Large Language Models in Mathematics Education: Practitioners as Pedagogical Prompt Engineers

Sima Caspari-Sadeghi
Østfold University of Applied Sciences

Submitted:
June 4, 2026
Accepted:
June 29, 2026
Revised:
July 13, 2026
Published:
September 12, 2026

Citation:
Caspari-Sadeghi, S. (2026). Large Language Models in Mathematics Education: Practitioners as Pedagogical Prompt Engineers. *Journal of Higher Education Theory and Practice*, *26*(4), 16–26.

Effective classroom integration of large language models (LLMs) requires domain-specific Prompt Engineering (PE) competence among teachers and students. Existing PE literature remains largely technical or cross-disciplinary, offering limited guidance for mathematics education, where LLMs continue to struggle with mathematical reasoning. This paper proposes a pedagogically grounded framework that positions educators as pedagogical prompt engineers who scaffold students' interactions with LLMs through discipline-specific prompting practices. The framework provides structured classroom exemplars to enhance mathematical reasoning, conceptual understanding, and critical evaluation of AI-generated outputs. By fostering critical AI literacy and human-centered inquiry, it supports deeper, reflective, and ethical engagement with intelligent technologies.

Keywords: critical AI Literacy, large language models, mathematics education, pedagogical prompt engineering

INTRODUCTION

Generative Artificial Intelligence (GenAI)—a branch of artificial intelligence that employs advanced machine learning algorithms to generate original, human-like outputs such as text, images, audio, and video—has significant potential to transform education. By enabling large-scale personalization, adaptive learning, automated assessment, and real-time feedback, GenAI can enhance both teaching and learning processes (Cain, 2024; Caspari-Sadeghi, 2025, 2026a; Kim et al., 2022; Mioa & Cukurova, 2024). It also facilitates the efficient development of educational materials, including lesson planning, interactive content creation, and formative assessment activities. In recent years, GenAI-driven conversational agents powered by Large Language Models (LLMs), such as ChatGPT, Copilot, and Gemini, have gained widespread popularity due to their accessibility, ease of use, user-friendliness, and ability to adapt dynamically to users' needs. Large-scale surveys indicate that students increasingly rely on AI-based chatbots for studying, completing assignments, and supporting their overall learning activities (von Garrel & Mayer, 2023).

Research in mathematics education indicates that LLMs can facilitate self-regulated learning, mathematical modeling, logical reasoning, and abstract problem-solving (Pepin et al., 2025). However, their performance varies across mathematical tasks, with stronger results in arithmetic and Word Math Problems (WMP) compared to Automated Theorem Proving (ATP), where effectiveness is more limited. Despite these advantages, integrating LLMs into mathematics education poses several persistent challenges that warrant further investigation. In the following section, we outline these challenges and delineate the specific objectives that guide the present study.

First, although LLMs have achieved impressive results across various NLP tasks, they often struggle with even basic mathematical tasks. The fragility of LLMs in mathematical reasoning becomes evident when identical questions yield different final answers across multiple trials, often produced through inconsistent reasoning paths. Increasing model size (e.g., GPT-3 to GPT-5) does not necessarily resolve issues such as hallucinations and inaccuracies, particularly in domains that require logical and formal reasoning. This limitation largely stems from the fact that LLMs are pre-trained on massive text corpora, which limits their ability to grasp abstract mathematical reasoning (Ahn et al., 2024; Wei et al., 2022).

Building on this technical limitation, a second challenge concerns the pedagogical implications, even when LLMs do generate correct solutions. It remains unclear how AI can meaningfully support mathematics learning, as mathematical problem-solving rarely follows predetermined procedures or algorithms; rather, it demands metacognitive engagement and deep conceptual understanding. Learners must make sense of problems by activating relevant prior knowledge and exploring alternative strategies (e.g., solving analogous simpler problems or reasoning backward). Moreover, effective problem-solving requires the ability to transfer learned strategies to novel contexts. Thus, it is still uncertain to what extent and under what conditions LLMs can adequately support the diverse cognitive and metacognitive needs of learners in mathematical problem-solving (Schoenfeld, 1992; Schorcht et al., 2024; Wardat et al., 2023).

