Research Square

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

Generative AI as Metacognitive Scaffolding in Undergraduate Mathematics: A Quasi-Experimental Study of Problem-Solving Performance and Engagement

Ahmed Metwally

dalsayed@ksu.edu.sa

King Saud University

Article

Keywords: generative artificial intelligence, mathematical problem solving, metacognitive scaffolding, self-regulated learning, Pólya's model, university calculus, quasi-experimental design

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11115839/v1

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

Additional Declarations: No competing interests reported.

Abstract

This study examines whether generative artificial intelligence (GenAI) can improve university students' mathematical problem-solving when it is designed not as an answer generator but as a metacognitive scaffold. Building on sociocultural scaffolding, self-regulated learning, and Pólya's problem-solving model, the study tested a 15-week structured GenAI-supported intervention in Calculus I. A cluster-assigned quasi-experimental pre-test/post-test design was implemented with 128 undergraduates distributed across four tutorial sections: two sections used a Socratic GenAI protocol (n = 60), and two sections received traditional instruction using the same weekly problem sets (n = 68). Students' solutions were scored with a validated analytic rubric aligned with the four Pólya stages, and outcomes were analyzed using ANCOVA with pre-test adjustment and a multilevel sensitivity model for section-level clustering. Results showed a statistically significant adjusted between-group difference favoring the GenAI-supported group, $F(1, 125) = 28.76$, $p < .001$, partial eta squared = .187, with the largest stage-specific effects for carrying out the plan and understanding the problem. Student perceptions were highly positive, and cognitive engagement showed the strongest association with performance gains. The study contributes to educational psychology by specifying a testable instructional mechanism: structured GenAI-supported instruction may support mathematical problem solving when prompts elicit planning, monitoring, explanation, and reflection while discouraging direct solution-seeking.

Introduction

Generative artificial intelligence (GenAI) has rapidly entered higher education and created a new instructional problem for educational psychology: whether adaptive language-based tools can support genuine learning processes or primarily encourage answer-seeking and superficial performance. Tools such as ChatGPT and Gemini can generate explanations, hints, alternative representations, and feedback in real time, but their educational value depends on how students are guided to interact with them [1,2]. Consequently, the central question is no longer whether students can access GenAI, but under what instructional conditions GenAI can promote durable cognitive and metacognitive development.

Mathematical problem solving is a particularly demanding context for testing this question. Successful problem solving requires more than procedural fluency; it requires interpreting problem conditions, selecting strategies, monitoring intermediate steps, evaluating the plausibility of results, and revising reasoning when necessary [3]. These processes are closely aligned with Pólya's [4] four stages: understanding the problem, devising a plan, carrying out the plan, and looking back. In university calculus, students often struggle not because they lack exposure to worked examples but because they do not consistently regulate their thinking across these stages.

From a sociocultural and constructivist perspective, learning improves when students receive scaffolded support within their zone of proximal development [5]. GenAI has the potential to provide such scaffolding by asking clarifying questions, offering hints, prompting students to justify steps, and encouraging reflection. However, the same technology can also undermine learning if it supplies complete solutions without requiring cognitive effort. This tension makes GenAI a theoretically important case for examining the boundary between productive scaffolding and cognitive outsourcing.

The present study therefore conceptualizes GenAI as a metacognitive scaffold rather than a technological treatment. This distinction is important for educational psychology [6,7]. If GenAI is treated simply as an additional digital tool, its contribution remains limited and context-specific. If, however, GenAI is designed to elicit planning, monitoring, self-explanation, and reflection, it becomes a mechanism for strengthening self-regulated mathematical problem solving [8,9,10]. This mechanism directly addresses concerns in the literature about over-reliance, hallucinated responses, and shallow learning [11,12,13].

Although reviews and exploratory studies have identified promising uses of GenAI in education, much of the evidence remains perception-based, qualitative, or insufficiently connected to specific cognitive processes [14,15,16]. Few studies have tested a structured GenAI intervention at the university level while measuring stage-specific problem-solving outcomes and accounting for classroom-level assignment. This gap limits the field's ability to move from general claims about AI-assisted learning to theory-informed evidence about how GenAI changes mathematical reasoning.

This study addresses that gap by evaluating a 15-week GenAI-supported intervention in Calculus I. The intervention positioned GenAI as a Socratic tutor embedded within a think-pair-GenAI-share routine. Students were trained to request hints, test reasoning, compare strategies, and reflect on errors rather than obtain direct answers. The study contributes in four ways: it offers a theoretically grounded model of GenAI as metacognitive scaffolding; it uses a controlled quasi-experimental design; it analyzes learning gains across Pólya's four stages rather than relying only on total scores; and it examines how students' perceptions of utility, ease of use, and engagement relate to measured learning gains.

Contribution Statement

In response to the need for stronger theoretical contribution, the study is positioned around a specific explanatory mechanism: structured GenAI prompts operate as adaptive metacognitive scaffolds that activate students' planning, monitoring, self-explanation, and reflection during mathematical problem solving. The expected contribution is therefore not that GenAI "improves achievement" in a general sense, but that pedagogically constrained GenAI use can transform students' regulation of the problem-solving process.

Research Objectives:

The general objective of this study is to test whether a structured 15-week GenAI-supported instructional intervention improves university students' mathematical problem-solving by functioning as a metacognitive scaffold rather than as a direct solution provider.

