INTERNATIONAL JOURNAL OF ENVIRONMENT, ENGINEERING & EDUCATION

ISSN 2656-8039 (ONLINE)
VOLUME 8 • ISSUE 1 • MARCH 2026

Received: 1 September 2025 | Accepted: 8 February 2026 | Published: 1 March 2026

Research Article

A Polya-Aligned Prompting Protocol for ChatGPT Scaffolding: Evidence from Eighth-Grade Systems-of-Linear-Equations Problem Solving

Marwati Abd. Malik [a,*] • Ranak Lince [b] • Husnaeni [b]

[a] Department of Mathematics Education, Universitas Muhammadiyah Parepare, Parepare, South Sulawesi, Indonesia | [b] Department of Mathematics Education, Universitas Terbuka, Makassar, South Sulawesi, Indonesia

ABSTRACT

Generative artificial intelligence offers new opportunities to scaffold students' mathematical reasoning, yet rigorous evidence of its impact on secondary students' problem-solving remains limited. This study examined whether ChatGPT-driven adaptive learning improves eighth-grade students' problem-solving performance on systems of linear equations in two variables (SPLDV) compared with conventional instruction. A pre–post-test control-group design was implemented with 47 eighth-grade students in Parepare, Indonesia (experimental $n = 24$; control $n = 23$) during the 2024/2025 academic year. The experimental group used ChatGPT as an adaptive tutor aligned with Polya's stages (understand, plan, execute, look back) through guided prompts, hints, and feedback. In contrast, the control group received a lecture and practice. Students completed a six-item contextual SPLDV test scored with a Polya-based rubric. Between-group differences were tested on post-test scores and normalized gains after verifying normality and homogeneity assumptions. The experimental group achieved higher post-test performance ($M = 68.33$) than the control group ($M = 59.57$), with a significant difference ($p = 0.019$; $\eta^2 = 0.117$). Learning gains were also larger in the experimental group (mean N-gain = 0.34, medium) than in the control group (0.21, low; $p = 0.001$; $\eta^2 = 0.372$). Indicator-level patterns suggested the greatest improvements in understanding the problem and carrying out the plan, whereas devising a plan remained the most challenging stage in both groups. These findings indicate that ChatGPT-based adaptive scaffolding can enhance students' mathematical problem-solving on SPLDV, but explicit teacher-guided routines are needed to strengthen strategic planning and the critical evaluation of AI outputs.

KEYWORDS AI-supported scaffolding • ChatGPT • large language model tutoring • normalized learning gain • inclusion and education • quasi-experimental design

ARTICLE CITATION
M. A. Malik, R. Lince, Husnaeni, "A Polya-Aligned Prompting Protocol for ChatGPT Scaffolding: Evidence from Eighth-Grade Systems-of-Linear-Equations Problem Solving," International Journal of Environment, Engineering and Education, Vol. 8, No. 1, pp. 35-50, 2026. https://doi.org/10.55151/ijeedu.v8i1.331

*CORRESPONDENCE
Marwati Abd. Malik marwatiabdmalik@gmail.com Department of Mathematics Education, Universitas Muhammadiyah Parepare, Jalan Jenderal Ahmad Yani Km. 6, Bukit Harapan, Parepare, South Sulawesi 91112, Indonesia. https://orcid.org/0009-0001-8378-1185

1. INTRODUCTION

Mathematics education is essential for cultivating students' logical, critical, and creative thinking; supporting real-world problem solving; and advancing science and technology [1], [2]. In the 21st century, problem solving is widely recognized as a core competency for addressing complex global challenges [3], [4]. Accordingly, mathematics instruction should move beyond conceptual and procedural mastery to promote higher-order thinking, including analysis, evaluation, and creativity [5], [6].

However, evidence indicates that students' problem-solving ability remains underdeveloped. Many students struggle to interpret problem statements, devise appropriate strategies, and verify their solutions [7]–[9]. An overreliance on rote memorization limits success on non-routine tasks [10], [11], and classroom practices often emphasize final answers rather than the reasoning processes that lead to them [12], [13]. These limitations underscore the need for instructional approaches that cultivate adaptive problem-solving skills.

