International Electronic Journal of Mathematics Education
2026, 21(1), em0862
e-ISSN: 1306-3030
https://www.iejme.com

MODESTUM

Research Article

OPEN ACCESS

Intelligent teaching design assistant for primary mathematics: A large language model-driven framework with retrieval-augmented generation and problem-chain pedagogy

Danna Tang [1], Ran Ding [2,3*], Meng He [4], Yushen Wang [5], Kaka Cheng [6]

[1] School of Mechanical Engineering, Suzhou University of Technology, CHINA
[2] School of Educational Science, Anhui Normal University, CHINA
[3] Xinhua Middle School, Gedian Economic & Technological Development Zone, Ezhou, CHINA
[4] Academy of Arts & Design, Tsinghua University, CHINA
[5] School of Engineering and Materials Science, Queen Mary University of London, UK
[6] Department of Mechanical Engineering, Imperial College London, UK
***Corresponding Author:** diagran@ahnu.edu.cn

Citation: Tang, D., Ding, R., He, M., Wnag, Y., & Cheng, K. (2026). Intelligent teaching design assistant for primary mathematics: A large language model-driven framework with retrieval-augmented generation and problem-chain pedagogy. *International Electronic Journal of Mathematics Education, 21*(1), em0862. https://doi.org/10.29333/iejme/17447

ARTICLE INFO

Received: 15 May 2025

Accepted: 15 Oct 2025

ABSTRACT

Primary mathematics education faces systemic challenges in translating curriculum reforms into classroom practice, exacerbated by teachers' cognitive overload and limited support for pedagogical innovation. This study develops an Intelligent Teaching Design Assistant grounded in socio-constructivist and cognitive load theories to address these challenges. Thirty-four primary mathematics teachers participated in a quasi-experimental study. The Intelligent Teaching Design Assistant integrates Large Language Models with multi-dimensional knowledge bases (curriculum standards, teaching strategies, student profiles) and a multi-agent architecture (process planner, student simulator). The Intelligent Teaching Design Assistant significantly outperformed generic Large Language Models, improving overall lesson plan quality. This work pioneers a replicable pathway for AI to empower teacher agency and advance 21st-century educational transformation.

Keywords: cognitive load theory, elementary education, teacher professional development, teaching/learning strategies

INTRODUCTION

Due to their core features in generating inspirational content, understanding conversational contexts, and executing sequential tasks. Large language models have had a significant impact and profound implications on the field of education. They may also promote and catalyze deep changes from educational philosophy to educational practice (Kinder et al., 2025). As representatives of the new generation of artificial intelligence technology, large language models have attracted widespread attention from practitioners in the education field due to their advantages in conversational fluency, task processing versatility, and logical reasoning (Yu, 2024). Large Language Models hold transformative potential for education by aligning with socio-constructivist principles (Vygotsky, 1978), where learning is mediated through social interaction and tool-based scaffolding. This study operationalizes Vygotsky's concept of the Zone of Proximal Development by positioning Large Language Models as cognitive partners that co-construct pedagogical content knowledge with teachers. Simultaneously, the design adheres to cognitive load theory (Sweller, 2011), aiming to reduce extraneous cognitive burden through retrieval-augmented generation, thereby enabling teachers to focus on higher-order design tasks.

Large Language Models enhance teaching outcomes by fostering creativity, enabling adaptive digital tutors, innovating pedagogical strategies, and personalizing feedback and assessment (Pandey et al., 2025). Large model technology has also shown great potential in addressing some of the current pain points in education and teaching. Through effective integration with primary and secondary education, Large Language Models can not only reduce teachers' workload but also enhance the innovation and personalization of teaching.

The principles and methods of teaching design have begun to be applied in educational institutions at all levels, as well as in training programs of corporate and public organizations. Teaching design competence has become an essential skill for teachers engaged in modern teaching (Liu et al., 2024; Weng & Chiu, 2023). A teaching plan is a detailed program developed by teachers based on curricula and teaching resources, including teaching content, methods, media, student activities, and assessment. It is

an important educational tool that reflects teachers' teaching design methods and teaching strategies (Farhang et al., 2023). Although teaching design is indispensable in teaching, primary school mathematics teachers face numerous challenges in the actual teaching design process.

Currently, the main issues in primary school mathematics teaching design include vague statements of teaching objectives, direct replication of textbook analysis, arbitrary student analysis, and formalistic design of teaching activities. Unfortunately, the current lesson plan format provides insufficient guidance for teaching design, merely categorizing teaching objectives, preparation, and process in a broad and simple manner, leaving teachers to decide the key points to consider and how to integrate technology into the classroom on their own (Iqbal et al., 2021). These studies reveal the existing problems in teaching design and provide important directions for future research and practice. To enhance the effectiveness of teaching design, future efforts can draw on theories and findings from existing research, in combination with large model technology, to better assist teachers in completing teaching design. This would not only improve teaching effectiveness and efficiency but also better cultivate students' overall qualities and abilities.

The integration of Large Language Models provides a promising pathway for further enhancing the efficiency and quality of teaching design for teachers. Some scholars have pointed out that ChatGPT, as a representative of Large Language Models, can help teachers develop teaching plans, sequence teaching content, and generate classroom questions, saving teachers a significant amount of time in seeking teaching inspiration. This is especially useful for novice teachers (Li et al., 2024; Ma, 2025). Despite the enthusiastic discussions by many scholars about the application prospects of Large Language Models in teaching design, the specific application modes and potential issues in practical use remain unclear. Frontline teachers face some deficiencies when directly using general Large Language Models for teaching design. The research team, involved in curriculum development for frontline teachers and pre-service teachers, found that while teachers show a strong interest in leveraging Large Language Models for education, the high technical threshold poses challenges in understanding and effectively applying this technology.

Further investigations and product design efforts revealed that some teachers reported that the content generated by Large Language Models might contain factual errors and lack clear explanations of teaching objectives and strategies, making it difficult for teachers to understand the logical basis of the generated content. This could lead to discrepancies between the generated teaching content and the teaching objectives, affecting teaching effectiveness. Moreover, the Large Language Models' understanding of specific teaching contexts, student characteristics, and curriculum standards is limited, making it challenging to directly apply the generated teaching schemes in classroom teaching.

In summary, effectively utilizing Large Language Models, designing application models for their use in teaching design, and developing teaching design assistants suitable for teachers are critical issues that need urgent resolution. This study deeply investigates the problems in using general Large Language Models for teaching design. Taking the fifth grade (11~12 years old) mathematics curriculum in primary schools as an example, it uses large model technology to design application models in the teaching design process. Based on this model, the authors adopt the approach of constructing intelligent agents to design a teaching design assistant for primary school mathematics teachers. This assistant aims to improve the efficiency and quality of teaching design, fully leveraging the advantages of large model technology in primary education to address the practical challenges teachers face in teaching design.

LITERATURE REVIEW

Issues in Primary School Mathematics Curriculum Design

The study of curriculum design theory highlights that it serves as a bridge between learning, teaching theory, and teaching practice, directly addressing educational problems. However, the current effectiveness of curriculum design theory is insufficient. Teachers still predominantly rely on experience and traditional methods when designing and preparing lessons, neglecting the application of learned curriculum design theories (Küchemann et al., 2023). There are four main issues in current primary school mathematics curriculum design: Ambiguous articulation of teaching objectives, direct copying of textbook analyses, arbitrary student analysis, and formalized design of teaching activities (DaCosta & Kinsell, 2024). With the introduction of the new curriculum standards, teaching processes, activities, and procedures have become more complex.

