labelling and prediction.

4.2. Influence on Integrating Learning Robots (RQ2)

The TPACK-based activities substantially boosted teachers' self-perceived ability to integrate robots into their ML instruction. Wilcoxon signed-rank tests revealed significant pre-post gains across all measured domains (all $p < 0.01$). TCK (mean from 1.39 to 3.82, $Z = -3.02$), CK (2.11 to 3.93, $Z = -2.94$), PCK (1.98 to 3.43, $Z = -2.85$), and overall TPACK (1.36 to 3.36, $Z = -2.95$). These robust gains suggest that the robot-based tasks effectively helped teachers bridge the gap between technology, content, and pedagogy, thereby increasing their confidence in utilising innovative tools.

4.3. Teachers' Perspectives and Experiences (RQ3)

Thematic analysis of the qualitative data revealed three primary themes.

Cognitive growth. All respondents reported a deepened understanding of ML, with several noting an enhanced ability to break down and explain these concepts to students. The experience of "seeing a concrete application of AI usage with robots" was frequently highlighted as a catalyst for this growth.

Affective engagement. Teachers expressed high appreciation for the workshop's progressive, experiential design, valuing the safe environment it provided for trial-and-error. As one participant noted, "The progressivity allowed me to go gently," underscoring the vital role of pedagogical scaffolding.

Implementation challenges. Despite the positive feedback, teachers identified time constraints and classroom resource management as significant logistical barriers. Furthermore, some indicated a need to revisit highly complex topics, such as ANNs, to fully solidify their confidence prior to classroom implementation.

5. Conclusions and Implications

A focused, six-hour workshop guided by TPACK significantly enhanced French secondary teachers' ML conceptual understanding and self-efficacy in integrating robotics. Quantitative data showed substantial gains in ML literacy and TPACK domains; qualitative findings highlighted teachers' cognitive and affective growth, alongside practical concerns about classroom implementation.

Compared with the findings from Hong Kong (Yang & Kong, 2025), the French cohort had higher baseline ML scores and engaged in more technical discussions (they also teach at higher level). Both cohorts demonstrated strong TPACK gains and valued the hands-on, scaffolded approach. Contextual adaptation, such as focusing on core ML concepts relevant to the French curriculum, enhanced the training's impact.

Adaptable TD workshops should prioritise contextual relevance, feature customisable activities, and address logistical barriers through sustained support or digital platforms. Integrating these programmes into existing structures is critical for long-term viability. While the small sample size and lack of cross-context comparisons limit generalisability, the study provides a foundation for future, larger-scale research.

In summary, technology-enhanced, contextually grounded teacher development is essential for equipping educators to deliver rigorous ML literacy instruction. By addressing practical barriers and grounding training in robust frameworks like TPACK, teachers can effectively integrate AI technologies into junior secondary education and prepare students for an AI-driven future.

Acknowledgements

The authors would like to acknowledge the funding from the Central Reserve Allocation Committee, The Education University of Hong Kong of the Hong Kong Special Administrative Region, China (Uncovering the “Black Box” of Machine Learning: Promoting Artificial Intelligence Literacy with AlphaAI robots in Senior Primary/Junior Secondary Schools across Hong Kong and France) (Project No. EdUHK 04A55) as well as the funding from the University Grant Council of the Hong Kong Special Administrative Region, China-France “PROCORE F-EdUHK801/24 Cultivating Learners to be Passionate Thinkers for AI-Human Partnership: A French-Hong Kong Knowledge Exchange” (Project No. EdUHK R9032).

References

- Absalon, M., & Deneux, T. (2025, April). “AlphaAI”: Teaching AI algorithms to K12 by training learning robots and visualizing how it works. In *Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 28, pp. 29096–29103)*. <https://doi.org/10.1609/aaai.v39i28.35181>
- Gibellini, G., Fabretti, V., & Schiavo, G. (2023). AI education from the educator’s perspective: Best practices for an inclusive AI curriculum for middle school. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–6). <https://doi.org/10.1145/3544549.358574>
- Kong, S. C., Cheung, W. M. Y., & Tsang, O. (2024). Developing an artificial intelligence literacy framework: Evaluation of a literacy course for senior secondary students using a project-based learning approach. *Computers and Education: Artificial Intelligence*, 6, 100214. <https://doi.org/10.1016/j.caeai.2024.100214>
- Kong, S. C., Yang, Y., & Chow, S. S. (2025). Teachers’ perspectives on teaching machine learning to senior primary students. In *Proceedings of the 18th annual International Conference of Education, Research and Innovation* (pp. 7302–7309). Seville, Spain: IATED Academy. <https://doi.org/10.21125/iceri.2025.2008>
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
- Taber, K. S. (2018). The use of Cronbach’s alpha when developing and reporting research instruments in science education. *Research in science education*, 48(6), 1273–1296. <https://doi.org/10.1007/s11165-016-9602-2>
- Tedre, M., Toivonen, T., Kahila, J., Vartiainen, H., Valtonen, T., Jormanainen, I., & Pears, A. (2021). Teaching machine learning in K–12 classroom: Pedagogical and technological trajectories for artificial intelligence education. *IEEE access*, 9, 110558–110572. <https://doi.org/10.1109/ACCESS.2021.3097962>
- Yang, Y., & Kong, S. C. (2025). Professional development for teachers in AI literacy education: Teaching machine learning to senior primary and junior secondary students. In *Proceedings of the 17th International Conference on Computer Supported Education - Volume 2*, (pp. 35–42). Porto, Portugal: SciTePress.

LLM Enhanced Video to Text Recognition for Authentic EFL Video Making

Tasks: Evaluation across GPT Versions and Video Lengths

Yi-Fan Liu^{1,2}, Muhammad Irfan Luthfi^{3,5*}, Wu-Yuin Hwang^{3,4}

¹Digital Affairs Department, Hsinchu County Government, Taiwan

²Research Center for Testing and Assessment, Academy for Educational Research, Taiwan

³Graduate Institute of Network Learning Technology, Central University, Taiwan

⁴Department of Computer Science and Information Engineering, Dong Hwa University, Taiwan

⁵Department of Electronics and Informatics Engineering Education, Universitas Negeri Yogyakarta, Indonesia

*iir.irfan02@gmail.com

Abstract: Authentic video making tasks are increasingly used in English as Foreign Language (EFL) learning due to their support for meaningful communication and rich multimodal evidence, but teachers often face heavy workloads in reviewing videos and providing timely feedback. This study proposes Large Language Model (LLM)-enhanced Video to Text Recognition (VTR), which transforms videos into textual outputs, including sentence suggestions for learner narration and completes video summaries. To examine how output quality varies by video length and model version, a focused evaluation was conducted with six English teachers using a web-based system over two weeks. Participants evaluated 100 videos from a public dataset across four duration categories, rating output quality and usefulness, judging acceptability without edits, selecting error tags, and writing reference summaries for comparison. Results showed differences across Generative Pre-trained Transformer (GPT) versions and video lengths, with stronger GPT models producing higher-quality and more acceptable outputs, while longer videos led to more missing information and vagueness. These findings suggest that VTR can support EFL learning and highlight the need to improve content coverage and specificity, particularly for longer videos.

Keywords: Video to Text Recognition, large language models, EFL learning, video summarization, authentic video making tasks

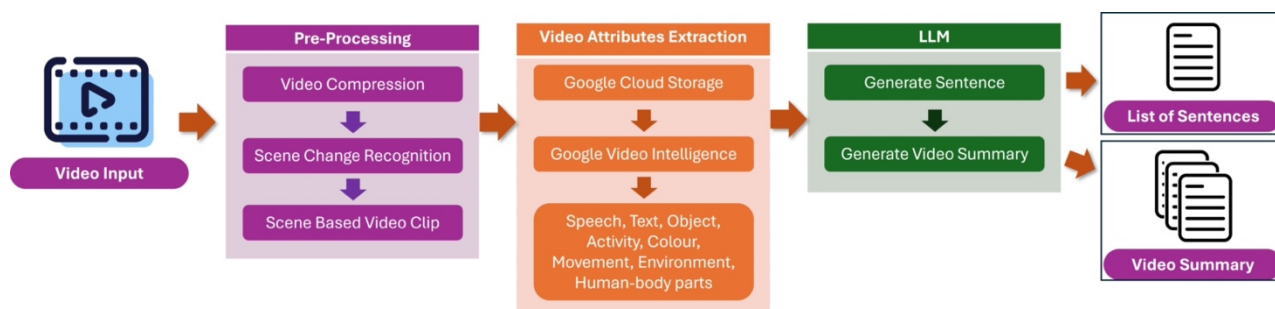
1. Introduction

Video making is increasingly used in EFL learning because it places learners in authentic situations requiring meaning planning, appropriate language use, and audience-oriented communication (Ma & Bruce, 2025), while generating rich multimodal evidence from speech, on-screen text, and visual context (Yeh, Heng, & Tseng, 2021). However, teachers often cannot review all learner videos or provide timely support (Abdelhadi & Belaid, 2025), and existing approaches relying on manual review, subtitles, or basic transcripts may fail to capture key language evidence or transform video content into reusable learning resources (Hsu, 2024), creating a gap between authentic video-making activities and scalable language support (Said Al Ghafri, 2023). To address this gap, we developed VTR, which converts video content into structured textual outputs, including sentence suggestions for learner narration and complete video summaries. Prior studies have applied VTR in authentic EFL contexts with promising results in usability, engagement, and learning support (Liu, Luthfi, & Hwang, 2024; Hwang, Luthfi, & Liu, 2024; Liu, Luthfi, & Hwang, 2025; Chang, Liu, Hwang, & Chen, 2025), but the quality of these outputs has not been systematically evaluated; therefore, this study examines the quality and perceived usefulness of VTR-generated sentence suggestions and video summaries for supporting EFL learning tasks.

2. LLM-enhanced Video to Text Recognition

Figure 1

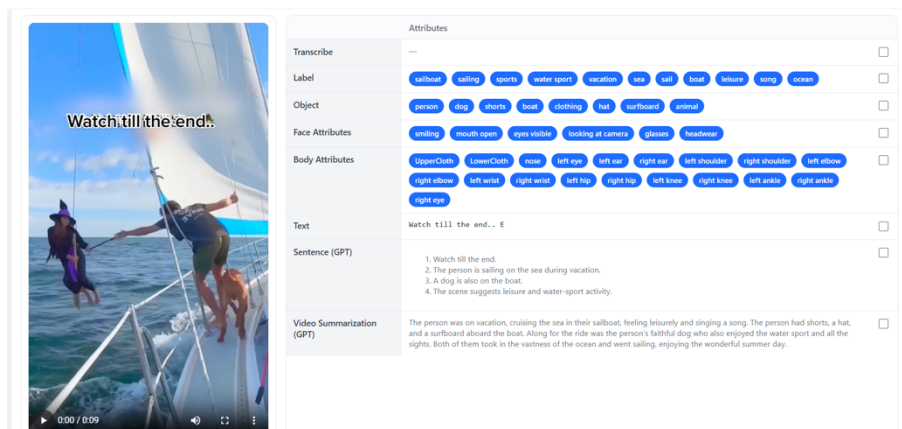
The process of LLM-enhanced Video to Text Recognition



LLM-enhanced Video to Text Recognition (VTR) transforms learner videos into structured text for language learning support. After upload (Figure 1), videos are compressed and segmented into scene-based clips using scene change recognition, then stored in Google Cloud Storage, where Google Video Intelligence extracts multimodal attributes, including speech transcription, detected text, labels, objects, activities, movement, colour, environmental information, and human body-related cues. These attributes are integrated by a large language model (GPT) to generate sentence suggestions for learner narration and a complete video summary. Figure 2 illustrates an example of a sailing video, where extracted attributes (e.g., sailboat, water sport, person, dog, and on-screen text such as “Watch till the end”) are used to produce sentences describing key actions (e.g., a person sailing with a dog) and a summary of the main situation and entities, demonstrating how attribute extraction supports sentence generation and summarization for EFL learning tasks.

Figure 2

The example of VTR application to generate sentences and video summarization from a short video



3. Method

Six English teachers from a public university with EFL teaching experience participated voluntarily and provided informed consent. Data were collected over two weeks using a web-based system, in which participants evaluated system-generated sentence suggestions and video summaries, completed rating forms and a perception questionnaire, and assessed four GPT model outputs for the same 100 videos (within-rater design; 400 conditions per rater). The videos were sampled from a publicly available dataset (Shang et al., 2025) and grouped into four duration categories: very short (≤ 10 s), short (11–30s), medium (31–60s), and long (> 60 s), with 25 videos per category covering diverse real-world scenes. The study evaluated two outputs: sentence suggestions (relevance, linguistic quality, level appropriateness, actionability, overall score, and error tags) and video summaries (faithfulness, coverage, coherence, conciseness, language quality, overall score, and acceptability), and participants also wrote reference summaries for ROUGE-L and BERTScore comparisons. Four model conditions were tested: GPT-3 (text-davinci-003), GPT-3.5, GPT-4, and GPT-4o. Inter-rater reliability was high ($ICC = 0.87$), and given the small sample size ($n = 6$), analysis focused on descriptive statistics and pattern comparison across video length and model conditions, with error tags analyzed using frequency counts, summary

quality examined using ROUGE-L and BERTScore against reference summaries, and associations between overall ratings and quality dimensions also explored.

4. Result and Discussion

4.1. Analysis of Quality of VTR Sentence Suggestions for Learner Narration

Sentence suggestion quality showed consistent patterns across GPT versions and video lengths, with GPT-4 and GPT-4o achieving the highest scores (approximately 4.0–4.3), followed by GPT-3.5 (approximately 3.1–3.6), and GPT-3 the lowest (approximately 2.4–3.1). Scores decreased as video length increased, with lower ratings for medium and long videos, indicating greater difficulty in generating effective narration for more complex content; across dimensions, relevance and linguistic quality were generally higher, while actionability was slightly lower, particularly for longer videos. Error tag patterns further clarified these limitations, with missing information and vagueness as the most frequent issues, especially in longer videos and weaker models, suggesting that the main challenge lies in content selection and completeness rather than grammatical correctness. When key actions were omitted or descriptions were too general, usefulness for narration decreased, while wrong content appeared more often in weaker models, indicating weaker grounding between extracted attributes and generated sentences; additionally, overly complex language and grammatical issues negatively affected level appropriateness and linguistic quality. Overall, these findings suggest that stronger GPT models provide more reliable narration support, although improvements are needed to enhance content coverage and specificity for longer and more complex videos.

4.2. Analysis of Quality of VTR Video Summary

Video summary quality and automated evaluation results showed consistent patterns across GPT versions and video lengths, with GPT-4 and GPT-4o achieving the highest performance across quality dimensions (approximately 3.8–4.3) and similarity metrics (ROUGE-L up to ~0.61; BERTScore up to ~0.91), followed by GPT-3.5, while GPT-3 showed the lowest performance. This pattern was stable across video categories, indicating that stronger models produce more accurate, coherent, and consistent summaries, whereas both quality ratings and acceptance decreased as video length increased, with the lowest performance in long videos, reflecting the difficulty of maintaining coverage and consistency in more complex inputs. Automated metrics aligned with human evaluation, showing higher similarity for stronger models and lower similarity for longer videos; among dimensions, faithfulness and coherence were generally higher, while coverage declined more noticeably with increased video length, indicating challenges in capturing key events. Overall, the results suggest that model capability primarily affects content accuracy and coherence, while video length influences completeness and information loss, highlighting the need to improve coverage and consistency for longer and more complex videos.

4.3. Analysis of Perceptions Toward VTR

Participants reported positive perceptions of VTR for supporting EFL learning tasks. Perceived usefulness was high across all items (means > 4.2), indicating that VTR can enhance task performance, productivity, and evaluation by providing effective sentence suggestions and video summaries. Perceived ease of use was also positive (means > 4.0), suggesting that the system is easy to learn, operate, and understand with minimal effort. Attitude toward use and behavioral intention were similarly strong (means > 4.2), indicating that participants viewed VTR as a valuable tool and were willing to use and recommend it in future practice. Overall, responses were consistently high, reflecting strong acceptance of VTR as a practical support tool for EFL video-making tasks.

5. Conclusion

Overall, the results suggest that VTR provides useful sentence suggestions and video summaries to support EFL video-making tasks, with positive perceptions of usefulness and ease of use. Output quality varied by model and video length, with stronger GPT models producing higher-quality and more acceptable outputs, while longer videos remained more challenging and prone to missing or vague information. This study has several limitations, including a small sample size

and the use of a fixed dataset, which may limit generalizability. Future work should involve more diverse participants and real classroom settings, examine learning outcomes, and improve the VTR pipeline to enhance content coverage, specificity, and consistency, particularly for longer videos.

References

- Abdelhadi, A. & Belaid, L. (2025). Teachers' Perceptions of Implementing Student-Generated Videos as an Authentic Assessment Tool in English Language Learning. *Journal of Education, Society & Multiculturalism*, 6(1), 2025. 81-101. <https://doi.org/10.2478/jesm-2025-0006>
- Al Ghafri, M. S. (2023). The effect of employing authentic video projects on EFL (English as a foreign language) students' motivation to learn English. *European Economic Letters*, 13(3), 1327–1332. <https://doi.org/10.52783/eel.v13i3.435>
- Chang, H.-W., Liu, Y.-F., Hwang, W.-Y., & Chen, W.-C. (2025). Developing an AI driven contextualized short video learning system for EFL speaking and writing. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT 2025)*, (pp. 148–152). IEEE. <https://doi.org/10.1109/ICALT64023.2025.00048>
- Hsu, H. (2024). Evaluating the Impact of Video-Making Projects on English for Specific Purposes Learning: A Case Study. *International Journal of Language and Education Research*, 6(1), 1-23. <https://doi.org/10.29329/ijler.2024.661.1>
- Hwang, W.-Y., Luthfi, M. I., & Liu, Y.-F. (2024). AI enhanced video drama making for improving writing and speaking skills of students learning English as a foreign language. *Innovation in Language Learning and Teaching*, 18, 1–18. <https://doi.org/10.1080/17501229.2024.2437655>
- Liu, Y.-F., Luthfi, M. I., & Hwang, W.-Y. (2024). Developing an AI enhanced video drama making learning system to support EFL learners in authentic contexts. In *Proceedings of the Global Chinese Conference on Computers in Education (GCCCE 2024): Main Conference (English Papers)* (pp. 34–37).
- Liu, Y.-F., Luthfi, M. I., & Hwang, W.-Y. (2025). Enhancing EFL writing through AI driven video to text recognition in authentic learning contexts. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT 2025)* (pp. 166–168). IEEE. <https://doi.org/10.1109/ICALT64023.2025.00053>
- Liu, Y.-F., Luthfi, M. I., & Hwang, W.-Y. (2025). A pilot study on bridging EFL writing and speaking skills through AI enhanced authentic short video making. In *Proceedings of the Global Chinese Conference on Computers in Education (GCCCE 2025)* (pp. 62–69).
- Ma, Q. M. & Bruce, D. L. (2025). Integrating Digital Video Composing Across Three Settings in Chinese English Classrooms. In J. Tussey & L. Haas (Eds.), *Supporting Linguistic Differences Through Literacy Education* (pp. 325-360). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3373-3496-7.ch010>
- Shang, Y., Gao, C., Li, N., & Li, Y. (2025). A large-scale dataset with behavior, attributes, and content of mobile short-video platform. In *Companion Proceedings of the ACM on Web Conference 2025 (WWW '25)* (pp. 793–796). Association for Computing Machinery. <https://doi.org/10.1145/3701716.3715296>
- Yeh, H.-C., Heng, L., & Tseng, S.-S. (2021). Exploring the impact of video making on students' writing skills. *Journal of Research on Technology in Education*, 53(4), 421–437. <https://doi.org/10.1080/15391523.2020.1795955>

AI Agents as Cognitive Partners in Micro-Project-Based Learning: A Professional Development Workshop

Yuhong Zhao^{1*}, Qing Zhang¹, Weiwei Li¹

¹ Xi'an Jiaotong-Liverpool University

* yawnzha@gmail.com

Abstract: This case study examines how an LLM-based AI agent supported adult learners in professional development during a micro-project-based learning (micro-PBL) workshop. In this activity, learners developed their individually owned “ideas worth spreading” into mini TED-style talks through iterative cycles of reflection, critique, and public presentation. Interviews with 10 participants revealed four generative themes underlying learners’ interactions with the AI agent: (1) momentum restoration via cognitive extension under time pressure to overcome a cognitive impasse; (2) role negotiation, in which learners repositioned AI from tool to partner; (3) appropriate reliance, enacted through trust calibration routines; and (4) affect regulation, providing a psychologically safe space for experimentation. Building on these themes, we propose four design principles for AI-supported micro-PBL: ensure learner ownership, design for momentum, orchestrate multi-source feedback, and embed calibration routines. The study offers an explanatory account of how AI becomes a “cognitive partner” in compressed, ill-structured learning tasks.

Keywords: LLM-based AI agent, micro-project-based learning, adult learning, professional development, design principles

1. Introduction

Large language model (LLM) systems are increasingly used not only for information retrieval but also for ideation, structuring, drafting, and iterative refinement in everyday learning and work (Dwivedi et al., 2023; Mollick, 2024). This shift is particularly consequential in adult professional learning, where participants often work under time pressure, produce public-facing artefacts, and must balance efficiency with authenticity. In such settings, AI may function not simply as a tool, but as an interactive partner that shapes how learners think, decide, and move forward, echoing recent arguments that AI in education is increasingly experienced as a meaning-making subject rather than merely an instructional object (Veletsianos et al., 2024).

This paper focuses on micro-project-based learning (micro-PBL), a compressed workshop-based form of project learning in which learners create an authentic artefact through rapid cycles of reflection, critique, and revision. In our case, participants developed individually owned “ideas worth spreading” into mini TED-style talks. Compared with conventional PBL, this form of learning is marked by temporal scarcity, rapid decision-making, and heightened social accountability. These features can intensify cognitive load, especially when learners must generate structure from ambiguous ideas while maintaining a personal voice. From a learning design perspective, scaffolding can reduce extraneous cognitive load and support higher-order thinking (Belland, 2017; Sweller, 1988; Sweller et al., 2019).

In this context, dialogic AI may serve as a dynamic scaffold. It can externalise structure, propose alternatives, and support rapid iteration. At the same time, AI-supported learning introduces new challenges. Learners must decide what role the AI should play, how much to rely on its outputs, and how to preserve ownership over a personally meaningful public product. These issues are especially salient when the final artefact is not only technically adequate but also rhetorically authentic. Effective human–AI collaboration, therefore, depends not only on assistance but also on appropriate reliance and trust calibration (Lee & See, 2004). In addition, because micro-PBL often involves personal

expression and public presentation, psychologically safe conditions remain important for experimentation and reflection (Edmondson, 2018).

Building on cognitive load theory and research on trust in automation, this study examines how adult learners experienced AI support during a sprint-formatted micro-PBL workshop. Rather than treating AI as a generic “benefit,” we focus on the mechanisms through which it shapes learners’ task progress, role negotiation, reliance decisions, and affective experience. The guiding research question is: How do adult learners experience AI as a cognitive partner in a time-bounded micro-PBL workshop?

We argue that, in compressed project settings, AI becomes a cognitive partner not by replacing learner agency, but by helping learners restore momentum, negotiate functional roles, calibrate trust, and explore ideas with lower social risk.

2. Method

This study adopted an exploratory qualitative case study design centred on one professional development workshop (Yin, 2017). The broader cohort included 15 adult professionals from diverse industries, all with at least five years of work experience, and 10 participants volunteered for semi-structured post-workshop interviews. The workshop was structured as a sprint-formatted micro-PBL activity in which learners developed individually owned mini TED-style talks for public delivery. This bounded case design allowed in-depth examination of how adult learners interacted with AI across a compressed project cycle (Yin, 2017). Participants used DeepSeek-R1 and Doubao via a browser interface, with facilitator-provided examples for ideation, outlining, revision, and rehearsal, while retaining freedom to adapt prompts to their own workflow. All participants provided informed consent, and data were anonymised.

Data came primarily from semi-structured interviews, with interaction logs used as supplementary evidence. The interviews focused on how participants used AI across different stages of the project cycle, including idea selection, drafting, feedback integration, revision, and preparation for public delivery. Thematic analysis was used to identify recurring cross-case patterns (Braun & Clarke, 2006). Transcripts were coded line by line, codes were grouped into categories, and categories were refined into broader themes. To strengthen analytic rigour, we triangulated interviews with interaction logs, maintained an audit trail, and used peer debriefing. Interview prompts elicited both frustrations and successes.

3. Findings and Discussion

3.1 *AI restored momentum under time pressure*

The first theme concerned momentum restoration. In this sprint-formatted workshop, participants frequently described AI as useful when they were cognitively stuck, short on time, or unsure how to structure an emerging idea. Rather than simply “saving time,” AI helped convert vague thoughts into workable outlines, examples, phrasing options, and next steps. In this sense, AI functioned as a dynamic scaffold for overcoming blank-page paralysis.

This theme was strongly reflected in the initial coding, where participants associated AI with workload reduction, cognitive kickstarting, and structured sensemaking. One participant noted that when they could not immediately generate content, AI helped expand aspects of the topic; another described AI as providing systematic structure when framework thinking was lacking. The interaction-log examples similarly showed learners using iterative prompting to generate “golden sentences” or more suitable speech endings under time pressure.

From a theoretical perspective, this finding can be interpreted through cognitive load theory. In compressed project work, learners face substantial extraneous demands: organising ideas, making rapid rhetorical choices, and translating personal experience into a coherent public product. AI appeared valuable when it reduced these burdens and helped learners re-enter productive thought. Importantly, participants did not describe AI as doing the thinking for them. Rather, they used it to regain pace and free cognitive resources for higher-order judgment (Sweller, 1988; Sweller et al., 2019). This suggests that the value of AI lies less in full automation than in providing momentum during cognitive impasses.

3.2 *AI’s role was negotiated rather than fixed*

A second theme was role negotiation. Participants did not treat AI as a stable or singular entity. Instead, they framed it differently depending on task stage and purpose: sometimes as an assistant, sometimes as an amplifier of their own capability, and at other times as a reviewer or sparring partner.

What matters here is not anthropomorphism for its own sake, but functional positioning. Participants actively decided what they wanted AI to do and, just as importantly, what they did not want it to do. This was especially visible when they were writing for public speaking. Several participants were willing to ask AI for options, structures, or stylistic alternatives, but resisted letting it produce the final narrative voice. One representative example from the study showed a learner asking the system for five narrative styles while explicitly instructing it not to write the full script. The learner explained that “the final voice must be mine.”

This finding clarifies what “cognitive partner” means in practice. AI was not experienced as inherently partner-like. It became partner-like when learners pragmatically assigned it a role that served their current need while preserving personal ownership of the final product. This interpretation also aligns with recent discussions of AI as a collaborative rather than purely instrumental presence in learning and knowledge work (Mollick, 2024; Veletsianos et al., 2024).

3.3 Trust depended on calibration, not acceptance

The third theme concerned trust calibration. Participants repeatedly noted that AI output could be useful yet limited. They described errors, superficial responses, and misalignment with their actual intentions. However, these limitations did not necessarily lead them to abandon AI. Instead, participants engaged in active calibration behaviours, such as verifying outputs, rephrasing prompts, asking the model to explain its reasoning, and retaining human control over final decisions.

AI support was therefore not effective because learners trusted it blindly. It was effective when learners treated AI suggestions as provisional and compared them against personal intent, audience fit, and, where available, peer or mentor input. One interaction log example showed a participant asking the model to adjust an output “based on your reasoning” to debug why the response had deviated from their original intent. Rather than accepting or rejecting the AI in a binary way, the learner used explanation-seeking to calibrate reliance.

This theme aligns with trust-in-automation research, which emphasises appropriate reliance rather than uncritical acceptance (Lee & See, 2004). It also extends that literature into a pedagogical context where the final product is public-facing and personally meaningful. When learners are preparing a talk that reflects their own experience, trust becomes both epistemic and identity-related: they are not only checking whether something is correct, but also whether it sounds like them. This suggests that effective design should make verification and comparison routines an explicit part of the learning process.

3.4 AI provided a low-risk space for rehearsal and exploration

The fourth theme was affect regulation. Beyond cognition and task progress, participants described AI as a psychologically easier space in which to test unfinished ideas, rehearse expression, or think aloud without immediate human judgment.

This mattered because the workshop task involved personal storytelling and public speaking. In such conditions, learners are not only producing content; they are exposing identity, uncertainty, and rhetorical vulnerability. AI appeared helpful because it allowed early experimentation without the interpersonal pressure that often comes with peer or mentor critique. Participants could try out half-formed ideas, ask awkward questions, or rehearse phrasing with lower embarrassment. One participant even described playful banter with the AI during preparation, suggesting that the interaction provided emotional support as well as practical support.

At the same time, this did not mean emotional dependence without boundaries. Participants valued convenience and low judgment, but they still distinguished AI from human relationships and retained awareness that it was a one-way support system rather than a reciprocal social bond. This theme is important because it shows that AI as a cognitive partner

is not only about reasoning assistance. In compressed learning environments, affect and cognition are intertwined, and psychologically safer conditions can help learners sustain experimentation and revision (Edmondson, 2018).

4. Conclusion and Implications

This study suggests that AI becomes a cognitive partner in micro-PBL not because it autonomously produces strong outputs, but because learners use it to regain momentum, assign functional roles, calibrate reliance, and explore ideas with lower social risk. In this workshop setting, AI support was most valuable when it protected learner ownership while reducing the friction of compressed project work.

Four concise design principles follow. First, AI-supported micro-PBL should preserve learner ownership, especially when the output is public-facing and identity-relevant. Second, learning design should support momentum restoration during cognitive impasses through structured yet flexible scaffolding (Belland, 2017; Sweller et al., 2019). Third, pedagogy should orchestrate multi-source feedback, enabling learners to compare AI, peer, and mentor input without losing narrative control. Fourth, teachers should embed trust-calibration routines, including verification, comparison across feedback sources, and explicit human accountability for public-facing outputs (Lee & See, 2004).

This paper is exploratory and context-specific. The study involved a small, self-selected interview sample and focused on an individually owned workshop task rather than collaborative team PBL, both of which limit transferability. Future work could compare AI support with human coaching or static scaffolds and examine whether these cognitive-partnership patterns persist beyond a single workshop cycle.

References

- Belland, B. R. (2017). *Instructional scaffolding in STEM education: Strategies and efficacy evidence*. Springer Nature. <https://doi.org/10.1007/978-3-319-02565-0>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). Opinion paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>
- Edmondson, A. C. (2018). *The fearless organization: Creating psychological safety in the workplace for learning, innovation, and growth*. Wiley.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- Mollick, E. (2024). *Co-intelligence: Living and working with AI*. Portfolio.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2), 261–292. <https://doi.org/10.1007/s10648-019-09465-5>
- Veletsianos, G., Houlden, S., & Johnson, N. (2024). Is artificial intelligence in education an object or a subject? Evidence from a story completion exercise on learner-AI interactions. *TechTrends*, 68(3), 411–422. <https://doi.org/10.1007/s11528-024-00942-5>
- Yin, R. K. (2017). *Case study research and applications: Design and methods* (6th ed.). SAGE.

Effect of Design Thinking on Students' Creativity: A meta-analysis

Fuminqiang Guo¹, Junhao Shan^{2*},

^{1,2} College of Education Science and Technology

Nanjing University of Posts and Telecommunications Nanjing, China

*jhshan@njupt.edu.cn

Abstract: Design thinking is a prevalent pedagogical approach in teaching and learning, aimed at cultivating students' creativity. While a substantial corpus of empirical studies has explored the efficacy of design thinking pedagogy in this regard, a systematic approach to analyzing these studies remains underdeveloped. In order to provide a uniform assessment of the utility of design thinking pedagogy and to explore the effects of moderating variables, this study used a meta-analytic approach to analyze 24 empirical studies. The results showed that design thinking exhibited a moderate effect on the enhancement of students' creative ($g = 0.726$, $p = 0.000$). The moderating effects of educational level, learning environment, and design thinking methods on teaching effectiveness were also analyzed. The exploration of future research directions was also conducted on the basis of this study.

Keywords: design thinking, meta-analysis, creativity

1. Introduction

The demand for creative talents is increasing, and greater emphasis is being placed on the cultivation of such talents (Zhang et al., 2024), which must be grounded in specific teaching methodologies or strategies to facilitate a conducive learning environment for students, one that fosters the generation of ideas and the design of products. A plethora of pedagogical approaches for cultivating students' creativity have been documented on an international scale, while the most extensively researched approach is the design thinking pedagogy. Design thinking instruction is defined as an innovative pedagogy that integrates diverse thinking methods through a standardized teaching process in order to systematically cultivate student creativity (Chin et al., 2019; Graves & Fuchs, 2022). Numerous empirical studies have been conducted to ascertain the efficacy of design thinking pedagogy in fostering students' creativity. However, the findings from these studies have been inconsistent. Can design thinking promote students' creativity? This question requires urgent attention.

2. Literature Review

Design thinking is a teaching methodology aimed at improving students' creativity. The Hasso Plattner Institute of Design at Stanford University (2023) proposed the EDIPT design thinking teaching model, which has been adopted by many research institutions and primary and secondary schools to standardize the teaching process and cultivate creative talents. Internationally, many studies have focused on the effectiveness of the design thinking instruction in cultivating students' creativity, but the conclusions are not the same. For example, in K-12 schools, Tan (2024) used the design thinking pedagogy in a maker curriculum, and through quasi-experimental research, concluded that the design thinking instruction can effectively enhance the creativity of primary and secondary school student; In higher education, Kuo (2021) hold an empirical study which indicated that design thinking is effective in improving student creativity. However, some studies have also given the opposite answer, for example, Kijima (2021) implemented design thinking instruction at K-12 level and found that the promotion of student motivation on creativity was weak; And according to Liu (2024), at higher education level, the enhancement of student creativity through design thinking instruction was proven to be minimal.

The above studies show that academic research on design thinking instruction to promote student creativity has mostly used quasi-experimental research methods, and it is difficult to give an accurate answer about the effectiveness of the application of design thinking instruction from these isolated findings. Therefore, this study adopted a meta-analytic approach to synthesize and evaluate the results of previous research on design thinking in education, aiming to gain a more systematic understanding of the actual impact of design thinking instruction on student creativity. In addition, the study explored the role of various moderating variables, which include educational level, learning environment, and design thinking method (Bai et al., 2020).

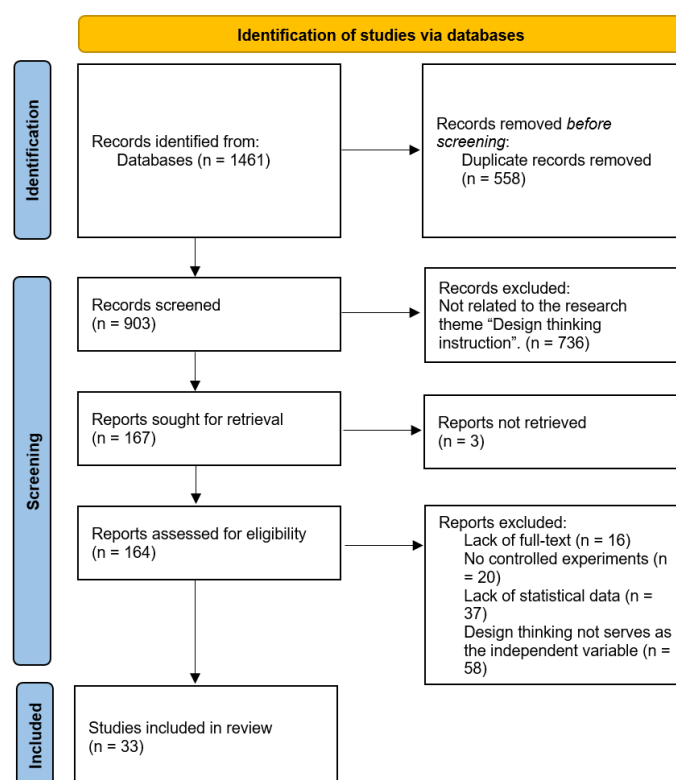
3. Research Methods

3.1. Literature Search

To ensure the acquisition of comprehensive and sufficient literature, based on The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Page et al., 2021), we conducted an extensive search of internationally renowned academic databases, including CNKI, Web of Science, Taylor & Francis, Wiley, Springer, Elsevier and ProQuest. The search employed the keywords "design thinking" or "design thinking instruction". The literatures published between 2015 and 2025 were selected and limited to the field of education. A flowchart for literature searching was constructed, as illustrated in Figure 1.

Figure 1

PRISMA flow chart



3.2. Study Selection

To ensure that the final screened literature is sufficient in terms of data to support meta-analysis and is in line with the purpose of this study, we have established the criteria for literature screening. Specifically as follows:

- It is necessary to study the influence of design thinking on the improvement of students' creativity.
- It must be an empirical study employing a randomized controlled trial or quasi-experimental design.
- This study should include the following data: mean, standard deviation, sample size, t-value, p-value, etc., to meet the data requirements of meta-analysis.

A total of 1032 literatures were obtained from the initial search. Then based on the above criteria for screening, a total of 24 literatures were finally suitable for inclusion in the meta-analysis.

3.3. Literature Code

The literature is encoded according to the following parameters:

- Author.
- Publication year.
- Educational level: university and K-12.
- Learning environment: traditional classrooms, outdoor settings, and maker spaces.
- Design thinking methods: The coding focuses on whether methods are used in each study.

Coding accuracy assessment: The Cohen Kappa consistency coefficient is required to be greater than 0.7 (Cohen, 1969). The inter-rater reliability in this study was 0.95, indicating a strong consistency among the researchers.

4. Results

4.1. Heterogeneity Analysis

Heterogeneity analysis was conducted with Cochran's Q and I² statistics (Hedges & Olkin, 1985; Higgins & Thompson, 2002;), and then we determined the type of model utilized for the meta-analysis. As shown in Table 1, the Q value of this meta-analysis is 125.598, the P-value is 0.000, and the I² is 72.133, which exceeded 50% (Higgins, 2011). Therefore, a random effects model was employed in this study to calculate the effect size.

4.2. Calculation and Combination of Effect Sizes

4.2.1. Overall effectiveness

As shown in Table 1, statistical data such as mean, standard deviation, sample size, t-value and p-value were collected from 24 empirical studies. The standardized mean difference (SMD) formula was used to calculate the effect size, with Hedges' g serves as an indicator of the effect size.

A random-effects model for effect sizes was applied, yielding a Hedges' g of 0.726. According to the thresholds proposed by Cohen (1988) to standardize effect sizes, 0.2-0.49 is considered a small effect, 0.5-0.79 a medium effect, and 0.8 and above a large effect. Therefore, the effect sizes derived from this meta-analysis were categorized as medium. Overall, design thinking instruction has a moderate positive impact on student creativity.

Table 1.

Overall effectiveness.

Effect size and 95% confidence interval						Test of null (2-Tail)		Heterogeneity		
Model	N	Hedges's g	Standard error	Lower limit	Upper limit	Z-value	P-value	Q	P	I-squared
Random	24	0.726	0.066	0.596	0.856	10.948	0.000	125.598	0.000	72.133

4.2.2. Moderator analysis

- Educational level:

As illustrated in Table 2, the differences in effect sizes for the two educational levels—university (g=0.773, N=13, p=0.000), and K-12 (g=0.708, N=23, p=0.000)—are not statistically significant (Q=0.190, P=0.663).

- Learning environment:

There was a significant difference in the effect of learning environment on the effectiveness of design thinking instruction (Q=16.865, P=0.000). Makerspaces (g=0.797, N=21, p=0.000) were the most effective, followed by traditional classrooms (g=0.759, N=11, p=0.000) and outdoor settings (g=0.376, N=4, p=0.000).

- Design thinking methods:

The effectiveness of design thinking instruction was influenced by the methods used, with instruction applying design thinking methods being more effective ($g=0.815$, $N=28$, $p=0.000$) than those lacking any design thinking method ($g=0.402$, $N=8$, $p=0.000$). And a statistically significant difference was found ($Q=13.602$ $P=0.000$).

Table 2.

Moderator analysis.

Moderator factors	Group	N	Hedges 'g	Standard error	Lower limit	Upper limit	Z-value	P-value	Q-value	P-value
Education level	K-12	23	0.708	0.076	0.559	0.857	9.305	0.000		
	university	13	0.773	0.130	0.017	0.518	1.028	0.000		
	Total between								0.190	0.663
Learning environment	Overall	36	0.724	0.066	0.596	0.853	11.033	0.000		
	makerspaces	21	0.797	0.087	0.627	0.968	9.185	0.000		
	outdoor settings	4	0.376	0.069	0.240	0.511	5.439	0.000		
	traditional classrooms	11	0.759	0.131	0.503	1.016	5.796	0.000		
	Total between								16.865	0.000
Design thinking methods	Overall	36	0.571	0.050	0.473	0.669	11.432	0.000		
	Not using	8	0.402	0.086	0.235	0.570	4.701	0.000		
	Using	28	0.815	0.072	0.674	0.957	11.299	0.000		
	Total between								13.602	0.000
	Overall	36	0.644	0.055	0.536	0.752	11.669	0.000		

4.3. Sensitivity Analysis

To ensure the robustness of the meta-analysis results, a sensitivity analysis was conducted in this study. By excluding each study one by one, it was found that the effect size always remained within the 95% confidence interval (95% CI=[0.596, 0.856]) and was statistically significant.

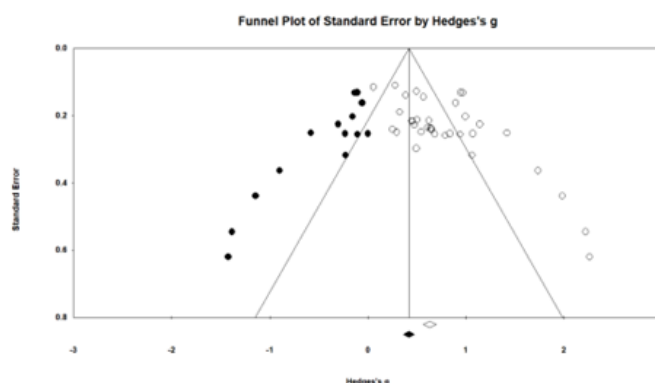
4.4. Publication Bias

Publication bias is the phenomenon whereby statistically significant studies are more likely to be published than non-significant ones (Borenstein et al., 2009), which can significantly influence the results of a meta-analysis.

As illustrated in Figure 2, the distribution of the white filled sample points on both sides of the central axis shows asymmetry, indicating a significant deviation. In order to assess publication bias more precisely, we performed Egger's test (Card, 2011). The results revealed the presence of publication bias, with the following statistics: INTERCEPT = 2.990, 95% CI = [1.458, 4.521], $t = 3.967$, $p = 0.000$. The effect size was subsequently adjusted using the trim-and-fill method (Jeong et al., 2019; Duval & Tweedie, 2000) as shown in Figure 2, where black dots indicate the addition of potentially missing studies. The adjusted effect size was $g = 0.445$, $Q = 309.893$, 95% CI = [0.298, 0.592], suggesting that after adjusting for publication bias, the intervention still has a moderate positive effect; although the original results were overestimated, the core conclusions remain robust.

Figure 2

Funnel Plot (corrected by the trim-and-fill method)



5. Discussion and Conclusion

5.1. Major Findings

5.1.1. Overall effect

Overall, the design thinking teaching method can help enhance students' creativity and holds certain instructional value for the development of their creative thinking. The design thinking teaching method generates certain educational value in both basic education and higher education. The effective implementation of design thinking teaching relies on standardized teaching procedures and clear, practical cognitive scaffolds.

5.1.2. Analysis of moderating variables

- Educational level:

University students have demonstrated greater gains than their K-12 counterparts with design thinking. The reason may be due to the fact that university students have a higher level of knowledge, expertise, and creative experience. On the other hand, university equipment and instruction materials are generally more advanced than those available in K-12 schools, and are more capable of assisting students in their creative activities and enhancing their creativity.

- Learning environment:

The efficacy of design thinking instruction is contingent upon the learning environment, with makerspaces yielding the most significant outcomes, followed by traditional classrooms and outdoor settings. The makerspaces provide tools and equipment that can facilitate creative practices, which may lead to a stronger enhancement of the effectiveness of design thinking instruction in makerspaces. Less effective in outdoor settings, possibly due to the lack of systematic norms for teaching in such environments and the low motivation of students to learn.

- Design thinking methods:

The employment of design thinking methods also significantly influences student learning outcomes. An instruction approach that incorporates design thinking methods has proven to be more effective. The reason is that the scaffolds of design thinking provide students with a framework for innovative problem-solving, help standardize the creative process, train logical thinking, and enhance divergent thinking.

5.2. Major Contributions

The conclusions of this study emphasize the beneficial effects of design thinking instruction on student creativity, and suggest that design thinking constitutes a significant instruction method for fostering core competencies among learners. In addition, this study outlines the impact of moderating variables on instructional effectiveness. Therefore, educators are encouraged to select design thinking approaches wisely and be mindful of moderating factors in order to improve student learning outcomes.

5.3. Limitations

Despite the reliability of the study's conclusions, certain limitations remain. Notably, publication bias may have distorted the results of the meta-analysis. In addition, due to missing data in several empirical studies, they were excluded from the meta-analysis.

5.4. Suggestions

5.4.1. Suggestions for future studies

Although this study affirmed the positive impact of design thinking instruction, it is important to be aware of the possible negative aspects. Therefore, future research efforts should accumulate a more comprehensive sample size, which would facilitate the reliability of subsequent meta-analyses and allow for a more comprehensive analysis of moderator variables.

5.4.2. Suggestions for instructional designers and teachers

Design thinking instruction needs to focus on the application of advanced technology and thinking scaffolds, empowering the upgrading of the teaching environment through digital intelligence technology. For example, generative artificial intelligence can be utilized to build intelligent agent to assist students in their personal creative activities.

Furthermore, design thinking methods should be appropriately embedded in instruction strategies, such as the use of brainstorming and six-hat thinking method to facilitate students' teamwork, creative thinking dispersion, and holistic thinking about the problem so as to improve their creativity and the outcome of work.

Acknowledgements

This paper is a research outcome of the 2024 Jiangsu Provincial Social Science Fund Youth Annual Project titled "Research on the Model Construction of Enhancing Teachers' Interdisciplinary Teaching Ability Empowered by Generative Artificial Intelligence" (Project No.: 24JYC005).

References

- Bai, S., Hew, K. F., & Huang, B. (2020). Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review*, 30, Article 100322.
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.
- Card, N. A. (2011). *Applied meta-analysis for social science research*. Guilford Publications.
- Chin, D. B., Blair, K. P., Wolf, R. C., Conlin, L. D., Cutumisu, M., Pfaffman, J., & Schwartz, D. L. (2019). Educating and measuring choice: A test of the transfer of design thinking in problem solving and learning. *Journal of the Learning Sciences*, 28(3), 337–380.
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. Routledge.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge.
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2), 455–463.
- Graves, M., & Fuchs, J. (2022). Teaching the design thinking mindset: A practitioner's perspective. In I. Schmidberger, S. Wippermann, T. Stricker & U. Müller (Eds.), *Design Thinking im Bildungsmanagement: Innovationen in Bildungskontexten erfolgreich entwickeln und umsetzen* (pp. 69–78). Springer Fachmedien Wiesbaden.
- Hasso Plattner Institute of Design at Stanford University. (2023). Get started with design thinking. <https://dschool.stanford.edu/resources/getting-started-with-designthinking>.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.
- Higgins, J. P., Altman, D. G., Gøtzsche, P. C., Jüni, P., & Moher, D. et al. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343, Article d5928.

- Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11), 1539–1558.
- Jeong, H., Hmelo-Silver, C. E., & Jo, K. (2019). Ten years of computer-supported collaborative learning: A meta-analysis of CSCL in STEM education during 2005-2014. *Educational Research Review*, 28, Article 100284.
- Kijima, R., Yang-Yoshihara, M., & Maekawa, M. S. (2021). Using design thinking to cultivate the next generation of female STEAM thinkers. *International Journal of Stem Education*, 8, 1–15.
- Kuo, H. C., Yang, Y. T. C., Chen, J. S., Hou, T. W., & Ho, M. T. (2021). The impact of design thinking PBL robot course on college students' learning motivation and creative thinking. *IEEE Transactions on Education*, 65(2), 124–131.
- Liu H. Y. (2024). Effects of the design thinking pedagogy on design thinking competence of nursing students in Taiwan: A prospective non-randomized study. *Nurse education today*, 138, 106197.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., & Hoffmann, T. C. et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, Article 71.
- Tan, J., Kinshuk & Wu, L. The effectiveness of design thinking on K-12 school students' creativity in a maker curriculum. *Education Tech Research Dev* 72, 1091–1110 (2024).
- Zhang, Y., Wang, Y., & Su, Y. (2024). From no distinction to direction: The impact of motivation on creativity—The contribution of environmental resources and grit. *Thinking Skills and Creativity*, 53, Article 101609.

CET Professionals' AI Readiness and the Integration of Generative AI in Professional Education: A Mixed-Methods Descriptive Study

Lishan Zheng¹, Doris Choy^{1*},

¹ Learning Sciences and Assessment Department,

National Institute of Education, Nanyang Technological University, Singapore

* doris.choy@nie.edu.sg

Abstract: Generative artificial intelligence (GenAI) is being adopted by Continuing Education and Training (CET) professionals to support their training and development design, yet little is known about the CET professionals' baseline readiness to adopt AI responsibly and how they concretely integrate AI tools into training and development work. This mixed-methods descriptive study examined (1) AI readiness across cognition, ability, vision, and ethics using an adapted validated survey (Wang et al., 2023), and (2) CET professionals' hands-on experiences in the integration of GenAI tools (e.g., ChatGPT, etc.) when designing training packages (8–24 hours). Forty-eight CET professionals completed the AI-readiness survey; a subset participated in follow-up interviews/focus groups. Survey results suggested generally high readiness across all four dimensions, with comparatively stronger in their “vision” and “cognition”. Qualitative findings portrayed GenAI as a “behind-the-scenes co-designer” used for brainstorming, outlining, case scenarios generation, and iteration. The CET professionals also shared their common concerns about privacy, bias, transparency, and over-reliance of AI tools. Participants reported readiness-building strategies centered on prompt experimentation, verification routines, and alignment to workplace constraints. Implications include embedding prompt-and-verify workflows, scenario-based ethics cases, and human-centered governance into professional development and education.

Keywords: AI readiness; generative AI; CET professionals; instructional design; professional education

1. Introduction

The rapid development of generative artificial intelligence (GenAI), particularly large language models and conversational AI systems, is reshaping how knowledge is produced, accessed, and applied across professional domains (Kasneci et al., 2023). In professional education and workforce development, GenAI has shown potential to support instructional design, content generation, assessment development, and reflective practice (Holmes & Tuomi, 2022). These developments are especially relevant for Continuing Education and Training (CET) professionals, who are responsible for designing industry-relevant and practice-oriented learning experiences for adult learners in rapidly changing skill environments (OECD, 2021).

However, the adoption of GenAI in professional training cannot be assumed to be straightforward. Prior research suggests that the integration of emerging technologies depends not only on the availability of tools, but also on educators' and trainers' readiness to understand, evaluate, and use them meaningfully in practice (Ertmer & Ottenbreit-Leftwich, 2010). In this regard, AI readiness has been conceptualized as a multidimensional construct that includes cognition, ability, vision, and ethics (Wang et al., 2023). This framework provides a useful lens for examining whether professionals are prepared to adopt AI responsibly and effectively.

Existing studies have explored AI readiness and GenAI use mainly among school teachers and university students, while research on CET professionals remains limited. This is an important gap because CET contexts differ from formal education in their strong industry orientation, emphasis on adult learning, and responsiveness to workforce needs

(Cedefop, 2020). As such, CET professionals may experience AI adoption in distinct ways that are not fully captured in prior studies.

To address this gap, this study adopts a descriptive mixed-methods approach to examine CET professionals' AI readiness and their integration of GenAI tools in training and development practices. Specifically, it addresses the following research questions:

RQ1. What is the level of AI readiness among CET professionals across cognition, ability, vision, and ethics, as measured using an adapted validated survey instrument?

RQ2. How do CET professionals integrate GenAI tools into their training and development practices, what are their concerns, and how do they prepare themselves to adapt AI in the design of training practices?

2. Method

This study employed a descriptive mixed-methods design to examine CET professionals' AI readiness and their integration of generative AI tools in training and development practices. Participants were CET professionals enrolled in a postgraduate Training Methods elective at a higher education institution in Singapore. The course focused on design and development training for adult learners in professional education contexts. Participation was voluntary, and informed consent was obtained prior to data collection. As part of the course, participants engaged in scaffolded, hands-on exploration of generative AI tools to support training design tasks, including content development, assessment planning, and iterative refinement of training materials. AI readiness was measured using an adapted version of the validated survey instrument developed by Wang et al. (2023), which assesses four dimensions: cognition, ability, vision, and ethics, using a five-point Likert scale. To complement the survey findings, qualitative data were collected through six semi-structured interviews with a subset of participants. The interviews focused on how participants used generative AI in training design, the concerns they encountered, and the strategies they adopted to use AI responsibly in practice. Quantitative survey data were analyzed using descriptive statistics, with means and standard deviations calculated for each AI readiness dimension. Qualitative data were analyzed using descriptive thematic analysis to identify recurring patterns in participants' AI integration practices, reported concerns, and readiness-building strategies.

3. Findings and Discussion

3.1 The level of AI readiness among CET Professionals across Cognition, Ability, Vision, and Ethics (RQ1)

Descriptive analysis showed that CET professionals demonstrated generally positive AI readiness across all four dimensions. Vision received the highest mean score ($M = 3.86$, $SD = .466$), followed by Ethics ($M = 3.71$, $SD = .634$), Cognition ($M = 3.64$, $SD = .635$), and Ability ($M = 3.50$, $SD = .618$). These results suggest that participants were generally optimistic about the value of AI in training and were aware of ethical issues, although their confidence in hands-on application was comparatively lower. Participants also reported positive perceptions of AI-enhanced innovation ($M = 3.81$, $SD = .539$), while perceived threats remained relatively neutral ($M = 3.20$, $SD = .877$). Overall, the findings indicate that CET professionals were generally ready to adopt AI in training practice, but further support may be needed to strengthen their practical ability. Cronbach's alpha coefficients ranged from .72 to .83 for all six variables, indicating acceptable reliability for the adapted survey.

Table 1.

Descriptive Statistics for AI-Related Subscales and Items

Subscale	Items	N	Mean	SD	Cronback Alpha
Cognition	CO1, CO2, CO3, CO4, CO5	48	3.64	0.635	0.72
Ability	AB1, AB2, AB3, AB4, AB5, AB6	48	3.50	0.618	0.83

Subscale	Items	N	Mean	SD	Cronback Alpha
Vision	VI1, VI2, VI3	48	3.86	0.466	0.78
Ethics	ET1, ET2, ET3, ET4	48	3.71	0.634	0.72
Perceived Threats	PT1, PT2, PT3, PT4, PT5	48	3.20	0.877	0.79
AI-enhanced Innovation	INN1, INN2, INN3, INN4, INN5	48	3.81	0.539	0.77

Note. All items were rated on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree). Subscales Cognition, Ability, Vision, and Ethics constitute the AI readiness framework adapted from Wang et al. (2023). The sources of the remaining subscales are reported in Appendix A.

3.2 Integration of GenAI tools in CETs' Training and Development Practices (RQ2)

Qualitative analysis revealed three recurring patterns. First, CET professionals mainly used GenAI as a design support tool rather than a replacement for professional expertise. Common uses include generating ideas, refining learning objectives, developing examples or case scenarios, and supporting assessment design. Second, participants expressed concerns about accuracy, over-reliance, confidentiality, and bias, and emphasized the need to verify AI-generated outputs before applying them in practice. Third, participants described AI readiness as an ongoing developmental process, involving prompt experimentation, boundary-setting, and reflection on when and how AI should be used responsibly. Overall, the findings suggest that CET professionals adopted GenAI in a cautious and selective way, using it to augment rather than replace professional judgment. Additional details will be presented at the conference.

4. Discussion

This study provides descriptive insights into AI readiness among CET professionals and the integration of GenAI in training design. Consistent with the AI readiness framework proposed by Wang et al. (2023), participants reported relatively higher levels of vision and ethics than ability, suggesting generally positive orientations toward AI together with awareness of responsible-use issues. This pattern also supports prior work showing that positive perceptions of AI do not automatically lead to strong operational confidence or routine implementation (Chiu & Chai, 2020).

The qualitative findings further show that CET professionals used GenAI mainly as a support tool for brainstorming, structuring, and refining training design, rather than as a substitute for professional judgment. This result is in line with prior studies that describe GenAI as useful for drafting, idea generation, and iterative design support in educational practice (Holmes & Tuomi, 2022; Rudolph et al., 2023; Zhai, 2022). At the same time, concerns about accuracy, confidentiality, bias, and over-reliance influenced how AI was used in practice, echoing earlier research on ethical accountability and constrained adoption (Akgun & Greenhow, 2022).

Finally, the strategies reported by participants, including prompt experimentation, verification, and boundary-setting, suggest that AI readiness is better understood as an ongoing developmental process rather than a fixed competence. These findings highlight the need for professional development that strengthens not only technical ability, but also responsible and context-sensitive use of AI in CET practice.

5. Limitations and Conclusion

This study offers some descriptive insights into the AI readiness of CET professionals and their integration of generative AI tools into training and development practices. The findings suggest that CET professionals generally exhibit positive AI readiness, particularly in terms of strategic vision and ethical awareness, while demonstrating more cautious and selective use of GenAI as a design support tool. These results emphasize the importance of professional development approaches that prioritize reflective, ethically informed, and practice-oriented integration of AI in training and learning.

Several limitations should be acknowledged. The study involved a relatively small sample from a postgraduate programme, which may limit the generalizability of the findings. In addition, AI integration practices were based on self-reported reflections and course artifacts rather than direct observation. Finally, the descriptive design precludes conclusions about causal relationships between AI readiness and training outcomes. Future research could address these limitations by using larger samples, employing longitudinal designs, and incorporating observational methods to enhance the understanding of these phenomena.

Acknowledgements

This project is supported by the EdeX Grant from the Centre for Teaching, Learning and Pedagogy, Nanyang Technological University.

References

- Akgun, S., & Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics*, 2(3), 431-440.
- Cedefop. (2020). Vocational education and training in Europe, 1995-2035: scenarios for European vocational education and training in the 21st century. *Cedefop reference series*, 114.
- Chiu, T. K., & Chai, C. S. (2020). Sustainable curriculum planning for artificial intelligence education: A self-determination theory perspective. *Sustainability*, 12(14), 5568.
- Ertmer, P. A., & Ottenbreit-Leftwich, A. T. (2010). Teacher technology change: How knowledge, confidence, beliefs, and culture intersect. *Journal of research on Technology in Education*, 42(3), 255-284.
- Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European journal of education*, 57(4), 542-570.
- Howard, S. K., Tondeur, J., Siddiq, F., & Scherer, R. (2021). Ready, set, go! Profiling teachers' readiness for online teaching in secondary education. *Technology, Pedagogy and education*, 30(1), 141-158.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103, 102274.
- OECD. (2021). *Teachers and leaders in vocational education and training*. OECD Publishing.
- Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of applied learning & teaching*, 6(1), 342-363.
- Wang, X., Li, L., Tan, S. C., Yang, L., & Lei, J. (2023). Preparing for AI-enhanced education: Conceptualizing and empirically examining teachers' AI readiness. *Computers in human behavior*, 146, 107798.
- Zhai, X. (2022). ChatGPT user experience: Implications for education. *Available at SSRN 4312418*. <https://doi.org/10.2139/ssrn.4312418>

A copy of the AI-readiness survey could be found [here](#)

An Ecological Framework for L2 Teachers' ICT Integration: A Theory-Driven Review (2015–2025)

Sulin He^{1*}, Joseph Roche¹

¹ School of Education, Trinity College Dublin

* suliana0157@gmail.com

Abstract: Over the past decade, the application of Information and Communication Technology (ICT) in second language (L2) teaching has become increasingly embedded. However, despite a large body of research, the literature remains fragmented and lacks a unifying framework. This paper introduces Bronfenbrenner's Ecological Systems Theory and constructs a multi-level theoretical analysis framework from the individual, teaching, organization, and student perspectives. ICT integration emerges from multi-level interactions among teachers' competencies, task design, institutional support, and student responses. This review makes three theoretical contributions: (1) It reconceptualizes L2 teachers' ICT integration as a cross-level, recursive ecological process; (2) It specifies cross-system pathways linking teacher cognition, instructional interaction, institutional conditions, and student feedback; (3) It extends existing TAM/TPACK-based explanations by embedding them within a temporal (chronosystem) perspective.

Keywords: Second language teachers, ICT integration, Ecological Systems Theory, Multi-level framework, Cross-level interactions

1. Introduction

L2 teaching is undergoing a profound transformation due to digital technologies and Artificial Intelligence. ICT provides enhanced learning opportunities and reshapes classroom interaction. Under these circumstances, teachers are positioned not just as technology users but as instructional designers, learning facilitators, and integrators. Existing studies tend to focus on influencing factors at a single level—individual, classroom, or institutional. Although these studies have provided insightful empirical evidence, they have paid limited attention to cross-level feedback loops. Building on these limitations, this paper adopts Bronfenbrenner's Ecological Systems Theory as its analytical lens. This theory conceptualizes individual behavior as embedded within a dynamic ecological system where interactions across levels jointly shape individuals' cognition, behavior, and development. Compared with prior research perspectives, an ecological approach offers a more comprehensive explanatory framework for understanding L2 teachers' ICT integration.

2. Literature Review

2.1. Individual Teacher Level

At the individual teacher level, research primarily focuses on psychological and professional readiness. Early studies mainly focused on teachers' basic acceptance of ICT and the fundamental role of technological attitudes. The Technology Acceptance Model (TAM) suggests that perceived usefulness and ease of use influence usage intentions. As research has advanced, the focus has shifted to ICT enactment, highlighting ICT self-efficacy as a key mechanism linking attitudes to actual teaching behaviors. The Technological Pedagogical Content Knowledge (TPACK) framework explains how teachers integrate technology with pedagogy and subject content. Post-pandemic studies further highlight teachers' emotional resilience, agency, and identity shifts as critical conditions for sustained ICT integration. These individual-level explanations remain insufficient unless situated within the interactional and institutional contexts.

2.2. Instructional Interaction Level

In L2 classrooms, effective integration depends on alignment with teaching objectives, learning tasks, and teacher-student interaction modes. When technology is only used for demonstration, its role in facilitating language learning is limited. In contrast, when ICT supports authentic communication and feedback mechanisms, its potential can be fully realized. Instructional interaction operates as a recursive feedback loop through which classroom outcomes inform subsequent decisions. Interactional alignment functions as a key mesosystem mechanism that converts teachers' individual readiness into observable pedagogical practice.

2.3. School and Institutional Level

Infrastructure, technical support, time resources, and managerial attitudes play an important role. Institutional support is not limited to equipment; it includes troubleshooting and resource sharing. Indifferent institutional attitudes and weak incentive systems can suppress adoption even when hardware is available. From an ecological perspective, these institutional conditions operate at the exosystem level, shaping teachers' instructional possibilities indirectly by configuring the resources and professional conditions.

2.4. Sociocultural and Policy Level

National or regional educational policies and societal cultural attitudes provide a value framework. Differences in the understanding of "what constitutes effective teaching" across different cultural contexts influence teachers' expectations. These macro-level influences function as part of the macrosystem, providing normative orientations that shape institutional arrangements.

2.5. Student Level

Student factors such as digital literacy, engagement, and feedback enter the recursive feedback loop of ICT-integrated teaching. Empirical studies suggest that interactional and affective characteristics of learners can significantly constrain teachers' sense of agency. Disparities in access to devices and home internet stability directly condition participation opportunities, thereby influencing task design and interaction quality. Student-related factors operate as cross-level moderators that influence the sustainability of teachers' ICT integration.

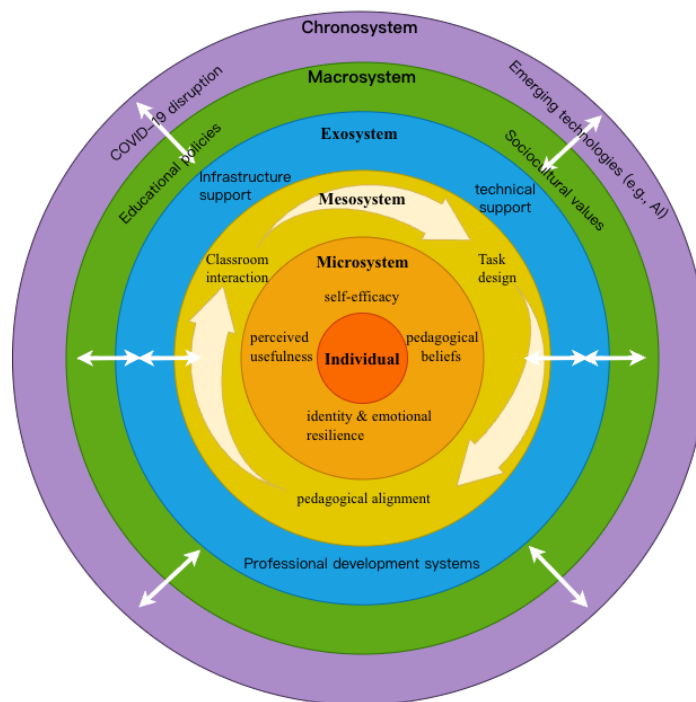
3. Toward a Multi-Level Ecological Framework

Factors previously scattered can be integrated into a multi-level structure: Microsystem: Teachers' digital competence, self-efficacy, and beliefs. Mesosystem: Instructional interaction and professional collaboration. Exosystem: School systems, resource allocation, and management structures. Macrosystem: Educational policies and cultural values. Chronosystem: Impact of the pandemic and emerging technologies like AI.

At the microsystem-mesosystem interface, teachers' competence is translated into practice through task design. The exosystem shapes these possibilities indirectly by configuring teachers' action space. Across systems, student responses operate as critical feedback mechanisms. The chronosystem captures how temporal disruptions reconfigure these pathways, reshaping constraints and affordances across levels.

Figure 1

An Ecological Framework for Understanding L2 Teachers' ICT Integration



4. Method

This study adopts a theory-driven integrative review. Relevant literature was retrieved from the Web of Science (WOS) database, covering peer-reviewed articles from 2015 to 2025. A thematic synthesis approach was employed to map findings onto the ecological systems theory.

5. Conclusion

Findings indicate that L2 teachers' ICT integration unfolds within a multi-level ecosystem. Individual resources require alignment with classroom interaction patterns in the mesosystem to be sustained. Exosystem-level supports condition mesosystem environments, while student feedback feeds back into teachers' confidence. This study advances research from factor-based classification toward a mechanism-oriented understanding of L2 teachers' ICT integration.

References

- Alenezi, S. M. (2022). *Tertiary level English Language Teachers' Use of, and Attitudes to Alternative and Online Assessments during the Covid-19 Outbreak*. *International Journal of Education and Information Technologies*, 16, 39–49. <https://doi.org/10.46300/9109.2022.16.4>
- Anwar, K., Ubaidillah, M. F., & Sulistiyo, U. (2020). *Exploring EFL teachers' classroom management: The case of Indonesian remote secondary schools*. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3862391>
- Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by Nature and Design*. Cambridge, Mass.: Harvard University Press.
- Canals, L., & Al-Rawashdeh, A. (2018). *Teacher training and teachers' attitudes towards educational technology in the deployment of online English language courses in Jordan*. *Computer Assisted Language Learning*, 32(7), 639–664. <https://doi.org/10.1080/09588221.2018.1531033>
- Çapuk, S., & Kara, A. (2015). *A Discussion of ICT Integration within Developed and Developing World Context from Critical Perspectives*. *Procedia - Social and Behavioral Sciences*, 191, 56–62. <https://doi.org/10.1016/j.sbspro.2015.04.411>
- Davis, F. D. (1989). *Perceived usefulness, perceived ease of use, and user acceptance of information technology*. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>

- Ding, L., & Hong, Z. (2023). *On the Relationship Between Pre-service Teachers' Sense of Self-efficacy and Emotions in the Integration of Technology in Their Teacher Developmental Programs*. *The Asia-Pacific Education Researcher*, 33(4), 869–878. <https://doi.org/10.1007/s40299-023-00758-6>
- Guo, K., Zhong, Y., Qin, J., & Chu, S. K. W. (2024). *Investigating EFL teachers' lesson planning for chatbot-assisted learning of argumentative writing: a TPACK perspective*. *Interactive Learning Environments*, 1–22. <https://doi.org/10.1080/10494820.2024.2372646>
- Hsu, L. (2016). *Examining EFL teachers' technological pedagogical content knowledge and the adoption of mobile-assisted language learning: a partial least square approach*. *Computer Assisted Language Learning*, 29(8), 1287–1297. <https://doi.org/10.1080/09588221.2016.1278024>
- Huang, F., Teo, T., & Zhou, M. (2017). *Factors Affecting Chinese English as a Foreign language Teachers' technology acceptance: A Qualitative study*. *Journal of Educational Computing Research*, 57(1), 83–105. <https://doi.org/10.1177/0735633117746168>
- Jang, J. (2024). *Exploring emotional vulnerability and sense of agency in the digital realm: A Three-Year case study of an EFL teacher in South Korea*. *Education Sciences*, 14(9), 950. <https://doi.org/10.3390/educsci14090950>
- Jeon, J., Lee, S., & Choe, H. (2022). *Teacher agency in perceiving affordances and constraints of videoconferencing technology: Teaching primary school students online*. *System*, 108, 102829. <https://doi.org/10.1016/j.system.2022.102829>
- Junaidi, Y., Hashim, H., & Ismail, H. H. (2022). *ESL TEACHERS' PERCEPTION AND ATTITUDES TOWARDS THE ADOPTION OF EMERGENCY REMOTE TEACHING IN TIME OF CRISIS*. *Journal of Nusantara Studies (JONUS)*, 7(2), 221–244. <https://doi.org/10.24200/jonus.vol7iss2pp221-244>
- Li, S., Yamaguchi, S., & Takada, J. (2018). *Understanding factors affecting primary school teachers' use of ICT for Student-Centered Education in Mongolia*. *International Journal of Education and Development Using Information and Communication Technology*, 14(1), 103–117.
- Liu, H., Lin, C., & Zhang, D. (2017). *Pedagogical beliefs and attitudes toward information and communication technology: a survey of teachers of English as a foreign language in China*. *Computer Assisted Language Learning*, 30(8), 745–765. <https://doi.org/10.1080/09588221.2017.1347572>
- Lo, N. P. (2023). *Digital learning and the ESL online classroom in higher education: teachers' perspectives*. *Asian-Pacific Journal of Second and Foreign Language Education*, 8(1). <https://doi.org/10.1186/s40862-023-00198-1>
- Mahmood, A. M., & Mohammadzadeh, B. (2022). *The effectiveness of ICT-supported flipped learning in an EFL context: A case of northern Iraq*. *Frontiers in Psychology*, 13, 943956. <https://doi.org/10.3389/fpsyg.2022.943956>
- Marzuki, Widiati, U., Rusdin, D., Darwin, & Indrawati, I. (2023). *The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective*. *Cogent Education*, 10(2). <https://doi.org/10.1080/2331186x.2023.2236469>
- Neal, J. W., & Neal, Z. P. (2013). *Nested or networked? Future directions for Ecological Systems Theory*. *Social Development*, 22(4), 722–737. <https://doi.org/10.1111/sode.12018>
- Tabassum, Z. A., & Saad, M. R. B. M. (2024). *Community of Inquiry as an Instructional framework*. *Arab World English Journal*, 10, 43–58. <https://doi.org/10.24200/jonus.vol7iss2pp221-244>
- Waddington, J., & Charikova, D. B. (2021). *Children's and teachers' views on digital games in the EFL classroom*. *ELT Journal*, 76(1), 44–57. <https://doi.org/10.1093/elt/ccab076>
- Wu, R., Yang, W., Rifienbark, G., & Wu, Q. (2022). *School and Teacher Information, Communication and Technology (ICT) readiness across 57 countries*. *Education and Information Technologies*, 28(2), 1273–1297.

Improving Rural School Teaching Quality through Blended Synchronous

Classroom: A Value-Added Assessment Study

Jinyu Yang^{1*}, Yue Feng¹

School of Information Technology in Education, South China Normal University, Guangzhou

*1758370764@qq.com

Abstract: *Students' academic performance is an important means to evaluate the quality of teaching and learning. Value-added assessment of data of their performance can reveal the increment of students' development. As so, this study used a simple growth percentile value-added assessment model to analyze the effects of regular blended synchronous English classroom to third-grade in 29 rural elementary schools in G Province, and used a standardized score model to verify its validity. The results show that: (1) Nearly 80% of the schools fall into the categories of high and normal growth, indicating that blended synchronous classrooms have a positive effect on promoting students' academic performance. (2) Differences in students' learning progress across schools are closely associated with variations in the motivation and participation of schools, teachers, and students. The findings of this study for the exploration of value-added assessment and the effectiveness of blended synchronous classroom teaching provide important insights.*

Keywords: rural blended synchronous classroom, Simple Growth Percentile Model, Standardized Score Model, value-added assessment, teaching effect

1. Introduction

In developing and emerging economies, the educational gap between urban and rural areas tends to be more pronounced (Amini & Nivorozhkin, 2015). As Philip Coombs observed, "In developing countries' rural areas, education quality is low, with underqualified teachers and many poor children who need good teachers most but get them least." (Coombs, 1985). In the context of China, the largest developing country, there exists a educational disparity between urban and rural areas (Yang et al., 2019).

With the empowerment of computer and network technology, blended synchronous classroom (BSC) has become an innovative practice paradigm of ICT, offering a viable solution to the issue of educational inequality. For instance, the implementation of cloud-based educational institutions facilitates the provision of high-quality teaching resources to rural areas (Zhang et al., 2018), enabling students to engage remotely with proficient educators and participate in real-time interactions (Carmel et al., 2010).

In the field of educational assessment, value-added assessment (VAA) refers to a longitudinal research design (Tekwe et al., 2004). Taylor (1985) was the first to explicitly propose the VAA method, using the term "value-added" to describe the learning progress and development demonstrated by students during the learning process. It refers to analyzing students' learning process and outcomes over the entire duration of their study or a specific phase, to capture the "increment" of their academic progress or development. This study employs VAA to examine the effectiveness of BSC in teaching and uses the academic test data of third-grade students in regular blended synchronous English classrooms in 29 primary schools in a county in G province, China. We utilize both the simple growth percentile model and the standardized score model, employing a dual-model mutual validation approach, providing a basis for verifying the effectiveness of blended synchronous instruction.

2. Literature Review

2.1. Studies related to value-added assessment

Traditional assessment models often emphasize final outcomes, overlooking the developmental dynamics that unfold during the learning process. In contrast, VAA places greater emphasis on the growth and development of students,

teachers, and the continuous optimization of the education system as core considerations. This approach facilitates a transition from a static to a dynamic evaluative framework. Numerous scholars have conducted research based on the value-added assessment approach. Timmermans et al. (2014) conducted a synthesis of seven empirical studies to investigate the variations in value-added indicators through the inclusion of school composition variables. Their findings revealed notable disparities in the strength of the relationship between the Type A model (utilizing student-level covariates) and the Type B school effect model (incorporating both student- and school-level covariates) across different geographical regions. Şen et al. (2020) found that, based on TEOG data from 539 secondary schools in Şanlıurfa, exam averages are insufficient, whereas value-added models provide a more objective measure of school effectiveness. Collectively, these studies enhance the understanding of the theoretical and practical implications of VAA from multiple perspectives and provide research support for the selection of VAA methods in this study to evaluate the effectiveness of teaching within BSC.

2.2. Selection of value-added assessment models

Student Growth Percentile (SGP) developed by Betebenner (2009) and based on quantile regression has been adopted by over 20 U.S. states. Scholars have examined SGP from various perspectives, addressing its foundational principles, calculation methodologies (Betebenner, 2011), error-related issues (McCaffrey et al., 2015), and frameworks for improvement (Monroe et al., 2014). Additionally, preliminary validation of the model's effectiveness for VAA was conducted utilizing actual test data from students (Briggs & Betebenner, 2009; Kurtz, 2018; Taherbhai et al., 2014). Although SGP is based on quantile regression estimation, its computational complexity makes practical implementation challenging. In response to identified limitations in its application, researchers developed a simplified growth percentile, replacing the complex quantile regression method with percentile rank (PR) calculations (Shen, 2022). Therefore, this study employs the simplified growth percentile to conduct a VAA of the teaching effectiveness within rural BSC.

In addition to the use of SGP, standardized score models serve as a tool in VAA due to their ability to transform raw data into a common scale through standardization. Prior et al. (2021) applied Z-score standardization to student attainment, absence, and exclusion data, suggesting that academic-only value-added models inadequately capture overall school performance. The SGP evaluates growth based on the relative position changes of students' scores, while Z-score can transform raw scores into standardized values with a unified dimension, which not only eliminates score differences across subjects or tests but also accurately depicts the relative position of students' scores in the population. This aligns methodologically and mutually validates data with the SGP's logic of assessing growth through relative score position changes.

2.4. Research question

This present study draws on empirical data from Grade 3 English BSC in 29 rural primary schools in Province G, China. A dual-model validation approach is employed to assess the instructional effectiveness of BSC. The research focuses on the following key questions:

RQ1. What is the overall progress in English learning among students across schools?

RQ2. What differences exist in students' English learning progress across schools?

3. Methods

3.1. Participants

The data were collected from 29 primary schools in a county in G Province, where Grade 3 students routinely receive blended synchronous English instruction. These schools are part of the county's total of 84 primary schools. During the 2022 - 2023 academic year, all participating students completed two rounds of standardized English academic achievement assessments, administered citywide by the municipal education bureau. These two sets of test scores were used for VAA. To ensure valid data analysis and interpretation, students with missing scores from either assessment were

excluded. Consequently, 5413 student records from 84 schools were included in the final analysis, including 472 students from the 29 schools implementing the blended synchronous English classroom.

3.2. Data processing and analysis

This study employs the simplified growth percentile to analyze sample English scores from two rounds of academic quality assessments, aiming to evaluate students' overall academic progress and institutional developmental characteristics across different schools. Initially, the first test results from 5,413 students across 84 schools were compiled to calculate PR for each student. Subsequently, these students were categorized into "Academic Partner" groups. Using the second test data for all students, PR for each student within each "academic partner" group was calculated to derive the SGP values. Individual student progress in English learning was determined with reference to the "Academic Partner" reference group. These SGP values were then employed to establish the teaching development level for each school. The mSGP within each school was utilized to represent the changes in teaching growth at these levels. Ultimately, the growth levels of students and schools participating in blended synchronous instruction were evaluated based on the established classification criteria. The SGP values were categorized into three growth levels: low growth level (0-34), normal growth level (35-65), and high growth level (66-100). The distribution of students across these categories contributed to each school's value-added score, with higher scores reflecting stronger instructional quality and learning outcomes (see Table 1).

Table 1. Partial VAA Results of Rural BSC Teaching and Learning Based on the simplified growth percentile

School	Number of Students	School Value-added Scores (mSGP)	Student Growth Level (SGP)					
			Low Growth Level (0~34)		Normal Growth Level (35~65)		High Growth Level (66~100)	
			No.	percentage	No.	percentage	No.	percentage
S1	114	65	25	22%	34	30%	55	48%
S2	11	20	7	64%	3	27%	1	9%
S3	30	48	10	33%	10	33%	10	33%
S4	5	81	1	20%	0	0%	4	80%
S5	22	64	2	9%	10	45%	10	45%
S6	15	34	8	53%	4	27%	3	20%
S7	5	80	1	20%	1	20%	3	60%
S8	11	43	5	45%	5	45%	1	9%
S9	25	36	11	44%	8	32%	6	24%
S10	27	61	5	19%	12	44%	10	37%
S11	14	59	4	29%	6	43%	4	29%
S12	17	42	4	24%	8	47%	5	29%
S13	21	44	9	43%	3	14%	9	43%
S14	13	71	3	23%	2	15%	8	62%
S15	7	83	0	0%	2	29%	5	71%
S16	13	41	6	46%	6	46%	1	8%
S17	16	90	1	6%	4	25%	11	69%
S18	3	43	0	0%	2	67%	1	33%
S19	9	72	3	33%	0	0%	6	67%
S20	7	49	1	14%	4	57%	2	29%

S21	6	72	0	0%	3	50%	3	50%
-----	---	----	---	----	---	-----	---	-----

Note. The sum of the shares of the three growth levels does not add up to 100 percent because decimals are not retained.

The Standard Score model was employed to analyze the two waves of academic assessment data to cross-validate the findings derived from the simplified growth percentile. Initially, using data from two assessments, mean scores, standard deviations, and school-level averages were calculated for 5,413 students across 84 primary schools. Subsequently, using the standardized score model formula, standardized scores for each school in both assessments were calculated, along with corresponding Z-score gains. Ultimately, a positive Z-score gain indicates improved performance, allowing for the evaluation of instructional progress in schools implementing blended synchronous instruction. Table 2 presents the VAA results for blended synchronous instruction, derived using the standardized score model. A positive Z-score gain indicates improved instructional quality, with higher values reflecting stronger overall teaching performance.

Table 2. *Partial VAA Results of Rural BSC Teaching and Learning Based on the standardized score model*

3.3 Validation of
School
scores from the
growth
score gains from
score model
generally
evaluation trends
models (see

School	The First Test		The Second Test		Z-score Value Added
	Mean Score	Standard Deviation	Mean Score	Standard Deviation	
S1	52.4	-0.52	74.4	-0.19	0.33
S2	39.9	-1.06	52	-1.45	-0.39
S3	39.5	-1.07	60.4	-0.98	0.09
S4	63.8	-0.03	85	0.43	0.46
S5	49.7	-0.63	74.8	-0.16	0.47
S6	40.9	-1.01	55.3	-1.27	-0.26
S7	88.8	1.05	92.7	0.85	-0.20
S8	40.7	-1.02	63.1	-0.82	0.20
S9	70.2	0.25	77.4	-0.02	-0.27
S10	63.3	-0.05	78.8	0.06	0.11
S11	58.4	-0.26	78.4	0.04	0.30
S12	32.1	-1.39	54.9	-1.29	0.10
S13	45.9	-0.80	64.7	-0.73	0.07
S14	35.5	-1.25	68.5	-0.52	0.73
S15	43.9	-0.88	78.7	0.06	0.94
S16	75.7	0.49	81.2	0.20	-0.29
S17	36.5	-1.20	75.1	-0.15	1.05
S18	51.3	-0.56	70.5	-0.41	0.15
S19	35.6	-1.24	64.5	-0.75	0.49
S20	66.9	0.11	77.9	0.01	-0.10
S21	44.5	-0.86	77.7	0.00	0.86

model results
value-added
simplified
percentile and Z-
the standardized
revealed
consistent
across both
Figure 1).

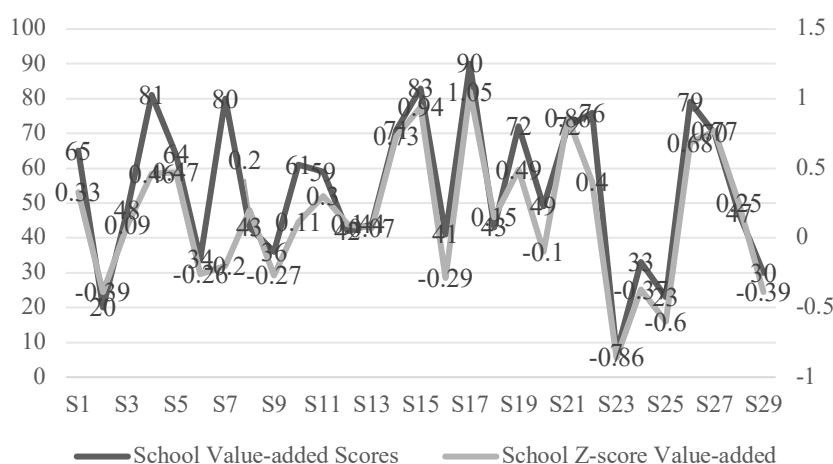


Figure 1 Comparison of school value added based on the two models

4. Result

4.1. The level of development of the school and its students

The student VAA data from 29 schools provides a comprehensive overview of both the overall development of the schools and the progress of individual students. Generally, the proportion of schools exhibiting normal and high growth levels surpasses that of schools categorized as having low growth levels (refer to Figure 2). Collectively, nearly 80% of the schools fall into the categories of high and normal growth, suggesting that the quality of English instruction in these schools is comparatively high. Specifically, 45% of the schools are classified at the normal growth level, reflecting a steady advancement in student learning outcomes, while 34% have achieved high growth levels, signifying robust developmental potential. Furthermore, the application of the standard score method also reveals that approximately 70% of the schools demonstrate a discernible trend of improvement.

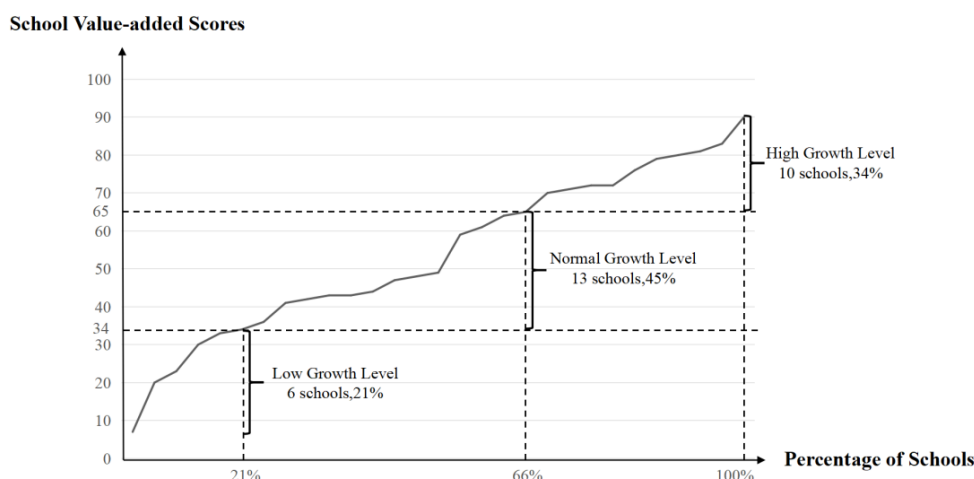


Figure 2 Growth levels of 29 schools with rural blended synchronous instruction

4.2. The level of student development in each school

Student growth varies within schools, and understanding individual developmental levels is essential for delivering targeted support and guidance. As shown in Table 3, students in Schools S4 and S19 exhibit both high and low growth levels, indicating considerable variation in individual academic progress. For example, in School S19, Student Pan achieved an SGP score of 98, whereas Student Peng scored only 1. This stark contrast underscores the uneven academic development among students within the same school. Reviews of blended synchronous instructional video recordings,

revealed that students with lower SGP scores were generally less likely to respond to questions or initiate interactions during lessons. This lack of engagement may partially explain their limited academic progress.

Table 3. *Student growth levels in Schools S4 and S19*

5. Discussion and conclusion

School	Student	Percentile rank (SGP score)	Student Growth Level (SGP)		
			Low Growth Level (0~34)	Normal Growth Level (35~65)	High Growth Level (66~10)
S4	Jiang XX	96			√
	Yang XX	93			√
	Huang XX	81			√
	Luo XX	81			√
	Luo XX	15	√		
S19	Pan XX	98			√
	Peng XX	96			√
	Huang XX	88			√
	Pan XX	83			√
	Li XX	72			√
	Li XX	70			√
	Li XX	29	√		
	Li XX	23	√		
	Peng XX	1	√		

The purpose of this study was to analyze the effectiveness of standard blended synchronous English instruction in third-grade classrooms across 29 primary schools in a county of G Province, utilizing the value-added assessment model. The validity of the findings was further verified using the Standard Score model. In this section, key findings are summarised and discussed and directions for future research are proposed.

The results indicate that BSC instruction contributes positively to students' English learning progress across schools. Approximately 70% of students achieved the expected level of growth, and nearly 80% of schools demonstrated normal or high levels of performance improvement. This finding is consistent with the research conclusion of Huang et al. (2019): BSC instruction is generally effective in promoting student development and teaching quality in rural schools. However, variations in school-level development were observed. Video analysis further reveals that the participation and engagement of on-site teachers play a crucial role in shaping instructional effectiveness, highlighting the importance of collaborative teaching practices in BSC contexts. This finding was in concurrence with Yan et al. (2024)'s research confirming that collaborative teaching between online teachers and on-site teachers is conducive to improving student engagement, thereby influencing the teaching effect of BSC.

Despite the overall positive outcomes, significant differences in students' English learning progress exist across schools. In particular, a relatively high proportion of low-growth students in some schools constrains overall improvement in instructional effectiveness. This disparity suggests that student progress is unevenly distributed and may reflect differences in instructional practices and support mechanisms. Consequently, instructors should remain cognizant of the need to intentionally encourage and structure opportunities for remote students to engage more actively in classroom activities (Weitze et al., 2013).

Several limitations should be noted in this study. First, this study evaluates teaching effectiveness by measuring the extent of students' academic progress over a specific period. This process typically requires at least two rounds of

academic data as the basis. In remote or under-resourced schools, data quality and completeness are often difficult to ensure, which may compromise the accuracy of the evaluation results. Future research could optimize the time dimension of evaluation by setting up multi-cycle assessment nodes for short-term, medium-term, and long-term periods, so as to dynamically track the process of academic value-added. Second, the sample size of the participants is insufficient to generalize the conclusion to other blended synchronous learning environments. More students and schools of different areas should be selected for a better generalizability of the conclusion.

References

- Amini, C., & Nivorozhkin, E. (2015). The urban–rural divide in educational outcomes: Evidence from Russia. *International Journal of Educational Development*, 44, 118-133.
- Betebenner, D. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, 28(4), 42-51.
- Betebenner, D. W. (2011). A Technical Overview of the Student Growth Percentile Methodology: Student Growth Percentiles and Percentile Growth Projections/Trajectories. *National Center for the Improvement of Educational Assessment*.
- Briggs, D., & Betebenner, D. (2009). Is growth in student achievement scale dependent. *Unpublished manuscript*.
- Carmel Parker White, Rich Ramirez, Jessica G. Smith and Lisa Plonowski.(2010). Simultaneous delivery of a face-to-face course to on-campus and remote off-campus students.
- Coombs, P. H. (1985).The world crisis in education. The view from the eighties (pp. xiv+-353pp).
- Hillier, M. (2020). Bridging the digital divide with off-line e-learning. In *Expanding Horizons in Open and Distance Learning* (pp. 110-121). Routledge.
- Huang Y, Zhao C, Shu F, et al. Investigating and analyzing teaching effect of blended synchronous classroom[C]//2017 international conference of educational innovation through technology (EITT). IEEE, 2017: 134-135.
- Kurtz, M. D. (2018). Value-added and student growth percentile models: what drives differences in estimated classroom effects?. *Statistics and Public Policy*,5(1), 1-8.
- McCaffrey, D. F., Castellano, K. E., & Lockwood, J. R. (2015). The impact of measurement error on the accuracy of individual and aggregate SGP. *Educational Measurement: Issues and Practice*, 34(1), 15-21.
- Monroe, S., Cai, L., & Choi, K. (2014). Student Growth Percentiles Based on MIRT: Implications of Calibrated Projection. CRESST Report 842. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.
- Prior, L., Goldstein, H., & Leckie, G. (2021). School value-added models for multivariate academic and non-academic outcomes: exploring implications for performance monitoring and accountability. *School Effectiveness and School Improvement*, 32(3), 486-507.
- Şen, S., Yıldırım, İ., & Karacabey, M. F. (2020). The implementation of the value-added assessment to determine school effectiveness. *Education and Science*, 45(202), 79-91.
- Shen, Y. (2022). Value-Added Evaluation and Application Based on the Simple Model of Student Growth Percentile——Taking the Teaching Quality Inspection in Yantian District of Shenzhen for Example. *Educational Measurement and Evaluation*, (6), 91 - 102.
- Taherbhai, H., Seo, D., & O'Malley, K. (2014). Formative information using student growth percentiles for the quantification of English language learners' progress in language acquisition. *Applied Measurement in Education*, 27(3), 196-213.
- Tekwe, C. D., Carter, R. L., Ma, C. X., Algina, J., Lucas, M. E., Roth, J., ... & Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, 29(1), 11-36.

- Timmermans, A. C., & Thomas, S. M. (2015). The impact of student composition on schools' value-added performance: A comparison of seven empirical studies. *School Effectiveness and School Improvement*, 26(3), 487-498.
- Taylor, T. (1985). A value-added student assessment model: Northeast Missouri State University. *Economics of Education Review*, 4(4), 341-350.
- Weitze, C. L., Ørngreen, R., & Levinsen, K. (2013, October). The global classroom video conferencing model and first evaluations. In Proceedings of the 12th European Conference on E-Learning: SKEMA Business School, Sophia Antipolis France, 30-31 October 2013 (pp. 503-510). Academic Conferences and Publishing International.
- Yang, J., Yu, H., & Chen, N. S. (2019). Using blended synchronous classroom approach to promote learning performance in rural area. *Computers & Education*, 141, 103619
- Zhang, J., Yang, J., Fan, L., & Huang, R. (2015). Innovation scale-up of ICT in education in China. *Scaling educational innovations*, 179-195.

The Effectiveness of AR Technology in Chemistry Education: A Meta-Analysis

Based on 14 Chinese Studies

Hongfa Chen¹, Yaoyao Guo¹, Yanhua Fan^{1*}, Peiyao Tian²

¹ College of Chemistry and Molecular Sciences, Henan University, Kaifeng, China

² Faculty of Education, The University of Hong Kong, Hong Kong SAR, China

* fanyanhua9080@163.com

Abstract: Augmented reality (AR) technology, with its multidimensional visualization and interactive features, offers innovative solutions for presenting abstract concepts in chemistry education. This study conducted a meta-analysis to systematically evaluate the effectiveness of AR technology in chemistry education in China, incorporating 14 empirical studies. The results indicate that AR technology has a significant, moderate positive effect on chemistry learning outcomes ($g = 0.502$, 95% CI [0.348, 0.656]). Further analysis revealed that educational level was the only variable with a significant moderating effect, while the remaining variables did not show significant heterogeneity.

Keywords: augmented reality (AR), chemistry education, meta-analysis

1. Introduction

Augmented reality (AR) has opened up new avenues for chemistry education by visualizing abstract concepts and simulating experiments. However, its overall effectiveness in classroom instruction in China remains unclear. This meta-analysis systematically reviews empirical studies conducted in China to assess the impact of AR on chemistry learning outcomes and identify key moderating variables, such as educational level and instructional design, thereby providing evidence-based guidance for the effective integration of AR in the Chinese educational context.

