npj | science of learning

Article

Published in partnership with The University of Queensland

https://doi.org/10.1038/s41539-025-00300-x

Exploring the potential of LLM to enhance teaching plans through teaching simulation

Bihao Hu [1,2], Jiayi Zhu[3], Yiying Pei[4] & Xiaoqing Gu [5]

The introduction of large language models (LLMs) may change future pedagogical practices. Current research mainly focuses on the use of LLMs to tutor students, while the exploration of LLMs' potential to assist teachers is limited. Taking high school mathematics as an example, we propose a method that utilizes LLMs to enhance the quality of teaching plans through guiding the LLM to simulate teacher-student interactions, generate teaching reflections, and subsequently direct the LLM to refine the teaching plan by integrating these teaching process and reflections. Human evaluation results show that this method significantly elevates the quality of the original teaching plans generated directly by LLM. The improved teaching plans are comparable to high-quality ones crafted by human teachers across various assessment dimensions and knowledge modules. This approach provides a pre-class rehearsal simulation and ideas for teaching plan refinement, offering practical evidence for the widespread application of LLMs in teaching preparation.

Large language models (LLMs), with their exceptional capabilities of natural language understanding and generation[1], have driven the development of educational applications in human-computer interaction, such as intelligent tutoring systems and teaching assistants, offering new opportunities for the transformation of teaching and learning paradigms[2]. Particularly in student tutoring, agents built on LLMs can undertake most of the teaching tasks of human teachers. By playing the roles of virtual tutors, assistants, and learning peers, these LLMs provide students with personalized, interactive, and engaging learning experiences[3–6]. While LLMs demonstrate potential in enhancing students' learning, there is little research that explores how LLMs may support teachers in their teaching[7], especially the extent to which the use of LLMs can influence various teaching processes, such as teaching preparation, classroom instruction, and post-class reflection and improvement[8]. Some researchers conducted preliminary research on using LLMs to support teaching, but the results were mixed. For instance, some used LLMs to evaluate teachers' classroom performance, but the accuracy of these evaluations and the coherence of the assessment content required further improvement[9]. Other researchers applied LLMs in teacher training programs, finding that these LLMs struggle to provide innovative, consistent, and valuable teaching guidance[10]. Researchers also used LLMs to generate teaching materials, including course outlines, teaching manuals, and exercises. Yet, these materials were not systematically evaluated and validated, and their content often lacked specificity and practicality[11]. Current research indicates that LLMs still require improvement in understanding the complexities of the teaching process and generating diverse teaching content. In this study, we propose a method that uses LLM to enhance the quality of teaching plans, which can improve the LLM's understanding of teaching, and better support teachers in teaching preparation. We asked the LLM to simulate teaching processes and generate teaching reflections based on original teaching plans. Drawing from these simulations and reflections, the LLM was guided to generate new teaching plans. Human evaluations indicate that the improved teaching plans can reach the high quality level of those written by skilled human teachers.

A teaching plan is a teacher's instructional design created during teaching preparation. It serves as a pre-set plan that guides classroom instruction, reflecting the teacher's teaching ideas and strategies for the specific lesson[12]. Experienced teachers, when crafting teaching plans, would take students' actual needs into consideration and design suitable teaching content and methods[13]. However, novice teachers, due to their limited practical teaching experience, often struggle to foresee students' needs and potential learning challenges[14]. Consequently, their teaching plans tend to be generic, sometimes assembled from exemplary teaching plans by other teachers, and often lack depth and coherence[15,16]. As a result, novice teachers often teach the same content in several classes to gain a more comprehensive understanding of students' real needs and challenges in the learning process. This practice enables them to accumulate teaching experience and enhance

[1]School of Computer Science and Technology, East China Normal University, Shanghai, China. [2]Shanghai Institute of Artificial Intelligence for Education, East China Normal University, Shanghai, China. [3]Shanghai Weiyu High School, Shanghai, China. [4]Department of Education, East China Normal University, Shanghai, China. [5]Department of Education Information Technology, East China Normal University, Shanghai, China. e-mail: xqgu@ses.ecnu.edu.cn

the specificity and quality of their teaching plans. Positive and harmonious teacher-student interactions during class can provide timely feedback on students' learning progress, and help teachers identify difficulties students encounter. However, a high student-to-teacher ratio makes it challenging for teachers to capture individualized learning needs and difficulties of every student[17]. This is particularly evident in middle and upper grades, where most students, due to shyness or unwillingness to reveal their lack of knowledge, refrain from actively seeking help from teachers. Only a few students approach teachers after class for assistance. Consequently, it becomes difficult for teachers to comprehensively identify specific obstacles faced by students and obtain adequate feedback on their learning, making the process of improving teaching plans and accumulating teaching experience more difficult. Additionally, teaching plans are mostly revised and optimized during post-class teaching reflections, which means that the improved, higher-quality teaching plans primarily benefits students in subsequent classes. Unless additional class time is allocated for supplemental instruction, the benefits for students in the current class remain limited. Therefore, it becomes particularly important to foresee potential learning difficulties that students may encounter during instruction and to adjust teaching plan content accordingly in advance[18].

Teachers often reflect on the challenges students encounter during a lesson and subsequently revise and refine their teaching plans[14], we propose that LLMs may be able to anticipate students' potential learning challenges before the actual teaching takes place. This approach could help teachers to optimize the content of their teaching plans in advance, reduce teaching risks, and enhance the quality, ensuring that all students benefit from the improved instructional materials. The key to achieving this lies in the role-playing capabilities of LLMs[19]. In this study, we designed prompt commands to make LLM simulate classroom interactions between a teacher and students of varying ability levels. LLM first simulated the teaching process based on the content of the teaching plan, including scenarios where students encountered learning challenges, such as providing incorrect answers due to conceptual confusion. Then, using LLM's reflection and error-correction capabilities[20], we used LLM to generate teaching reflections based on the simulated teaching process. Afterward, based on the teaching process and reflection texts, the LLM was instructed to make corresponding improvements and optimizations to the original teaching plan. Finally, by evaluating the enhanced teaching plans, we explored the potential of using LLM simulations and reflections to improve the quality of teaching plans. This study addresses two main research questions: (1) Can LLM enhance the quality of teaching plans by simulating the teaching process and generating reflections? (2) How do the teaching plans generated through this approach perform across various evaluation dimensions and different knowledge modules?

When designing teaching plans, teachers usually set up instructional objectives, analyze teaching content, select teaching methods, design teaching contexts and activities, and consider ways to evaluate students' learning outcomes based on curriculum standards[21,22]. This process reflects the teacher's Content Knowledge (CK) and Pedagogical Knowledge (PK), as well as their ability to effectively integrate them. This corresponds to the theory of Pedagogical Content Knowledge (PCK), which emphasizes that teacher transform subject-specific knowledge into forms that are accessible and comprehensible for students[23]. Also, teachers often incorporate pre-designed mathematical problems into their teaching plans, providing opportunities for students to construct knowledge and concepts[24]. Well-designed problems are organized into a sequence of interconnected questions, forming a mathematical problem chain. This chain links various knowledge points throughout the lesson, which enables students to understand and master mathematical concepts in a gradual way[25]. In our previous research[26], we incorporated content knowledge and pedagogical knowledge into prompts based on the PCK theory[27], and established an output format for teaching plans generated by LLMs using mathematical problem chains. Subsequently, we utilized GPT-4 to generate teaching plans for all mathematics lessons in high school[28]. Furthermore, an evaluation framework was developed to comprehensively assess the quality of the generated teaching plans. The results demonstrate that the teaching plans generated by GPT-4 perform well in areas such as establishing instructional objectives, identifying teaching priorities and challenges, designing teaching activities, and summarizing classroom teaching knowledge. However, there remains a gap between GPT-4 generated teaching plans and those high-