Artificial intelligence (AI) is among the most transformative contemporary technologies, reshaping diverse domains such as health [14], urban development [15], economics [16], and geospatial analysis [17]. AI systems are designed to emulate aspects of human intelligence, including recognition, reasoning, learning, and action [18], [19]. In recent years, AI has advanced rapidly and has become increasingly integrated into education [20]. Within this domain, AI is recognized for its potential to personalize learning, enhance student engagement, and support competency development [21]. More broadly, advances in information and communication technology (ICT), particularly AI, offer promising opportunities to address persistent challenges in teaching and learning by enabling more adaptive, efficient, and inclusive learning environments [22], [23].

A major recent development in AI for education is the emergence of large language models (LLMs), such as ChatGPT [24]. Built on transformer-based neural network architectures [25], ChatGPT can clarify mathematical concepts, generate practice problems, provide step-by-step solutions, and suggest multiple solution strategies [26], [27]. Importantly, its potential extends beyond answer generation by offering digital scaffolding aligned with Polya's [12] four stages of problem solving: understanding the problem, devising a plan, carrying out the plan, and evaluating the result. Through interactive dialogue, ChatGPT can support students in exploring alternative approaches, asking follow-up questions, and strengthening mathematical reasoning [28]–[30]. Thus, ChatGPT may function not only as a technological tool but also as a facilitator of deeper mathematical thinking and the development of higher-order problem-solving competencies [31].

Recent studies have begun to examine the role of ChatGPT in mathematics education. Noster, Gerber, and Siller [26] investigated how pre-service teachers use ChatGPT to solve mathematical problems and found that response quality depends strongly on prompting techniques. Although ChatGPT can provide step-by-step explanations, pre-service teachers may accept incorrect solutions when they are presented persuasively. These findings highlight the importance of mathematical fidelity that is, the capacity to critically evaluate the correctness of AI-generated mathematical outputs.

Similarly, Busuttil and Calleja examined teachers' beliefs and practices following a brief professional development (PD) program on ChatGPT [28]. Teachers with discovery-oriented or connectionist pedagogical philosophies were more inclined to incorporate ChatGPT into instruction. Nevertheless, concerns persisted regarding output accuracy, limitations in visual representation, and the risk of student dependency. These results suggest the need for systematic PD to support teachers in integrating ChatGPT within the Technological Pedagogical Content Knowledge (TPACK) framework.

In addition, a scoping review by Pepin, Buchholtz, and Salinas-Hernández [29] synthesized reported applications of ChatGPT in mathematics education, including problem generation, immediate feedback, and personalized learning support. However, the review also identified persistent challenges, including accuracy, ethical concerns, limited visualization, and the need for prompt-engineering skills. Moreover, empirical evidence regarding ChatGPT's effects on student achievement remains limited and inconsistent. Related work echoes these concerns: Kasneci et al. [27] discussed both opportunities and risks of LLMs in education, and Becker et al. [32] emphasized the importance of cultivating critical engagement with AI outputs rather than passive acceptance.

Taken together, these findings indicate that although ChatGPT shows promise for supporting problem-solving oriented mathematics learning, its implementation is not straightforward. First, outputs must be critically verified by both students and teachers to prevent uncritical acceptance of erroneous solutions. Second, effective use of ChatGPT requires skill in designing prompts that elicit relevant and accurate responses. Third, teachers need sufficient technological and pedagogical knowledge to integrate ChatGPT in ways that support rather than undermine the goals of mathematics education [33], [34].

Despite a growing literature base, several gaps remain. First, few controlled experimental studies have evaluated the impact of ChatGPT-based adaptive learning on students' problem-solving performance; much of the existing research is exploratory or based on small-scale case studies [27]. Second, evidence on the effectiveness of PD programs that prepare teachers to use ChatGPT in mathematics classrooms remains limited [28]. Third, the integration of multimodality combining text, symbolic notation, and visual representations in ChatGPT-assisted learning, particularly for geometry and spatial reasoning, remains underexplored [32].

Evaluating ChatGPT as an AI-based adaptive tutor for mathematical problem solving can yield both theoretical and practical contributions. Theoretically, such work extends scholarship on AI integration in mathematics education, particularly within the Knowledge for Teaching Mathematics with Technology (KTMT) framework [35]. Practically, findings may inform teachers, policymakers, and educational technology developers in designing innovative, effective, and ethical approaches to mathematics instruction.

