Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Teacher–Agent–Robot Collaborative Model for Bilingual Hands-on STEAM Learning: Integrating an LLM-based Pedagogical Agent with Educational Robots

Hsien-Sheng Hsiao
National Taiwan Normal University
Nian-Shing Chen
National Taiwan Normal University
Yi-Ying Lin
National Taiwan Normal University
Guang-Han Chung
gramchung@gmail.com

National Taiwan Normal University

Research Article

Keywords: LLM-based Pedagogical Agent, Educational Robots, Prompt-guided Instruction, R&T (Robots & Tangible Objects) Learning System, Bilingual Hands-on STEAM Activities

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11199056/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Teacher–Agent–Robot Collaborative Model for Bilingual Hands-on STEAM Learning: Integrating an LLM-based Pedagogical Agent with Educational Robots

Hsien-Sheng Hsiao[1] Nian-Shing Chen[2]; Yi-Ying Lin[2]; Guang-Han Chung[1]*;

hssiu@ntnu.edu.tw; nianshing@ntnu.edu.tw ; s9551830@gmail.com; gramchung@gmail.com

[1] Department of Technology Application and Human Resource Development, National Taiwan Normal University, Taipei, Taiwan.
No.162, Sec. 1, Heping E. Rd., Da'an Dist., Taipei City 106, Taiwan (R.O.C.)

[2] Institute for Research Excellence in Learning Sciences and Program of Learning Sciences, National Taiwan Normal University
No.162, Sec. 1, Heping E. Rd., Da'an Dist., Taipei City 106, Taiwan (R.O.C.)

*Corresponding author: Guang-Han Chung
Telephone: +886-2-7749-3471
Fax: +886-2-2392-1015
E-mail: gramchung@gmail.com

1. Authors

Hsien-Sheng Hsiao is a chair professor at the Department of Technology Application and Human Resource Development, National Taiwan Normal University. His research interests include database system, computer networks, pedagogy and learning design and STEAM education.

Nian-Shing Chen is a Chair Professor in the Institute for Research Excellence in Learning Sciences and Program of Learning Sciences at the National Taiwan Normal University. He is ranked number 8 in Asia and is ranked number 103 in the world in Education Research. He is listed as World's top 2% scientists in 2021, 2022, 2023, 2024 and in the list of "Lifelong Science Influence Rankings (1960-2023) by Stanford University.

Yi-Ying Lin is a research assistant at the Institute for Research Excellence in Learning Sciences and the Program of Learning Sciences at National Taiwan Normal University. Her work primarily focuses on research related to artificial intelligence and robotics.

Guang-Han Chung is a Postdoctoral Fellow at the Department of Technology Application and Human Resource Development, National Taiwan Normal University. His research interests include information technology, digital learning, flipped learning and STEM education. Address for correspondence: Guang-Han Chung, Department of Technology Application and Human Resource Development, National Taiwan Normal University, No.162, Sec. 1, Heping E. Rd., Da'an Dist., Taipei City 106, Taiwan. Email: gramchung@gmail.com

2. Consent to Participate

Written informed consent to participate in this study was obtained from the parents or legal guardians of all participants. Participants were informed that they had the right to withdraw from the study at any time without penalty.

3. Funding

This work was financially supported by the "Institute for Research Excellence in Learning Sciences" of National Taiwan Normal University (NTNU) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and sponsored by the National Science and Technology Council(NSTC), Taiwan, R.O.C. under Grant no. 112-2410-H-003 -098 -MY3, 112-2622-H-003 -007 -,

113-2410-H-003 -033 -MY3, 113-2622-H-003 -004 - NSTC 114-2622-H-003-007-

4. Ethics Approval and Consent to Participate

This study involving human participants was approved by the Research Ethics Committee of National Taiwan Normal University (REC No. 202401HS003) on February 05, 2024. The research was carried out in accordance with the guidelines and regulations of the Research Ethics Committee of National Taiwan Normal University. Informed consent was obtained from all participants and/or their legal guardians prior to participation.

5. Clinical Trial Number

not applicable.

Abstract

In recent years, Large Language Models (LLMs) have shown potential in providing dialog-based guidance and learning prompts in educational contexts. However, systematic empirical research remains limited on approaches that position educational robots as the interaction medium and LLMs as the instructional support engine to orchestrate classroom guidance using prompts and instructional workflow scripts. In response, this study adopted a bilingual STEAM (Science, Technology, Engineering, the Arts, and Mathematics) summer camp for lower elementary students as the research context, and developed and implemented an R&T (Robots & Tangible Objects) learning system that integrated educational robots with an LLM-based pedagogical agent to foster embodied interaction and bilingual learning. The system used predesigned prompts and procedural rules to organize instructional steps, question prompts, operational checks, and feedback statements, enabling the students to receive consistent and repeatable support at key task-progression checkpoints. A total of 34 elementary students participated in a five-day program (120 minutes per day). Data sources included STEAM knowledge tests, English vocabulary assessments, a curiosity scale, hands-on performance ratings, and system interaction logs. Wilcoxon signed-rank tests indicated significant improvements in the students' STEAM knowledge, curiosity, and English vocabulary, while group-level analyses showed significant differences in hands-on performance. Process log analyses further suggested that the prompt-guided workflow scripts helped to maintain consistency in classroom guidance and supported task completion through an action–feedback–revision cycle. Nevertheless, classroom implementation was constrained by factors such as token limits, speech-recognition latency, and network/API stability, which indicated the need to incorporate exception handling and timely teacher support mechanisms to ensure instructional continuity. Overall, this study provided initial empirical support for a feasible design framework for integrating educational robots × an LLM-based pedagogical agent into bilingual hands-on STEAM activities, and its findings offer practical recommendations for curriculum adoption and system deployment.

Keywords: LLM-based Pedagogical Agent, Educational Robots, Prompt-guided Instruction, R&T (Robots & Tangible Objects) Learning System, Bilingual Hands-on STEAM Activities

1. Introduction

With the rapid development of generative artificial intelligence and Large Language Models (LLMs), language-centered instructional support has become an increasingly important focus in educational technology research. LLM-based pedagogical agents can provide real-time feedback, adaptive dialogue, and personalized guidance, creating new possibilities for classroom support. Prior studies have also emphasized that such agents are intended to collaborate with teachers rather than replace them (Chen, 2025; Lan & Chen, 2024). However, deploying these systems in authentic classrooms remains challenging because pedagogical appropriateness, interaction reliability, and instructional controllability must all be maintained simultaneously (Chapelle, 2024; Relmasira et al., 2023).

