Integration of generative AI in creative mathematics learning: A correlational analysis between the quality of prompts and the logical reasoning of student teachers in the education degree programme

Integración de la IA generativa en el aprendizaje creativo de las matemáticas: un análisis correlacional entre la calidad de las indicaciones y el razonamiento lógico de los estudiantes de prácticas del grado en educación

Nanang Supriadi
Universitas Islam Negeri Raden Intan Lampung, Bandar Lampung, Indonesia
nanangsupriadi@radenintan.ac.id

Suherman Suherman
Universitas Islam Negeri Raden Intan Lampung, Bandar Lampung, Indonesia
suherman@radenintan.ac.id

Abstract
The emergence of Generative Artificial Intelligence (GenAI) has transformed the paradigm of mathematics education, while simultaneously introducing new pedagogical challenges in the form of "AI hallucinations" and the tendency of students to passively accept machine-generated answers. This study aims to investigate the correlation between prompt engineering quality (textual instructions) and students' mathematical logical reasoning accuracy when solving non-routine integral problems assisted by GenAI. Using an explanatory sequential mixed-methods design, this research involved 85 first-year university students. Quantitative data were extracted from assessment rubrics and worksheets and then analysed using Spearman's Rank Correlation test. Qualitative data were analysed using thematic analysis. The statistical results indicate a strong and highly significant positive correlation ($\rho = 0.784$, $p < 0.001$) between prompt quality and precision of logical reasoning. Thematic analysis reveals that students with low-quality prompts (limited to copy-paste behaviour) experienced blind epistemic trust, which led to a breakdown of mathematical reasoning. In contrast, students who applied iterative and algorithmic prompting were able to use AI as cognitive scaffolding to verify and challenge machine hallucinations. The study concludes that the effectiveness of AI is highly dependent on prompt literacy and highlights the need for educational institutions to train students to transition from passive users to critical-logical collaborators.
Keywords: Generative AI, prompt engineering, mathematical logical reasoning, AI hallucinations, calculus.

Resumen
La aparición de la Inteligencia Artificial Generativa (GenAI) ha transformado el paradigma de la educación matemática, al mismo tiempo que ha introducido nuevos desafíos pedagógicos en forma de "alucinaciones de la IA" y la tendencia de los estudiantes a aceptar pasivamente las respuestas generadas por la máquina. Este estudio tiene como objetivo investigar la correlación entre la calidad de la ingeniería de prompts (instrucciones textuales) y la precisión del razonamiento lógico matemático de los estudiantes al resolver problemas de integrales no rutinarios asistidos por GenAI. Utilizando un diseño de métodos mixtos secuencial explicativo, esta investigación involucró a 85 estudiantes de primer año universitario. Los datos cuantitativos se extrajeron de rúbricas de evaluación y hojas de trabajo, y

posteriormente se analizaron mediante la prueba de correlación de rangos de Spearman. Los datos cualitativos se analizaron mediante análisis temático. Los resultados estadísticos indican una correlación positiva fuerte y altamente significativa ($\rho = 0.784$, $p < 0.001$) entre la calidad del prompt y la precisión del razonamiento lógico. El análisis temático revela que los estudiantes con prompts de baja calidad (limitados al comportamiento de copiar y pegar) experimentaron una confianza epistémica ciega, lo que condujo al colapso del razonamiento matemático. En contraste, los estudiantes que aplicaron estrategias de prompting iterativas y algorítmicas fueron capaces de utilizar la IA como andamiaje cognitivo para verificar y cuestionar las alucinaciones del sistema. El estudio concluye que la efectividad de la IA depende en gran medida de la alfabetización en prompts y destaca la necesidad de que las instituciones educativas formen a los estudiantes para pasar de usuarios pasivos a colaboradores crítico-lógicos.
Palabras clave: IA generativa, ingeniería de prompts, razonamiento lógico matemático, alucinaciones de la IA, cálculo.

1. Introduction

The presence of Generative Artificial Intelligence (GenAI) based on Large Language Models (LLMs) such as ChatGPT, Gemini, and Claude has triggered a massive paradigm shift in mathematics education. GenAI no longer functions as an information search engine (Luo et al., 2025); it has evolved into a cognitive partner capable of performing complex mathematical reasoning. The integration of GenAI into creative mathematics learning is particularly important because it enables students to explore multiple solution strategies, develop original mathematical ideas, and refine their reasoning through continuous interaction and feedback. As creativity in mathematics requires both divergent thinking and logical justification, GenAI provides opportunities for students to articulate, evaluate, and improve their mathematical arguments in real time. Various recent studies show that the use of artificial intelligence in STEM education can enhance student engagement and offer personalised learning (Huang et al., 2025; Nikhil, 2025). However, the transition from conventional learning to AI-based interaction presents new pedagogical challenges, particularly regarding how students communicate with machines to structure their mathematical thinking.