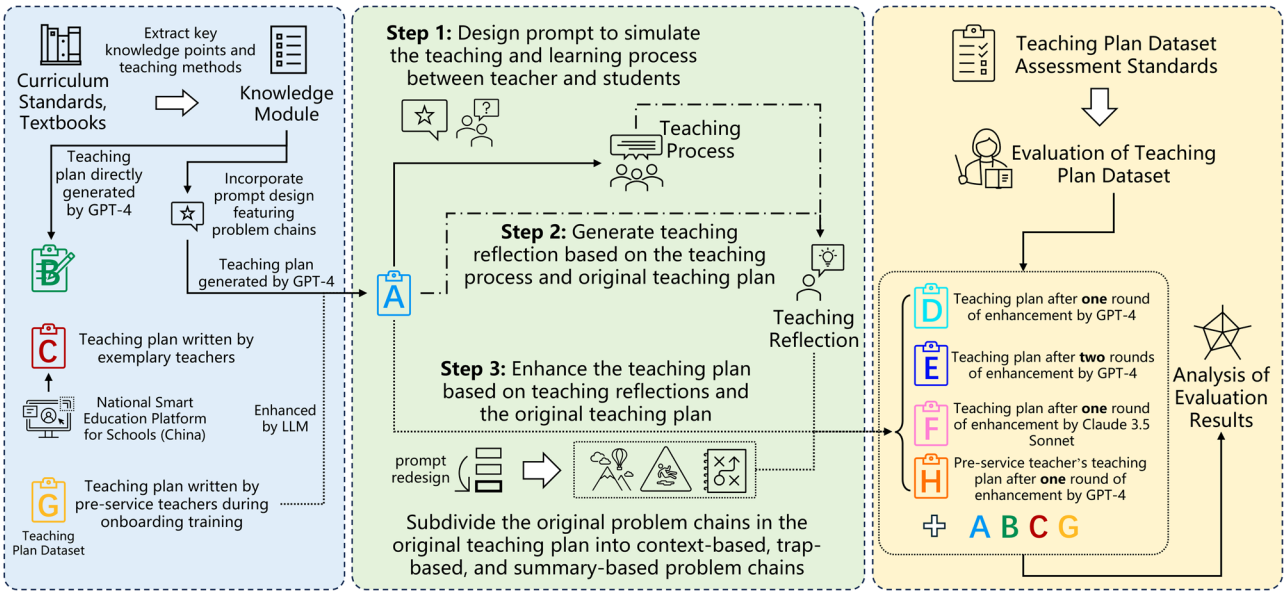


Fig. 1 | Construction and Evaluation of the Teaching Plan Dataset. This figure illustrates the process involving three main phases: dataset construction, quality enhancement, and effectiveness evaluation. During the dataset construction phase, four baseline teaching plan datasets (Datasets A, B, C, and G), comprising a total of 240 teaching plans, were created. In the quality enhancement phase, the process unfolded in three steps. First, the LLM simulated classroom interactions between teacher and students based on the content of the teaching plans. Second, the LLM generated teaching reflections by integrating the teaching plans with the simulated classroom interaction texts. Third, the LLM refined the teaching plans by incorporating the original plans and the generated teaching reflections. Four improved teaching plan datasets were obtained in this phase (Datasets D, E, F, and H), with 240 teaching plans in total. Finally, in the evaluation phase, manual assessment and comparative analysis were conducted on 8 teaching plan datasets.

Original Teaching Plan	**Knowledge Point : How to Conduct Stratified Random Sampling?** **[Problem Chain]:** Now that we understand the basic concept of stratified random sampling, how should we proceed with it? If we categorize supermarket customers by gender and age, how many people should be sampled from each category? **[Variant Question]:** Suppose we categorize male customers into age groups: 18-25, 26-35, 36-45, and over 46 years old, and female customers are categorized similarly. How many people should we sample from each category? How should we decide?
Simulated Teacher-Student Interaction Dialogue	**Student 1 (correct response):** I think we should sample according to the proportion of people in each category. If one category has more people, we should draw more samples; if another has fewer, we should draw fewer samples. **Teacher:** Very good! You're correct. Typically, we sample according to the proportion of each category, a common method in stratified random sampling. Let's try a specific example. Suppose we categorize male customers into age groups: 18-25, 26-35, 36-45, and over 46 years old, and female customers similarly. How many should we sample from each category? How do we decide? **Student 2 (struggling with previous knowledge):** Teacher, I'm not sure how to determine the specific number of samples. **Teacher:** No worries. We usually decide the number of samples based on the proportion of each category. For example, if male customers aged 18-25 make up 10% of all customers, we should sample 10% from this category. Everyone, try practicing this method. **Student 3 (posing their own question):** Teacher, if a category has very few people, do we still need to sample from it?
Teaching Reflection	**Deficiency:** Although classroom discussions and interactive Q&A sessions enhance student engagement, the teaching plan does not adequately address the needs of students with varying levels of cognitive abilities. For students who are slower in understanding concepts, the plan should provide layered questions or more detailed explanations to ensure that all students can keep pace with the lesson.
Enhanced Teaching Plan	**Knowledge Point : How to Conduct Stratified Random Sampling?** **[Context-Based Problem Chain]:** **[Problem Chain 1]:** Assuming we already understand the basic concept of stratified random sampling, how do we proceed with stratified random sampling? If we categorize supermarket customers by gender and age, how many people should we sample from each category? **[Problem Chain 2]:** Assuming we categorize male customers by age groups: 18-25 years old, 26-35 years old, 36-45 years old, and over 46 years old, and do the same for female customers. The supermarket has 1000 customers, with the following distribution by age and gender: Males 18-25 years old: 10%, Males 26-35 years old: 15%, Males 36-45 years old: 10%, Males over 46 years old: 15% Females 18-25 years old: 10%, Females 26-35 years old: 20%, Females 36-45 years old: 10%, Females over 46 years old: 10% If we need to sample 100 people, how many should we sample from each category? **[Trap-Based Problem Chain]:** **[Problem Chain 1]:** If we group male customers by employment type (full-time, part-time, student, retired) and then randomly sample from each group, do you think this grouping is reasonable? Why or why not? **[Problem Chain 2]:** If, in practice, the number of customers in certain categories is too small, how should we adjust our sampling strategy?

Fig. 2 | Example of Using GPT-4 to Improve Teaching Plan. This figure illustrates the process within the "Statistics" knowledge module, focusing on the lesson about "Stratified Random Sampling." It shows the teaching phase for "explaining new knowledge" on "how to conduct stratified random sampling." Utilizing GPT-4, the figure demonstrates how the original teaching plan content is used to simulate teacher-student interactive dialogs, generate teaching reflections, and subsequently, the content of the enhanced teaching plan.

quality ones created by human teachers. The teaching plans generated by GPT-4 still needs to be improved in the depth of teaching content, the complexity of problem chains, the breadth of knowledge coverage, the design of differentiated learning materials, interdisciplinary and subject culture integration, and practical value. These findings provide a data foundation for utilizing LLMs to enhance the quality of teaching plans in this study.

The research design of this study is illustrated in Fig. 1. During the teaching plan dataset construction phase, we used high school mathematics curriculum as an example. The content of 21 chapters in high school mathematics curriculum was categorized into four main knowledge modules: Statistics, Functions, Algebra, and Geometry[29,30]. From each module, 15 lessons were randomly selected, resulting in a total of 60 lessons. The teaching plans corresponding to these 60 lessons were used as evaluation subjects. We first constructed four baseline teaching plan datasets, comprising a total of 240 teaching plans, including: (1) Dataset A, derived from our previous research, in which finely designed prompt instructions based on PCK theory and mathematical problem chains were used to generate teaching plans; (2) Dataset B, consisting of teaching plans generated directly by GPT-4 using prompts that did not incorporate the structure of mathematical problem chains; (3) Dataset C, a high-quality teaching plan dataset sourced from China's National Primary and Secondary School Smart Education Platform, authored by experienced teachers with over ten years of teaching experience; and (4) Dataset G, consisting of teaching plans written by pre-service teachers during a two-week induction training program.

During the teaching plan quality enhancement phase, we designed prompt instructions and directed GPT-4 to simulate interactions between a teacher and students of varying ability levels during the teaching process, based on the content of each teaching plan from Dataset A, as illustrated in Fig. 2. Next, we instructed GPT-4 to generate teaching reflections by integrating the original teaching plans from Dataset A with the simulated teaching process texts. Subsequently, we commanded GPT-4 to improve the original teaching plans by incorporating the teaching reflections generated, resulting in the enhanced teaching plan dataset, Dataset D. To comprehensively evaluate the effectiveness of this enhancement approach, we conducted a second round of improvements using GPT-4 based on the initial enhancement, resulting in teaching plan dataset E. Additionally, we replaced the GPT-4 model with the Claude 3.5 Sonnet model to improve the teaching plans in dataset A, producing teaching plan dataset F. Furthermore, we utilized GPT-4 to enhance the teaching plans created by pre-service teachers, yielding teaching plan dataset H. In total, four enhanced teaching plan datasets were generated, comprising 240 teaching plans. Throughout the processes of simulating the teaching process, generating teaching reflections, and improving the original teaching plans, we included in the prompt instructions corresponding knowledge points for each lesson, along with instructional objectives and teaching priority and challenge from the original teaching plans. Additionally, to improve the quality of the mathematical problem chains designed by the LLM, we refined the original output format of problem chains when designing the prompt instructions. Instead of a standardized, generic problem chain format (e.g., "Problem Chain 1, Problem Chain 2, Problem Chain 3…"), we categorized the problem chain into three types: "context-based," "trap-based," and "summary-based" problem chains[31]. A "context-based" problem chain requires the LLM to create scenarios and knowledge-introduction questions related to the

lesson's knowledge points, ensuring logical progression and continuity across teaching phases. A "trap-based" problem chain refers to a series of questions designed by the LLM that are likely to cause students' mistakes, facilitate conceptual clarification, and present significant challenges. Finally, a "summary-based" problem chain involves questions that guide students to review and summarize the knowledge learned in the lesson, and stimulate their interest in further learning.