Additionally, the integration of interactive screens and mobile terminals has made classroom interactions and student feedback more diverse and flexible. The exponential increase in classroom design elements now compels teachers to meticulously design and plan every segment, activity, and even every instruction. Unfortunately, the current lesson plan format provides insufficient guidance on curriculum design, merely categorizing teaching objectives, preparations, and processes in a broad and simplistic manner. The critical points that need consideration during the design process and the integration of technology into the classroom are left entirely to the teacher's discretion (Hashem et al., 2024).

Survey results on curriculum design for primary and secondary schools indicate that teachers generally view curriculum design as a critical guarantee for improving teaching quality and conducting curriculum design before lessons is a common practice (Meron & Tekmen Araci, 2023). However, teachers typically consider curriculum design as a technical preparation process, a procedural and technical task with highly standardized and operational steps and sequences, resulting in a uniform lesson plan format. While there is an emphasis on knowledge transmission, the cultivation of students' abilities, interests, emotions, and morals is often overlooked, with teachers continuing to rely primarily on teaching experience rather than theory for design. Teachers usually prioritize the preset "basic knowledge and skills" as their main goals, with little adjustment to teaching objectives and content, reflecting insufficiently on curriculum standards and societal developments. In their curriculum design, teachers use fixed procedures, down to the exact minute, aiming for maximum certainty and attempting to eliminate uncertainties.

Early investigations into AI-assisted curriculum design have primarily centered on region-specific systems—particularly Chinese prototypes that leverage retrieval-augmented LLMs to draft learning objectives, align tasks with local syllabi, and generate formative assessments. While these efforts demonstrate the potential of large language models within a single educational context, they leave open questions about how such tools adapt to different pedagogical traditions and regulatory environments.

Recent international studies have begun to fill this gap. In Germany, a GPT-4–based Physics Feedback Bot enhanced undergraduates' hypothesis articulation and experimental planning (Steinert et al., 2024); in Spain, an Instructional Design Matrix systematically mapped ChatGPT-generated content to sustainability learning outcomes with high teacher fidelity (Ruiz-Rojas et al., 2023); in the United States, postgraduate teams treated ChatGPT as a "virtual design colleague," co-authoring syllabi that outperformed instructor-only versions on innovation and real-world relevance (Meron & Tekmen Araci, 2023); and in Japan, EduBot dynamically aligned its prompts to national language standards and student proficiency, yielding significant gains in K–12 oral fluency and grammatical accuracy (Li et al., 2024). Together, these examples highlight four key insights for AI-enhanced curriculum design: the necessity of localization and standards alignment, the power of iterative feedback loops, the promise of multi-agent or "virtual colleague" configurations, and the versatility of LLM tools across disciplines and educational levels.

Realistic Dilemmas of Applying Large Language Models in Curriculum Design

While LLMs can reduce workload and accelerate ideation, their application in curriculum design presents several realistic dilemmas observed in frontline practice. First, content accuracy is not guaranteed, and outputs may include factual or curricular misalignments that require teacher verification. Second, without context rich prompting, generated content tends to be generic and insufficiently tailored to specific learners or standards. Third, the cognitive demand of prompt engineering can be high, especially for teachers new to these tools. Fourth, there is a risk of overreliance on model outputs, potentially dampening teachers' creative decision making and critical reflection. Fifth, general models show limited sensitivity to local curriculum standards, classroom routines, and student characteristics unless explicitly grounded in domain knowledge bases. These dilemmas motivated the research assistant's design choices: Retrieval augmented generation, multidimensional knowledge bases, and a process planner paired with a student simulator to keep design decisions transparent and situated.

Ruiz-Rojas et al. (2023) conducted a survey of 42 university teachers, highlighting that the combination of generative artificial intelligence (AI) tools with curriculum design matrices is crucial for developing large-scale MOOC virtual classrooms. The results demonstrated the potential of generative AI tools in higher education. Moundridou et al. (2024) evaluated the impact of increased prompt specificity on the structure and content of curriculum plans generated by generative AI for language teachers. The study found that the relationship between prompt specificity and the quality of the output or generated curriculum plans was not linear, indicating that higher specificity does not always enhance curriculum design quality. Overall, the research suggests that ChatGPT can simplify curriculum planning, thereby reducing professional workload.

Research on the application of Large Language Models in curriculum design indicates that while these models have significant potential to improve efficiency and quality, they also face numerous challenges. Large language models have limitations in producing highly precise content, requiring teachers to supplement with their expertise to complete detailed educational tasks. Additionally, the generated content is often generic, and creating context-specific content demands substantial manual prompts and editing to achieve the desired effect. Overreliance on these models by teachers could hinder their creativity and critical thinking development. Most studies focus on higher education and specific disciplines, with fewer studies on primary and secondary education, especially across different subject areas.

Current Research on AI Agents in Curriculum Design

The Intelligent Teaching Design Assistant's multi-agent architecture embodies distributed cognition, where pedagogical knowledge is externalized across human-AI interactions (Hutchins, 1995). This approach mitigates individual cognitive load while fostering collective expertise—a critical mechanism for teacher professional development (TPD) in technology-integrated contexts (Matsumoto et al., 2024a). Furthermore, the problem-chain pedagogy is grounded in Bruner's spiral curriculum theory, which emphasizes iterative scaffolding of mathematical concepts to promote deep understanding. By simulating a 'more knowledgeable other' (e.g., through problem-chain prompts), the assistant scaffolds teachers' design process, enabling them to internalize expert strategies. Additionally, the integration of student learning styles into the memory module reflects distributed cognition theory, as the system externalizes pedagogical knowledge, reducing individual cognitive burden.

Research has shown that AI agents, leveraging the core capability of self-reflection in Large Language Models, can use these models as their brains to enhance environmental perception and task-solving abilities, effectively promoting collaborative learning, online learning, and the simulation of learning scenarios in educational settings (Matsumoto et al., 2024b). This study has limitations. First, the small sample size (N = 34) and focus on fifth-grade mathematics may limit generalizability. Future work should validate the framework across subjects and grade levels. Second, the assistant's dependency on pre-constructed knowledge bases requires ongoing updates to align with curriculum reforms.