Although GenAI has high computational capabilities, this model is prone to "AI hallucinations" a phenomenon where the machine generates answers that appear procedurally convincing but are conceptually and logically incorrect (Serrano, 2026). In the context of advanced mathematics such as Calculus, particularly the topic of Integrals, this error often occurs in solutions that require visual geometric interpretation (such as the area between curves) or understanding the meaning of the integration constant (+C). Ironically, the majority of students tend to exhibit passive behaviour by copy-pasting the problems directly into the AI system without providing clear contextual boundaries (Adnan et al., 2025). This results in students receiving incorrect answers, which ultimately degrades the accuracy of their mathematical logical reasoning.

The literature review mostly still focuses on evaluating the accuracy of AI answers or the impact of AI on maths anxiety (Chen et al., 2025; Polydoros et al., 2025; Wang, 2025). Several recent studies have also examined students' acceptance of AI, perceptions of ChatGPT, and the effectiveness of AI-assisted learning environments. However, these studies generally treat prompts as a fixed mechanism for obtaining responses, rather than

as a student-generated cognitive process that can influence the outcomes of mathematical reasoning. As a result, the role of prompt construction quality remains underexplored, particularly in mathematics education, where problem solving requires precise logical structures, mathematical representations, and contextual information. There are still very few studies that explore the input side of that interaction, namely the Prompt Engineering skills performed by students. Existing research on prompt engineering has emerged largely from computer science, natural language processing, and professional AI applications, with limited attention given to how learners formulate prompts in educational contexts. Even fewer studies have investigated how variations in prompt quality are associated with students' logical reasoning performance while solving mathematical problems. Consequently, it remains unclear whether effective prompting merely improves AI-generated answers or also reflects and supports students' underlying mathematical thinking processes. In fact, prompt engineering in mathematics is not just a technical typing skill but a new form of mathematical literacy. The process of composing prompts requires students to translate mathematical abstractions into a structured instructional language. Understanding how students naturally structure their first prompts is a critical area that has not been extensively mapped. This gap is particularly important because students' initial prompts may reveal how they conceptualise mathematical problems, organise relevant information, and communicate reasoning strategies to AI systems. Investigating these processes can contribute to a deeper understanding of the cognitive mechanisms underlying human–AI collaboration in mathematics learning.

Based on this literature gap, this research aims to explore and analyse the natural prompt engineering strategies used by students in solving non-routine mathematics problems on the topic of Integrals. By focusing on the input side of AI interaction, this study directly addresses the limitation of previous research that has primarily concentrated on AI outputs, learning outcomes, or students' attitudes toward AI. Rather than evaluating whether AI provides correct answers, the present study examines how the quality and complexity of student-generated prompts relate to the development of logical mathematical reasoning. Non-routine integral problems were selected because they require students to combine conceptual understanding, graphical interpretation, and multi-step reasoning rather than merely applying memorised procedures. Such problems provide a suitable context for examining whether students can communicate mathematical information effectively when interacting with AI systems. Specifically, this study will investigate the correlation between the complexity of the prompts given to students and the accuracy of the mathematical logical reasoning produced after interacting with AI. Student teachers constitute an important group of participants because they represent future mathematics teachers who are expected not only to solve mathematical problems, but also to critically evaluate and integrate AI technologies into instructional practice. Understanding how they interact with GenAI can provide valuable information on the development of AI-supported pedagogical competencies in teacher education. Thru naturalistic observation and think-aloud protocol analysis based on student Worksheets, the findings of this research are expected to provide a new pedagogical framework on how to train students to transition from passive users (copy-paste users) to logical collaborators in the era of artificial intelligence. The findings of this study are expected to contribute to both mathematics education and teacher education by providing empirical evidence on how prompt quality influences logical reasoning, thus informing the effective integration of generative AI into creative mathematics learning.

Accordingly, this study addresses the following research questions: (RQ1) What prompt engineering strategies do student teachers employ when using generative AI to solve non-routine integral problems? (RQ2) What is the relationship between prompt quality and mathematical logical reasoning accuracy in AI-assisted mathematics problem solving? and (RQ3) What cognitive mechanisms explain the relationship between prompt quality and mathematical logical reasoning accuracy among student teachers?

2. Method

Research design

This research used an explanatory sequential mixed-methods design. The first phase used a correlational quantitative approach to measure the relationship between the level of prompt engineering skills and the precision of students' mathematical logical reasoning. The second phase used a qualitative approach through thematic analysis of chat logs and metacognitive reflection questionnaires (as supporting instruments) to fully explain the cognitive phenomena and epistemic trust (blind trust in AI) underlying the correlation findings from the first phase. Qualitative analysis was informed by quantitative results, with particular attention given to students who represented contrasting levels of prompt quality and logical reasoning performance, allowing researchers to explore the cognitive processes underlying the observed statistical relationship.