Extending beyond model capability and pedagogical concerns, a third challenge relates to the low level of AI literacy among both teachers and students. The *UNESCO AI Competency Framework for Teachers* outlines AI literacy as comprising five key dimensions: (a) understanding AI foundations and applications, (b) applying AI in pedagogy, (c) considering ethical implications of AI, (d) using AI for professional development, and (e) adopting a human-centred mindset (Miao & Cukurova, 2024). To effectively harness the potential of LLMs in mathematics education, educators and learners must be able to interact with AI systems purposefully and critically. This includes designing precise instructions for the AI and evaluating the accuracy and validity of AI-generated outputs—competencies often referred to collectively as Prompt Engineering (PE). However, several studies have shown that teachers and students often lack sufficient knowledge and skills in PE, leading to either neglect or overreliance on AI tools. Consequently, this limits their ability to employ AI effectively and critically to support mathematical understanding (Ng et al., 2023; Walter, 2024).

Objectives & Questions

This paper seeks to advance mathematics educators' prompting literacy and enhance their skills as prompt engineers. Specifically, it aims to support teachers in creating effective prompts, employing diverse prompting techniques suited to mathematical tasks, and designing pedagogical activities that embed PE within classroom practice. Furthermore, the paper offers approaches to help students critically evaluate AI-generated mathematical solutions and reflect on their own reasoning processes. This paper aims to address the following research question:

RQ: *How can domain-specific PE techniques be pedagogically integrated into mathematics instruction to both enhance students' mathematical reasoning and critical evaluation of AI-generated outputs?*

This paper is organized into three main sections. First, we provide an overview of the fundamentals of prompt engineering (PE) and review existing frameworks for designing effective prompts in educational contexts. The second section introduces four pedagogical PE strategies specifically suited to mathematics teaching and learning. For each strategy, we present its conceptual foundation, strengths and limitations, and provide examples of classroom implementation. We also propose evaluative and reflective practices to assess the effectiveness of these strategies in supporting student learning outcomes. Collectively, these sections aim to offer both theoretical insight and practical guidance for the purposeful and critical integration of PE into mathematics education.

The aim of this paper is neither to provide a comprehensive review of existing approaches to prompt engineering (PE) nor to report empirical results. Rather, the goal is to offer an easy-to-follow, educator-friendly PE framework that mathematicians can readily use in their classrooms. By presenting a clear structure and practical guidance, we invite instructors to implement the proposed PE and reflect on its usability in real teaching contexts.

The suggested PE techniques were selected for two main reasons. They align with the demands of mathematical methodology, i.e., precision, logical rigor, and error awareness, and reflect key cognitive processes like step-by-step reasoning, task decomposition, and the development of higher-order concepts in areas such as algebra and geometry.

Prompt Engineering (PE)

Prompt Engineering (PE) refers to crafting, designing, and refining a careful instruction or text to guide LLMs to produce desired, high-quality outputs (Park & Choo, 2025; Velasquez-Henao et al., 2023). It is a steering mechanism (Cain, 2023), consisting of two stages: (1) prompting, or designing a well-crafted input (e.g., question, problem, instruction), and (2) engineering, which focuses on iteratively refining, optimizing, and evaluating the interaction with AI systems until the desired output is achieved.

As a critical dimension of AI literacy, PE represents the key to effective human–AI interaction and can be understood from two different perspectives: technical and practical. From a technical perspective, PE constitutes a sub-field of machine learning in which data scientists employ techniques such as fine-tuning, embedding, and specialized prompting strategies (e.g., Retrieval-Augmented Generation, or RAG) to optimize model parameters for specific applications—such as developing customer-support chatbots. From a practical perspective, PE is the skills and strategies that non-expert users employ in conversational interactions with user-friendly LLM-based systems, such as ChatGPT, Copilot, or Gemini.

In this article, we adopt the practical perspective on PE, focusing on its pedagogical potential within educational contexts. Specifically, we examine how teachers and learners—without prior expertise in machine learning or programming—can engage in productive, dialogic interactions with AI systems to support learning. Our approach aligns with the notion of Pedagogical PE (Xiao et al., 2025), which emphasizes designing prompts that instruct AI models to facilitate learning rather than merely provide correct answers. Such prompts aim to scaffold problem-solving, critical thinking, deeper conceptual understanding, and metacognitive awareness.

Review of Literature: PE Frameworks

Empirical studies, especially in the field of human–computer interaction, have shown that non-expert users—those without formal training in AI or with limited AI literacy—engage with LLMs as if they were conversing with a human. This anthropomorphic mode of interaction often leads to frustration, confusion, and difficulties in formulating effective prompts or iteratively refining AI-generated outputs (Dang et al., 2022; Knoth et al., 2024; Zamfirescu-Pereira et al., 2023).