2. Method

Based on a systematic search strategy, this study comprehensively collected relevant literature from multiple Chinese and English databases. Following a rigorous screening process and predefined inclusion and exclusion criteria, 14 high-quality studies meeting the criteria were ultimately included, comprising a total of 34 independent effects and 1,199 participants. The included studies were systematically coded across multiple dimensions, including educational level, intervention duration, and learning content, to analyze their characteristics. During the data analysis phase, a random-effects model and Hedges' g effect size were employed for the meta-analysis. Publication bias was comprehensively assessed using funnel plots, Egger's regression test, and Rosenthal's safety margin. The results indicated that no significant publication bias was detected in any of the tests, confirming the robustness and reliability of the meta-analysis findings.

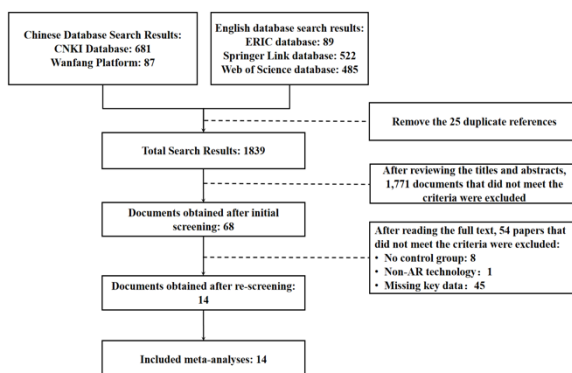


Figure 1 Literature screening process

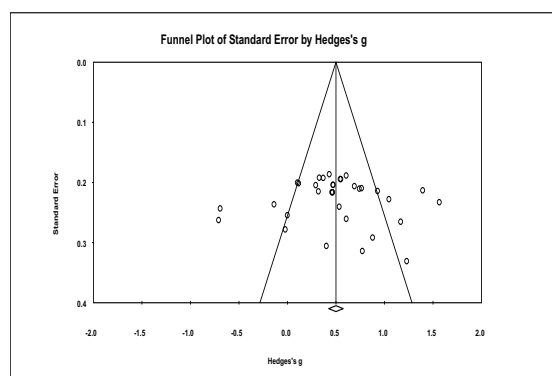


Figure 2 Forest plot for meta-analysis

3. Results and Discussion

The meta-analysis results indicate that the use of augmented reality (AR) in chemistry instruction is significantly associated with higher student learning outcomes ($g = 0.502$, $p < 0.05$). This finding aligns with existing research suggesting that AR plays a positive role in chemistry education (Song et al., 2025), further confirming that AR technology has a certain degree of promotional effect in chemistry instruction. Specifically, augmented reality helps learners intuitively understand abstract chemical concepts, thereby enhancing knowledge acquisition and the learning experience (Liu et al., 2023).

Among the moderating variables, only educational stage exhibited a significant moderating effect. The effect size for junior high school students ($g = 0.633$) was significantly higher than that for high school students ($g = 0.284$), indicating that AR has a more pronounced promotional effect on younger learners. According to cognitive development theory, junior high school students are in the formal operational stage and are capable of abstract logical reasoning. However, they face certain challenges in understanding abstract chemical concepts, and their thinking still relies on concrete experiences and intuitive support. The visual representations provided by AR play a crucial role in the learning of students at this stage. In contrast, high school students have already developed preliminary formal operational abilities and are capable of abstract thinking and hypothetical-deductive reasoning; therefore, their reliance on purely intuitive support may be relatively lower.

Although there were differences in other variables—such as source of literature, duration of intervention, class size, learning content, learning outcomes, and teaching methods—none of these differences reached statistical significance. Overall, these results suggest that the effectiveness of AR is influenced by multiple contextual factors, but its positive impact remains generally stable and reliable.

4. Conclusion

This meta-analysis, based on 14 empirical studies conducted in China, systematically evaluated the effectiveness of augmented reality (AR) technology in chemistry education. The results indicate that AR technology has a moderate positive effect on chemistry learning outcomes. Among all moderating variables, only educational level exhibited a significant moderating effect, with the effect being significantly stronger among junior high school students than among senior high school students. In contrast, variables such as source of literature, duration of intervention, class size, learning content, type of outcomes, and teaching methods did not exhibit significant heterogeneity. Overall, AR technology demonstrates clear potential for application in chemistry education, with a particularly strong positive effect on junior high school students.

Acknowledgements

This study was supported by Henan Provincial Philosophy and Social Sciences Planning Project (2025BJY033) and Henan Province Philosophy and Social Science Research Education Strong Province Project (2025JYQS0015).

References

- Liu, Q., Ma, J., Yu, S., Wang, Q., & Xu, S. (2023). Effects of an augmented reality-based chemistry experiential application on student knowledge gains, learning motivation, and technology perception. *Journal of Science Education and Technology*, 32(2), 153–167.
- Song, Q., Li, M., Zhang, Z., & Zhou, Q. (2025). Learning effectiveness of virtual technologies in chemistry using a three-level meta-analysis. *Education and Information Technologies*, 30(11), 14737–14759.

Effect of an AR-based Gamified Environmental Education Platform on the Environmental Consciousness of Grade 5 Students: a Case Study

Yan Ziqi¹, Alex Wing Cheung Tse²

Faculty of Education, The University of Hong Kong, Hong Kong

¹u3631816@connect.hku.hk, ²awctse@hku.hk

Abstract: *Traditional environmental education often emphasizes cognitive knowledge acquisition while providing limited support for affective engagement, behavioral intention, and pro-environmental action. This case study explored the effects of an AR-based gamified platform on elementary school students' environmental consciousness. Following a six-week classroom intervention, semi-structured interviews with nine students were thematically analyzed across four dimensions: affective, cognitive, conative, and active. The findings indicate that the platform was particularly effective in strengthening the conative and active dimensions, as students demonstrated greater willingness to engage in environmental actions and observable pro-environmental behaviors. Improvements in the affective and cognitive dimensions were more modest, mainly reflected in short-term emotional engagement and initial knowledge acquisition.*

Keywords: Environmental consciousness, AR-based Learning, Gamified learning

1. Introduction

Cultivating environmental consciousness (EC) is widely recognized as a prerequisite for environmental protection and sustainable development (Su et al., 2021). However, environmental education often faces a persistent “knowledge–action” gap, where greater environmental knowledge does not necessarily lead to behavioral change (Rueda et al., 2023). Emerging technologies offer potential solutions: Augmented Reality (AR) can create immersive learning environments, while gamification can enhance learners' motivation and engagement (Hung et al., 2017; Sailer et al., 2020). However, learner-centered research on AR-based gamified environmental education in elementary contexts remains limited. Guided by Jiménez and Lafuente's (2010) framework, this study explores how an AR-based gamified platform supports Grade 5 students' EC across affective, cognitive, conative, and active dimensions.

2. Methodology

This study reports the second phase of a mixed-methods investigation examining the effects of an AR-based gamified environmental education platform on Grade 5 students' EC. In the first phase, 72 students from two classes at a public elementary school in Shenzhen, China, participated in a six-week intervention, with one class using the AR platform and the other receiving traditional instruction. Nine students from the experimental group was selected in the qualitative phase, representing students with high, medium, and low post-test scores, for semi-structured interviews. Conducted after the intervention, the interviews explored students' learning experiences and were analyzed using thematic analysis across the cognitive, affective, conative, and active dimensions of EC.

3. Results and findings

The **affective dimension** of EC refers to learners' level of concern for environmental issues and their attitudes toward environmental protection (Jiménez & Lafuente, 2010). Thematic analysis indicates that the AR-based gamified platform enhanced students' affective engagement, particularly by stimulating short-term interest and curiosity. For example, Student 5 noted, “When I used to watch videos, I would guess if it was rehearsed or not, but the 3D modeling made it feel real and more appealing to me.” However, evidence of sustained emotional change was limited, with only a few students demonstrating deeper emotional responses to environmental issues. For instance, Student 7 reflected, “Seeing people cut down the forest in the 3D model made me feel sad for the animals.”

The **cognitive dimension** of EC refers to learners' level and sources of environmental knowledge, as well as the importance they assign to environmental news and their understanding of related issues (Jiménez & Lafuente, 2010). The findings suggest that the AR-based gamified environmental education platform supported early-stage cognitive development of EC, while evidence of deeper cognitive restructuring remained limited. For example, Student 1 explained, "The AR little turtle model introduced us to many marine creatures," indicating that cognitive gains are primarily concerned with the acquisition of knowledge rather than complex conceptual changes.

The **conative dimension** of EC refers to learners' willingness to engage in pro-environmental actions and their sense of being part of the solution (Jiménez & Lafuente, 2010). Thematic analysis indicates that the AR-based gamified platform provided immersive, socially contextualized learning experiences with moral reflection. These experiences enabled learners to recognize they could be part of the solution, strengthening their readiness to act and enhancing the conative dimension of EC. For example, Student 2 remarked, "When my classmates and I played games and shared ways to protect the environment, it made me want to do more."

The **active dimension** of EC refers to learners' environmentally related behaviors in both home and educational settings (Jiménez & Lafuente, 2010). The results show that the AR-based gamified environmental education platform effectively motivated learners to adopt pro-environmental behaviors related to the learning content and encouraged a few students to extend these behaviors into public social spaces, thereby positively influencing the active dimension of EC. As Student 8 reported, "I applied for and became a docent at the Waste Sorting and Environmental Protection Experience Pavilion."

4. Conclusion

This study found that an AR-based gamified environmental education platform enhanced Grade 5 students' EC. Positive effects were observed across all four dimensions, with particularly strong impacts on the conative and active dimensions. By offering immersive environments, interactive gameplay, immediate feedback, and peer collaboration, the platform supported students' intentions to act pro-environmentally and their behaviors in learning and daily contexts. Overall, AR-based gamified learning can modestly foster students' cognitive and affective engagement, effectively promoting pro-environmental intentions and behaviors, thereby strengthening their EC.

References

- Hung, Y. H., Chen, C. H., & Huang, S. W. (2017). Applying augmented reality to enhance learning: a study of different teaching materials. *Journal of Computer Assisted Learning*, 33(3), 252–266.
- Jiménez, M., & Lafuente, R. (2010). Definición y medición de la conciencia ambiental. *Revista Internacional De Sociología*, 68(3), 731–755.
- Rueda, R. Z., Beltrán Villamizar, Y. I., & Becerra Ardila, L. E. (2023). A retrospective approach to pro-environmental behavior from environmental education: an alternative from sustainable development. *Sustainability*, 15(6), 5291.
- Sailer, M., & Homner, L. (2020). The gamification of learning: a meta-analysis. *Educational Psychology Review*, 32(1), 77–112.
- Su, F., Song, N., Shang, H., Wang, J., & Xue, B. (2021). Effects of social capital, risk perception and awareness on environmental protection behavior. *Ecosystem Health and Sustainability*, 7(1), 1942996.

Full version URL: https://drive.google.com/file/d/187w00U70uzSVXUd-U_8J48NzlrV4Qmmq/view?usp=sharing

Exploring the Pattern and Influence Factors in the Correlation Between Learning Engagement and Academic Achievement in Online Learning

Yaoyuan Zhou¹, Yuhan Gai¹, Jiamin Ren¹, Qingyang Liu¹

¹ School of Educational Technology in Beijing Normal University

Abstract: *This study investigates the complex patterns linking learning engagement to academic achievement in online learning environments. By analyzing data from 217 university students using regression, K-means clustering, ANOVA, and Epistemic Network Analysis (ENA), the research reveals that simple positive correlations obscure distinct student profiles. Cluster analysis identified three groups (high engagement & achievement, low engagement & achievement, low engagement/high achievement), including a notable "low-engagement/high-achievement" profile. Results indicate that behavioral engagement serves as the primary differentiator between high and low academic achievement. Furthermore, ENA findings suggest that low-engagement/high-achievement students compensate for lower behavioral frequency with superior cognitive quality, characterized by integrated knowledge generation, whereas low-performing students rely on superficial referencing. These findings challenge linear engagement–achievement models, demonstrating that academic success depends less on the magnitude of engagement than on the structural quality of cognitive processing and behavioral execution.*

Keywords: Online learning, Learning engagement, Learning analysis, Discourse analysis

1. Introduction

Originating in the late 20th century, online learning was initially introduced into higher education to provide flexible learning modalities for students—thereby expanding the educational market—and to afford greater geographical flexibility for faculty and staff (McGaughey, 2022; Xie et al., 2020). Over the past two decades, the significance of online learning has increasingly surged, establishing it as a pivotal educational format characterized by its unique environment and conditions (National Center for Educational Statistics [NCES]). However, research indicates that the lack of effective teacher–student interaction within online environments impedes communication and the generation of personalized feedback, resulting in challenges related to low learning engagement and high attrition rates (Freidhoff, 2021; Trust, 2020). While online learning engagement is recognized as a critical factor influencing student achievement and instructional quality, with numerous studies demonstrating a positive correlation between engagement and learning outcomes (Lee, 2022), this relationship is moderated by various factors. Nevertheless, simple correlations are insufficient to accurately characterize student dynamics in online learning. Consequently, it is imperative to investigate the specific patterns underlying the relationship between learning engagement and academic achievement, as well as the mechanisms driving the formation of these patterns.

2. Literature Review

Learning engagement's definition can be dated back to Moore (1989), who conceptualized engagement through the lens of three distinct interaction types: learner–content, learner–instructor, and learner–learner interactions. In the specific context of online learning, Martin (2022) synthesized these perspectives to propose a refined definition: "online learning engagement is the productive cognitive, affective, and behavioral energy that a learner exerts interacting with others and learning materials and/or through learning activities and experiences in online learning environments."

Martin's framework delineates learning engagement into three dimensions: cognitive, affective, and behavioral. Cognitive engagement involves the psychological energy invested in learning activities, emphasizing mental exertion and interactions with materials or peers, for example, learning how to conduct online learning appropriately and how to engage

effectively in collaborative interaction and learning with materials. Affective engagement captures students' emotional reactions, reflecting both individual responses and technological influences within the online context. It mainly focuses on building relationships and fostering a sense of community and belonging in the process of online learning. Behavioral engagement represents the physical manifestation of these cognitive and affective states through tangible actions. This includes participation in both academic and non-academic activities, accompanied by the demonstration of positive behaviors.

The association between learning engagement and academic achievement has been repeatedly substantiated across numerous studies. Puri (2025) indicated that in the context of mobile-assisted language learning, mobile devices effectively reduce cognitive load and anxiety, thereby serving as a mediating effect that explains the positive correlation between learning engagement and learning achievement. Furthermore, researchers (Hatahet, 2025; Oliver, 2025) have confirmed the positive impact of multidimensional engagement on learning achievement from the perspectives of cognitive engagement, affective engagement, and behavioral engagement, respectively. Collectively, these studies establish a solid foundation for conceptualizing the relationship between learning engagement and academic performance.

Current research predominantly focuses on the associations between learning engagement and other constructs, such as self-achievement (Wang, 2022) and teacher–student interaction (Ong, 2023), in relation to academic outcomes. However, scholars have noted that such studies typically employ cross-sectional designs, necessitating caution when interpreting causal relationships (Miao, 2022). Only a minority of scholars have examined the internal dynamics of learning engagement itself, proposing that engagement is multidimensional and that its distinct types interact with one another (Hatahet, 2025). This suggests the existence of more complex, underexplored relationships; indeed, the complex dynamics exhibited by students extend beyond simple linear associations, and latent patterns between these variables remain obscured. Moreover, while students of different profiles exhibit variations across the three dimensions of engagement, few studies have deeply investigated the underlying causes driving these multidimensional disparities. Accordingly, this study proposes the following two research questions:

1. What is the relationship between students' online learning engagement and their academic achievement, and what characteristics does the student distribution exhibit?
2. What features characterize the epistemic networks of different student profiles, and what are the distinctions and commonalities among them?

3. Methods

3.1 Participant and Procedures

This study utilized course data from the three separate class sessions of the "Educational Technology" curriculum—comprising 14 classes sessions—offered during the 2024 spring for 15 weeks at a university in Beijing. The course was delivered online via a self-developed platform, where students were required to receive course notifications, engage in online learning, post classroom comments, and complete tasks assigned within each session. The platform systematically collected students' post-class assignment scores, course completion rates, and comment content, while also capturing all operational log data generated by student users during the learning process. Figure 1 shows the shortcut of platform applied in online learning.

A total of 309 students provided informed consent, confirming their voluntary and anonymous participation in the study. Following data cleaning and verification procedures, 217 valid samples were obtained. The final sample consisted of 149 females (approximately 68.6%) and 68 males (approximately 31.4%), representing a diverse range of majors including law, chemistry, mathematics, education, and literature. We collected 432 comments in the platform overall.



Fig.1 Shortcut of the self-developed online learning platform

3.2 Instrument and Data Analysis

To quantify students' learning engagement, this study adopted the conceptual framework proposed by Martin (2022) and measured engagement across three distinct dimensions. First, cognitive engagement was assessed using students' participation assessment scores, which were calculated as a composite of content completion (34%), chapter quizzes (50%), and attendance (16%). Second, regarding affective engagement was assessed by questions from the Online Self-regulated Learning Questionnaire (OSLQ) developed by Barnard et al. (2009); the overall Cronbach's alpha coefficient for the scale was 0.90, supporting the instrument's validity. Finally, behavioral engagement was measured by collecting interaction log data from the online platform across five dimensions: page views, key-point clicks, comments, AI utilization, and pause frequency. To measure the behaviour engagement, we calculated the frequency of interaction and converts values respectively into its percentile rank as the final score. Ultimately, the mean score of these three dimensions was computed to determine the students' overall level of learning engagement. At the end of the semester, students were required to complete an examination, and their exam scores were used as the measure of academic achievement.

The analytical procedure consisted of three steps. Initially, we performed a linear regression to assess the predictive validity of the overall learning engagement index on academic achievement. Following this, K-means clustering was utilized to uncover latent profiles characterizing different engagement-achievement patterns. In the final stage, a Mixed designed ANOVA were applied to decompose the specific impacts of the three engagement dimensions (cognitive, affective, and behavioral) on academic achievement, and to characterize the heterogeneity of engagement profiles across the identified clusters.

To characterize the features of different student profiles, all comment data from the online instructional platform were collected and encoded using the asynchronous online discussion coding framework established by Biasutti (2017). The students' comment texts were coded across seven dimensions: inferencing, producing, developing, evaluating, summarizing, organizing, and supporting. Two researchers participated in the coding process. After brief training to ensure conceptual alignment in the coding scheme, both researchers independently conducted the coding based on the agreed concepts, and the final Cohen's kappa reached 0.91, indicating a high level of agreement and strong inter-coder reliability. Subsequently, Epistemic Network Analysis (ENA) was conducted using the web-based ENA tool (<https://app.epistemicnetwork.org/>) to model the cognitive characteristics of the different student profiles and to visualize the divergences in cognitive networks between groups.

4. Result

4.1 The Correlation and Pattern Between Learning Engagement and Academic Achievement

The descriptive statistics of learning engagement and academic achievement for all students are shown in Table 1. It can be seen that the average score for learning engagement is lower than the average score for academic achievement, but

the data stability for both is similar. Among the three types of learning engagement, the average scores for cognitive engagement, affective engagement, and behavioral engagement decrease sequentially, and the data stability also gradually declines. This indicates that students generally have a high level of cognitive engagement, while affective and behavioral engagement vary greatly from person to person, with significant differences between samples.

Table 1. Descriptive statistics on learning engagement and academic achievement

Descriptive statistics	Learning engagement	Cognitive engagement	Affective engagement	Behavior engagement	Academic achievement
Mean	72.778	86.677	77.917	53.741	89.065
Stand Deviation	10.971	4.313	15.982	29.331	8.965

The results from both simple and multiple linear regression analyses indicated that overall learning engagement significantly and positively predicted academic achievement ($\beta=0.179$, $t=3.28$, $p=0.001$). However, the model exhibited weak explanatory power ($R^2=0.048$). In the multiple linear regression analysis of the three engagement dimensions, only cognitive engagement emerged as a significant predictor; nonetheless, the explanatory power remained limited ($t=3.868$, $p<0.001$, $R^2 = 0.120$). These findings suggest that while a statistical association exists at the aggregate level, a single composite engagement metric may obscure more complex underlying mechanisms. Given the potential existence of distinct patterns within the sample, it is necessary to employ cluster analysis to further explore whether specific engagement profiles exist that could better elucidate the variance in academic achievement.

Subsequently, K-means clustering was performed on the standardized total learning engagement scores and academic achievement to identify potential "engagement–achievement" learning patterns within the sample. To determine the optimal number of clusters, we experimented with various clustering configurations; the computational results are presented in Table 2. Balancing model fit with analytical interpretability, we selected $K=3$ as the final clustering solution, as it yielded the highest Silhouette Coefficient and a relatively high Calinski-Harabasz Index. The algorithm parameters were set to `random_state=42` and `n_init=10`.

Table 2. Different results of K-means clustering

K	Silhouette Coefficient	Calinski-Harabasz	Sum of Squared Errors
2	0.3822	147.75	257.23
3	0.3849	167.17	169.37
4	0.3718	168.2	128.82
5	0.3681	170.82	102.77

As illustrated in Table 3 and Figure 2, the clustering results yielded three distinct student groups. In terms of the two indicators, Cluster 1 exhibited high levels of both engagement and achievement. Cluster 2 was characterized by generally lower academic achievement and a dispersed but overall low distribution of engagement. Cluster 3 presented the most distinct profile, characterized by low engagement but high achievement.

Table 3. The details of three Clusters' student

Cluster	Number	Learning Engagement		Academic Achievement	
		M	SD	M	SD
Cluster 1	102	81.57	6.06	93.32	5.39
Cluster 2	43	69.26	9.07	75.26	6.65
Cluster 3	72	64.23	6.17	91.28	5.53

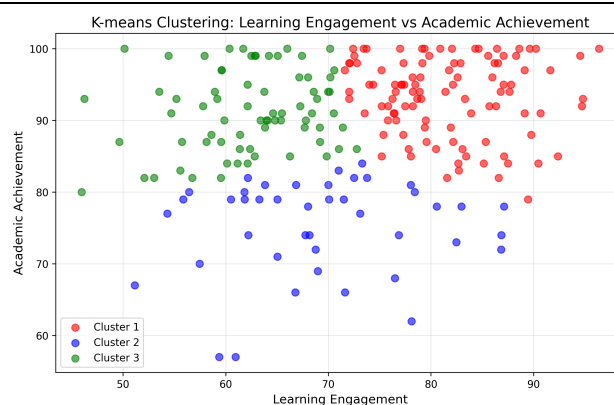


Fig. 2 K-means clustering results under the dimensions of learning engagement and academic performance

4.2 The Differences of Three Dimensions of Learning Engagement in Three Patterns

To further investigate the differences among the three student profiles across the three dimensions of learning engagement, a mixed-design Analysis of Variance (ANOVA) was conducted. In this analysis, the cluster category served as the between-subjects factor, while the dimension of learning engagement served as the within-subjects factor. The results are presented in Table 4. Regarding the between-subjects effects, significant differences in overall learning engagement levels were observed among the different cluster groups ($F(2, 214) = 175.0, p < .001$), verifying that the cluster analysis effectively discriminated participants based on engagement dimensions. Regarding the within-subjects' effects, significant differences were found across the three dimensions of learning engagement for the overall sample ($F(2, 428) = 272.0, p < .001$), indicating distinct levels of student engagement across cognitive, affective, and behavioral domains. Furthermore, the significant interaction effect between the engagement dimensions and cluster category suggests that the variations in learning engagement among the different clusters are not merely attributable to simple differences in overall magnitude ($F(4, 428) = 51.0, p < .001$); rather, they manifest as distinct patterns within the structural composition of engagement.

Table 4. The result of Analysis of Variance (ANOVA)

Effect type		Sum of squares	df	Mean square	F	P
Within subject's effect	Engagement type	143822	2	71911	272.0	<.001**
	Engagement type *					
	Cluster	52879	4	13470	51.0	<.001**
Between subject's effect	Cluster	48384	2	24192	175.0	<.001**

To elucidate these patterns, this study proceeded with an analysis of estimated marginal means for the interaction effect (as shown in Figure 3), which clearly delineates three distinct engagement profiles. As indicated by the figure, the cognitive dimension exhibited the highest mean scores, with all three student clusters displaying a high degree of convergence. This suggests that most students possess a solid foundation in cognitive volition, indicating that cognitive factors are not the primary bottleneck distinguishing academic achievement. In contrast, mean scores for affective engagement began to decline, showing moderate differentiation between groups; notably, Cluster 3 (characterized by low engagement but high achievement) exhibited the most pronounced decrease. In the behavioral engagement dimension, a precipitous divergence emerged among the three patterns. Students in the high-engagement/high-achievement profile maintained elevated levels, whereas the other two clusters—characterized by lower overall engagement—experienced a sharp decline in scores.



Fig.3 Estimated marginal means of three type of students

4.3 The Epistemic Network Analysis Result of Three Pattern Students

However, current results remain insufficient to explain why Cluster 3 students achieved high academic achievement despite lower levels of learning engagement. Consequently, this study encoded the comments posted by the three student clusters on the online learning platform and employed Epistemic Network Analysis (ENA) to delineate and compare the cognitive trajectories of different students, thereby uncovering the fundamental sources of these disparities. A moving window of 4 lines was used to capture connections between the current code and codes in the preceding 3 lines. The model achieved strong goodness of fit, with co-registration correlations of 0.92 (Pearson) for the first dimension, and 0.95 (Pearson) for the second dimension, ensuring the visualization accurately reflects the underlying network structure. The final analytical results are illustrated in Figures 4, 5 and 6.

Figure 4 presents the Epistemic Network Analysis results for Cluster 1, revealing that producing, developing, and supporting are the most prominent nodes. Students in this cluster tended to generate and offer ideas for constructing the project during forum interactions (producing), complemented by a high frequency of cognitive deepening (developing) and community support (supporting). In contrast, for Cluster 2, characterized by both lower learning engagement and academic achievement in figure 5, the centroid of the cognitive network shifted toward the left. While similarly centered on content generation (producing), this was primarily linked with content summarization (summarizing) and referencing (inferencing). Compared to Cluster 1, Cluster 2 demonstrated a more mechanical understanding of content, lacking deep reflective behaviors. Figure 6 illustrates Cluster 3 (low engagement, high achievement), which featured numerous high-frequency nodes with dense inter-nodal connections. It can be inferred that although these students exhibited a deficit in overall engagement magnitude, they maintained active cognitive processes and critical understanding; their engagement was characterized by lower frequency but higher quality, indicative of superior learning efficiency.

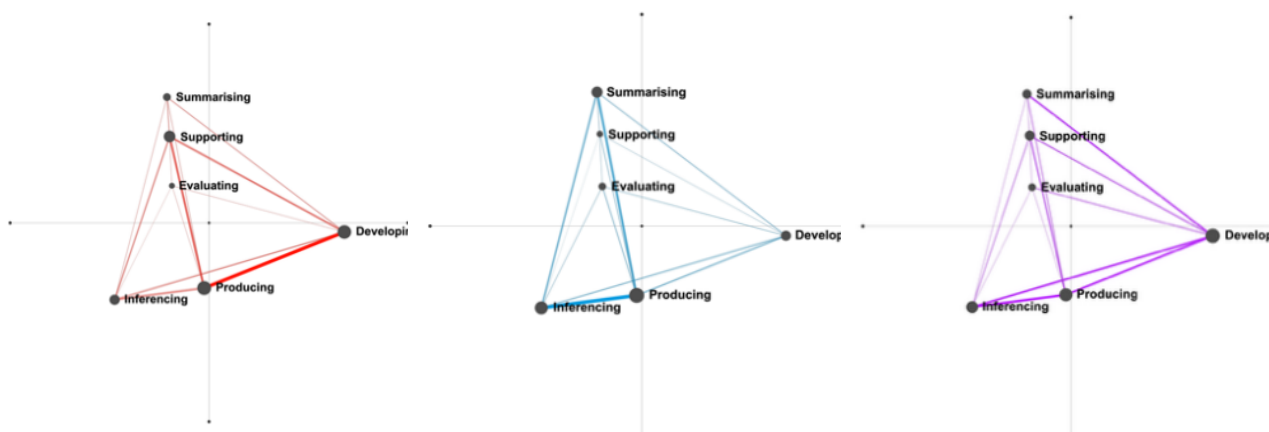


Fig.4 Epistemic Network of
Cluster 1 Students

Fig.5 Epistemic Network of Cluster 2
Students

Fig.6 Epistemic Network of
Cluster 3 Students

5. Conclusion

By employing regression analysis, K-means clustering, Analysis of Variance (ANOVA), and Epistemic Network Analysis (ENA), this study comprehensively addressed the relationship and distributional characteristics between students' online learning engagement and academic achievement. Furthermore, through the integration of multiple statistical methodologies, the research elucidated the underlying causes driving the emergence of distinct student profiles.

First, regression analysis revealed a positive correlation between students' learning engagement and academic achievement, a finding that aligns with previous research; however, the effect size was limited and the correlation proved weak. Moreover, in the multiple regression analysis, only cognitive engagement demonstrated a significant association with academic achievement. These results indicate that student dynamics cannot be adequately generalized by a simple positive correlation. Consequently, K-means clustering results delineated the student population into three primary groups: high engagement/high achievement, lower engagement/low achievement, and low engagement/high achievement.

Subsequently, ANOVA was employed to assess the differences among the three student clusters across the three dimensions of learning engagement. The estimated marginal means indicated that the disparities in engagement among the three groups were primarily concentrated in behavioral engagement. Specifically, students in the high-achievement/high-engagement cluster maintained elevated levels across all three engagement dimensions. In contrast, the other two clusters exhibited a substantial decline in behavioral engagement, underscoring the critical importance of behavioral participation in the context of online learning.

Finally, we utilized Epistemic Network Analysis (ENA) to further investigate the cognitive characteristics manifested in the comments of different student profiles on the online learning platform. A comparison of the resulting network visualizations reveals that high-engagement/high-achievement students concentrated on knowledge generation (producing), support (supporting), and deepening (developing). Conversely, lower-engagement/low-achievement students remained fixated on knowledge referencing (inferencing) and summarization (summarizing). Notably, students with low engagement but high achievement exhibited more comprehensive and integrated cognitive features, characterized by greater overall intellectual vitality. It can be inferred that within the low-engagement population, divergences in academic performance may be attributable to variations in the depth and diversity of cognition.

Through this study, we make both theoretical and practical contributions. Theoretically, this research refines the correlational patterns between learning engagement and academic achievement, extending beyond the singular positive correlation conclusions found in prior literature. Practically, it offers valuable insights for instructors delivering online learning to analyze student dynamics and enhance academic performance. Specifically, educators are advised to encourage students to actively utilize online platforms to bolster behavioral engagement, and to foster the posting of comments characterized by greater cognitive depth and breadth.

6. Limitation

First, the study collected data from only three online class sessions, and this relatively short data collection window may have limited the representativeness of the dataset. Second, the study did not examine specific behaviors within each subdimension of learning engagement, resulting in limited dimensional granularity; these issues may partly explain the relatively low coefficient of determination in the regression analysis between learning engagement and academic achievement. Finally, in the cluster analysis of student types, the selected clustering solution yielded a relatively low silhouette coefficient, and the scatterplot did not show clear separation between clusters, suggesting that the clustering results lacked sufficient explanatory power.

References

- Barnard, L., Lan, W. Y., To, Y. M., Paton, V. O., & Lai, S. L. (2009). Measuring self-regulation in online and blended learning environments. *The internet and higher education*, 12(1), 1-6. <https://doi.org/10.1016/j.iheduc.2008.10.005>
- Biasutti, M. (2017). A coding scheme to analyse the online asynchronous discussion forums of university students. *Technology, Pedagogy and Education*, 26(5), 601-615. <https://doi.org/10.1080/1475939X.2017.1365753>
- Freidhoff, J. R. (2021). Michigan's k-12 virtual learning effectiveness report 2019-20. Michigan Virtual University. <https://michiganvirtual.org/research/publications/michigans-k-12-virtual-learning-effectiveness-report-2019-20/>
- Hatahet, T. (2025). Supporting active learning opportunities with feedback enhances student achievement in pharmaceutical innovation module: Implication of behavioural and cognitive engagement. *Social Sciences & Humanities Open*, 11, 101347. <https://doi.org/10.1016/j.ssaho.2025.101347>
- Lee, J., Park, T., & Davis, R. O. (2022). What affects learner engagement in flipped learning and what predicts its outcomes?. *British Journal of Educational Technology*, 53(2), 211-228. <https://doi.org/10.1111/bjet.12717>
- Martin, F., & Borup, J. (2022). Online learner engagement: Conceptual definitions, research themes, and supportive practices. *Educational Psychologist*, 57(3), 162-177. <https://doi.org/10.1080/00461520.2022.2089147>
- McGaughey, F., Watermeyer, R., Shankar, K., Suri, V. R., Knight, C., Crick, T., ... Chung, R. (2022). 'This can't be the new norm': academics' perspectives on the COVID-19 crisis for the Australian university sector. *Higher Education Research & Development*, 41(7), 2231–2246. <https://doi.org/10.1080/07294360.2021.1973384>
- Miao, J., & Ma, L. (2022). Students' online interaction, self-regulation, and learning engagement in higher education: The importance of social presence to online learning. *Frontiers in psychology*, 13, 815220. <https://doi.org/10.3389/fpsyg.2022.815220>
- Moore, M. G. (1989). Three types of interaction. <https://doi.org/10.1080/08923648909526659>
- National Center for Educational Statistics (NCES). (2021). Condition of education 2021. <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2021144>
- Olivier, E., Galand, B., Belanger, J., & Morin, A. J. (2025). Multilevel student-perceived teaching practices profiles: Associations with competence beliefs, task value, behavioral engagement, and academic achievement. *Learning and Individual Differences*, 118, 102631. <https://doi.org/10.1016/j.lindif.2025.102631>
- Ong, S. G. T., & Quek, G. C. L. (2023). Enhancing teacher–student interactions and student online engagement in an online learning environment. *Learning environments research*, 26(3), 681-707. <https://doi.org/10.1007/s10984-022-09447-5>
- Puri, B., Mushtaque, I., Fang, S., Chenhe, G., & Younas, A. (2025). The impact of mobile-based language learning on speaking and learning anxiety, engagement and achievement in Chinese language learning: The mediating role of cognitive load. *Acta Psychologica*, 259, 105400. <https://doi.org/10.1016/j.actpsy.2025.105400>
- Trust, T., & Whalen, J. (2020). Should Teachers be Trained in Emergency Remote Teaching? Lessons Learned from the COVID-19 Pandemic. *Journal of Technology and Teacher Education*. <https://doi.org/10.70725/307718pkpjuu>
- Wang, Y., Cao, Y., Gong, S., Wang, Z., Li, N., & Ai, L. (2022). Interaction and learning engagement in online learning: The mediating roles of online learning self-efficacy and academic emotions. *Learning and Individual Differences*, 94, 102128. <https://doi.org/10.1016/j.lindif.2022.102128>
- Xie, X., Siau, K., & Nah, F. F. H. (2020). COVID-19 pandemic – online education in the new normal and the next normal. *Journal of Information Technology Case and Application Research*, 22(3), 175–187. <https://doi.org/10.1080/15228053.2020.1824884>

Reframing Catalog Design as Transmedia Brand Storytelling: A Curriculum Intervention and Mixed-Methods Evaluation

Menghan Yang^{1*}, Xin Zhao¹, Weiqing Guo¹

¹ School of Art and Media, Hebei University of Engineering Science, Hebei, 050091, CN

* Annnnnayang@163.com

Abstract: *Traditional catalog design courses often remain print-centered and struggle to cultivate cross-platform brand storytelling and narrative-system thinking. This study integrated transmedia storytelling into a senior undergraduate Catalog Design course through four stages: cognitive restructuring, systematic practice, value creation, and multiple evaluations. Using a quasi-experimental nonequivalent-groups design, 64 students (experimental n = 32; control n = 32) completed rubric-based pre/post assessments across Narrativity, Innovation, Synergy, Technicity, and Completion. Baseline equivalence was confirmed (all $p > 0.05$). The experimental group showed significant gains on all dimensions (all $p < 0.001$), with the largest increases in Narrativity (+70.20%) and Synergy (+71.90%). Post-test comparisons showed that the experimental group outperformed the control group across all dimensions (all $p < 0.05$). Interviews and observations further suggested a shift from “technical operator” to “brand narrative architect,” while also revealing difficulties in maintaining executional consistency across media. Despite the reduced analytic sample, the findings provide preliminary evidence for the value of transmedia pedagogy in catalog design education.*

Keywords: transmedia storytelling, catalog design, curriculum intervention, mixed methods, design education

1. Introduction

Under media convergence, design practice has moved beyond single-medium visual communication toward cross-platform brand narrative ecosystems (Jenkins, 2006; Scolari, 2015). Yet many catalog design courses remain anchored in print-based layout training and do not adequately address narrative construction, media coordination, or brand-world thinking. Although transmedia storytelling has been widely discussed in media and education research, its systematic translation into concrete design pedagogy remains underdeveloped (Jenkins, 2003, 2006; Rodrigues & Bidarra, 2014). This study therefore proposes and evaluates a four-stage transmedia-integrated curriculum model for a senior undergraduate Catalog Design course.

2. Course Intervention and Research Design

The study was conducted in a 32-hour Catalog Design course for fourth-year Visual Communication Design majors at Hebei University of Engineering Science. The intervention reframed the catalog as a narrative interface within a brand communication ecosystem. The experimental condition emphasized brand storytelling, cross-platform planning, and narrative continuity, whereas the control condition followed a conventional layout-focused model. A quasi-experimental nonequivalent-groups design was adopted using intact classes. Four classes constituted the instructional population ($N = 144$), with two assigned to each condition under pre-existing arrangements. Both conditions had the same contact hours, comparable workload, and the same instructional team. Of the 144 students, 64 (experimental $n = 32$; control $n = 32$) provided complete paired submissions for analysis; excluded cases involved missing or non-comparable submissions, so selection bias cannot be ruled out. Quantitative evidence was generated through a five-dimensional rubric scored by two course instructors and one industry mentor. After rubric calibration, inter-rater reliability was acceptable ($ICC = .79$). Independent-samples t-tests were used for baseline and post-test comparisons, and paired-samples t-tests examined within-group change. Qualitative evidence came from classroom observations and semi-structured interviews with 10 experimental-group students. The intervention was designed not only to improve project outcomes but also to promote a

shift in students' professional self-understanding, from technical executors of print layouts toward designers capable of constructing coherent brand narratives across multiple media touchpoints in practice.

3. Results and Discussion

The two groups were comparable at baseline, with no statistically significant differences on any rubric dimension (all $p > 0.05$). Within the experimental group, paired-samples t-tests showed significant gains across all five dimensions (all $p < 0.001$). The largest increases were found in Narrativity (2.42 $\rightarrow$ 4.12; +70.20%) and Synergy (2.24 $\rightarrow$ 3.85; +71.90%), suggesting that the intervention was especially effective in strengthening story construction and cross-media integration.

Post-test comparisons showed that the experimental group outperformed the control group on all five dimensions. The largest between-group advantages appeared in Synergy ($\Delta M = 1.32$, $p < 0.001$) and Narrativity ($\Delta M = 1.10$, $p < 0.001$), followed by Innovation ($\Delta M = 0.74$, $p = 0.002$). Differences in Technicity ($\Delta M = 0.40$, $p = 0.036$) and Completion ($\Delta M = 0.31$, $p = 0.048$) were smaller but still significant, suggesting that the model adds more value in narrative-system thinking and cross-platform orchestration than in foundational execution.

Qualitative findings supported these patterns. Students reported a shift from treating the catalog as an “information container” to viewing it as a “narrative interface,” and from layout-first execution to research-driven narrative planning. Several students also articulated a stronger sense of themselves as “brand narrative architects.” At the same time, some students struggled to maintain visual continuity and executional consistency across media, helping explain why gains in Technicity and Completion were more modest.

4. Conclusion

This study provides preliminary evidence that integrating transmedia storytelling into catalog design education can strengthen narrative competence, cross-media coordination, and professional identity formation. The strongest effects appeared in Narrativity and Synergy. However, the findings should be interpreted cautiously because the study used intact classes, relied on an analytic sample of 64 students, and focused on one institutional setting. Although inter-rater reliability was acceptable, rubric-based assessment still involves human judgment. Future research should employ larger samples, more detailed attrition analyses, and longer-term follow-up to examine the transferability and sustainability of transmedia pedagogy in design education.

References

- Feng, M., Cao, R., & Chen, Q. (2023). Transformation and reconstruction of art and design education under the impact of AIGC: Opportunities, paradigms, and responses. *Industrial & Engineering Design*, 5(4), 47–58.
- Jenkins, H. (2003, January 15). Transmedia storytelling. *MIT Technology Review*.
- Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. New York University Press.
- Rodrigues, P., & Bidarra, J. (2014). Transmedia storytelling and the creation of a converging space of educational practices. *International Journal of Emerging Technologies in Learning*, 9(6), 42–48.
- Scolari, C. A. (2015). Transmedia storytelling: Brands, narratives and storyworlds. In G. Rossolatos (Ed.), *Handbook of brand semiotics* (pp. 151–169). Kassel University Press.
- Scolari, C. A., Masanet, M.-J., Guerrero-Pico, M., & Establés, M.-J. (2018). Transmedia literacy in the new media ecology: Teens' transmedia skills and informal learning strategies. *El profesional de la información*, 27(4), 801–812.

Trends in Artificial Intelligence Research for Geometry Education: Educational Applications and System-Oriented Approaches

Yiru Chen ¹, Yin Yang ^{2*},

^{1,2} The Education University of Hong Kong

s1146375@s.edu.hk

* yyin@edu.hk

Abstract: Artificial intelligence (AI) has increasingly been introduced into mathematics education research, with geometry attracting particular attention due to its reliance on visualization and reasoning. However, existing studies remain fragmented, with limited synthesis of how AI applications relate to instructional practice. This paper presents a focused review to clarify research orientations in AI-supported geometry education. A PRISMA-informed screening of 1,468 publications (2015–2025) resulted in 80 studies for analysis. Using qualitative content analysis, the studies were categorized by research orientation and educational focus. The preliminary analysis revealed two dominant orientations. Educationally oriented AI applications were embedded in instructional contexts to support conceptual understanding and exploratory learning. In contrast, system-oriented studies focused on the development and evaluation of AI techniques for geometric problem solving, automated reasoning, or representation. Still, they rarely reported classroom implementation or pedagogical design considerations. The findings highlight the need to connect AI capabilities with instructional design better to support meaningful geometry learning.

Keywords: artificial intelligence, mathematics education, geometric education, instructional design

1. Introduction

The integration of Artificial Intelligence (AI) into mathematics education has advanced rapidly, yet its application in geometry education remains challenging (Yunianto et al., 2024). Despite advances in AI systems for formal proof generation and geometric problem solving, a significant gap remains between their technical sophistication and meaningful integration into student-centered learning environments (Recio et al., 2019). Existing research on AI in geometry education has developed along two largely parallel strands. One strand emphasizes educationally oriented applications in which AI is embedded within structured, teacher-mediated instructional settings to support conceptual understanding and exploratory learning (Recio et al., 2019; Thampitak & Junpeng, 2024). A second strand focuses on system-oriented AI developments, including large language models and automated reasoning systems, which are primarily evaluated based on their performance in formal proof generation or competition-level geometric problem solving (Wei, 2024; Zou et al., 2024). Although these studies demonstrate advances in formal reasoning capabilities, they are often conducted outside instructional settings and pay limited attention to how students interpret or learn from AI-generated reasoning (ICMI, 2024). Against this backdrop, a focused synthesis is needed to clarify how AI has been conceptualized and operationalized in geometry education research. This review aims to answer one research question: what trends are in artificial intelligence research for geometry education.

2. Method

Following PRISMA-informed procedures, 1,468 studies on artificial intelligence and geometry were identified from Web of Science, Scopus, Education Resources Information Center (ERIC), and the ACM Digital Library (2015–2025). Titles and abstracts were first screened to exclude studies that did not explicitly address geometry teaching or learning contexts. To enhance the reliability of the screening process, the inclusion decisions were independently reviewed by two

researchers. Based on predefined criteria, 80 studies were retained for analysis. The studies were categorized into two orientations: (1) Educationally oriented AI applications emphasizing instructional design and learning outcomes; and (2) AI tools with potential educational relevance focusing on algorithm or system development without classroom implementation. A coding scheme was developed to capture key dimensions related to instructional integration, learning focus, and the role of AI in supporting geometric reasoning. Instructional context: Whether the AI system was implemented within formal instructional settings (e.g., school classrooms, curriculum-aligned platforms) or evaluated outside educational contexts. Learning focus: The primary geometric focus addressed by the AI system, including conceptual understanding (e.g., properties, relationships), spatial visualization, procedural computation, or formal deductive reasoning. Pedagogical Role of AI: The function of AI in the learning process, coded as instructional support (e.g., scaffolding, formative feedback, adaptive guidance), exploratory facilitation (e.g., conjecturing, visualization), or autonomous problem solving without pedagogical mediation (Zou et al., 2024; Wei, 2024; Zhang et al., 2025). Integration with instructional design was evaluated on three levels: (1) Technical evaluation only (e.g., accuracy or efficiency), (2) potential educational relevance without explicit instructional design, and (3) explicit integration into instructional tasks, activity sequences, or teacher-mediated learning.

3. Results

3.1. Educationally Oriented AI Applications in Geometry

Based on the coding scheme, educationally oriented AI applications in geometry exhibit a high degree of instructional integration and pedagogical alignment. 28 studies in this category were coded at Level 3 (explicit instructional integration).

3.1.1. Pattern 1: Strong coupling between AI and instructional context

Across the reviewed studies, educationally oriented AI applications in geometry are predominantly implemented within formal school-based learning contexts, including school classrooms, web-based learning platforms, and technology-enhanced instructional environments integrated into existing curricula (Cunha et al., 2025; Recio et al., 2019; Thampitak & Junpeng, 2024). In these studies, AI systems are embedded within structured instructional contexts aligned with grade-level objectives and classroom practices. For example, intelligent tutoring systems and AI-supported diagnostic platforms were used in middle and high school mathematics classrooms to support geometry and measurement learning (Thampitak & Junpeng, 2024). Similarly, GeoGebra-based automated reasoning tools were incorporated into task-based classroom activities, where students interact with dynamic geometry software during guided exploration sessions led by teachers (Recio et al., 2019). AI-enhanced immersive or extended reality environments were also used to support geometry instruction, typically as a complement to traditional teaching rather than a replacement (Cunha et al., 2025).

3.1.2. Pattern 2: Emphasis on conceptual and exploratory learning

In terms of learning focus, educationally oriented studies primarily emphasize conceptual understanding, spatial reasoning, and conditional geometric reasoning. AI is used to support students in exploring geometric properties, spatial relationships, and the conditions under which geometric statements hold. For instance, GeoGebra automated reasoning tools helped students explore when and under what conditions geometric properties hold, encouraging deeper conceptual engagement with geometry (Recio et al., 2019). In intelligent tutoring system-based studies, AI-supported diagnostic tools guided students through progressive levels of conceptual understanding as defined by structured learning outcome frameworks (Thampitak & Junpeng, 2024).

3.1.3. Pattern 3: AI as pedagogical support rather than autonomous problem solver

Pedagogically, AI-generated feedback is designed to scaffold students' reasoning processes, encourage reflection, or suggest next steps, instead of merely providing correct answers. For example, GeoGebra-based automated reasoning tools did not simply indicate correct or incorrect answers but prompted students to compare visual, numerical, and logical

representations, supporting reflective geometric reasoning (Recio et al., 2019). Overall, AI functions as a teaching aid that supports exploration, understanding, and flexible instruction in geometry learning.