In language education, bilingual learning emphasizes the integration of content learning and second-language development (Dicks & Genesee, 2016). For lower elementary learners, a key design issue is how to balance comprehensibility and language exposure: instruction should provide sufficient second-language input while also offering first-language support, visual cues, and repeated practice to reduce cognitive load and sustain engagement (Fung et al., 2025). Prior studies further suggested that when bilingual learning is embedded within task- and inquiry-oriented activities with appropriate scaffolding, it can promote both language use and content understanding (Corrigan et al., 2025; Quigley et al., 2020; Yang, 2021).

Within this broader context, STEAM (Science, Technology, Engineering, the Arts, and Mathematics) education emphasizes interdisciplinary integration, inquiry, hands-on practice, and creativity, and it has become an important direction in contemporary education (Honey et al., 2014; Mohsen et al., 2025). However, for younger elementary learners, designing bilingual STEAM activities that are engaging, concrete, and manageable, while also addressing diverse learning and language needs, places substantial demands on teachers. In such contexts, teachers must often coordinate conceptual understanding, task progression, second-language input, and classroom facilitation at the same time, making it difficult to respond to students' comprehension and language-expression differences in real time (Hsieh et al., 2025).

In such settings, educational robots can serve as an embodied interaction channel that provides repetitive guidance, verbal prompts, and immediate feedback under teacher supervision (Belpaeme et al., 2018; Belpaeme & Tanaka, 2021). For younger learners, the multimodal and tangible features of educational robots can enhance engagement, reduce anxiety, and support both cognitive and affective learning through embodied interaction (Lindgren & Johnson-Glenberg, 2013; Mubin et al., 2013; Wilson, 2002). When integrated with LLM-based pedagogical agents and controlled through scripted prompts, educational robots offer a scalable way to deliver more consistent bilingual input and interactive support while preserving classroom controllability (Chen, 2025; Lan & Chen, 2024; Pai et al., 2024).

However, effective classroom implementation of language-driven systems and robot-mediated interaction still depends on bridging instructional needs and system capabilities. Educational and technical teams often prioritize different concerns, which can lead to misalignment between pedagogical goals and system designs (Feng et al., 2023). Therefore, establishing repeatable interdisciplinary collaboration models is

crucial for bringing pedagogical agents × educational robot learning systems into authentic classrooms (Chen & Yeh, 2025a; Lan & Chen, 2024).

Against this background, the present study adopted a bilingual hands-on STEAM program for lower elementary students to examine how an LLM-based pedagogical agent and educational robots worked together as an executable and controllable instructional system. Specifically, the study focused on three issues: how the interdisciplinary team translated lesson plans into scripted workflows with bilingual support; what deployment constraints occurred in the authentic classroom; and how the system related to the students' STEAM knowledge, curiosity, English vocabulary, and hands-on performance.

This study pursued four objectives: (1) to identify principles for transforming lesson plans into instructional workflows executable by an LLM-based pedagogical agent and educational robots; (2) to clarify how the interdisciplinary team collaborated in the design and implementation processes; (3) to summarize key challenges in classroom deployment and corresponding improvement strategies; and (4) to examine the students' multimodal participation patterns together with their STEAM knowledge, curiosity, English vocabulary, and hands-on performance.

Accordingly, the study addressed the following research questions:
RQ1. How does the interdisciplinary team collaborate to develop a pedagogical agent–robot learning system for lower elementary students?
RQ2. What challenges occur when the pedagogical agent and educational robots are implemented in the authentic classroom, and how do these challenges affect classroom interaction and student learning?
RQ3. To what extent does the pedagogical agent–robot system improve the students' STEAM knowledge, curiosity, English vocabulary, hands-on performance, and participation patterns in bilingual STEAM learning?

2. Method

2.1 Research Context

This study was conducted in a regular classroom at an elementary school during a five-day bilingual STEAM summer camp. The curriculum centered on the project theme "Little Octopus Vacuum Cleaner," which covered vacuum-cleaner concepts, batteries and circuits, parts assembly, casing design, and decoration. The activities emphasized conceptual understanding and hands-on construction through a staged workflow. Instruction was supported by an LLM-based pedagogical agent, educational robots, tablets, and QR-code-labeled tangible teaching aids, which together structured task progression and reduced ambiguity during assembly.

The camp integrated STEAM instruction with target English vocabulary and sentence patterns. The students' STEAM knowledge, curiosity, English vocabulary, and hands-on performance were assessed by tests, questionnaires, rubrics, and system logs. The teacher and teaching assistants also provided in-person scaffolding when repeated errors or technical interruptions affected the scripted workflow.

2.2 Participants

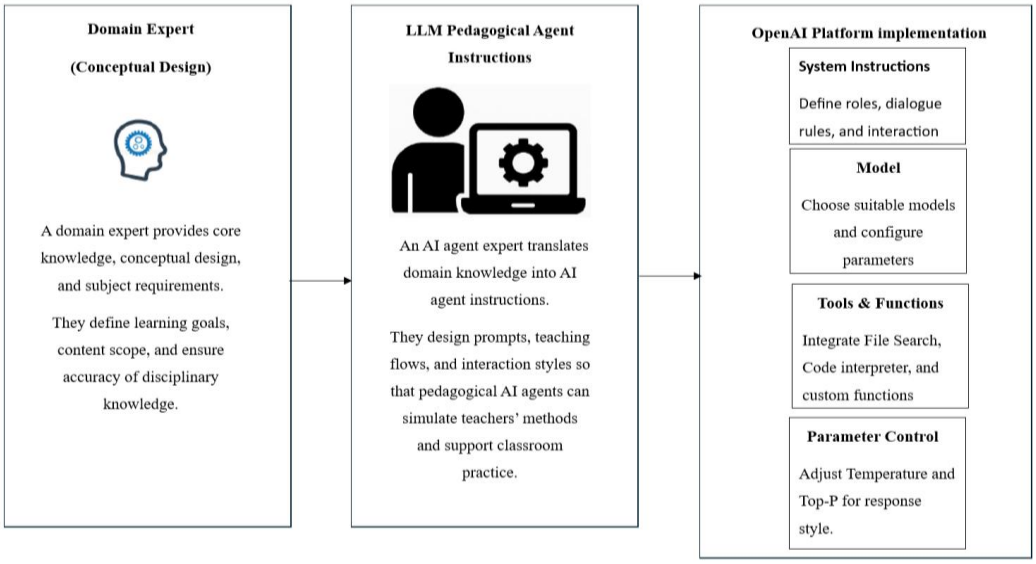
This pilot study initially recruited 40 first- and second-grade elementary students aged 7 to 9. Students who missed two or more sessions or had incomplete pre-/post-test data were excluded, resulting in a final sample of 34 participants. The camp was delivered in two cohorts (morning and afternoon), and all sessions were taught by the researcher. The students worked in pairs, with each pair using one educational robot, one set of QR-code-labeled tangible teaching aids, and one tablet, while the teaching assistants provided support during the activities. Participants were recruited through school announcements and parent registration. Parents were informed of the study's purpose and course arrangements before the camp, and written parental consent was obtained for all participants. The study procedures complied with established research ethics requirements.