Specific Objectives

1. 1. To estimate the effect of the GenAI-supported intervention on overall mathematical problem-solving performance after controlling for pre-test scores and section-level assignment.
2. 2. To determine whether the intervention differentially improves the four Pólya-based stages of problem solving: understanding the problem, devising a plan, carrying out the plan, and looking back/reflecting.
3. 3. To examine whether students' perceived utility, ease of use, cognitive engagement, and behavioral engagement are associated with improvements in problem-solving performance, with particular attention to the role of cognitive engagement as an indicator of deep rather than superficial GenAI use.

Research Questions:

The study was guided by the following overarching question: Under what conditions does structured GenAI-supported instruction improve university students' mathematical problem-solving in Calculus I?

Specific Research Questions

1. 1. After controlling for pre-test performance and section-level assignment, do students receiving structured GenAI-supported instruction demonstrate higher post-test mathematical problem-solving scores than students receiving traditional instruction?
2. 2. Which stages of Pólya's problem-solving model show the strongest intervention-related gains, and what do these gains suggest about the cognitive processes supported by GenAI scaffolding?
3. 3. How do students' perceptions of the GenAI tool—especially perceived utility and cognitive engagement—relate to their measured gains in mathematical problem-solving performance?

Research Hypotheses

Based on the theoretical framing and research questions, the following hypotheses were formulated:

H1. Students receiving the 15-week GenAI-supported instructional intervention will demonstrate significantly higher adjusted post-test scores on overall mathematical problem-solving than students receiving traditional instruction, after controlling for pre-test scores and section-level assignment.

H2. Students receiving the GenAI-supported intervention will demonstrate significantly higher adjusted post-test scores across the four Pólya-based problem-solving stages than students in the control group.

H3. Students' positive perceptions of the GenAI tool, particularly cognitive engagement, will be positively associated with gains in mathematical problem-solving performance.

Literature Review

Mathematical Problem Solving as a Metacognitive Process

Mathematical problem solving is a central construct in educational psychology because it requires students to coordinate conceptual knowledge, procedural fluency, strategic planning, and metacognitive regulation. Pólya's model remains useful because it separates problem solving into observable phases: understanding the problem, devising a plan, carrying out the plan, and looking back. These phases are not merely procedural steps; they represent different forms of cognitive and metacognitive activity. Understanding requires constructing a representation of the problem, devising a plan requires strategy selection, carrying out the plan requires monitoring and execution, and looking back requires evaluation and transfer.

This stage-based view is important for evaluating GenAI. A tool that increases final scores may not necessarily improve problem solving if students bypass planning or reflection. Therefore, evaluating GenAI through Pólya's stages allows the study to identify whether the intervention supports deeper cognitive regulation or merely facilitates answer production.

Generative AI as Metacognitive Scaffolding

From a sociocultural perspective, scaffolding refers to temporary support that enables learners to perform beyond their current independent level while gradually internalizing effective strategies. GenAI can approximate some scaffolding functions by generating prompts, hints, counterexamples, explanations, and feedback. However, these functions are educationally productive only when they require students to remain cognitively active. In this study, GenAI was intentionally constrained through a Socratic prompting protocol that asked students to explain their reasoning, request hints instead of solutions, verify intermediate steps, and compare strategies.

This design positions GenAI as a metacognitive scaffold. Its expected function is to improve students' self-regulation during problem-solving by externalizing planning questions, supporting error monitoring, and encouraging reflection. This theoretical positioning strengthens the study's contribution because the intervention is linked to a defined learning mechanism rather than to the mere presence of a new technology.

The Pedagogical Condition: Structured Use Versus Unstructured Use

The literature on GenAI in education presents a productive tension. On one hand, GenAI can provide personalized explanations and immediate feedback, which may support conceptual understanding and persistence [1,2]. On the other hand, unstructured use can encourage over-reliance, inaccurate reasoning, and reduced cognitive effort [11,12,13]. The difference between these outcomes is not the technology alone but the instructional design surrounding the technology.

The present study responds to this tension by embedding GenAI within a structured think-pair-GenAI-share routine. Students first attempted problems independently, then discussed strategies with peers, then used GenAI to obtain hints or evaluate reasoning, and finally shared solution paths. This sequence was intended to preserve student agency and cognitive effort while using GenAI to support monitoring and refinement.

Empirical Gap and Rationale for the Current Study

Existing studies have offered valuable insights into students' perceptions of GenAI and its potential for supporting mathematics learning. Nevertheless, several limitations remain. Many studies rely primarily on self-report data, focus on general academic outcomes, or do not separate problem-solving performance into cognitive stages. In addition, fewer studies have tested university-level mathematics interventions using a control group and analytic models that address pre-test differences and classroom-level clustering.

The current study addresses these limitations by combining a quasi-experimental design, a validated analytic rubric, Polya-stage analysis, and a multilevel sensitivity model. Its rationale is that a carefully structured GenAI intervention should produce measurable gains in the cognitive phases most directly supported by scaffolding: understanding, planning, execution monitoring, and reflection.

Conceptual Mechanism Tested in the Study

The proposed mechanism is as follows: structured GenAI prompting encourages students to articulate what is known and unknown, select a solution strategy, monitor each step, and evaluate whether the final answer is reasonable. These metacognitive behaviors should lead to higher performance across Pólya's stages. Student perceptions, particularly cognitive engagement, are expected to be associated with gains because they indicate that students used GenAI as a thinking partner rather than as a shortcut to answers.