Participants

This study involved 85 first-year students from the Mathematics Education programme at the Universitas Islam Negeri Raden Intan Lampung. The participants had a mean age of 18.44 years ($SD = 0.50$). Participants were selected using purpose-sampling techniques with the inclusion criteria: (1) currently taking a Calculus course (intra-core material) and (2) having basic experience using Generative AI interface interactions. Purposive sampling was used to ensure that all participants had the prerequisite mathematical background and AI experience required to engage meaningfully with the integral problem-solving tasks. However, because the sample was drawn from a single cohort at one institution, the findings should be interpreted with caution and may not be fully generalisable to other student populations or educational contexts. Demographic characteristics were designed to capture variations in the natural communication styles of students of diverse cultural backgrounds and secondary education histories. The details of the demographic profiles of the participants are presented in Table 1.

Table 1.
Demographics of the sample

Category	Classifications	Frequency	Percentage
Gender	Girls	55	64.7%
	Boys	30	35.3%
Age	18 years old	48	56.5%
	19 years old	37	43.5%
Previous high school	Public	60	70.6%
	Private	12	14.1%
	Vocational/Islamic	13	15.3%
Ethnics	Lampung	38	44.7%
	Java	28	32.9%
	Minangkabau	12	14.1%
	Others	7	8.3%

Instrument

To precisely measure independent and dependent variables, as well as correlational phenomena (thematic analysis), the research instrument is divided into three main components that are integrated and modified in the context of mathematical interaction with GenAI.

Integral Problem-Based student worksheet

This instrument consisted of three Higher-Order Thinking Skills (HOTS) questions on integral calculus topics: (1) Area Trap, (2) Integration Constant, and (3) Leaky Tank Modelling. The characteristics of the questions were adopted from the Creative Mathematically Based Reasoning (CMR) framework developed by Lithner (2017), which emphasises conceptual justification rather than procedural memorisation. The construction of the questions was further adapted based on the guidelines for testing AI hallucinations in mathematics proposed by Wardat (2023), in which the problems were intentionally designed to trigger computational errors when students failed to provide precise and well-structured instructions to AI systems. To ensure the accuracy and appropriateness of the instrument, three experts conducted content validity tests, consisting of two mathematics education specialists and one educational technology expert. The expert evaluation produced a Content Validity Index (CVI) score of 0.92, exceeding the recommended acceptance threshold of 0.79 and indicating that the instrument possessed a high level of content validity for use as an AI interaction trigger instrument.

Prompt Quality Coding Rubric

This instrument was used to assess screenshots and digital chat logs documenting interactions between students and AI systems. The rubric was adapted from the prompt engineering taxonomy in STEM education proposed by Wang et al. (2023) and Mollick and Mollick (2023). Accordingly, the rubric operationalised the construct of prompt quality by translating the theoretical dimensions of prompt engineering into five observable levels of students' prompting behaviour. The researchers modified the framework into a five-level ordinal scale tailored to the cognitive domain of mathematics:

Level 1 (Naive Copy-Paste), Level 2 (Basic Instructional), Level 3 (Contextual), Level 4 (Algorithmic/Chain-of-Thought) and Level 5 (Iterative/Debugging). Because the coding process involved qualitative interpretation of textual interactions, the rubric was evaluated using inter-rater reliability testing. Two researchers independently coded 30% of the chat log samples, resulting in a Cohen's Kappa coefficient of $\kappa = 0.86$ ($p < 0.001$). This result indicates a very strong level of agreement between raters and supports the reliability of the coding instrument.

Rubric for Assessing the Accuracy of Mathematical Logical Reasoning

This instrument was used to assess the final written responses of the students on the worksheet after interacting with AI. The rubric was developed based on the mathematical argumentation assessment guidelines proposed by Bieda (2010) and the mathematical critical thinking framework developed by Aizikovitsh-Udi & Cheng (2015), which emphasize the ability to identify logical fallacies. Accordingly, the rubric operationalised the construct of mathematical logical reasoning by translating these theoretical dimensions into observable performance indicators reflecting the quality of students' mathematical arguments. The assessment used a scoring scale of 0 to 3 for each question: Score 0 (Illogical/Blind Acceptance) indicated verbal acceptance of AI-generated but conceptually incorrect answers without manual verification; Score 1 (Partial Reasoning/Flawed Logic) indicated recognition of AI errors, although the independently constructed calculus solution still contained logical flaws; Score 2 (Logical but Incomplete) indicated correct reasoning and conclusions, although the proof argument was not adequately articulated; and Score 3 (Accurate and Rigorous) indicated a fully accurate conclusion supported by a rigorous proof argument that successfully refuted the hallucination of AI. Similarly to the prompt coding rubric, this reasoning assessment rubric also underwent an independent inter-rater reliability test. Statistical analysis yielded a Cohen's Kappa coefficient of $\kappa = 0.89$ ($p < 0.001$), indicating a high level of agreement and supporting the objectivity and consistency of the evaluation between the assessors.