2.3 Design of the R&T Learning System and LLM-based Pedagogical Agent

Developing an LLM-based pedagogical agent requires the integration of educational theory, technical implementation, and content design, and therefore depends on interdisciplinary collaboration. In this study, such collaboration was essential not only for building the final system but also for aligning instructional goals, interaction design, and classroom implementation.

This study integrated an LLM-based pedagogical agent with an R&T (Robots & Tangible Objects) learning system to create an interactive bilingual STEAM learning environment for lower elementary students. The design was grounded in embodied learning perspectives that emphasized physical manipulation and multisensory interaction. The teacher-facing perspective illustrated in Figure 1 shows how the interdisciplinary team translated instructional intentions into executable processes through model configuration, system instructions, and integrated tools, whereas the learner-facing perspective shows how the students interacted with the R&T learning system through educational robots, tablets, and tangible teaching aids. The dual-perspective design highlighted the coordinated relationships between the teacher, the pedagogical agent, and the learners in implementing the R&T learning system in the classroom.

Figure 1.
System Architecture of the Teacher–Agent–Learner Coordinated Learning Model
(a) Teacher-facing perspective



(b) Learner-facing perspective

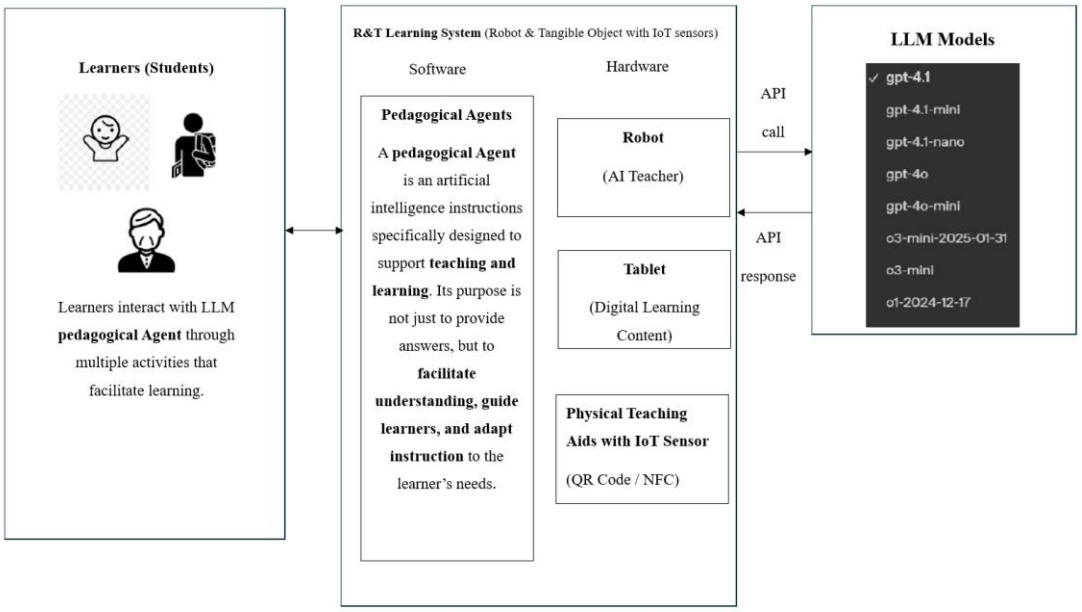
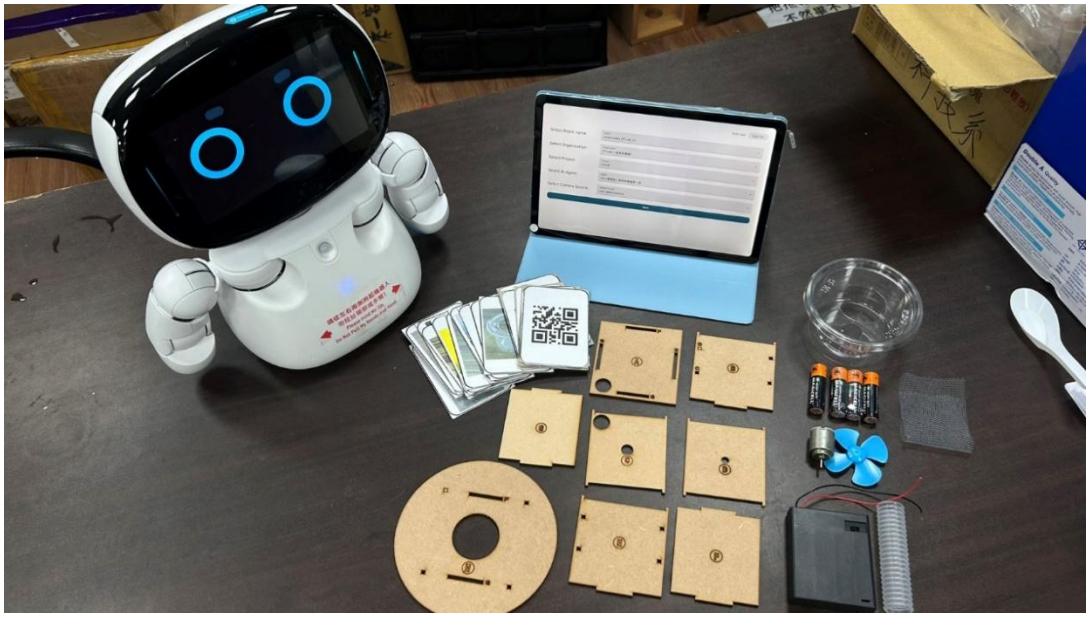


Figure 2 presents the prototype of the R&T learning system, which comprised three main components: educational robots, tablet devices, and QR-code-labeled tangible teaching aids. The educational robots provided speech recognition and output, facial-expression and motion control, and touch sensing, serving as the embodied interface of the pedagogical agent. The tablet devices supported content display, response collection, QR-code scanning, uploads, selection tasks, and feedback delivery. QR-code-labeled tangible teaching aids were used for object identification and step verification, thereby reducing reliance on high-precision image recognition while supporting intuitive understanding through visible and tangible interaction. In the "Little Octopus Vacuum Cleaner" activity, for example, the students selected labeled components, scanned them for identification, and verified the assembly steps through the R&T learning system.