Methodology

This study employed a cluster-assigned quasi-experimental pre-test/post-test control-group design to examine whether a structured GenAI-supported instructional intervention was associated with differences in undergraduates' mathematical problem-solving in Calculus I. Assignment occurred at the tutorial-section level rather than at the individual level to reduce contamination and preserve the ecological validity of a real university course. Because the number of clusters was small, section-level clustering was addressed analytically through a multilevel sensitivity model and is acknowledged as a design limitation.

Population and Sample

The target population consisted of undergraduate students enrolled in Calculus I at a large public university in Saudi Arabia. The study involved 128 students distributed across four intact tutorial sections during the second semester of the 2024/2025 academic year. Two sections were assigned to the GenAI-supported condition (n = 60), and two sections were assigned to the traditional-instruction control condition (n = 68).

Baseline equivalence was assessed at both individual and cluster levels. Demographic variables (age, gender, high school GPA, and prior mathematics achievement) did not differ significantly between groups. We calculated and reported the intra-cluster correlation coefficient (ICC) to quantify the degree of clustering in the outcome variable.

Table 1
Comparison of Experimental and Control Groups

Aspect	Experimental Group (GenAI-supported)	Control Group (Traditional Instruction)
Number of Participants (n)	60	68
Number of Tutorial Sections	2	2
Intervention Duration	15 weeks	15 weeks
Instructional Content	Same weekly problem sets	Same weekly problem sets
Delivery Mode	Think-pair-GenAI-share routine + textbooks	Face-to-face lectures + textbooks
Role of GenAI	Socratic tutor (structured prompting)	Not used
Instructor	Same core lecturer for the section after training	Same core lecturer for the section
Baseline Equivalence	Comparable (no significant differences)	Comparable (no significant differences)

Table 2
Participant Flow Through the Study

Stage	Experimental Group	Control Group	Total
Assessed for Eligibility	-	-	128
Excluded (did not meet criteria or declined)	0	0	0
Assigned to study condition	2 sections (n = 60)	2 sections (n = 68)	4 sections (n = 128)
Received Allocated Intervention	60 (100%)	68 (100%)	128 (100%)
Discontinued Intervention	0	0	0
Lost to Follow-up	0	0	0
Analyzed (Post-test)	60	68	128
Excluded from Analysis	0	0	0

Instruments

The study utilized two main instruments for data collection:

1. Mathematical Problem-Solving Test (MPST). The MPST was a researcher-developed assessment consisting of 10 open-ended Calculus I problems aligned with course objectives, including limits, derivatives, and integration. Several items were embedded in applied contexts relevant to students' program of study, but the scoring emphasized general mathematical reasoning rather than domain-specific knowledge. The test was designed to assess performance across the four stages of Pólya's problem-solving model: understanding the problem, devising a plan, carrying out the plan, and looking back/reflecting.

Each problem was scored using a detailed analytic rubric (total score out of 100, with 10 points per problem, and 25% allocated to each stage).

Validity and Reliability:

- Content validity was established through expert review by two experienced mathematics professors.
- Internal consistency reliability was high, with Cronbach's alpha = 0.88.

The MPST was administered as a pre-test (one week before the intervention) and as a post-test (immediately after the 15-week intervention).

2. Perception Survey:

Perception Survey: A 16-item questionnaire using a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree) was developed to measure students' perceptions of the generative AI tool. The survey assessed four main dimensions: perceived utility, perceived ease of use, cognitive engagement, and behavioral engagement.

The items were adapted from established instruments, primarily the Technology Acceptance Model (TAM) and student engagement scales. The survey was piloted for clarity and reliability before the main study. It was administered only to the experimental group (n = 60) at the end of the intervention.

Additional Tools and Materials

Generative AI Tool: The experimental group used an institutionally approved large-language-model interface configured for instructional support. The tool was introduced as a Socratic tutor rather than a solution generator. Students were trained to use prompts such as "Give me a hint without solving the problem," "Ask me a question that helps me identify the next step," and "Check whether my reasoning is valid without giving the final answer."

- Demographic and Background Questionnaire: A short form collected data on age, gender, high school GPA, and prior mathematics achievement to verify group equivalence.

Table 3
Cronbach's Alpha for Dimension Perception Survey

Dimension	N Items	Sources	Alpha Cronbach
Perceived Utility	4	Davis [21]	0.93
Perceived Ease of Use	4	Davis [21]	0.92
Cognitive Engagement	4	Fredricks et al. [20]	0.91
Behavioral Engagement	4	Appleton et al. [22]	0.92

All instruments were validated, and ethical approval was obtained from the university's Institutional Review Board (IRB). Data collection procedures ensured standardization across groups while maintaining the integrity of the cluster-assigned design.

Intervention Procedures

To ensure comparability, the intervention lasted exactly 15 weeks for both groups, during which all students worked on identical weekly problem sets drawn from the course syllabus.

Experimental Group (GenAI-supported Instruction): Participants used GenAI within a structured think-pair-GenAI-share routine. They first attempted each problem independently, then discussed their reasoning with a peer, then used GenAI to obtain hints or check reasoning, and finally shared solution strategies. Weekly sessions were monitored to reinforce the rule that GenAI should support thinking rather than provide complete solutions.