Georgia Tech pioneered the virtual teaching assistant Jill Watson, based on ChatGPT, which can answer questions related to course content. Jill Watson uses a modular design to integrate new APIs, handle multi-document content, and perform excellently in classrooms while minimizing hallucinations and harmful outputs through safety measures (Taneja et al., 2024). Researchers have proposed the von-Neumann multi-agent system framework, combining four modules and four types of operations to explore the capability enhancement cycle of multi-agent systems in education. Through human-computer collaboration and reflection, this framework promotes learners' knowledge construction and enhancement of teaching abilities. Viswanathan et al. (2022) developed a virtual intelligent teaching assistant (TA) system framework based on a powerful language model (i.e., GPT-3). This

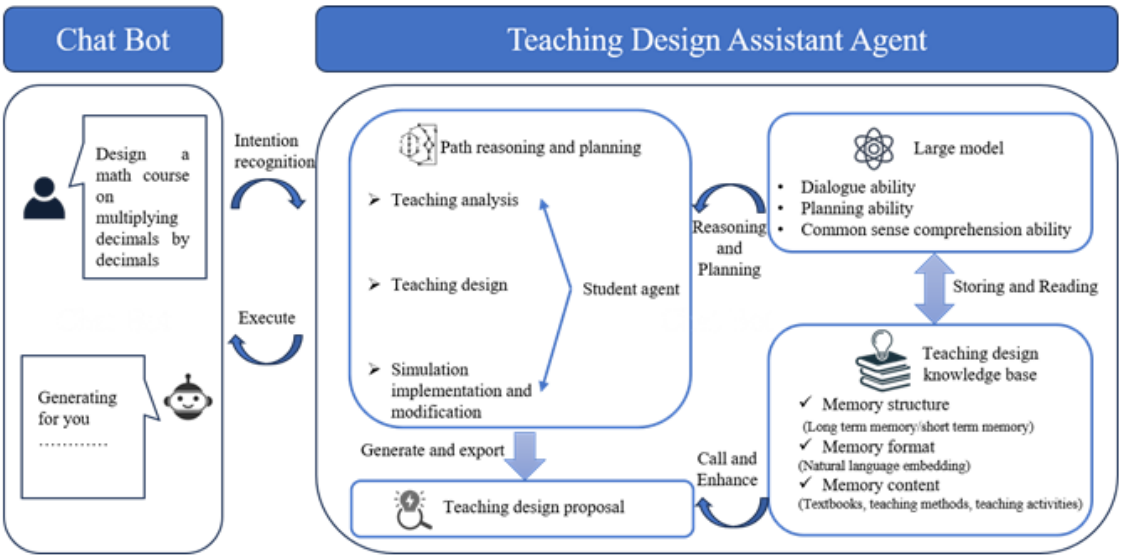

Figure 1. A large model-based teaching design assistant framework (Source: Authors' own elaboration)

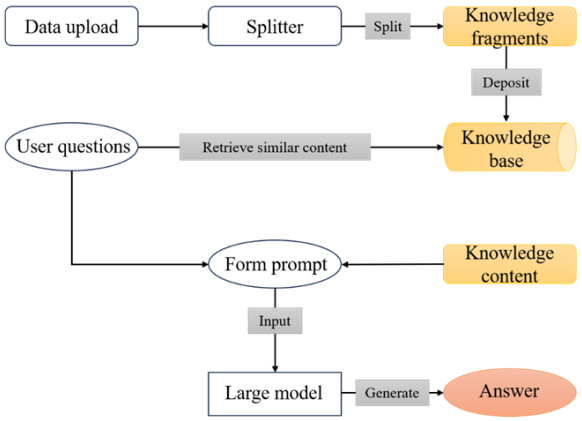

Figure 2. Flowchart of a retrieval-augmented generation based on the Coze platform (Source: Authors' own elaboration)

framework can automatically generate intelligent assistants for specific courses without being limited by subject or academic level. These assistants, equipped with voice functionalities, can answer various course-specific questions, from course content to logistics and course policies.

Additionally, researchers have attempted to use multiple general-purpose agents to simulate classroom interactions between teachers and students. Experimental results show high similarity to real classroom environments in teaching methods, curricula, and student performance (Jinxin et al., 2023) However, there is a lack of in-depth exploration into the targeted design of teaching agents for actual curriculum design by primary and secondary school teachers, failing to comprehensively reflect the needs and challenges faced by primary school mathematics teachers in practice.

METHOD

System Architecture of the Teaching Design Assistant

This study addresses the main issues and needs discovered in the research on primary school mathematics teaching design. By leveraging the rapid development and application of large model technology in the current educational field, this study proposes a teaching design assistant developed through the construction of intelligent agents based on Large Language Models. The assistant aims to improve the precision, professionalism, and efficiency of primary school mathematics teaching design through the deep integration of Large Language Models and knowledge bases. The system architecture is shown in **Figure 1**. The teaching design assistant consists of four core modules: The Analysis and Planning Module, the Memory Module (teaching design content memory), the Execution Module (teaching design generation and export), and the Chatbot Input and Output Module.

The Memory Module ensures the depth and accuracy of the generated teaching designs by constructing multiple knowledge bases, including the Curriculum Knowledge Base, Teaching Strategy Base, and Student Learning Style Base (**Figure 2**). Each module not only relies on these knowledge bases for content retrieval but also utilizes the large model to generate and optimize

Table 1. Composition of experimental subjects (unit: number of people)

Gender (number of people)		Length of teaching (number of people)		
Male (8)	Female (26)	Less than 5 years (25)	5-10 years (8)	More than 10 years (1)

Table 2. Basic information of graded teachers

Number	Gender	The age of the students taught	Length of teaching
Teacher W	Female	11~12 years old	7 years
Teacher L	Female	11~12 years old	7 years
Teacher X	Male	11~12 years old	6 years

Table 3. List of questionnaire items

Dimension	The questions included	Number of questions
Perceived ease of use	4,5,6	3
Perceived usefulness	7,8,9,10,11,12	6
Technical acceptance	13,14,15	3
Satisfaction level	16,17	2

results based on the retrieved content. The content of the knowledge bases is integrated throughout the entire process of teaching analysis, teaching design, and feedback optimization, providing data support and guidance for each stage of teaching.

The large model plays a foundational supporting role in the system, facilitating efficient collaboration between modules through data flows supported by the large model. The knowledge points and student learning information input by teachers are processed in the Teaching Analysis Module, which calls upon the knowledge bases to generate analysis results. These results are then transmitted to the Teaching Design Module, supporting the setting of teaching objectives, the construction of problem chains, and the design of the teaching process.

Platform Usage Detection and Participants

This study recruited elementary school mathematics teachers who had undergone large model training to participate in product experience. The preparatory work before the experiment mainly focused on three aspects:

1) Recruiting teachers,
2) Assigning Task 1, and
3) Preparing tools and environment.

First, 34 elementary school mathematics teachers were recruited (**Table 1**). Task 1 and Task 2 were posted in an online communication group. Task 1 used a general large model, while Task 2 used the teaching assistant designed for this study. In Task 1, the control condition was a general-purpose large language model accessed via its standard web interface, without retrieval-augmented generation and without integration of curriculum standards, teaching strategies, or student profile knowledge bases. Both tasks were identical, requiring participating teachers to complete the lesson design for “Multiplying Decimals by Whole Numbers,” the first unit of Chapter 5 in elementary mathematics, using two different Large Language Models. The specific requirements for the lesson design included teaching objectives, key and difficult points of teaching, and the design of the teaching process. The content had to meet the standards of the elementary mathematics curriculum; the activities had to be practical and applicable in real classrooms. Finally, the teachers output their designs in a Word document, recording the number of dialogue rounds with the large model, the completion time, the optimization steps, and the encountered problems. All participating teachers provided written informed consent after receiving a detailed explanation of the research objectives and data handling procedures. To ensure confidentiality, participant identities were anonymized during data collection, and all records were stored on password-protected servers accessible exclusively to the research team.

Additionally, three elementary school fifth-grade mathematics teachers with at least five years of teaching experience were invited to compare, evaluate, and score the lesson designs submitted by the teachers for Task 1 and Task 2, as shown in **Table 1**. The basic information is listed in **Table 2**. Descriptive statistical analysis was then conducted.