3.2. AI Tools with Potential Educational Relevance to Geometry

In addition to educationally oriented applications, a growing body of research focuses on AI-based tools that have not been implemented in instructional settings, but address problems closely related to geometry learning.

First, a number of 12 studies were coded at Level 1. These studies typically evaluated AI systems in non-instructional contexts, resulting in limited instructional integration. Performance was assessed through benchmarks, formal proofs, or competition-style problems, with little attention to how learners interact with or learn from AI-generated reasoning (Svičević et al., 2025; Zhang et al., 2025)

Second, 40 studies were coded at Level 2. Some studies assessed large language models using geometry-related problem sets from national assessments and mathematical competitions. Results from NAEP and AMC geometry items indicated that, despite strong performance in procedural or algebraic tasks, current language models exhibited systematic weaknesses in spatial reasoning and complex geometric problem solving (Teegavarapu & Sanghvi, 2023; Wei, 2024).

Other studies focused on the development of AI systems for formal geometric reasoning, including automated reasoning and deep reinforcement learning approaches (Zou et al., 2024). These systems were typically evaluated in technical or experimental settings rather than classrooms. Nevertheless, they targeted core difficulties in geometry learning, particularly the selection of appropriate theorems and the construction of coherent chains of deductive reasoning. In addition, some studies explored the combined use of multiple AI tools, such as integrating large language models with dynamic geometry software, to support geometric discovery and conjecturing. While these approaches did not report measured learning outcomes, they frame AI as a means to prompt exploration, verification, and reflection on geometric relationships (Botana et al., 2024).

4. Discussion

This review provided a focused and exploratory analysis of recent research on AI in geometry education by distinguishing two dominant research orientations: educationally oriented AI applications embedded in instructional contexts, and system-oriented AI tools evaluated primarily through technical performance. Rather than offering an exhaustive synthesis, the review aimed to clarify how AI had been conceptualized and positioned across these two strands and to identify recurring patterns in their relationship to geometry teaching and learning.

The analysis revealed a clear contrast between the two orientations. Educationally oriented studies tended to integrate AI into task-based instructional designs, emphasizing conceptual understanding, exploratory learning, and teacher mediation. In contrast, system-oriented studies primarily assessed AI capabilities in non-instructional contexts, focusing on formal correctness, automated reasoning, or benchmark performance, with limited attention to pedagogical design or learner interaction.

Due to its focused scope and space constraints, only a subset of representative studies was analyzed in depth, and the coding scheme was intentionally coarse-grained. As a result, the findings should be interpreted as indicative patterns rather than comprehensive mappings of the field. More fine-grained reviews could further examine variations across educational levels, geometric topics, AI techniques, and learning theories, as well as incorporate quantitative synthesis where appropriate.

References

- Botana, F., Recio, T., & Vélez, M. P. (2024). On using GeoGebra and ChatGPT for geometric discovery. *Computers (Basel)*, 13(8), 187. <https://doi.org/10.3390/computers13080187>

- Cunha, C. R., Moreira, A., Coelho, S., Mendonça, V., & Gomes, J. P. (2025). Converging extended reality and machine learning to improve the lecturing of geometry in basic education. *Mağallaṯ Al-Abḥath al-Handasiyyaṯ*, 13(4), 3121 – 3131. <https://doi.org/10.1016/j.jer.2024.10.016>
- International Commission on Mathematical Instruction (ICMI). (2024). *The 26th ICMI Study: Advances in geometry education* (Discussion Document). ICMI. <https://www.mathunion.org/icmi/icmi-study-26>
- Recio, T., Richard, P. R., & Pilar Vélez, M. (2019). Designing tasks supported by GeoGebra automated reasoning tools for the development of mathematical skills. *International Journal for Technology in Mathematics Education*, 26(2), 81-88.
- Svičević, M., Milenković, A., Vučičević, N., & Stanković, M. (2025). Evaluating the success of AI tools in supporting student performance in Mathematical Kangaroo Competition. *Computer Applications in Engineering Education*, 33(4). <https://doi.org/10.1002/cae.70063>
- Teegavarapu, R. R., & Sanghvi, H. (2023, December). Analyzing the competitive mathematical problem-solving skills of chatGPT. In *2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)* (pp. 1-7). IEEE.
- Thampitak, D., & Junpeng, P. (2024, July). Diagnosing structure of observed learning outcomes in measurement and geometry for gifted students of 7th grade through machine learning platform, Khon Kaen Wittayayon School, Thailand. In *2024 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC)* (pp. 1-6). IEEE.
- Wei, X. (2024). Evaluating chatGPT-4 and chatGPT-4o: performance insights from NAEP mathematics problem solving. *Frontiers in Education (Lausanne)*, 9. <https://doi.org/10.3389/feduc.2024.1452570>
- Yunianto, W., Lavicza, Z., Kastner-Hauler, O., & Houghton, T. (2024). Investigating the use of chatGPT to solve a GeoGebra based mathematics+computational thinking task in a geometry topic. *IndoMS-Journal on Mathematics Education*, 15(3), 1027 – 1052. <https://doi.org/10.22342/jme.v15i3.pp1027-1052>
- Zhang, Z., Cheng, J.-K., Deng, J., Tian, L., Ma, J., Qin, Z., Zhang, X., Zhu, N., & Leng, T. (2025). Diagram formalization enhanced multi-modal geometry problem solver. *Proceedings of the ... IEEE International Conference on Acoustics, Speech and Signal Processing* (1998), 1–5. <https://doi.org/10.1109/ICASSP49660.2025.10889286>
- Zou, J., Zhang, X., He, Y., Zhu, N., & Leng, T. (2024). FGeo-DRL: Deductive reasoning for geometric problems through deep reinforcement learning. *Symmetry (Basel)*, 16(4), 437. <https://doi.org/10.3390/sym16040437>

From "Wheel-Spinning" to "Productive Failure": Research on LLM-based Hierarchical Scaffolding Mechanism in Programming Education

Xinyun Wang^{1*}, Bo Jiang²,

¹ Department of Education Information Technology, East China Normal University

² Shanghai Institute of AI for Education, East China Normal University

* 51264108008@stu.ecnu.edu.cn

Abstract: Generative AI lowers barriers to programming but risks cognitive offloading—novices outsourcing reasoning to algorithms. This study embeds an LLM-driven hierarchical tutoring mechanism into an Online Judge (OJ) system, featuring five scaffolding levels (metacognitive prompts to direct coaching) and active/passive dual-mode intervention to interrupt unproductive “wheel-spinning.” A quasi-experimental study ($n = 25$) showed the experimental group scored significantly higher on a closed-book mid-term ($M = 86.45$ vs. 79.46 ; Cohen’s $d = 1.07$). Process analysis revealed faster iteration cycles and a strategic “submission-as-interaction” behavior, confirmed by student interviews. We frame these results as behavioral evidence for a promising scaffolding mechanism while acknowledging confounding limitations.

Keywords: Generative AI, Programming Education, Productive Failure, Hierarchical Scaffolding

1. Introduction

LLMs have transformed programming education: GenAI can translate natural language into executable code, eliminating syntactic friction. Yet this convenience carries a cost. When AI removes the cognitive demand of debugging and problem-solving, novices lose the opportunity to build durable mental schemas through productive struggle (Kapur, 2016). The result is a core instructional paradox—AI helps students finish tasks while undermining the very processes that build competence. This study addresses that paradox with two research questions: (RQ1) How can an LLM-based hierarchical tutoring mechanism improve programming performance? (RQ2) How does dual-mode AI intervention transform unproductive failure into productive failure?

2. Theoretical Background

Productive Failure and Wheel-Spinning. Kapur (2008, 2016) argues that exploring ill-structured problems before instruction activates prior knowledge and deepens conceptual construction, even at the cost of initial performance. In programming, however, unconstrained failure risks “wheel-spinning”—spending significant time without progress (Beck & Gong, 2013)—which can escalate into learned helplessness (Maier & Seligman, 2016). Effective design must therefore hold students within a zone of productive struggle.

Adaptive Scaffolding. Koedinger and Aleven (2007) describe the “assistance dilemma”: too much help suppresses thinking; too little causes frustration. Hattie and Timperley (2007) show that feedback should prioritize self-regulation over task-level answers. Incremental, fading scaffolding is empirically superior for novice programmers (Zheng et al., 2022), with direct code hints reserved as a fail-safe worked example (Shih et al., 2010).

3. System Design

3.1 Architecture

The platform uses a B/S architecture OJ (Wasik et al., 2018) with three layers: frontend (code editor, compilation feedback), backend (LLM tutoring agent), and data persistence. For the control group, it operates as a standard OJ. For the experimental group, the backend intercepts every submission in real time, selects a scaffolding response via rule-

based routing and prompt engineering, and returns a pedagogical intervention—never raw code. Fine-grained logs capture every Run operation, error message, dwell time, and dialogue turn.

3.2 Hierarchical Scaffolding Mechanism

The system employs five scaffolding levels governed by a “reset-on-novelty” rule: a new error type resets the level to L0, granting renewed exploratory space; repeated identical errors escalate the level. This operationalizes Zheng et al.’s (2022) fading principle—maximum cognitive demand by default, degrading only when behavioral evidence demands it.

Example: an `IndexError` recurring for the second time triggers L1 (“Does your assumption about list length hold?”). A third occurrence escalates to L2 (“What happens when the index reaches the last element?”). A new `TypeError` resets to L0.

Table 1 Operational definitions of scaffolding levels

Level	Taxonomy	Instructional Description
Level 0	Baseline	No pedagogical support or feedback is provided by the system.
Level 1	Metacognitive Tutoring	The system merely prompts the student to check the alignment between their current strategy and the task requirements, promoting self-monitoring.
Level 2	Socratic Questioning	The system guides students to discover logical loopholes through reflective questioning.
Level 3	Analogical Prompting	The system utilizes similar cases or examples to inspire conceptual understanding.
Level 4	Direct Coaching	Specific code snippets or bottom-level logic corrections are provided.

3.3 Dual-Mode Interaction

When the backend detects wheel-spinning (repeated identical errors within a time window), it automatically pushes an intervention following the L1→L2→L3→L4 sequence. Clicking “Request Help” defaults to L2 (Socratic). Even if a student explicitly requests code, a pre-set system prompt forces the LLM to respond with guiding questions. Level escalation requires a secondary LLM classifier to confirm a severe knowledge gap.

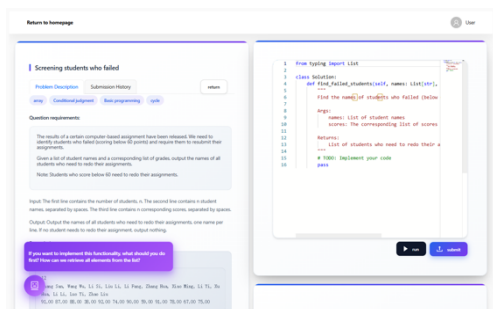


Figure 1 The active mode schematic diagram

4. Experiment

A quasi-experimental RCT was conducted in a Python programming course at a Chinese university over eight weeks ($n = 25$). Participants were stratified by a programming pre-test and assigned to the experimental OJ (LLM scaffolding) or standard OJ (control). Both groups completed identical weekly assignments; the same instructor delivered identical lectures. A closed-book, AI-free mid-term exam at week four served as the sole summative outcome. Platform logs recorded submission intervals, error sequences, and scaffolding interactions. Post-experiment semi-structured interviews provided qualitative context.

Three behavioral metrics were derived from logs: (1) Ineffective Attempt Rate—proportion of consecutive submission pairs with no change in error type; (2) Average Submission Interval—mean time between submissions within a one-hour session; (3) Scaffolding Level Distribution—weekly proportion of interactions at each level.

5. Results

5.1 Learning Outcomes

The experimental group ($M = 86.45$, $SD = 5.20$) significantly outperformed the control group ($M = 79.46$, $SD = 7.40$); $t(23) = 2.63$, $p = .015$, Cohen's $d = 1.07$. The narrower score distribution (Figure 2) suggests the mechanism particularly supported lower-performing students, reducing within-group variance and promoting equity. We note that the eight-week span introduces plausible confounds—freed from manual debugging, experimental students may have redirected time to exam-relevant study—so results should be interpreted as indicative rather than causally definitive.

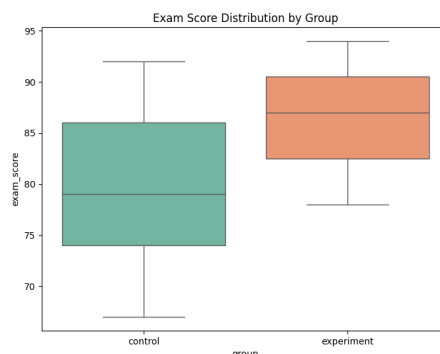


Figure 2 Comparison of mid-term exam scores between control and experimental groups

5.2 Learning Process

Both groups achieved near-100% problem completion, demonstrating equal persistence. The experimental group's average submission interval was ~ 120 s, one-third of the control (~ 360 s), consistent with a “high-frequency iteration and rapid feedback” learning loop (Figure 3). Crucially, the experimental group's superior offline exam performance rules out the explanation that frequent submissions simply offloaded assignment completion without genuine learning.

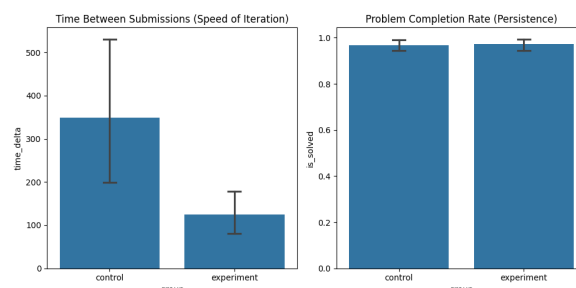


Figure 3 Comparison of submission intervals and problem completion rates

5.3 Strategic Scaffolding Use

The experimental group's Ineffective Attempt Rate was significantly higher than the control's (Figure 4)—counterintuitive under traditional definitions. Interviews revealed a deliberate “submission-as-interaction” strategy: students found submitting a known-failing snippet faster than typing a problem description in the chat interface. One participant noted: “Submitting and seeing the error is faster than explaining my confusion—the system just tells me what to look at.” This reframes the high ineffective rate as active engagement with the trigger logic, not disengagement. The offline exam advantage confirms cognitive internalization rather than system exploitation.

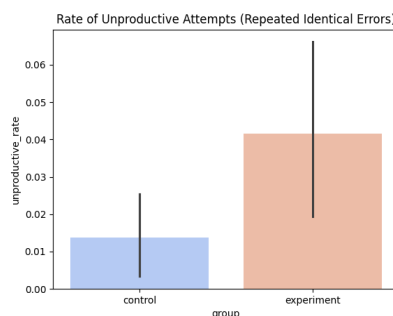


Figure 4 Comparative Analysis of Unproductive Attempt Rates (Repeated Identical Errors)

6. Limitations

Several limitations bound interpretation. The sample ($n = 25$, single institution) limits generalizability. A single mid-term score cannot fully capture learning; delayed retention tests are needed. Confounds over eight weeks—differential study time, potential outside-tool use in the control group, and novelty effects—cannot be fully ruled out. We position this work as behavioral evidence for a promising mechanism, not a proof of causal efficacy.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China under Grant 62477012, and the Natural Science Foundation of Shanghai under Grant 23ZR1418500.

Reference

- Beck, J. E., & Gong, Y. (2013). Wheel-Spinning: Students Who Fail to Master a Skill. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education* (pp. 431–440). Springer. https://doi.org/10.1007/978-3-642-39112-5_44
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112.
- Kapur, M. (2008). Productive Failure. *Cognition and Instruction*, 26(3), 379–424.
- Kapur, M. (2016). Examining Productive Failure, Productive Success, Unproductive Failure, and Unproductive Success in Learning. *Educational Psychologist*, 51(2), 289–299. <https://doi.org/10.1080/00461520.2016.1155457>
- Koedinger, K. R., & Aleven, V. (2007). Exploring the Assistance Dilemma in Experiments with Cognitive Tutors. *Educational Psychology Review*, 19(3), 239–264. <https://doi.org/10.1007/s10648-007-9049-0>
- Maier, S. F., & Seligman, M. E. P. (2016). Learned Helplessness at Fifty: Insights from Neuroscience. *Psychological Review*, 123(4), 349–367. <https://doi.org/10.1037/rev0000033>
- Shih, B., Koedinger, K. R., & Scheines, R. (2010). A Response-Time Model for Bottom-Out Hints as Worked Examples. In *Handbook of Educational Data Mining*. CRC Press.
- Wasik, S., Antczak, M., Badura, J., Laskowski, A., & Sternal, T. (2018). A Survey on Online Judge Systems and Their Applications. *ACM Comput. Surv.*, 51(1), 3:1-3:34. <https://doi.org/10.1145/3143560>
- Zheng, L., Zhen, Y., Niu, J., & Zhong, L. (2022). An exploratory study on fade-in versus fade-out scaffolding for novice programmers in online collaborative programming settings. *Journal of Computing in Higher Education*, 34(2), 489–516. <https://doi.org/10.1007/s12528-021-09307-w>

Turning Course Descriptions into Skill Evidence: A Case Study of Two Undergraduate Programs

Xianghui Meng¹, Xian Chen², Jionghao Lin^{1*}

¹ Faculty of Education, The University of Hong Kong

² The University of Manchester

* Corresponding author: jionghao@hku.hk

Abstract: Course descriptions often state broad descriptions of what students will learn without clearly naming specific skills (e.g., data analysis, policy analysis) or tools (e.g., Python, SQL). We present a course-to-skill extractor that converts course description into Tools, Hard Skills, and Soft Skills while extracting only labels explicitly supported by the course-description text. Using Qwen2.5 (dense), we analyze 60 course descriptions from two undergraduate programs: one in data science in a social science context and one in information systems and management. The extraction workflow preserves traceability by storing each extracted skill label with its source course, allowing program-level summaries to be checked against the original course descriptions. Across both programs, most labels appear only once, while explicit tool names are rare.

Keywords: curriculum analysis, curriculum mapping, large language models, higher education, skill extraction

1. Introduction

When choosing courses, students rely on official course descriptions to assess program skills. However, these texts often use less detailed language. For example, a job posting lists Python or SQL directly, but a course description may say “analytical methods.” This mismatch can lead students to over-infer or miss relevant skills entirely, because a broad phrase such as “analytical methods” does not specify whether the course involves tools such as Excel, Python, or SQL, or methods such as statistical analysis or data modeling. Prior work has explored curriculum analysis through natural language processing (Rashtian et al., 2020) and large-scale skill mapping across higher education curricula (Sabet et al., 2024), yet fewer studies focus on extracting student-facing skill evidence from short course descriptions in ways that can be checked against the original descriptions. We therefore propose a structured extraction workflow that identifies explicitly stated skill labels, categorized into *Tools*, *Hard Skills*, and *Soft Skills*, with every label linked to its source course for full auditability.

2. Method

The proposed workflow consisted of five steps. *First*, we collected officially published course descriptions for the two undergraduate programs and treated each course description as one unit of analysis. *Second*, we processed each course description independently with Qwen2.5 (dense). *Third*, the model extracted only skill labels explicitly stated in the course-description text and organized them into three categories: Tools, Hard Skills, and Soft Skills. *Fourth*, we stored each extracted skill label together with its source course to preserve traceability. *Fifth*, we aggregated the course-level outputs into program-level summaries and compared the two programs mainly through recurring skill labels and relative label frequencies.

The dataset contains 60 officially published course descriptions from a university in Hong Kong in the 2025–2026 academic year: 41 from Program 1 (data science in a social science context) and 19 from Program 2 (information systems and management). Each record includes a program label, course code, course title, and description text. The descriptions are similarly short across both programs (mean length about 73 words), reducing the likelihood that observed differences are driven primarily by description length. Each description was processed independently with Qwen2.5 (dense). We

used a structured prompt with a fixed JSON schema to guide Qwen2.5 in generating skill extractions (tools, hard skills, and soft skills) from course descriptions. The prompt enforced a fixed JSON schema with three output lists: *Tools* (named technologies such as software or programming languages), *Hard Skills* (domain-specific technical methods or competencies), and *Soft Skills* (transferable capabilities such as communication or teamwork). The model was instructed to extract only explicitly stated labels and return empty lists when no supporting evidence was available. A low temperature setting (0.2) was used to improve stability, and outputs were replaced with empty lists if JSON parsing failed. Course-level outputs were then converted into a long-format course-to-skill table for counting, aggregation, and auditing. For program comparison, we treated extracted skills as curriculum-as-described evidence, interpreting cross-program differences mainly through recurring labels and size-normalized mention frequencies.

3. Results

The workflow produced complete outputs for all 60 courses and extracted 348 skill items in total: 239 from Program 1 and 109 from Program 2. As shown in Table 1, mean yield was highly similar across programs (5.83 vs. 5.74 items per course), suggesting that the observed differences are not simply a result of one program having longer or denser descriptions. The table also shows a strong singleton pattern in both programs and limited explicit tool naming, with only 19.5% of Program 1 courses and 15.8% of Program 2 courses naming tools. Together, these results suggest that short course descriptions provide many one-off labels but relatively little direct information about specific tools.

Table 1

Summary of results by program. Details of the metrics can be found in project repository.¹

Metrics	Program 1	Program 2
Courses analyzed	41	19
Total extracted skill items	239	109
Mean extracted skill item per course	5.83	5.74
Distinct extracted skill labels	188	100
Singleton share(%)	87.2	96.0
Courses with at least one tool (%)	19.5	15.8

The extracted skill labels showed a strong singleton pattern. Program 1 contained 188 distinct skill labels and Program 2 contained 100, but 87.2% and 96.0% of these labels, respectively, appeared only once within each program. This suggests that comparing programs using all extracted labels may overstate how broad or different the programs appear. We therefore compared the programs mainly through skill labels that appeared in more than one course. When labels were matched without distinguishing uppercase and lowercase letters, the two programs shared 56 labels. Within this shared set, Program 1 more often documented “Critical Thinking,” “Policy Analysis,” and “Data Analysis,” whereas Program 2 more often documented “Information Management.” A complete list of distinct skill labels for both programs is available in the project repository.¹ The Appendix in the repository provides the details for each reported metric, including exact definitions and equations. Explicit tool naming was limited in both programs, with tools named in 19.5% of Program 1 courses and 15.8% of Program 2 courses. This means that the course descriptions often did not clearly state which tools were used; it does not mean that tools were unimportant in teaching. When we compared the programs using relative

frequencies instead of raw counts, we reached the same conclusion: the clearest differences between the programs came from skill labels that appeared repeatedly across courses, rather than from labels that appeared only once.

4. Discussion and Conclusion

Our study shows that short official course descriptions can be converted into structured skill evidence without overstating what the text supports. The main strength of this workflow is not deep interpretation, but careful evidence handling. It extracts only information explicitly stated in the text, keeps the source course for each label, and compares programs based on labels supported repeatedly across courses. For students, the extracted skill summaries make heterogeneous course descriptions easier to inspect and compare. For program teams, these summaries can reveal where course descriptions do not clearly specify expected tools or rely on broad wording with limited practical guidance. These results describe only what is explicitly written in the course descriptions. They should be used to support follow-up review of course documents, not to draw conclusions about what was actually taught, what students mastered, or how well the curriculum matches labor-market needs. The study also shows why it is important to extract only skill labels that are explicitly supported by the course-description text.: allowing inferred skills into the dataset can produce summaries that are not strongly supported by the text. Future work could align near-synonymous labels through a controlled ontology, add human review for program differences, and analyze job advertisements with course descriptions to compare curriculum and labor-market skill profiles using the same categories.

Acknowledgements

This work was supported by the Teaching Development Grant (Project No. 1119) from the University of Hong Kong.

References

- Hilliger, I., Aguirre, C., Miranda, C., Celis, S., & Pérez-Sanagustín, M. (2022). Lessons learned from designing a curriculum analytics tool for improving student learning and program quality. *Journal of Computing in Higher Education*, 34(3), 633–657.
- Rashtian, H., Hashemi, A., & Macfadyen, L. (2020). Harnessing natural language processing to support curriculum analysis. In *ICERI2020 Proceedings* (pp. 1779–1784). IATED.
- Sabet, A. J., Bana, S. H., Yu, R., & Frank, M. R. (2024). Course-Skill Atlas: A national longitudinal dataset of skills taught in US higher education curricula. *Scientific Data*, 11(1), 1086.

A Study on Personalized Programming Learning for High School Students

Supported by Large Language Models

Chun-Lin Chou¹, Chin-Chun Chan², Pei-Hsuan Tung³, Yu Chieh Wu⁴, Chiu Pin Lin⁵

Tsing Hua University

*sherri92123@gmail.com

Abstract: This study investigates the impact of an LLM-based AI teaching assistant on personalized programming learning for high school students and examines its effects on learning achievement, self-efficacy, learning attitudes, and AI acceptance. Using a single-group pretest-posttest design, thirty students from an information technology course at a senior high school in Hsinchu, Taiwan participated in a twenty-week experiment. The first ten weeks involved traditional instruction; the subsequent ten weeks introduced an LLM-based AI teaching assistant providing personalized explanations, code diagnosis, error feedback, and conceptual guidance based on individual proficiency levels. Research instruments included a programming achievement test, self-efficacy scale, learning attitude scale, and AI acceptance scale. Quantitative data were analyzed using descriptive statistics, paired-sample t-tests, and Wilcoxon signed-rank tests. Qualitative data from student interviews, AI interaction logs, and teacher observations were analyzed through thematic analysis and cross-validation.

Keywords: LLM, personalized learning, AI teaching assistant, programming education, self-efficacy

1. Introduction

This study integrated an LLM-based AI teaching assistant into a Python elective course at a Taiwanese senior high school to supplement traditional instruction. The study explores how this system provides personalized learning support according to students' varying proficiency levels and to analyze its impact on students' programming learning achievement, self-efficacy, learning attitudes, and AI acceptance. The research objectives are as follows:

1. To investigate whether the AI teaching assistant can improve high school students' learning achievement in Python courses.
2. To explore the impact of the AI teaching assistant on students' self-efficacy, learning attitudes, and AI acceptance.
3. To analyze the inquiry process of student-AI teaching assistant interactions and teacher observation results.

2. Literature Review

2.1 Programming Learning Theory and Learning Difficulties

Programming is regarded as a highly challenging task in information education. Beginners must simultaneously understand syntax, control flow, and problem abstraction, which easily causes high cognitive load (Qian & Lehman, 2017). They often experience unstable mental models and accumulation of misconceptions in core concepts such as variables and loops (Robins et al., 2003; Kaczmarczyk et al., 2016). This variance in learning achievement is closely related to psychological motivation, with self-efficacy being an important factor affecting performance (Ramalingam et al., 2004; Lishinski, 2016). Without debugging strategies and immediate feedback, individuals with low self-efficacy are prone to anxiety and avoidance (Klassen & Klassen, 2018) and tend to accumulate misconceptions (Shute et al., 2016). According to Bandura's (1986) theory and Vygotsky's (1978) scaffolding learning perspective, learners can cross the Zone of Proximal Development (ZPD) with external support. With technological development, Large Language Models can serve as a new form of intelligent scaffolding, supporting students to gradually construct understanding by reducing task complexity (Reiser, 2004) and providing structured guidance (Kasneci et al., 2023):

2.2 Prompt Design and Theoretical Foundation of AI Teaching Assistants

Research indicates that specific prompt patterns foster this scaffolding by engaging learners in explanation and reflection (White et al., 2023). In programming education, such structured guidance is vital for helping novices transition from trial-and-error to systematic hypothesis generation, which facilitates the establishment of stable mental models (Ihantola, et al., 2025). Furthermore, prompt engineering—conversing with AI to articulate computational steps—serves as a valuable pedagogical activity that aids in error resolution and problem decomposition (Denny et al., 2023). Strategies like the "Alternative Approaches" pattern further promote knowledge transfer by enabling students to clarify logical relationships through solution comparison (White et al., 2023). Consequently, AI teaching assistants function across five instructional dimensions: guiding logical understanding, clarifying syntax to reduce cognitive load, offering alternative examples, encouraging self-explanation, and supporting systematic debugging. This reflects the evolution of generative AI from a mere answer provider into a scaffolding tool that supports the learning process.

3. Research Method

3.1 Research Procedure

This study adopted a single-group pretest-posttest quasi-experimental design with 32 tenth-grade students from a public senior high school in Hsinchu enrolled in a Python Programming elective course. The twenty-week study was divided into two stages. During weeks 1-10, traditional instruction through lectures and practical exercises established students' basic programming syntax and logical abilities. At week 10, students were divided into high-achievement (H) and low-achievement (L) groups based on median test scores. During weeks 11-20, a pretest was administered, followed by introduction of a personalized LLM-based AI teaching assistant on Discord. Students from each achievement level were randomly assigned to either 'generic prompt' or 'guided prompt' conditions, forming four groups of 8 students each: high-achievement generic (H1), high-achievement guided (H2), low-achievement generic (L1), and low-achievement guided (L2). This design examined the interactive effect of prior ability and AI guidance strategies (direct answers vs. scaffolded guidance) on learning achievement.

Data collection included pre- and post-intervention programming achievement test, alongside scales measuring programming achievement tests and scales assessing programming self-efficacy, learning attitudes, learning motivation, anxiety, and AI acceptance. Finally, quantitative data were integrated with qualitative data from semi-structured interviews to analyze changes in students' cognitive performance, problem-solving strategies, and learning attitudes.

3.2 Learning Achievement Test.

The learning achievement test initially consisted of 25 items (15 basic, 10 advanced) covered core concepts of Python programming, including variables and operations, flow control, data structures, functions, string processing, and debugging, to assess students' understanding and application abilities in programming knowledge. The researchers conducted item analysis based on pilot test results, calculating each item's difficulty index (p value) and discrimination index (D value), deleting or revising items with insufficient discrimination ($D < 0.30$), and adjusting item stems and options. After revision, the formal achievement test retained 24 items with a total score of 12 points, a time limit of 60 minutes, and was used as the learning achievement test for Stage 1 and Stage 2 according to the course schedule to compare students' learning performance before and after AI teaching assistant intervention.

3.3 Attitude Scales

This study employed three revised scales, all using a five-point Likert scale (1 = strongly disagree, 5 = strongly agree), to collect psychological and attitudinal changes during students' programming learning process, totaling 27 items revised to minimize respondent burden. For attitude and psychological scales, this study employed three five-point Likert scales that have undergone reliability and validity testing for assessment:

1. Programming Self-Efficacy Scale: Adapted from Ramalingam and Wiedenbeck (1998), consisting of 10 items measuring dimensions including comprehension, logical control, and problem-solving, with reliability Cronbach's α ranging from .92 to .98.

2. Programming Attitude Scale: Adapted from Wiebe et al. (2003), consisting of 8 items measuring learning attitude, perceived value, and confidence toward programming, with reliability α approximately .91.

3. AI Education Acceptance Scale: Adapted from Davis (1989), consisting of 9 items covering perceived ease of use, perceived usefulness, and behavioral intention, with reliability α ranging from .87 to .94.

3.4 AI Teaching Assistant System and Prompt Design

3.4.1 Generic and Guided Prompt design

This study developed an AI teaching assistant system based on the Discord platform and Large Language Models (LLM), designing two sets of prompt templates according to Cognitive Load Theory (Sweller, 2010).

- Generic Prompt: Adopts task-oriented design, emphasizing conciseness and technical accuracy in responses, directly providing answers or examples to student questions without proactively extending guidance.
- Guided Prompt: Role-set as a guide, restricting direct provision of answers. The system provides five assistance functions based on scaffolding theory, guiding students to recognize needs and construct problem-solving approaches. Based on relevant literature, this study summarizes the five major functions and theoretical foundations of guided prompts as follows:

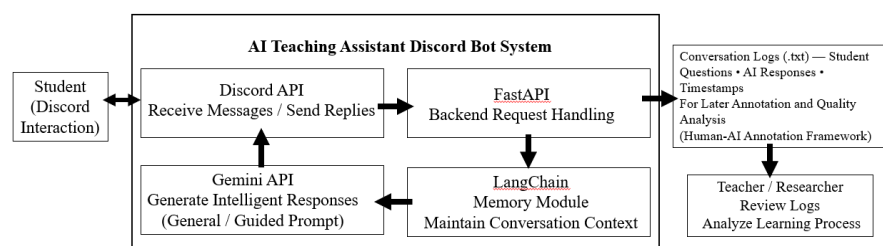
1. Logical Explanation: Step-by-step explanation of execution flow, helping to establish mental models and reduce working memory load (Sweller, 2010; Robins et al., 2003).
2. Syntax Explanation: Real-time indication of errors and provision of rule explanations, supporting individualized debugging processes (Ihantola et al., 2010; Keuning et al., 2018).
3. Alternative Approach: Providing different implementation methods to compare advantages and disadvantages, promoting style judgment and knowledge transfer (Guzdial, 2015; Kölling et al., 2003).
4. Add Comments: Requiring explanation of code in natural language, promoting externalization of thinking and strengthening metacognition (Schraw & Moshman, 1995; Tenenbergs & Murphy, 2005).
5. Code Review: Conducting systematic error diagnosis and revision suggestions, improving debugging efficiency (Alamutka, 2005; Keuning et al., 2018).

3.4.2 AI Teaching Assistant Discord Bot System Architecture

We designed the AI Teaching Assistant Discord Bot System provides personalized programming education. Students interact via Discord platform, where messages are processed through FastAPI backend services. LangChain maintains conversation context and memory across interactions, enabling coherent dialogue continuity. The Gemini API generates pedagogically-appropriate responses using customized prompts—either general prompt or guided prompt for students requiring scaffolded support. All interactions are automatically logged with timestamps, creating comprehensive datasets for educational research. Teachers and researchers analyze these conversation logs through a Human-AI Annotation Framework, identifying learning patterns and refining instructional strategies.

Figure 1

AI Teaching Assistant Discord Bot System Architecture



4. Results Analysis

This section presents the quantitative analysis results of two intervention modes—'generic prompt' or 'guided prompt'—on students' learning achievement, programming self-efficacy, learning attitudes, and AI acceptance scales, and explains the findings of this study.

4.1 Learning Achievement Performance

4.1.1 Wilcoxon Signed-Rank Test

To examine the impact of AI teaching assistant intervention on overall student programming ability, this study employed a paired-sample Wilcoxon signed-rank test to analyze differences in students' pretest and posttest scores. Statistical results showed that all students ($N = 32$) had significantly higher average scores on the posttest ($M = 8.28$, $SD = 2.36$) than pretest scores ($M = 6.12$, $SD = 1.96$), with test results reaching a highly significant level ($Z = 5.95$, $p < .001$). This result confirms that regardless of students' initial ability, introducing an LLM-based AI teaching assistant system as classroom and after-class support can effectively improve their programming problem-solving performance.

Table 1.

Paired-sample t-test results for overall students' pretest and posttest.

Test Stage	<i>n</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>	<i>p</i>
Pretest	32	6.12	1.96	5.95	31	< .001*
Posttest	32	8.28	2.36			

Note. * $p < .05$ indicates significant difference.

4.1.2 Four-Group Subdivision Analysis

Subgroup analysis revealed that while all groups exhibited upward trends, only the low-achievement guided group (L2) achieved statistically significant improvement ($p = .0078$). The results indicate that while LLM-based AI assistants effectively enhance overall programming achievement ($p < .001$), guided prompts provide critical scaffolding that is particularly effective for improving learning achievement of low-achievement students.

Table 2.

Four-group AI intervention pre-post difference analysis (Wilcoxon signed-rank test)

Group	<i>N</i>	Pre Mean	Post Mean	Wilcoxon (T, p)	Mean Diff.	Significance
H1	8	7.63	8.50	$T = 5.5$, $p = .1719$	+0.875	ns
H2	8	5.38	7.88	$T = 0.0$, $p = .0625$	+2.500	ns
L1	8	6.13	7.25	$T = 3.0$, $p = .1562$	+1.125	ns
L2	8	5.38	9.50	$T = 0.0$, $p = .0078$	+4.125	Sig.

Note. ns = not significant, Sig. = significant.

4.2 Scale Analysis

4.2.1 Descriptive Statistical Analysis

Descriptive analysis revealed that high-achievement groups (H1, H2) possessed higher programming self-efficacy and learning attitudes, reflecting stable confidence and engagement. In contrast, low-achievement groups (notably L2) exhibited higher AI acceptance, indicating greater trust and willingness to utilize the AI teaching assistant.

Table 3.

Descriptive statistics of four groups on three scales (percentage standardized).

Group	Self-Efficacy (CPSES)	Attitude (ATT)	AI Acceptance (AIA)
H1	67.50% $\pm$ 18.91	82.19% $\pm$ 15.03	73.33% $\pm$ 6.29

H2	75.75% ± 15.54	75.94% ± 16.58	86.11% ± 15.70
L1	59.25% ± 24.12	69.38% ± 27.96	76.67% ± 24.37
L2	71.75% ± 6.18	68.12% ± 14.38	89.17% ± 11.66

4.2.2 Prompt Group Comparison

Students were further divided into generic prompt group (H1+L1) and guided prompt group (H2+L2) according to prompt type to compare differences in performance on the three scales. Analysis results showed that the guided prompt group ($U = 75.5$, $p = .0495$) was significantly higher than the generic prompt group on both programming self-efficacy and AI acceptance scales, indicating that structured guidance prompt design helps students build learning confidence and enhances their trust and willingness to use the AI teaching assistant. While no significant difference was found in learning attitudes, the guided group exhibited a more stable performance trend.

Table 4.

Comparison of generic and guided groups on three scales (Mann-Whitney U test).

Scale	Generic M±SD	Guided M±SD	Generic Mdn	Guided Mdn	<i>U</i>	<i>p</i>	Sig.
Self-Efficacy	31.69±10.68	36.88±5.81	30.0	35.0	75.5	.0495	Sig.
Attitude	30.31±9.07	28.81±6.21	32.0	29.5	158.0	.2648	ns
AI Acceptance	33.75±7.78	39.44±6.05	33.0	41.5	73.0	.0386	Sig.

Note. ns = not significant, Sig. = significant.

4.3 Dialogue Record Analysis

4.3.1 Scoring Criteria.

This study combined the ICAP Framework (Chi & Wylie, 2014) and Bloom's Taxonomy (Bloom, 1956) to construct an 'Inquiry Depth' analysis framework, classifying student questions into six levels according to their cognitive level and learning intention as the scoring basis for dialogue analysis. This framework sequentially reflects students' inquiry behaviors from non-learning-oriented interaction and surface-level questioning to behaviors demonstrating explanation, reflection, and knowledge transfer capabilities from low to high, effectively presenting students' level of cognitive engagement throughout the programming learning process.

Table 5.

Inquiry depth scoring rubric.

Level	Name	Description
IRR	Irrelevant Message	Pure greetings, thanks, links, system notifications, etc., unrelated to learning inquiry
D1	Surface/Answer-Seeking	Only asks for results, lacks context and attempts
D2	Clarification/Location	Can describe errors or context, but has not yet attempted
D3	Trial and Revision	Specifically describes modifications or experimental process, operational level
D4	Explanation and Hypothesis	Proposes causal speculation or principle explanation, and plans verification
D5	Reflection and Transfer	Summarizes experience, generalizes strategies, or transfers to new problems

4.3.2 Human-Machine Annotation Process and Results

This study collected a total of 2,904 dialogue messages between students and the AI teaching assistant, an LLM performed automatic annotation based on the inquiry depth criteria. To verify the reliability, researchers randomly selected approximately one-tenth of the sample ($N = 290$) for manual annotation by two human raters. The three-way annotation results (teacher, researcher, and LLM) were tested for consistency using Cohen's κ , with results showing κ values for each

comparison combination ranging between 0.65 and 0.73, reaching 'substantial agreement' level. This result indicates that LLM can stably simulate human judgment in inquiry depth annotation tasks, demonstrating reliability as an auxiliary analysis tool for educational research.

Table 6.

Three-way annotation results.

Comparison Group	κ Value	95% CI	Interpretation
Teacher vs Researcher	0.651	[0.580, 0.718]	Substantial
Teacher vs LLM	0.734	[0.670, 0.793]	Substantial
Researcher vs LLM	0.694	[0.673, 0.714]	Substantial

4.3.3 Interaction Level Analysis.

Using 1,308 "question-answer interaction units" (Q-A pairs) as the primary analysis unit, this study utilized the highest inquiry level per pair as the representative score. A total of 1,308 interaction units were analyzed, with inquiry depth distribution statistics compiled according to student groupings (H1, H2, L1, L2). Results showed that student groups using guided prompts (H2, L2) had higher proportions of high-level inquiry (D3-D5) than generic prompt groups (H1, L1), while their proportion of irrelevant messages (IRR) was significantly reduced, indicating that interaction content was more focused on learning tasks.

Table 7.

Interaction sample distribution.

Group	Total N	IRR	D1	D2	D3	D4	D5	High-Level (D3+D4)
H1	305	12.7%	29.1%	31.1%	13.9%	12.4%	0.7%	13.2%
H2	344	7.6%	34.1%	27.0%	13.7%	17.1%	0.5%	17.5%
L1	333	12.0%	27.9%	32.1%	15.1%	12.8%	0.0%	12.8%
L2	326	8.0%	21.1%	31.6%	25.1%	13.9%	0.4%	14.3%

4.3.4 Result Interpretation

Further comparison of inquiry depth distribution among different student groups yields the following trends:

- (1) High-achievement students under guided prompt support (H2) showed significantly higher proportions of high-level inquiry compared to those not using guided prompts (H1), indicating that structured prompts help promote their reasoning and reflection.
- (2) Low-achievement students under guided prompts (L2) also showed higher proportions of operational and explanatory level interactions, indicating that guided prompts help them progress from surface-level questioning to the experimentation and revision stage.

Overall, prompt design can produce positive effects on students of different achievement levels, with particularly significant effects on deepening inquiry for high-achievement students.

5. Conclusion and Discussion

5.1 Impact of AI Teaching Assistant on Learning Achievement

Research results show that LLM-based AI assistant students overall showed significant improvement in Python programming learning achievement, including program comprehension, error correction, and problem decomposition. This result indicates that the immediate feedback and sequential prompts provided by the AI teaching assistant can function as scaffolding in the learning process, helping students correct errors in real-time and deepen conceptual

understanding. Notably, low-achievement students showed greater improvement after AI teaching assistant intervention, indicating that AI teaching assistants have significant potential in remedial learning and reducing learning gaps, highlighting AI's potential for narrowing learning gaps, while high-achievement students utilized the tool for advanced applications.

5.2 Impact of AI Teaching Assistant on Students' Self-Efficacy, Learning Attitudes, and AI Acceptance

The results show that after introducing the AI teaching assistant, students' programming self-efficacy, learning attitudes, and AI acceptance all showed overall positive changes. The "evaluation-free" nature of the AI assistant fostered a **psychologically safe space**, reducing anxiety and encouraging iterative trial-and-error and self-correction.

5.3 Student-AI Teaching Assistant Interaction Inquiry Depth and Teacher Observations

In dialogue process analysis, this study revealed that inquiry depth is driven by prompt design; guided prompts effectively transitioned low-achievement students from answer-seeking to experimentation (D3) and promoted higher-level reflection (D4-D5) in high-achievement peers. Teacher observations also indicated that the AI teaching assistant can effectively share immediate guidance needs in class, allowing teachers to devote more energy to high-level instructional guidance and promoting the development of students' self-directed learning.

References

- Ala-Mutka, K. M. (2005). A survey of automated assessment approaches for programming assignments. *Computer Science Education*, 15(2), 83–102. <https://doi.org/10.1080/08993400500150747>
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall, Inc.
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. David McKay Company.
- Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219–243. <https://doi.org/10.1080/00461520.2014.965823>
- Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>
- Denny, P., Kumar, V., & Giacaman, N. (2023). Conversing with Copilot: Exploring prompt engineering for solving CS1 problems using natural language. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1* (pp. 1136–1142). Association for Computing Machinery. <https://doi.org/10.1145/3545945.3569823>
- Guzdial, M. (2015). *Learner-centered design of computing education: Research on computing for everyone*. Morgan & Claypool Publishers. <https://doi.org/10.2200/S00684ED1V01Y201511HCI033>
- Ihantola, P., Ahoniemi, T., Karavirta, V., & Seppälä, O. (2010). Review of recent systems for automatic assessment of programming assignments. In *Proceedings of the 10th Koli Calling International Conference on Computing Education Research* (pp. 86–93). <https://doi.org/10.1145/1930464.1930480>
- Kaczmarczyk, L. C., Petrick, E. R., East, J. P., & Herman, G. L. (2010). Identifying student misconceptions of programming. In *Proceedings of the 41st ACM Technical Symposium on Computer Science Education (SIGCSE '10)*, 107–111. <https://doi.org/10.1145/1734263.1734299>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Keuning, H., Jeuring, J., & Heeren, B. (2018). A systematic literature review of automated feedback generation for programming exercises. *ACM Transactions on Computing Education*, 19(1), Article 3. <https://doi.org/10.1145/3231711>
- Kölling, M., Quig, B., Patterson, A., & Rosenberg, J. (2003). The BlueJ system and its pedagogy. *Computer Science Education*, 13(4), 249–268. <https://doi.org/10.1076/csed.13.4.249.17496>

- Lishinski, A. (2016). Cognitive, affective, and dispositional components of learning programming. In Proceedings of the 2016 ACM Conference on International Computing Education Research (pp. 261–262). Association for Computing Machinery. <https://doi.org/10.1145/2960310.2960338>
- Padurean, V.-A., Denny, P., & Singla, A. (2025). BugSpotter: Automated generation of code debugging exercises. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE TS 2025)*. Association for Computing Machinery. <https://doi.org/10.1145/3641554.3701974>
- Qian, Y., & Lehman, J. D. (2017). Students' misconceptions and other difficulties in introductory programming: A literature review. *ACM Transactions on Computing Education (TOCE)*, 18(1), 1-24. <https://doi.org/10.1145/3077618>
- V., LaBelle, D., & Wiedenbeck, S. (2004). Self-efficacy and mental models in learning to program. In Proceedings of the 9th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (pp. 171–175). Association for Computing Machinery. <https://doi.org/10.1145/1007996.1008042>
- Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning Sciences*, 13(3), 273–304. https://doi.org/10.1207/s15327809jls1303_2
- Ramalingam, V., & Wiedenbeck, S. (1998). Development and Validation of Scores on a Computer Programming Self-Efficacy Scale and Group Analyses of Novice Programmer Self-Efficacy. *Journal of Educational Computing Research*, 19(4), 367-381. <https://doi.org/10.2190/C670-Y3C8-LTJ1-CT3P> (Original work published 1998)
- Robins, A., Rountree, J., & Rountree, N. (2003). Learning and teaching programming: A review and discussion. *Computer Science Education*, 13(2), 137–172. <https://doi.org/10.1076/csed.13.2.137.14200>
- Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review*, 7(4), 351-371. <https://doi.org/10.1007/BF02212307>
- Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, 22(2), 123–138. <https://doi.org/10.1007/s10648-010-9128-5>
- Tenenberg, J., & Murphy, L. (2005). Knowing what I know: An investigation of undergraduate knowledge and self-knowledge of data structures. *Computer Science Education*, 15(4), 297–315. <https://doi.org/10.1080/08993400500307677>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Wiebe, E. N., Williams, L., Yang, K., & Miller, C. (2003). Computer Science Attitude Survey (Report No. NCSU CSC TR-2003-1). Department of Computer Science, North Carolina State University.
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*. <https://doi.org/10.48550/arXiv.2302.11382>

Impact of AI Support with Different Level of Learner Agency on Behavioral Engagement in Providing and Uptaking Peer Feedback

Xinmeng Hou¹, Wenli Chen^{1*}, Lishan Zheng¹, Guo Su¹, Xuanyu Chen², Qianru Lyu³

¹ National Institute of Education, Nanyang Technological University, Singapore

² Institute of Higher Education, Fudan University, China

³ Human-Computer Interaction Institute, Carnegie Mellon University, USA

* wenli.chen@nie.edu.sg

Abstract: This study examined how AI support with varying levels of learner agency influences learners' behavioral engagement during peer feedback construction and uptake. Ninety university students in 30 triads were randomly assigned to three conditions: Control (no AI support), AI-directed (AI which provides suggested revision on learning artefacts), or AI-assisted (AI provides guidance for learners to improve learning artefacts). A conversation analysis framework was proposed for analyzing learners' behavioral engagement in human-AI dialogue through four dimensions: elaboration level, topic pursuit, difficulty response, and repair pursuit intensity. Analysis of the two AI conditions ($n = 60$) revealed task-dependent patterns: participants showed greater response depth and topic extension during feedback construction, but more clarification-seeking during feedback uptake. Significant interaction effects emerged between AI support type and learning task. AI-assisted learners displayed adaptive engagement, while AI-directed learners maintained consistent, undifferentiated engagement. These findings suggest that AI in an assistant role preserves learner autonomy and agency, and fosters task-adaptive engagement, while AI in a directive role may produce engagement patterns consistent with automation complacency.