- Control Group (Traditional Instruction): Students received standard face-to-face instruction covering the same content and problem sets, utilizing textbooks, lecture notes, and board work. No access to GenAI tools was provided.
- To reduce differential effects, all procedures were standardized to the extent possible within the cluster design.

Data Analysis

Quantitative data were analyzed using SPSS (Version 28) and R. Descriptive statistics were computed, and assumptions relevant to the planned analyses were examined before inferential testing. Baseline equivalence was assessed using an independent-samples t-test. The primary analysis used ANCOVA with pre-test MPST score as the covariate. Because assignment occurred at the tutorial-section level, a multilevel mixed-effects model was used as a sensitivity analysis to assess whether the direction and magnitude of the group effect were robust to section-level clustering.

For the four Pólya stages, separate ANCOVA models were estimated using the corresponding pre-test stage score as a covariate. Effect sizes were reported using partial η^2, and the multilevel sensitivity models reported the group coefficient and its standard error. Pearson correlations were used to examine associations between post-intervention perception dimensions and MPST gain scores within the experimental group ($n = 60$). A two-sided $\alpha = 0.05$ criterion was used for statistical significance.

Open-ended survey responses were analyzed thematically as complementary qualitative evidence. Because the study included only four tutorial sections, cluster-level variance estimates and inferential results from the multilevel sensitivity models were interpreted cautiously.

Threats to Internal Validity and Mitigation

The following threats were acknowledged and mitigated where possible:

- Novelty effect: Structured training and gradual integration of the tool helped reduce initial excitement bias.
- Hawthorne effect: Both groups were aware of participation; however, the control condition mirrored standard course delivery.
- Teacher effect and differential interaction: Core lectures were unified, and tutorial protocols were standardized.
- Contamination: Cluster-level assignment and explicit instructions against sharing materials minimized cross-group influence.

Ethical Considerations

The study was reviewed and approved by the Institutional Review Board at King Saud University (KSU-IRB; Approval No. 363–2025). All methods involving human participants were performed in accordance with the relevant institutional and ethical guidelines and regulations, including the principles of the Declaration of Helsinki. Participation was voluntary, and written informed consent was obtained from all participants before data collection. Participants were informed about the purpose and procedures of the study, their right to withdraw without academic penalty, and the measures taken to protect privacy and confidentiality. Participant data were handled confidentially and stored securely in accordance with applicable institutional requirements and data-protection regulations.

Results

This section presents the findings of the data analysis in relation to the research objectives and hypotheses. The primary analyses used ANCOVA, with a multilevel mixed-effects model treated as a sensitivity analysis because assignment occurred at the tutorial-section level.

1. Initial Equivalence of Groups

An independent samples t-test was performed on the pre-test scores of the Mathematical Problem-Solving Test (MPST) to verify baseline equivalence between groups. As shown in Table 4, there was no statistically significant difference in pre-test mean scores between the experimental group (M = 65.8, SD = 8.2) and the control group (M = 64.5, SD = 8.9), t(126) = 0.86, p = .394. The observed baseline difference was not statistically significant, indicating no evidence of a pre-test difference between groups.

Table 4
Descriptive Statistics and Independent Samples t-test for Pretest Scores

Group	N	Mean	SD	t-value	df	p-value
Experimental	60	65.8	8.2	—	—	—
Control	68	64.5	8.9			

Note. The reported t and p values were recalculated from the displayed rounded group means, standard deviations, and sample sizes using the independent-samples t-test with equal-variance pooling; the exact values should be confirmed against the original participant-level dataset before publication.

2. Effect of the GenAI-Supported Intervention on Mathematical Problem-Solving Skills

To test the primary hypothesis (H1), an analysis of covariance (ANCOVA) was conducted on the post-test MPST scores, using pre-test scores as a covariate. Because assignment occurred at the tutorial-section level, a multilevel mixed-effects model was also used as a sensitivity analysis to account for students being grouped in tutorial sections and to estimate the intra-cluster correlation (ICC = 0.12).

Table 5 **presents the descriptive statistics for the adjusted post-test means (after controlling for pre-test scores). The experimental group demonstrated substantially higher adjusted mean scores (M_adj = 84.7) compared to the control group (M_adj = 69.1).**

Table 5
Adjusted Post-Test Means and Standard Errors for Mathematical Problem-Solving Skills

Group	N	Adjusted Mean	SE	95% CI Lower	95% CI Upper
Experimental	60	84.7	1.12	82.48	86.92
Control	68	69.1	1.05	67.02	71.18

The ANCOVA results (Table 6) showed a statistically significant adjusted between-group difference in post-test MPST scores, $F(1, 125) = 28.76$, $p < .001$, partial $\eta^2 = .187$. The multilevel sensitivity analysis showed a group coefficient in the same direction ($\beta = 15.6$, $SE = 2.91$, $p < .001$). The partial η^2 indicates a large standardized effect under commonly used benchmarks; however, this estimate should be interpreted in the context of the quasi-experimental design and the small number of tutorial sections.