Data Analysis

A post-use survey was conducted among the teachers listed in **Table 1** who completed Tasks 1 and 2. To ensure a comprehensive evaluation of this study, both questionnaires and semi-structured interviews were administered to verify the practicality of the Intelligent Teaching Design Assistant. The questionnaire employed a Likert five-point scale, where respondents selected the option that best matched their views from the following: A (strongly agree), B (agree), C (neutral), D (disagree), E (strongly disagree). The survey focused on four dimensions to assess the assistant’s usage: Perceived ease of use, perceived usefulness, technology acceptance, and satisfaction. These dimensions were specifically designed to evaluate the support and functionality provided by the assistant for teachers’ lesson planning. There was a total of 14 questions, and the four dimensions along with their corresponding questions are shown in **Table 3**. During this process, feedback and issues encountered by teachers while using the assistant were collected to identify any shortcomings of the Intelligent Teaching Design Assistant in the context of elementary mathematics education.

Table 4. List of dimensions for teaching design evaluation

First level indicator	Secondary indicators	Number of indicators
Teaching objectives	T1, T2	2
Teaching situation and textbook analysis	S1, S2	2
Problem Chain	Q1, Q2, Q3	3
Teaching activities	A1, A2	2
Knowledge Content	C1, C2, C3	3
Teaching Methods and Strategies	M1, M2	2
Teaching evaluation	E1, E2	2
Usability	V1	1
Range	R1	1
Overall rating	F	1

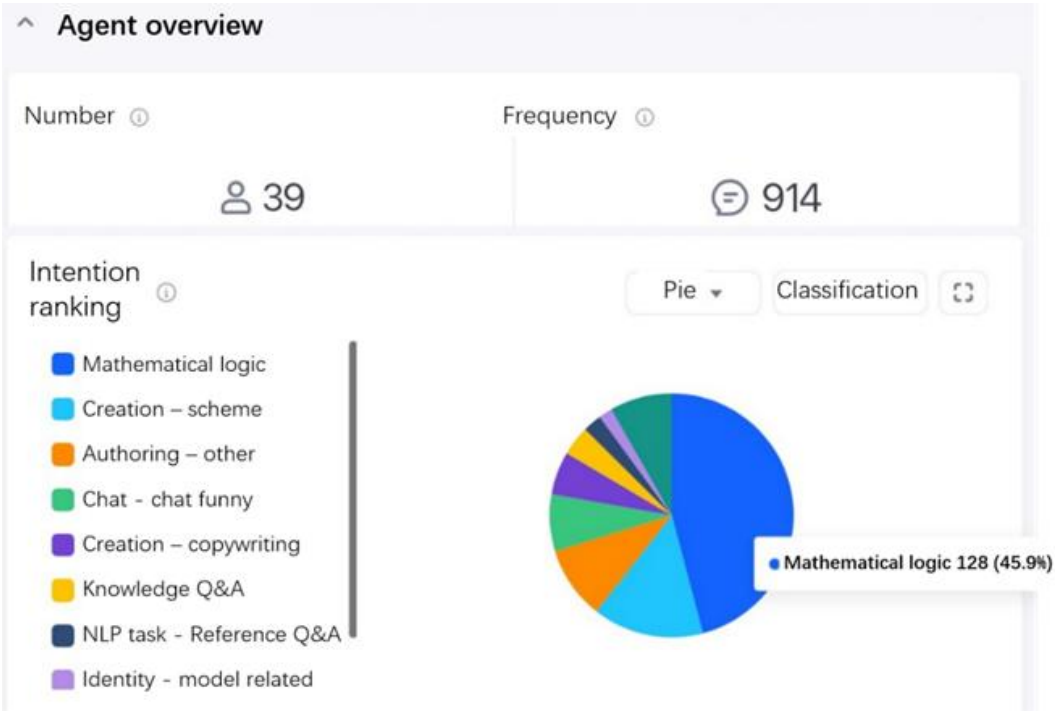

Figure 3. Background data overview (Source: Authors' own elaboration)

To assess whether the teaching design assistant had an impact on the quality of lesson plans created by the teachers in **Table 2**, an evaluation scale was developed, drawing on established evaluation tools and referencing the lesson design evaluation criteria from the Tian Jiabing Teaching Skills Competition. This evaluation tool encompassed a total of 19 items across 10 dimensions. The scoring used an eight-point Likert scale, where 1 represented "strongly disagree" and 8 represented "strongly agree." The evaluation dimensions and their corresponding items are shown in **Table 4**. The evaluation rubric was developed by drawing on established tools and explicitly referencing the lesson design evaluation criteria from the Tian Jiabing Teaching Skills Competition, and the rating process followed shared scoring dimensions and examples provided to raters prior to evaluation. Formal statistical assumption checks for the paired t tests (e.g., distributional normality of paired differences and homogeneity of variances) were not conducted or reported in the present study; we acknowledge this as a methodological limitation and address it in the Discussion.

RESULTS

Platform Usage Back-End Data

Back-end data showed that during the experiment, a total of 39 teachers registered to use the teaching design assistant (**Figure 3**). Excluding the author and one teacher user who initially tested the system, 34 users were invited through the experiment, and there were 3 new users, indicating a good promotional effect of the assistant among teachers. Teachers conducted a total of 914 dialogues, with the intent of generating a complete teaching design, accounting for 45.9% of the dialogues. Taken together, these usage patterns—multiple short sessions per user, modest rounds per session, and a high proportion of dialogues aimed at producing complete plans—suggest sustained, workflow integrated engagement rather than sporadic or exploratory use. Teachers appeared to rely on the assistant to iteratively clarify objectives, refine problem chains, and finalize exportable lesson plans, consistent with time boxed cycles of classroom preparation.

During the week of Task 2, the average number of active users was 6, representing 35.3% of the total users, reflecting the practicality and acceptance of the assistant in daily teaching design (**Figure 4**). In terms of interaction data, the average number