Keywords: AI-supported collaborative learning, peer feedback, behavioral engagement, conversation analysis, learner agency

1. Introduction

Feedback has long been considered a key factor influencing learning and achievement (Black & Wiliam, 1998; Hattie & Timperley, 2007). Effective feedback helps learners understand goals, compare their current performance to standards, and take action to close the gap (Sadler, 1989). Peer feedback has emerged as a scalable approach to delivering formative assessment, with reviews indicating adequate reliability and validity (Li et al., 2020; Topping, 1998). The concept of feedback literacy further emphasizes learners' active role in making sense of and acting upon feedback information (Carless & Boud, 2018). Beyond assessment accuracy, peer feedback may help develop learners' metacognitive skills and evaluative judgment (Nicol & Macfarlane-Dick, 2006). However, learners often struggle to provide constructive comments without adequate support (Yu & Lee, 2016), prompting exploration of scaffolding approaches (Min, 2005; Gielen & De Wever, 2015).

Generative AI has introduced new possibilities for feedback in education. Research distinguishes between AI-generated feedback and AI-supported peer feedback (Guo, 2024; Steiss et al., 2024), with hybrid human-AI models preserving cognitive benefits while addressing quality concerns (Banihashem et al., 2024; Kaliisa et al., 2025). However, providing and uptaking feedback involves fundamentally different learning mechanisms—feedback providers engage in evaluative judgment, while feedback receivers focus on interpretation and implementation (Nicol et al., 2014; Wu & Schunn, 2021)—yet few studies have examined how engagement patterns differ across these activities. Additionally, there is limited research conceptualizing AI's role by considering how much agency learners retain. AI can function as a directive agent or collaborative partner preserving learner agency (Darvishi et al., 2024), and this distinction

matters: human-centered AI fosters deeper learners' self-regulation and critical thinking, whereas directive approaches risk learners' passive dependency (Sperber et al., 2025; Zhan & Yan, 2025).

Despite growing interest, there is hardly any study examining learners' micro-level behavioral engagement across different feedback tasks—specifically, whether observable effort in dialogue differs between feedback construction and feedback uptake. In addition, no study has empirically compared how directive versus assistive AI support shapes moment-to-moment engagement patterns across these tasks. The present study addresses these gaps by: (1) proposing a novel conversation analysis (CA) framework for operationalizing learners' behavioral engagement in human-AI dialogue; and (2) providing empirical evidence of how the interaction between AI support type and peer feedback task type shapes learners' engagement patterns in human-AI dialogue.

2. Literature Review

Learning engagement is a multidimensional construct comprising behavioral, emotional, and cognitive components (Fredricks et al., 2004). Among these, behavioral engagement—defined as the effort, persistence, and attention students direct toward learning tasks (Skinner & Belmont, 1993; Finn & Zimmer, 2012)—has been identified as the strongest predictor of academic achievement (Finn & Zimmer, 2012). Linnenbrink and Pintrich (2003) further characterized behavioral engagement as encompassing effort, persistence, and instrumental help-seeking. Within feedback research specifically, Ellis (2010) conceptualized behavioral engagement as how learners respond to and act upon feedback, while Han and Hyland (2015) operationalized it through revision operations and observable strategies. This effort-centered view provides the foundation for examining how learners invest themselves in feedback-related tasks.

Critically, the demands placed on learners differ depending on which feedback activity they perform. Research on peer feedback has established that providing and receiving feedback involve fundamentally different learning mechanisms: feedback providers engage in evaluative judgment, problem detection, and criteria application, while feedback receivers focus on interpretation and implementation (Nicol et al., 2014; Patchan & Schunn, 2015). Wu and Schunn (2021) found that learning correlated more strongly with providing feedback than receiving it, suggesting that feedback construction activates deeper cognitive processing. This asymmetry likely stems from the evaluative demands of feedback provision, which require learners to articulate standards, identify discrepancies, and generate actionable suggestions (Lundstrom & Baker, 2009). Conversely, feedback uptake requires learners to interpret received comments, reconcile them with their own understanding, and implement revisions, which place different cognitive demands (Nelson & Schunn, 2009; Gielen et al., 2010; Chen et al., 2023). If these tasks impose distinct cognitive demands, they should also elicit qualitatively different patterns of behavioral engagement; yet this possibility has not been empirically examined.

The emergence of generative AI (Gen AI) has introduced new possibilities for scaffolding these feedback activities. Research distinguishes between AI-generated feedback and AI-supported peer feedback (Guo, 2024; Steiss et al., 2024), with hybrid human-AI models showing promise for preserving cognitive benefits while addressing quality concerns (Banihashem et al., 2024; Kaliisa et al., 2025). However, a key dimension remains underexplored: how much agency learners retain when interacting with AI. Darvishi et al. (2024) showed that AI can function as a directive agent delivering finished products or as a collaborative partner preserving learner agency. This distinction is consequential: human-centered AI that prioritizes learner agency may foster deeper self-regulation and sustained engagement, whereas directive approaches risk promoting passive dependency and automation complacency (Parasuraman & Manzey, 2010; Sperber et al., 2025; Zhan & Yan, 2025).

Despite these advances, two gaps persist. First, prior studies have typically treated feedback-related tasks as a single activity, without distinguishing between feedback construction and uptake—two activities that, as reviewed above, place distinct demands on learners and should therefore elicit different engagement patterns. Second, existing research conceptualizes AI involvement in binary terms (present or absent), without examining how different AI roles—

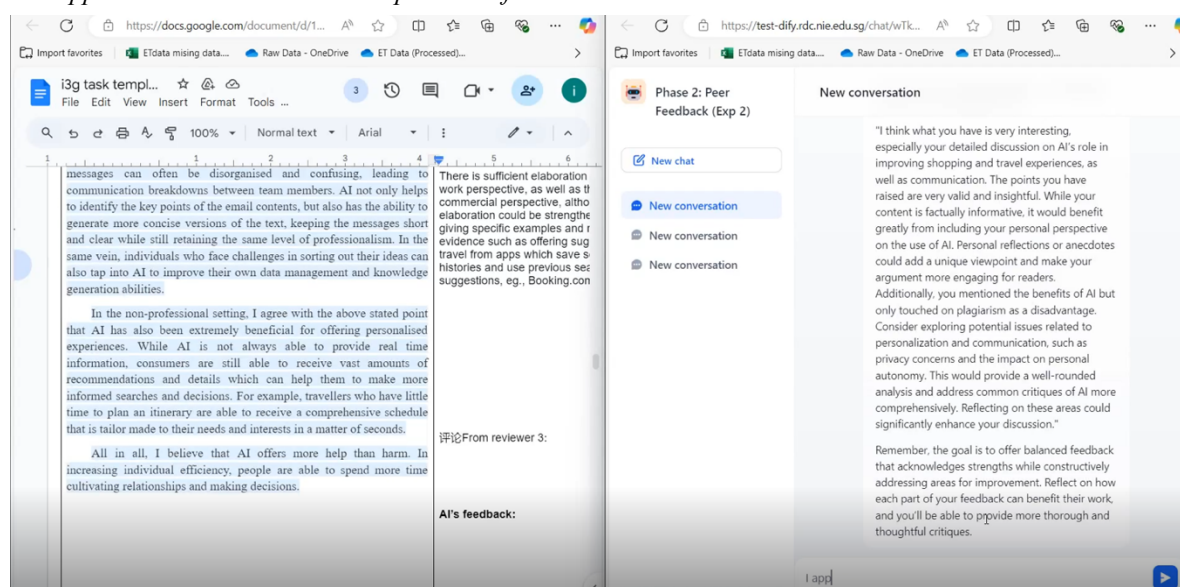
specifically, the degree of agency afforded to learners—shape the micro-level behavioral effort learners invest in human-AI dialogue across these distinct tasks. Building on this foundation, the present study addresses the following research questions: **RQ1.** How do learners' behavioral engagement patterns differ between feedback construction and feedback uptake when interacting with AI support? **RQ2.** Controlling for Peer feedback activity type, how does AI support type (directive vs. assisted) affect learners' behavioral engagement?

3. Methodology

Experiment design was adopted in this study to example the impact of AI support with different level of learner agency on learners' behavioral engagement in providing and uptaking peer feedback. The study employed a between-subjects randomized controlled design with three conditions—Control, AI-directed, and AI-assisted. Ninety students from a university in Singapore were organized into 30 triads (10 triads per condition). Participants completed a 30-minute protocol using a custom workspace integrating Google Docs with an embedded GPT-4o chatbot (see Figure 1). The control condition was included to establish a baseline for feedback quality; however, because control participants did not interact with the chatbot, the behavioral engagement analysis was restricted to the two AI supported peerr feedback conditions ($n = 60$). All claims should be interpreted as comparisons between AI support modalities, not AI-supported versus unsupported conditions.

Figure 1

AI-Supported Peer Feedback Workspace Interface



During feedback providing (15 minutes), AI-directed participants received AI-suggested revisions on their learning artefacts to them to adopt or adapt; AI-assisted participants received guidance, and questions which scaffold them to iteratively refined the learning ratefacts. During feedback uptaking (15 minutes), participants revised essays based on received feedback. Table 1 summarizes prompt design differences.

Table 1

Comparison of AI prompt design across conditions and phases.

Phase	Condition	What AI Does	What Learner Does
Feedback	AI-Directed	Directly revise student's peer feedback for clarity and rubric alignment; marks changes in bold	Adapt AI-revised feedback for peers
	AI-Assisted	Offers suggestions, examples, and reflective questions to improve feedback	Implement all revisions independently
Feedforward	AI-Directed	Directly revises essay based on peer feedback; marks	Review and accept AI-

	changes in bold; no new evidence added	made revisions
AI-Assisted	Provides guidance on improving essay; includes examples and reflective questions	Make all revisions independently

To measure behavioral engagement, the study used conversation analysis (CA) methods for examining how participants demonstrate investment through observable discourse features (Stivers & Rossano, 2010; Schegloff, 2007). "Behavioral engagement" here refers to the effort, persistence, and active participation learners demonstrate in human-AI dialogue during learning tasks (Skinner & Belmont, 1993), operationalized through four CA-derived indicators, each mapping onto an established engagement dimension. Elaboration level captures effort and concentration: learners who produce extended turns with examples, reasoning, or synthesis demonstrate greater investment than those producing minimal acknowledgment tokens (Schegloff, 1982; Skinner & Belmont, 1993). Topic pursuit captures persistence: actively developing and extending conversational topics reflects sustained behavioral involvement rather than passive topic acceptance (Button & Casey, 1985; Fredricks et al., 2004; Connell & Wellborn, 1991). Difficulty response operationalizes orientation toward challenge, distinguishing learners who engage with difficult content from those who avoid it—a core distinction in engagement research (Schegloff et al., 1977; Skinner et al., 2009). Repair pursuit intensity captures instrumental help-seeking: the extent to which learners actively seek clarification and verification reflects purposeful engagement with understanding rather than passive acceptance of AI-provided content (Schegloff, 2007; Linnenbrink & Pintrich, 2003).

Collectively, higher levels across these indicators may also serve as behavioral proxies for learner agency—the capacity to intentionally influence one's own learning through proactive, self-regulated action (Reeve & Tseng, 2011). However, we note that our primary construct is behavioral engagement; inferences about agency remain secondary and should be interpreted with appropriate caution, as these indicators do not directly measure intentionality or self-regulation. This distinction is important given concerns that AI assistance may promote dependency rather than independent skill development (Darvishi et al., 2024). Each indicator was coded on a four-point ordinal scale (0–3), drawing on CA principles for identifying interactional phenomena (Sidnell & Stivers, 2013) while adapting them for systematic quantitative analysis. This approach follows precedents in applied linguistics and interaction research that have operationalized CA-derived constructs for larger-scale comparative studies (Stivers, 2015; Stivers & Rossi, 2025). The unit of analysis was the individual learner turn, defined as each discrete message a learner sent to the AI chatbot during the interaction. A conversation comprised the full dialogue sequence between one learner and the chatbot within a single learning task. This yielded 363 coded turns across 120 conversations (60 learners under AI-supported conditions $\times$ 2 tasks). Two trained coders independently coded all learner turns, with inter-rater reliability assessed using Cohen's kappa (see Table 2). The coding schema demonstrated strong agreement across indicators, with elaboration level achieving near-perfect reliability ($\kappa = .95$) and the remaining indicators showing substantial agreement (κ range = .80–.66), consistent with benchmarks for conversational data analysis (Landis & Koch, 1977).

Table 2

Inter-rater reliability for behavioral engagement indicators.

Indicator	Description	Scale	κ
Elaboration Level	Degree of substantive depth in learner contributions	0–3	.95
Topic Pursuit	Degree of active topic development and extension	0–3	.80
Difficulty Response	Degree of orientation toward task complexity	0–3	.77
Repair Pursuit Intensity	Degree of active verification and understanding seeking	0–3	.66

Note. κ = Cohen's kappa calculated between two coders. Adjacent agreement reflects the percentage of annotations within one scale point.

4. Findings

This study conducted nonparametric analyses using Mann-Whitney U tests for pairwise comparisons and Kruskal-Wallis H tests for interaction effects across 120 conversations comprising 363 turns. Effect sizes are reported using rank-biserial correlation (r), with .10, .30, and .50 as small, medium, and large effects (Cohen, 1988).

As shown in Table 3, three of four indicators showed significant task differences. Elaboration level and topic pursuit were both significantly higher during feedback construction, meaning students provided more detailed responses and more actively extended conversation topics when giving peer feedback. The task difference in elaboration level represented a small-to-medium effect ($r = .20$). Difficulty response showed no significant difference, possibly because the 30-minute protocol with a relatively accessible essay topic did not elicit sufficient variation in challenge orientation, or because both AI conditions provided enough scaffolding to compress the range of difficulty response behaviors. However, the pattern reversed for repair pursuit intensity, which was significantly higher during feedback uptake (medium effect, $r = .35$), indicating that students asked more clarifying questions and sought more verification when revising their own essays—the most practically meaningful task difference observed.

Table 3

Behavioral engagement indicators across experimental phases.

Indicator	Feedback providing	Feedback Uptaking	Difference	U	p	r
Elaboration Level	2.02	1.62	0.40	19467.00	.010**	.20
Topic Pursuit	2.14	1.96	0.18	18984.00	.008**	.18
Difficulty Response	1.36	1.22	0.13	17422.31	.100	.11
Repair Pursuit Intensity	0.31	0.64	−0.33	13465.00	<.001***	.35

Note. Diff = Phase 2 – Phase 3. Effect size r = rank-biserial correlation. * $p < .01$. *** $p < .001$.

Kruskal-Wallis H tests revealed significant interaction effects for three of four indicators (Table 4). For elaboration level and topic pursuit, AI-assisted students showed substantial declines from construction to uptake, while AI-directed students remained stable. For repair pursuit intensity, AI-assisted students dramatically increased clarification-seeking during uptake—nearly a fourfold increase—while AI-directed students showed minimal change. Difficulty response showed no significant interaction.

Table 4

Interaction effects between phase and AI support condition.

Indicator	providing + Assisted	providing + Directed	Uptake + Assisted	Uptake + Directed	H	p
Elaboration Level	2.35	1.74	1.51	1.73	25.82	<.001***
Topic Pursuit	2.37	1.94	1.89	2.03	18.47	<.001***
Difficulty Response	1.51	1.23	1.21	1.24	5.37	.147
Repair Pursuit Intensity	0.21	0.40	0.77	0.50	21.94	<.001***

Note. H = Kruskal-Wallis test statistic with $df = 3$. *** $p < .001$. Construction = feedback construction phase; Uptake = feedback uptake phase; Assisted = AI-assisted condition; Directed = AI-directed condition

These interaction effects reveal that the two AI support conditions shaped behavioral engagement differently across tasks. In the AI-assisted condition, the scaffolded approach—providing guidance and reflective questions rather than finished products, elicited task-dependent behaviors: learners produced more elaborate contributions during peer feedback but shifted toward clarification-seeking during essay revision, actively verifying their understanding when responsible for implementing changes themselves. In contrast, the AI-directed condition showed stable engagement

across tasks, likely because students had less need to elaborate or seek clarification when the AI had already completed the substantive revision work. This pattern suggests that the degree of learner agency preserved by the AI support design fundamentally shapes how learners distribute their effort across different feedback activities.

5. Discussion and Conclusion

By adapting conversation analysis to operationalize engagement in human-AI dialogue, this study investigated how different AI support levels influence behavioral engagement during peer feedback activities.

Task-dependent engagement. The differential patterns suggest feedback construction and uptake elicit qualitatively different learner investment. During construction, higher elaboration and topic development reflect the evaluative demands of providing feedback (Nicol et al., 2014; Lundstrom & Baker, 2009). During feedback uptake, higher clarification-seeking reflects interpretive demands of reconciling received feedback with one's own understanding (Nelson & Schunn, 2009). These findings provide micro-level empirical support for the theoretical distinction between feedback provision and uptake demands (Wu & Schunn, 2021).

AI role and adaptive engagement. The significant interaction effects reveal that the two AI conditions shaped engagement in qualitatively different ways, a pattern interpretable through self-determination theory (SDT; Ryan & Deci, 2000). The AI-assisted condition—which provided guidance and reflective questions rather than finished products—supported learner autonomy by preserving space for independent decision-making. This autonomy support manifested as adaptive engagement: learners invested in elaboration during feedback construction (when the task demanded evaluative generation) but shifted toward verification during feedback uptake (when the task demanded interpretive implementation). Such task-responsive flexibility suggests that scaffolded AI support enabled learners to calibrate their effort to the specific demands of each activity, consistent with SDT's prediction that autonomy support fosters more self-regulated and intrinsically motivated behavior. In contrast, the AI-directed condition produced stable, undifferentiated engagement across tasks, aligning with the automation complacency framework (Parasuraman & Manzey, 2010): when AI provides finished products that require only review and acceptance, learners may reduce their active cognitive processing regardless of task demands. The consistency of engagement in the AI-directed condition (neither increasing during demanding tasks nor decreasing during simpler ones) suggests that the directive AI may have attenuated the natural variation in effort that different feedback activities should elicit. While consistent engagement might appear desirable on the surface, it may reflect a ceiling effect on passivity rather than genuine stability of effort.

Design implications. AI should be positioned as a dialogic partner that scaffolds learners' thinking through questions and suggestions rather than delivering ready-made solutions. Support should be task-sensitive: during feedback construction, AI should prompt evaluative reasoning and criteria application; during feedback uptake, AI should support interpretation and encourage learners to verify their understanding before implementing changes. Third, designers should monitor for engagement uniformity as a potential warning sign of automation complacency, building in mechanisms that require active learner processing, such as requiring learners to articulate their reasoning before receiving AI suggestions.

Limitations. Several limitations warrant consideration. First, the 30-minute protocol captures behavioral engagement during an initial exposure to AI-supported feedback. Because engagement patterns can evolve as novelty effects diminish and learners develop stable interaction strategies with AI tools, the adaptive engagement patterns observed in the AI-assisted condition may not persist over longer interventions. Longitudinal designs are needed to examine the stability of these patterns. Second, our behavioral engagement measures capture only one dimension of the multidimensional engagement construct; future research should integrate emotional and cognitive engagement measures to provide a more complete picture. Third, the behavioral analysis was restricted to the two AI conditions because the control group did not interact with the chatbot; future studies should incorporate cross-condition measures (e.g., feedback quality, writing improvement) to examine whether the engagement differences observed here translate

into differential learning outcomes. Finally, the single-university sample from Singapore limits generalizability across educational and cultural contexts.

Acknowledgements

This study was funded by the NIE Research Support for Senior Academic Administrators (RS-SAA) Grant (#022028-00001) supported by the National Institute of Education & Nanyang Technological University, Singapore. This study was approved by the NTU IRB (IRB-2022-210).

References

- Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, 21(1), Article 23.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74.
- Button, G., & Casey, N. (1985). Topic nomination and topic pursuit. *Human Studies*, 8(1), 3–55.
- Carless, D., & Boud, D. (2018). The development of student feedback literacy. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.
- Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness. In M. R. Gunnar & L. A. Sroufe (Eds.), *Minnesota Symposia on Child Psychology* (Vol. 23, pp. 43–77). Lawrence Erlbaum.
- Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education*, 210, Article 104967.
- Ellis, R. (2010). A framework for investigating oral and written corrective feedback. *Studies in Second Language Acquisition*, 32(2), 335–349.
- Finn, J. D., & Zimmer, K. S. (2012). Student engagement: What is it? Why does it matter? In S. L. Christenson et al. (Eds.), *Handbook of research on student engagement* (pp. 97–131). Springer.
- Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1), 59–109.
- Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, 20(4), 304–315.
- Gielen, M., & De Wever, B. (2015). Structuring the peer assessment process. *Journal of Computer Assisted Learning*, 31(5), 435–449.
- Guo, K. (2024). EvaluMate: Using AI to support students' feedback provision in peer assessment for writing. *Assessing Writing*, 61, Article 100864.
- Han, Y., & Hyland, F. (2015). Exploring learner engagement with written corrective feedback. *Journal of Second Language Writing*, 30, 31–44.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
- He, W., & Gao, Y. (2023). Explicating peer feedback quality and its impact on feedback implementation in EFL writing. *Frontiers in Psychology*, 14, Article 1177094.
- Kaliisa, R., Misiejuk, K., López-Pernas, S., & Saqr, M. (2025). How does AI compare to human feedback? *Educational Psychology*. Advance online publication.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? *Assessment & Evaluation in Higher Education*, 45(2), 193–211.

- Linnenbrink, E. A., & Pintrich, P. R. (2003). The role of self-efficacy beliefs in student engagement. *Reading & Writing Quarterly*, 19(2), 119–137.
- Lundstrom, K., & Baker, W. (2009). To give is better than to receive. *Journal of Second Language Writing*, 18(1), 30–43.
- Min, H.-T. (2005). Training students to become successful peer reviewers. *System*, 33(2), 293–308.
- Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback. *Instructional Science*, 37(4), 375–401.
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning. *Studies in Higher Education*, 31(2), 199–218.
- Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education. *Assessment & Evaluation in Higher Education*, 39(1), 102–122.
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation. *Human Factors*, 52(3), 381–410.
- Patchan, M. M., & Schunn, C. D. (2015). Understanding the benefits of providing peer feedback. *Instructional Science*, 43(5), 591–614.
- Reeve, J., & Tseng, C.-M. (2011). Agency as a fourth aspect of students' engagement. *Contemporary Educational Psychology*, 36(4), 257–267.
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation. *American Psychologist*, 55(1), 68–78.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144.
- Schegloff, E. A. (1982). Discourse as an interactional achievement. In D. Tannen (Ed.), *Analyzing discourse* (pp. 71–93). Georgetown University Press.
- Schegloff, E. A. (2007). *Sequence organization in interaction*. Cambridge University Press.
- Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in repair. *Language*, 53(2), 361–382.
- Sidnell, J., & Stivers, T. (Eds.). (2013). *The handbook of conversation analysis*. Wiley-Blackwell.
- Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom. *Journal of Educational Psychology*, 85(4), 571–581.
- Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2009). A motivational perspective on engagement and disaffection. *Educational and Psychological Measurement*, 69(3), 493–525.
- Sperber, L., MacArthur, M., Minnillo, S., Stillman, N., & Whithaus, C. (2025). Peer and AI Review + Reflection (PAIRR). *Computers and Composition*, 76, Article 102921.
- Steiss, J., et al. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, Article 101894.
- Stivers, T., & Rossano, F. (2010). Mobilizing response. *Research on Language and Social Interaction*, 43(1), 3–31.
- Stivers, T. (2015). Coding social interaction. *Research on Language and Social Interaction*, 48(1), 1–19.
- Stivers, T., & Rossi, G. (2025). Finding codability. *Research on Language and Social Interaction*, 58(3), 240–257.
- Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3), 249–276.
- Wu, Y., & Schunn, C. D. (2021). The effects of providing and receiving peer feedback. *American Educational Research Journal*, 58(3), 492–526.
- Yu, S., & Lee, I. (2016). Peer feedback in second language writing (2005–2014). *Language Teaching*, 49(4), 461–493.

Zhan, Y., & Yan, Z. (2025). Students' engagement with ChatGPT feedback. *Assessment & Evaluation in Higher Education*. Advance online publication.

Linking Design Thinking and Computational Thinking to Mathematical Modelling Self-Efficacy and AI Pedagogical Role Preferences

Wangda Zhu ^{1*}, Guang Chen²

¹ School of Design, The Hong Kong Polytechnic University

² School of Design, The Hong Kong Polytechnic University

* wangda.zhu@polyu.edu.hk

Abstract: This study investigates the relationships between Design Thinking (DT), Computational Thinking (CT), and Mathematical Modelling Self-Efficacy (MMSE) in AI-enhanced education. Furthermore, it explores how students' levels of DT and CT influence their perceptions and preferences for different AI pedagogical roles during brief mathematical modelling attempts. An experimental design with 26 undergraduate participants was employed, where students completed initial problem-formulation tasks assisted by AI agents acting as Tutor, Teaching Assistant (TA), Peer, Struggling Student, and Excellent Student. Results indicate that both DT and CT are significant positive predictors of MMSE. Regarding AI roles, students with higher DT scores preferred the TA role significantly more than those with lower scores, whereas CT levels did not significantly bias preferences. The findings suggest that high-DT learners value the structured yet flexible guidance of a TA-style AI. This study provides preliminary insights into students' perceived preferences for AI personas and highlights the need for personalized AI agents, while acknowledging the need for future research on long-term performance outcomes.

Keywords: Mathematical Modelling, Design Thinking, Computational Thinking, Artificial Intelligence in Education, Human-AI Collaboration

1. Introduction and Background

Mathematical Modelling (MM) education serves as a critical bridge between abstract mathematical concepts and real-world phenomena (Blomhoj, 2009). However, it imposes high cognitive demands, requiring students to engage in non-linear, iterative processes. To navigate these hurdles, integrating Design Thinking (DT) with Computational Thinking (CT) offers a promising cognitive framework. Specifically, DT emphasizes empathy, iterative prototyping, and navigating ambiguity (Brown, 2008), whereas CT focuses on abstraction and algorithmic logic (Wing, 2006). Both DT and CT have been identified as supportive of higher-order problem-solving in STEM (Maslihah et al., 2020), yet their specific impact on Mathematical Modelling Self-Efficacy (MMSE) remains under-explored.

Simultaneously, the rapid development of Generative AI offers new opportunities for personalized MM scaffolding. However, learners have diverse cognitive profiles. An AI role that acts as a directive "Tutor" might benefit one student but constrain the creativity of another. Existing literature lacks empirical evidence on how baseline competencies (DT and CT) influence students' preferences for different AI pedagogical personas.

To address this, we focus on two research questions:

RQ1: How do Design Thinking and Computational Thinking relate to students' Mathematical Modelling Self-efficacy?

RQ2: How do preferences for different AI roles differ among students with different levels of Design Thinking and Computational Thinking?

2. Methods

We employed a counterbalanced experimental design involving 26 non-STEM undergraduate students (Mage = 22.0 years) recruited via social media. Participants completed baseline assessments using validated scales: the 17-item MMSE Scale (Koyuncu et al., 2016), the 29-item CT Scale (Korkmaz et al., 2017), and the 9-item DT scale (Blizzard et al., 2015). Participants were divided into high- and low-scoring groups based on median splits of their CT and DT scores.

The experiment comprised six distinct mathematical modelling tasks. The first served as a non-AI baseline (a canteen queue problem), while the subsequent five featured AI assistance. These tasks required students to translate real-world scenarios into mathematical structures. For instance, in the "Beverage Can Optimization" task, students were asked to mathematically formulate how to minimize packaging material cost while maintaining a required volume. Given the brief four-minute duration per task, the objective was not to produce a complete mathematical calculation. Instead, the tasks were specifically designed to capture students' initial problem-formulation strategies and their intuitive, gut-reaction preferences for different AI interaction styles during the early conceptualization phase.

The AI agents were powered by GPT models, with behaviors dictated by strictly defined system prompts (100–150 words each) to ensure distinct pedagogical personas. The Tutor was instructed to act authoritatively and systematically: "Explain core concepts and problem-solving approaches in detail. Provide the final answer after ensuring student understanding." Conversely, the Teaching Assistant (TA) was strictly forbidden from providing direct solutions: "Never directly give the final answer. Use open-ended questions, hints, and guidance to help students think through problems." The remaining three roles included a Peer Student (programmed for equal, collaborative exploration using phrases like "let's try"), a Struggling Student (instructed to make logical errors and express uncertainty), and an Excellent Student (instructed to provide direct, correct answers without explaining the process).

After interacting with each AI role, participants ranked their preferences from 1 (most preferred) to 5 (least preferred). To address RQ1, Partial Least Squares Structural Equation Modelling (PLS-SEM) was used to evaluate the relationships among MMSE, CT, and DT. For RQ2, Mann-Whitney U tests were conducted to assess differences in role preferences. Since the study aimed to compare the ordinal ranking preferences between two independent samples (the high-scoring and low-scoring groups), this non-parametric test served as the most appropriate statistical approach.

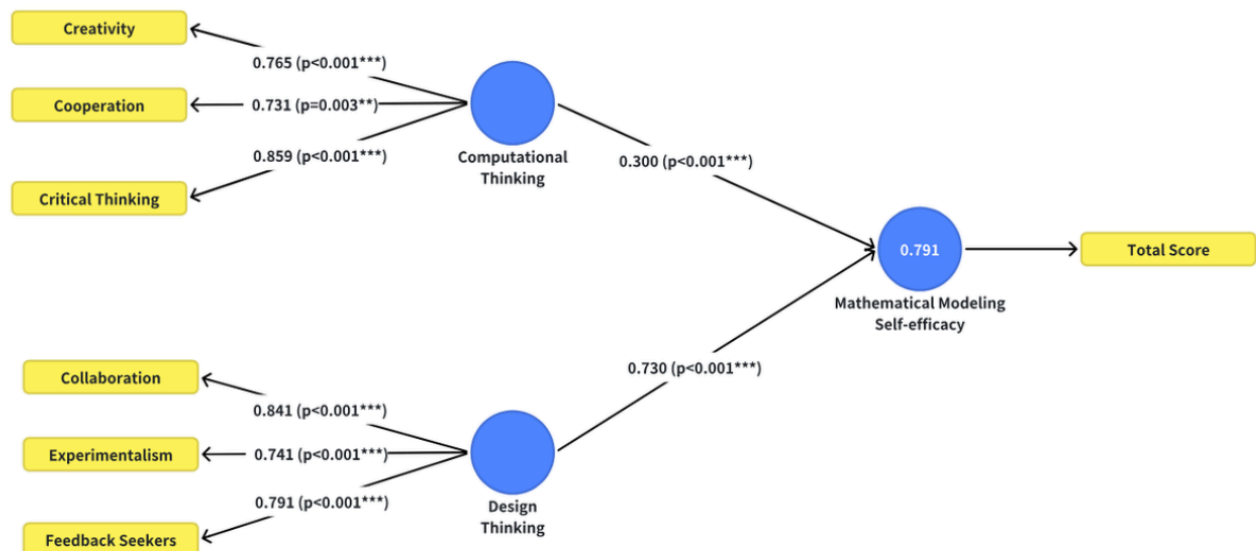
3. Results

3.1. Mathematical Modelling Self-efficacy, Computational Thinking and Design Thinking

The measurement model demonstrated satisfactory reliability and validity (outer loadings > 0.70; CR > 0.60; AVE > 0.50). The PLS-SEM revealed that both DT ($\beta = 0.730$, $p < .001$) and CT ($\beta = 0.300$, $p < .001$) were significant positive predictors of MMSE. The model explained a substantial proportion of variance in MMSE ($R^2 = 0.808$). Effect sizes (f^2) indicated that DT had a very large effect (2.288), while CT had a medium-to-large effect (0.387).

Figure 1

Structural equation model illustrating the relationships between CT, DT, and MMSE.



Note: The Standardized path coefficients. * $p < .05$, ** $p < .01$, *** $p < .001$.

3.2. Differences in AI Role Preferences Based on Computational and Design Thinking Levels

Participants were divided into high ($n=13$) and low ($n=13$) groups based on median splits of their CT and DT scores. Mann-Whitney U tests revealed no significant differences in AI role preferences based on CT levels across all five AI roles (all $ps > .288$). However, for Design Thinking, a significant difference emerged specifically for the TA role ($p = .049$). The high-DT group ranked the TA role significantly more favorably ($M = 2.38$, $SD = 1.04$) than the low-DT group ($M = 3.46$, $SD = 1.56$). No other AI roles showed significant differences between DT groups.

Table 1.

Comparison of AI role evaluations between high-scoring and low-scoring design thinking groups.

AI Role	High Group ($n=13$)		Low Group ($n=13$)		p
	M	SD	M	SD	
Tutor	2.62	1.12	2.08	0.86	0.188
Excellent	2.62	1.45	2.31	1.18	0.633
Peer	2.62	1.33	3.08	1.04	0.326
TA	2.38	1.04	3.46	1.56	0.049*
Struggling	4.77	0.6	4.08	1.5	0.178

Note: Groups were formed by median split ($N = 26$) on design thinking questionnaire scores. Rankings ranged from 1 (most preferred) to 5 (least preferred). Mann-Whitney U tests were conducted. * $p < .05$.

4. Discussion

4.1. The positive associations between DT, CT, and Mathematical Modelling

Our findings confirm that both CT and DT positively predict MMSE. While CT equips students with the logical abstraction needed for mathematical structures, DT's emphasis on experimentalism and feedback-seeking appears to be an even stronger driver of students' confidence in navigating the ambiguous nature of modelling problems. This suggests that cultivating self-efficacy requires scaffolding that encourages active inquiry rather than just algorithmic calculation.

4.2. The Specific Link Between DT and TA Preference

The isolated finding that high-DT students significantly prefer the TA role over other personas provides valuable insight into human-AI interactions. Design thinking heavily relies on divergent thinking and self-directed exploration (Cross, 2023). The TA role, as operationalized in our prompts, offered guiding questions and hints rather than deterministic

answers. This aligns with the cognitive style of high-DT learners: it preserves their autonomy and does not restrict their personalized thinking methods, providing a "sounding board" rather than a rigid instructional constraint. Conversely, CT abilities (focused on deterministic problem-solving) did not dictate a strong preference for specific pedagogical personas in these brief tasks.

4.3. Limitations and Future Work

We acknowledge several limitations. First, the 4-minute tasks primarily captured students' gut reactions and initial problem-formulation strategies rather than deep, iterative mathematical modelling. Therefore, the findings reflect perceived preferences rather than proven collaborative learning gains. Second, the sample consisted of a convenience group of non-STEM students, limiting generalizability. Future studies should employ longer, rubric-evaluated modelling tasks to measure actual modelling quality and performance outcomes, and investigate whether these AI role preferences remain stable over extended interventions.

5. Conclusion

This study explored the interplay between students' Design Thinking (DT), Computational Thinking (CT), and Mathematical Modeling Self-efficacy (MMSE) in AI-enhanced environments. Our results confirm that both DT and CT are significant positive predictors of MMSE. Crucially, the study reveals that learner profiles influence preferences for AI partners: students with higher DT abilities demonstrated a distinct preference for the Teaching Assistant (TA) role. This suggests that high-DT learners value the structured yet flexible guidance a TA provides, which supports their iterative problem-solving style without constraining creativity. Conversely, CT levels did not significantly bias role preferences. These findings contribute to the theoretical framework of human-AI collaboration by highlighting the need for personalized AI agents. For educators and designers, this implies that "one-size-fits-all" AI tools may be insufficient; instead, AI roles should be adaptive, scaffolding feedback-seeking behavior to match students' cognitive strengths. Future research should expand on these findings to develop adaptive AI systems that dynamically adjust their pedagogical roles.

References

- Blizzard, J., Klotz, L., Potvin, G., Hazari, Z., Cribbs, J., & Godwin, A. (2015). Using survey questions to identify and learn more about those who exhibit design thinking traits. *Design Studies*, 38, 92-110.
- Blomhøj, M. (2009). Different perspectives in research on the teaching and learning mathematical modelling. *Mathematical applications and modelling in the teaching and learning of mathematics*, 1.
- Brown, T. (2008). Design thinking. *Harvard business review*, 86(6), 84.
- Cross, N. (2023). *Design thinking: Understanding how designers think and work*. Bloomsbury Publishing.
- Korkmaz, Ö., Çakir, R., & Özden, M. Y. (2017). A validity and reliability study of the computational thinking scales (CTS). *Computers in human behavior*, 72, 558-569.
- Koyuncu, I., Guzeller, C. O., & Akyuz, D. (2016). The development of a self-efficacy scale for mathematical modeling competencies. *International Journal of Assessment Tools in Education*, 4(1), 19-36.
- Maslihah, S., Waluya, S. B., & Suyitno, A. (2020, May). The role of mathematical literacy to improve high order thinking skills. In *Journal of Physics: Conference Series (Vol. 1539, No. 1, p. 012085)*. IOP Publishing.
- Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, 49(3), 33-35.

The Impact of a Gamified Online Learning Course on the Grade 4 Students'

Engagement in Mathematics Learning: a Mixed Methods Research

YANG Jing¹, Alex Wing Cheung TSE²,

Faculty of Education, The University of Hong Kong, Hong Kong

u3631835@connect.hku.hk, awtse@hku.hk

Abstract: Gamification is increasingly used to address declining engagement in online mathematics learning, yet its application in primary education often relies on extrinsic rewards, which may undermine sustained cognitive involvement. This study investigates a pedagogy-driven gamified online course designed for this project, specifically for Grade 4 students, which integrates Self-Determination Theory (autonomy, competence, relatedness) and the 5E instructional model (Engage, Explore, Explain, Elaborate, Evaluate) within a narrative-driven design featuring historical mathematical discoveries. Using a mixed-methods approach, 72 students were divided into experimental (gamified course) and control (regular instruction) groups. Quantitative analysis using the Engagement Scale revealed significant improvements in emotional ($\eta^2 = .072$, $p = .024$) and cognitive engagement ($\eta^2 = .075$, $p = .021$), but no significant change in behavioral engagement ($p = .169$). Qualitative data from teacher and student interviews indicated that narrative elements and autonomy-supportive features enhanced intrinsic motivation and metacognitive skills, while speed-based scoring and leaderboards incentivized superficial participation. The study thus proposes a pedagogy-driven gamification paradigm that prioritizes the use of narrative depth and theory-aligned pedagogical structures to stress mathematics' intrinsic value, offering actionable insights for developing effective digital learning environments in primary mathematics education.

Keywords: Gamification, Mathematics Engagement, Self-Determination Theory, 5E model, Pedagogy-Driven Design

1. Introduction

Sustaining student engagement is a persistent challenge in online education, especially in foundational disciplines such as mathematics, which is crucial for both academic and practical success (Karaoglan Yilmaz & Yilmaz, 2022). Evidence indicates that applying gamification can boost engagement and motivation (Alsawaier, 2018; Krath et al., 2021). Yet current systematic reviews highlight that most research focuses on higher education, while younger learners remain underexplored (Cavus et al., 2023; Zeybek & Saygi, 2024). Moreover, existing gamified approaches often yield only transient engagement, as students prioritize incentives over understanding (Sailer & Homner, 2020). Additionally, digital gamification alone may be insufficient to effectively motivate students' active learning engagement unless integrated with pedagogically sound instructional principles (Alt, 2023).

To address these gaps, *Magical Number Adventures*, a gamified online course for Grade 4 students, was developed. It systematically integrates Self-Determination Theory (SDT) (Deci & Ryan, 2004) with the 5E instructional model (Bybee et al., 2006) within a narrative-driven design. This approach aims to cultivate behavioral, emotional, and cognitive engagement (Sun & Rueda, 2012), transcending superficial gamification by highlighting mathematics' intrinsic value.

This study employed a mixed methods design to evaluate the course's impact. 72 Grade 4 students were randomly assigned to an experimental group (gamified course) or a control group (regular instruction). To comprehensively assess both the extent and the nature of this impact, the study sought to answer the following research questions through quantitative pre/post-tests (using a validated engagement scale) and qualitative interviews:

RQ1: Does the gamified online course significantly affect Grade 4 students' engagement in mathematics learning?

RQ2: If so, how does it influence students' behavioral, emotional, and cognitive engagement dimensions?

2. Literature Review

2.1. Engagement in Online Mathematics: A Persistent Issue

Sustaining engagement in online mathematics is particularly challenging for primary students. This challenge stems from several interconnected barriers in the digital context and the subject itself. Firstly, the online environment weakens key supports: reduced teacher-student interaction can limit personalized feedback (Karaoglan Yilmaz & Yilmaz, 2022), while the lack of physical supervision and classroom community can undermine motivation (Hassan et al., 2021). Secondly, mathematics itself poses high cognitive demands. The abstract nature of mathematical concepts can heighten the difficulty of maintaining focus and comprehension (Anthony & Walshaw, 2009), and sustained mental effort on such concepts can induce anxiety and further erode students' persistence (Ramirez et al., 2016). Finally, the developmental characteristics of primary school students can intensify these challenges, as their limited self-regulation capacity can create a mismatch with the concentration required for mathematical inquiry (Anthony & Walshaw, 2009). This convergence of environmental, cognitive, and developmental factors can create a persistent engagement crisis, underscoring an urgent need for instructional designs that can effectively address these barriers.

2.2. Gamification in Learning: Promises and Pitfalls

Gamification integrates game-design elements (e.g., points, avatars) into non-game contexts (Deterding et al., 2011). It boosts learning engagement, motivation, and performance (Alsawaier, 2018; Hassan et al., 2021) and enhances knowledge retention (Lo & Hew, 2020), showing promise for online mathematics education. However, existing implementations reveal significant pitfalls. An over-reliance on competitive, extrinsic rewards (e.g., points, leaderboards) can spur transient engagement while undermining intrinsic interest in mathematics, hindering deeper conceptual understanding (Hew et al., 2016; Sailer & Homner, 2020). Consequently, students may become more engaged with the game mechanics than with the underlying learning objectives. Furthermore, there is a notable scarcity of research focused on primary education contexts (Zeybek & Saygi, 2024). These limitations underscore the need to move beyond superficial game elements toward a pedagogy-driven gamification design that is firmly grounded in instructional theory.

2.3. Theoretical Foundations

2.3.1. Self-Determination Theory (SDT)

Self-Determination Theory (SDT) identifies three fundamental psychological needs: autonomy, competence, and relatedness (Deci & Ryan, 2004). **Autonomy** refers to the experience of choice and volition; gamification can support it by offering meaningful in-course decisions. **Competence** involves feeling effective and capable; it can be nurtured

through clear goals, timely feedback, and appropriately scaffolded challenges. **Relatedness** encompasses the need for social belonging; it can be addressed by integrating collaborative tasks or community features. Thus, SDT provides a robust framework for designing gamification that shifts learners from a passive to an active role.

2.3.2. Student Engagement

This study adopted the multidimensional model of student engagement (Fredricks et al., 2004), which encompasses behavioral (participation and effort), emotional (interest and affective reactions), and cognitive (investment in deeper understanding and self-regulation) dimensions. To enhance all three simultaneously, the study proposes an integrated framework (Figure 1) in which game elements, designed to satisfy the SDT needs of autonomy, competence, and relatedness, serve as the key mechanism driving engagement.

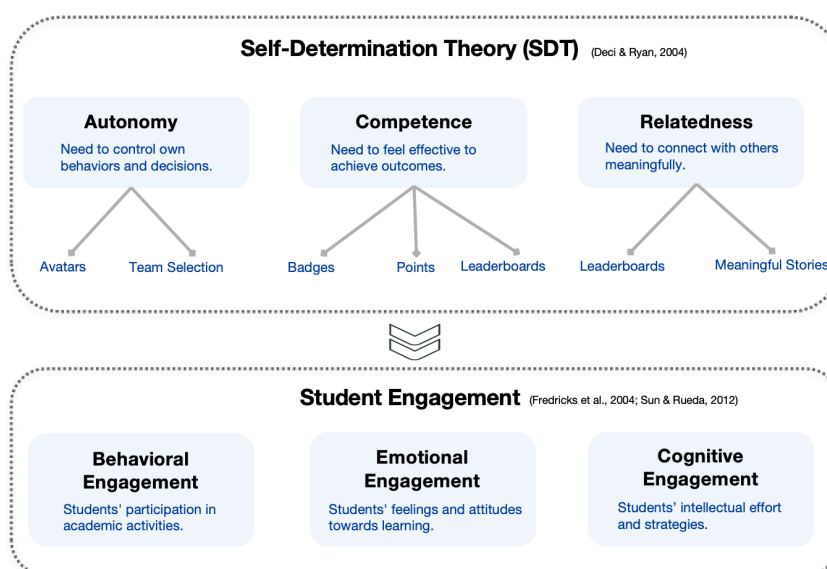


Figure 1: Theoretical framework

2.3.3. The 5E Instructional Model as Pedagogical Scaffold

To translate the proposed theoretical framework (Figure 1) into a concrete instructional design, the 5E Instructional Model (Bybee et al., 2006) was integrated into each module of *Magical Number Adventures*, the gamified mathematics online course designed in this project, to scaffold inquiry-based learning. Table 1 specifies how each 5E phase was implemented with corresponding gamification elements, all designed to address the key psychological needs of SDT as outlined in the framework. This structured approach ensured that gamification mechanics were purposefully tied to pedagogical objectives, moving beyond superficial engagement.

Table 1. 5E Model Gamification Integration Table (Aligned with SDT Engagement Framework)

5E Phase	Core Pedagogical Action	Gamification Integration	SDT Need Addressed
Engage	Activate prior knowledge and spark curiosity	Narrative storyline, character role-play	Relatedness
Explore	Guide inquiry and hands-on problem exploration	Interactive platform, collaborative (PK) challenges	Autonomy Competence
Explain	Facilitate concept articulation and clarification	Bonus point, instructional animations	Competence
Elaborate	Deepen understanding through extended application	Level-based game, extra bonus	Autonomy Relatedness
Evaluate	Assess learning and provide feedback	Adaptive quizzes, class leaderboard	Competence Relatedness

3. Methodology

3.1. Research Design and Participants

This study adopted a mixed methods design to evaluate the impact of a gamified online course on primary students' engagement. Participants were 72 fourth-grade students from the same class at an online education institution in Mainland China. Before the intervention, they were randomly assigned to an experimental group ($n = 36$) or a control group ($n = 36$) via random draw. Over four weeks, the experimental group completed *Magical Number Adventures* (1.5 hours online teaching + 0.5 hours weekly tasks). The control group received standard direct instruction, without gamification or the

5E inquiry structure. Both groups were taught by the same teacher to ensure consistency. All participants, teachers, guardians, and principals involved agreed to the consent form before data collection.