Table 6
ANCOVA Results for Post-Test Scores (Controlling for Pre-Test)

Source	SS	df	MS	F	p-value	Partial η^2
Pre-test (Covariate)	1856.4	1	1856.4	12.85	< 0.001	0.093
Group (Intervention)	4152.7	1	4152.7	28.76	< 0.001	0.187
Error	18045.2	125	144.36	-	-	-

These results support H1: after adjustment for baseline performance, the GenAI-supported condition had a higher adjusted post-test MPST score than the control condition.

3. Differences Across Pólya's Problem-Solving Stages

To address the second research objective and H2 in greater depth, a detailed stage-specific analysis was conducted on the four Pólya-based stages of mathematical problem-solving. Each stage was scored independently (out of 25 points) using the validated analytic rubric. For each stage, we performed separate ANCOVA models (with pre-test stage scores as covariates) and multilevel mixed-effects models to account for clustering.

Table 7 **shows larger adjusted post-test scores for the experimental group in all four stages. The largest standardized effects were observed for "Carrying Out the Plan" (partial $\eta^2 = .231$) and "Understanding the Problem" (partial $\eta^2 = .217$). The multilevel models yielded effects in the same direction; however, because only four tutorial sections were available, the cluster-level estimates should be interpreted cautiously.**

Table 7
Detailed Statistics for Pre- and Post-Intervention Scores Across Pólya's Four Stages (Adjusted Means)

Stage	Group	Pre-test Mean (SD)	Post-test Adjusted Mean (SE)	Difference (adjusted post-test − pre-test mean)	F-value	df	p-value	Partial η^2	Multilevel β (SE)
Understanding the Problem	Experimental	16.2 (3.1)	22.8 (0.68)	+ 6.6	34.72	1, 125	< 0.001	0.217	6.71 (1.14)
Understanding the Problem	Control	15.9 (3.4)	17.5 (0.64)	+ 1.6					
Devising a Plan	Experimental	15.8 (3.0)	21.9 (0.71)	+ 6.1	29.85	1, 125	< 0.001	0.192	6.05 (1.11)
Devising a Plan	Control	15.5 (3.2)	17.2 (0.67)	+ 1.7	-	-	-	-	-
Carrying Out the Plan	Experimental	17.1 (2.8)	23.4 (0.65)	+ 6.3	37.41	1, 125	< 0.001	0.231	6.45 (1.05)
Carrying Out the Plan	Control	16.8 (3.0)	18.9 (0.61)	+ 2.1	-	-	-	-	-
Looking Back / Reflecting	Experimental	16.7 (3.3)	22.1 (0.74)	+ 5.4	26.14	1, 125	< 0.001	0.173	5.38 (1.05)
Looking Back / Reflecting	Control	16.3 (3.5)	17.8 (0.69)	+ 1.5	-	-	-	-	-

Table 7 **presents stage-specific adjusted means for the four Pólya stages. The experimental group showed larger adjusted gains than the control group in each stage, with the strongest effects in carrying out the plan and understanding the problem. This pattern suggests that the intervention supported both procedural execution and the construction of an initial problem representation.**

The multilevel sensitivity analyses showed group effects in the same direction for all four stages (all $p < .001$). These results are consistent with H2, while the small number of clusters warrants cautious interpretation of cluster-level variance estimates.

4. Students' Perceptions of the Generative AI Tool

The perception survey administered to the experimental group (n = 60) after the 15-week intervention showed generally positive ratings of the GenAI tool. As shown in Table 8, mean scores across all dimensions exceeded 4.2 on the 5-point Likert scale.

Table 8
Descriptive Statistics of Student Perceptions of the GenAI Tool (N = 60)

No.	Survey Item	Mean	SD	Agree/Strongly Agree (%)
1	The GenAI tool helped me understand difficult calculus concepts more clearly.	4.55	0.57	90%
2	Using the GenAI tool improved my ability to solve mathematical problems effectively.	4.48	0.61	87%
3	The GenAI tool was a valuable resource for my learning in this course.	4.50	0.55	88%
4	The support provided by the GenAI tool was relevant to my learning needs.	4.47	0.59	85%
5	It was easy for me to interact with the GenAI tool using natural language prompts.	4.38	0.68	82%
6	The GenAI tool was simple to use during problem-solving activities.	4.42	0.62	85%
7	I could quickly get the type of assistance I needed from the GenAI tool.	4.35	0.70	80%
8	Learning to use the GenAI tool effectively did not require much effort.	4.40	0.65	83%
9	The GenAI tool encouraged me to think more deeply about the mathematical problems.	4.53	0.54	88%
10	Using the GenAI tool made me reflect on my problem-solving strategies.	4.45	0.58	85%
11	Interacting with the GenAI tool stimulated my critical thinking skills.	4.48	0.60	87%
12	The GenAI tool helped me make connections between different mathematical concepts.	4.42	0.63	83%
13	I spent more time and effort on difficult problems when using the GenAI tool.	4.33	0.72	78%
14	I felt motivated and actively involved while using the GenAI tool.	4.50	0.56	88%
15	The GenAI tool encouraged me to persist when I encountered challenges in solving problems.	4.38	0.67	82%
16	The GenAI tool encouraged me to try different approaches to solve problems.	4.35	0.69	80%
	Total	4.43	0.42	85%

Students in the experimental group reported highly positive perceptions toward the structured GenAI tool across all four measured dimensions. The overall mean score for the Perception Survey was 4.43 (SD = 0.42), with 85% of respondents selecting "Agree" or "Strongly Agree" on average. This indicates generally favorable perceptions of the structured GenAI activities.