Data Analysis

Quantitative data obtained from the total prompt quality scores (1–5) and the overall logical reasoning accuracy scores (0–3), assigned based on students' performance across the three integral problems, were tested for normality. Although three reasoning tasks were completed, the correlation analysis used the final overall Logical Reasoning Accuracy score (0–3), representing the holistic evaluation of students' performance across all three tasks. Because the data were measured using ordinal scales, hypothesis testing to examine the significance of the correlation between the two variables was conducted using the non-parametric Spearman rank correlation (rho) test with the assistance of SPSS statistical software. Furthermore, qualitative data obtained from chat histories and open-ended questionnaire responses were analysed using thematic analysis to narratively describe the cognitive processes that mediate the correlational relationship (Braun & Clarke, 2006). The qualitative phase followed the quantitative analysis using an explanatory sequential mixed-methods approach. Participants representing different levels of prompt quality and logical reasoning accuracy (low, moderate, and high) were purposively selected to illustrate and explain the quantitative correlation patterns. The

qualitative findings were therefore used to provide explanatory insights into the statistical results rather than to test causal relationships.

3. Results

Before hypothesis testing, total prompt quality scores (1–5) and overall Mathematical Logical Reasoning Accuracy scores (0–3), assigned based on students' performance across the three integral problems, were calculated for a sample of $n = 85$ students. Given that both variables were measured on ordinal scales and the data were not normally distributed, hypothesis testing was conducted using Spearman's Rank Correlation (ρ). Table 2 presents the statistical results used to examine the direction, strength, and significance of the relationship between the two variables.

Table 2.
Spearman's Rank Correlation between Prompt Quality and Logical Reasoning Accuracy

Variables	Spearman Correlation Coefficient (ρ)	*p-values*
Prompt Quality (X) – Logical Reasoning Accuracy (Y)	0.784**	< 0.001

Note: ** indicates that the correlation is significant at the 0.01 level (2-tailed).

Based on Table 2, the statistical analysis indicates a strong and highly significant positive relationship between the students' prompt engineering skills and their mathematical logical reasoning accuracy after interacting with AI ($\rho = 0.784$, $p < 0.001$), suggesting that higher levels or better quality of the prompts produced by the students (e.g., reaching Level 4: Algorithmic or Level 5: Iterative) are associated with higher scores in mathematical logical reasoning accuracy on the worksheet. From a practical perspective, these quantitative findings are consistent with the initial hypothesis that students who remained at a passive level (Level 1: Naive Copy-Paste) consistently demonstrate reasoning failure (scores of 0 or 1) as a result of blindly accepting AI hallucinations. In contrast, students who are able to construct well-defined mathematical prompts are more capable of verifying output, detecting logical flaws in machine responses, and producing accurate mathematical proofs (score of 3).

To gain a deeper understanding of how student interaction patterns influence their mathematical outcomes, the analysis was extended by mapping the distribution of students using a Crosstabulation Table. Table 3 presents the frequency distribution of the students based on their Prompt Quality level and their mathematical linear reasoning accuracy scores.

Table 3.
Cross-tabulation of prompt quality and accuracy of mathematical logistic reasoning

Prompt Quality (X)	Illogical/Blind Acceptance	Partial Reasoning	Logical but Incomplete	Accurate and Strong
Level 1 (Naive Copy-Paste)	15	5	0	0
Level 2 (Basic Instructional)	5	15	5	0
Level 3 (Contextual)	0	5	10	5
Level 4 (Algorithmic / Chain-of-Thought)	0	0	4	8
Level 5 (Iterative / Debugging)	0	0	1	7

Based on Table 3, the frequency distribution reveals a clear diagonal pattern, confirming a strong positive association between the two variables. The most striking finding of this mapping is the high incidence of logical failure among students with low prompt quality. A total of 20 students were identified as simply performing mechanical question transfer without meaningful transformation (Level 1: Naive Copy-Paste). Within this group, the vast majority (15 students) obtained a score of 0, indicating that they blindly accepted and reproduced AI-generated hallucinations (e.g., ignoring curve intersection points in area problems) without any cognitive intervention. Not a single student at Level 1 or Level 2 achieved the highest logical reasoning score (Score 3).

In contrast, a substantial shift is observed when students begin to apply mathematical problem decomposition in their prompts. Among the 20 students who reached the structured prompt at a higher-level (Level 4 and Level 5), none obtained scores of 0 or 1. Furthermore, 7 of 8 students at Level 5 (Iterative/Debugging), who were capable of challenging and correcting AI logical errors, successfully produced comprehensive and accurate mathematical arguments free from machine bias (Score 3). These findings suggest that higher prompt quality is associated with more productive interactions with AI, whereas lower prompt quality is more frequently associated with inaccurate logical reasoning outcomes.

Explanation of the Correlational Phenomenon (Thematic Analysis)

To understand the cognitive mechanisms underlying the strong correlation between prompt quality and precision of logical reasoning, a thematic analysis was conducted on chat logs and responses to open-ended reflective questionnaires. Following the six-phase framework of Braun and Clarke (2006), the data extraction process yielded three main themes that describe students' behavioural transition from passive acceptance to critical collaboration with AI.