Despite increasing interest in generative AI for mathematics learning, controlled evidence remains limited regarding process-level outcomes aligned with Polya's stages and replicable intervention protocols specifying prompting rules, teacher mediation, and verification routines. This study addresses these needs by (1) testing a Polya-aligned prompting protocol for ChatGPT scaffolding in an eighth-grade unit on systems of linear equations, (2) estimating post-test and learning-gain effects relative to conventional instruction while accounting for baseline performance, and (3) identifying which Polya stages exhibit the largest improvements under structured AI scaffolding.

Accordingly, the present study examines the effect of ChatGPT-driven adaptive scaffolding on eighth-grade students' problem-solving performance in systems of linear equations in two variables (SPLDV) and profiles learning improvements across Polya's stages. Specifically, the study addresses the following research questions: (RQ1) Does ChatGPT-driven adaptive scaffolding improve students' SPLDV problem-solving post-test scores relative to conventional instruction after controlling for pre-test performance? (RQ2) Do students who receive ChatGPT scaffolding demonstrate higher normalized learning gains (N-gain) than students in the control group? (RQ3) Which Polya stages show the largest and smallest improvements under ChatGPT-based scaffolding?

2. LITERATURE REVIEW

2.1. Adaptive Scaffolding and Formative Feedback as Mechanisms for Learning

Instructional scaffolding is commonly defined as contingent support that is dynamically adjusted to learners' needs and gradually withdrawn (faded) as competence increases. Seminal work emphasizes that effective scaffolding involves diagnosing learners' current understanding, providing targeted assistance, and progressively transferring responsibility to learners, rather than offering complete solutions from the outset [36]. In mathematics learning, scaffolding is closely linked to metacognitive regulation because students must plan, monitor, and evaluate solution pathways while navigating procedural and conceptual demands [37], [38].

From an assessment-for-learning perspective, scaffolding operates through feedback processes that reduce the discrepancy between current performance and learning goals. Foundational research on feedback indicates that its impact depends strongly on the level and function of feedback (e.g., task-level, process-level, and self-regulation–level) and that feedback can be ineffective or counterproductive when it is poorly targeted or promotes superficial compliance [39]. A comprehensive review further suggests that formative feedback is most effective when it is timely, specific, supportive, and actionable, and when it provides information that learners can immediately use to adjust strategies [40].

Meta-analytic evidence also shows that feedback effects vary widely across contexts and research designs, reinforcing the need to specify what feedback is provided, when it is provided, and how it is implemented within an intervention [41]. In computer-based environments, a meta-analysis indicates that elaborated feedback (i.e., explanations and guidance) tends to produce stronger learning effects than simple correctness feedback, particularly for higher-order outcomes [42]. Taken together, these perspectives support viewing ChatGPT-driven adaptive scaffolding as a mechanism that can operationalize formative feedback principles at scale—provided that the scaffolding is contingent, explanation-oriented, and designed to strengthen learners' self-regulation rather than replace it.

2.2. Large Language Models in Mathematics Learning

Large language models (LLMs) such as ChatGPT introduce a new form of interactive support that can generate on-demand explanations, hints, worked steps, and reflective prompts. Recent experimental evidence comparing ChatGPT-generated algebra hints with human tutor hints suggests that both conditions can yield learning gains; however, human-authored hints produced significantly larger improvements. Moreover, a nontrivial proportion of ChatGPT hints had to be discarded because they contained incorrect steps or answers [43], [44]. These findings underscore two implications for mathematics learning: (i) LLM-based scaffolding can be instructionally valuable as an always-available support channel, but (ii) it requires safeguards to minimize incorrect guidance and ensure pedagogical alignment.

Systematic syntheses in school mathematics education similarly conclude that ChatGPT's potential is accompanied by risks, including inaccuracies, overreliance, and variability in the quality of mathematical reasoning across tasks, which necessitate careful instructional orchestration and explicit prompting strategies [45]. At the level of model behavior, research in natural language processing (NLP) documents that LLMs may generate hallucinated outputs—plausible but false statements—particularly when the model's internal dynamics favor fluent continuation over factual correctness [46]. In mathematics classrooms, such failures are not merely technical limitations but pedagogical threats because incorrect steps can entrench misconceptions and distort students' procedural schemas. Accordingly, any ChatGPT-supported intervention should clearly specify whether ChatGPT functions primarily as (a) a hint generator, (b) a feedback provider, or (c) a reflective coach, and should articulate verification routines (e.g., teacher validation, rubric-based checking, or constrained prompting) to prevent high-fluency, low-validity guidance from shaping learning.