In this section, we review existing frameworks and best-practice principles which guide users, particularly educators and learners, in crafting and refining effective prompts to get the desired output. Therefore, this review does not seek to analyse prompt taxonomies (such as Sahoo et al., 2025), but instead examines pedagogical frameworks that can meaningfully support mathematics educators in implementing prompt engineering practices in their classrooms.

The "*Five S*" Model (AI for Education, 2024) provides a structured approach for educators aiming to enhance the pedagogical utility of LLMs. It consists of five key elements:

1. *S*et the *Scene* – provide context and assign a role to the AI;

2. Be *Simple* – formulate clear and concise instructions;
3. *Specify* the Task – define precisely what the AI should do;
4. *Structure* the Output – indicate the desired length, level, or format; and
5. *Share* Feedback – iteratively refine the response through feedback and revision.

Similarly, Park and Choo (2024) suggested the *PARTS* framework—*Persona, Aim, Recipients, Theme, and Structure*—to support users in composing effective prompts. According to this model, the user should assign a role to the AI (Persona, e.g., "You are a geometry professor"), define the objective and topic (Aim and Theme), identify the intended audience (Recipients), and specify the expected characteristics of the response (Structure, e.g., length, tone, or difficulty level). Furthermore, Park and Choo emphasize the importance of clarity, logical coherence, and critical reflection on the quality of AI-generated responses.

In a parallel contribution, Eager & Brunton (2023) outlined a six-component framework to guide prompt formulation: Verb, Focus, Context, Condition, Alignment, and Constraint. For instance, if a teacher wishes to explore the use of AI in mathematics education, an effective prompt could include:

- Verb: Analyze
- Focus: Survey research
- Context: The field of mathematics
- Condition: Latest trends in AI research
- Alignment: Positive impact on mathematics learning
- Constraint: Only peer-reviewed sources from 2022 on.

Despite differences in terminology, the reviewed frameworks share several core elements, e.g., assigning a clear role to the AI or specifying the task precisely, which can directly support educators by improving the pedagogical usefulness of AI-generated responses. The primary difference is in the level of granularity—while the *Five S* model offers broad, accessible guidance suitable for classroom practice, *PARTS* and the Eager & Brunton (2023) framework introduce more detailed categories (e.g., Recipients, Alignment, Constraints) that allow for more fine-grained alignment when needed.

While such frameworks provide valuable scaffolds for designing prompts, they should be viewed as flexible guidelines and tips rather than rigid, prescriptive formulas. There is no universally superior framework or one-size-fits-all prompting that guarantees the desired output. We therefore suggest that educators consider the following practical strategies when composing prompts:

a) Assign a clear role to the AI.
b) Specify the task explicitly.
c) Describe the desired format or structure of the output.
d) Provide contextual information about the topic and intended audience.

Although contextual information generally enhances model accuracy, excessive detail may reduce relevance when presented with overly superfluous input. Hence, educators and students should approach prompt design with an experimental mindset, i.e., embracing tinkering, patience, trial-and-error, and creativity (Correia, Hickey & Xu, 2025)

To illustrate the difference between weak and effective prompting, we present two examples:

- *Simple prompt: Explain the Pythagorean theorem.*

(This prompt is likely to elicit a general or unfocused explanation, potentially containing irrelevant details).

- *Effective Prompt (5-S model): <u>Act as a university professor of Geometry</u>. Explain the Pythagorean theorem <u>to a freshman in mathematics</u> in academic, yet <u>beginner-friendly</u> language in three paragraphs, <u>including</u> its definition, formula, a numerical example, and a real-life application.*

(This prompt provides clear context, role assignment, structure, and scope, thereby improving the precision, coherence, and educational relevance of the AI-generated response).