Theme 1: Blind Epistemic Trust and the Illusion of Understanding

This theme emerges dominantly among students with Level 1 and Level 2 prompt quality who achieved a Logical Reasoning Score of 0. Qualitative data indicate that their reasoning failure is not primarily caused by a lack of basic calculus competence, but rather by an excessive level of epistemic trust in the machine. The GenAI outputs, which are presented in a well-structured step-by-step format and delivered in a confident tone, create an "illusion of understanding" among students.

This phenomenon is explicitly reflected in the student responses (P-12, Prompt Quality Level 1, Reasoning Score 0), who were misled by AI hallucinations in the "Area Calculation Trap" problem:

"When I entered the problem, AI provided a very long and neatly structured substitution-based solution. I directly copied it into my worksheet because I assumed that a system this intelligent could not possibly make mistakes in basic integral calculations. I did not recheck whether the curves actually intersect or not."

This excerpt confirms that poor-quality prompting (simple copy-paste behaviour) is associated with a breakdown in logical reasoning, driven by the absence of critical scepticism toward machine-generated output.

Theme 2: Barriers in Translating Mathematical Representations from Visual to Textual Form

For students at Level 2 (Basic Instructional), questionnaire data indicate the presence of cognitive overload when they attempt to translate visual–geometric representations (such as function graphs) into text-based prompts. Although these students are often aware that the AI-generated answers are logically incorrect (e.g. obtaining zero or negative values for tank leakage area problems), they are unable to guide the AI toward correction due to limited prompt literacy.

A student reflection (P-45, Prompt Quality Level 2, Reasoning Score 1) illustrates this difficulty:

"I knew the result for the leaking tank area did not make sense. But I was confused about how to 'tell' the AI in text form. I tried typing 'recalculate, the graph is wrong,' but the AI just repeated the same formula. In the end, I solved it manually as best I could."

This theme explains why the intermediate group recognises computational errors (partial reasoning), but does not achieve maximum scores because their prompts are not sufficiently algorithmic to modify the AI's reasoning state.

Theme 3: Debugging as Cognitive Scaffolding

In contrast, students at Level 4 and Level 5 Prompt Quality (Reasoning Score 3) demonstrate an evolution of prompting behaviour from simple instruction-giving to mathematical argumentation and critique. Chat log analysis reveals that the ability to detect AI hallucinations and formulate corrective (debugging) prompts functions as a form of cognitive scaffolding, which in turn strengthens the final output of logical reasoning of the students in their worksheets.

Excerpt from Digital Chat Log (Student P-78, Prompt Quality Level 5) in Solving Challenge 1 (Area Trap of $\sin(x)$ and $\cos(x)$)

Below is an excerpt from the digital chat log of a student (P-78, Prompt Quality Level 5) while completing Challenge 1 (the "Area Trap" problem involving sin(x) and cos(x)):

1. Prompt Iteration 1
   Student: Calculate the area between the curves $f(x) = \sin(x)$ and $g(x) = \cos(x)$ on the interval [0, π].
2. GenAI Output
   AI: Provides a direct integral computation $\int_0^{\pi} \sin(x) - \cos(x)\, dx = 2$
3. Prompt Iteration 2
   Student (Correction): The answer seems incorrect. The curves $\sin(x)$ and $\cos(x)$ intersect at $x = \frac{\pi}{4}$. On the interval [0, $\frac{\pi}{4}$], $\cos(x)$ is above $\sin(x)$. Please, recompute by splitting the integral into two Recomputecompute by splitting the integral into two parts!

4. GenAI Output
   AI: Apologizes for the mistake, acknowledges the correction, and provides a revised piecewise integral computation: $\left[\int_0^{\pi/4} \sin(x) - \cos(x)\, dx\right] + \left[\int_{\pi/4}^{\pi} \sin(x) - \cos(x)\, dx\right] = 2\sqrt{2}$

The above interaction demonstrates that high-level prompt construction (Level 5) requires equally high-level logical reasoning. To challenge and correct the AI, students must mentally visualise the graph and apply the definite integral concept with precision. This explains why the quantitative correlation observed in the earlier phase is strong and positive.

4. Discussion

The main findings of this study indicate that prompt quality is strongly associated with mathematical logical reasoning in AI-assisted mathematics learning, suggesting that the way students interact with GenAI may be an important factor in conjunction with the capabilities of the AI system itself. These results are consistent with the quantitative finding of a strong positive association between prompt quality and mathematical logical reasoning accuracy. In principle, this aligns with the views of Kasneci et al. (2023) and Lo (2023), who argue that students' quality of thinking does not automatically improve through the implementation of LLMs in education, but rather depends heavily on users' ability to guide the machine's cognitive process. These findings may be interpreted as consistent with the notion of GenAI as a potential cognitive amplifier, whereby higher-quality prompts tend to be associated with higher levels of mathematical reasoning. However, this interpretation should be viewed with caution given the correlational nature of the study. Nevertheless, the observed relationship should not be interpreted as evidence of causality, as other factors such as prior mathematical knowledge, metacognitive competence, and previous experience with AI systems may also contribute to both prompt quality and logical reasoning performance (Constantinou et al., 2023; Fan et al., 2025; Tankelevitch et al., 2024; Xavier et al., 2026). The findings suggest that high-quality prompts may be associated with several key components of mathematical logic. First, they are associated with more explicit strategy selection, as students specify appropriate mathematical procedures, constraints, and solution pathways before engaging with AI. Second, they are associated with richer representation construction, as students