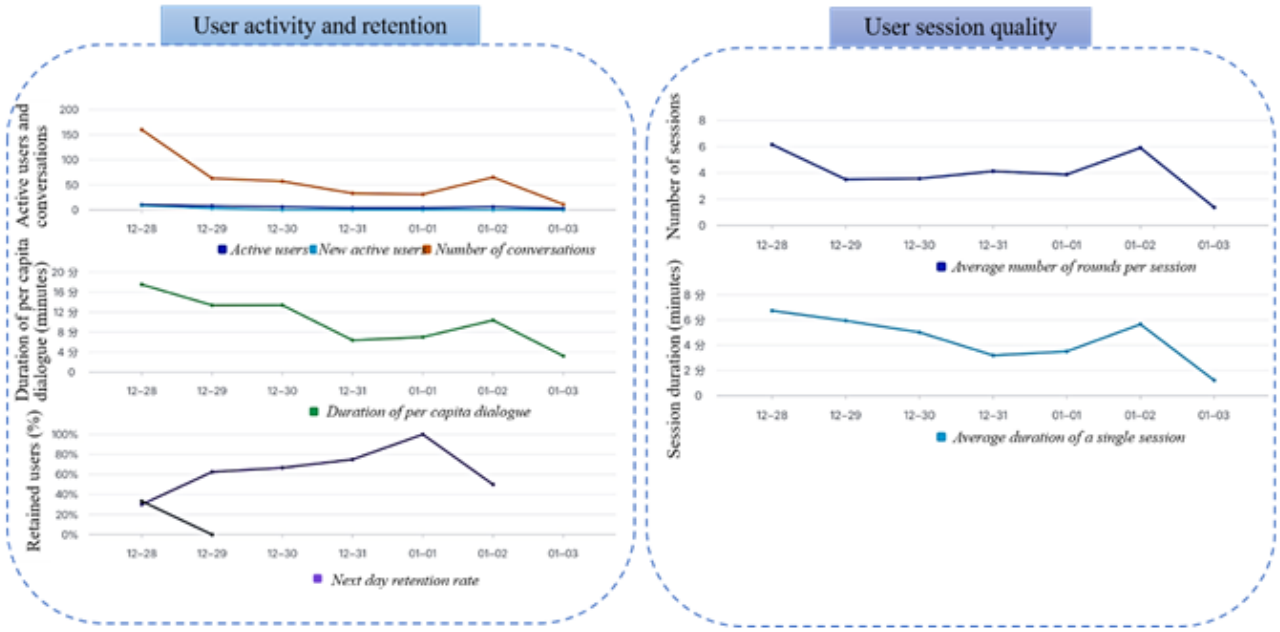

Figure 4. User usage data (Source: Authors' own elaboration)

Table 5. Cronbach reliability analysis

Number of items	Sample sizes	Cronbach α coefficient
14	34	0.881

of dialogues per user was 10.2, the average dialogue duration per user was 8.8 minutes, the average number of rounds per session was 3.7, and the average session duration was 3 minutes and 55 seconds. This data indicates that teachers interacted with the system multiple times during their use of the assistant, with each interaction time being reasonably controlled. Each session, consisting of four dialogue rounds, took about 4 minutes to solve a problem. The back-end data showed positive usage of the assistant's features and a high level of teacher engagement.

Evaluation of Platform Usage Experience

The survey questionnaire mainly focused on whether the designed teaching assistant met the daily teaching design needs of teachers and whether the system's functionality was scientifically sound. After Task 2 was completed, the author distributed the "Teaching Design Assistant Usage Experience Survey" to the 34 elementary mathematics teachers who participated in this trial. A total of 34 questionnaires were collected, with a return rate and validity rate of 100%. The reliability coefficient (Cronbach's alpha) of the questionnaire was found to be 0.881 through reliability and validity tests, indicating a high level of reliability (**Table 5**).

The evaluation of platform usage experience was conducted from four dimensions: perceived ease of use, perceived usefulness, technology acceptance, and satisfaction. The data were categorized and statistically analyzed based on effective percentage, mean, and standard deviation. The scores for each option were entered as follows: Option A (strongly agree) and Option B (agree) represented positive feedback attitudes, with A scoring 5 points and B scoring 4 points; Option C (neutral) indicated a neutral attitude, scoring 3 points; Option D (disagree) and Option E (strongly disagree) represented negative feedback attitudes, with D scoring 2 points and E scoring 1 point. Data was entered, organized, and analyzed using Excel and SPSS. The specific statistical results are shown in **Table 6**.

Table 6. Trial experience

Dimesion	Question items	Effective percentage					Mean (M)	Standard deviation (SD)
		A (%)	B (%)	C (%)	D (%)	E (%)		
Perceived ease of use	4	82.4	17.6	0	0	0	4.824	0.387
	5	50	50	0	0	0	4.500	0.508
	6	82.4	17.6	0	0	0	4.824	0.387
Perceived usefulness	7	50%	50	0	0	0	4.500	0.508
	8	32.4	67.6	0	0	0	4.324	0.475
	9	32.4	67.6	0	0	0	4.324	0.475
	10	50.0	50.0	0	0	0	4.500	0.508
	11	58.8	41.2	0	0	0	4.588	0.500
	12	55.9	44.1	0	0	0	4.559	0.504
Technical acceptance	13	41.2	26.5	33.3	0	0	4.147	0.821
	14	35.3	64.7	0	0	0	4.353	0.485
	15	26.5	41.2	32.4	0	0	3.941	0.776
Satisfaction level	16	23.5	76.5	0	0	0	4.235	0.431
	17	32.4	67.6	0	0	0	4.324	0.475

Table 7. Correlation coefficient results within ICC groups

Bidirectional mixing/random consistency	ICC intra group correlation coefficient	95% CI
Single metric ICC (C, 1)	0.885	0.875 ~ 0.895
Average metric ICC (C, K)	0.958	0.954 ~ 0.962

Note: C represents consistency, 1 represents single measure, K represents average measure

Table 8. Scores of teachers' works

Name (Task 2)	Sample sizes	Minimum	Maximum	M	SD	Median
Teaching objectives	34	5.833	7.000	6.279	0.306	6.333
Teaching focus and difficulties	34	6.833	7.667	7.221	0.271	7.167
Problem chain	34	4.222	6.778	6.026	0.655	6.111
Teaching activities	34	4.833	6.500	5.691	0.532	5.750
Knowledge content	34	3.778	5.000	4.389	0.383	4.333
Teaching methods and strategies	34	5.167	6.833	6.000	0.516	5.833
Teaching evaluation	34	4.333	5.833	5.029	0.497	5.000
Availability	34	5.333	7.000	5.990	0.583	6.167
Knowledge scope	34	4.333	6.333	5.480	0.610	5.500
Total rating	34	4.333	6.667	5.804	0.598	5.667

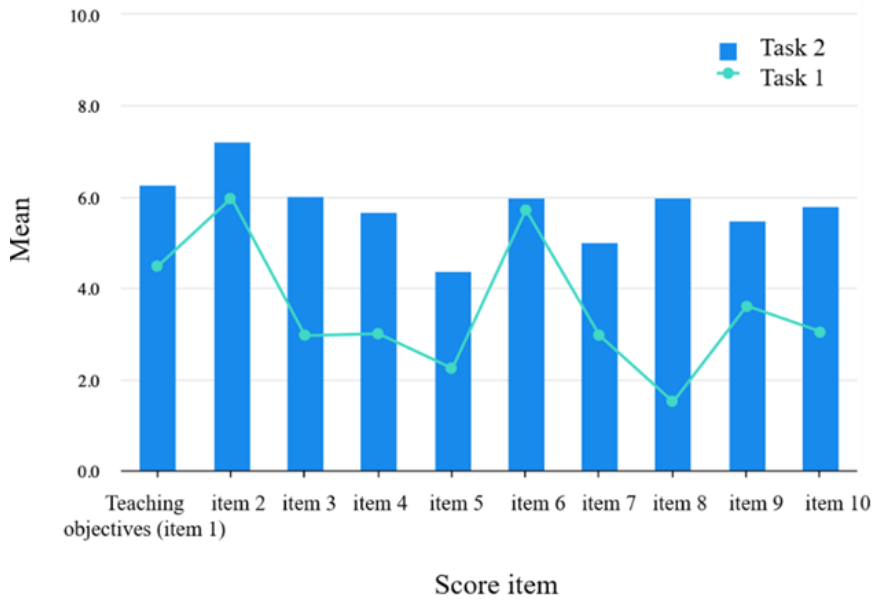

Figure 5. Comparison of the mean values of task 1 and task 2 (Source: Authors' own elaboration)

Evaluation of Platform-Generated Lesson Plans

The analysis of the platform-generated teaching designs aimed to examine whether the teaching design assistant effectively enhanced the teachers' lesson planning. To ensure uniformity in the evaluation process, the study provided all participating teachers from **Table 1** with detailed explanations of the scoring dimensions and scoring examples. To support consistent scoring, raters reviewed the common scoring dimensions and worked through provided scoring examples; each rater then devoted a minimum of eight minutes to each lesson plan prior to assigning scores according to the shared criteria. The inter-rater reliability was calculated using the two-way random-effects model intraclass correlation coefficient for absolute agreement using average measures (ICC [2, 3]) among the three raters to ensure the credibility of the scoring results. The single measure ICC (C, 1) was 0.885, and the average measure ICC (C, K) was 0.958, indicating good consistency in the teachers' evaluations of the generated lesson plans (**Table 7**).