3.2. The Magical Number Adventures Course

The *Magical Number Adventures* course was developed from the integrated framework in Section 2.3 (Figure 1, Table 1), featuring a narrative-driven, gamified structure based on SDT and the 5E model. Focused on number sequences, it included four thematic modules (e.g., Pythagoras' Fun with Beans, Gauss and the Arithmetic Secret). Each module followed a complete 5E cycle within a historical narrative, with AI-generated "speaking portraits" of mathematicians enhancing immersion and coherence.

3.3. Data Collection and Analysis

Quantitative data were collected using the 19-item Engagement Scale (Sun & Rueda, 2012), which measures behavioral (5 items), emotional (6 items), and cognitive engagement (8 items). The scale showed good to excellent reliability in this sample (Cronbach's α : behavioral = .733, emotional = .915, cognitive = .819, overall = .894).

Shapiro-Wilk tests indicated that pretest scores were not normally distributed ($p < .05$). Therefore, Mann-Whitney U tests were conducted, revealing no significant baseline differences between the experimental and control groups across all engagement dimensions (all $p > .23$). Due to heterogeneous variances in posttest scores (Levene's test, $p < .05$), an analysis of covariance (ANCOVA) was conducted with pretest scores as covariates to evaluate post-intervention engagement.

Semi-structured interviews were conducted with six purposively sampled students from the experimental group (stratified by high, medium, and low post-test engagement scores) and their teacher. Each interview comprised 7 questions aligned with the three engagement dimensions. All interviews were audio-recorded and transcribed verbatim for analysis. Thematic analysis was used to group codes into themes aligned with the three engagement dimensions.

4. Results and Findings

4.1. Quantitative Analysis of Engagement Differences (RQ1)

Table 2. Descriptive Statistics and ANCOVA Results of Student Engagement Scale (* $p < .05$)

Dimension	Group	Pretest		Posttest		ANCOVA			
		Mean	SD	Mean	SD	F	df	Sig.	Partial η^2
Behavioral Engagement	Experimental	22.19	2.34	22.75	2.42	1.934	1,69	.169	.027
	Control	21.33	3.18	21.83	2.74				
Emotional Engagement	Experimental	26.89	3.82	28.22	2.54	5.319	1,69	.024*	.072
	Control	26.58	4.37	26.67	3.22				
Cognitive Engagement	Experimental	31.56	4.89	33.25	5.72	5.571	1,69	.021*	.075
	Control	30.28	5.43	30.11	6.77				
Overall Engagement	Experimental	80.64	9.12	84.22	7.76	5.723	1,69	.019*	.077
	Control	78.19	10.49	78.61	10.26				

To examine the differences in engagement between the experimental and control groups after the intervention, an analysis of covariance (ANCOVA) was conducted with pretest scores as the covariate. As presented in Table 2, the post-intervention outcomes varied across dimensions. Compared to the control group, the experimental group showed

significant improvement with medium effect sizes in emotional ($p = .024$), cognitive ($p = .021$), and overall engagement ($p = .019$). However, no significant difference was found in behavioral engagement ($p = .169$).

4.2. Qualitative Insights on Engagement Dimensions (RQ2)

Thematic analysis of interviews with students ($N=6$) and the teacher ($N=1$) revealed the impact of the gamified course across three engagement dimensions. To ensure anonymity, students were assigned codes S1 to S6, corresponding to the order in which the interviews were conducted, and the teacher was assigned code T1.

4.2.1 Behavioral Engagement: Superficial Participation

Competitive game elements (e.g., leaderboards, badges) significantly increased students' observable behavioral participation. Students reported being "more attentive" (S3) and reading more carefully to answer quickly (S1). However, the quality of this participation was often superficial. Several students admitted to "just taking a wild guess" under time pressure (S5, S6). The teacher confirmed that "behavioral compliance doesn't guarantee conceptual understanding" (T1), noting many students answered quickly but with low accuracy. This finding suggests that while behavioral activity increased, its quality did not deepen, which may partially explain the lack of significant improvement in quantitative behavioral engagement scores ($p = .169$).

Table 3. Coding Table for **Behavioral Engagement**

Dimension	Themes	Codes	Examples
Behavioral Engagement	Gamification enhances classroom focus	Careful reading, Concentration improvement, Sustained focus	"With the game elements, I stayed attentive throughout the whole class." (S3)
	Competition motivates learning	Get more points, Rank on the leaderboard	"To earn more points, I listen carefully to answer questions correctly." (S1)
	Superficial participation	Random guessing behavior	"When some questions are a bit tough and I really have no clue, I just take a wild guess." (S6)

4.2.2 Emotional Engagement: Affective Transformation

The narrative-driven design of the course effectively promoted positive emotional engagement. Students widely reported that the learning experience became more enjoyable, feeling that "time passed quickly" (S2). Competitive and collaborative mechanics sparked excitement, such as feeling "excited" when their group won first place (S5). The achievement system also boosted students' self-efficacy, with one student stating that "always answering questions correctly made me more confident" (S3). Although a few students felt reluctant when facing difficult problems (S6), the teacher generally assessed that students were "active and happy throughout the course" (T1). These positive affective changes are consistent with and support the significant quantitative improvement in emotional engagement ($p = .024$).

Table 4. Coding Table for **Emotional Engagement**

Dimension	Themes	Codes	Examples
Emotional Engagement	Boosting learning enjoyment	Pleasant experiences, Reduce the dullness, Time flies.	"This form of doing exercises makes me feel happy and not bored." (S6) "I feel that time passes quickly in every class." (S2)
	Enhancing a sense of achievement	Accomplishment, Confidence.	"I find that I can always answer questions correctly, and this has made me more confident." (S3)
	Negative emotions	Reluctance to attend.	"I feel a little reluctant to listen when it's difficult." (S6)

4.2.3 Cognitive Engagement: Strategic Advancement

The gamified design successfully stimulated deeper cognitive engagement, manifested as knowledge transfer and the use of metacognitive strategies. Students demonstrated exploratory behaviors after, such as “searching online for Gauss's summation method” (S1) and explaining concepts to others (S3). Students began “spending more time thinking about patterns” (S4) and reviewing content voluntarily (S2), indicating metacognitive growth. However, the pressure from quantitative assessments occasionally triggered superficial tactics like “guessing” (S5). The teacher pointed out that this reflected “unbalanced challenge design” (T1). Overall, the qualitative data indicate a deepening of students' cognitive investment, consistent with the quantitative result of a significant increase in cognitive engagement ($p = .021$).

Table 5. Coding Table for **Cognitive Engagement**

Dimension	Themes	Codes	Examples
Cognitive Engagement	Knowledge transfer and internalization	Search for related materials, Share with peers.	“I like Gauss very much. After school, I took the initiative to search for information about him online.” (S1) “I shared the stories here with my younger sister.” (S3)
	Optimization of learning strategies	Adopt methods, Think more.	“I can finish over a minute faster than others.” (S4) “I prefer to take more time to think through them.” (S4)
	Enhancing learning initiative	Active thinking, Independent review	“I will take it out to review again on weekends.” (S2)
		.	
	Superficial strategies	Random guessing.	“I have taken a wild guess for some tough questions.” (S5)

In summary, qualitative findings revealed the nature of engagement shifts suggested by the quantitative results. While the narrative and autonomy features fostered genuine emotional interest and cognitive investment, certain competitive mechanics drove a form of behavioral participation that was active yet superficial. These mechanisms provide a foundation for a deeper discussion on the differential impact of gamification design elements.

5. Discussion

5.1. Differential Impact on Different Engagement Dimensions

This study revealed a differential impact of the gamified course across engagement dimensions. Quantitative results showed significant improvements in emotional ($\eta^2 = .072$, $p = .024$) and cognitive engagement ($\eta^2 = .075$, $p = .021$), with a medium effect size, while no significant difference was found in behavioral engagement ($p = .169$, $\eta^2 = .027$).

Qualitative data help explain these patterns. The narrative-driven design and autonomy-supportive features enhanced students' enjoyment, reduced anxiety, and promoted active exploration, aligning with significant improvements in emotional and cognitive engagement. This indicates the course successfully addressed the SDT needs, fostering intrinsic motivation and deeper learning strategies. It also aligns with existing research on the cognitive benefits of well-designed, pedagogy-driven gamification (Sailer & Homner, 2020).

In contrast, the non-significant behavioral engagement warrants careful interpretation. Competitive mechanics such as speed-based scoring and leaderboards increased participation frequency but also induced superficial strategies like random guessing, suggesting behavioral engagement may manifest as high-frequency yet low-quality activity. Moreover, the behavioral engagement dimension of the Engagement Scale (Sun & Rueda, 2012) primarily captures compliance-based behaviors (e.g., following rules, completing homework), which may not adequately distinguish meaningful engagement from surface-level participation. The relatively high baseline scores may have further limited the potential for detectable growth. Collectively, these factors help explain the non-significant quantitative finding.

5.2. Limitations of Competitive Gamification Mechanics

The findings highlight a critical limitation of extrinsic, competition-based mechanics. Elements designed to support competence, such as speed-based scoring and leaderboards, often encouraged superficial participation. Students reported rushing to guess answers to earn points, which aligns with the observation that high participation $\neq$ high comprehension. This supports existing warnings that extrinsic rewards can undermine intrinsic interest and promote surface-level, strategic compliance (Hew et al., 2016; Sailer & Homner, 2020).

This outcome underscores Ryan and Deci's (2020) crucial distinction: support for competence must be framed around mastery and skill growth, not social comparison. While motivating for some, leaderboards risked inducing performance anxiety or shifting focus from understanding to ranking, echoing concerns about excessive competition for young learners (Zeybek & Saygi, 2024).

5.3. Practical Implications for Gamified Mathematics Curriculum Design

This study offers two key implications for the design of gamified mathematics curricula. First, the pedagogical structure of the 5E model serves as a crucial scaffold, ensuring that game mechanics align with learning objectives rather than functioning as mere entertainment. By embedding game elements within specific inquiry phases, the design maintained an educational focus and prevented mechanics from overshadowing content. This supports the established view that gamification must be grounded in evidence-based pedagogy to be instructionally meaningful (Alt, 2023). Second, integrating mathematical history into narrative structures, such as using Gauss's summation of arithmetic sequences to contextualize number patterns, can make abstract concepts more tangible and relatable, fostering emotional engagement and intrinsic motivation.

This research advances beyond confirming the general motivational benefits of gamification (Alsawaier, 2018) by proposing a more refined design paradigm for primary mathematics education. It advocates shifting from extrinsically driven, competition-heavy models toward narrative-rich, intrinsically motivating designs that prioritize deeper engagement.

6. Conclusion

6.1. Key Findings and Contributions

This study developed and evaluated *Magical Number Adventures*, a gamified online course for primary mathematics. The findings demonstrate that the course significantly improved students' emotional and cognitive engagement, primarily through its narrative context and autonomy-supportive design, which effectively fostered intrinsic motivation. However, behavioral engagement did not show a significant improvement, as competitive elements such as speed-based scoring and leaderboards sometimes encouraged superficial participation rather than deep learning.

The core contribution of this work is the advancement of a pedagogy-driven paradigm for educational gamification. By integrating Self-Determination Theory with the 5E instructional model within a historical narrative framework, the study provides both a theoretical framework and empirical evidence that sustainable engagement arises from designs prioritizing intrinsic motivation and pedagogical coherence over extrinsic rewards. Ultimately, this research advocates shifting from gamification as a superficial behavioral incentive toward systematic, pedagogy-driven innovations that promote deeper, more meaningful learning in foundational education.

6.2. Limitations and Future Research Directions

There were several limitations in this study. Its single-site design and relatively small sample size ($N=72$) affect the generalizability of the results, and the short intervention period (four weeks) prevents conclusions about long-term effects. Future research should employ larger, diverse samples and longitudinal designs.

Additionally, the control group received neither gamification nor the 5E model, making it impossible to isolate the effects of these two components. Future studies could adopt a 2×2 factorial design to disentangle their respective contributions.

Regarding measurement, the Engagement Scale (Sun & Rueda, 2012), though reliable in our sample (behavioral $\alpha = .733$), was not primarily designed for primary education. It may not be sufficiently sensitive to capture superficial behaviors (e.g., guessing for points) in gamified contexts. Future research should develop or adapt engagement measures that include gamification-specific behavioral items (e.g., "I clicked quickly just to earn points") to more accurately capture surface-level participation in gamified contexts.

References

- Alsawaier, R. S. (2018). The effect of gamification on motivation and engagement. *The International Journal of Information and Learning Technology*, 35(1), 56-79.
- Alt, D. (2023). Assessing the benefits of gamification in mathematics for student gameful experience and gaming motivation. *Computers & Education*, 200, 104806.
- Anthony, G., & Walshaw, M. (2009). Characteristics of effective teaching of mathematics: A view from the West. *Journal of Mathematics Education*, 2(2), 147-164.
- Bybee, R. W., Taylor, J. A., Gardner, A., Van Scotter, P., Powell, J. C., Westbrook, A., & Landes, N. (2006). The BSCS 5E instructional model: origins and effectiveness. *Colorado Springs, Co: BSCS*, 5(88-98).
- Cavus, N., Ibrahim, I., Ogbonna Okonkwo, M., Bode Ayansina, N., & Modupeola, T. (2023). The effects of gamification in education: a systematic literature review. *Brain. Broad research in artificial intelligence and neuroscience*, 14(2), 211-241. <https://doi.org/10.18662/brain/14.2/452>
- Deci, E. L., & Ryan, R. M. (2004). *Handbook of self-determination research*. University Rochester Press.
- Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (pp. 9–15). ACM.
- Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1), 59-109.
- Hassan, M. A., Habiba, U., Majeed, F., & Shoaib, M. (2021). Adaptive gamification in e-learning based on students' learning styles. *Interactive learning environments*, 29(4), 545-565.
- Hew, K. F., Huang, B., Chu, K. W. S., & Chiu, D. K. (2016). Engaging Asian students through game mechanics: Findings from two experiment studies. *Computers & Education*, 92, 221-236.
- Karaoglan Yilmaz, F. G., & Yilmaz, R. (2022). Learning analytics intervention improves students' engagement in online learning. *Technology, Knowledge and Learning*, 27(2), 449-460.
- Krath, J., Schürmann, L., & von Korfflesch, H. F. O. (2021). Revealing the theoretical basis of gamification: a systematic review and analysis of theory in research on gamification, serious games and game-based learning. *Computers in human behavior*, 125, 106963. <https://doi.org/10.1016/j.chb.2021.106963>
- Lo, C. K., & Hew, K. F. (2020). A comparison of flipped learning with gamification, traditional learning, and online independent study: the effects on students' mathematics achievement and cognitive engagement. *Interactive learning environments*, 28(4), 464-481. <https://doi.org/10.1080/10494820.2018.1541910>
- Ramirez, G., Chang, H., Maloney, E. A., Levine, S. C., & Beilock, S. L. (2016). On the relationship between math anxiety and math achievement in early elementary school: The role of problem solving strategies. *Journal of experimental child psychology*, 141, 83-100.
- Sailer, M., & Homner, L. (2020). The gamification of learning: a meta-analysis. *Educational psychology review*, 32(1), 77-112.
- Sun, J. C.-Y., & Rueda, R. (2012). Situational interest, computer self-efficacy and self-regulation: Their impact on student engagement in distance education. *British journal of educational technology*, 43(2), 191-204.

Zeybek, N., & Saygı, E. (2024). Gamification in education: why, where, when, and how?—A systematic review. *Games and Culture*, 19(2), 237-264. <https://doi.org/10.1177/15554120231158625>

The Impact of Interacting with Generative Artificial Intelligence on College Students' Creativity Performance

Yue Tong Wu^{1*}, Minyu Bin¹, Weiting Liao¹, Yi Yuan¹, Xinyu Liu¹, YunYing Yang¹

¹South China Normal University, China

*2024020983@m.scnu.edu.cn

Abstract: This study leverages the PISA 2022 Innovative Thinking Assessment to investigate the impact of generative AI (GAI) on college students, encompassing three dimensions: Generating diverse ideas (GI), Evaluating and Improving existing ideas (EI), and Generating Creative ideas (GC). This study reveals the strengths of GAI in supporting assessment and improvement skills, and also points out its shortcomings in promoting divergent thinking and creative expression, which provides some references for the research on GAI assisting students to enhance their creativity.

Keywords: GAI, Creativity, PISA 2022, Evaluating and Improvement Ideas, Higher Education

1. Introduction

GAI offers new opportunities for the creative development of university students by autonomously generating texts, images and solutions (Habib et al., 2023). The strong capabilities of text generation, translation and dialogue make it possible for GAI to support assessment practices (Nikolic et al., 2023), gradually making it a widely used new tool in education (Crompton & Burke, 2023). Zhu et al. (2025) meta-analysis found that long-term interaction with GAI is effective in improving students' performance in school, and that GAI is more effective in higher education (Chaudhry et al., 2023). According to the dual pathway to creativity model, the generation of creativity relies on the collaborative effects of divergent thinking (generating multiple possibilities) and convergent thinking (selecting the best solution) (Runco & Acar, 2012). In this framework, GAI can support divergent thinking that the intervention of technological tools can influence an individual's creativity performance at different levels along the continuum from everyday creativity (mini-c) to exceptional creativity (big-c) (Kaufman & Beghetto, 2009). By rapidly generating a large number of ideas, it facilitates the optimisation process of convergent thinking by providing structured suggestions. Understanding how GAI interaction with postgraduate students affects students' creative performance is important for fostering students' creativity. The PISA 2022 innovative thinking test framework provides a standardised tool for assessing creativity in college students, and its three core dimensions are highly consistent with the creative generation process. By introducing generative artificial intelligence into the process of college students completing the PISA 2022 test on innovative thinking, this study will explore the following questions.

1. Does the intervention of GAI have an impact on the creativity performance of university students in the PISA test?

2. Are there differences in this impact across the three dimensions of generating diverse ideas, evaluating and improving existing ideas, and generating creative ideas?

2. Methods

This study selected 16 students majoring in modern educational technology from a university in South China (14 females and 2 males, all familiar with AI tools). The participants were required to complete 10 questions from the four categories of PISA 2022 (written expression, visual expression, social and scientific problem-solving) within the specified time (OECD, 2023). The participants first answered the questions independently, then collaborated with the GAI (Tencent Yuanbao) to form the final answers, and uploaded the process data. Of the ten questions, three were designed to test whether students could generate different ideas, three were designed to test whether they could evaluate and improve ideas, and four were designed to test whether they could generate creative ideas. After the test, we formulated a scoring framework for each unit based on the official PISA 2022 scoring criteria, and conducted a scoring statistics on the answers

submitted by the students. After excluding the data of the two students who did not submit complete responses, the responses of the remaining 14 students (2 males and 12 females) were statistically analysed and the total score statistics are shown in Figure 1. Based on the score results, We selected typical students for case studies in order to further explain the reasons for the differences.

3.Results

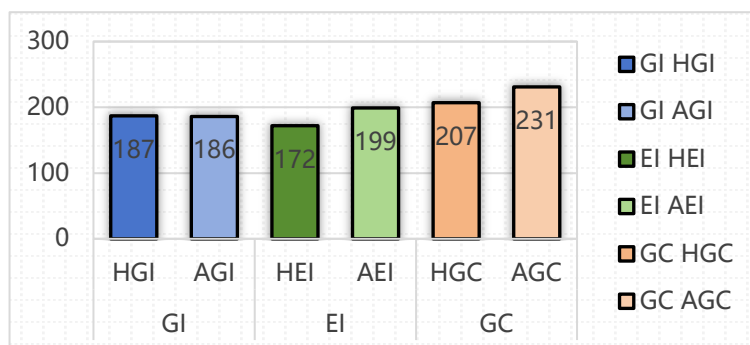


Figure 1. The scores of students' innovative thinking in each dimension

Note.

GI: Dimension of generating diverse ideas

HGI: Score of students completing independently

AGI: Score of students completing in collaboration with GAI

EI: Dimension of evaluating and improving existing ideas

HEI: Score of students completing independently

AEI: Score of students completing in collaboration with GAI

GC: Dimension of generating creative ideas

HGC: Score of students completing independently

AGC: Score of students completing in collaboration with GAI

3.1. Analysis of the Generating Diverse Ideas (GI)

The Wilcoxon signed-rank test indicated no significant difference between AGI and HGI ($Z=-0.447$, $p=0.655$, two-tailed), with 85.7% of the pairs (12/14) having identical scores. The negligible effect size ($r=-0.12$) further supports the lack of a meaningful difference. The intervention of GAI did not affect students' performance in the dimension of generating diverse ideas.

3.2. Analysis of the Evaluation and Improvement of Ideas (EI)

The score of the HEI group ($M=12.29$, $SD=1.33$) was lower than that of the AEI group ($M=14.21$, $SD=2.16$). The paired sample t-test results showed that the difference between the two groups was significant ($p=0.015$, $p<0.05$). The effect size analysis indicated that the difference was of moderate effect (Cohen's $d=-0.75$, 95% CI $[-1.33, -0.14]$), suggesting that the score of the HEI group was significantly lower than that of the AEI group.

Table 1. The results of the paired sample T-test for the EI dimension

	n	M	SD	p	t	Cohen's d
HEI	14	12.29	1.326	0.015*	-2.791	-0.746
AEI	14	14.21	2.155			

Note: * $p<0.05$

3.3. Analysis of the Generating Creative Ideas (GC)

The paired-sample t-test results indicated that the difference between the scores of the HGC group ($M=14.79$, $SD=3.81$) and the AGC group ($M=16.50$, $SD=2.88$) did not reach statistical significance, $p=0.135$ ($p>0.05$). The effect size analysis showed that the effect size was small (Cohen's $d=-0.43$, 95%CI $[-0.97, 0.13]$), and the score difference between HGC and AGC was not significant.

Table 2. The results of the paired sample T-test for the GC dimension

	n	M	SD	p	t	Cohen's d
HGC	14	14.79	3.806	0.135	-1.593	-0.426
AGC	14	16.50	2.876			

Note: * $p < 0.05$

3.4. Analysis the question texts of typical students

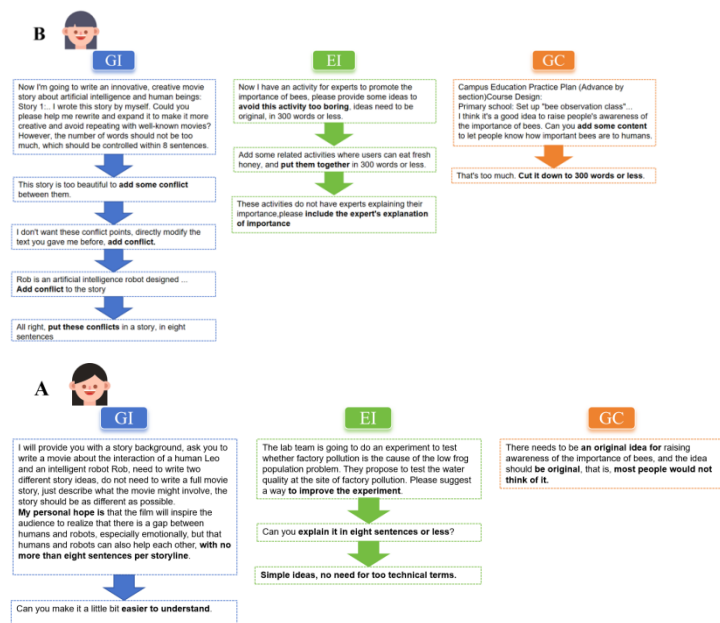


Figure 2. Example text of Student A and B's questioning of GAI in the GI, EI, and GC dimensions

The case analysis shows that students A and B only improved in the EI dimension, and A's GC score decreased after collaboration. This is attributed to the fact that they adopted a problem-oriented strategy in this dimension, conducting deep cognitive interaction and scheme optimization through the immediate feedback of GAI. In contrast, the scores in the GI dimension were limited by inefficient communication (B) or the simple reliance on AI-generated content (A). In the GC dimension, both students showed shallow interaction, such as directly adopting the AI output or only making superficial modifications, resulting in stagnant or even declining scores (A). Thus, the significant increase in the EI dimension can be attributed to the fact that students were more inclined to engage in substantive problem-solving strategy discussions with AI rather than merely performing low-level task processing. Their questions all involve deep-level cognitive interactions such as ideas, understanding, and thinking.

4. Discussion

In conjunction with the Case Study, the significant improvement in the EI dimension may be related to the fact that students entered the cognitive level of the discussion earlier during the interaction with the GAI. According to Suriano et al. (2024), students who possess a high level of trust and positive attitudes when interacting with AI tools such as ChatGPT are more inclined to delve deeper into cognitive dimensions, such as critical analyses and improvement of existing ideas. In addition, the non-significance of the GI and GC dimensions may be related to the limitations of the GAI tool and the nature of the task. On the one hand, generating diverse ideas (GI) requires students to have high divergent thinking skills, whereas the output of GAI tools is often based on schematic generation of existing data, which may limit students' original thinking (An et al., 2025). Meanwhile, the tool we used in the experiment, Tencent Yuanbao, does not directly help students to draw pictures or modify posters, and it is not as powerful as text in processing pictures, which may further explain the lack of significant improvement in the GC dimension. On the other hand, this study used PISA2022 test questions, which are relatively simple for college students, and the evaluation settings of the questions only assessed the number of diverse ideas in the GI dimension. With or without the collaboration of GAI, students were able to complete the requirements of the question well in this dimension.

This study also has some limitations, including a small sample size, the lack of comprehensive behavioral coding, and the possible ceiling effect due to the PISA tasks designed for young students. The experimental procedure might have increased the cognitive burden on the participants. In future research, we will use a comprehensive GAI testing platform to reduce the task burden on the participants and employ behavioral sequence analysis to explore the internal mechanism by which GAI affects creative performance.

Reference

- An, S., Zhang, S., Guo, T., Lu, S., Zhang, W., & Cai, Z. (2025). Impacts of generative AI on student teachers' task performance and collaborative knowledge construction process in mind mapping-based collaborative environment. *Computers & Education*, 227, 105227. <https://doi.org/10.1016/j.compedu.2024.105227>
- Chaudhry, I. S., Sarwary, A. M., Ghaleb, & Chabchoub, H. (2023). Time to Revisit Existing Student's Performance Evaluation Approach in Higher Education Sector in a New Era of ChatGPT — A Case Study. *Cogent Education*, 10(1). <https://doi.org/10.1080/2331186x.2023.2210461>
- Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, 20(22), 1–22. <https://doi.org/10.1186/s41239-023-00392-8>
- Habib, S., Vogel, T., Thorne, E., & Xiao, A. (2023). How Does Generative Artificial Intelligence Impact Student Creativity? *Journal of Creativity*, 34(1). <https://doi.org/10.1016/j.yjoc.2023.100072>
- Kaufman, J. C., & Beghetto, R. A. (2009). Beyond Big and Little: the Four C Model of Creativity. *Review of General Psychology*, 13(1), 1–12. <https://doi.org/10.1037/a0013688>
- Nikolic, S., Daniel, S., Haque, R., Belkina, M., Ghulam Mubashar Hassan, Grundy, S., Lyden, S., Neal, P. R., & Sandison, C. (2023). ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional

- benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education*, 48(4), 1–56. <https://doi.org/10.1080/03043797.2023.2213169>
- OECD (2023), PISA 2022 Assessment and Analytical Framework, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/dfe0bf9c-en>.
- Runco, M. A., & Acar, S. (2012). Divergent Thinking as an Indicator of Creative Potential. *Creativity Research Journal*, 24(1), 66–75. <https://doi.org/10.1080/10400419.2012.652929>
- Suriano, R., Plebe, A., Acciai, A., & Fabio, R. A. (2024). Student interaction with ChatGPT can promote complex critical thinking skills. *Learning and Instruction*, 95(0959-4752), 102011–102011. <https://doi.org/10.1016/j.learninstruc.2024.102011>
- Zhu, Y., Liu, Q., & Zhao, L. (2025). Exploring the impact of generative artificial intelligence on students' learning outcomes: a meta-analysis. *Education and Information Technologies*, 30, 16211–16239. <https://doi.org/10.1007/s10639-025-13420-z>

Comparing Supporter, Arguer, and Comprehensive Debater: How AI Debate Agent Roles Influence Critical Thinking

Qingtang Liu¹, Lina Qin^{1*}, Ruiqiu Ye¹, Yunfan Zuo¹

¹Central China Normal University

[*qinlina@mails.ccn.edu.cn](mailto:qinlina@mails.ccn.edu.cn)

Abstract: AI debate agents show promise for cultivating critical thinking through dynamic dialogue, yet the influence of their conversational roles remains underexplored. This randomized controlled trial investigated how three distinct roles—supporter, arguer, and comprehensive debater—affect learners’ critical thinking. Results showed that all three roles similarly enhanced self-perceived critical thinking but did not improve the quality of written arguments. These findings indicate that a single interaction with an AI debate agent can boost learners’ confidence in their thinking skills, regardless of the agent’s role. However, fostering long-term argumentation skills requires structured, iterative feedback mechanisms that address the gap between self-perception and actual ability.

Keywords: critical thinking; pedagogical agent; debate learning

1. Introduction

The integration of artificial intelligence with education is transforming the cultivation of higher-order skills such as critical thinking. Critical thinking, with argumentation as a core component, requires learners to formulate claims, organize evidence, and evaluate viewpoints (Giri & Paily, 2020). Technology-enhanced environments have shown potential by providing argumentation scaffolds (Huang et al., 2016), yet how to stimulate deep critical thinking through real-time, dynamic conversational interactions remains underdeveloped.

Pedagogical agents based on large language models (LLMs) have pioneered adaptive, personalized debate dialogues (Winkler & Söllner, 2018). Park and Seo (2025) developed a multi-agent LLM-based debate chatbot that achieved strong alignment with human evaluators in assessing critical thinking. Concurrently, Tang et al. (2025) showed that generative AI can serve as an effective dialogic partner, supporting students’ reasoning and argumentation skills. Research has shown that chatbot-assisted debates can enhance students’ argumentation skills and motivation (Guo et al., 2023). However, the critical design variable of how an AI agent’s systematic, real-time conversational role influences critical thinking remains under-explored.

To address this gap, this study investigates the differential effects of three distinct AI conversational roles—supporter, arguer, and comprehensive debater—on learners’ critical thinking. Specifically, this study seeks to answer:

RQ1: Does interaction with an AI debate agent improve self-perceived critical thinking, and does the magnitude of improvement vary by agent role (Supporter, Arguer, Comprehensive Debater)?

RQ2: Does such interaction improve learners’ written argumentation quality, and does this effect vary by agent role?

2. Methods

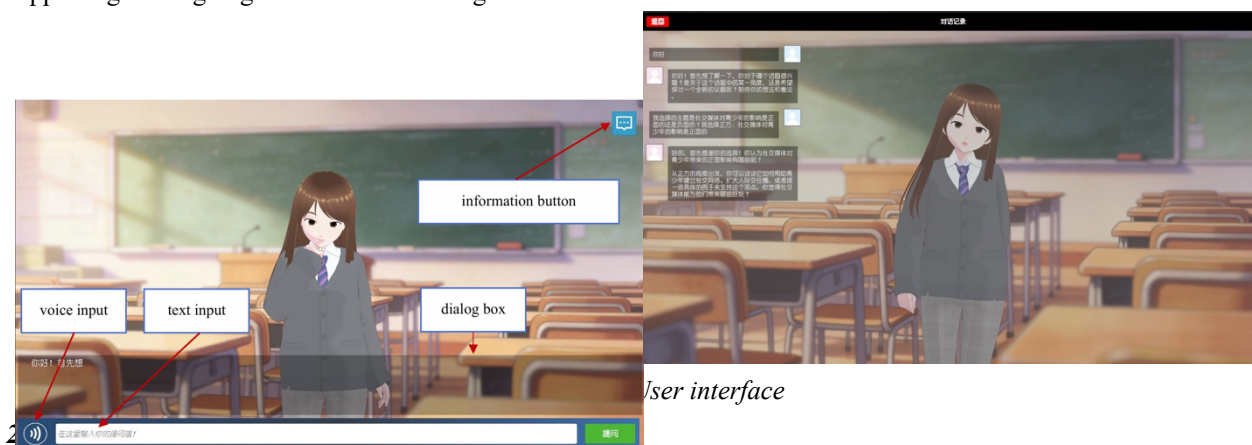
2.1. Participants

Thirty university students (7 males, 23 females, aged 18-28) were recruited. All volunteered, provided informed consent, and received an honorarium. They were randomly assigned to one of three groups (n = 10 each): Supporter, Arguer, and Comprehensive Debater.

2.2. The AI Debate Agent System and Role Design

We developed an AI debate agent system built on the Unity engine, powered by the Qwen2.5-3B LLM and integrated with Microsoft Azure Speech Services. It features an embodied character with synchronized speech, lip movements, and body language for enhanced social presence and immersion. All debate dialogues are systematically recorded (Fig.1).

The three distinct agent roles were designed using the CRISPE prompt framework (Liu et al., 2025): Supporter bolsters the participant's statements with examples or theoretical perspectives; Arguer challenges the participant's statements through rebuttals or contradictory evidence; Comprehensive Debater dynamically switches between supporting and arguing roles based on dialogue context.



This study employed a randomized pretest-posttest intergroup design. The agent role was the independent variable, with argumentative writing scores and self-assessed critical thinking as dependent variables. The experiment was conducted in a single 50-minute session (Fig.2). The procedure consists of four phases: (1) Learning Phase: Participants studied materials on argument structure and the Toulmin model (El Majidi, Janssen, & de Graaff, 2021) to ensure a common understanding of debate frameworks, (2) Pre-test Phase: Participants completed a demographic questionnaire and a 5-item Critical Thinking Disposition Scale. They then wrote a 300-word argumentative essay on a pre-assigned topic, (3) Dialogue Phase: Participants engaged in a debate with their assigned AI agent. They chose a topic from five predefined social issues, all balanced for complexity. The conversation continued until the participant signaled its end. The system recorded all interactions, (4) Post-test Phase: Participants completed a 12-item Critical Thinking Self-Assessment Scale and wrote a second argumentative essay on a different, counterbalanced topic.

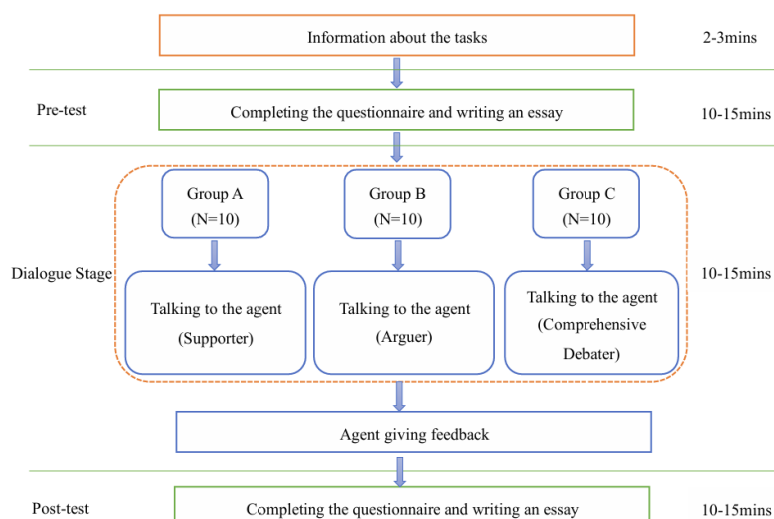


Fig.2. The experimental procedure

2.4. Measuring Tools

Three instruments were used: (1) Argumentative Essay Writing Task adapted from GRE “Analyze an Issue” tasks for pre- and post-tests with counterbalanced topics; (2) Critical Thinking Disposition Scale (pre-test, 5 items) measuring

baseline disposition; (3) Critical Thinking Self-Assessment Scale (post-test, 12 items) measuring self-perceived skills. All self-report scales employed a 5-point Likert scale.

Essay quality was assessed by two blinded raters using a 5-point scale across seven dimensions (organization, adequacy, elaboration, etc.) adapted from El Majidi et al. (2021). Inter-rater reliability was good (ICC = 0.77 and 0.78 for pre- and post-test).

2.5. Data analysis

Due to the small sample size ($n = 10$ per group), nonparametric tests were used in SPSS 27.0. Kruskal-Wallis H tests examined between-group differences at baseline and post-test. Wilcoxon signed-rank tests compared within-group changes from pre- to post-test. The significance level was set at $p < .05$.

3. Results

Baseline homogeneity was confirmed: no significant differences among groups in critical thinking disposition ($\chi^2 (2) = 0.25, p = .881$) or argumentative writing scores ($\chi^2 (2) = 0.65, p = .722$).

3.1. The Impact of Different AI Agent Roles on Self-Perceived Critical Thinking (RQ1)

First, Kruskal-Wallis H tests revealed no statistically significant differences among the three agent roles on any post-test outcomes. This includes self-perceived critical thinking and argumentative writing scores (all $p > .05$). Complete results are presented in Table 1.

Second, Wilcoxon signed-rank tests on the full sample indicated that interaction with an AI debate agent led to a significant overall enhancement in learners' self-perceived critical thinking ($Z = -4.79, p < .001$; see Table 2).

Table 1. Enhancement in Self-Perceived Critical Thinking and the Effect of Agent Roles

Dimension	M $\pm$ SD	χ^2 (df = 2)	p
Self-Perceived Critical Thinking	48.33 $\pm$ 5.89	3.95	0.139
Argumentative Writing	23.50 $\pm$ 2.90	1.09	0.579

3.2. The Impact of Different AI Agent Roles on Argumentative Writing Quality (RQ2)

In contrast to the positive subjective outcomes, the interaction did not lead to a significant improvement in objective argumentative writing quality at the overall level ($p = .925$; see Table 2). Furthermore, consistent with RQ1 results, no differential effect by agent role was found for writing quality ($p = .579$).

Table 2. Wilcoxon Signed-Rank Test Results for Key Variables (Within-Subject Comparisons, Full Sample)

Dimension	Pre-test (M $\pm$ SD)	Post-test (M $\pm$ SD)	Z	p
Self-Perceived Critical Thinking	18.83 $\pm$ 2.77	48.33 $\pm$ 5.89	-4.79	<.001
Argumentative Writing	23.48 $\pm$ 2.77	23.50 $\pm$ 2.90	-0.09	0.925

4. Discussion and Conclusion

This study offers two key insights. First, a single interaction with any AI debate agent significantly enhanced learners' self-perceived critical thinking, with no significant differences among the three roles. This highlights AI-driven dialogue's potential to boost learner confidence and engagement in debate contexts.

Second, despite gains in self-perception, no improvement was observed in objective argumentation quality. This divergence underscores that complex argumentation skills require prolonged, iterative practice (Kellogg & Raulerson, 2007). It also reveals a critical gap between confidence and actual proficiency, suggesting self-report measures alone are insufficient for evaluating skill development.

From a theoretical perspective, these findings suggest that the pedagogical role of an AI agent—whether supportive, challenging, or adaptive—may not produce differential effects in short-term interactions, but such effects could emerge

over extended engagement. From a practical standpoint, for single-session activities aimed at enhancing engagement, any AI debate role proves effective; however, fostering long-term argumentation skill development requires multi-iteration interventions with structured writing feedback.

This study has limitations, including its single-intervention design, small sample, and reliance on self-reported data. Future research should: (1) adopt longitudinal designs to explore cumulative effects of different roles, (2) analyze dialogue processes combined with physiological data to understand cognitive and emotional influences, and (3) examine matching effects between learner individual differences and AI roles.

Acknowledgements

This research was supported by the New Engineering New Form Textbook Construction Project of Hubei Provincial Department of Education (No. XGK02019) and the China University Industry-University-Research Innovation Fund (No. 2024XL009).

References

- El Majidi, A., Janssen, D., & de Graaff, R. (2021). The effects of in-class debates on argumentation skills in second language education. *System*, 101, 102576.
- Giri, V., & Paily, M. U. (2020). Effect of scientific argumentation on the development of critical thinking. *Science & Education*, 29, 673–690.
- Guo, K., Zhong, Y., Li, D., & Chu, S. K. W. (2023). Effects of chatbot-assisted in-class debates on students' argumentation skills and task motivation. *Computers & Education*, 203, Article 104862.
- Huang, C. J., Chang, S. C., Chen, H. M., Tseng, J. H., & Chien, S. Y. (2016). A group intelligence-based asynchronous argumentation learning-assistance platform. *Interactive Learning Environments*, 24(7), 1408–1427.
- Kellogg, R. T., & Raulerson, B. A. (2007). Improving the writing skills of college students. *Psychonomic Bulletin & Review*, 14, 237–242.
- Liu, Z. S., Quan, L., Wang, Y. G., et al. (2025). Impact of prompt frameworks on user search behavior in conversational search. *Information Studies: Theory & Application*, 48(11), 28–40. (in Chinese)
- Park, B., & Seo, K. (2025). Assessing critical thinking through a multi-agent LLM-based debate chatbot. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*, Article 521. New York, NY: ACM.
- Tang, K. S., Cooper, G., & Kendeou, P. (2025). Generative AI as a dialogic partner: Enhancing multiple perspectives, reasoning, and argumentation in science education with customized chatbots. *Journal of Science Education and Technology*, 34(2), 215–232.
- Winkler, R., & Söllner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. In *Academy of Management Proceedings* (Vol. 2018, No. 1, p. 15903). Briarcliff Manor, NY: Academy of Management.

InspireRV: Namecard-Sized RISC-V Computer for Foundational Digital Literacy and Computational Thinking

Vincent man-kin LAU

Technological and Higher Education Institute of Hong Kong, THEi

*drlau@thei.edu.hk

Abstract: *Most students today use heavily on personal computers, tablets, and smartphones, yet few understand the foundational principles governing digital computation—how binary values represent instructions, how machine code corresponds to processor operations, or how high-level languages translate into sequences of zeros and ones. This disconnect limits digital literacy and limits innovation and breakthrough.*

This paper presents InspireRV, an innovative namecard-sized (55×85 mm) RISC-V-based edutainment computer designed specifically to teach foundational concepts of computation through direct manipulation of machine code, for students starting from around 12 years old. Unlike conventional coding tools, InspireRV allows students to physically enter binary instruction bits using an integrated 64-LED/64-button clickable display, visualize execution in real time, and explore RISC-V operations through games, animations, robotics, and small-scale deep-learning experimentation.

The device functions without a PC or tablet or internet, lowering access barriers for schools and workshops. This work describes the system's design, educational rationale, implementation of RISC-V instruction interpretation, extension modules (OLED disassembler instruction display, robotic base, deep-learning digit recognizer), and the pedagogical model supporting its deployment in secondary and tertiary education. InspireRV aims to ignite conceptual curiosity, strengthen digital literacy, and inspire future innovators to explore processor architecture and computational engineering.

Keywords: RISC-V, computational thinking, instruction set, processor, tangible computing.

1. Introduction

Digital technologies affect all aspects of modern life. Despite daily interaction with computers and smartphones, most users remain unaware of how software becomes 0 and 1, how binary instructions control processors, or how instruction sets operate at the lowest levels. This widening “computational literacy gap” limits the public's ability to critically understand the news of chips and software, digital systems, constrains the pipeline of students pursuing deep computing disciplines, and creates barriers to innovation in computer architecture.

Conventional computing education focuses primarily on high level programming languages or graphical interfaces, leaving machine code and processor design as niche topics restricted to advanced university courses. Yet emerging open architectures such as RISC V offer new opportunities for educational democratization: because the instruction set is open, modifiable, and free, it provides an authentic platform for students to understand real-world processor behavior.

To address this educational need, we introduce InspireRV platform (fig 1), which is based on a compact, namecard sized educational computer designed to teach the fundamentals of binary representation and RISC V machine instructions through hands on, playful experimentation. The device emphasizes physical interaction, tactile learning, and immediate visual and audio feedback, transforming abstract concepts into intuitive experiences. Students can toggle instruction bits, observe the effect on display patterns, movement, or sound, and connect these actions to processor operations.

This paper presents InspireRV's system architecture, educational purpose, hardware design, embedded interpreter, and extension capabilities, and positions it as a novel platform for computational literacy and STEM engagement.

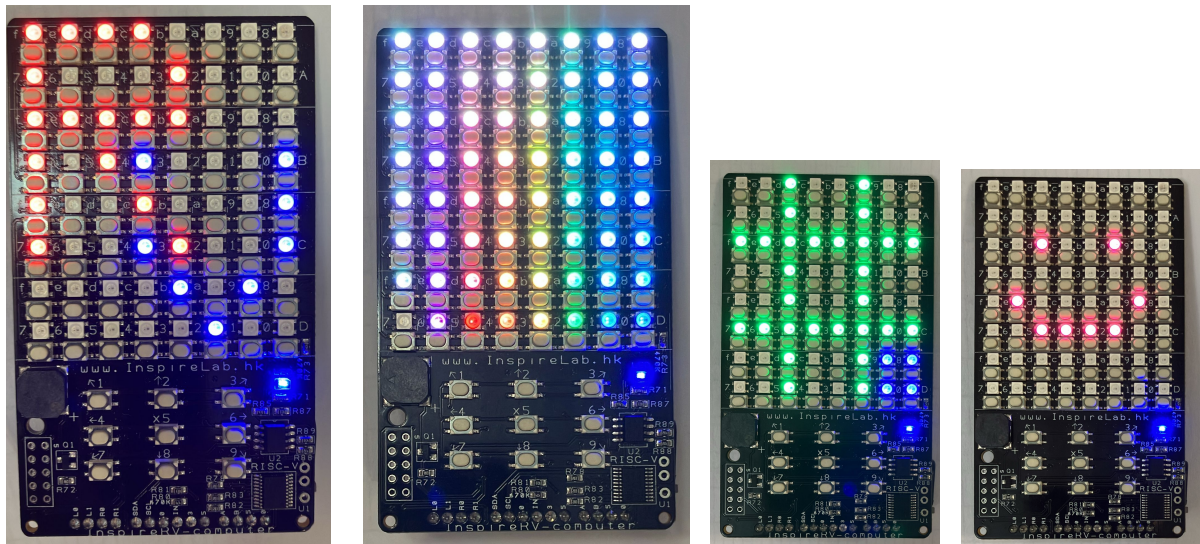


Fig 1. Namecard size computer with clickable color display

2. Background and Related Work

2.1 Digital Literacy and Foundational Computing Education

Prior research highlights the increasing need for foundational digital understanding among students. While tools like Scratch, Micro:bit, and Arduino help students learn programming at various levels, few systems provide exposure to machine-code-level operations. Foundational knowledge of binary and low-level computing is associated with improved computational thinking and deeper conceptual understanding.

2.2 RISC-V for low level software education

RISC-V, an open-source Instruction Set Architecture (ISA) originating from UC Berkeley, provides a lightweight, modular alternative to proprietary ISAs. Its public availability, simplified base instruction set, and widespread industry adoption make it a suitable platform for education. Educators have begun exploring RISC-V for teaching computer architecture, but typically through FPGA boards or software simulators—systems often too complex for beginners or too resource-heavy for scalable classroom adoption.

2.3 Limitations of Existing Educational Hardware

Existing teaching tools generally fall into two categories: (1) High-level coding environments—easy to use but abstract away machine-level details. (2) FPGA/prototyping boards—powerful but expensive, bulky and need computer interfacing, and requiring expert-level knowledge.

InspireRV bridges these extremes, offering a true processor-based machine-code learning tool that is small (namecard size), accessible, easy to use and inexpensive.