Pedagogical PE in Mathematics

In this section, we introduce four mathematics-specific PE techniques, each of which alters the nature of the input provided to the AI model and, consequently, produces a distinct type of output. These techniques are: *Few-shot, Chain-of-Thought (CoT), Self-Consistency,* and *Take-a-Step-Back.* Each approach enables a unique form of reasoning and scaffolding when engaging with mathematical problems. Although a wide range of prompt engineering (PE) techniques exist, e.g., Sahoo et al. identified 46 strategies, and the list continues to grow, the techniques selected for this study (*Few-shot, Chain-of-Thought, Self-Consistency,* and *Take-a-Step-Back* prompting) were chosen for two main reasons. First, they align well with the epistemological and methodological demands specific to mathematics, including precision, logical validity, rigor of proof, and sensitivity to error. Second, they closely reflect the cognitive processes involved in mathematical learning. Mathematical problem-solving typically relies on step-by-step reasoning, the decomposition of complex tasks, and the construction of higher-order conceptual structures, particularly in areas such as algebra and geometry. The chosen techniques directly support these processes: *Chain-of-Thought* and *Self-Consistency* promote explicit reasoning, *Few-shot* offers example-based scaffolding, and *Take -a-Step-Back* facilitates abstraction and conceptual reframing before generating the solution. Consequently, these methods are especially well suited to the pedagogical requirements of mathematics education.

It is important to note that different AI platforms employ varying model architectures, training data, and computational capacities. As a result, the same prompt may elicit different outputs across platforms, and the effectiveness of a given prompting technique can vary depending on the underlying model and its training context.

In the following subsections, we define each prompting technique in detail, discuss their pedagogical strengths and limitations, and offer practical guidance for classroom implementation. This includes examples of effective prompting text, suggested learning activities, and strategies for formative evaluation. The target participants are university-level educators and students, including both mathematics majors and pre-service mathematics teachers.

The successful implementation of these prompting techniques in mathematics education depends on teachers and students developing three interrelated forms of competency or literacy, namely:

1. *Mathematics Literacy*: a deep understanding of disciplinary knowledge in mathematics, in this study geometry, is essential for crafting meaningful prompts and for engaging productively with AI-generated reasoning. Teachers and students must be able to interpret, critique, and extend the mathematical content that emerges from AI interactions.
2. *Prompt Literacy*: This refers to the ability to design, refine, and adapt prompts to elicit desired responses from AI systems. It includes awareness of how to frame problems, apply constraints, and iteratively modify the AI's focus to enhance reasoning accuracy and relevance. Developing prompt literacy requires both an acceptable understanding of AI interaction and creative experimentation.
3. *Critical Thinking*: Effective engagement with AI tools necessitates analytical reasoning, evaluative judgment, and reflective practice. Teachers and students must be able to assess the validity and reliability of AI-generated responses and recognize limitations, such as hallucinations or embedded biases.

Together, these literacies provide a foundation for integrating AI responsibly and effectively into mathematics education, enabling AI to serve not as a replacement for human reasoning but as a cognitive tool that supports conceptual development, problem-solving, and metacognitive awareness. To implement the following proposed practical PE techniques in classroom settings, we propose three methodological pathways, each aligned with different levels of research expertise and classroom constraints:

(a) *Design-Based Research (DBR*) can offer a robust way for iterative development and refinement of pedagogical prompt engineering practices, provided that the mathematics educator collaborates with an instructional designer or an expert in educational research capable of guiding systematic cycles of design, enactment, analysis, and redesign.
(b) *Action Research* can be another viable option for mathematics educators who are familiar with conducting small-scale, practitioner-led classroom inquiries; through reflective cycles of planning, acting, observing, and reflecting, teachers can adapt and study the framework within the realities of their own instructional context.

(c) Finally, *Formative Assessment* presents a low-effort, highly accessible method in which teachers can try the proposed practical strategies during everyday instruction in a naturalistic manner, gathering ongoing evidence of student understanding and adjusting PE activities responsively. The pedagogical activities for integrating PE techniques proposed in this article are therefore best suited for formative assessment.

Together, these methodological options offer flexible yet rigorous routes for implementing, evaluating, and refining the proposed pedagogical approach across diverse mathematics classrooms.

Few-Shot Prompting (FS)

When users simply ask a question and give chatbots a problem to solve, without any examples or tailor-made instructions, it is called *Zero-shot* prompting. Although zero-shot prompting is efficient and straightforward, especially for exploring a new concept or topic, it often produces overly broad, vague, and sometimes inconsistent or irrelevant responses. A powerful alternative PE technique is *Few-Shot* (FS) prompting, in which few examples or instances (known as shots) are provided to LLMs to guide them toward more accurate, relevant, and contextually appropriate responses (Brown et al., 2020). This approach resembles reasoning by analogy, in which understanding is developed through generalization from specific instances, and it aligns with observational learning theory, which suggests that humans learn more effectively by observing, imitating, and modeling another person in practice (Bandura & Jeffrey, 1973). By presenting well-chosen examples, the model gains a clearer understanding of the user's expectations and can more reliably follow the intended reasoning patterns. While even a single example (one-shot prompting) can substantially improve zero-shot performance, research indicates that providing approximately 3 examples typically yields the most robust results.