2.3. Polya-Aligned Problem-Solving Stages as Process Outcomes

Problem-solving performance is not only an outcome score but also a process competency that can be profiled across stages. Polya's classic framework conceptualizes problem solving as a four-stage cycle understanding the problem, devising a plan, carrying out the plan, and looking back and is widely used to structure mathematics instruction and assessment [47]. When operationalized as analytic indicators, these stages allow researchers to examine where learning gains occur, not merely whether gains occur. This distinction is particularly important for algebra topics such as systems of linear equations in two variables (SPLDV), where students may execute procedures successfully while remaining weak in representation, planning, and verification.

However, stage-based measurement requires explicit alignment between the intervention's instructional moves and the targeted stage indicators. If support primarily emphasizes step-by-step execution, improvements may concentrate in "carrying out the plan," while "devising a plan" remains underdeveloped. Conversely, prompts that emphasize justification, monitoring, and error checking may yield larger gains in "looking back." Theoretical work on scaffolding and metacognition strengthens the expectation that interventions can differentially influence Polya stages because metacognitive prompts and self-regulation supports are more proximal to planning, monitoring, and evaluation behaviors than to routine execution alone [36], [37].

Polya's [48] four-step framework (1) understanding the problem, (2) devising a plan, (3) carrying out the plan, and (4) looking back remains one of the most influential models of mathematical problem solving and continues to underpin instructional practice worldwide. Empirical studies also support its instructional value. For example, Fitriani [49] reported that structured implementation of Polya's stages improves students' mathematical reasoning. More broadly, systematic problem-solving models in classroom instruction have been associated with gains in students' higher-order thinking skills [50].

Table 1. Problem-Solving Assessment Rubric (Based on Polya's Framework)

No	Indicator	Score	Description
1	Understanding the Problem	5	Clearly restates the problem in words, identifying all essential data and objectives completely.
		4	Restates the problem with minor omission of less critical data.
		3	Partially restates the problem; some essential information is missing.
		2	Shows effort to understand, but with many mistakes or confusion.
		1	No attempt to understand or a completely incorrect statement.
2	Devising a Plan	5	Provides an appropriate and rational strategy (e.g., equation, comparison, diagram) and explains the reason for choosing it.
		4	Provides a suitable strategy but with a brief or incomplete explanation.
		3	Provides a less suitable strategy but demonstrates some relevant thinking.
		2	Provides an incorrect or inadequate strategy.
		1	No strategy provided / random guessing.
3	Carrying Out the Plan (Calculation & Logic)	5	Executes steps correctly, performs accurate calculations, uses neat notation, and maintains consistent logic.
		4	Significant steps correct, with minor calculation/presentation errors that do not affect the final solution.
		3	Some steps are correct, but significant errors lead to inaccurate results.
		2	Execution is disorganized; calculations are incorrect, or the sequence of steps is inconsistent.
		1	No meaningful execution.
4	Checking/Looking Back	5	Reviews the answer (e.g., substitution), interprets the result in the problem's context, and mentions rounding or assumptions if necessary.
		4	Provides brief checking or basic interpretation.
		3	Minimal effort required for checking, but not convincing.

No	Indicator	Score	Description
		2	No checking or incorrect interpretation.
		1	No attempt made.

In addition to Polya's framework, inquiry-based approaches have been shown to complement problem-solving instruction. In particular, problem-posing activities through which students generate mathematical question scan foster creativity and flexibility in approaching nonroutine tasks [51]. Menezes et al. [52] likewise advocate inquiry-oriented learning environments, arguing that such approaches cultivate both problem-solving and problem-posing competencies. Collectively, these findings suggest that problem solving is not only a pedagogical technique but also a broader learning philosophy that emphasizes reasoning, creativity, and adaptability.