Finally, during the teaching plan evaluation phase, we conducted manual scoring and analysis of the eight teaching plan datasets. An evaluation framework was designed, comprising nine categories and nineteen dimensions, including problem chains, teaching activities, content knowledge, teaching methods and strategies, teaching evaluation, interdisciplinarity, practical value, scope beyond the syllabus, and overall rating. A Likert 8-point scale was used for the manual assessment. Detailed descriptions of each evaluation dimension (see Table 2), the basic information of the evaluators, the design of prompt instructions for simulating the teaching process and generating teaching reflections, as well as the detailed construction of the evaluation dataset, will be thoroughly discussed in the Methods section of this paper. As far as we know, this study is the first to explore and evaluate the potential of LLMs in enhancing the quality of teaching plans. This is expected to have a significant impact on how teachers engage in human-AI collaborative instructional design during teaching preparation and how pre-service teachers learn to design teaching plans in the future.

Results

Descriptive Statistical Analysis of the Enhanced Teaching Plans Across Evaluation Dimensions

We first conducted a descriptive statistical analysis of 16 evaluation dimensions, using a total sample size of $N = 480$ teaching plans. This included $N = 240$ teaching plans before enhancement and $N = 240$ teaching plans after enhancement. The evaluation results for each dimension are depicted in Fig. 3.

From the perspective of various dimensions, the enhanced teaching plans achieved an average score exceeding 7.0 in Dimension Q1 (designing a rich context for problems). In Dimensions Q2 (designing sequentially cohesive and challenging learning tasks) and Q3 (designing variant exercises to consolidate learned knowledge), the average scores were generally above 6.5. This indicates that the LLM effectively utilized the teaching process and teaching reflections to improve the quality of teaching plans, and further enhanced the effectiveness of problem chain design. However, in Dimension A1 (designing learning activities that promote teacher-student interaction), the enhanced teaching plans still fell short compared to the high-quality teaching plans written by human teachers. Meanwhile, the average scores for Dimensions A2 (designing teaching activities aligned with instructional objectives), C1 (accurately explaining and summarizing disciplinary knowledge and concepts), C2 (refining and summarizing lesson content), M1 (selecting appropriate teaching methods), and M2 (applying a variety of teaching strategies) fluctuated around 6.8. This demonstrates that the enhanced teaching plans showed significant improvements in areas such as designing teaching activities aligned with objectives, explaining and summarizing subject knowledge, and employing suitable teaching methods, reaching or even surpassing the level of high-quality teaching plans authored by human teachers. Notably, in Dimension C3 (introduction of disciplinary history and culture), the enhanced teaching plans outperformed those created by human teachers, with the Claude model demonstrating particularly strong results. However, the average scores remained below 5, indicating room for further exploration and improvement. Similarly, in Dimension D1 (design of interdisciplinary content), the enhanced teaching plans also surpassed those of human teachers, with additional improvements observed after two rounds of refinement. This suggests that LLMs, through reflective analysis of the teaching process, can effectively make use of their multidisciplinary knowledge to design learning content that is more closely aligned with disciplinary culture and interdisciplinary themes. In Dimensions E1 and E2, the enhanced teaching plans showed significant improvements in identifying differences among students and employing diverse methods for assessment. In Dimension R1 (designing content aligned with the scope of the lesson's knowledge), the enhanced teaching plans reached the level of high-quality teaching plans created by human teachers. Moreover, the average score for the enhanced teaching plans generated from pre-service teachers' teaching plans exceeded 7. This indicates that the method proposed in this study effectively addresses the issue observed in previous research, in which teaching plans generated by LLMs often included content beyond the scope of the lesson.

In terms of overall scores, among the four baseline teaching plan datasets, the average scores of teaching plans generated directly by GPT-4

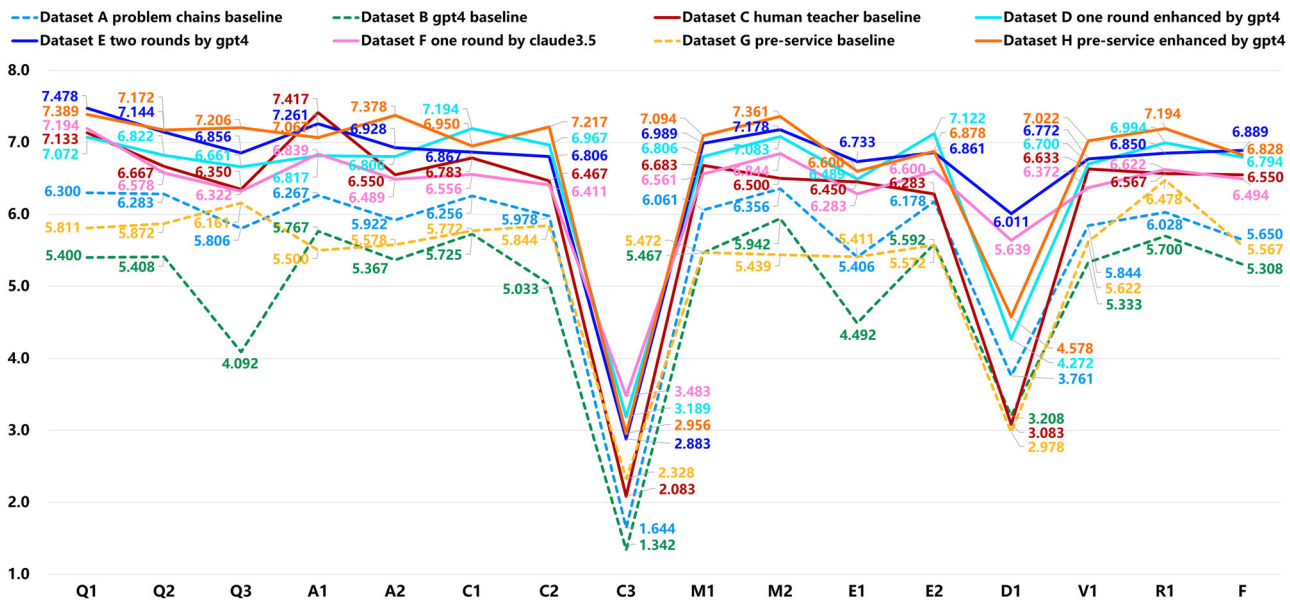


Fig. 3 | Average Values of Assessment Dimensions. This figure presents the average scores in 16 assessment dimensions for the eight teaching plan datasets, each comprising $N = 60$ samples. The evaluation employed an 8-point Likert scale, where 8 indicates "strongly agree" and 1 indicates "strongly disagree." The assessment dimensions encompass nine categories with a total of 16 evaluation criteria: problem chains, teaching activities, knowledge content, teaching methods and strategies, teaching evaluation, interdisciplinarity, practical value, scope, and overall score.

(Dataset B) in most dimensions were lower than those of teaching plans written by pre-service teachers during induction training (Dataset G). However, the teaching plans generated by GPT-4 with integrated mathematical problem chains (Dataset A) slightly outperformed those written by pre-service teachers. After enhancement, the four improved teaching plan datasets (Datasets D, E, F, and H) had higher average scores across most dimensions compared to the baseline teaching plans (Datasets A, B, and G). Among these, the teaching plans improved by the Claude model in one round (Dataset F) had average scores around 6.5, slightly below those of high-quality teaching plans written by exemplary human teachers (Dataset C). In contrast, the teaching plans enhanced by GPT-4 in one round (Dataset D) scored slightly higher than the high-quality teaching plans written by human teachers. Notably, after two rounds of improvement with GPT-4 (Dataset E), the average scores across most dimensions surpassed those of the high-quality human-written teaching plans, with 12 dimensions achieving average scores exceeding 6.8. Importantly, the teaching plans derived from pre-service teachers' original plans and enhanced through one round of improvement by the GPT-4 model (Dataset H) achieved average scores across dimensions hovering around 7, exceeding those of high-quality teaching plans written by experienced human teachers and closely approximating the scores of teaching plans improved through two rounds of GPT-4 enhancement.

These results indicate that the directly generated teaching plans by GPT-4 in previous studies was less effective, aligning with findings from other research that the content generated by LLMs often tends to be broad and lacks depth[32]. However, the method proposed in this study effectively addresses these issues. By utilizing LLMs to simulate teaching processes and generate teaching reflections, and then integrating these insights to refine teaching plans, this approach has demonstrated considerable success in improving the quality of teaching plans. Furthermore, due to their lack of teaching experience, pre-service teachers tend to produce teaching plans that are relatively generic and lack specificity. Under the guidance of prompts incorporating mathematical problem chains, the quality of teaching plans generated by GPT-4 showed improvement, reaching a level comparable to those written by pre-service teachers. However, they still fell short of the high-quality teaching plans created by exemplary teachers. Building on this foundation, the application of the method proposed in this study resulted in noticeable improvements in the quality of teaching plans, bringing them to a level comparable to high-quality teaching plans. Additionally, after two rounds of enhancement, the quality of the teaching plans exhibited a modest yet further improvement.