Despite its potential, the major challenge in FS prompting is selecting high-quality, diverse, and well-written examples. A small mistake can confuse and undermine the model (Boonstra, 2025). Moreover, FS prompting is less effective for logical or highly abstract problems that require multi-step or symbolic reasoning. In such cases, *Chain-of-Thought* prompting performs better. Below, we provide an example of how FS prompting can be used in a geometry classroom:

- *FS Prompt: <u>You are an expert</u> in Geometry. Here are <u>two examples</u> of how the Pythagorean Theorem works:*
  - *A triangle with sides 3, 4, and 5 is right-angled.*
  - *A triangle with sides 5, 12, and 13 is also right-angled.*

Now, <u>use these two examples to check whether</u> a triangle with sides 7, 24, and 25 is right-angled.

<u>Pedagogical activity</u>: Students work in pairs to design two well-written and mathematically correct examples to serve as input (shots) for ChatGPT. Then, they write a short reflection on how providing shots or examples improved the model's performance. Students work will be assessed based on two criteria: (a) the quality of the examples in prompt and (b) the depth of critical thinking on the AI-generated responses.

Chain-of-Thought (CoT) Prompting

Chain-of-Thought (CoT) prompt mimics human logical reasoning by decomposing a complex problem into logical intermediate steps (Wei et al., 2022). It is similar to concurrent think-aloud protocols, in which problem solvers explicitly verbalize the sequential steps or strategies they adopt to reach a solution (Federiakin et al., 2024). *CoT* is shown to perform particularly well in mathematics, especially on arithmetic and multi-step tasks that require commonsense reasoning. Compared to FS prompting, *CoT* offers practical advantages: it is low-effort and does not need any curated exemplars. Moreover, LLMs are often criticized for their lack of explainability and interpretability (known as black box models). Requesting intermediate reasoning steps involved in arriving at the final solution makes the model's internal process visible and open to verification, i.e., when malfunctions or errors occur, they can be detected and corrected. This transparency is particularly relevant in mathematics, where visible stepwise solutions provide direct, formative feedback on students' misconceptions and misunderstandings *CoT* is therefore well-suited to word math problems and other tasks that require multiple, clearly articulated reasoning steps.

- *CoT Prompt: You are a geometry professor explaining to a first-year student. Determine whether a triangle with sides 3 cm, 4 cm, and 5 cm is right-angled? explain your reasoning step by step.*

Pedagogical Activity: Students work individually to solve a problem independently and then with CoT prompting. They compare their work with a CoT solution and answer a questionnaire with three questions: (a) Did the AI get all steps correct? (b) Was the final answer correct? (c) Was the AI's reasoning in line with your expectations? In small groups, students discuss their reflections on the effectiveness and pitfalls of CoT solutions.

Self-Consistency Prompting

Self-Consistency prompting (Wang et al., 2022) is an advanced extension of CoT prompting that enhances model accuracy and coherence in complex mathematical reasoning tasks. This approach involves generating multiple reasoning paths, similar to producing several CoTs, and selecting the most frequently occurring solution. Unlike CoT, which relies on greedy decoding to produce a single, preferred answer, Self-Consistency generates diverse, valid reasoning paths and determines the final answer through majority voting, thereby improving robustness and reliability. This technique uses "wisdom of the crowd" principle, assuming collective judgments from a group often produce more accurate decisions than those of a single individual. In this context, each reasoning path generated by the model is treated as an independent solution, similar to an individual contribution (Federiakin et al., 2024; Surowiecki, 2004).