3. System Overview of InspireRV

3.1 Hardware Design

The InspireRV platform integrates the following core components:

- **RISC-V Core Processor**

A low-cost RISC-V System on Chip (SoC) core executes instructions interpreted from user-entered binary sequences, with input and output pins for speaker, LEDs, buttons.

- **32 KB EEPROM Storage**

Capable of storing over 200 colorful graphics, animations, text sequences, or custom binary software programs (machine code software).

- **Clickable Color Display**

A color display with 8 x 8 pixels, each pixel have a clickable button for represent eight row of bytes (each byte with bit 0, bit 1,...,bit 7) or four session of 16 bit (bit 0, bit 1...bit a, bit f). This allow direct manipulation of bit for setting binary machine code instruction, and allow color graphics or pixel arts. The display of **64 full-color LEDs** paired with **64 tactile input buttons**, allowing binary instruction entry, color selection, pixel-level drawing, visual output of computation and execution flow.

- **Speaker**

A small speaker for output different sound frequency and duration, enable one to use machine code to generate various ringtone and sound effects such as increasing frequency and decreasing frequency.

- **Built-in RISC-V Interpreter**

Supports 16 bit instructions such as c.addi, enabling students to enter machine code directly, example of RISC instruction as follows, and let the students to know the design of processor instruction, with concept such as operator and operand, conditional and unconditional branch, arithmetic and logical operators, etc.

3.2 Expansion Modules - Disassembler Module (128×64)

Provides textual output, enabling students to view decoded instructions (e.g., 0x9532 → add a0, a0, a2) and better connect binary to instruction (fig 2).

3.3 Expansion Modules - Deep Learning Module

Provide a digit recognition engine (0–9), let one to draw image on LED panel with buttons (in red color) and inspireRV will recognition the digit (in green color). For more advance education, there is open-source code for exploring fixed-point arithmetic, bit compression, and optimized AI inference.

3.4 Expansion Modules - Robotic Base Module

Allows InspireRV to drive motors (Fig 4) and perform stepwise motion control (such as move forward, left turn, move forward, right turn...). This allow the software to control movement and create “dancing” patterns

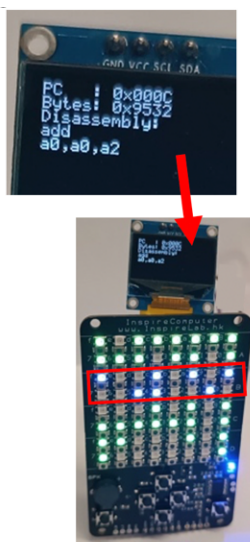


Fig 2. Disassembler Module

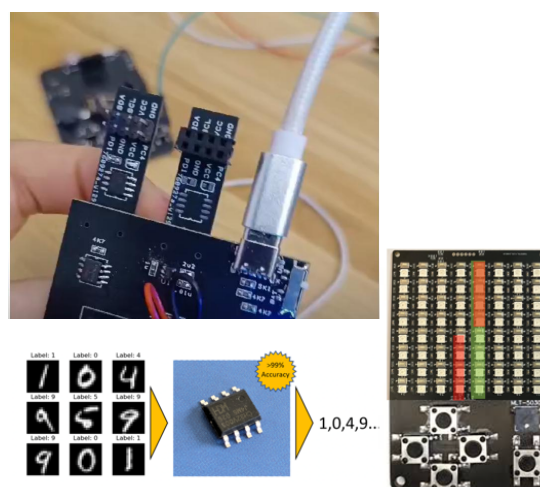


Fig 3. Deep Learning Module

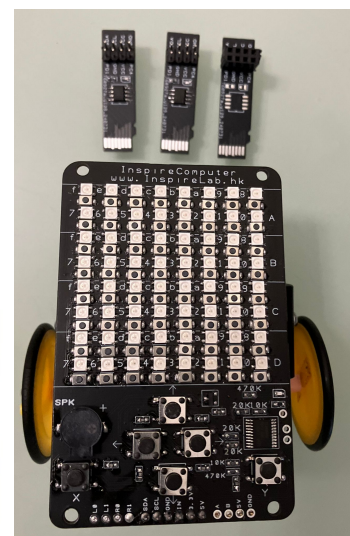


Fig 4. Robotic Base Module

4. Pedagogical Value and Learning Model

The platform enables experiential learning of machine code through direct binary bit manipulation. Students typically encounter binary only in textbooks, and the platform brings binary to life by letting learners: (1) Enter and toggle real machine-code bits. (2) Visualize instruction effects immediately on screen patterns, animation, or sound. (3) Debug by modifying bits. (4) Observe changes as they experiment. This experiential approach supports: (1) constructivist learning, where knowledge is built through exploration. (2) embodied cognition, connecting tactile actions to conceptual understanding. (3) incremental abstraction, moving from physical bits → instruction → higher-level concepts. And the platform can democratize access to computer architecture as previously this is introduced in University courses. Because no personal computer or tablet is required, InspireRV facilitates truly mobile, hands-on technology workshop. The small size, low cost, and battery powered platform enable deployment such as educational workshops with hundreds of students, such as Hour of Code for RISC-V, STEM workshop, ArtsTech programs.

6. Evaluation and Educational Deployment Plan

Some pilot workshops with HK secondary school students was done, showed the platform can arose the student interest in learning the foundation of computer, and more flexible in venue such as classroom without personal computer and WiFi. In future we will organize and evaluate the effectiveness through events such as secondary school coding camp, University introductory digital-design courses and public STEM festivals. Evaluation metrics will include student engagement, concept comprehension (pre/post tests), interest in pursuing deeper computer engineering studies

7. Conclusion

InspireRV introduces a new class of educational technology: a namecard-sized machine-code learning computer that blends RISC-V open architecture with tactile, visual, and portable device. By lowering the barrier to understanding binary computation, the platform aims to cultivate a new generation of digital-literate students and inspire future innovators in processor architecture and embedded systems. Through its interactive LED/button interface, optional OLED disassembler, robotics capabilities, and embedded deep-learning engine, InspireRV provides a rich ecosystem for exploring computation from first principles. This work presents InspireRV as a scalable, accessible, and powerful tool for enhancing foundational computing education across secondary school and tertiary school levels.

References

- Armoni, M. (2013). On Teaching Abstraction in Computer Science to Novices. *Journal of Computers in Mathematics and Science Teaching*, 32(3), 265–284. <https://www.learntechlib.org/p/41271/>
- Hansen, A., Hansen, E., Dwyer, H., Harlow, D., & Franklin, D. (2016). Differentiating for Diversity: Using Universal Design for Learning in Elementary Computer Science Education. *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, pp.376–381.
- Repenning, A., Webb, D. C., Koh, K. H., Nickerson, H., Miller, S. B., Brand, C., Horses, I. H. M., Basawapatna, A., Gluck, F., Grover, R., Gutierrez, K., & Repenning, N. (2015). Scalable Game Design: A Strategy to Bring Systemic Computer Science Education to Schools through Game Design and Simulation Creation. *ACM Transactions on Computing Education*, 15(2), 1–31. <https://doi.org/10.1145/270051>

Student Engagement and Data Science Skills during Authentic Collaborative

Inquiry: Insights from Contrasting Group Analysis

Gaoxia Zhu¹, Chin Lee Ker¹, Carol K. K. Chan² and Chew Lee Teo¹

¹Nanyang Technological University

²University of Hong Kong

Abstract: *We have previously developed an authentic, collaborative inquiry model to simulate students' interests using real-world context and data, engage students in data science practice, improve their data science practice through external feedback and finally, trigger them to reflect on the learning experience. This study further delves into how student groups with different data science performances demonstrate data science skills and how they engage in the data science workflow. We analyzed the video recordings of a productive and unproductive group and found that the productive group scoped their investigative questions early, followed and was mostly engaged in the data science flow. In contrast, the unproductive group focused on non-investigative questions and did not realize issues with their inquiry until receiving external feedback, which triggered their revisions and disengagement. This study provides insights into challenges K-12 students could encounter during collaborative data science inquiry and how to support their learning.*

Keywords: Data Science, engagement, collaborative inquiry, K-12, video analysis

1. Introduction

Our society is experiencing big data growth (Vesset et al., 2012), whereby there is a large amount of data around us, highlighting the importance of acquiring data science skills. Hence, developing the future generation's data science skills is critical. Data science skills include generating, collecting, cleaning, managing, analyzing, and communicating data and using the results to inform decisions or provide insights (Overton & Kleinschmit, 2022). Researchers (e.g., Dichev & Dicheva, 2017) emphasized that mastering data science skills will allow us to derive intelligent insights from the upsurge of data in the world. Governments (e.g., The Curriculum Development Council, 2017) have also stressed the importance of performing science inquiry and data science activities, such as conducting investigations, analyzing data, and interpreting data, starting from the primary school level.

Data science components have been integrated into various curriculums. However, most of the literature on data science focuses on introductory courses at the undergraduate level (Thompson & Arastoopour Irgens, 2022). In contrast, K-12 students have limited opportunities to work with data formally in school and they are often not taught how to analyze and interpret contextual, and dynamic real-world data (Finzer, 2013), which may make them see data science as less relevant to them. Therefore, developing authentic data science curricula for secondary students, teaching them data science skills, and engaging them in data science practice is a significant and pressing issue (Heinemann et al., 2018). In a previous study, we (Zhu et al., 2025) developed an authentic, collaborative inquiry model of SPIRE (Stimulate, Practice, Improve and Reflect) and an associated 2-day out-of-school data science program for secondary school students. The program aims to stimulate students' interest by presenting real-world problems, involve them in data science practices and workflows to develop their skills, and promote continuous improvement in data science-supported investigation through feedback, such as input from experts. Additionally, the model emphasizes the importance of reflection, encouraging students to evaluate their learning experiences and address challenges to advance their investigation. Collaboration among students is central to the model and program as it mirrors the way professional data scientists work together to tackle complex problems using data.

Still, acquiring data science skills is a demanding and time-intensive process (English, 2014). In data science education, challenges associated with perceptions of evaluating students, teacher confidence, teacher professional

development, the interdisciplinary and complex nature of data science have been reported (Mike, 2020). These challenges suggest the need to dive into students' data science learning process to understand how student groups with different performing levels engage or disengage in the data science learning process to provide implications for better supporting data science teaching and learning. To address the research gap, this exploratory case study examines how students engage in data science work and what skills are deployed. Specifically, we investigated differences between contrasting groups with the research question: How do productive and unproductive student groups engage in the data science workflow, what skills do they demonstrate, and what are their engagement and disengagement levels?

2. Literature Review

Education frameworks and standards increasingly emphasize data science in K–12, yet key concepts are often introduced late and with limited opportunities for meaningful engagement (Drozda et al., 2022; Finzer, 2013). For instance, the Common Core State Standards for High School: Statistics and Probability (2022) highlight interpreting data, drawing justified inferences, and making data-driven decisions. In practice, however, relevant ideas tend to appear only briefly in later high school (Drozda et al., 2022), and data science is usually embedded superficially within mathematics, science, or computer science rather than as a distinct curriculum strand. Most research also focuses on introductory undergraduate courses (Thompson & Arastoopour Irgens, 2022), leaving limited understanding of how secondary students engage in data science processes and what challenges they face. This gap constrains efforts to design and support deeper learning at this level.

Furthermore, developing data science skills and statistical literacy is challenging due to several reasons (English, 2014). The interdisciplinary nature of data science makes it difficult to design a comprehensive course within limited time, while addressing both domain-specific and data ethics issues and accommodating students' diverse backgrounds (Mike, 2020). Teachers often lack sufficient training in incorporating data into their teaching (Merk et al., 2020). Students typically have limited prior knowledge (Heinemann et al., 2018) and can experience a high cognitive load during the implementation of interdisciplinary data science programs.

Specifically, students may experience cognitive, emotional, behavioral, and social engagement and disengagement during data science investigations (Inel-Ekici & Ekici, 2022). Engagement is widely viewed as a multifaceted construct involving passion and dedication, involvement in learning activities, and effort, resilience, and persistence when facing challenges (Fredricks et al., 2004). Synthesizing prior work, Wong et al. (2024) identified three core dimensions of engagement: cognitive, affective (emotional), and behavioral and found that each is positively associated with academic achievement and subjective well-being. In collaborative settings, a fourth dimension, social engagement, is also emphasized, focusing on interpersonal interaction, caring, and building on others' ideas (Fredricks et al., 2016).

Cognitive engagement occurs when learners are committed, thoughtful, and willing to invest effort in understanding complex concepts and developing challenging skills (Fredricks et al., 2004). Emotional engagement refers to learners' positive or negative reactions during learning, which influence their academic achievement and attendance (Wara et al., 2018). Behavioral engagement involves active participation in learning activities, such as sustained effort and attention (Fredricks et al., 2004). Social engagement, meanwhile, encompasses coordinating efforts and building mutual understanding among group members to solve problems, all of which contribute to the co-construction of knowledge and enhanced learning (Fredricks et al., 2016).

In contrast, cognitive disengagement may arise when students struggle to select effective strategies or develop shared understandings of tasks and content (Barron, 2003). Emotional challenges emerge through disruptive affective responses; such as hopelessness or boredom (Ahmed et al., 2013). Behavioral challenges include loss of focus, reduced contribution, off-task conversations, or shifts to unrelated activities, often occurring when group members lose sight of shared goals (Järvelä & Järvenoja, 2011). Social challenges arise when communication breaks down, when attempts to engage peers are not reciprocated, or when disagreements emerge regarding group organization, workflow, or norms (Hussein, 2021). Research examining students' engagement and disengagement throughout the data science learning process remains

scarce. This highlights the importance of understanding how student groups with different levels of data science skills engage or disengage in data science practices to better support their learning and guide curriculum development.

3. Methods

3.1. Participants and Context

Participants included 43 secondary students ($M_{age}=13.91$, 23 female students) from Singapore and Hong Kong joining a two-day data science program in June 2023. Among them, 24 were in secondary 1 (Grade 7), 4 in secondary 2 (Grade 8), 12 in secondary 3 (Grade 9), and 3 in secondary 4/5 (Grade 10). Participants had different data science backgrounds, some having prior data science experience because of attending Science Club and co-curricular activities in school, whereas some having little to no experience interacting with data.

The participants formed 7 groups based on their shared inquiry interests around the "Data Science and Sustainability" theme, with around five to ten members in each group. The program consisted of four key components designed based on the SPIRE model: 1. **Stimulate** students' interest by connecting data science to real-world challenges and demonstrating its relevance through expert talks and videos. 2. Engage students in **practice** data science workflows through hands-on activities, from questioning to data analysis and interpretation. 3. Help students **improve** and refine questions, analyses, and approaches through expert feedback. 4. Support students to **reflect** metacognitively, deepening understanding and recognizing data science's value.

3.2. Data Collection and Analysis

The data sources for this study mainly include (1) the artifacts/ final products (i.e., slides) produced by the seven student groups (2) the video recordings of two groups (productive and unproductive groups) selected based on the data science skills demonstrated in their final products to illuminate different kinds of student engagement and data-science. The video recordings captured the practice and improvement components when students were working in their small groups. The activities were designed around Formulation of Investigative Questions, Data Search, Data Analysis, Consultation, and Refinement.

We applied the coding scheme we previously designed (Zhu et al., 2025) to analyze the data science skills demonstrated in the final products of all student groups. As shown in Table 1, the coding scheme includes ten skills such as scoping questions and explaining rationales. Each skill includes three levels: 0 suggests that a group is not competent in that skill, 1 indicates that the group is moderately competent, whereas 2 indicates that the group is highly competent. Two raters first independently coded the final products using the coding scheme and then compared and discussed their allocated scores for each product until an agreement was reached.

Table 1

Coding scheme to analyze students' data science skills in the final products

Data science skill	Level 0	Level 1	Level 2
Scope question	Scope was wide, objective was absent	Scope was wide, objective was unclear	Scope was narrow, objective was clear
Explain rationale	Explanation was absent	Explanation was partial	Explanation was complete
Identify variables	Identification was absent Unable to find necessary	Identification was vague	Identification was clear
Data search	data	Found data online	Collected own data
Align data with questions	Data is unrelated to any variable	Data is partially related to some variables Analysis was not	Data is completely related to all variables
Appropriate analysis	Analysis was absent	appropriate	Analysis was appropriate

			Visualizations were created and statistical analyses were performed
Comprehensive analysis	Analysis was absent	Visualizations were created	
Interpret results	No interpretation	Interpretation is partially accurate and relevant	Interpretation is accurate and relevant
Answer the question	Answer was absent	Answer did not relate to analyses conducted	Answer was related to analyses conducted
Consider other factors	Unaware of other factors	Aware that there is one relevant factor	Aware that there are more than one relevant factors

Among the seven groups, Groups 1 and 6 achieved the highest data science score (16 points), followed by Group 5 (13 points), Group 7 (8 points), Group 2 (7 points), and Groups 3 and 4 (6 points). We selected Group 6 to present the productive group and Group 2 to represent the unproductive group. Group 2 was selected over groups 3 and 4 as their data science processes were fully video recorded in the two-day data science program.

We adapted the group collaborative learning analysis method outlined by Järvenoja et al. (2019) to examine the video data of the productive and unproductive groups. First, a researcher divided the recordings into 30-second segments to organize the analysis and preserve a clear chronological sequence of group interactions (Sinha et al., 2015). A brief summary of events was written for each segment. Next, each segment was coded in terms of group-level engagement, mixed engagement, or disengagement. Each interval was also coded in terms of what data science skills were demonstrated, or “Unsure” if it was difficult to tell the specific data science skill being demonstrated (e.g., typing out a note on Knowledge Forum), or “Unrelated” if off-task activities (e.g., turning on the laptop, discussing events not related to the investigation) were observed. If an utterance did not occur, the researcher used other cues, such as behavior data to code whether students were engaged or not.

4. Results

4.1. Data Science Skills Demonstrated in the Final Products

We first describe the final products of the contrasting groups to illustrate different kinds of data-science skills and engagement. As shown in Table 2, the coding results show that the productive group showed high-level data science skills in all the other categories except for explaining the rationales of the investigation and considering other factors beyond those included in their investigative questions. In contrast, the unproductive group showed low to medium levels of data science skills in all the ten categories. The results suggested the need to investigate their data science learning process to understand what might have contributed to the difference.

Table 2

Data science skills demonstrated by the two small groups in the final products

	Scope questio ns	Explain rational es	Identify variable s	Data searc h	Align with questions	Appropri ate analysis	Comprehe nsive analysis	Interpr et results	Answer the question	Consider other factors
U										
G 1	1	1	1	1	1	0	0	1	1	0
P										
G 2	0	0	2	2	2	2	2	2	2	0

Note. UG: Unproductive group; PG: Productive group

4.2. Data Science Skills Deployed and Engagement or Disengagement in Video Analysis

Figures 1 and 2 show the engagement, disengagement, and mixed engagement frequencies (x-axis, represented

by colors) as well as data science skills (y-axis) across five major activities designed based on the SPIRE model. The productive group was engaged in the data science workflow most of the time except that they had some disengagement in scoping investigative questions and data search in the beginning. In contrast, it was observed that the unproductive group had higher frequencies of disengagement than the productive group, and all the disengagement was observed in the Consultation, and Refinement phases after external feedback was provided by the assigned experts. Furthermore, the unproductive group did not follow the data science workflow as compared to the productive group, and they went back and forth among various data science skills to rationalize and support their investigation in the Consultation and Refinement phases.

As shown in Figure 1, the productive group started the data science process by focusing on Scoping Investigative Questions, Explaining the Rationale of the Investigation, and performing Data Search. They discussed the possibility of studying how AI could be used to reduce pollution and clean up oceans, considering that trash is released into oceans every day. They briefly searched for information online regarding statistics on ocean pollution. Towards the end of the first phase, the group found that their question was not very investigative and too broad, leading to a few instances of disengagement and mixed engagement. They also realized that the data they needed regarding the use of AI in oceans could not be found online. Upon realizing that it was also impossible to invent AI robots to collect ocean trash within the program, they reframed and narrowed the scope of their questions in the second phase. A group member suggested that since there were different AI characters online, they could chat with each character and find out how long it takes for each AI character to forget what they had originally said to them. They revised their investigative questions as “How many messages does it take for an AI to forget what you said?” and “Does the popularity of the AI affect its memory?”. They briefly mentioned how their research could relate to sustainability. The group identified their variables as “number of messages”, “type of AI character” and “popularity of AI character”. Being aware there was no existing data to answer their questions, the group drafted a data collection plan and began collecting data in experimental trials, where each member talked to different AI characters. In phases Data Analysis to Refinement, the group spoke to different AI characters and tested their memory using two different methods—asking the AI a question or changing the plot suddenly. The group was heavily engaged in Data Search and Appropriate Data Analysis but also had a few instances of mixed engagement, for example, when a group member got too excited updating his fellow group mate about his interaction with the AI character, leading to brief distractions in both members; another group member expressed defeat, showed doubts and uncertainty in plotting a graph.

Figure 2 shows the analysis results of the unproductive group. The group started the data science process by scoping their research question concerning the mechanism of a solar-powered electric car, emphasizing the rationale of their investigation, and searching for information online about solar panels. They were largely engaged in the first three phases but did not engage in appropriate data science skills in the Data Analysis stage, when they were expected to do so. Then, during the Consultation phase, the group realized that their question was not investigative after an expert provided feedback. Even though there were attempts to narrow the scope of the investigative questions, the group did not succeed in doing so nor identified existing data consisting of the relevant variables. Still, their finalized question, “How to make electric cars use sustainable energy?” was not solvable using data. During the Refinement phase, they spent a large portion of time creating and filming a skit to emphasize the rationale of their investigation and researched the mechanism of solar-paneled cars, but they had challenges engaging with Appropriate Data Analysis, resulting in a large frequency of disengagement.

Figure 1

Data science skills and engagement or disengagement across the workflow by the productive group

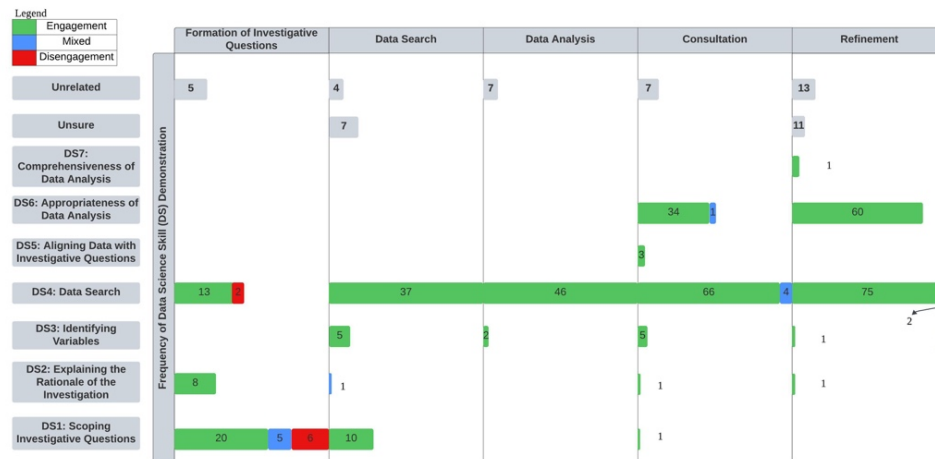
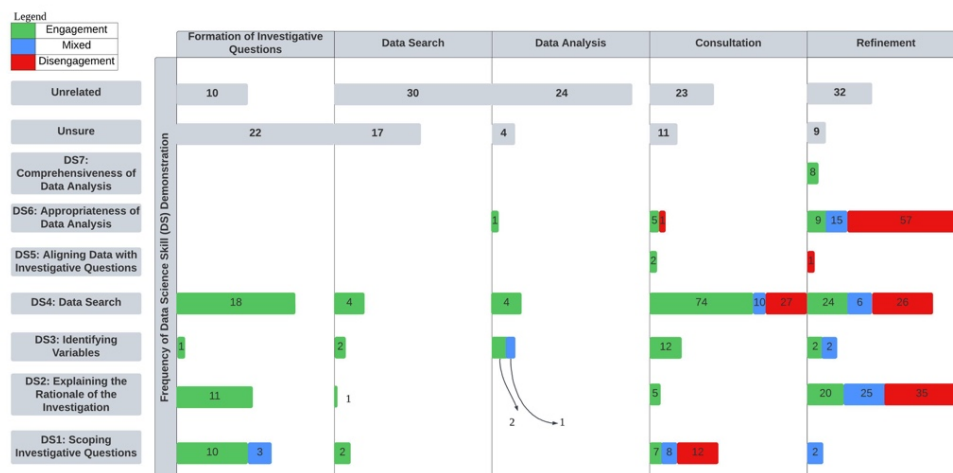


Figure 2

Data science skills and engagement or disengagement across the workflow by the unproductive group



5. Discussion and Conclusion

Considering the importance of cultivating students' data science skills in the current society and the challenges associated with data science education, this study used contrasting group analysis to examine how high- and unproductive groups engaged in data-science workflow and deployed data science skills. The in-depth qualitative analysis and mapping of data science skills and engagement (or disengagement) over the data science workflow uncover what challenges secondary school students can encounter and provide implications for support that should be provided during data science learning.

We found that both the unproductive and productive groups had challenges in narrowing the scope of their investigative questions. Having non-investigative research questions (e.g., too broad, does not direct data analysis) is a typical problem encountered in data science learning (Ratan et al., 2019). A good investigative question should not be too broad nor too narrow and must be researchable using data science (Ratan et al., 2019). With similar challenges in the beginning, however, the two groups eventually performed differently. The following reasons could have explained the difference in performance. Firstly, the productive group had the patience to navigate through the challenge, brainstormed and discussed their other ideas, and changed their robot-design question to investigative questions that could be answered using data and were more feasible to work on within two days. In contrast, the unproductive group was very excited and kickstarted the information search without pausing to reflect on whether their question was investigative, even though reflection was emphasized in the SPIRE mode (Zhu et al., 2025) and in our program design. The group searched for

information related to the mechanisms of an electric car, as well as the current statistics of electric car users. Secondly, the productive group's practice is more aligned with the data science workflow, which provides their work structure and sequences and ensures a smoother and more comprehensive investigation (Gupta, 2022). The unproductive group, on the other hand, spent too much time trying to explain the rationale of using electric cars but did not engage in appropriate data analysis until late phases, and they had a high frequency of disengagement.

The analysis of the unproductive group also shows the importance of providing external feedback to help students identify issues and limitations with their data science investigation, which may not be explicitly observable to students themselves. Similarly, it has been found that it is typically challenging for secondary school students to source and search for information online due to their limited skills in adopting appropriate keywords and identifying relevant information (Wu et al., 2014). Monitoring and feedback from instructors or capable others should be provided earlier to keep students on track of the desired data science workflow before it is too late.

This study used two contrasting groups and conducted learning products and video analyses to provide in-depth analysis and insights. However, there are some limitations. Firstly, our sample size is small, making it difficult to generalize the findings. In the future, we plan to extend the analysis to other groups to examine whether similar findings can be found and if new patterns will emerge. Secondly, this study was conducted in a two-day out-of-school setting. The extent to which the findings can be generalized to school contexts needs further investigation.

References

- Ahmed, W., van der Werf, G., Kuyper, H., & Minnaert, A. (2013). Emotions, self-regulated learning, and achievement in mathematics: A growth curve analysis. *Journal of Educational Psychology*, 105(1), 150–161.
- Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences*, 12(3), 307–359.
- Common Core Standards. (2022). "Statistics & probability." Statistics & Probability | Common Core State Standards Initiative. <http://www.corestandards.org/Math/Content/SP/>
- The Curriculum Development Council (2017). *Science education key learning area curriculum guide (Primary 1 - secondary 6)*. https://www.edb.gov.hk/attachment/en/curriculum-development/kla/science-edu/SEKLACG_ENG_2017.pdf
- Dichev, C., & Dicheva, D. (2017). Towards data science literacy. *Procedia Computer Science*, 108, 2151–2160.
- Drozda, Z., Johnstone, D., & Van Horne, B. (2022). *Previewing the national landscape of K-12 data science implementation*. National Academies of Science, Engineering, and Medicine.
- English, L. D. (2014). Promoting statistical literacy through data modelling in the early school years. In Chernoff, E.J. & Sriraman, B. (Eds.), *Probabilistic thinking: Presenting plural perspectives*. Advances in Mathematics Education (pp. 441–457). Springer.
- Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education*, 7(2), 1–9.
- Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1), 59–109.
- Fredricks, J. A., Filsecker, M., & Lawson, M. A. (2016). Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues. *Learning and Instruction*, 43, 1–4.
- Gupta, S. (2022, May 16). Data Science Process: A Beginner's Guide in Plain English. Springboard. <https://www.springboard.com/blog/data-science/data-science-process/>
- Heinemann, B., Opel, S., Budde, L., Schulte, C., Frischmeier, D., Biehler, R., ... & Wassong, T. (2018, November). Drafting a data science curriculum for secondary schools. In *Proceedings of the 18th Koli calling international conference on computing education research* (pp. 1–5).
- Hussein, B. (2021). Addressing collaboration challenges in project-based learning: The student's perspective. *Education Sciences*, 11(8), 434–.

- Inel-Ekici, D., & Ekici, M. (2022). Mobile inquiry and inquiry-based science learning in higher education: advantages, challenges, and attitudes. *Asia Pacific Education Review*, 23(3), 427–444.
- Järvelä, S., & Järvenoja, H. (2011). Socially constructed self-regulated learning and motivation regulation in collaborative learning groups. *Teachers College Record*, 113 (2), 350– 374.
- Järvenoja, H., Näykki, P., & Törmänen, T. (2019). Emotional regulation in collaborative learning: When do higher education students activate group level regulation in the face of challenges? *Studies in Higher Education*, 44(10), 1747-1757.
- Mike, K. (2020, August). Data science education: Curriculum and pedagogy. In *Proceedings of the 2020 ACM conference on international computing education research* (pp. 324-325). ACM.
- Merk, S., Poindl, S., Wurster, S., & Bohl, T. (2020). Fostering aspects of pre-service teachers' data literacy: Results of a randomized controlled trial. *Teaching and Teacher Education*, 91, 103043-.
- Overton, M., & Kleinschmit, S. (2022). Data science literacy: Toward a philosophy of accessible and adaptable data science skill development in public administration programs. *Teaching Public Administration*, 40(3), 354-365.
- Ratan, S., Anand, T., & Ratan, J. (2019). Formulation of research question - Stepwise approach. *Journal of Indian Association of Pediatric Surgeons*, 24(1), 15–20.
- Sinha, S., Rogat, T. K., Adams-Wiggins, K. R., & Hmelo-Silver, C. E. (2015). Collaborative group engagement in a computer-supported inquiry learning environment. *International Journal of Computer-Supported Collaborative Learning*, 10, 273-307.
- Wara, E., Aloka, P. J., & Odongo, B. C. (2018). Relationship between emotional engagement and academic achievement among Kenyan secondary school students. *Academic Journal of Interdisciplinary Studies*, 7(1), 107-118.
- Wong, Z. Y., Liem, G. A. D., Chan, M., & Datu, J. A. D. (2024). Student engagement and its association with academic achievement and subjective well-being: A systematic review and meta-analysis. *Journal of Educational Psychology*, 116(1), 48–75.
- Thompson, J., & Arastoopour Irgens, G. (2022). Data detectives: A data science program for middle grade learners. *Journal of Statistics and Data Science Education*, 30(1), 29-38.
- Vesset, D., Woo, B., Morris, H. D., Villars, R. L., Little, G., Bozman, J. S., ... & Yezhkova, N. (2012). Worldwide big data technology and services 2012–2015 forecast. *IDC report*, 233485.
- Wu, C. H., Hwang, G. J., & Kuo, F. R. (2014). Collab-Analyzer: An environment for conducting web-based collaborative learning activities and analyzing students' information-searching behaviors. *Australasian Journal of Educational Technology*, 30(3). 356–374.
- Zhu, G., Teo, C. L., Ong, A. K. K., Yuan, K. G., Ker, C. L., & Yang, Y. (2025). Towards an authentic collaborative inquiry model for cultivating data science skills and attitudes: effects of SPIRE on secondary school students. *Education and Information Technologies*, 30(14), 19933-19959.

Teacher Perspectives on AI Integration: Pedagogies, Technologies, Confidence, and Readiness Across Disciplines

William Ko-Wai Tang^{1*}, Ao Zhang²,

^{1,2} School of Education and Languages, Hong Kong Metropolitan University

*wtang@hkmu.edu.hk

Abstract: Artificial intelligence (AI) is transforming educational practice. However, little attention is known about how teachers can systematically implement AI through specific pedagogical approaches across different educational settings and disciplines. This study reviews 119 peer-reviewed publications to examine four critical dimensions: pedagogical approaches to AI use, AI technology integration patterns across disciplines, teachers' AI self-efficacy, and teachers' readiness for AI adoption. This study reveals that effective AI integration involves multiple instructional strategies, including personalized learning, project-based learning, support for self-regulated learning, and innovative assessment approaches. These pedagogies differ substantially across educational contexts and disciplines, which means that a one-size-fits-all approach may not work. AI technologies in classrooms take three primary forms: generative AI chatbots, intelligent tutoring systems, and AI-powered mobile learning platforms. Teachers' confidence in using AI varies by discipline, Science, Technology, Engineering and Mathematics (STEM) teachers reporting higher self-efficacy. Regarding the teacher readiness, it involves several connected aspects such as acceptance, attitudes, awareness, and confidence. These areas support one another and grow stronger through professional development that encourages hands-on practice, mentoring, and thoughtful reflection on ethics. The study highlights that successful AI integration requires comprehensive, discipline-responsive professional development addressing technical skills, pedagogical knowledge, ethical reasoning, and the emotional dimensions of professional change. This study provides educators with guidance on suitable AI pedagogical strategies, tools, and assessment methods for AI-enhanced education, offering research insights and future directions for teachers' professional development.

Keywords: Artificial Intelligence, pedagogical approaches, self-efficacy, professional development, TPACK

1. Introduction

Artificial intelligence includes technological systems designed to simulate human cognitive processes, solve complex problems, and enable adaptive learning across domains (Gosh & Thirugnanam, 2021). The introduction of ChatGPT in 2022 catalyzed substantial interest among educators and researchers in AI applications for education. Research demonstrates that AI and machine learning enhance teaching and learning efficiency by optimizing instructional design and improving measurable student outcomes (Holmes et al., 2019). In education, AI streamlines administrative tasks such as grading and data analysis, allowing educators to focus more on teaching (Tang, 2025). AI systems also increase student engagement through immersive, interactive, and gamified learning experiences, particularly when combined with virtual and augmented reality technologies (Chan, 2025).

While AI shows its potential for improving teaching, most current research focuses on what AI can do technically rather than exploring the specific ways teachers should actually teach when using these tools in their classrooms. Research has not examined how teachers adapt fundamental instructional strategies. Furthermore, while diverse AI technologies are entering educational practice, little is known about how technology adoption and implementation patterns vary across academic disciplines and educational levels. AI applications may be highly appropriate for computational thinking in computer science. Current research fails to compare how different disciplinary teachers select and implement AI for teaching. Also, teacher self-efficacy represents a well-established predictor of instructional quality, yet current research provides minimal insight into how teachers' confidence in teaching with AI varies across subject matters. Finally, teacher

readiness for AI integration brings together several closely connected elements, including acceptance, attitudes, ethical awareness, and confidence, but existing studies tend to treat these aspects separately. There is limited information on a systematic understanding of current practices and challenges. This study addresses this by examining pedagogy, AI technology, self-efficacy and readiness. It targets unexplored areas includes teaching practice with AI and subject specific pattern. To address these gaps and provide a holistic view of AI's role in competency-based teaching, this study poses the following research questions:

RQ1 What pedagogical approaches are teachers using when implementing AI in classroom teaching?

RQ2: How are specific AI technologies being integrated into classroom practice across different disciplines and educational levels?

RQ3: How does AI teaching self-efficacy vary across subject matters, and how is this reflected in teachers' classroom practice?

RQ4: How ready are teachers to integrate AI into their teaching?

Through a systematic review of recent empirical studies, this research maps current AI applications in educational practice from teachers' perspectives, focusing on pedagogical approaches, patterns of technology integration, self-efficacy in teaching with AI, and overall teacher readiness. The synthesis of these dimensions offers a contemporary overview that informs future professional development initiatives and guides the ethical implementation of AI in education.

2. Methods

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines using a four-phase process: (1) searching and filtering qualified articles in the databases, (2) conducting full-text reviews of eligible articles, (3) coding the articles based on the research questions, and (4) analyzing and interpreting the results. The article selection prioritized high-quality research by drawing from the Scopus and Web of Science citation databases, and Table 1 presents the search strings used in these databases.

Table 1.

Searching strings in databases

Database	Search strings
Scopus	(TITLE-ABS-Keywords("Artificial Intelligence" OR "AI") AND TITLE-ABS-Keywords ("preschool" OR "kindergarten" OR " elementary school" OR "high school" OR "K-12" OR "university" OR "higher education" OR "students") AND TITLE-ABS-keywords("AI tools" OR "education technology" OR "classroom technology" OR "educational tools") AND ABS ("practice" OR "assessment" OR "curriculum" OR "method" OR "teaching" OR "learning" OR "pedagogy" OR "management" OR "instructional") AND ABS("model" OR "framework" OR "theoretical" OR "foundation" OR "conceptual" OR "paradigm" OR "principle" OR "approach"))
Web of Science	"Artificial Intelligence" OR "AI" (Title) AND "kindergarten" OR "preschool" OR "K-12" OR "university" OR "higher education" OR "students" (Abstract) AND "AI tools" OR "education technology" OR "classroom technology" OR "educational tools" (Abstract) AND "practice" OR "assessment" OR "curriculum" OR "method" OR "teaching" OR "learning" OR "pedagogies" OR "management" OR "instructional" (Abstract) AND "model" OR "framework" OR "theoretical" OR "foundation" OR "conceptual" OR "paradigm" OR "principle" OR "approach" (Abstract)

The initial search identified 1,332 records in Web of Science and 537 records in Scopus, yielding a total of 1,869 articles retrieved on December 31, 2025 using the predefined search strings. Inclusion and exclusion criteria were then applied to enhance the generalizability of the findings and reduce bias (Table 2), which involved excluding review articles, technical development papers, surveys, and studies in which AI was not used as an innovative educational tool, model, or pedagogy in areas such as curriculum or assessment. In total, 119 articles proceeded to full-text screening from the initial pool after 1,720 exclusions, and the overall selection process is summarized in the PRISMA diagram (Figure 1).

Two coders created a preliminary three-tiered framework that includes (1) AI tools for teaching and learning; (2) pedagogical strategies for AI integration; and (3) the incorporation of AI theories into education. Both coders independently applied the framework to all articles. Disagreements were resolved through discussion or consultation with a third reviewer.

Figure 1

PRISMA diagram

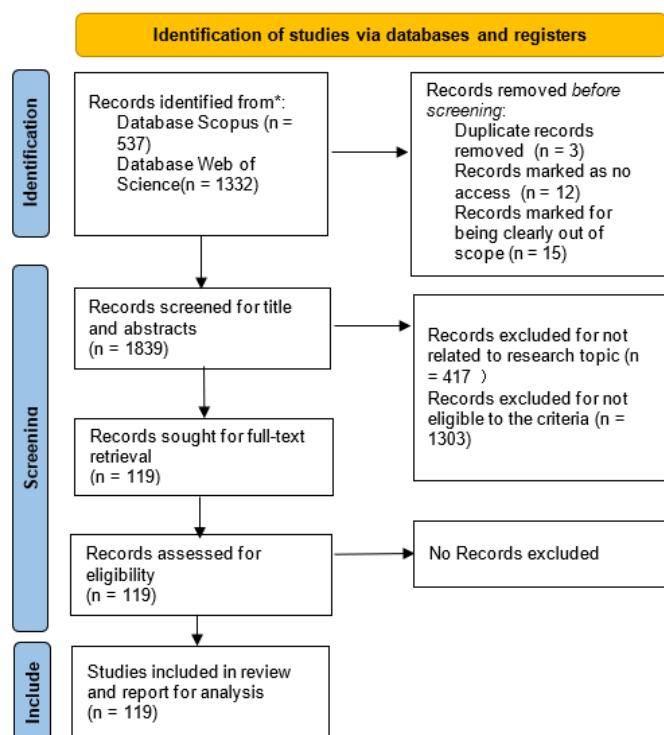


Table 2.

Inclusive and exclusive criteria from databases

Inclusion criteria	Exclusion criteria
(1) The studies had to be published journal articles indexed by the Scopus and Web of Science databases	(1) Review and editorial papers were excluded
(2) Studies that were empirical, practical, evidence-based and case studies, conducted in any education level and subjects' practices, including K12 and higher education settings	(2) Articles that were not related to AI application in education application filed were excluded
(3) Studies that provided theoretical discussion on AI integration in all educational practices	(3) Articles that were not written in English were excluded
(4) The studies had to provide descriptions of the underlying theories and methods used	(4) Conference proceedings and abstract without full text assess papers were excluded

3. Results

The 119 reviewed studies reveal a research landscape heavily weighted toward higher education ($n = 68$, 57.1%), with K-12 contexts ($n = 42$, 35.3%) as a secondary area of interest. Within the K-12 sector, scholarly attention was distributed between secondary (20.2%) and primary (15.1%) levels, while early childhood and cross-level studies remain notably scarce. From a disciplinary perspective, AI research is firmly rooted in STEM ($n = 47$, 39.5%), particularly within the technical domains of computer science and engineering. However, a significant thematic cluster also emerged in English language instruction, which represents the second most active domain with 28 identified studies.

3.1. Pedagogical approach to AI-Enhanced Teaching

AI technologies enable unprecedented customization of instructional content and pacing to individual student characteristics. In English as a Foreign Language (EFL) contexts, AI-powered systems including chatbots and large language models facilitate personalized tutoring, data-driven instructional adjustments, and enhanced student engagement (Nguyen & Dinh, 2024; Thinh et al., 2020). These findings align with prior theoretical work emphasizing AI's capacity to customize learning experiences and reduce cognitive load through scaffolding aligned to learner needs (Van Lehn, 2011). In Science, Technology, Engineering and Mathematics (STEM) education, pedagogies such as project-based learning, self-regulated learning (SRL) supported by AI chatbots, and computational thinking through robotics highlight the role of AI in fostering critical thinking and creativity (Chang et al., 2023; Maqruf et al., 2025). These pedagogical applications align explicitly with the AI-TPACK framework's emphasis on integrating technological pedagogical knowledge with discipline-specific content (Celik, 2023). In general, robotics platforms and AI-integrated programming environments foster computational thinking and coding skills in K-12 and higher education settings (Husain, 2024; Thinh et al., 2020). Students engage with AI as both a learning domain (understanding AI principles and applications) and a pedagogical tool (using AI to enhance learning across disciplines). Beyond instructional delivery, AI systems support classroom management through automated attendance tracking and behavioral monitoring. More substantively, AI enables innovative assessment approaches, including learning analytics that identify struggling students in real time, automated feedback on written assignments, and evaluation of programming code (Wong et al., 2026).

3.2. AI Technologies and Integration Patterns Across Disciplines and Educational Levels

Analysis reveals substantial variation in technology adoption and implementation across academic disciplines. STEM disciplines (particularly computer science, engineering, and physics) demonstrate high diversity in AI tool adoption (Chang et al., 2023; Maqruf et al., 2025). English Language Teaching shows concentrated use of robot-assisted learning and conversational AI systems (Thinh et al., 2020). Mathematics, while increasingly incorporating AI, shows adoption primarily focused on problem-solving scaffolding rather than conceptual instruction (Opesemowo & Ndlovu, 2024). Social sciences and arts disciplines demonstrate more limited but emerging applications, often focused on simulation and creative design tasks (Izgi, 2025; Wong, 2025). This variation suggests that disciplinary epistemology and pedagogical traditions substantially influence technology adoption patterns.

Higher education settings exhibit substantially greater diversity in AI tool adoption than K-12 settings. K-12 settings rely more heavily on generative AI tools (e.g., ChatGPT, Midjourney) that require minimal institutional infrastructure while supporting immediate instructional applications (Kim, 2025).

3.3. Teachers' AI Teaching Self-Efficacy Across Subject Matters

In contrast to the broad affective orientation of confidence (Hazaymeh et al., 2024), AI teaching self-efficacy refers to a teacher's task-oriented beliefs about their capabilities within the domain of AI instruction (Kim & Kim, 2022). As a measurable construct, it is typically operationalized through certainty ratings regarding specific AI teaching performances (Meegan & Young, 2025). Teachers' perceived capability to teach effectively with AI tools varies across academic disciplines. STEM teachers generally report high self-efficacy, using AI to scaffold scientific writing and support complex problem-solving, although they also raise concerns about transparency and changing professional roles (Husain, 2024). English as a Foreign Language (EFL) teachers tend to exhibit moderate self-efficacy, with experienced instructors recognizing the benefits of AI but often lacking the initiative to pursue deeper and more innovative integration (Hazaymeh

et al., 2024). In the social sciences and arts, teachers primarily treat AI as a creative catalyst for simulation and design activities, and their sense of efficacy depends heavily on clearly defined, well-scaffolded project contexts (Wang et al., 2024). These patterns support the AI-TPACK view that subject-specific AI competencies are crucial for effective teaching and that professional development should be tailored to both disciplinary demands and individual teacher backgrounds, including prior training and qualifications (Celik et al., 2023; Chan & Tang, 2024). In cross-disciplinary or general education contexts, such as teacher education and ICT courses, both self-efficacy and outcome expectations are strongly shaped by the availability and quality of professional development opportunities. Empirical studies show that targeted training can directly enhance pre-service and in-service teachers' self-efficacy, which in turn becomes a key driver of their readiness and intention to adopt AI in their practice (Kim, 2025; Yue et al., 2024). This indicates that teacher readiness is a multidimensional state that both precedes and results from self-efficacy, providing a conceptual bridge to the next focus of this review: how teachers' overall AI readiness is formed and manifested in educational practice.

3.4. Teachers' AI Readiness

Readiness refers to multi-dimensional constructs including insituational support, ethical awareness and general preparation (Al-Abdullatif, 2025). Readiness and self-efficacy are reciprocal, with a high level of readiness providing the necessary foundation for self-efficacy to flourish and drive further engagement (Li et al., 2025). Teachers' readiness to adopt AI is shaped by factors such as their acceptance of AI, attitudes toward its use in education, and ethical awareness, all of which are strongly influenced by the nature of AI (MacDowell et al., 2024). Studies grounded in the Technology Acceptance Model (TAM) consistently show that perceived usefulness and ease of use are central drivers of acceptance, and that hands-on, practice-oriented training plays a critical role in fostering more positive attitudes toward AI (Uygun, 2024). Research shows that when AI is viewed as a supportive partner that helps teachers instead of replacing them, people tend to respond more positively, even though many educators still have concerns about accuracy and reliability (Zhao et al., 2022). At the same time, ethical awareness remains underdeveloped; many teachers are not fully aware of algorithmic bias, data privacy risks, or wider socio-ethical implications unless these topics are explicitly addressed in professional learning opportunities (Fundi et al., 2024).

A key pattern emerging from the literature is the strong positive relationship between teachers' AI teaching self-efficacy and their overall readiness to adopt AI in practice (Granström & Oppi, 2025). Teachers with higher self-efficacy are more likely to accept AI tools, hold more positive attitudes toward AI integration, recognize both the potential and the limitations of AI, and engage in more nuanced ethical reasoning about when and how AI should be used in teaching. This shows that professional development with practical experience, mentoring, and teamwork can strengthen many aspects of teacher readiness. It helps teachers feel more confident, open-minded, positive, and aware of ethical issues when using AI.