2.4. ChatGPT-Based Adaptive Scaffolding

Integrating the strands above, this study conceptualizes ChatGPT-driven adaptive scaffolding as contingent, interactive support capable of delivering elaborated formative feedback and metacognitive prompts that are aligned with learners' evolving needs [39], [40], [42]. In practice, ChatGPT can be used to (i) clarify givens and unknowns to strengthen understanding the problem, (ii) propose strategic options and representations to support devising a plan, (iii) provide stepwise checks and targeted hints to improve carrying out the plan, and (iv) prompt verification and reflection to strengthen looking back [47]. At the same time, the literature suggests important boundary conditions: excessive guidance or premature provision of complete solution plans may reduce productive struggle and weaken planning autonomy, and incorrect or hallucinated steps may undermine conceptual integrity if not explicitly controlled [43], [45], [46]. Accordingly, effects may differ across Polya stages. The intervention is expected to produce the strongest gains in execution and reflection, whereas gains in planning likely depend on how prompts are engineered (e.g., requiring students to propose an initial plan before receiving adaptive refinement).

2.4.1. Independent variable (X): ChatGPT-driven adaptive scaffolding.

This construct refers to structured student–ChatGPT interactions designed to provide: (a) stage-specific prompts aligned with Polya's framework, (b) tiered hints that progress from minimal cues to more explicit guidance, and (c) feedback and verification prompts (e.g., checking assumptions, revisiting intermediate steps, and validating final results). In this study, AI support is implemented through a predefined prompting protocol and teacher-mediated rules that discourage direct requests for final answers and require strategy generation before AI consultation.

2.4.2. Dependent variable (Y): Mathematical problem-solving performance.

Problem-solving performance is measured using a contextual SPLDV test scored with a Polya-based rubric, yielding (1) a total score and (2) stage-level scores corresponding to each Polya component (understanding, planning, executing, and looking back/checking).

2.4.3. Covariate: Baseline ability (pre-test).

Students' pre-intervention problem-solving performance (pre-test total score and/or stage-level scores) is treated as a covariate to control for initial group differences in subsequent analyses.

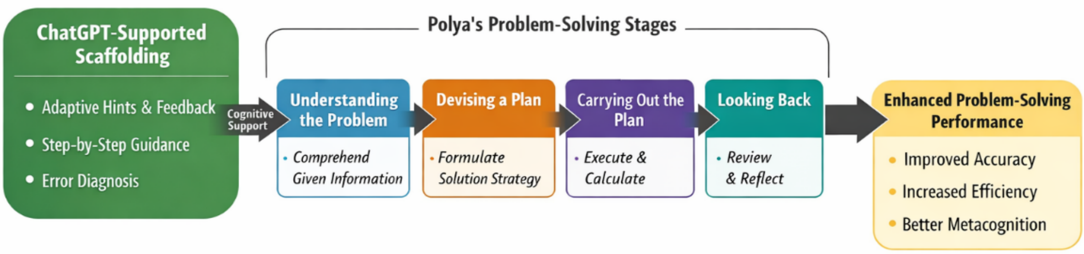


Figure 1. Conceptual framework of the study

The framework predicts that ChatGPT scaffolding will most strongly influence stages that benefit from immediate feedback and procedural guidance. Specifically, step-by-step hints and error diagnosis are expected to improve carrying out the plan, while metacognitive prompts (e.g., "verify your result," "check consistency," and "consider alternative solutions") are expected to strengthen looking back. By contrast, devising a plan may exhibit smaller gains if students become overly reliant on AI-generated strategies or if classroom routines do not require independent planning before help-seeking.

Hypotheses 1: After controlling for pre-test performance, students in the experimental group will achieve higher post-test problem-solving scores than students in the control group.

Hypotheses 2: Students receiving ChatGPT-driven adaptive scaffolding will demonstrate higher normalized learning gains (N-gain) than students receiving conventional instruction.

Hypotheses 3: Stage-level improvements will be larger for understanding the problem and carrying out the plan than for devising a plan, unless explicit instructional constraints require independent planning prior to AI use.

3. MATERIALS AND METHODS

3.1. Research Design

A pre-test–post-test non-equivalent control group quasi-experimental design was employed to examine the effect of ChatGPT-based adaptive scaffolding on eighth-grade students' mathematical problem-solving performance in systems of linear equations in two variables (SPLDV). Two intact Grade 8 classes from the same school participated: one class served as the experimental group and the other as the control group. Group assignment was conducted at the class level to preserve the existing classroom structure and instructional schedule.