Interestingly, the teaching plans written by pre-service teachers showed significant improvement after being enhanced by the LLM, with their scores in Dimension V1 (practical value) even approaching the teaching plans improved through two rounds of enhancement. Analyzing the reasons behind this, we posit that teaching plans created by pre-service teachers are inherently designed to meet the practical needs of real classroom teaching preparation. As a result, the LLM's first round enhancement on these plans aligns more closely with actual teaching preparation requirements. In contrast, the Dataset A teaching plans used in the two-round enhancement process were originally generated by the LLM itself, and therefore still differ from teaching plans designed grounded in real-world teaching preparation needs. After two rounds of content enhancement, although significant improvements were observed across most dimensions, the improved teaching plans were built upon and extended from a "fabricated" teaching plan. In contrast, the results of LLM enhanced teaching plans based on real-world examples demonstrated greater practical value, with scores in Dimension V1 (practical value) exceeding 7. This indicates that the method proposed in this study holds considerable applicability and practicality, providing valuable instructional insights and references to support teachers in their instructional design.

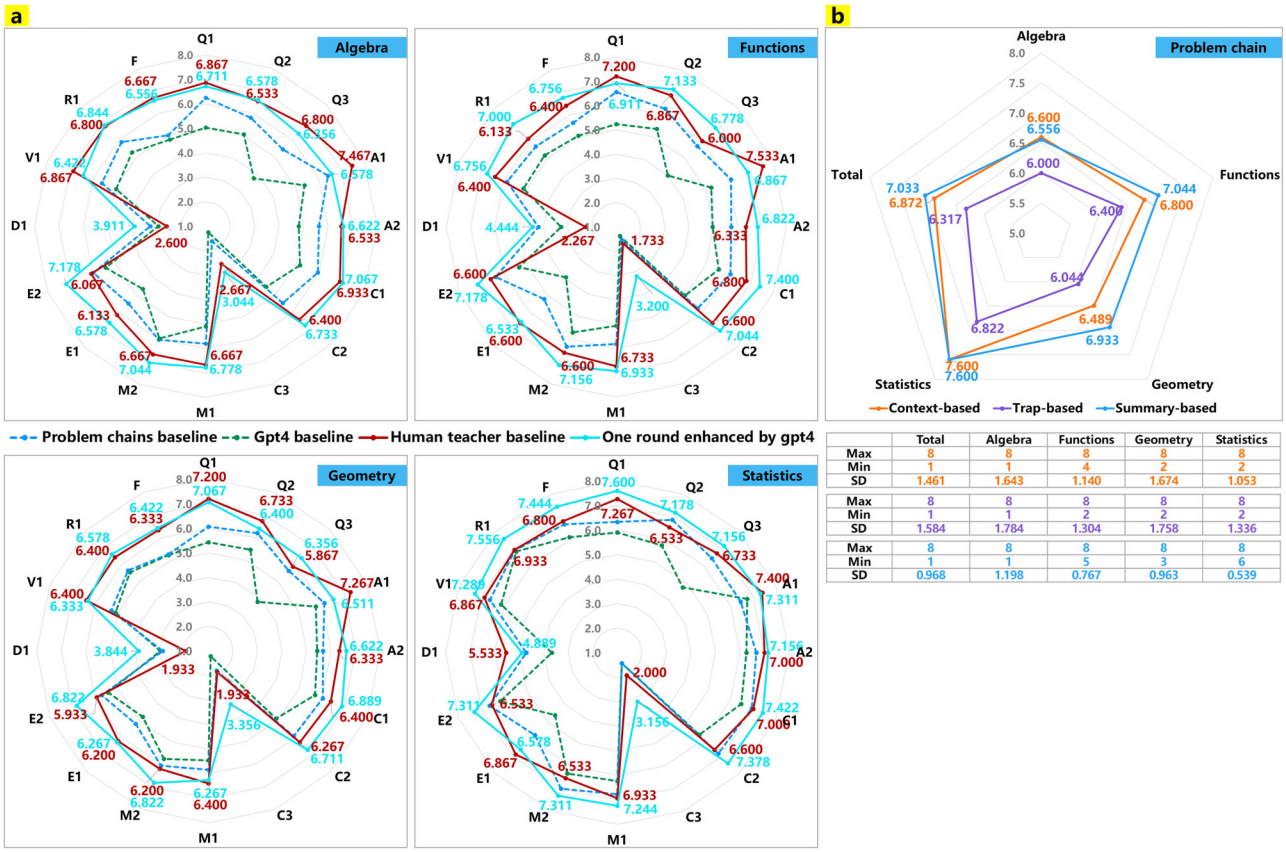
Cross-Analysis of the Enhanced Teaching Plans by Dimensions and Knowledge Modules

Additionally, taking the enhanced teaching plan dataset D (improved through one round of GPT-4 refinement) as an example, we conducted a more in-depth exploration of its performance across various evaluation dimensions within different knowledge modules. Comparisons were made against three baseline teaching plan datasets (Datasets A, B, and C). Each knowledge module—Algebra, Functions, Geometry, and Statistics—had a sample size of $N = 60$. The evaluation results for each dimension across the different knowledge modules are presented in Fig. 4a.

Across the knowledge modules, in the "Algebra" module, the teaching plans enhanced through one round of GPT-4 refinement outperformed high-quality teaching plans written by human teachers in 11 evaluation dimensions. They achieved scores exceeding 7 points in areas such as accurately explaining disciplinary concepts, employing diverse teaching methods and strategies, and utilizing varied approaches to evaluate students' learning. Most scores within this module ranged between 6.5 and 6.8. In the "Functions" module, the improved teaching plans surpassed high-quality human-crafted teaching plans in 13 evaluation dimensions. They demonstrated outstanding performance in designing challenging learning tasks, accurately explaining disciplinary theories, summarizing subject knowledge, employing diverse teaching methods and strategies, evaluating learning situations diversely, and designing content within the curriculum's scope, with average scores exceeding 7 points in these areas. Most scores in this module ranged between 6.7 and 7.0. The overall performance of the "Geometry" module was slightly lower than that of the other three modules, with scores exceeding 7 points only in the dimension of designing effective and scientifically sound problem contexts. While the improved teaching plans in this module matched the level of high-quality teaching plans in 11 evaluation dimensions, most scores hovered around 6.4. In contrast, the "Statistics" module demonstrated the best performance among all knowledge modules, with 13 evaluation dimensions scoring above 7 points. Scores in dimensions such as designing effective and scientifically sound problem contexts, creating teaching activities that foster teacher-student interaction, accurately explaining disciplinary theoretical concepts, summarizing subject knowledge, employing diverse teaching methods and strategies, evaluating learning situations diversely, and designing content within the curriculum's scope exceeded 7.3 points, reaching the level of high-quality teaching plans. These results indicate that the method proposed in this study effectively improved the performance of enhanced teaching plans across all knowledge modules.

Additionally, we evaluated the practicality and applicability of the refined context-based, trap-based, and summary-based problem chains, as illustrated in Fig. 4b. Overall, the evaluation results showed that these three types of problem chains performed best in the "Statistics" module. The context-based and summary-based problem chains achieved average scores above 6.5 across all knowledge modules. The trap-based problem chains scored above 6.5 in the "Statistics" module, while their scores in other modules exceeded 6.0, indicating room for further improvement.

By comparing the performance scores of the improved teaching plans across various knowledge modules, we identified disparities among the knowledge modules across different evaluation dimensions. Given the heteroscedasticity of the sample datasets, we conducted the Kruskal-Wallis H test, a non-parametric rank-sum test, to further investigate whether these differences were statistically significant. The pairwise comparison results for the knowledge modules across each evaluation dimension are presented in Table 1. We have listed comparisons where the adjusted significance levels, after Bonferroni correction, were less than 0.05. From the perspective of various evaluation dimensions, the "Statistics" module significantly outperformed the other modules, while the "Functions" module slightly exceeded the "Algebra" and "Geometry" modules. The "Statistics" module excelled in designing effective and scientifically sound problem contexts, challenging learning tasks, variant exercises to consolidate knowledge, teaching activities that promote teacher-student interaction discussions, and maintaining content within the curriculum scope. Additionally, it demonstrated superior performance in summarizing subject knowledge and selecting appropriate teaching methods and strategies. The context-based and summary-based problem chains within the "Statistics" module also showed considerable practical value compared to those in other modules. Furthermore, although there was improvement in the design of



	Total	Algebra	Functions	Geometry	Statistics
Max	8	8	8	8	8
Min	1	1	4	2	2
SD	1.461	1.643	1.140	1.674	1.053
Max	8	8	8	8	8
Min	1	1	2	2	2
SD	1.584	1.784	1.304	1.758	1.336
Max	8	8	8	8	8
Min	1	1	5	3	6
SD	0.968	1.198	0.767	0.963	0.539

Fig. 4 | Average Values of Each Assessment Dimension under Different Knowledge Modules. In **a**, the average scores of the four knowledge modules—Algebra, Functions, Geometry, and Statistics—are presented for each of the four datasets (Datasets A, B, C and D) across all assessment dimensions. In **b**, the detailed analysis of refined problem chains (context-based, trap-based, and summary-based) shows their average scores, maximum, minimum, and standard deviations across these modules and overall.

disciplinary culture and interdisciplinary content across all knowledge modules compared to previous studies, there remains considerable room for further enhancement.