The sample size for the dataset "Teaching Designs Using the Intelligent Teaching Assistant" for Task 2 is N = 34. The score of each option is entered into the following way: 1-4 points for low availability and 5-8 points for high availability. The average score is calculated as the final score of teachers' teaching design. The specific details are shown in **Table 8** and **Figure 5**.

To further verify whether the differences in scores between Task 1 and Task 2 are statistically significant, this study used paired sample t-tests to analyze the significance of scores across various dimensions. The significantly higher overall score of Task 2 ($M = 5.83$ vs. $M = 3.67$, $p < 0.01$) suggests that the LLM-based assistant effectively bridges the gap between generic AI outputs and domain-specific teaching needs. The alignment with national curriculum standards was assessed through the Teaching situation and textbook analysis dimension, where Intelligent Teaching Design Assistant-generated designs achieved significantly higher scores ($M = 7.22$, $SD = 0.28$) compared to generic Large Language Models ($M = 4.79$, $SD = 1.22$), $t(33) = -6.328$, $p < 0.01$ (**Table 9**). This improvement aligns with prior findings on retrieval-augmented generation enhancing content accuracy (Ruiz-Rojas et al., 2023). Moreover, Task 2 demonstrated an overall advantage over Task 1 across all dimensions. However, the improvement effects varied among different dimensions.

Table 9. Paired t-test analysis results

Task 1 pairing Task 2	Pairing (mean ± standard deviation)		Difference (Pairing 1-Pairing 2)	*t*	*p*
	Pairing 1	Pairing 2			
Teaching objectives	4.19±0.97	6.26±0.31	-1.97	-5.896	0.000**
Teaching situation and textbook analysis	4.79±1.22	7.22±0.28	-2.42	-6.328	0.000**
Problem chain	3.58±0.79	6.05±0.66	-2.43	-8.938	0.000**
Teaching activities	3.60±0.95	5.69±0.54	-2.07	-10.218	0.000**
Knowledge content	3.29±0.57	4.40±0.40	-1.08	-7.916	0.000**
Teaching methods and strategies	4.69±0.74	5.99±0.52	-1.29	-6.455	0.000**
Teaching evaluation	3.58±0.72	5.01±0.50	-1.39	-5.820	0.000**
Availability	3.00±1.10	6.00±0.59	-2.94	-9.781	0.000**
Knowledge scope	3.84±1.17	5.50±0.63	-1.61	-5.119	0.000**
Total rating	3.62±1.00	5.83±0.61	-2.17	-8.006	0.000**

DISCUSSION

Good Experience and Acceptance of the Teaching Design Assistant

This study has several methodological limitations. The sample size was modest (N = 34) and focused on a single subject and grade level (fifth grade mathematics), which constrains generalizability. In addition, although we report strong inter-rater reliability, the authors did not conduct or report formal statistical assumption checks for the paired t tests, nor did we compute effect sizes from paired difference data. These omissions, together with the assistant's lower scores in knowledge content and evaluation design, indicate that future work should broaden samples, incorporate assumption checks and standardized effect sizes, and strengthen support for interdisciplinary content and formative assessment. Relative to other AI assisted teaching tools cited in the literature—such as modular virtual teaching assistants emphasizing safety and multi document handling, or multi agent frameworks that automate course specific Q&A—our approach is distinctive in its tight coupling of retrieval augmented generation with curriculum standards and in its multi agent design that includes a process planner and student simulator. This coupling explicitly foregrounds teacher agency in goal setting, problem chain construction, and activity design. At the same time, potential unintended consequences must be considered: overreliance on AI suggestions could dampen teachers' creative exploration; lagging updates to knowledge bases could misalign plans with evolving standards; and biases in source materials could be reflected in outputs. We mitigate these risks through transparent grounding to standards, teachers facing rationales for generated elements, prompts that require teacher critique before export, and a maintenance workflow for knowledge base updates aligned to curriculum reforms.

The survey conducted among teachers who completed Tasks 1 and 2 indicates that the overall experience of using the teaching design assistant among elementary mathematics teachers was quite positive. The perceived ease of use dimension received the highest average score (M = 4.72, SD = 0.39), indicating that teachers found the assistant intuitive to operate. Teachers generally found the system easy to operate, with high consistency in their evaluations (the lowest standard deviation was 0.387). The Intelligent Teaching Design Assistant's effectiveness in reducing cognitive load (M = 4.72 for perceived ease of use) aligns with Sweller's cognitive load theory, as offloading routine tasks (e.g., curriculum standard alignment) allowed teachers to prioritize creative pedagogical decisions—a finding critical for teacher professional development in technology-dense environments.

The dimension of perceived usefulness had an average score of 4.472, reflecting a high level of recognition from teachers regarding the practical utility of the teaching design assistant, particularly in supporting the setting of teaching objectives and the design of activities (with an average score of 4.574). Although the overall scores for technology acceptance were relatively high (averaging between 3.941 and 4.353), the standard deviations for items 13 and 15 were relatively larger (0.853 and 0.793, respectively), indicating some variability among teachers in accepting and using this technology. This may reflect a lower adaptability to new technology or a lack of deep understanding of system functions for some teachers.

The dimension of satisfaction had an average score of 4.292, indicating overall satisfaction with the teaching design assistant. Overall, teachers expressed high satisfaction with the assistant, believing it met their teaching design needs. However, lower scores from some teachers in the technology acceptance dimension and the variability in feedback suggest that there is still room for improvement in terms of technical operation or understanding of system functions.

Significant Quality Improvement Effect of the Teaching Design Assistant

Based on the evaluation results, the overall score for instructional design is 5.80, indicating that the quality of instructional design has reached a usable level after the use of the instructional design assistant. Among the various dimensions, "Teaching focus and difficulties " received the highest score of 7.22, followed by "Teaching objectives" with a score of 6.26, demonstrating significant support for teachers in these two aspects. In contrast, the dimensions of "Knowledge content" and "Teaching evaluation" received relatively lower scores, 4.39 and 5.02 respectively, suggesting that there is still room for improvement in the instructional design assistant's support for expanding knowledge content and designing evaluations.

A further analysis of the lower score in "Knowledge content" reflects the assistant's shortcomings in integrating interdisciplinary content. Additionally, the standard deviations for "Problem chain" and "Knowledge scope" are relatively large, at 0.66 and 0.63 respectively, indicating substantial variance in scores among teachers in these dimensions. This suggests that the effectiveness of the instructional design assistant in these areas may be influenced by teachers' proficiency and individual

differences. The high usability scores (M = 5.99) can be explained through technology acceptance model (TAM) (Davis, 1989). Teachers perceived the assistant as both useful (e.g., reducing design time) and easy to use (e.g., intuitive interface), which are key determinants of TAM's perceived usefulness and ease of use. However, the variability in technology acceptance scores (SD = 0.82) suggests that individual differences in teacher self-efficacy may influence adoption (Bandura, 1997). Less confident teachers might require additional training to leverage the assistant's full potential.