The Perceived Utility dimension received the highest ratings (M = 4.50), with 85–90% of students agreeing that the GenAI tool helped them understand difficult calculus concepts and improved their problem-solving. These findings are notably higher than those reported in previous studies that used unstructured GenAI. For example, Lagon (2025) and

Liu et al. (2025) found moderate-to-high perceived usefulness but with greater variability and concerns about accuracy and over-reliance. The consistently high scores are compatible with favorable perceptions of the structured scaffolding approach, although the perception data alone do not establish that the protocol caused those ratings.

In terms of perceived ease of use, the mean score reached 4.39, with 80–85% of students finding the tool easy to interact with through natural language prompts. This result aligns with Venkatesh et al.'s (2003) Technology Acceptance Model, which posits that perceived ease of use is a strong predictor of technology adoption. The relatively high scores here indicate that the custom-configured interface and prior training effectively reduced the cognitive load associated with using GenAI, addressing one of the common barriers identified in earlier research [11].

Cognitive engagement was also rated highly (M = 4.47), with 83–88% of students reporting that the GenAI tool encouraged deeper thinking, reflection on strategies, and connections among mathematical ideas. This finding is important because it suggests that the intervention promoted active reasoning rather than passive answer-seeking.

Behavioral Engagement recorded a slightly lower but still positive mean (M = 4.39), with 78–82% of students indicating increased persistence and motivation when using the tool. While this dimension scored the lowest among the four, the results remain substantially higher than those typically reported in studies using unstructured GenAI [17,18]. This pattern is consistent with favorable self-reported engagement during challenging problem-solving activities.

Overall, the pattern of results—high perceived utility and cognitive engagement coupled with ease of use—provides evidence of favorable student perceptions of the structured GenAI activities. These perception data complement the performance findings and show positive associations between students' subjective experiences and observed learning gains.

Thematic analysis of open-ended responses further indicated that students perceived the GenAI tool primarily as an effective scaffolding mechanism that supported metacognitive processes, particularly in the "planning" and "reflecting" stages of problem-solving. Several students noted that structured prompting helped them "think step-by-step" rather than relying on direct answers.

These perceptual findings support H3, showing strong positive correlations between perceived utility/engagement and improvement in problem-solving performance (r ranging from 0.48 to 0.67, all $p < .01$).

5. Students' Perceptions of the GenAI Tool and Their Relationship with Performance

In addition to the descriptive statistics, Pearson correlation coefficients were calculated to examine the relationship between students' perceptions of the GenAI tool (as measured by the Perception Survey) and their actual performance gains on the Mathematical Problem-Solving Test (MPST) from pre-test to post-test.

Table 9
Pearson Correlations Between Perception Dimensions and MPST Performance Gains (N = 60)

Perception Dimension	Correlation with Overall MPST Gain (r)	p-value	Interpretation
Perceived Utility	0.62	< .001	Strong positive
Perceived Ease of Use	0.48	< .001	Moderate positive
Cognitive Engagement	0.67	< .001	Strong positive
Behavioral Engagement	0.55	< .001	Moderate to strong positive
Overall Perception Score	0.71	< .001	Very strong positive

Note. All correlations are statistically significant at p < .001. MPST Gain = Post-test score minus pre-test score.

The results revealed statistically significant positive correlations between all perception dimensions and improved mathematical problem-solving skills. The strongest relationships were found with cognitive engagement (r = 0.67) and the overall perception score (r = 0.71). This indicates that higher perceived cognitive engagement and utility were associated with larger observed gains in problem-solving performance.

These findings are consistent with the Technology Acceptance Model [19] and student engagement theory [20], which posit that positive perceptions and high levels of engagement are important predictors of learning outcomes. These correlations are consistent with an association between positive perceptions of structured GenAI use and observed performance gains; because the perception data were collected after the intervention, they do not establish that perceptions caused the gains.

The pattern of correlations is consistent with the interpretation that performance gains were linked to cognitively engaged use of GenAI. The strong association between cognitive engagement and MPST gains suggests that the structured prompting protocol supported deep processing and metacognitive activity rather than simple reliance on generated answers.

Discussion

Interpreting the Main Effect: GenAI as Designed Scaffolding

The findings indicate that students in the structured GenAI-supported condition had higher post-test problem-solving scores when the tool was embedded in a pedagogical design that preserved student reasoning. The experimental group outperformed the control group after controlling for pre-test scores, and the multilevel sensitivity analysis produced a group effect in the same direction after accounting for section-level clustering. The important interpretation is not that GenAI alone improves achievement, but that GenAI can be educationally powerful when it is constrained to function as a scaffold for planning, monitoring, and reflection.

This interpretation directly addresses the concern that GenAI may produce superficial learning. In the present intervention, students were trained to request hints, verify reasoning, and compare strategies rather than obtain direct solutions. Thus, the observed difference is consistent with the possibility that structured interaction patterns helped keep students cognitively responsible for the solution process.

Stage-Specific Effects Across Pólya's Model

The stage-specific analysis strengthens the theoretical contribution of the study. The largest effects appeared in carrying out the plan and understanding the problem. The effect on carrying out the plan suggests that GenAI-

supported prompts helped students monitor procedural steps, identify errors, and maintain progress through multi-step calculus solutions. The effect on understanding the problem suggests that GenAI also supported the construction of problem representations, including clarification of givens, goals, constraints, and relationships among quantities.