Figure 2.
Prototype of the R&T Learning System



This system design did not intend to replace the classroom teacher with the pedagogical agent; rather, the agent functioned as a bounded instructional partner that guided the learning processes, provided real-time feedback, and sustained classroom interaction through speech, expressions, and brief task prompts. To maintain instructional consistency and controllability, the instructional flow was translated into prompt-scripted workflows specifying objectives, staged procedures, allowable response formats, feedback rules, and bilingual support. The agent therefore operated within a predefined instructional scope rather than as an open-ended chatbot.

From a technical perspective, LLMs served as the computational core of the pedagogical agent and supported bilingual prompts as well as context-sensitive responses. Combined with process control, the system enabled the step-by-step execution of instructional tasks according to the planned sequence. Overall, the design of the pedagogical agent followed three core principles: clear role differentiation between the teacher, the agent, and the students; standardized instructional processes; and scalability across the activity themes.

2.4 Platform Architecture and Technical Design

This study used the OpenAI platform as the development and deployment environment, with LLMs serving as the computational core of the LLM-based pedagogical agent. The platform supported multimodal classroom interaction, including speech, touch, image-based input, QR-code scanning, and real-time data recording. The platform's architecture emphasized modularity and workflow control so that the instructional activities could be executed in a predictable and scalable manner under classroom conditions.

To ensure stable task progression, activity-specific system instructions and structured prompts were developed to define the stage-based tasks, interaction methods, expected response formats, and feedback criteria. Rather than allowing unrestricted dialogue, the system relied on prompt-scripted workflow control to guide the students through the predefined instructional steps. In addition, interaction records and student operation events were logged to support both instructional monitoring and later process analysis.

A representative interaction involved the students scanning QR-code-labeled parts with the tablets, after which the system recorded the event, triggered the next scripted step, and delivered immediate feedback through the robots. In this way, the platform supported step-by-step instructional execution, multimodal interaction, and process-data collection within a unified classroom environment.

2.5 Instructional Design Template

To translate the lesson plans into executable instructional workflows, this study developed a standardized instructional design template for the LLM-based pedagogical agent. The template specified the learning objectives, role definitions, learning context, task prompts, interaction rules, and recovery strategies, thereby enabling the lesson content to be systematically converted into prompt-scripted classroom procedures. Through this template, the instructional activities were aligned with the pedagogical goals and system-executable interaction flows.

As an illustrative example, Table 1 presents the battery installation step, in which the lesson activity was transformed into a scripted prompt with corresponding response mappings, feedback patterns, and teacher-support procedures. This example demonstrates how the template supported step-by-step guidance, error correction, and instructional consistency while preserving the teacher's role in providing additional support when needed. Overall, the template served as a bridge between the expert-designed lesson plans and the executable pedagogical agent workflows.

Table 1.
Converted Instructional Template for Battery Installation Step

Field	Template Content
Student Objectives	To understand the positive (+) and negative (–) terminals of a battery and install it correctly into the holder.
Roles	Agent Instructor (LLM-based pedagogical agent) guides (scripted prompts) → Students operate → Classroom Teacher supplements when needed (e.g., demonstration or troubleshooting).
Context	In the "Little Octopus Oliver" storyline, the battery must be installed correctly to start the vacuum cleaner and power the motor.
Task (Scripted Prompt)	Agent Instructor prompt: "Please put the battery into the holder. Remember, the yellow end is the positive (+) terminal, and the black end is the negative (–) terminal."
Rules (Response Mapping)	If correct → Encourage: "Great job! The battery is ready!" If incorrect → Hint: "Look again—the yellow end should be placed on the + side."
Tools / Interfaces	QR-code-labeled tangible teaching aids: battery and holder (QR-code scanning for identification/step-checking); Tablets: display diagrams and step cues; Educational robots: deliver voice guidance and facial-expression feedback.
Answer Samples	Correct: Battery orientation matches polarity markers and is fully inserted. Incorrect: Battery reversed and/or not fully inserted.
Guidance (Recovery Strategy)	If incorrect → Provide a scripted hint: "Check the positive and negative ends again." If repeatedly incorrect → Classroom Teacher provides a brief demonstration and prompts the student to retry.
Closure & Extension	Agent Instructor wraps up: "Excellent! Now the battery is installed, and we can get the motor spinning. Can you think of other appliances that need batteries?"

Field	Template Content
Prompt-based Constraints (Optional Note)	This step follows the predefined prompts and response rules; the agent's output is restricted to the scripted instructional scope to ensure consistent, step-by-step progression.

2.6 Design Process of the LLM-based Pedagogical Agent

The LLM-based pedagogical agent and the R&T learning system were developed through iterative collaboration between educational specialists, content experts, and technical developers. The educational specialists designed the lesson plans and instructional goals, the content experts transformed them into prompt-scripted interaction flows, and the technical developers integrated the educational robots, tablets, and tangible teaching aids into a unified classroom system. Through repeated simulation, testing, and revision, the interdisciplinary team progressively refined the system to improve classroom usability, interaction stability, and instructional alignment.

The bilingual learning strategy was refined early in the design process. The initial approach relied mainly on Chinese storytelling, with English presented only as key terms. After testing, this was revised into a format of short English sentences accompanied by Chinese translation and voice support, allowing the students to receive bilingual input while maintaining comprehension. As shown in Figure 3, the educational robot delivered a story-based bilingual introduction as part of the learning activity.

Figure 3.
Story-based Bilingual Introduction Delivered by the Educational Robot



To strengthen classroom usability and process-data collection, artifact-upload and interaction-logging functions were incorporated into the system. Through the tablet interface, the students could submit responses and receive immediate feedback, and these interactions also served as a source of learning-process data. This response-and-feedback mechanism is illustrated in Figure 4. In addition, the assessment framework was revised so that it would better reflect the younger learners' characteristics and more appropriately capture their creativity, problem-solving relevance, and construction performance.

Figure 4.
Tablet-based Response, Feedback, and Process-data Collection



The interaction design was further improved during testing. Some static procedural demonstrations were replaced with instructional videos, and selected tasks were redesigned as collaborative activities to encourage discussion and participation. As shown in Figure 5, the students assembled the "Little Octopus Vacuum Cleaner" prototype as part of the staged operate–feedback–revise process. These revisions helped to make the task procedures clearer and supported more active peer interaction during hands-on learning.

Figure 5.
Students Assembling the "Little Octopus Vacuum Cleaner" Prototype



Technical adjustments were also made to improve implementation reliability. Preliminary testing showed that image recognition was not sufficiently accurate for fine-grained operations such as battery installation. Accordingly, selected verification steps were simplified, and touch-based confirmation was adopted in some activities. In addition, the initial plan to use NFC-enabled teaching aids was replaced by QR-code-labeled objects to reduce implementation complexity and better support repeated

classroom use. As shown in Figure 6, the students scanned the QR-code-labeled teaching aids for identification and step-checking, which helped to reduce ambiguity and support workflow control.