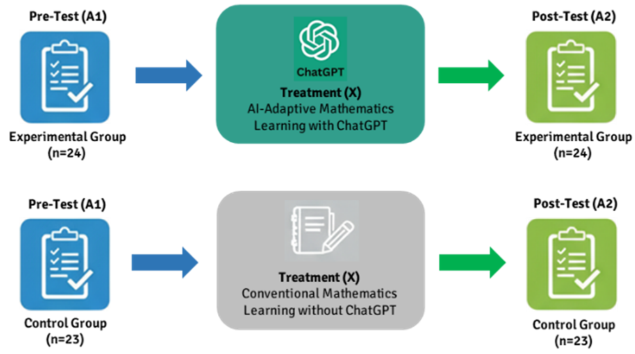


Figure 2. Research Design

Both groups received the same curriculum content, learning objectives, instructional time, and assessment schedule. The key difference was the mode of instructional support: (1) the experimental group engaged in ChatGPT-driven adaptive learning structured around Polya's problem-solving stages (understanding the problem, devising a plan, carrying out the plan, and looking back), and (2) the control group learned through conventional instruction (teacher explanations, guided discussion, and practice exercises) without AI support.

Table 2. Participant Demographics and Role in the Study

Characteristics	Experimental Group (n = 24)	Control Group (n = 23)	Total (N = 47)
Age			
13-years	11 (45.83%)	10 (43.48%)	21 (44.68%)
14-years	13 (54.17%)	13 (56.52%)	26 (55.32%)
Gender			
Male	10 (41.67%)	9 (39.13%)	19(40.43%)
Female	14 (58.33%)	14 (60.87%)	28 (59.57%)
Mathematics Achievement Level*			
High	6 (25.0%)	5 (21.7%)	11 (23.4%)
Medium	12 (50.0%)	11 (47.8%)	23 (48.9%)
Low	6 (25.0%)	7 (30.5%)	13 (27.7%)

*Mathematics achievement levels were categorized based on students' prior semester mathematics scores.

3.2. Participants

The study was conducted in Parepare City, South Sulawesi, Indonesia, involving 47 eighth-grade students in the second semester of the 2024/2025 school year. Of these, 24 were placed in the experimental group, which used AI-based adaptive mathematics learning. At the same time, 23 formed the control group, which underwent conventional teaching.

To ensure data validity, inclusion criteria were applied: (1) students had to attend at least 80% of the learning sessions, and (2) students had to complete both the pre-test and the post-test. The main research instrument was a mathematics problem-solving ability test, designed to measure students' reasoning and solution strategies before and after the intervention.

3.3. Instruments

This study used a mathematics problem-solving ability test as the primary instrument. The test was explicitly structured around Polya's four stages of problem solving: (1) understanding the problem, (2) devising a plan, (3) carrying out the plan, and (4) looking back. This structure was intended to capture both the correctness of students' final answers and the quality of their reasoning processes. The instrument consisted of six contextual SPLDV (systems of linear equations in two variables) problems designed to elicit students' ability to: (a) extract and organize relevant information from real-world situations, (b) define variables and formulate a system of two linear equations, (c) apply an appropriate solution strategy (e.g., elimination or substitution) accurately and coherently, and (d) interpret and verify solutions with respect to the original context. Representative items and contexts are shown in Table 3.

To ensure instrument quality, an analytic rubric aligned with Polya's four indicators was developed. Each indicator was rated on a 1–5 scale, so each student's final score reflected overall mastery of mathematical problem-solving. The instrument underwent expert-based content validation by mathematics education specialists to confirm alignment among the SPLDV items, Polya-stage indicators, language clarity, and contextual realism, followed by a pilot test with students outside the main sample to evaluate administration feasibility, instruction clarity, and reliability.