must translate symbolic, graphical, and contextual information into structured textual instructions that can be interpreted accurately by AI systems. Third, they are associated with more rigorous logical verification, as students critically evaluate AI-generated solutions, identify inconsistencies, and compare alternative reasoning pathways before accepting a final answer. These processes were particularly evident in the integral tasks examined in this study, where successful problem solving required students to identify curve intersections, interpret graphical representations, and verify the validity of intermediate and final mathematical conclusions.

Furthermore, student behaviour patterns in this study reveal an epistemological shift from answer consumption to collaborative knowledge production. Students with low prompt literacy are vulnerable to developing blind epistemic trust, namely the tendency to accept AI output as mathematical truth without verification. This phenomenon is consistent with the findings of Hendriks et al. (2015) and Pandey et al. (2026), who found that epistemic trust in digital systems tends to become excessive when information is presented in a systematic and persuasive manner. In mathematics, this condition leads to what is known as the "illusion of understanding," where students believe that they have understood the solution, but in fact commit conceptual logical errors, particularly in integral problems that require visual and analytical interpretation (Amo Filva et al., 2026). In addition, a variety of learning activities and opportunities for interaction with peers and instructors is needed in mathematical learning activities to support learning creativity (Supriadi et al., 2025).

In contrast, students with high-level prompting demonstrate a more dialogical and critical interaction pattern, which this study terms debugging-based reasoning. This process may be interpreted as a form of cognitive scaffolding associated with more sophisticated mathematical reasoning through iterative verification and correction of AI outputs. This finding aligns with previous literature, which emphasises that prompt engineering in STEM is not merely a technical skill, but a new form of mathematical metacognition (Qian, 2025; Samtani, 2026). From a metacognitive perspective, the prompting behaviours observed in this study reflect the core processes of planning, monitoring, and evaluation during mathematical problem-solving. Planning is evident when students formulate structured prompts that specify mathematical goals, relevant information, and solution procedures before engaging with AI. Monitoring occurs when students continuously examine AI-generated responses, identify inconsistencies, and recognise potential logical or computational errors. Evaluation is reflected in students' decisions to accept, review, or reject AI output based on mathematical justification and conceptual correctness. The iterative prompting and debugging practices demonstrated by high-performing students therefore suggest that effective prompt engineering may function as an externalised form of metacognitive regulation, enabling learners to make their reasoning processes explicit while interacting with AI systems. Furthermore, these results, which highlighted the importance of monitoring and control in mathematical problem-solving and creativity (Suherman & Vidákovich, 2024; Supriadi et al., 2024), show that metacognitive processes are now also observable in human–AI interaction (Jamaluddin Z et al., 2025), reflecting stronger beliefs about mathematical ability (Hidayatullah et al., 2026).

Pedagogical Implications

The findings of this study carry substantial paradigm implications for STEM education, particularly in mathematics. In the era of AI disruption, mathematical literacy can no longer be defined solely by the ability to execute computational procedures, as machines can perform such tasks in seconds, supporting the structured integration of pedagogically informed AI competencies (Nabhan & Habók, 2026). Instead, the focus must shift toward textual articulation and critical verification skills. The ability to translate mathematical constraints from visual or geometric representations into precise textual prompts has now become an essential competency. Consequently, it is no longer pedagogically sufficient for educators to simply prohibit the use of AI in classroom settings. Rather, educators are expected to guide students in evolving from passive consumers of answers to active logical collaborators who can effectively direct, constrain, and critically evaluate AI outputs for the design of training strategies (Amo Filva et al., 2026).

5. Limitations and future research

Although this study provides novel information on the relationship between prompt quality and mathematical logical reasoning, several limitations should be acknowledged. First, mathematical tasks were restricted to the domain of Calculus (integral topics), which may limit the generalisability of the findings. The cognitive overload observed in translating mathematical representations into prompts may manifest differently in other more abstract mathematical domains. Second, this study only examined student interactions with a standard text-based AI system. AI systems capable of processing multimodal input, such as graphs or images of functions, were not included in the analysis and may produce different interaction patterns.

Based on these limitations, several directions for future research are recommended. At the institutional level, mathematics curricula in secondary and higher education should explicitly integrate Mathematical Prompt Engineering as a core competency. Students should be systematically trained to engage with intentionally flawed AI outputs in order to stimulate critical mathematical reasoning and epistemic vigilance. For future research, experimental designs are encouraged to evaluate the effectiveness of structured prompting frameworks, such as Chain-of-Thought prompting, in reducing mathematical anxiety and improving reasoning accuracy. Furthermore, future studies should explore the use of multimodal AI platforms to address the visual-to-text translation barrier identified in this study, particularly in geometry and graph-based mathematical tasks. Thus, this study contributes to the literature by positioning prompt engineering as a central cognitive variable rather than merely an additional technical skill, and demonstrates that the quality of interaction determines whether AI becomes a tool for enhancing thinking or, conversely, a source of conceptual errors.