Figure 6.
Students Scanning QR-code-labeled Teaching Aids for Identification and Step-checking



To support classroom implementation, the instructional staff received phased training before and during the camp. This included one-on-one familiarization with the system workflow, workshop-based training for the instructional staff, supporting manuals, and on-site technical assistance during the lessons. Through these design, testing, and support processes, the interdisciplinary team progressively completed the development and classroom implementation of the LLM-based pedagogical agent and the R&T learning system.

2.7 Data Collection and Analysis

To capture both the interdisciplinary development process and the instructional effectiveness of the LLM-based pedagogical agent in the lower elementary bilingual hands-on STEAM activities, this study adopted a mixed qualitative–quantitative approach. The qualitative data documented how the system was designed, refined, and enacted in the classroom setting, whereas the quantitative data were used to assess changes in the students' learning outcomes and affective responses.

The qualitative data were collected from three sources. First, expert collaboration records, including meeting notes, digital communication logs, and development documents, were used to trace how the instructional requirements were translated into executable workflows, and how the major design decisions were negotiated across the educational specialists, content experts, and technical developers. Second, learning-system operation logs recorded the platform interactions, database entries, and student operation events during the five-day camp, providing evidence of how the pedagogical agent and the R&T learning system functioned under classroom constraints. Third, supplementary classroom observations, including field notes, student artifacts, and teacher reflections, were used to contextualize the logged interaction records and support the interpretation of classroom implementation.

Four quantitative instruments were administered as pre-/post-measures. First, a researcher-developed, 10-item, multiple-choice STEAM knowledge test assessed the students' conceptual understanding of the curriculum content, including motors, circuits, suction principles, and material use. The items were presented in Chinese with visual support, and two experienced elementary science teachers reviewed the test for content validity. Second, curiosity was measured using a 10-item adapted scale based on the CEI-II (Kashdan et al., 2009) and the CQ-STEM (Ahmad & Siew, 2021), covering the two dimensions of Exploration and Acceptance using a 5-point Likert scale. The items were linguistically simplified for young learners and reviewed for age appropriateness, and the teacher read each item aloud during administration. Third, a 10-item, picture-based English vocabulary test, developed with reference to the PPVT-4 format (Dunn et al., 2007), assessed the students' acquisition of target curriculum vocabulary, such as octopus, battery, fan, motor, and suction. Each item required the students to select the correct picture after hearing the word read aloud. Fourth, hands-on performance was evaluated using the Creative Product Assessment Matrix (CPAM; Besemer & Treffinger, 1981), which measures Novelty, Resolution, and Elaboration & Synthesis across nine indicators using a 5-point scale. Two experienced teachers jointly rated the products, and inter-rater agreement was examined using Kendall's W.

2.8 Data Analysis Procedures

The qualitative data were analyzed through thematic and content-based procedures. The expert collaboration records were coded to identify recurring issues and collaborative strategies, such as lesson-to-script transformation, bilingual alignment, and the handling of technical constraints. A chronological mapping approach was then used to reconstruct how the major design decisions evolved from initial planning to classroom implementation. The learning-system logs were analyzed through content categorization of technical and instructional issues, including recognition errors, interaction breakdowns, and workflow interruptions, and were interpreted together with classroom observations and teacher reflections.

The quantitative data were analyzed using both descriptive and inferential statistics. The internal consistency of the questionnaire-based measures was examined with Cronbach's α, and expert review was used to support the content validity of the researcher-developed instruments. Normality was assessed using the Shapiro–Wilk test. For non-normally distributed variables, pre-/post-test differences were examined using Wilcoxon signed-rank tests. For group-level differences in hands-on performance, one-way analysis of variance (ANOVA) and Scheffé post-hoc comparisons were conducted. Effect sizes were reported for selected analyses where appropriate, and Kendall's W was used to confirm inter-rater agreement for the CPAM scoring.

For analyses involving dyad composition and hands-on performance, the dyad was treated as the unit of analysis because the students completed the construction tasks collaboratively. For pre-/post-test learning outcomes, individual student scores were analyzed, while dyad grouping was used as a classification variable.

3. Results

3.1 Outcomes of Interdisciplinary Collaboration

Interdisciplinary collaboration played a central role in the design and implementation of the LLM-based pedagogical agent × R&T learning system. The interdisciplinary team consisted of educational experts, content specialists, and technical experts, who contributed complementary expertise throughout the development process. The educational experts defined the learning goals and curriculum structure according to the cognitive and learning characteristics of lower elementary students. The content specialists translated the lesson plans into executable instructional flows and prompt-scripted interactions, whereas the technical experts developed the platform architecture and integrated the pedagogical agent, educational robots, and tangible teaching aids into the learning system.

The collaboration process followed an iterative design–test–revise cycle. Through recurring workshops and meetings, the team established a shared instructional framework and developed a standardized instructional design template that improved cross-domain coordination and workflow efficiency. As a result, the team produced scripted workflows capable of guiding QR-code scanning, supporting voice-based interaction, and handling student work uploads. Classroom trials further showed that the pedagogical agent could support bounded instructional tasks such as story introduction, component recognition, and step-by-step task guidance. In addition, the lesson-plan conversion process enhanced the reusability of the learning system across the activities and topics.

3.2 Technical and Instructional Challenges

During classroom implementation, the LLM-based pedagogical agent × R&T learning system demonstrated clear potential for supporting interactive learning, but also revealed several technical and instructional challenges. Technically, the system was constrained by token consumption, speech-recognition errors, and response latency, as well as instability in vision-based judgment and device connectivity. To mitigate these issues, token-intensive operations were routed to higher-tier accounts, audio input settings were refined, fallback interaction modes were prepared, and selected verification procedures were simplified to tactile confirmation. Visual designs and retry mechanisms were also adjusted to reduce ambiguity and interruption during task progression.

Instructionally, the younger students often engaged in trial-and-error exploration without fully following the prompts, making it more difficult for the scripted workflows to infer their learning states. In addition, the system could not effectively manage multiple simultaneous voice inputs or regulate group-level pacing. As a result, teacher and assistant intervention was essential for maintaining order, clarifying prompts, and ensuring smooth transitions. Overall, these findings indicated that the pedagogical agent and educational robots functioned as instructional support rather than teacher substitutes, and that effective classroom implementation still depended on timely human coordination. These challenges affected interaction continuity, increased the need for teacher intervention, and occasionally interrupted the students' task progression, thereby influencing both classroom interaction and learning support.

3.3 Quantitative Analysis of Learning Outcomes

To complement the qualitative findings, quantitative analyses were conducted to examine the students' learning outcomes after their participation in the bilingual STEAM intervention. Because the Shapiro–Wilk test indicated that several variables deviated from normality (see Table 2), Wilcoxon signed-rank tests were used for pre-/post-test comparisons.