However, Self-Consistency is not without limitations. For instance, when several flawed reasoning paths share the same misconception (e.g., applying a theorem that requires a right triangle to a non-right triangle), their agreement does not ensure correctness. Moreover, due to its higher computational resource and energy demands, Self-Consistency should be applied judiciously. Below, we provide two different ways to use Self-Consistency prompting, once alone and once in combination with CoT:

- *Self-Consistency Prompting (combined with CoT): You are my geometry professor in an introductory course, guide me through solving the problem: "In a right-angled triangle with legs of length 5 cm and 12 cm, find the hypotenuse." Complete the following tasks: 1) Generate three independent solutions using different paths, (2) For each path, provide detailed step-by-step reasoning. 3) Compare the results and select the most consistent final answer. 4) Provide a one-sentence explanation justifying your final answer.*

Pedagogical activity: The main pedagogical advantage of using the Self-Consistency technique is its ability to promote critical thinking and creativity simultaneously, as it allows students to see different perspectives and alternative solutions to the same problem. Students work individually, use Self-Consistency technique to solve a problem. Then, they rank alternative suggested solutions with rubric in terms of their (a) mathematical soundness, (b) completeness, and (c) clarity. This structured comparison helps students (i) diagnose misconceptions, (ii) recognize the strength and limits of different methods, and (iii) articulate principled preferences among solution strategies.

Take-a-Step-Back Prompting

Take-a-Step-Back prompting is a technique designed to enhance a model's critical self-reflection and metacognitive reasoning during complex, multi-step problem solving (Zheng et al., 2023). The approach involves two sequential stages: Abstraction and Reasoning. In the Abstraction stage, the model identifies and extracts the abstract concepts, fundamental principles, and underlying high-level relationships relevant to the question. In the Reasoning stage, the model applies these high-level abstractions to reason through a specific problem. Using "step back" before problem-solving activates relevant background knowledge, reducing cognitive biases, hallucinations, and distorted reasoning. Empirical evidence indicates that Take-a-Step-Back prompting outperforms traditional CoT prompting by up to 27% on reasoning-intensive STEM tasks (Sahoo et al., 2025). This improvement is attributed to the method's strategic emphasis on conceptual understanding before procedural reasoning: instead of proceeding directly to linear, logic-based thinking, Take-a-Step-Back prompting facilitates a deeper engagement with the abstract principles that underlie problem-solving processes.

In practical application, Step-Back prompting involves a two-phase structure. First, the original question followed by a query such as "What are the fundamental principles involved?" is inserted as a prompt. Once the model's conceptual explanation is generated and validated, the second prompt is: "Now use these laws and principles to solve this problem."

Example of Take-a-Step-Back prompting in mathematics:

- *Step-back Prompting (Abstraction): You are a Geometry professor, explaining the following topic to a first-year university student in mathematics. First, take a step back and explain the high-level concepts or first principles related to the Pythagorean theorem.*
- *AI Response: The model explains foundational mathematical ideas such as Euclidean geometry, orthogonality, and the relationship between right-angled triangles and distance measurement.*
- *Step-back Prompting (Abstraction): Now, use these principles to explain why the theorem is inapplicable to a spherical surface.*

Pedagogical activity: students engage in Think-Pair-Share, where they first try Take-a-Step-Back prompting individually. Then, they compare their results to their peers. This peer evaluation focuses on (a) effectiveness of the prompting strategy, and (b) accuracy of the AI's responses at both the abstraction and reasoning stages. Finally, students synthesize their collaborative reflections and share their conclusions with the teacher, promoting collective metacognition and deeper conceptual understanding.

Summary and Limitations

Table 1 summarizes all prompting techniques and their corresponding pedagogical integration strategies discussed in the preceding sections.

TABLE 1
SUMMARY OF PE TECHNIQUES WITH THEIR PEDAGOGICAL DESIGN

Prompting Technique	**Learning Goals**	**Teacher Activities**
Few-shot prompting (FS)	Create clear examples; understand how examples shape AI output; reflect on model performance.	Explain technique; guide pairs in designing two examples; scaffold brief reflection.
Chain-of-Thought prompting (CoT)	Evaluate step-by-step reasoning; compare human vs. AI solutions; reflect on accuracy and reasoning quality.	Assign a problem; instruct students to solve independently and with CoT; provide comparison questions; facilitate small-group reflection.
Self-Consistency Prompting	Evaluate multiple solution paths; diagnose misconceptions; recognize strengths/limits of methods; articulate justified preferences.	Introduce technique; assign problem; provide rubric for ranking solutions; guide reflection on alternative solution strategies.
Take-a-Step-Back prompting	Strengthen metacognition; assess abstraction vs. reasoning accuracy; evaluate prompting effectiveness through peer comparison.	Introduce technique; facilitate Think–Pair–Share; guide peer evaluation; coordinate synthesis of group conclusions.