Overall, this study demonstrates that the improvement of teaching plans using LLMs can lead to enhancements across all knowledge modules, with the evaluation performance closely resembling that of high-quality teaching plans created by exemplary human teachers. This indicates that guiding LLMs to enhance teaching plan quality through a process of simulation, reflection, and refinement can yield favorable results. Regarding the performance differences observed among the various knowledge modules, we attribute these to the nature of the content in the "Geometry" module of high school mathematics. Certain concepts in this module require the use of geometric diagrams to aid understanding. Relying solely on textual descriptions generated by the LLMs may not effectively achieve the ideal "integration of algebra and geometry" (numerical and graphical relationships), which resulted in its evaluation performance being lower than that of other knowledge modules. For the "Algebra" and "Functions" knowledge modules, the LLMs still have room for improvement, particularly when tasked with designing and solving medium-to-high-difficulty algebraic and functional problems. This limitation stems from the model's inherent capacity to handle complex mathematical tasks. In contrast, the "Statistics" module focuses more on providing students with rich statistical problem contexts to enhance their understanding and application of statistical knowledge[33]. This aligns well with the strengths of LLMs, which excel in supporting multidisciplinary knowledge and vast information, leading to better performance in the evaluation for this module.

Discussion

The results of this study make significant contributions to the application and use of LLMs in teaching, particularly in supporting teachers with pre-class preparation. The rationale for adopting a simulation and reflection approach in this research is twofold. Firstly, in our preliminary studies, we aimed to use LLMs to grade and provide feedback for teaching plans. However, we found that the scores provided by LLMs were inconsistent and lacked a scientific basis. Regardless of the quality of the teaching plans, LLMs tended to provide positive ratings. One possible explanation is that, as language models, LLMs may not be sufficiently sensitive to numerical values. Therefore, the focus of the study should not be on the numerical scores, but rather on exploring how LLMs' capabilities in language understanding, role imitation, and self-reflection can be used for educational practice. Secondly, novice teachers, early in their teaching careers, have not established solidified teaching content, teaching strategies, or teaching styles. Much of their instructional material can only be refined and enhanced through reflection and revision after several trial lectures[34]. Additionally, due to their lack of teaching experience and the reluctance of some students to reveal their own ignorance, these teachers often struggle to anticipate the learning difficulties students might encounter, making it challenging to adequately prepare in advance[35]. As a result, students in prior classes may not benefit from the subsequent improvements and refinements made to the teaching content. This study aims to leverage LLMs to enhance teaching plans prior to actual teaching, providing novice teachers with higher-quality teaching content and instructional ideas, thereby reducing potential educational risks arising from inadequate teaching preparation.

As for research questions 1 and 2, the analysis reveals that the method proposed in this study, which uses LLMs for simulation and reflection to enhance teaching plans, effectively improves the quality of teaching plans and increases the practical value of LLM-generated teaching plans. The improved teaching plans perform at or above the level of high-quality teaching plans authored by experienced human teachers across various

Table 1 | Pairwise Comparisons of Knowledge Modules Across Assessment Dimensions

Dimension	Knowledge Modules Avg(SD)	Pairwise Comparison	H statistic	Adj. p
Q1	A: 6.711(1.014) F: 6.911(0.821) G: 7.067(1.116) S: 7.600(0.580)	A < S*** F < S**	-47.300 -40.500	0.000 0.001
Q2	A: 6.578(1.270) F: 7.133(0.919) G: 6.400(1.405) S: 7.178(0.886)	G < S* G < F*	-29.733 -27.444	0.024 0.047
Q3	A: 6.356(1.448) F: 6.778(1.064) G: 6.356(1.554) S: 7.156(1.065)	A < S* G < S*	-32.356 -28.567	0.012 0.039
A1	A: 6.578(1.158) F: 6.867(0.661) G: 6.511(1.359) S: 7.311(1.062)	A < S*** G < S*** F < S**	-42.678 -39.756 -34.011	0.000 0.000 0.004
C2	A: 6.733(1.572) F: 7.044(1.043) G: 6.711(1.660) S: 7.378(1.173)	A < S*	-29.800	0.022
M1	A: 6.778(1.166) F: 6.933(0.963) G: 6.267(1.629) S: 7.244(0.981)	G < S**	-33.389	0.009
V1	A: 6.422(1.305) F: 6.756(0.933) G: 6.333(1.398) S: 7.289(1.014)	G < S*** A < S*** F < S*	-42.844 -41.711 -31.622	0.000 0.000 0.015
Context-based	A: 6.600(1.643) F: 6.800(1.140) G: 6.489(1.674) S: 7.600(1.053)	G < S*** A < S*** F < S***	-44.767 -41.900 -41.867	0.000 0.000 0.000
Summary-based	A: 6.556(1.198) F: 7.044(0.767) G: 6.933(0.963) S: 7.600(0.539)	A < S*** G < S** F < S**	-55.922 -39.5178 -35.300	0.000 0.001 0.003
R1	A: 6.844(1.413) F: 7.000(0.977) G: 6.578(1.699) S: 7.556(1.139)	G < S** F < S** A < S**	-39.822 -37.767 -36.944	0.001 0.001 0.002
F	A: 6.556(1.358) F: 6.756(0.957) G: 6.422(1.469) S:7.444(1.035)	G < S*** A < S*** F < S**	-44.911 -43.122 -39.789	0.000 0.000 0.001

Note. This table presents post-hoc pairwise comparisons of different knowledge modules across various assessment dimensions. It tests the null hypothesis that "the distribution of the two knowledge modules is the same." Progressive significance is displayed using a two-sided test, with the significance level set at 0.05. The Bonferroni correction has been applied to adjust the significance values for multiple comparisons. In the table, the modules are denoted as A for Algebra, F for Functions, G for Geometry, and S for Statistics. Significance levels are indicated with asterisks: * for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$. Only results showing significant differences after adjustment are listed.

evaluation dimensions and knowledge modules, confirming the practicality and potential of LLMs in supporting teaching preparation. A deeper exploration of the reasons for the improvement in teaching plan quality points to two main factors: the simulation of the classroom teaching process and the classification of mathematical problem chains.

Simulation has long played a pivotal role in the field of education[19]. It emerged from learning systems that feature teachable student agents. In these systems, human students learn by teaching simulated student agents with lower ability levels. However, these student agents are only capable of responding to questions but cannot generate their own, which lacks real experience of classroom teaching[36]. The advancement of LLM technology has significantly enhanced the effectiveness of simulations. Agents based on LLMs not only can assume various professional roles[37] and simulate human behaviors[38], but also hold the potential to predict future societal developments[39]. In the field of education, simulations based on LLMs primarily focus on student roles, learning styles, and personality traits, with fewer simulations addressing the varying cognitive levels of students[40]. For student simulation, some researchers have created roles such as learning partners, competitors, and troublemakers to stimulate the social and emotional development of real students[41]. Other researchers have used data on students' learning behaviors, knowledge levels, and memory capacities to simulate real students[42], replacing traditional knowledge tracing methods to predict learning performance[43]. However, these simulations still lack a comprehensive evaluation of their effectiveness[19,44]. In terms of simulating teachers, some researchers have developed teacher agents from the perspective of instructional decision-making, enabling them to design the next study plans based on course materials and student backgrounds[45].

In this study, we focused on students' common learning difficulties at various cognitive levels in the classroom. Rather than simulating a single type of student, we simulated scenarios where students encounter obstacles in their learning, such as being unable to answer due to gaps in prior knowledge, or providing incorrect answers due to misconceptions. This simulation of students' incorrect responses in the classroom helps the LLM reflect on the original teaching plan, thereby facilitating its revision and improvement. However, LLMs tend to directly provide correct answers[18]. Therefore, it is necessary to adjust the design of the prompt instructions by utilizing chains of thought or reasoning[36,46], guiding the LLM to generate incorrect responses that align with the cognitive level of high school students[39]. In addition to designing prompts, some researchers have used fine-tuning of LLMs to simulate the language style of specific roles[47]. However, this method presents a high technical barrier for ordinary teachers, making it difficult be widely used in primary and secondary education. Reducing the difficulty, time, and effort for teachers to use LLM-based tools[48] is crucial for advancing the widespread application of LLMs in teaching. Therefore, in this study, we adopted a prompt-based design approach, which allows ordinary teachers to easily implement the method. By using the role simulation and reflection-improvement capabilities of LLMs, we depicted the process of teacher-student interaction in the classroom.