When comparing the average scores of Task 2 and Task 1 using a combination chart (see **Figure 5**), it becomes evident that the instructional design assistant has significantly improved the quality of teachers' instructional designs. The overall score for Task 2 (5.83) is notably higher than Task 1 (3.67), and this trend is observed across all dimensions. However, the degree of improvement varies among the dimensions. The largest difference is seen in usability, with a difference of 2.94, followed by question design and analysis of key and difficult points of teaching, indicating that the instructional design intelligent assistant has a significant optimization effect in these areas.

While the assistant improved teaching strategy selection (ΔM = 1.29), its impact on interdisciplinary integration (ΔM = 1.08) and teaching evaluation (ΔM = 1.39) was less pronounced. This aligns with prior findings that Large Language Models struggle to autonomously bridge domain-specific knowledge gaps without explicit guidance (Küchemann et al., 2023). To address this, future systems could adopt a scaffolding framework, where the assistant prompts teachers to articulate connections between mathematical concepts and other disciplines (e.g., 'How might decimal multiplication relate to measuring ingredients in a recipe?').

Further Optimization Directions for the Teaching Design Assistant

The results in **Table 9** show that the p-values for all dimensions are 0.000 ($p < 0.01$), indicating that the differences between Task 1 and Task 2 are statistically significant. Notably, the t-values and score differences for the "Question Chain (Q)" (t = -8.938) and "Teaching Activities (A)" (t = -10.218) are prominent, demonstrating the teaching design assistant's strong impact in supporting teachers in designing high-quality question chains and teaching activities.

The score difference for "Usability (V)" is 2.94, highlighting the high applicability of the assistant-generated plans for practical use. Although the score differences for "Knowledge Content (C)" and "Teaching Methods and Strategies (M)" are relatively small (1.08 and 1.29, respectively), these differences are statistically significant, indicating that there is still room for improvement in the assistant's support for knowledge selection and teaching strategy design.

The assistant demonstrates strongest efficacy in supporting teaching situation analysis (M = 7.22) and problem chain design (M = 6.03), likely due to the retrieval-augmented generation mechanism that grounds outputs in curriculum standards. However, its limited support for interdisciplinary knowledge integration (M = 4.39) suggests that future iterations should incorporate cross-domain knowledge graphs (e.g., linking mathematics to real-world scenarios) to enhance contextual relevance. (Task 2: M = 4.39) The lower scores in knowledge content and teaching evaluation (Task 2: M = 5.02) suggest two limitations. First, the current knowledge base lacks interdisciplinary connections (e.g., linking decimal multiplication to real-world financial literacy), which is critical for fostering students' transferable skills. Second, the assistant's evaluation templates overemphasizing summative assessments (e.g., quizzes), neglecting formative strategies (e.g., peer feedback). Future iterations should integrate situated learning theory by embedding contextualized evaluation prompts (e.g., 'Design a peer review rubric for group problem-solving').

Through descriptive statistics, comparative analysis, and paired t-test verification, this study fully demonstrates the practical value of the teaching design assistant in enhancing the quality of lesson plans. This study also contributes to AI-in-education theory by demonstrating how Large Language Models can be pedagogically grounded through retrieval-augmented generation and problem-chain scaffolding. Unlike generic AI tools, our framework operationalizes socio-constructivist principles, proving that AI systems can act as cognitive partners rather than mere content generators. This aligns with recent calls for 'theory-driven AIED', where technology design is rooted in learning sciences.

To operationalize these directions, implementation should include scheduled knowledge base updates aligned with curriculum revision cycles; teacher facing scaffolds that prompt explicit cross disciplinary connections during problem chain design (e.g., linking decimal multiplication to measurement and financial literacy contexts); embedded formative assessment templates and peer review rubrics within the assistant's evaluation module; and an iterative improvement loop in which teacher feedback on generated objectives, problem chains, activities, and evaluations is collected and used to refine retrieval prompts and generation patterns across successive design cycles.

CONCLUSION

This study demonstrates the significant potential of integrating large language models with pedagogical frameworks to enhance primary mathematics teaching design. The proposed Intelligent Teaching Design Assistant, grounded in retrieval-augmented generation and problem-chain pedagogy, effectively addresses the limitations of generic Large Language Models in accuracy and contextual adaptability. Empirical results from 34 teachers revealed that the Intelligent Teaching Design Assistant substantially improved lesson plan quality, with overall scores increasing from 3.67 (Task 1: Generic Large Language Models) to 5.83 (Task 2: Intelligent Teaching Design Assistant) ($p < 0.01$). Key improvements were observed in critical dimensions: Teaching objectives (ΔM = 2.07, $p < 0.01$), problem-chain design (ΔM = 2.45, $p < 0.01$), and usability (ΔM = 2.99, $p < 0.01$). Teachers reported high perceived ease of use (M = 4.72) and satisfaction (M = 4.29), reflecting reduced cognitive load through AI-human collaboration, consistent with Sweller's cognitive load theory.

The Intelligent Teaching Design Assistant's plugin functionality streamlined administrative tasks by automating Word document generation, showcasing practical scalability. However, challenges persist in interdisciplinary knowledge integration (M = 4.39) and formative assessment design. The framework's alignment with the national curriculum standards underscores its potential as a policy-responsive tool for curriculum reform. Theoretically, this work repositions Large Language Models as pedagogical partners within a socio-constructivist framework, emphasizing collaborative knowledge construction over automation. Limitations include the focus on fifth-grade content and small sample size.

Future work will prioritize three directions. First, the assistant will be extended to additional subjects and grade levels to assess the generalizability of retrieval-augmented, multi agent design supports beyond fifth grade mathematics. Second, interdisciplinary knowledge bases will be expanded and continuously curated to connect mathematical concepts to real-world contexts, thereby strengthening support for knowledge content breadth and formative evaluation design. Third, the framework will be tested across diverse school contexts to examine its robustness and equity implications. For practical adoption, a replicable pathway grounded in the present architecture is offered. Schools should assemble a curriculum standards base, a teaching strategy base, and a student learning profile base; configure the process planner and student simulator agents to surface rationales and anticipate learner responses; and establish a review loop in which teachers articulate objectives and constraints, receive grounded drafts, critique and adjust problem chains and activities, and only then export Word lesson plans. Professional development should focus on aligning objectives to standards, critically appraising AI outputs, and embedding the assistant into lesson study cycles so that teachers can collectively reflect on and improve generated designs while maintaining pedagogical ownership.

Author contributions: DT: conceptualization, methodology, funding acquisition, writing - original draft; **RD:** supervision, project administration, writing - review & editing; **MH:** software, formal analysis; **YW:** data curation, investigation; **KC:** validation, visualization. All authors have agreed with the results and conclusions.

Acknowledgment: The authors would like to express their sincere gratitude to the universities in Jiangsu Province, the National Natural Science Foundation of China, and the Natural Science Foundation of the Jiangsu Higher Education Institutions for their support and contribution to the completion of this research.