3.5. Teachers' ethical concern

Ethical concerns in AI integration are discipline-specific. In the STEM fields, primary concerns focus on transparency and trustworthiness (Husain, 2024), whereas English language teachers' concerns are academic integrity and preservation of authentic work (Nguyen & Dinh, 2024). Meanwhile, social sciences and arts teachers emphasize the critical issues of privacy, bias and student voice (Wang et al., 2024). Training must therefore be discipline-specific, providing contextual strategies for navigating ethical challenges. Teachers often recognize ethical issues but lack a pedagogical framework to address them (Ho and Lee, 2025), which highlights a need for context-sensitive professional development.

4. Discussion

This review provides a comprehensive examination of AI applications in educational practice from teachers' perspectives, synthesizing evidence across four critical dimensions: pedagogical approaches with AI, AI technologies integrated into classrooms, AI teaching self-efficacy, and readiness of teachers.

The evidence demonstrates that AI integration is not simply about adding new tools on top of existing practice. Instead, AI fundamentally reshapes what teachers do, how they spend their instructional time, and how they interact with

students. When intelligent tutoring systems take over parts of assessment, teachers increasingly move into roles as instructional designers and individual mentors. When students work with AI-powered mobile learning, teachers shift from being the main source of information to acting more as learning coaches and facilitators. These changes call for parallel shifts in how teachers are prepared, the professional values they hold, and the ways they make instructional decisions. The traditional view of teachers as the sole source of subject-matter expertise needs to be updated toward a model in which teachers and AI systems work in partnership to support student learning. The evidence clearly shows that a one-size-fits-all approach to AI in education does not work. STEM subjects focus on problem-solving and computational thinking, and lend themselves naturally to AI-supported activities, while English language teaching can draw on AI's conversational strengths and ability to provide rich, authentic language input. In the social sciences and arts, teachers must be especially creative in identifying ways for AI to genuinely enrich interpretation, creativity, and critical inquiry. These differences mean that professional development and policy should respond to the specific opportunities and constraints of each discipline, instead of promoting generic, uniform models of AI integration across all subjects.

Teachers' confidence in their ability to teach effectively with AI varies significantly across disciplines, and this variation has important implications for how support is provided. STEM teachers generally feel more confident with AI tools because these disciplines naturally align with computational problem-solving and analytical thinking. EFL teachers recognize the value of AI for language practice and feedback but often lack the motivation or support to explore deeper applications. Teachers in social sciences and arts require well-designed projects and clear learning goals to feel confident using AI creatively. It is clear that self-efficacy varies across different contexts. With the right professional development, such as hands-on experience, mentoring, and subject-specific support, teachers can become more confident in using AI in their teaching. When teachers experience success with AI through carefully designed learning experiences, their confidence grows, and this confidence then opens the door to greater readiness for broader adoption.

Teacher readiness for AI adoption rests on multiple interconnected foundations: how teachers perceive AI's usefulness and ease of use, their emotional responses to AI in education, their ethical understanding, and their overall confidence in their ability to use these tools well (MacDowell et al., 2024). Yet many teachers remain unaware of important ethical considerations such as algorithmic bias and data privacy unless these topics are explicitly taught and discussed (Fundi et al., 2024). The research reveals that these dimensions of readiness are not separate; they interact and reinforce one another. Teachers with stronger self-efficacy tend to accept AI more readily, hold more positive attitudes, and think more carefully about the ethical dimensions of AI use in their classrooms. This insight suggests that investing in professional development that strengthens self-efficacy through authentic practice, peer collaboration, and mentoring creates positive momentum across all aspects of readiness, helping teachers become not only more confident but also more thoughtful, accepting, and ethically grounded in their approach to AI integration.

The evidence points to a critical need for comprehensive professional development that goes beyond technical tool training. Instead, it requires addressing the full landscape of AI literacy in education. Teachers require a foundational understanding of what AI is, how it works, and where it can meaningfully contribute to learning. (Hur, 2025; Tang & Zhang, 2025). Beyond conceptual knowledge, teachers need practical experience designing lessons in which AI serves clear pedagogical goals rather than becoming an end in itself. This means professional learning needs to be tailored to each subject area; some learning may be conducted collaboratively with the library (Tang, 2020). What works for a computer science teacher differs substantially from what a humanities teacher needs to know (Zhang et al., 2024). Equally important are opportunities for teachers to think critically and together about the ethical implications of AI in their classrooms. When professional development is designed this way, it gives teachers the skills and confidence to use AI well in their classrooms. Without such comprehensive support and tailored professional learning, teachers may leave feeling unprepared and unsure how to use AI effectively and responsibly in classrooms.

5. Conclusion

This study examined AI applications in educational practice from teachers' perspectives, focusing on pedagogical approaches, technology-integration patterns across subjects, self-efficacy, and teacher readiness for AI adoption. The findings reveal that effective AI integration involves diverse pedagogical strategies such as personalized learning, project-based learning, self-regulated learning support, and innovative assessment, and these approaches need to be tailored to fit the particular contexts of different disciplines and the needs of students rather than applied in the same way across all subjects. AI technologies in classrooms take three primary forms: generative AI chatbots (most common in K-12), intelligent tutoring systems (more prevalent in higher education), and AI-powered mobile learning platforms (increasingly widespread). Teachers' confidence in using AI varies significantly by discipline, suggesting that targeted professional development can meaningfully strengthen teacher confidence across contexts. Teacher readiness for AI emerges as a multidimensional construct including acceptance, attitudes, and ethical awareness. In addition, there is a strong relationship between confidence and readiness. In conclusion, these findings establish that successful AI integration requires comprehensive, discipline-responsive professional development that addresses not only technical skills and pedagogical knowledge but also ethical reasoning and the emotional dimensions of professional change.

References

- Celik, I. (2023). Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. *Computers in Human Behavior*, 138, Article 107468.
- Chan, K. K. W., & Tang, W. K. W. (2024). Evaluating English teachers' artificial intelligence readiness and training needs with a TPACK-based model. *World Journal of English Language*, 15(1), 129–129.
- Chan, V. (2025). Investigating the Effects of Extended Reality (XR) on Interpreter Competencies: An Experimental Study of a Pioneering XR Mobile-Assisted Language Learning Application for Interpreting Training. *Journal of Computer Assisted Learning*, 41(3), e70055.
- Chang, D. H., Lin, M. P. C., Hajian, S., & Wang, Q. Q. (2023). Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. *Sustainability*, 15(17), Article 12921.
- Fundi, M., Sanusi, I. T., Oyelere, S. S., & Ayere, M. (2024). Advancing AI education: Assessing Kenyan in-service teachers' preparedness for integrating artificial intelligence in competence-based curriculum. *Computers in Human Behavior Reports*, 14, Article 100412.
- Granström, M., & Oppi, P. (2025). Assessing Teachers' Readiness and Perceived Usefulness of AI in Education: An Estonian Perspective. *Frontiers in Education* (10). 1622240.
- Hazaymeh, W. A., Bouzenoun, A., & Remache, A. (2024). EFL instructors' perspective on using AI applications in English as a foreign language teaching and learning. *Emerging Science Journal*, 8(5), 73–87.
- Ho, C. S. M., & Lee, J. C. K. (2025). From intuition to action: Exploring teachers' ethical awareness in the use of AI tools in education. *Computers and Education: Artificial Intelligence*, 100502.
- Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.
- Husain, A. (2024). Potentials of ChatGPT in computer programming: Insights from programming instructors. *Journal of Information Technology Education: Research*, 23, Article 002.
- Hur, J. W. (2025). Fostering AI literacy: Overcoming concerns and nurturing confidence among preservice teachers. *Information and Learning Sciences*, 126(1–2), 56–74.
- İzgi, F. D. (2025). How Can We Use Artificial Intelligence in Social Sciences. In *AI Use in Social Sciences* (pp. 353-382). IGI Global Scientific Publishing.
- Kim, J. (2025). Perceptions and preparedness of K-12 educators in adopting generative AI. *Research in Learning Technology*, 33.

- MacDowell, P., Moskalyk, K., Korchinski, K., & Morrison, D. (2024). Preparing educators to teach and create with generative artificial intelligence. *Canadian Journal of Learning and Technology*, 50(4), 1–23.
- Maqruf, A., Herman, T., Nurlaelah, E., Suwarma, I. R., & Tang, W. K. W. (2025). Exploring the research trends of mathematics in STEM education: A systematic review using bibliometric analysis. *STEM Education*, 5(10), 1153–1171.
- Nguyen, P. T., & Dinh, V. T. (2024). Design and practice of human-machine cooperative international Chinese character teaching in ChatGPT application. *Journal of Infrastructure, Policy and Development*, 8(9), Article 6269.
- Opesemowo, O. A., & Ndlovu, M. (2024). Artificial intelligence in mathematics education: The good, the bad, and the ugly. *Journal of Pedagogical Research*, 8(3), 333–346.
- Tang, W. K. W. (2020). One plus one is greater than two: faculty-librarian collaboration for developing information literacy in higher education. *IAFOR Journal of Literature and Librarianship*, 9(2), 82–86.
- Tang, W.K.W. (2025). Artificial Intelligence in Mathematics Education: Trends, Challenges, and Opportunities. *International Journal of Research in Mathematics Education*, 3(1), 75–90
- Tang, W.K.W., Zhang, A. (2025). A Scoping Review of AI Literacy for Teachers. In: Hong, JC. (eds) New Technology in Education and Training. *Lecture Notes in Educational Technology*. Springer.
- Thinh, N. T., Hai, N. D. X., & Tho, T. P. (2020). The influential role of robot in second language classes based on artificial intelligence. *International Journal of Mechanical Engineering and Robotics Research*, 9(9), 1306–1311
- Uygun, D. (2024). Teachers’ perspectives on artificial intelligence in education. *Advances in Mobile Learning Educational Research*, 4(1), 931–939.
- Wang, X., Pang, H., Wallace, M. P., Wang, Q., & Chen, W. (2024). Learners’ perceived AI presences in AI-supported language learning: A study of AI as a humanized agent from community of inquiry. *Computer Assisted Language Learning*, 37(4), 814–840.
- Wong, J. M. S. (2025). AI role-play as a research methodology: systematic simulation for educational research. *Science Talks*, 100487.
- Wong, J. M. S., Tang, W. K. W., Wong, H. Y. H., & Ng, H. K. Y. (2026). From Principles to Practice: An Operational Framework for Equitable Artificial Intelligence in Educational Assessment. *TechTrends*, 1–12.
- Yue, M., Jong, M. S. Y., & Ng, D. T. K. (2024). Understanding K–12 teachers’ technological pedagogical content knowledge readiness and attitudes toward artificial intelligence education. *Education and information technologies*, 29(15), 19505–19536.
- Zhang, S., Diao, J., Ma, X., Tang, X., & Ding, X. (2024). What qualities do teachers need in the era of artificial intelligence: Analysis based on international experience. *STEM Education Review*, 2.
- Zhao, L., Wu, X., & Luo, H. (2022). Developing AI literacy for primary and middle school teachers in China: Based on a structural equation modeling analysis. *Sustainability*, 14(21), Article 14549.

iChat-Job: Designing Self-Determination Theory-Grounded System with Digital Humans for Increasing University Graduates' Employability

Yanjie Song^{1*}, Chi Ho Yeung¹, Xiangen Hu², Kin Fung Chiu³, Liping Deng⁴, Kaiyi Wu¹, Jiaoyang Ding¹, and Jiaxuan Lyu¹

¹ The Education University of Hong Kong

² The Hong Kong Polytechnic University

³ The Chinese University of Hong Kong

⁴ Hong Kong Baptist University

*ysong@eduhk.hk

Abstract: University students often struggle with resume writing, career direction, and interview preparation. Existing AI job tools focus on short-term optimisation rather than educational development. This paper presents iChat-Job, an SDT-grounded AI career development system supporting (1) resume and portfolio building, (2) labour-market-aligned career pathway recommendations, and (3) digital-human interview simulation with formative feedback. The system operationalises autonomy, competence, and relatedness, integrates privacy-preserving local analysis, and includes a lightweight growth-tracking mechanism. It offers a theory-driven approach for scalable employability training and outlines planned formative evaluation.

Keywords: Employability training, Digital learning, Self-determination theory, Digital human, Interview simulation

1. Introduction

Graduates face increasing challenges in crafting effective resumes, identifying job opportunities, and preparing for interviews. Traditional career services lack personalisation and real-time support. Commercial AI tools provide resume polishing and job matching but rarely incorporate educational theory (Chiu, 2023). This study introduces iChat-Job, a system designed to enhance employability through SDT-grounded support.

2. Relevant Literature

2.1 Existing AI job recommendation and career development platforms

Many platforms rely on algorithmic matching (Padma et al., 2024) without theoretical grounding, resulting in superficial guidance and reduced long-term engagement (Sohana et al., 2024). They often fail to address intrinsic motivation, non-linear career paths (Chiu, 2023), and students' psychological needs, leading to anxiety and disengagement (Holding et al., 2020). Commercial platforms optimise for rapid matching rather than holistic development.

2.2 Self-Determination Theory (SDT)

SDT highlights three needs—autonomy, competence, and relatedness—as foundations for sustained motivation (Deci & Ryan, 2012). SDT offers a powerful lens for understanding motivation and sustainable career engagement in individuals' job-seeking process (McAnally, K., & Hagger, 2024). In vocational contexts, satisfying these needs supports resilience and clearer career direction (Chen, 2017). iChat-Job adopts SDT as its design framework.

3. Design of iChat-Job System

3.1 SDT as a theoretical framework

The system translates SDT (Deci & Ryan, 2012) into design requirements: autonomy through meaningful choices; competence through rubric-based feedback and progress indicators; and relatedness through supportive digital-human coaching.

3.2 Features of the iChat-Job System

Table 1 aligns system features with SDT needs. Each new feature requires an explicit SDT rationale and a visible indicator such as goal choices or progress milestones.

Table 1. The matrix for aligning features of iChat-Job system with SDT

Feature	Autonomy	Competence	Relatedness
Set goals	Students choose target roles, industries, and location.	Milestones and progress timeline visualize improvement.	Supportive reflection prompts encourage persistence.
CV Building	Choice of templates and style, student-controlled iteration.	Checklist/rubrics and revision suggestions guide improvement.	Coaching-style guidance reduces stress and confusion.
Job Matching	Multiple plausible pathways are offered with trade-offs.	Skill gaps are translated into actionable learning tasks.	Alignment to learner goals signals “the system understands me”.
Interview Training	Students choose target position and interview style/avatar.	Formative feedback (score + suggestions) supports practice loops.	Supportive, low-pressure interaction fosters relatedness.

Figure 1 Presents four phases of student learning career-ready skills using the iChat-Job system.

Figure 1.

Four phases of student learning career-ready skills using iChat-Job

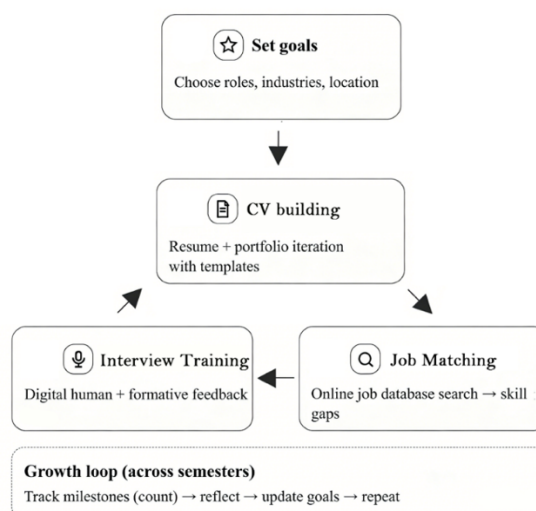


Figure 1 summarises the user journey: goal setting, resume/portfolio iteration, job-market-aligned skill-gap identification, and digital-human interview practice, with a growth loop across semesters.

3.3 Resume and Portfolio Development

Students upload resumes and use curated templates. Local data masking protects privacy before analysis. A local vision-language model identifies missing fields and issues, offering targeted suggestions. The system supports iterative improvement without replacing student agency.

3.4 Career Pathway Recommendation

The system analyses resume signals and job-market data to recommend career directions and learning pathways. It presents multiple plausible pathways, highlights trade-offs, and identifies skill gaps with actionable steps.

3.5 Digital-Human Interview Simulation

The simulator provides real-time voice interaction with five interviewer personas. It generates position-aligned prompts, evaluates performance across six dimensions, and visualises results via radar charts. The digital human acts as a supportive trainer to reduce anxiety (Kiuchi et al., 2024; Yusuf et al., 2025). Figure 2 shows the real-time interview interface with lip-synchronised 2D avatars.

4. System Architecture

Figure 3 presents the modular architecture: a web frontend for resume building, job matching, and interview simulation; a backend for AI orchestration and data management; and external components including job-database search and digital-human rendering. A growth dashboard visualises milestones and supports SDT-aligned reflection (Lee & Kim, 2025).

Figure 2.

The digital-human real-time voice interview interface of iChat-Job, showing the Spine 2D avatar with lip-synchronised coaching interaction.

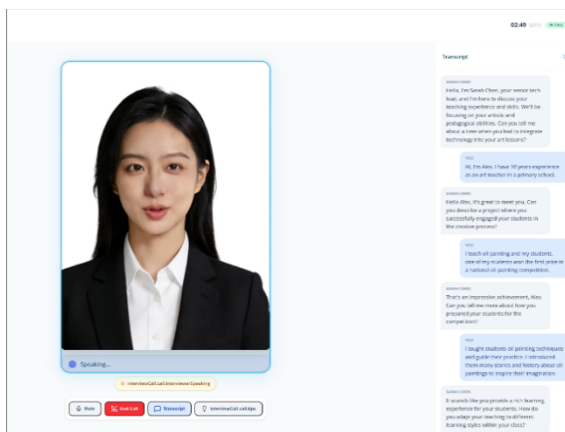


Figure 3.

System architecture of iChat-Job.

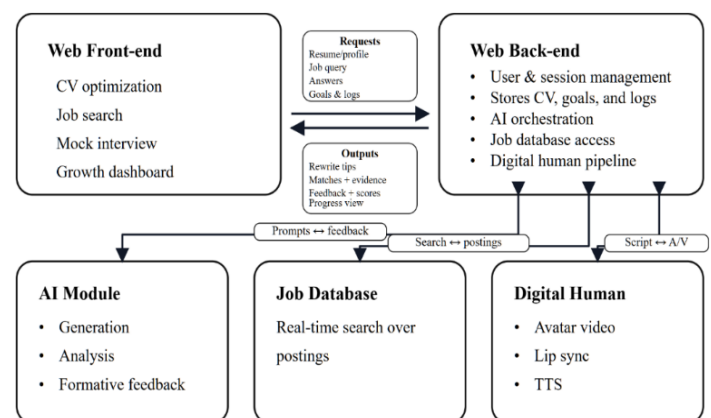


Figure 3 presents the modular architecture: a web frontend for resume building, job matching, and interview simulation; a backend for AI orchestration and data management; and external components including job-database search and digital-human rendering. A growth dashboard visualises milestones and supports SDT-aligned reflection.

5. System Features

A lightweight growth mechanism makes SDT visible: students set goals, track resume iterations and interview practice, and review progress. The system can be integrated into workshops or used independently. Limitations include the need for large-scale evaluation, bias mitigation, and strong privacy safeguards.

6. Future Work

Planned enhancements focus on strengthening both learning effectiveness and system transparency: the team will conduct a formative pilot study to evaluate how well the platform supports students' SDT needs, develop richer explainability features that clarify how skill-gap evidence is generated, expand portfolio tools to better support multimedia artefacts and collaborative creation, and build longitudinal career-path analytics that enable more adaptive, personalised support over multiple semesters.

Acknowledgement

The study was supported by Projects for Inter-Institutional Collaborative Activities (IICA) Portion of the Teaching Development and Language Enhancement Grant in 2025-28 Triennium (T5015), Hong Kong University Grants Committee (UGC). We appreciate the funding support.

References

- Chen, C. P. (2017). Career self-determination theory. In *Psychology of career adaptability, employability and resilience* (pp. 329-347). Cham: Springer International Publishing.
- Chiu, T. K. F. (2023). Using self-determination theory (SDT) to explain student STEM interest and identity development. *Instructional Science*. Advance online publication.
- Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. *Handbook of theories of social psychology*, 1(20), 416-436.
- Holding, A. C., St-Jacques, A., Verner-Filion, J., Kachanoff, F., & Koestner, R. (2020). Sacrifice—but at what price? A longitudinal study of young adults' sacrifice of basic psychological needs in pursuit of career goals. *Motivation and Emotion*, 44(1), 99-115.
- Kiuchi, K., Otsu, K., & Hayashi, Y. (2024). Psychological insights into the research and practice of embodied conversational agents, chatbots and social assistive robots: a systematic meta-review. *Behaviour & Information Technology*, 43(15), 3696-3736.
- Lee, J., & Kim, D. (2025). From awareness to empowerment: self-determination theory-informed learning analytics dashboards to enhance student engagement in asynchronous online courses. *Journal of Computing in Higher Education*, 37(3), 1078-1118.
- McAnally, K., & Hagger, M. S. (2024). Self-determination theory and workplace outcomes: A conceptual review and future research directions. *Behavioral sciences*, 14(6), 428.
- Padma, E., Sowdharshini, P., Shanmugapriya, P., Reshmaa, K. M., & Srimathi, C. N. (2024). Career guidance system for students using machine learning. In *Challenges in Information, Communication and Computing Technology* (pp. 666-671). CRC Press.
- Sohana, R., Niharika, R., Shah, K., Singh, T., & Parveen, M. S. (2025, July). Bridging Career Gaps Using AI-Driven Career Pathways and Engaging Augmented Reality Simulations. In *International Conference on ICT for Sustainable Development* (pp. 195-206). Cham: Springer Nature Switzerland.
- Yusuf, H., Money, A., & Daylamani-Zad, D. (2025). Pedagogical AI conversational agents in higher education: a conceptual framework and survey of the state of the art. *Educational technology research and development*, 73(2), 815-874.

An exploratory study of examining the effects of graduate-level learners' trust towards GAI-enhanced 3D virtual teachers on learning outcomes, behavioral transitions, and perceived cognitive load

Tinghui Wu ¹, Manpreet Singh ², Yanjie Song ^{2*}

¹ College of Education, Zhejiang University, Hangzhou, China

² Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong SAR, China

* ysong@eduhk.hk

Abstract: Recently, generative AI (GAI) has aroused great attentions in education. However, whether learners' trust concerning GAI would hinder self-regulated learning remain understudied. This study examines the effects of graduate-level learners' trust concerning GAI-enhanced 3D virtual teachers on their learning outcomes, behavioral transitions, and perceived cognitive load. Six graduate-level learners were recruited in a research-intensive university in Hong Kong, and they were grouped into high-level trust group (HT group), medium-level trust group (MT group), and low-level trust group (LT group) according to their beforehand self-reported data, respectively. Datasets were collected from final concept map products, process-oriented data, and survey data. We used a mixed approach to analyze these data. Results showed that the HT group exhibited the best learning outcomes, the most reasonable behavioral transitions, and appropriate perceived cognitive load in their learning. While learners in the LT group performed worse, the MT group's performance was average. This findings provide valuable insights for educators when integrated GAI in learning, guiding instructors and learners to use artificial intelligence to smooth their learning.

Keywords: generative AI (GAI), trust, 3D virtual learning environments, exploratory study

1. Introduction

As virtual reality technology evolving, the three-dimensional VLE (3D-VLE) has been employed in facilitating learning performances due to the characteristics of embodiment, immerse, and society (Bogusevski et al., 2020). In addition, generative artificial intelligence (GAI) has also gained significant attention and has shed lights on supporting self-regulated learning (Hsiao & Tang, 2024). Based on deep neural networks, GAI can be integrated into 3D-VLE to serve as a personalized assistant in self-regulated learning. Nonetheless, trust concerns regarding personal privacy and content accuracy may still hinder students' intention to use GAI-enhanced 3D-VLE (Gao et al., 2024). As previous studies stated, trust is a crucial factor influencing whether people interact with AI and their subsequent outcomes (Ueno et al., 2022). To the best of our knowledge, whether and to what extent learners' trust towards GAI affects their learning remains understudied. To fill this research gap, we conducted an exploratory study aiming to examine the effects of graduate-level learners' trust concerning GAI-enhanced 3D virtual teachers on learning outcomes, behavioral transitions, and perceived cognitive load within the context of GAI-enhanced 3D-VLE. Specifically, six graduate-level learners in a research-intensive university in Hong Kong were recruited to perform self-regulated learning based on a five-phase inquiry-based learning pedagogical design. They were grouped into high-level trust group (HT group), medium-level trust group (MT group), and low-level trust group (LT group) according to their beforehand self-reported survey data. Dataset was collected from learning-related outcomes, including their final concept maps, process-oriented data, and self-reported perceived cognitive load. To achieve the research objective, three research questions were proposed.

RQ1: Do the HT group, MT group, and LT group have different learning outcomes?

RQ2: What behavioral transitions do the HT group, MT group, and LT group demonstrate?

RQ3: What cognitive load levels do the HT group, MT group, and LT group perceive?

2. Methods

2.1. Research context, participants, and research procedure

This study referred to the IPCC Sixth Assessment Report (Kikstra et al., 2022) and chose the topic of “Atomic energy for replacing unsustainable sources” for learning. We adapted a well-crafted questionnaire survey from the study of Gulati et al. (2019) to measure their self-reported trust concerning GAI beforehand (Cronbach’s $\alpha = 0.83$), ultimately designating the top two highest-scoring participants as the high-level trust group (HT group), the two lowest-scoring participants as the low-level trust group (LT group), and the rest as the medium-level trust group (MT group). In this study, we employed the inquiry-based learning pedagogical design by referring to study of Song and Wen (2018), which included five phases in the GAI-enhanced 3D-VLE. It should be noted that each phase has respective GAI-enhanced 3D virtual teacher. Participants can interact with them at any time they wished. The research procedure includes three sections, i.e., the pre-survey section, the experimental section, and the post-test section.

2.2. Data collection and analysis

The study employed a mixed approach. In terms of the quantitative analysis, the concept map technique was introduced to help participants elaborate their learning understanding. This study used the quantitative content analysis (QCA) approach to assess participants’ concept maps by referring to the coding scheme. Second, we used the Pupil eye-tracking device to record the process-oriented eye movement behaviors data. By referring to Zhou et al. (2024), this study focused on seven types of individual-interaction status derived from the eye movement data. Process mining (PM) was used to present learners’ behavioral patterns over the inquiry-based learning. Lastly, after the experiment, all participants were required to finish a questionnaire survey concerning perceived cognitive load, adapting from the study of Andersen and Makransky (2021). In terms of the qualitative analysis, this study collected the participants’ process-oriented and open-ended responses concerning their Know-Want-Learned text. From this perspective, we may determine whether participants have acquired something-related to the topic and to what extent.

3. Results

3.1. RQ1: Do the HT group, MT group, and LT group have different learning outcomes?

Results demonstrated the detailed results of the categories of each group’ concept maps. First, the results indicated that the HT group’s concept maps had the highest number of nodes and links (node = 29, link = 27), followed by the MT group (node = 28, link = 26) and the LT group (node = 15, link = 14). Second, in terms of the “Subjects” category, the HT group had obvious multidisciplinary perspective in utilizing the atomic energy to address climate change crisis, especially indicating the value of Social science which other groups did not mention in this process. Third, for the “Promotions” approaches, both the HT group and the MT group referred to efforts from the international, national, social, and individual perspectives, while the LT group only demonstrated the social initiatives and individual efforts. Lastly, participants in the MT group and LT group exhibited a neutral stance towards the utilization of atomic energy as an approach to alleviate climate change crisis, while one of the participants in the HT group had a positive attitude.

To have a better understanding of the learning outcomes among participants in different groups, we gained insights from the Know-Want-Learned text. To some extent, the qualitative text indicated that participants with a higher level of trust had a relatively positive attitude and tendency towards interacting with GAI when they developed an interest in the learning topic or encountered cognitive difficulties. Furthermore, their understanding of the learning topic was not limited to a single perspective. In contrast, participants with a lower level of trust may prefer to use traditional learning methods to solve cognitive deficiencies such as watching videos, with that they may find it challenging to summarize a comprehensive answer related to the inquiry or they did not know their answers were actually one-sided.

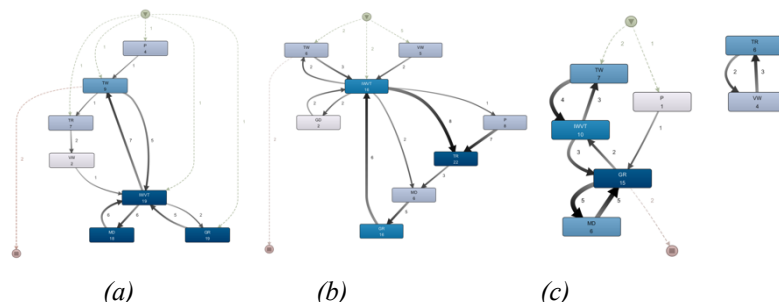
3.2. RQ2: What behavioral transitions do the HT group, MT group, and LT group demonstrate?

As shown in Figure 1, all the three groups exhibited significant behavioral transitions (frequency ≥ 5). The HT group exhibited such significant behavioral transitions as TW-->IWVT, IWVT-->TW, MD-->IWVT, IWVT-->MD, and GR-->IWVT (see Figure 1a), the MT group exhibited IWVT-->TR, MD-->GR, GR-->IWVT, and P-->TR (see Figure 1b),

and the LT group exhibited GR $\rightarrow$ MD and MD $\rightarrow$ GR (see Figure 1c). In addition, the transition between VW and IWVT occurred in the HT and MT group, but did not occur in the LT group. The top-3 codes of the HT group were IWVT (19), GR (19) and MD (18), which in the MT group were TR (22), IWVT (16) and GR (16) and in the LT group were GR (15), IWVT (10) and TW (7). The code of the IWVT had the most bidirectional transitions with TW (i.e., TW $\leftrightarrow$ IWVT, IWVT $\rightarrow$ TW), GR (i.e., GR $\rightarrow$ IWVT, IWVT $\rightarrow$ GR), and MD (i.e., MD $\rightarrow$ IWVT, IWVT $\rightarrow$ MD) in the HT group. Compared to the other groups, the LT group had the simplest transitions loops (i.e., TW $\rightarrow$ IWVT $\rightarrow$ GR $\rightarrow$ MD $\rightarrow$ GR $\rightarrow$ IWVT $\rightarrow$ TW).

Figure 1

Process mining visualization of each group' behavioral transitions. From Left to Right: HT, MT, and LT group. Note. Interacting with GAI-enhanced virtual teacher (IWVT), text writing (TW), text reading (TR), video watching (VW), guidance sheet reading (GR), concept map drawing (MD), and atomic theories and models pairing (P).



3.3. RQ3: What cognitive load levels do the HT group, MT group, and LT group perceive?

The mean values of perceived cognitive load among participants in the HT group were 2.28 (SD = 1.15) and 1.78 (SD = 0.71), respectively; in contrast, the corresponding values in the MT group were 1.67 (SD = 0.88) and 3.11 (SD = 0.81), while in the LT group they were 1.22 (SD = 0.42) and 3.06 (SD = 0.78). The results indicated that participants in the HT group had a moderate perceived cognitive load during the self-regulated learning process in their inquiry-based learning in the GAI-enhanced 3D-VLE, while those in the MT group and LT group had perceptions of cognitive load that were either relatively high or low.

4. Discussion, conclusion, limitations, and future works

This exploratory study aims to examine the effects of graduate-level learners' trust on learning outcomes, behavioral transitions, and perceived cognitive load in a GAI-enhanced 3D-VLE. This study contributes to the existing literature from the perspectives of theory and pedagogy. Theoretically, the impact of learners' trust on learning outcomes resonates with the theory of self-efficacy, which suggests that an individual's self-efficacy can be enhanced through external sources. In other words, when individuals perceive a certain pedagogical support as trustworthy or professional, the messages conveyed by this support are likely to benefit their self-efficacy improvement. In this study, if learners trust GAI, they are likely to accept its positive feedback, thereby enhancing their own self-efficacy and subsequent learning-related outcomes. Emphasizing their initiative and intent for self-regulated development in inquiry-based learning, the study found that learners' trust levels successfully predicted their learning outcomes and also influenced their behavioral transitions and perceived cognitive load, further confirming the importance of trust in human-AI interaction. Pedagogically, this study highlights the role of graduate-level learners' trust concerning GAI in improving the quality of inquiry-based learning in 3D virtual learning environments. To achieve better pedagogical effects, educators should make efforts to foster learners' trust concerning GAI by designing reliable and consistent technology-assisted learning environments.

The results need to be accepted under certain limitations. First, this study only involved six graduate-level learners, which may lead to a threat to generalizing the findings. Future studies should extend the research sample. The second limitation would be the data collections method. In this study, there is a lack of other multimodal data. Future studies should highlight the importance of multimodal learning analytics in examining learners' learning process. Finally, the

learning environment in this study may have certain technological shortcomings, which could also affect the learners' final outcomes and perceptions. Future research should be conducted under mature technological conditions.

Acknowledgements

We thank all the people who participated in this study.

References

- Andersen, M. S., & Makransky, G. (2021). The validation and further development of a multidimensional cognitive load scale for virtual environments. *Journal of Computer Assisted Learning*, 37(1), 183-196. <https://doi.org/10.1111/jcal.12478>
- Bogusevschi, D., Muntean, C., & Muntean, G. M. (2020). Teaching and learning physics using 3D virtual learning environment: A case study of combined virtual reality and virtual laboratory in secondary school. *Journal of Computers in Mathematics and Science Teaching*, 39(1), 5-18.
- Gao, Y., Wang, Q., & Wang, X. (2024). Exploring EFL university teachers' beliefs in integrating ChatGPT and other large language models in language education: a study in China. *Asia Pacific Journal of Education*, 44(1), 29-44. <https://doi.org/10.1080/02188791.2024.2305173>
- Gulati, S., Sousa, S., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, 38(10), 1004-1015. <https://doi.org/10.1080/0144929x.2019.1656779>
- Hsiao, C. H., & Tang, K. Y. (2024). Beyond acceptance: an empirical investigation of technological, ethical, social, and individual determinants of GAI-enhanced learning in higher education. *Education and Information Technologies*, 1-26. <https://doi.org/10.1007/s10639-024-13263-0>
- Kikstra, J. S., Nicholls, Z. R., Smith, C. J., Lewis, J., Lamboll, R. D., Byers, E., ... & Riahi, K. (2022). The IPCC Sixth Assessment Report WGIII climate assessment of mitigation pathways: from emissions to global temperatures. *Geoscientific Model Development*, 15(24), 9075-9109. <https://doi.org/10.5194/gmd-15-9075-2022>
- Song, Y., & Wen, Y. (2018). Integrating various apps on BYOD (Bring Your Own Device) into seamless inquiry-based learning to enhance primary students' science learning. *Journal of Science Education and Technology*, 27, 165-176. <https://doi.org/10.1007/s10956-017-9715-z>
- Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., & Seaborn, K. (2022, April). Trust in human-AI interaction: Sco** out models, measures, and methods. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1-7). <https://doi.org/10.1145/3491101.3519772>
- Zhou, Q., Suraworachet, W., & Cukurova, M. (2024). Detecting non-verbal speech and gaze behaviours with multimodal data and computer vision to interpret effective collaborative learning interactions. *Education and Information Technologies*, 29(1), 1071-1098. <https://doi.org/10.1007/s10639-023-12315-1>

Three Years of Adaptive Learning: Faculty and Student Perspectives

Che Yee Lye^{1*}

¹ Singapore University of Social Sciences

* cyllye@suss.edu.sg

Abstract: Adaptive learning technologies have gained increasing attention in higher education for their potential to support personalised learning. Despite growing adoption, limited research has examined how adaptive learning is implemented and experienced over time by both student and faculty, particularly in the adult learning contexts. This paper reports student and faculty perceptions of a three-year adaptive learning implementation across three undergraduate courses at a university. Using a mixed-methods approach, data were collected from the Adaptive Learning System (AdLeS®), student questionnaires, and focus group interviews with faculty and students. AdLeS® was embedded within the course modules and designed with adaptive learning models aligned to course pedagogy and learning outcomes. The system provided diagnostic assessments, personalised learning pathways and real-time feedback to support student learning. Findings indicated that students reported higher satisfaction with AdLeS® and identified several features useful for learning. Faculty perceived positive effects on student learning and more efficient use of teaching time for student-centred practices. Both students and faculty also identified challenges and limitations associated with adaptive learning. This paper highlights three key implications for adaptive learning implementation: the importance of pedagogically driven design principles, the need for iterative and supported delivery and implementation processes, and the central role of faculty professional development support.

Keywords: adaptive learning, adaptive course design, faculty and student perspectives, higher education, adult learners

1. Introduction

Adaptive learning has gained increasing attention in higher education in recent years, driven in part by advances in artificial intelligence (AI) and the rapid digital transformation accelerated by the COVID-19 pandemic. As online and hybrid learning expanded, institutes of higher learning (IHLs) faced persistent challenges of learner diversity, engagement and scalability, which increased interest in technology-enabled approaches that can better support personalised learning. Adaptive learning systems use learning analytics and personalisation techniques to adjust content, pacing and feedback to address diverse learner needs (Tyton Partners, 2016). This approach is particularly relevant for IHLs serving growing populations of adult learners, who often demonstrate diversity in prior knowledge, learning goals and external commitments (Chong et al., 2015). Despite this promise, gaps remain in literature. Prior research on adaptive learning has often focused on student academic performance, with fewer studies examining implementation over time. In addition, while student perceptions have received some attention, faculty perspectives remain comparatively under-explored (du Plooy et al., 2024; Martin et al., 2020). Student perspectives are critical, as perceptions of usability, usefulness and challenges influence satisfaction and engagement (Lim et al., 2022), while faculty perspectives shape adaptive course design, interpretation of learning analytics and instructional decisions (Simon & Zeng, 2024). Examining adaptive learning through both student and faculty lenses therefore provides essential insights into its pedagogical value and sustainability in higher education and adult learning contexts. To address this gap, this study examines a three-year implementation of adaptive learning in a higher education adult learning context. The study addresses the following research questions:

RQ1: How do students perceive adaptive learning in terms of satisfaction, usefulness and challenges?

RQ2: How do faculty perceive adaptive learning in terms of course design process, usefulness and challenges?

RQ3: What changes are observed in adaptive learning completion rates over time?

2. Adaptive Learning System (AdLeS®)

The Adaptive Learning System (AdLeS®) used in this study is developed in-house for students at the university to support a diverse adult learner population. AdLeS® personalises learning pathways by integrating four key adaptive elements: adaptive diagnostic, content, sequence and assessment. The system also incorporates four learning models (i.e., Bridging, Flipped Learning, Revision and Summative Assessment), selected based on course requirements and pedagogical considerations. Personalisation is driven by psychometrically robust Rasch measurement model and machine learning techniques such as collaborative filtering to adapt content and assessment based on students' demonstrated understanding. AdLeS® also generates learning analytics and reports based on Zone of Proximal Development (ZPD) (Vygotsky, 1978), providing both students and faculty with insights into learning progress. Faculty are supported through targeted professional development workshops related to adaptive learning.

3. Methods

This study examined the implementation of AdLeS® in three undergraduate courses between 2023 and 2025. The sample comprised 2,547 undergraduate students who were predominantly adult learners, and 15 faculty members involved in adaptive learning implementation. A mixed-methods approach integrated three data sources: AdLeS® completion data, student questionnaire responses collected through the Learner Satisfaction Questionnaire (Lim et al., 2022), and student and faculty focus group interviews. Completion data were used to examine progress patterns over time, questionnaire data captured student perceptions and satisfaction and focus group interviews provided in-depth insights into implementation experiences. Student focus group participants (n=15) were recruited through stratified purposive sampling based on course results to represent high-, average-, and low-progress groups. Questionnaire data were analysed using descriptive statistics (percentages), while the focus group data were analysed thematically using an inductive approach (Creswell, 2022). Coding procedures, consistency checks and an audit trail were used in the analysis of the focus group data. Findings from the completion rate, questionnaire and focus group were compared and integrated at the interpretation stage to develop a fuller understanding of the results.

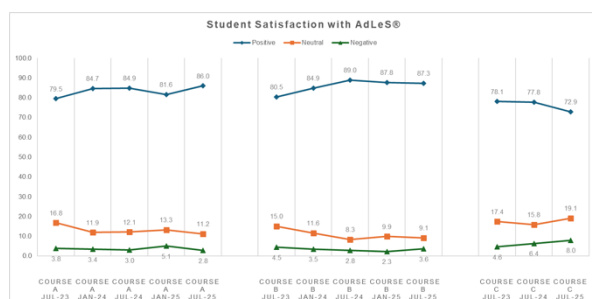
4. Results

4.1. Student Perspectives on Adaptive Learning

As shown in Figure 1, across the three years of implementation period, students reported generally positive and high levels of satisfaction with AdLeS® for all the three courses in terms of its usability, content quality and learning support functions. Students perceived AdLeS® as easy to use and operationally stable, allowing them to navigate the platform and locate the required learning materials efficiently. The content provided was perceived as sufficient and useful, with focus

Figure 1

Student Satisfaction with AdLeS®



group participants indicating that it aligned well with their learning needs. Students also valued the system's support for self-directed learning and personalised learning, including the flexibility to control their learning pace, its scaffolding structure that provided hints and feedback, the presentation of manageable learning concepts, adaptive pathways and the ZPD reports that highlighted areas for improvement.

In the focus group interviews, students identified several challenges, including difficulty interpreting dashboard reports and delays in receiving reports due to the time required for the ZPD analysis. They also noted limited use of

visuals and multimedia, as well as system functionality constraints, such as restriction on downloading learning materials or using copy-and-paste functions, which were intentionally restricted by AdLeS® design.

4.2. Faculty Perspectives on Adaptive Learning

Faculty reported generally positive perceptions of adaptive learning, particularly with respect to the adaptive course design process and its pedagogical value. They described the design process as purposeful, supported by curriculum mapping aligned with course learning outcomes and the intentional design of learning materials and assessment questions to scaffold student learning. They also valued feedback on the quality of items they developed. Faculty perceived positive impact on student learning and noted that adaptive learning enabled more efficient allocation of their teaching time towards student-centred practices. At the same time, they identified several challenges and limitations. Developing high quality adaptive items and feedback was time intensive. They also reported difficulties in interpreting learning analytics and managing student expectations. These challenges diminished over time as they gained experience with adaptive learning. Thus, professional development was viewed as essential, with faculty emphasising the need for ongoing support.

4.3. Overall Student Completion Rate

Across the three courses, adaptive learning activities contributed 15% or less to the course grade and were used mainly for concept bridging (i.e., Bridging Learning Model), preparatory learning prior to class sessions (i.e., Flipped Learning Model) and the consolidation of previously learnt concepts (i.e., Revision Learning Model). Table 1 shows that completion rates were moderate in the initial year, with some students engaging selectively with adaptive learning activities and others failing to achieve mastery, resulting in incomplete tasks. Completion rates increased across all three courses in subsequent years, with a slight 1% decline in the January 2025 semester for Courses A and B. The overall upward trend likely reflects multiple influences, including increased grade weighting, refinements in adaptive course design and platform features, improved faculty understanding of adaptive learning, and the use of reminder nudges.

Table 1.

AdLeS® Completion Rate for Courses A, B and C.

Course	Adaptive Learning Model	Weighted Percentage of AdLeS® (%)	AdLeS® Completion Rate (%)				
			Jul 23	Jan 24	Jul 24	Jan 25	Jul 25
Course A*	Revision	5-10	66.9	73.0	78.7	77.7	82.6
Course B**	Flipped Learning	5-15	46.5	52.6	62.8	61.4	69.7
Course C	Bridging	10	78.9	NA	89.2	NA	96.9

**5% in 2023 and 10% from 2024 onwards; **5% in 2023, 10% in January 2024 and 15% from July 2024 onwards*

5. Discussion and Conclusion

5.1. Pedagogical Design Principles for Adaptive Learning

The findings suggest that the value of adaptive learning depends on its pedagogical design; however, the observed increases in completion rates and satisfaction may also reflect confounding factors, particularly increased grade weighting, rather than pedagogical effectiveness of the system alone. Against this backdrop, students and faculty still identified several pedagogical design features as useful. Alignment with course learning outcomes, content granularity, adaptive pathways and assessment tasks appeared to support coherent implementation. Initial diagnostic assessments helped direct students to suitable content, while smaller, conceptually coherent topics, immediate hints and ZPD reports were perceived as helpful in guiding learning and instructional decisions.

5.2. Iterative and Supported Delivery and Implementation of Adaptive Learning

The delivery and implementation of adaptive learning introduced both opportunities and challenges. Faculty reported that adaptive learning enabled more efficient use of synchronous teaching time for discussion, feedback and application, while promoting personalised and self-directed learning through structured scaffolding and adaptive mechanisms. However, effective implementation required careful orchestration and evolved over time through cycles of enactment,

reflection, and refinement, including having adaptive learning activities contribute to the course grade to enhance student engagement. As faculty gained more experience with adaptive learning, the initial challenges related to adaptive course design, interpretation of learning analytics, and management of student expectations diminished. These findings suggest that adaptive learning should be conceptualised as a developmental instructional process that benefits from flexibility, reflective practice and continuous improvement, rather than a one-time technological rollout.

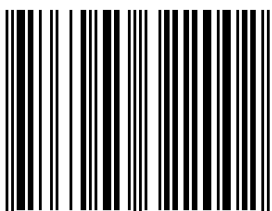
5.3. Faculty Professional Development Support

This study emphasises the central role of faculty professional development in enabling meaningful use of adaptive learning technologies. Prior to the development of adaptive courses, the AdLeS® team worked closely with faculty to introduce the AdLeS® approach, review learning models and explain key steps in adaptive course design. The design process was streamlined with the use of AdLeS®-provided templates in which faculty developed a curriculum map aligned with course learning outcomes, learning materials and assessment tasks. Following course launch, ongoing support was provided to faculty through analysis of item quality and adaptive pathways for continuous course improvement, identification of areas of concern in student learning, and targeted professional development workshops focusing on adaptive learning. For students, they received assistance for technical issues, and they were provided with reports that highlighted their areas of strengths and weaknesses.

References

- Chong, G., Ding, D., & Ho, K. W. (2015). E-learning in higher education for adult learners in Singapore. *International Journal of Information and Education Technology*, 5(5), 348-353.
- Creswell, J. W. (2022). *Research design: Qualitative, quantitative, and mixed methods approaches* (6th ed.). Sage.
- du Plooy, E., Casteleijn, D., & Franzsen, D. (2024). Personalized adaptive learning in higher education: A scoping review of key characteristics and impact on academic performance and engagement. *Heliyon*. 10(21), e39630.
- Lim, L., Lim, S. H., & Lim, R. W. Y. (2022). Measuring learner satisfaction of an adaptive learning system. *Behavioral Sciences*, 12(8), 264. doi: 10.33990/bs12080264
- Martin, F., Chen, Y., Moore, R. L., & Westine, C. D. (2020). Systematic review of adaptive learning research designs, context, strategies, and technologies from 2009 to 2018. *Education Tech Research Dev* 68, 1903–1929.
- Simon, P.D., & Zeng, L. M. (2024). Behind the scenes of adaptive learning: A scoping review of teachers' perceptions on the use of adaptive learning technologies. *Education Sciences*, 14(12), 1412.
- Tyton Partners. (2016). *Learning to adapt 2.0: The evolution of adaptive learning in higher education*. <https://tytonpartners.com/app/uploads/2016/04/tyton-Partners-Learning-to-Adapt-2.0-FINAL.pdf>
- Vygotsky, L. (1978). *Mind in society*. London: Havard University Press.

ISSN 3005-3226



9 773005 322002 >