Classroom questions pre-designed by teachers can be categorized into types such as introduction, progression, diagnosis, inquiry, and conclusion[16]. A combination of these types can facilitate students' understanding of subject knowledge[49,50]. However, due to their lack of teaching experience, novice teachers often struggle to design high-quality classroom questions[51,52]. To address this challenge, we attempted to utilize LLMs to provide novice teachers with ideas for designing questions. In previous research, we focused on the design of problem contexts by LLMs, which yielded promising results. However, when it comes to designing more

advanced questions, such as progressively challenging problems and variation exercises, LLMs struggled to design coherent questions from easy to difficult. The reason was that LLMs did not fully comprehend the layered progression of problem chains, such as "Problem Chain 1, Problem Chain 2, Problem Chain 3," and thus failed to generate questions that were appropriately challenging. Instead, it produced over simplistic content with limited practical utility.

To address this problem, when designing prompts for this study, we refined the generic problem chains from the original output format into three types: context-based problem chains, trap-based problem chains, and summary-based problem chains. The context-based problem chains correspond to problem contexts and relatively easy questions, which are used to introduce and explain new knowledge. Trap-based problem chains correspond to medium- to high-difficulty questions, aimed at provoking students' critical thinking and analysis of the key concepts of the lesson. The design of trap-based problems, as a means of facilitating productive failure[53], helps deepen students' understanding and application of the core knowledge. For the summary-based problem chain, we aim to design questions that facilitate students' review of the knowledge learned in class, which can help teachers integrate them into their teaching plan[54]. The results indicate that this subdivision approach enhances the LLM's understanding of the teaching process, progressing from basic to advanced levels, and leads to better outcomes when designing and improving teaching plans. However, the analysis also reveals that the usability scores for trap-based problem chains are slightly lower than those for other types of questions, suggesting that LLMs still need improvement in designing deep and challenging questions centered around core knowledge. In future research, we will explore how to further enhance the effectiveness of LLMs in designing trap-based problems.

In conclusion, as the capabilities of LLMs continue to develop, future approaches of instructional design and teaching preparation may surpass traditional methods[55]. Specifically, the approach employed in this study involves using a single LLM to simultaneously simulate dialogs between different roles via a web interface, thereby simulating a complete classroom teaching process. Future research can further explore the use of multi-agent frameworks, such as AutoGen[56], to integrate multiple LLMs to create distinct agents. These agents could represent various tutor or expert roles—such as engineering experts, mathematics experts, and literature experts—along with different student roles in group cooperation, such as recorders, problem creators, and data analysts[57]. This approach would enable the simulation of more diverse teaching scenarios across various disciplines[44], thereby enhancing and refining teaching plans to better suit specific teaching environments and learning needs[3]. The method we propose for improving teaching plans through the simulation and reflection of LLMs is innovative in that it semi-automates the teacher's daily preparation, teaching, and reflection processes through inquiring on the web interface. This approach lowers the technical barrier for teachers, eliminating the need to write code to access LLMs. Additionally, the simulation and reflection process of the LLMs, as a form of "thinking aloud," reveals the model's reasoning and decision-making processes, which are fully visible on the web interface. This enhances the interpretability and transparency of the process of improving teaching plans[58]. As a result, the design, optimization, and iteration of teaching plans become more efficient and can be accomplished prior to actual classroom instruction. Future research could incorporate the judgmental role of human teachers to evaluate the rationale behind the simulated teaching process and the generated teaching reflection content. This would facilitate a "human-in-the-loop" approach, where LLMs assist teachers in their teaching preparation. Lastly, one possible direction for subsequent research is to explore whether, with the overall improvement of teaching plan quality, future intelligent tutoring systems could use these high-quality plans to enable true "machine teaching" rather than having students passively watch teaching videos, and complete adaptive exercises as part of a monotonous learning process.

Methods

In this study, we propose a method for enhancing the quality of teaching plans by utilizing LLMs through simulation and reflection from the perspective of teachers' teaching preparation. We guide the LLM to design and generate teaching plans (as explored in our previous research), simulate formal lessons (achieved in this study through the simulation of the teaching process), and generate post-lesson reflections and enhanced teaching plans. Finally, we analyze and evaluate the effectiveness of the enhanced teaching plans through teacher-led manual assessments. The evaluation dimensions, evaluators, teaching plan framework, prompt design, baseline data, and dataset used in this study are outlined in detail below.

Evaluation Dimensions and Evaluators

In the design of the evaluation dimensions, we incorporated research findings and assessment indicators related to PCK theory, teachers' instructional design abilities, and the quality evaluation of mathematical problem design[59–61]. Based on the evaluation dimensions developed in previous research[28,62–64], we eliminated the assessment of instructional objectives and teaching priorities and challenges, while introducing new evaluation dimensions focused on the usability of the context-based, trap-based, and summary-based problem chains generated by LLMs, as shown in Table 2. Ultimately, the evaluation framework in this study includes 9 categories and 19 assessment indicators. The 9 categories include problem chain, teaching activities, knowledge content, teaching methods and strategies, teaching evaluation, interdisciplinarity, practical value, scope, and overall score.

To better analyze the effects of teaching plan enhancement, this study invited three mathematics teachers with more than five years of teaching experience and teaching credentials to participate in the evaluation. Additionally, two experienced teachers with over ten years of teaching experience were invited to balance discrepancies in the scores. First, we provided each mathematics teacher with a detailed explanation of the meaning and examples of each assessment dimension, and asked them to review each teaching plan for at least ten minutes. Subsequently, we randomly selected 15 teaching plans from both Dataset C and Dataset D, and mixed them together for the teachers to conduct pre-evaluation training. The formal evaluation was conducted after the three teachers were familiar with the assessment dimensions. The consistency test for the pre-evaluation had an ICC (2, 3) value of 0.716, which exceeds the 0.6 threshold, indicating a high degree of consistency in the teachers' evaluations after becoming familiar with the assessment dimensions. Before the formal evaluation, we informed the three teachers of the sources of each evaluation dataset, while emphasizing the importance of maintaining objective and impartial evaluation standards. For dimensions with significant discrepancies in scores, consultations with the two expert teachers were held to achieve as much consistency as possible. Furthermore, to facilitate comparison with prior studies, the Likert 8-point scale was used for scoring, in which 8 represents "strongly agree" and 1 represents "strongly disagree," with no intermediate values, to better distinguish between the quality of the teaching plans.

Framework of Teaching Plans and Design of Prompts

To better compare with the teaching plans generated using the LLM in our previous study, we maintained the content structure of the enhanced teaching plans, which includes: (1) the instructional analysis section (instructional objectives, teaching priorities and challenges), and (2) the teaching process section (pre-class introduction, explanation of new knowledge, consolidation and improvement, comprehensive exercise, and lesson summary, with a total of five teaching phases).

When designing prompt instructions for simulating the teaching process, we included few-shot examples of teacher-student interactive dialogs as well as the teaching plan content for the lesson in the prompts. In student simulation, the LLM was instructed to realistically depict various characteristics of students in classroom[65], particularly negative behaviors[39]. For instance, it was tasked with imitating students with weaker learning abilities who are unable to answer questions. This approach aims to guide the

Table 2 | Dimensions of Teaching Plan Evaluation

Category	Label	Dimension
Problem Chain	Q1	Capable of designing and introducing rich, effective, and scientifically contextualized problem scenarios
	Q2	Able to design coherent and progressively challenging learning problems and tasks
	Q3	Capable of designing appropriate variation exercises to consolidate learned knowledge
Teaching Activities	A1	Capable of designing content that fosters teacher-student and peer interactions
	A2	Able to design instructional activities that align with instructional objectives and content
Knowledge Content	C1	Able to accurately and appropriately explain the fundamental theories and concepts of the subject
	C2	Capable of distilling and summarizing the subject matter taught in the lesson
	C3	Able to present the historical development of important theories and the mathematical culture of the discipline
Teaching Methods and Strategies	M1	Able to choose suitable teaching methods and strategies
	M2	Capable of employing a variety of teaching methods and strategies
Teaching Evaluation	E1	Able to recognize individual and learning differences among students
	E2	Capable of employing various methods to assess students' learning
Interdisciplinary	D1	Able to establish relevant connections with other disciplines and design activities that cultivate students' interdisciplinary skills
Practical Value	V1	This teaching plan offers value for revision and provides a reference for instructional design
	Context-based	The designed context-based problem chains offer value for revision and provide a reference for instructional design
	Trap-based	The designed trap-based problem chains offer value for revision and provide a reference for instructional design
	Summary-based	The designed summary-based problem chains offer value for revision and provide a reference for instructional design
Scope	R1	Able to design content within the scope of the lesson
Overall Score	F	Overall evaluation of the teaching plan

Note. This table outlines the evaluation dimensions for teaching plans using an 8-point Likert scale, where 8 indicates "strongly agree" and 1 signifies "strongly disagree." The evaluation encompasses 480 teaching plans across eight datasets.

LLM to reflect on issues students encountered during the teaching process, enabling more targeted improvements to the original teaching plan. In this study, we focused specifically on cognitive differences among students. From the perspective of students' classroom learning behaviors, we identified five common performance types, including: (1) answering incorrectly due to conceptual confusion; (2) being unable to answer due to gaps in prior knowledge; (3) answering incorrectly due to misconceptions; (4) asking their own questions; and (5) answering correctly. These performance types were integrated into the prompts, instructing the LLM to select and simulate appropriate types of student behavior based on the knowledge being taught.