The gains in devising a plan and looking back were also statistically significant, although somewhat smaller. This pattern is compatible with the possibility that planning and reflection may require sustained practice to develop fully; however, the present design does not establish a developmental sequence across the four stages.

Student Perceptions and Cognitive Engagement

Students' positive perceptions of the GenAI tool complement the performance results, but the most important perception finding is the strong association between cognitive engagement and MPST gains. This pattern is consistent with the interpretation that students who reported greater cognitive engagement also tended to show larger performance gains. In this respect, the findings extend technology-acceptance perspectives by linking perceived usefulness and engagement to an educational psychology mechanism: self-regulated problem solving.

Theoretical Contributions

The study contributes to educational psychology in three ways. First, it frames GenAI as a potential metacognitive scaffolding mechanism rather than as a generic educational technology. Second, it specifies learning processes that may be supported by structured prompts, including problem representation, planning, execution monitoring, and evaluation. Third, it provides a testable instructional configuration that is designed to reduce cognitive outsourcing and promote self-explanation.

This positioning responds directly to the need for a stronger and more lasting contribution. The study does not simply report that students liked GenAI or scored higher after using it. Rather, it proposes and tests a mechanism that can be examined in future educational psychology research across domains, age groups, and mathematical topics.

Practical and Methodological Implications

Practically, the findings suggest that universities should not introduce GenAI as an unrestricted assistance tool in mathematics courses. Instead, instructors should provide prompt protocols, model appropriate use, prohibit direct solution-seeking during learning activities, and integrate GenAI within peer discussion and reflection routines. Faculty development should therefore focus on designing prompts and tasks that require students to explain, justify, and evaluate reasoning.

Methodologically, the study illustrates the value of moving beyond overall achievement scores. By using a Pólya-aligned analytic rubric, the analysis identified where the intervention had the strongest impact. This approach can help future studies distinguish between performance differences associated with answer access and those associated with changes in problem-solving regulation.

Limitations and Future Research

Several limitations should be considered. First, the study was conducted in a single institutional context and one undergraduate Calculus I course, which limits generalizability to other institutions, disciplines, and student populations. Second, only four tutorial sections were included, with two sections per condition; consequently, the multilevel analysis should be treated as a sensitivity analysis rather than definitive evidence about section-level variance. Third, the intervention combined GenAI with a structured think-pair-GenAI-share routine, so the design cannot completely isolate the contribution of GenAI from the broader pedagogical package. Fourth, outcomes were measured immediately after the 15-week intervention, so the study cannot establish long-term retention or transfer to unfamiliar mathematical

problems or other STEM domains. Fifth, perception and engagement measures were self-reported and may be affected by response bias. Finally, the intervention used a specific structured prompting protocol and an institutionally approved GenAI environment, so the findings should not be generalized to unrestricted or differently configured GenAI use. Future research should replicate the study across multiple institutions and larger numbers of clusters, include delayed post-tests and transfer measures, compare alternative prompting protocols, and incorporate behavioral or interaction-log measures of GenAI use.

Conclusion

This study examined whether GenAI can support university students' mathematical problem-solving when it is deliberately designed as a metacognitive scaffold. The results showed higher adjusted post-test scores for students in the GenAI-supported condition than for students in the control condition, with statistically significant between-group differences across all four Pólya stages. The largest standardized effects were observed for understanding the problem and carrying out the plan. These findings are consistent with the possibility that structured GenAI-supported activities can support representation and execution processes in calculus problem solving, although the quasi-experimental design does not establish a definitive causal effect attributable to GenAI alone.

The central conclusion is that the educational value of GenAI depends on instructional design. Unrestricted use may encourage over-reliance, but structured use can prompt students to plan, monitor, explain, and reflect. By positioning GenAI as a Socratic and metacognitive scaffold, this study offers a theoretically grounded model for responsible AI integration in mathematics education.

For educational psychology, the study provides a mechanism that can be tested and refined in future research: structured GenAI use may be most educationally useful when it supports students' regulation of the problem-solving process rather than replacing that process. Future work should examine retention, transfer, scalability, and alternative instructional configurations across diverse educational settings.

Abbreviations

ANCOVA
analysis of covariance
GenAI
generative artificial intelligence
ICC
intra-cluster correlation coefficient
MPST
Mathematical Problem-Solving Test
TAM
Technology Acceptance Model.

Declarations

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available because they contain participant-level educational performance data and are subject to privacy and ethical restrictions. De-identified data may be made available from the corresponding author upon reasonable request, subject to applicable ethical,

institutional, and data-protection requirements. The data can also be made available to the journal editors and referees for the purpose of evaluating the manuscript, subject to the same confidentiality and ethical requirements.

Competing interests

The author declares that there are no competing interests.

Acknowledgements

This work was supported by the Ongoing Research Funding program, (ORF-RICSP-2026-9), king Saud University, Riyadh, Saudi Arabia. The funder had no role in the study design, data collection, data analysis, interpretation of results, decision to publish, or preparation of the manuscript.