Funding: This project was funded by the General Project of Philosophy and Social Science Research in Colleges and Universities of Jiangsu Province (Grant No. 2024SJYB1050), the National Natural Science Foundation of China (Grant No. 52405340), and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 24KJB460001).

Ethical statement: The authors stated that this study followed ethical guidelines for educational research. The study was considered a low-risk study and ethics committee approval was waived by the institutional ethics committee of Anhui Normal University 1 May 2024 (Document number: AHNU-ERB-2024-0105). Participation was anonymous and voluntary, with the right to withdraw at any time. Data was used solely for academic purposes and stored securely.

AI statement: The authors stated that, during the preparation of this work, the authors used ChatGPT in order to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W H Freeman/Times Books/ Henry Holt & Co.

DaCosta, B., & Kinsell, C. (2024). Investigating media selection through ChatGPT: An exploratory study on generative artificial intelligence in the aid of instructional design. *Open Journal of Social Sciences*, *12*(4), 187-227. https://doi.org/10.4236/jss.2024.124014

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

Farhang, A., Hashemi, A., & Ghorianfar, A. (2023). Lesson plan and its importance in teaching process. *International Journal of Current Science Research Review*, *06*(08), 5901-5913. https://doi.org/10.47191/ijcsrr/V6-i8-57

Hashem, R., Ali, N., El Zein, F., Fidalgo, P., & Abu Khurma, O. (2024). AI to the rescue: Exploring the potential of ChatGPT as a teacher ally for workload relief and burnout prevention. *Research and Practice in Technology Enhanced Learning*, *19*, Article 023. https://doi.org/10.58459/rptel.2024.19023

Hutchins, E. (1995). *Cognition in the Wild*. The MIT Press. https://doi.org/10.7551/mitpress/1881.001.0001

Iqbal, M. H., Siddiqie, S. A., & Mazid, M. A. (2021). Rethinking theories of lesson plan for effective teaching and learning. *Social Sciences & Humanities Open*, *4*(1), Article 100172. https://doi.org/10.1016/j.ssaho.2021.100172

Jinxin, S., Jiabao, Z., Yilei, W., Xingjiao, W., Jiawen, L., & Liang, H. (2023). *CGMI: Configurable general multi-agent interaction framework*. Cornell University.

Kinder, A., Briese, F. J., Jacobs, M., Dern, N., Glodny, N., Jacobs, S., & Leßmann, S. (2025). Effects of adaptive feedback generated by a large language model: A case study in teacher education. *Computers and Education: Artificial Intelligence*, *8*, Article 100349. https://doi.org/10.1016/j.caeai.2024.100349

Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K. E., & Kuhn, J. (2023). Can ChatGPT support prospective teachers in physics task development? *Physical Review Physics Education Research*, *19*(2), Article 020128. https://doi.org/10.1103/PhysRevPhysEducRes.19.020128

Li, Y., Qu, S., Shen, J., Min, S., & Yu, Z. (2024). Curriculum-driven edubot: A framework for developing language learning Chatbots through synthesizing conversational data. In T. Kawahara, V. Demberg, S. Ultes, K. Inoue, S. Mehri, D. Howcroft, & K. Komatani (Eds.), *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue* Kyoto. https://doi.org/10.18653/v1/2024.sigdial-1.35

Liu, L., Hew, K. F., & Du, J. (2024). Design principles for supporting self-regulated learning in flipped classrooms: A systematic review. *International Journal of Educational Research*, *124*, Article 102319. https://doi.org/https://doi.org/10.1016/j.ijer.2024.102319

Ma, T. (2025). Systematically visualizing ChatGPT used in higher education: Publication trend, disciplinary domains, research themes, adoption and acceptance. *Computers and Education: Artificial Intelligence*, *8*, Article 100336. https://doi.org/https://doi.org/10.1016/j.caeai.2024.100336

Matsumoto, M., Polli, J. R., Swaminathan, S. K., Datta, K., Kampershroer, C., Fortin, M. C., Salian-Mehta, S., Dave, R., Yang, Z., Arora, P., Hiura, M., Suzuki, M., Brennan, F. R., & Sathish, J. (2024a). Beyond MABEL: AniIntegrative approach to first in human dose selection of immunomodulators by the health and environmental sciences institute (HESI) immuno-safety technical committee (ITC). *Clinical Pharmacology & Therapeutics*, *116*(3), 546-562. https://doi.org/10.1002/cpt.3316

Matsumoto, T., Nishikawa, R., & Morimoto, C. (2024b). Reflection through interaction with digital twin AI in the human-AI-collaboration SECI Model. *Procedia Computer Science*, *246*, 3743-3752. https://doi.org/10.1016/j.procs.2024.09.182

Meron, Y., & Tekmen Araci, Y. (2023). Artificial intelligence in design education: Evaluating ChatGPT as a virtual colleague for post-graduate course development. *Design Science*, *9*, Article e30. https://doi.org/10.1017/dsj.2023.28

Moundridou, M., Matzakos, N., & Doukakis, S. (2024). Generative AI tools as educators' assistants: Designing and implementing inquiry-based lesson plans. *Computers and Education: Artificial Intelligence*, *7*, Article 100277. https://doi.org/10.1016/j.caeai.2024.100277

Pandey, H. L., Bhusal, P. C., & Niraula, S. (2025). Large language models and digital multimodal composition in the first-year composition classrooms: An encroachment and/or enhancement dilemma. *Computers and Composition*, *75*, Article 102892. https://doi.org/10.1016/j.compcom.2024.102892

Ruiz-Rojas, L. I., Acosta-Vargas, P., De-Moreta-Llovet, J., & Gonzalez-Rodriguez, M. (2023). Empowering education with generative artificial intelligence tools: Approach with an instructional design matrix. *Sustainability*, *15*(15), Article 11524. https://doi.org/10.3390/su151511524

Steinert, S., Avila, K. E., Ruzika, S., Kuhn, J., & Küchemann, S. (2024). Harnessing large language models to develop research-based learning assistants for formative feedback. *Smart Learning Environments*, *11*(1), Article 62. https://doi.org/10.1186/s40561-024-00354-1

Sweller, J. (2011). Cognitive load theory. In *The psychology of learning and motivation: Cognition in education,* Vol. 55 (pp. 37-76). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Taneja, K., Maiti, P., Kakar, S., Guruprasad, P., Rao, S., & Goel, A. K. (2024). Jill Watson: A virtual teaching assistant powered by ChatGPT. In A. M. Onley, I. A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial Intelligence in Education* (Vol. 14829). Springer. https://doi.org/10.1007/978-3-031-64302-6_23

Viswanathan, N., Meacham, S., & Adedoyin, F. F. (2022). Enhancement of online education system by using a multi-agent approach. *Computers and Education: Artificial Intelligence*, *3*, Article 100057. https://doi.org/10.1016/j.caeai.2022.100057

Vygotsky, L. S. (1978). *Mind in society development of higher psychological processes*. Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4

Weng, X., & Chiu, T. K. F. (2023). Instructional design and learning outcomes of intelligent computer assisted language learning: Systematic review in the field. *Computers and Education: Artificial Intelligence*, *4*, Article 100117. https://doi.org/https://doi.org/10.1016/j.caeai.2022.100117

Yu, H. (2024). The application and challenges of ChatGPT in educational transformation: New demands for teachers' roles. *Heliyon*, *10*(2), Article e24289. https://doi.org/10.1016/j.heliyon.2024.e24289