Table 2.
Results of the Shapiro–Wilk Normality Test

Variables	W Statistics	*p* Values	Normal Distribution
Pre_STEAM	0.8694	0.0008	No
Post_STEAM	0.5228	< 0.001	No
Pre_Curiosity	0.9381	0.0632	Yes
Post_Curiosity	0.9247	0.0401	No
Pre_PPVT	0.9495	0.1226	Yes
Post_PPVT	0.9159	0.0368	No

As shown in Table 3, the students demonstrated significant improvements in STEAM knowledge, curiosity, and English vocabulary. STEAM knowledge increased significantly from pre-test to post-test (Z = -5.086, $p < 0.001$, r = 0.872). Overall curiosity also improved significantly (Z = -2.889, p = 0.019, r = 0.496). In addition, English vocabulary measured by the PPVT increased significantly (Z = -4.283, $p < 0.001$, r = 0.734). These results indicated that the R&T learning system with the integrated pedagogical agent supported both cognitive and language-related learning outcomes in the bilingual hands-on STEAM activities.

Table 3.
Wilcoxon Signed-rank Test Results (Overall Sample)

Variables	Pre-test Means	Post-test Means	*Z*	*p*	*r*	Significant
STEAM	7.62	9.74	-5.086	< 0.001	0.872	Yes
Curiosity	37.21	40.12	-2.889	0.019	0.496	Yes
PPVT	6.38	8.26	-4.283	< 0.001	0.734	Yes

Subgroup analyses based on dyad composition further revealed differentiated patterns of improvement (see Table 4). All three dyad types—High–High (HH), High–Low (HL), and Low–Low (LL)—showed significant gains in STEAM knowledge, which indicated robust improvement in conceptual understanding across the proficiency pairings. In addition, the HL dyad group showed a significant increase in English vocabulary (p = 0.011), and the LL dyad group also showed significant vocabulary improvement (p = 0.007). Although overall curiosity did not differ significantly across the proficiency pairings, the findings suggested that the intervention supported STEAM learning across all three dyad types, while producing variations in the language- and curiosity-related outcomes depending on the initial English-proficiency pairing.

Table 4.
Wilcoxon Signed-rank Test Results by Dyad Type

Dyads	Variables	Pre-test Means	Post-test Means	*Z*	*p*	*r*	Significant
1	STEAM	8.50	10.00	-2.521	0.039	0.891	Yes
(HH)	Curiosity	39.88	41.75	-1.400	0.599	0.495	No
	PPVT	8.38	9.00	-1.610	0.188	0.569	No
2	STEAM	7.50	9.60	-2.803	0.007	0.886	Yes
(HL)	Curiosity	36.50	40.30	-2.039	0.074	0.645	No
	PPVT	7.00	8.90	-2.803	0.011	0.886	Yes
3	STEAM	7.25	9.69	-3.516	< 0.001	0.879	Yes
(LL)	Curiosity	36.31	39.19	-1.706	0.093	0.427	No
	PPVT	5.00	7.50	-2.844	0.007	0.711	Yes

Group-level hands-on performance also differed significantly across the dyad types (see Table 5). One-way ANOVA showed significant group differences in Novelty ($F = 5.377$, $p = 0.0099$); Resolution ($F = 38.086$, $p < 0.001$); Elaboration & Synthesis ($F = 11.731$, $p < 0.001$); and Overall Performance ($F = 23.098$, $p < 0.001$). Scheffé post-hoc comparisons (see Table 6) further indicated that the HH dyad group significantly outperformed the LL dyad group across all four dimensions, whereas the HL dyad group also performed significantly better than the LL dyad group on most measures. Overall, these findings suggested that the dyads with higher or mixed-English proficiency demonstrated stronger creative and problem-solving performance in the bilingual STEAM context.

Table 5.
One-way ANOVA Results for Creative Product Performance

Dimensions	HH Means	HL Means	LL Means	*F*	*p*	Significant
Novelty	7.42	7.10	5.92	5.377	0.0099	Yes
Resolution	19.00	14.10	11.92	38.086	< 0.001	Yes
Elaboration & Synthesis	12.50	10.00	8.67	11.731	< 0.001	Yes
Overall Performance	38.92	31.20	26.50	23.098	< 0.001	Yes

Table 6.
Scheffé Post-hoc Comparisons

Dimensions	Dyad Comparisons	Mean Diff.	*F*	*p*	Significant
Novelty	HH vs HL	0.317	0.435	0.651	No
	HH vs LL	1.500	10.741	0.0002	Yes
	HL vs LL	1.183	6.077	0.0055	Yes
Resolution	HH vs HL	4.900	34.837	< 0.001	Yes
	HH vs LL	7.083	80.079	< 0.001	Yes
	HL vs LL	2.183	5.413	0.0078	Yes
Elaboration & Synthesis	HH vs HL	1.000	6.121	0.0054	Yes
	HH vs LL	2.167	17.992	< 0.001	Yes
	HL vs LL	1.167	6.673	0.0041	Yes
Overall Performance	HH vs HL	7.717	25.781	< 0.001	Yes
	HH vs LL	11.717	56.493	< 0.001	Yes
	HL vs LL	4.000	3.640	0.032	Yes

3.4 Validation of Results and Learning Log Analysis

The learning-log data were further analyzed to triangulate the quantitative findings by examining the differences in task duration, response accuracy, and interaction patterns across the three dyad types. As shown in Table 7, the HH dyad group completed tasks in the shortest total time and achieved the highest average accuracy, whereas the HL and LL groups required longer completion times and showed lower average accuracy. These patterns suggested that the HH dyad group engaged in a relatively efficient and accurate mode of task completion, whereas the other two dyads appeared to require more time and support during the activities.

At the phase level, the QR-code scanning phase showed relatively stable performance across the three dyad types, whereas the drawing and voice-feedback phases revealed clearer differentiation. The HL dyad group spent more time on the listening and drawing-related phases, which suggested more exploratory engagement during task completion. The LL dyad group showed longer voice-feedback interaction, consistent with the interpretation that lower-proficiency learners benefited more from the system's supportive verbal scaffolding. By contrast, the HH dyad group showed the strongest accuracy and the most efficient task completion, reflecting stronger task management and the integration of prompts into performance.

The learning-log patterns were broadly consistent with the statistical findings reported above. The HH dyad group showed the strongest hands-on performance, the HL dyad group showed an intermediate performance with greater exploratory engagement, and the LL dyad group demonstrated lower efficiency but stronger reliance on system support. Taken together, the convergence of the learning-log evidence and the statistical results suggested that the pedagogical agent × R&T learning system were able to support different learner profiles while revealing meaningful differences in how the dyads engaged with the bilingual STEAM tasks.