Ethics approval and consent to participate

The study was reviewed and approved by the Institutional Review Board at King Saud University (KSU-IRB; Approval No. 363-2025). All methods involving human participants were performed in accordance with the relevant institutional and ethical guidelines and regulations, including the principles of the Declaration of Helsinki. Participation was voluntary, and informed consent was obtained from all participants before data collection. Participants were informed about the purpose and procedures of the study, the voluntary nature of participation, their right to withdraw at any time without academic penalty, and the measures taken to protect their privacy and confidentiality. Participant data were handled confidentially and stored securely in accordance with applicable institutional requirements and data-protection regulations.

Prior to participation, all participants were informed about the purpose of the study, the nature of their involvement, the voluntary nature of participation, their right to withdraw from the study at any time without penalty, and the measures taken to protect confidentiality and anonymity. Informed consent was obtained from all participants before data collection.

Consent for publication

Not applicable.

Use of generative AI tools in manuscript preparation

The study reports an educational intervention involving GenAI. No generative AI tool is listed as an author. Any AI-assisted language or formatting support during manuscript preparation was reviewed, verified, and approved by the author, who takes full responsibility for the final content.

References

1. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., and Kasneci, G.. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274
2. Lo, C. K.. What is the impact of ChatGPT on education? A rapid review of the literature. Education Sciences, 13(4), Article 410. https://doi.org/10.3390/educsci13040410

3. Artzt, A., and Armour-Thomas, E.. Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. Cognition and Instruction, 9(2), 137–175. https://doi.org/10.1207/s1532690xci0902_3
4. Pólya, G.. How to solve it: A new aspect of mathematical method (2nd ed.). Princeton University Press.
5. Vygotsky, L. S.. Mind in society: The development of higher psychological processes. Harvard University Press.
6. Abolkasim, E., and M. Hasan.. Integrating ChatGPT in Education and Learning: A Case Study on Libyan Universities. Journal of Pure and Applied Sciences 23, no. 2: 19–24. https://doi.org/10.51984/jopas.v23i2.3082.
7. Ahmadi Safa, M., and F. Motaghi.. Cognitive vs. Metacognitive Scaffolding Strategies and EFL Learners' Listening Comprehension Development. Language Teaching Research 28, no. 3: 987–1010. https://doi.org/10.1177/13621688211021821.
8. Song, Y., Huang, L., Zheng, L. et al.. Interactions with generative AI chatbots: unveiling dialogic dynamics, students' perceptions, and practical competencies in creative problem-solving. Int J Educ Technol High Educ 22, 12. https://doi.org/10.1186/s41239-025-00508-2
9. Strzelecki, A.. Students' acceptance of ChatGPT in higher education: An extended technology acceptance model. Interactive Learning Environments. Advance online publication. https://doi.org/10.1080/10494820.2024.2309560
10. Sullivan, M., Kelly, A., and McLaughlan, P. ChatGPT in higher education: Considerations for academic integrity and student learning. Journal of Applied Learning and Teaching, 6(1), 31–40. https://doi.org/10.37074/jalt.2023.6.1.17
11. Kuhail, M. A., Alturki, N., Alramlawi, S., and Alhejori, K. Interacting with educational chatbots: A systematic review. Education and Information Technologies, 28, 973–1018. https://doi.org/10.1007/s10639-022-11177-3
12. Schorcht, S., Buchholtz, N., and Baumanns, L. Prompt the problem: Investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. Frontiers in Education, 9, Article 1386075. https://doi.org/10.3389/feduc.2024.1386075
13. Stolz, S. A., Winterburn, A. L., and Palmer, E. Is learning with ChatGPT really learning? Educational Philosophy and Theory, 56(12), 1253–1264. https://doi.org/10.1080/00131857.2024.2376641
14. AlHatmi, B. S. K. Exploring generative AI's role as a learning supplement tool for higher education students in mathematics: A focus on solving exams and assignments (Order No. 31972837) [Doctoral dissertation, ProQuest Dissertations and Theses Global].
15. Crompton, H., and Burke, D. Artificial intelligence in higher education: The state of the field. International Journal of Educational Technology in Higher Education, 20, Article 22. https://doi.org/10.1186/s41239-023-00392-8
16. Lin, W., and Jiang, P. Factors influencing college students' generative artificial intelligence usage behavior in mathematics learning: A case from China. Behavioral Sciences, 15(3), Article 295. https://doi.org/10.3390/bs15030295
17. Lagon, H. M. Integrating ChatGPT in math education: A narrative inquiry into mathematical investigation, problem solving, and mathematical modeling. Asian Journal of Education and Social Studies.
18. Liu, J., Sun, D., Sun, J., Wang, J., and Yu, P. L. H. Designing a generative AI-enabled learning environment for mathematics word problem solving in primary schools: Learning performance, attitudes, and interaction. Computers and Education: Artificial Intelligence, 9, Article 100438. https://doi.org/10.1016/j.caeai.2025.100438
19. Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3), 425–478. https://doi.org/10.2307/30036540
20. Fredricks, J. A., Blumenfeld, P. C., and Paris, A. H. School engagement: Potential of the concept, state of the evidence. Review of Educational Research, 74(1), 59–109. https://doi.org/10.3102/00346543074001059
21. Davis, F. D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340. https://doi.org/10.2307/249008

22. Appleton, J. J., Christenson, S. L., and Furlong, M. J. Student engagement with school: Critical conceptual and methodological issues of the construct. Psychology in the Schools, 45(5), 369–386. https://doi.org/10.1002/pits.20303