When designing prompts for generating teaching reflections, we included few-shot examples of teaching reflections, along with the teaching plan and corresponding text of teaching process for the lesson. Building on findings from prior research, we focused the reflection process on addressing key limitations commonly observed in teaching plans generated by LLMs, such as exceeding the knowledge scope of the curriculum, difficulties in designing differentiated teaching activities, and the lack of alignment in problem chain design with core knowledge points. This led to the identification of nine categories of reflective dimensions, which include: (1) whether students' prior knowledge was considered; (2) whether the design of the problem chain was overly context-based; (3) whether the problem chain design accounted for students of varying levels; (4) whether the problem chain reflected the definition and analysis of core knowledge concepts; (5) whether the teaching content exceeded the scope of the current lesson; (6) whether the content incorporated relevant mathematical history and culture; (7) whether the summary of core knowledge concepts was appropriate; (8) whether the transitions between knowledge points were effectively connected; and (9) whether there were problems in the choice of teaching methods and the design of teaching activities. These reflective dimensions were incorporated into the prompts, instructing the LLM to systematically perform teaching reflections for each dimension by synthesizing the teaching plan and the simulated teaching process text.

When designing prompts for improving teaching plans, we incorporated the original teaching plan and the corresponding teaching reflection text for the lesson into the prompts. The output format of the enhanced teaching plan remained consistent with that used in previous studies. Prior research demonstrated that LLMs performed well in establishing instructional objectives and identifying teaching priorities and challenges. Therefore, this study retained the instructional objectives and teaching priorities and challenges generated in earlier studies, without prompting the LLM to regenerate these elements. In the prompts, the LLM was instructed to focus solely on improving the teaching process section of the teaching plan. Additionally, the output format for problem chains in the prompts was refined. The previous generic format (e.g., "Problem Chain 1, Problem Chain 2, Problem Chain 3…") was categorized into three distinct types: context-based, trap-based, and summary-based problem chains. Each type was supplemented with corresponding examples to guide the LLM.

Assessment Baselines and Datasets

The evaluation dataset in this study comprises a total of 480 teaching plans, including four baseline teaching plan datasets and four enhanced teaching plan datasets. These were derived from the high school mathematics curriculum, which consists of 21 chapters divided into four primary knowledge modules. Fifteen lessons were randomly selected from each module, resulting in 60 lessons used for evaluation. Specifically, the baseline datasets include: (1) Dataset A, teaching plans generated by GPT-4 using prompts meticulously designed based on PCK theory and the mathematical problem chain format; (2) Dataset B, teaching plans generated directly by GPT-4 without incorporating mathematical problem chains; (3) Dataset C, high-quality teaching plans authored by experienced high school mathematics teachers (with over ten years of teaching experience) and sourced from China's National Primary and Secondary School Smart Education Platform; and (4) Dataset G, teaching plans written by five pre-service teachers during a two-week induction training program following one year of internship

experience. The enhanced teaching plan datasets include: (1) Dataset D, teaching plans improved in one round by GPT-4 based on Dataset A; (2) Dataset E, teaching plans further refined through a second round of improvement by GPT-4 based on Dataset D; (3) Dataset F, teaching plans improved in one round by Claude 3.5 Sonnet based on Dataset A; and (4) Dataset H, teaching plans improved in one round by GPT-4 based on Dataset G.

For the teaching plan datasets generated by LLMs, this study adopted a step-by-step approach where the five teaching phases were generated sequentially and then merged into a complete teaching plan. This method aimed to ensure that the teaching plans generated were as detailed and comprehensive as possible. Similarly, for the enhanced teaching plan datasets, the same step-by-step generation and merging approach was applied during all three stages: simulating the teaching process, generating teaching reflections, and improving the teaching plans. During the simulation of the teaching process, the LLM was required to mimic at least two types of student responses for each teaching phase other than giving correct answers. In generating teaching reflections, each teaching phase needs to produce 4–5 reflective statements. If the content or format generated by the LLM at any teaching phase did not meet the specified requirements, we would re-prompt the model and regenerate the content. This study aimed to employ a method accessible to teachers for improving teaching plan quality using LLMs. Thus, all interactions with the LLM, including queries and content generation, were conducted via a web-based interface (both GPT-4 and Claude 3.5 models).

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Code availability

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

Received: 28 June 2024; Accepted: 23 January 2025;
Published online: 06 February 2025

References

1. Brown, T. B. et al. in *Proceedings of the 34th International Conference on Neural Information Processing Systems* Article 159 (Curran Associates Inc., Vancouver, BC, Canada, 2020).
2. Bahrami, M. R., Bahrami, B., Behboodi, F. & Pourrafie, S. in *Agents and Multi-agent Systems: Technologies and Applications 2023*. (eds Gordan Jezic et al.) 393–402 (Springer Nature Singapore).
3. Kim, J., Lee, H. & Cho, Y. H. Learning design to support student-AI collaboration: perspectives of leading teachers for AI in education. *Educ. Inf. Technol.* **27**, 6069–6104 (2022).
4. Yildirim-Erbasli, S. N. & Bulut, O. Conversation-based assessment: a novel approach to boosting test-taking effort in digital formative assessment. *Comput. Educ. Artif. Intell.* **4**, 100135, https://doi.org/10.1016/j.caeai.2023.100135 (2023).
5. Ericsson, E., Sofkova Hashemi, S. & Lundin, J. Fun and frustrating: Students' perspectives on practising speaking English with virtual humans. *Cogent. Educ.* **10**, 2170088 (2023).
6. Vallis, C., Wilson, S., Gozman, D. & Buchanan, J. Student perceptions of AI-generated avatars in teaching business ethics: we might not be impressed. *Postdigit. Sci. Educ.* **6**, 537–555 (2024).
7. Pereira, D. S. M. et al. Here's to the future: conversational agents in higher education-a scoping review. *Int. J. Educ. Res.* **122**, 102233 (2023).
8. Ouyang, F. & Jiao, P. Artificial intelligence in education: the three paradigms. *Comput. Educ. Artif. Intell.* **2**, 100020, https://doi.org/10.1016/j.caeai.2021.100020 (2021).
9. Gandolfi, A. GPT-4 in Education: Evaluating Aptness, Reliability, and Loss of Coherence in Solving Calculus Problems and Grading Submissions. *Int. J. Artif. Intell. Educ*., 1–31; https://doi.org/10.1007/s40593-024-00403-3 (2024).
10. Wang, R. & Demszky, D. in *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* 626-667 (Toronto, Canada. Association for Computational Linguistics, 2023).
11. Koraishi, O. Teaching english in the age of AI: embracing ChatGPT to optimize EFL materials and assessment. *Lang. Educ. Tech.* **3**, 55–72 (2023).
12. Mutton, T., Hagger, H. & Burn, K. Learning to plan, planning to learn: the developing expertise of beginning teachers. *Teach Teach* **17**, 399–416 (2011).
13. McNamara, D. S. AIED: from cognitive simulations to learning engineering, with humans in the middle. *Int. J. Artif. Intell. Educ.* **34**, 42–54 (2024).
14. Nilsson, P. From lesson plan to new comprehension: exploring student teachers' pedagogical reasoning in learning about teaching. *Eur. J. Teach. Educ.* **32**, 239–258 (2009).
15. Abdelghani, R. et al. GPT-3-driven pedagogical agents to train children's curious question-asking skills. *Int. J. Artif. Intell. Educ.* **34**, 483–518 (2024).
16. Li, Q. The Application of "Problem Chain" Teaching Mode in Senior High School English Reading Teaching. *Educ. Rev. USA* **7**, 1083–1087, https://doi.org/10.26855/er.2023.08.007 (2023).
17. Chen, Y., Jensen, S., Albert, L. J., Gupta, S. & Lee, T. Artificial intelligence (AI) student assistants in the classroom: designing chatbots to support student success. *Inf. Syst. Front.* **25**, 161–182 (2023).
18. Essel, H. B., Vlachopoulos, D., Tachie-Menson, A., Johnson, E. E. & Baah, P. K. The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education. *Int. J. Educ. Technol. High. Educ.* **19**, 57 (2022).
19. Käser, T. & Alexandron, G. Simulated learners in educational technology: a systematic literature review and a turing-like test. *Int. J. Artif. Intell. Educ.* **34**, 545–585 (2024).
20. Pan, L. et al. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *T. Assoc. Comput. Ling.* **12**, 484–506 (2024).
21. Taylan, R. D. Th\asks. *Int. J. Sci. Math. Educ.* **16**, 337–356 (2018).
22. Santagata, R., Zannoni, C. & Stigler, J. W. The role of lesson analysis in pre-service teacher education: an empirical investigation of teacher learning from a virtual video-based field experience. *J. Math. Teach. Educ.* **10**, 123–140 (2007).
23. Hill, H. C., Ball, D. L. & Schilling, S. G. Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *J. Res. Math. Educ.* **39**, 372–400 (2008).
24. Crespo, S. & Sinclair, N. What makes a problem mathematically interesting? Inviting prospective teachers to pose better problems. *J. Math. Teach. Educ.* **11**, 395–415 (2008).
25. Zhu, Y. & Fan, L. Focus on the representation of problem types in intended curriculum: a comparison of selected mathematics textbooks from Mainland China and the United States. *Int. J. Sci. Math. Educ.* **4**, 609–626 (2006).
26. Hu, B. et al. Teaching Plan Generation and Evaluation With GPT-4: Unleashing the Potential of LLM in Instructional Design. *IEEE T. Learn. Technol.* **17**, 1471–1485 (2024).
27. Tichá, M. & Hošpesová, A. Developing teachers' subject didactic competence through problem posing. *Educ. Stud. Math.* **83**, 133–143 (2013).
28. Kabakci Yurdakul, I. et al. The development, validity and reliability of TPACK-deep: A technological pedagogical content knowledge scale. *Comput. Educ.* **58**, 964–977 (2012).
29. Yang, X., Kaiser, G., König, J. & Blömeke, S. Relationship between pre-service mathematics teachers' knowledge, beliefs and instructional practices in China. *ZDM Math. Educ.* **52**, 281–294 (2020).