Table 7.
Analysis of the AI Agent Learning-log Patterns across the Learning Phases

Indicators	HH Means	SDs	HL Means	SDs	LL Means	SDs	Statistical Trends
Listening to the Story	835.97	1945.41	921.45	2313.39	789.18	1761.99	The HL dyad group spent the longest time; the LL dyad group spent the shortest time.
Accuracy (Listening)	0.02	0.14	0.00	0.00	0.06	0.24	Small differences among dyads; considered an input-oriented task.
Answering Story Questions	291.78	635.28	293.26	608.43	307.81	688.25	Similar across dyads, indicating effective AI scaffolding.
Accuracy (Question Response)	2.29	4.91	2.29	4.67	2.26	4.98	Slightly higher in the HL dyad group , but not significant.
Scanning QR-codes	349.30	761.01	354.98	732.46	362.75	818.49	Stable performance across dyads; no significant differences.
Drawing	826.00	901.70	1380.80	1086.20	1232.60	945.30	The HL and LL groups spent considerably more time.
Voice Feedback	913.40	1008.50	318.80	420.30	512.80	637.70	The HH dyad group showed the most active interaction.
Total Task Time	1840.80	750.20	2718.60	923.40	2632.00	905.30	The HL and LL groups spent more time overall.
Total Correct Score	3.00	0.00	2.80	0.40	2.60	0.50	The HH dyad group achieved the highest accuracy.
Average Accuracy per Lesson	0.60	0.00	0.56	0.08	0.52	0.10	Reflected consistent ability-level differences.

4. Discussion

4.1 Addressing the Research Findings

This study examined how interdisciplinary collaboration supported the integration of an LLM-based pedagogical agent, educational robots, and an R&T (Robots & Tangible Objects) learning system into bilingual hands-on STEAM activities for lower elementary students. Overall, the findings suggested that the pedagogical agent × R&T learning system was feasible for classroom use when supported by a teacher–agent–robot coordinated model. Rather than functioning independently, the learning system was most effective when the pedagogical agent provided prompt-scripted guidance, the robots served as embodied interaction channels, and the teacher and teaching assistants offered timely support for clarification, pacing, and recovery from breakdowns.

The subgroup findings further indicated that the learning system supported different learner profiles in different ways. The dyads with higher or mixed-English proficiency demonstrated stronger hands-on and creative performance, whereas the lower-proficiency dyads appeared to benefit more from structured verbal support. These

patterns suggested that the pedagogical agent × R&T learning system not only supported bilingual STEAM learning across all three dyad types but also revealed meaningful variations in how the students engaged with the task procedures, language input, and interaction support.

4.2 Comparison with Related Research

The present findings are consistent with those in prior studies that showed that pedagogical agents can enhance engagement, guidance, and motivation in learning environments (Baylor & Kim, 2005; Johnson et al., 2000). At the same time, this study extends previous work by showing that in a lower elementary bilingual STEAM setting, the value of a pedagogical agent is more evident when it is combined with embodied interaction and tangible task structures. In this sense, the present study adds empirical support for the integration of language-mediated guidance with robot-mediated and tangible learning processes.

From an embodied learning perspective, physical manipulation and multimodal interaction helped the learners to connect language, action, and conceptual understanding (Lindgren & Johnson-Glenberg, 2013; Wilson, 2002). In the present study, QR-code scanning, component identification, and stepwise assembly created repeated do–observe–revise cycles that supported procedural clarity and sustained attention. These findings are also consistent with research suggesting that structured tangible activities can promote interaction, inquiry, and feedback in learning environments (Avila-Garzon et al., 2021). The current study contributes to this literature by showing that such support can be implemented in a lower elementary bilingual STEAM context through relatively low-cost tangible structures and scripted guidance.

The findings also align with prior work emphasizing that classroom technologies should support, rather than replace, teachers (Holmes et al., 2022). In this study, classroom implementation still depended on teacher and assistant support to manage simultaneous interactions, maintain pacing, and recover from workflow interruptions. In addition, the development process echoed the co-creation perspectives in STEM education design (Honey et al., 2014), as classroom viability depended on iterative alignment between pedagogical intentions and technical implementation. The study therefore also complements recent R&T research by showing that robot-mediated interaction and tangible task structures can jointly support both conceptual learning and vocabulary development in bilingual activities (Chen & Yeh, 2025a, 2025b; Hsieh et al., 2025).

4.3 Integrated Interpretation of Group Differences

The subgroup findings suggested that the pedagogical agent × R&T learning system supported different learner profiles in different ways. The HH dyad group showed the most efficient task completion and the strongest hands-on performance, which suggested stronger self-monitoring and integration of bilingual and procedural cues. The HL dyad group appeared to benefit from more exploratory engagement and peer negotiation, which explained its stronger language-related gains and intermediate performance outcomes. By contrast, the LL dyad group showed lower efficiency overall but greater reliance on voice-based support, which suggested that structured prompts and supportive verbal feedback were particularly important for sustaining participation

among lower-proficiency learners.

Taken together, the convergence between the statistical results and the learning-log evidence suggested that the system not only supported learning improvement but also captured meaningful differences in how the students engaged with the bilingual STEAM tasks. This pattern highlights the importance of considering learner pairing and interaction support when designing pedagogical agent learning systems for young learners.

5. Limitations

This study has several limitations. First, as a five-day pilot study with a small sample from a single school, its findings have limited generalizability. Second, the pre–post design without random assignment or a control group restricted strong causal interpretation. Third, because the intervention combined multiple components, the contribution of any single element could not be isolated. In addition, the measures may still have been affected by the young learners' attention and comprehension variability, and classroom implementation was constrained by technical instability, such as recognition errors, latency, and connectivity problems. Future research should include larger samples, comparison conditions, and delayed or transfer measures to strengthen the evidence base.

6. Conclusion

This study has contributed to the growing literature on AI-supported education by showing how an LLM-based pedagogical agent, educational robots, and tangible learning tools can be integrated into bilingual hands-on STEAM activities for lower elementary students. The findings suggest that when prompt-scripted guidance, robot-mediated interaction, and tangible task structures are coherently coordinated with teacher support, such learning systems can promote STEAM knowledge development, curiosity, vocabulary learning, and hands-on performance in classroom settings. Rather than functioning as teacher substitutes, pedagogical agents and educational robots are most effective when positioned as instructional collaborators within a broader teacher–agent–tool ecosystem.