6. Conclusions

This mixed-methods study indicates that the successful use of GenAI in advanced mathematical problem-solving is more closely associated with the quality of users' instructional input (prompt engineering) than with the sophistication of the technology itself. A strong and highly significant positive correlation was found between students'

prompt engineering skills and their mathematical logical reasoning accuracy. The findings indicate that passive behaviour in the form of direct, uncritical question transfer (Level 1: Naive Copy-Paste) was consistently associated with reasoning failure (Score 0). This pattern may reflect blind epistemic trust, in which students experience an "illusion of understanding" due to the seemingly coherent and confident narrative of AI-generated responses despite underlying logical inaccuracies. In contrast, students who applied algorithmic thinking and formulated corrective or adversarial instructions (Level 5: Iterative/Debugging) demonstrated interaction patterns that may function as a form of cognitive scaffolding. Such interactions were associated with more frequent concept verification, problem decomposition, and the development of more rigorous mathematical justifications during AI-assisted problem solving.

Article submission date: May 24, 2026
Approval date: Junee 26, 2026
Publication date: July 1, 2026

Supriadi, N. & Suherman, S. (2026). Integration of generative AI in creative mathematics learning: A correlational analysis between the quality of prompts and the logical reasoning of student teachers in the education degree programme. *Revista de Educación a Distancia, 26* (84). http://dx.doi.org/10.6018/red.716941

Author statement on the use of LLMs

During the preparation of this work, the author(s) used LLM (i.e., ChatGPT and Open Writefull) in order to improve scientific English. After using this tool/service, the author(s) reviewed and edited the content as needed and assume(s) full responsibility for the content of the published article.

Authors' Contributions

Conceptualisation, N.S.; data curation, N.S.; software, N.S.; formal analysis, N.S. and S.S.; investigation, N.S. and S.S.; methodology, N.S., and S.S; visualisation, N.S., and S.S.; writing – original draught, N.S. and S.S.; writing – review & editing, N.S., and S.S.

Funding

This work has not received specific grants from funding agencies in the public, commercial, or non-profit sectors.

References

Adnan, A. H. M., Salim, M. S. A. M., Shah, D. S. M., Yusuf, A. H. M., Salim, M. N. F. M., & Tahir, M. H. M. (2025). Cheating Using AI and Copy-Pasting from LLMs: New Realities in Higher Education. *International Conference on Business and Technology*, 399–410. https://doi.org/10.1007/978-3-032-00250-1_36

Aizikovitsh-Udi, E., & Cheng, D. (2015). Developing critical thinking skills from dispositions to abilities: Mathematics education from early childhood to high school. *Creative Education*, *6*(04), 455. https://doi.org/10.4236/ce.2015.64045

Amo Filva, D., Guàrdia Ortiz, L., Donate Beby, B., Bautista Pérez, G., & Fanni, L. (2026). *Integración de la Inteligencia Artificial y la Alfabetización de Datos en la ESO: Análisis de percepciones y condiciones de adopción*. *26*(83). https://doi.org/10.6018/red.690641

Bieda, K. N. (2010). Enacting proof-related tasks in middle school mathematics: Challenges and opportunities. *Journal for Research in Mathematics Education*, *41*(4), 351–382. https://doi.org/10.5951/jresematheduc.41.4.0351

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Chen, F., Chen, J., & Xu, Y. (2025). The More Anxious, the More Dependent? The Impact of Math Anxiety on AI-Assisted Problem-Solving. *Psychology in the Schools*, *62*(8), 2685–2701. Scopus. https://doi.org/10.1002/pits.23500

Constantinou, A. C., Guo, Z., & Kitson, N. K. (2023). The impact of prior knowledge on causal structure learning. *Knowledge and Information Systems*, *65*(8), 3385–3434. https://doi.org/10.1007/s10115-023-01858-x

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, *56*(2), 489–530. https://doi.org/10.1111/bjet.13544

Hendriks, F., Kienhues, D., & Bromme, R. (2015). Measuring laypeople's trust in experts in a digital age: The Muenster Epistemic Trustworthiness Inventory (METI). *PloS One*, *10*(10), e0139309. https://doi.org/10.1371/journal.pone.0139309

Hidayatullah, A., Setiyawan, R., & Syarifuddin. (2026). Primary education students' beliefs about the nature of mathematics, self-efficacy, and mathematics educators; a descriptive study from Indonesia. *Education 3-13*, *54*(4), 958–973. https://doi.org/10.1080/03004279.2024.2380468

Huang, J., Zhong, Y., & Chen, X. (2025). Adaptive and personalized learning in STEM education using high-performance computing and artificial intelligence. *The Journal of Supercomputing*, *81*(8), 981. https://doi.org/10.1007/s11227-025-07481-7