A key limitation of this study is that the proposed pedagogical framework and the suggested mathematics-specific PE activities remain conceptual and have not yet been empirically implemented or evaluated in real classroom settings. Without classroom-based trials, the actual feasibility of these strategies for educators with varying levels of AI literacy, how students interact with LLM-generated outputs during mathematical problem solving, or whether the suggested techniques meaningfully improve reasoning, collaboration, or conceptual understanding remains unclear.

Additionally, we have not addressed practical constraints such as time, training requirements, or technological access. Therefore, the effectiveness, scalability, and potential unintended consequences of the proposed human-centered PE framework remain open for future empirical investigation.

Finally, the effectiveness of the proposed prompt engineering techniques may vary depending on the AI model, platform, and context, limiting the generalizability of the findings across different educational environments.

CONCLUSION

Prompt Engineering (PE) represents a transformative development in the educational use of Large Language Models (LLMs), as it eliminates the need for extensive data training, model construction, and fine-tuning. Instead, it relies on user input to activate and direct the relevant knowledge embedded in pre-trained models (Sahoo et al., 2025). Although the existing literature acknowledges potentials of LLMs to enhance teaching and learning, their integration into mathematics education remains challenging due to relatively poor mathematical reasoning performance of current LLMs and the limited level of PE literacy among both educators and students.

Rather than focusing on improving LLM performance on mathematical tasks, this paper adopts a human-centered perspective on PE, emphasizing how educators can strategically integrate LLMs into mathematics instruction to foster student learning. We proposed a set of practical, classroom-ready pedagogical activities for implementing four mathematics-specific PE techniques, namely *Chain-of-Thought (CoT), Few-Shot (FS), Self-Consistency, and Take-a-Step-Back* prompting, as tools to foster students' problem-solving, logical reasoning, and critical thinking skills in mathematics. Each activity is tailored to engage learners actively and to promote collaborative exploration, reflection, and professional judgment. Developing PE competence among educators and students enhances their agency to act as active co-designers in their interactions with AI, rather than as passive recipients of AI-generated outputs.

For professional teacher development, these findings highlight the need for embedded learning opportunities where educators practice domain-specific PE techniques, e.g., *CoT, FS, Self-Consistency,* and *Take-a-Step-Back* prompting, within authentic mathematical tasks. Such activities help teachers develop the technical fluency and pedagogical judgment needed to integrate these techniques effectively into instruction (Caspari-Sadeghi, 2026b).

Nonetheless, LLMs exhibit several limitations. Beyond general concerns such as data privacy, user protection, and cognitive overreliance, there are domain-specific pedagogical risks. A particularly critical issue is the reinforcement of misconceptions: when AI-generated mathematical explanations contain inaccuracies that are accepted uncritically, they may introduce new misunderstandings or reinforce existing errors in students' conceptual frameworks (Yen & Hsu, 2023).

To mitigate these risks, one promising pedagogical approach involves asynchronous collaborative learning, in which students discuss, critically evaluate, and compare diverse AI-generated responses before submitting their final solutions for teacher verification. Additional strategies include raising educators' and students' critical awareness of AI's limitations, maintaining human judgment in the decision-making loop, and ensuring human accountability for pedagogical actions based on AI-assisted reasoning. Ultimately, developing educators' and students' PE literacy should not be treated as an instrumental goal; rather, it is essential for shaping a responsible, human-centered AI pedagogy.

REFERENCES

Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R., & Yin, W. (2024). *Large language models for mathematical reasoning: Progresses and challenges*. *arXiv* preprint arXiv:2402.00157

AI for Education. (2024). *Prompt framework for educators: The Five "S" model*. Retrieved from https://www.aiforeducation.io/ai-resources/the-five-s-model

Bandura, A., & Jeffrey, R.W. (1973). Role of symbolic coding and rehearsal processes in observational learning. *Journal of Personality and Social Psychology*, *26*(1), 122–130.

Boonstra, L. (2025). *Prompt engineering* [White paper]. Innopreneur. Retrieved from https://www.innopreneur.io/wp-content/uploads/2025/04/22365_3_Prompt-Engineering_v7-1.pdf

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Caspari-Sadeghi, S. (2026a). Adaptive learning systems: Leveraging generative artificial intelligence for optimal adaptivity. In B. Wasson & C.E. Tømte (Eds.), *Handbook of learning analytics and adaptive learning in compulsory education* (pp. 21–34). Edward Elgar Publishing. https://doi.org/10.4337/9781035330676.00009

Caspari-Sadeghi, S. (2026b). AI literacy for teacher educators: A holistic curriculum for capacity-building in higher education. *Frontiers in Education*.