Building on these findings, the study has made three main contributions. First, it has demonstrated the feasibility of a teacher–agent–robot coordinated model for bilingual STEAM learning in a lower elementary classroom. Second, it has provided a replicable process for converting lesson plans into executable pedagogical agent workflows through interdisciplinary collaboration. Third, it has offered practical evidence that embodied interaction, tangible task structures, and scripted bilingual support can be combined to support young learners' participation and learning. Overall, the study offers both conceptual and practical implications for the design of classroom-ready AI-supported learning environments in elementary education. The study also contributes theoretically by extending the discussion of pedagogical agents from language-based support to coordinated teacher–agent–robot interaction in embodied bilingual STEAM learning.

References

Ahmad, C. N. C., & Siew, N. M. (2021). Development of a STEM curiosity questionnaire for primary school students. *Journal of Baltic Science Education, 20*(1), 5-19. https://doi.org/10.33225/jbse/21.20.05

Avila-Garzon, C., Bacca-Acosta, J., Kinshuk, Duarte, J., & Betancourt, J. (2021). Augmented reality in education: An overview of twenty-five years of research. *Contemporary Educational Technology, 13*(1), ep295.

Baylor, A. L., & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education, 15*(2), 95-115.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science Robotics, 3*(21), eaat5954. https://doi.org/10.1126/scirobotics.aat5954

Belpaeme, T., & Tanaka, F. (2021). Social robots as educators. *OECD Digital Education Outlook 2021: Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots,* 143-162. https://doi.org/10.1787/589b283f-en

Besemer, S. P., & Treffinger, D. J. (1981). Analysis of creative products: Review and synthesis. *The Journal of Creative Behavior, 15*(3), 158-178. https://doi.org/10.1002/j.2162-6057.1981.tb00287.x

Chapelle, C. (2024). Open generative AI changes a lot, but not everything. *The Modern Language Journal*. https://doi.org/10.1111/modl.12927

Chen, N. S. (2025). Harnessing large language models for education: A framework for designing effective pedagogical AI agents. In *E-learning and enhancing soft skills: Contemporary models of education in the era of artificial intelligence* (pp. 13-35). Springer.

Chen, N. S., & Yeh, H. C. (2025a). Designing educational robots for preschool classrooms: Challenges, solutions, and collaborative insights from content, pedagogy, and technology experts. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-025-01297-4

Chen, N. S., & Yeh, H. C. (2025b). Emulating real world environment for contextual language learning using educational robot and IoT-enabled tangible objects. In *AI mediated language education in the metaverse era* (Chinese Language Learning Sciences, pp. 147-169). Springer. https://doi.org/10.1007/978-981-95-0245-5_7

Corrigan, M., Wong, J., Grove, D., Andersen, S., & Hughes, B. (2025). Enhancing elementary students' conceptual understandings of scientific phenomena: The impact of STEAM-first and STEM-first approaches. *Science Education.* https://doi.org/10.1002/sce.21942

Dicks, J., & Genesee, F. (2016). Bilingual education in Canada. In O. García, A. M. Y. Lin, & S. May (Eds.), *Bilingual and multilingual education* (pp. 453-467). Springer. https://doi.org/10.1007/978-3-319-02324-3_32-1

Dunn, L. M., Dunn, D. M., Pearson Assessments, & Psych Corp. (2007). *PPVT-4: Peabody picture vocabulary test* (4th ed.). Pearson Assessments.

Feng, X., Chen, N. S., Wang, D., & Peng, B. (2023). Effects of mobile-assisted emotional robot on EFL learners' oral proficiency: Perspectives from the cognitive and affective domains. *Computer Assisted Language Learning, 36*(7), 1302-1333. https://doi.org/10.1080/09588221.2021.2006024

Fung, H. Y., Yeh, H. C., & Chen, N. S. (2025). Effects of a social robot-supported tangible learning approach on learners' foreign language vocabulary acquisition, motivation, and willingness to communicate. *Interactive Learning Environments, 33*(1), 56-75. https://doi.org/10.1080/10494820.2023.2167274

Holmes, W., Bialik, M., & Fadel, C. (2022). *Artificial intelligence in education:*

Promises and implications for teaching and learning. Brookings Institution Press.
Honey, M., Pearson, G., & Schweingruber, H. (Eds.). (2014). *STEM integration in K-12 education: Status, prospects, and an agenda for research*. National Academies Press.
Hsieh, W. M., Yeh, H. C., & Chen, N. S. (2025). Impact of a robot and tangible object (R&T) integrated learning system on elementary EFL learners' English pronunciation and willingness to communicate. *Computer Assisted Language Learning, 38*(4), 773-798. https://doi.org/10.1080/09588221.2023.2186608
Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education, 11*(1), 47-78.
Kashdan, T. B., Gallagher, M. W., Silvia, P. J., Winterstein, B. P., Breen, W. E., Terhar, D., & Steger, M. F. (2009). The curiosity and exploration inventory-II: Development, factor structure, and psychometrics. *Journal of Research in Personality, 43*(6), 987-998. https://doi.org/10.1016/j.jrp.2009.04.011
Lan, Y. J., & Chen, N. S. (2024). Teachers' agency in the era of LLM and generative AI. *Educational Technology & Society, 27*(1), I-XVIII. https://doi.org/10.30191/ETS.202401_27(1).0017
Lindgren, R., & Johnson-Glenberg, M. (2013). Emboldened by embodiment: Six precepts for research on embodied learning and mixed reality. *Educational Researcher, 42*(8), 445-452.
Mohsen, M., Khalil, I., Alzahrani, A., Awaji, B., & Abdullah, K. (2025). Synthesizing STEAM integration in language education literature: Trends, themes, and outcomes. *International Journal of Multilingualism.* Advance online publication. https://doi.org/10.1080/14790718.2024.2423664
Mubin, O., Stevens, C. J., Shahid, S., Al Mahmud, A., & Dong, J. J. (2013). A review of the applicability of robots in education. *Technology for Education and Learning, 1*(1), 1-7.
Pai, K. C., Kuo, B. C., Liao, C. H., & Liu, Y. M. (2024). An application of Chinese robot-assisted language learning: Development and learner authentication. *Computer Applications in Engineering Education, 32*(1), e22667. https://doi.org/10.1002/cae.22667
Quigley, C., Herro, D., King, E., & Plank, H. (2020). STEAM designed and enacted: Understanding the process of design and implementation of STEAM curriculum in an elementary school. *Journal of Science Education and Technology, 29*(4), 499-518. https://doi.org/10.1007/s10956-020-09832-w
Relmasira, S., Lai, Y., & Donaldson, J. (2023). Fostering AI literacy in elementary science, technology, engineering, art, and mathematics (STEAM) education in the age of generative AI. *Sustainability*. https://doi.org/10.3390/su151813595
Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9*(4), 625-636.
Yang, W. (2021). Problem-based CLIL: An efficacious pathway to enhancing students' STEM performance, STEM motivation, and 21st century skills. *SAGE Open, 11*(4), 21582440211061363. https://doi.org/10.1177/21582440211061363