Jamaluddin Z, W., Supriadi, N., & Suherman, S. (2025). Creative AI in education: The role of technological dependence, motivation, and student participation. *Revista de Estudios e Investigación En Psicología y Educación*, *12*(2), e11997–e11997. https://doi.org/10.17979/reipe.2025.12.2.11997

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., & Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Lithner, J. (2017). Principles for designing mathematical tasks that enhance imitative and creative reasoning. *Zdm*, *49*(6), 937–949. https://doi.org/10.1007/s11858-017-0867-3

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), 410. https://doi.org/10.3390/educsci13040410

Luo, X., Xu, D., Li, Y., & Wan, L. C. (2025). Advancing information search through GenAI: the roles of search type, travel motive and GenAI customization level. *International Journal of Contemporary Hospitality Management*, *37*(5), 1725–1743. https://doi.org/10.1108/IJCHM-06-2024-0941

Mollick, E., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. *arXiv Preprint arXiv:2306.10052*. https://doi.org/10.2139/ssrn.4475995

Nabhan, S., & Habók, A. (2026). Language teachers' AI literacy: A psychometric study based on the ED-AI framework. *Computers and Education: Artificial Intelligence*, *10*, 100583. https://doi.org/10.1016/j.caeai.2026.100583

Nikhil, V. (2025). A Comprehensive Study on AI-Enhanced Personalized Learning in STEM Courses. In *Adopting Artificial Intelligence Tools in Higher Education* (pp. 149–170). CRC Press.

Pandey, C. S., Mishra, P., Pandey, S. R., & Pandey, S. (2026). Epistemic trust in generative AI for higher education scale (ETGAI-HE scale). *AI & SOCIETY*, *41*(2), 1387–1400. https://doi.org/10.1007/s00146-025-02566-6

Polydoros, G., Galitskaya, V., Pergantis, P., Drigas, A., Antoniou, A.-S., & Beazidou, E. (2025). Innovative AI-driven approaches to mitigate math anxiety and enhance resilience among students with persistently low performance in mathematics. *Psychology International*, *7*(2), 46. https://doi.org/10.3390/psycholint7020046

Qian, Y. (2025). Prompt engineering in education: A systematic review of approaches and educational applications. *Journal of Educational Computing Research*, *63*(7–8), 1782–1818. https://doi.org/10.1177/07356331251365189

Samtani, P. (2026). Self-explanation Prompts in STEM: Comparing Human and AI Metacognitive Accuracy. *American Journal of Computer Science and Technology*, *9*(1), 19–29. https://doi.org/10.11648/j.ajcst.20260901.13

Serrano, S. M. (2026). Critical Generative AI Literacy for Social Studies Educators: A Typology of GenAI Errors and Their Impacts on Epistemology. *Journal of Geography*, 1–12. https://doi.org/10.1080/00221341.2026.2621014

Suherman, S., & Vidákovich, T. (2024). Relationship between ethnic identity, attitude, and mathematical creative thinking among secondary school students. *Thinking Skills and Creativity*, *51*, 101448. https://doi.org/10.1016/j.tsc.2023.101448

Supriadi, N., Jamaluddin, W., Suherman, S., & Komarudin, K. (2025). The Role of Blended Learning in Improving Students' Numerical Ability and Learning Creativity. *Revista de Educación a Distancia (RED)*, *25*(81). https://doi.org/10.6018/red.619061

Supriadi, N., Jamaluddin Z, W., & Suherman, S. (2024). The role of learning anxiety and mathematical reasoning as predictor of promoting learning motivation: The mediating role of mathematical problem solving. *Thinking Skills and Creativity*, *52*, 101497. https://doi.org/10.1016/j.tsc.2024.101497

Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The Metacognitive Demands and Opportunities of Generative AI.

Proceedings of the CHI Conference on Human Factors in Computing Systems, 1–24. https://doi.org/10.1145/3613904.3642902

Wang, X. (2025). The Influence of Gen-AI Assisted Learning on Primary School Students' Math Anxiety: An Intervention Study. *Applied Cognitive Psychology*, *39*(4). https://doi.org/10.1002/acp.70088

Wang, X., Liu, Q., Pang, H., Tan, S. C., Lei, J., Wallace, M. P., & Li, L. (2023). What matters in AI-supported learning: A study of human-AI interactions in language learning using cluster analysis and epistemic network analysis. *Computers & Education*, *194*, 104703. https://doi.org/10.1016/j.caeai.2023.100120

Wardat, Y. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Available at SSRN 5653030*. https://doi.org/10.29333/ejmste/13272

Xavier, A., Naeem, S. S., Rizwi, W., & Rabha, H. (2026). Comparing AI-Assisted Problem-Solving Ability With Internet Search Engine and e-Books in Medical Students With Variable Prior Subject Knowledge: Cross-Sectional Study. *JMIR Medical Education*, *12*(1), e81264. https://doi.org/10.2196/81264