Caspari-Sadeghi, S. (2025). *Technology-enhanced and formative assessment* (Unpublished habilitation thesis). University of Passau.

Cain, W. (2024). Prompting change: Exploring prompt engineering in large language model AI and its potential to transform education. *TechTrends*, *68*, 47–57.

Correia, A.-P., Hickey, S., & Xu, F. (2025). Realizing the possibilities of large language models: Strategies for prompt engineering in educational inquiries. *Theory Into Practice*, *64*(4), 434–447.

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). *How to prompt? Opportunities and challenges of zero- and few-shot learning for human-AI interaction in creative applications of generative models*. arXiv preprint arXiv:2209.01390.

Eager, B., & Brunton, R. (2023). Prompting higher education towards AI-augmented teaching and learning practice. *Journal of University Teaching and Learning Practice*, *20*(5).

Federiakin, D., Molerov, D., Zlatkin-Troitschanskaia, O., & Maur, A. (2024). Prompt engineering as a new 21st century skill. *Frontiers in Education*, *9*, 1366434.

von Garrel, J., & Mayer, J. (2023). Artificial intelligence in studies—Use of ChatGPT and AI-based tools among students in Germany. *Humanities and Social Sciences Communications*, *10*, 799.

Kim, J., Lee, H., & Cho, Y.H. (2022). Learning design to support student-AI collaboration: Perspectives of leading teachers for AI in education. *Education and Information Technologies*, *27*(5), 6069–6104.

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J.M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, *6*, 100225.

Lo, L.S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, *49*(4), 102720.

Miao, F., & Cukurova, M. (2024). *AI competency framework for teachers*. UNESCO.

Ng, D.T.K., Su, J., Leung, J.K.L., & Chu, S.K.W. (2023). Artificial intelligence (AI) literacy education in secondary schools: A review. *Interactive Learning Environments*, 1–21.

Park, J., & Choo, S. (2024). Generative AI prompt engineering for educators: Practical strategies. *Journal of Special Education Technology*, *40*(3), 411–417.

Pepin, B., Buchholtz, N., & Salinas-Hernández, U. (2025). A scoping survey of ChatGPT in mathematics education. *Digital Experiences in Mathematics Education*, *11*(1), 9–41.

Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S.S., & Chadha, A. (2024). *A systematic survey of prompt engineering in large language models: Techniques and applications*. *arXiv* preprint arXiv:2402.07927.

Schoenfeld, A.H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D.A. Grouws (Ed.), *Handbook for research on mathematics teaching and learning* (pp. 334–370). Macmillan.
Schorcht, S., Buchholtz, N., & Baumanns, L. (2024). Prompt the problem: Investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. *Frontiers in Education*, *9*, 1386075.
Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. Doubleday.
Tassoti, S. (2025). Assessment of students' use of generative artificial intelligence: Prompting strategies and prompt engineering in chemistry education. *Journal of Chemical Education*, *101*(6), 2475–2482.
Velásquez-Henao, J.D., Franco-Cardona, C.J., & Cadavid-Higuita, L. (2023). Prompt engineering: A methodology for optimizing interactions with AI-language models in the field of engineering. *Dyna*, *90*(230), 9–17.
Walter, Y. (2024). Embracing the future of artificial intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, *21*, 1–29.
Wang, B., Rau, P.L.P., & Yuan, T. (2023a). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, *42*(11), 1324–1337.
Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., & Zhou, D. (2022). *Self-consistency improves chain of thought reasoning in language models*. arXiv preprint arXiv:2203.11171.
Wardat, Y., Tashtoush, M., Alali, R., & Jarrah, A. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*(7), 1–18.
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.
Xiao, Y., Hou, X., Ye, R., Kazemitabar, M., Diana, N., . . . Stamper, J. (2025). *Improving student–AI interaction through pedagogical prompting: An example in computer science education*. *arXiv* preprint arXiv:2506.19107.
Yen, A.-Z., & Hsu, W.L. (2023). *Three questions concerning the use of large language models to facilitate mathematics learning*. arXiv preprint arXiv:2310.13615.
Zamfirescu-Pereira, J.D., Wong, R.Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–21.
Zheng, H.S., Mishra, S., Chen, X., Cheng, H., Chi, E.H., Le, Q.V., & Zhou, D. (2023). *Take a step back: Evoking reasoning via abstraction in large language models*. arXiv preprint arXiv:2310.06117