30. Kwon, O. N. & Ju, M. K. Standards for professionalization of mathematics teachers: policy, curricula, and national teacher employment test in Korea. *ZDM Math. Educ.* **44**, 211–222 (2012).
31. Silver, E. A., Mamona-Downs, J., Leung, S. S. & Kenney, P. A. Posing mathematical problems: an exploratory study. *J. Res. Math. Educ.* **27**, 293–309 (1996).
32. Cooper, G. Examining science education in ChatGPT: an exploratory study of generative artificial intelligence. *J. Sci. Educ. Technol.* **32**, 444–452 (2023).
33. Gutierrez, A. D. C., Denny, P. & Luxton-Reilly, A. in *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* 289–295 (Association for Computing Machinery, Portland, OR, USA, 2024).
34. Smith, J. P. Efficacy and teaching mathematics by telling: a challenge for reform. *J. Res. Math. Educ.* **27**, 387–402 (1996).
35. Hill, H. C. et al. Mathematical knowledge for teaching and the mathematical quality of instruction: an exploratory study. *Cognition Instr.* **26**, 430–511 (2008).
36. Biswas, G. et al. Learning by teaching: a new agent paradigm for educational software. *Appl. Artif. Intell.* **19**, 363–392 (2005).
37. Kim, H., Cho, C.-Y. & Hong, S. W. Impact of agent-based simulation on novice architects' workplace design exploration and trade-offs. *Autom. Constr.* **145**, 104635 (2023).
38. Aher, G., Arriaga, R. I. & Kalai, A. T. in *Proceedings of the 40th International Conference on Machine Learning* Vol. 202 Article 17 (JMLR.org, Honolulu, Hawaii, USA, 2023).
39. Wang, L. et al. A survey on large language model based autonomous agents. *Front. Comput. Sci.* **18**, 186345 (2024).
40. Liffiton, M., Sheese, B. E., Savelka, J. & Denny, P. in *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*, Article 8, 1–11 (Association for Computing Machinery, New York, NY, USA, 2024).
41. Kim, Y. & Baylor, A. L. Research-based design of pedagogical agent roles: a review, progress, and recommendations. *Int. J. Artif. Intell. Educ.* **26**, 160–169 (2016).
42. Virvou, M., Manos, K. & Katsionis, G. in *SMC'03 Conference Proceedings*. 2003 *IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No.03CH37483)*, 4872-4877 (Washington, DC, USA, 2003).
43. Susnjak, T. Beyond predictive learning analytics modelling and onto explainable artificial intelligence with prescriptive analytics and ChatGPT. *Int. J. Artif. Intell. Educ.* **34**, 452–482 (2024).
44. Schroeder, N. L., Romine, W. L. & Craig, S. D. Measuring pedagogical agent persona and the influence of agent persona on learning. *Comput. Educ.* **109**, 176–186 (2017).
45. Shi, H., Shang, Y. & Chen, S.-S. in *Proceedings of the 5th annual SIGCSE/SIGCUE ITiCSE conference on Innovation and technology in computer science education*, 1-4 (Association for Computing Machinery, Helsinki, Finland, 2000).
46. Wei, J. et al. in *Proceedings of the 36th International Conference on Neural Information Processing Systems* Article 1800 (Curran Associates Inc., New Orleans, LA, USA, 2024).
47. Shao, Y., Li, L., Dai, J. & Qiu, X. in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 13153–13187 (Singapore. Association for Computational Linguistics, 2023).
48. Chou, C.-Y., Chan, T.-W. & Lin, C.-J. Redefining the learning companion: the past, present, and future of educational agents. *Comput. Educ.* **40**, 255–269 (2003).
49. Hu, X., Li, W., Geng, X. & Zhao, L. Exploring the effects of different interventions of the problem-oriented teaching model on students' creativity in STEM education. *Res. Sci. Technol. Educ.*, 1–20; https://doi.org/10.1080/02635143.2023.2219622 (2023).
50. Muskens, M., Frankenhuis, W. E. & Borghans, L. Math items about real-world content lower test-scores of students from families with low socioeconomic status. *npj Sci. Learn.* **9**, 19 (2024).
51. Snell-Rood, E. C. et al. Bioinspiration as a method of problem-based STEM education: A case study with a class structured around the COVID-19 crisis. *Ecol. Evol.* **11**, 16374–16386 (2021).
52. Johnson, B. G. et al. Automatic Question Generation for Spanish Textbooks: Evaluating Spanish Questions Generated with the Parallel Construction Method. *Int. J. Artif. Intell. Educ.*, 1-20; https://doi.org/10.1007/s40593-024-00394-1 (2024).
53. Chowrira, S. G., Smith, K. M., Dubois, P. J. & Roll, I. DIY productive failure: boosting performance in a large undergraduate biology course. *npj Sci. Learn.* **4**, 1 (2019).
54. van Kesteren, M. T. R., Krabbendam, L. & Meeter, M. Integrating educational knowledge: reactivation of prior knowledge during educational learning enhances memory integration. *npj Sci. Learn.* **3**, 11 (2018).
55. Tlili, A. et al. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learn. Environ.* **10**, 15 (2023).
56. Dibia, V. et al. in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 72–79.
57. Abouelenein, Y. A. M., Selim, S. A. S. & Elmaadaway, M. A. N. Impact of a virtual chemistry lab in chemistry teaching on scientific practices and digital competence for pre-service science teachers. *Educ. Inf. Technol.* **29**, 2805–2840 (2024).
58. Yuan, L. Where does AI-driven education, in the Chinese Context and Beyond, go next? *Int. J. Artif. Intell. Educ.* **34**, 31–41 (2024).
59. Prescott, A., Bausch, I. & Bruder, R. TELPS: a method for analysing mathematics pre-service teachers' Pedagogical Content Knowledge. *Teach. Teach. Educ.* **35**, 43–50 (2013).
60. Wang, A. Y. Understanding levels of technology integration: A TPACK scale for EFL teachers to promote 21st-century learning. *Educ. Inf. Technol.* **27**, 9935–9952 (2022).
61. Koichu, B. & Kontorovich, I. Dissecting success stories on mathematical problem posing: a case of the Billiard Task. *Educ. Stud. Math.* **83**, 71–86 (2013).
62. Schmid, M., Brianza, E. & Petko, D. Self-reported technological pedagogical content knowledge (TPACK) of pre-service teachers in relation to digital technology use in lesson plans. *Comput. Hum. Behav.* **115**, 106586 (2021).
63. Bostancıoğlu, A. & Handley, Z. Developing and validating a questionnaire for evaluating the EFL 'Total PACKage': Technological Pedagogical Content Knowledge (TPACK) for English as a Foreign Language (EFL). *Comput. Assist. Lang. L.* **31**, 572–598 (2018).
64. Kadıoğlu-Akbulut, C., Çetin-Dindar, A., Küçük, S. & Acar-Şeşen, B. Development and validation of the ICT-TPACK-science scale. *J. Sci. Educ. Technol.* **29**, 355–368 (2020).
65. Nguyen, H. Role design considerations of conversational agents to facilitate discussion and systems thinking. *Comput. Educ.* **192**, 104661 (2023).

Acknowledgements

This work was supported by the National Natural Science Foundation of China [grant numbers 62477013].

Author contributions

B.H. designed and authored the study, generated and processed the teaching plan datasets, revised the assessment dimensions, and analyzed the evaluation results of the teaching plans. J.Z. collected the teaching plan datasets and organized the teaching plan evaluation. Y.P. contributed to data processing for this study and revised its academic expressions. X.G. supervised and reviewed the content of this study.

Competing interests

X.G. serves as the Associate Editor for npj Science of Learning. X.G. did not participate in the review of this manuscript or in any related decision-making for this journal. Beyond this, the authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41539-025-00300-x.

Correspondence and requests for materials should be addressed to Xiaoqing Gu.