Table 3. Problem-Solving Ability Test – Topic SPLDV

No.	Context	Questions
1	Education	At a school, 72 students join extracurricular activities: futsal and scouting. Each futsal member pays Rp 20,000, while each scouting member pays Rp 15,000. The school collects a total of Rp1,220,000. How many students joined futsal, and how many joined scouting?
2	Liquid Mixture	A seller mixes syrup and mineral water. In the first mix, 3 bottles of syrup and 2 bottles of water make 5 liters of drink. In the second mix, 2 bottles of syrup and 4 bottles of water make 4.4 liters. How many liters are in one bottle of syrup and one bottle of water?

Table 4. Polya-Aligned Prompt Templates

Template	Polya Stages	Prompts
Template 1	Understand the problem	"Here is the problem: [paste]. Please help me identify the given information and what is being asked. Do not solve it; only restate the problem and list variables."
Template 2	Devise a plan	"I propose this plan: [write plan]. Can you check whether my plan is appropriate and suggest two alternative strategies (e.g., elimination vs substitution) without giving the final answer?"
Template 3	Carry out the plan—step check.	"I have done these steps: [paste steps]. Please check for errors and tell me the next step only. Do not complete the whole solution."
Template 4	Looking back—verification	"I obtained solution (x, y) = Please verify it by substitution and tell me whether it makes sense in the context. If incorrect, point out where my reasoning likely went wrong."
Template 5	Metacognitive reflection	"What is the key idea behind this problem? What common mistake should I avoid next time? Give me a short checklist to self-check my work."

To support students' work at each Polya stage while maintaining procedural consistency across participants, ChatGPT scaffolding was standardized using fixed, Polya-aligned prompt templates. Students selected templates based on the stage they were working on, and the prompts were designed to provide metacognitive and procedural support—such as identifying given/asked information, evaluating plans, checking steps for errors, and verifying solutions in context—while explicitly restricting ChatGPT from generating complete final solutions (Table 4).

Internal consistency of the total test score was acceptable (Cronbach's alpha α = 0.82). Two raters independently scored 25% of student responses using the Polya-based rubric; interrater reliability was high (ICC = 0.91). Discrepancies were resolved through discussion.

3.4. Intervention Procedure

3.4.1. Model specification and technical setting

The experimental group received ChatGPT-driven adaptive scaffolding during SPLDV instruction. The intervention used ChatGPT version 4.5 (accessed in 2025). Interactions were conducted in Indonesian to match the language of classroom instruction. Students accessed ChatGPT on their smartphones via school Wi-Fi using individual accounts. The teacher ensured stable connectivity and standardized device access to minimize technology-related confounds across sessions.

3.4.2. Duration, dose, and instructional timeline

The intervention lasted 4 weeks and comprised eight 45-minute sessions aligned with the SPLDV unit. Both the experimental and control groups received the same curricular objectives, worksheets, and total instructional time. The only difference was that the experimental group used ChatGPT as an adaptive scaffold during problem solving, whereas the control group received conventional teacher-led instruction without AI support.

3.4.3. Session structure

Each 45-minute session followed a standardized sequence: (a) briefing and rule reminders (5 min), (b) individual attempt and written plan (10 min), (c) ChatGPT interaction using Polya-aligned templates (10 min), (d) peer comparison of strategies (12 min), (v) whole-class verification and reflection prompts (6 min), and (e) exit-ticket reflection (2 min).

3.4.4. Rules and constraints for responsible ChatGPT use

To prevent answer copying and overreliance, ChatGPT use was governed by the following constraints:

- No final-answer requests: Students were prohibited from prompting ChatGPT to provide a complete solution (e.g., “Give me the full answer”).
- Plan-first requirement: Students were required to write an initial plan (strategy) before consulting ChatGPT.
- Hint-based interaction: Students were instructed to request hints, checks, or alternative strategies rather than fully worked solutions.
- Mandatory verification: Students were required to verify the final result using at least one method (e.g., substitution or comparison with an alternative strategy) before submitting their work.
- Teacher mediation: The teacher intervened when ChatGPT responses were incorrect, incomplete, or misleading and used these instances as teachable moments to highlight AI limitations.

3.5. Data Analysis

Data were analyzed using SPSS. Preliminary assumption checks included normality testing (Shapiro–Wilk) and homogeneity of variance testing (Levene’s test). Independent-samples t-tests were used when assumptions were met. When pre-test differences between groups were significant, analysis of covariance (ANCOVA) was conducted with pre-test scores included as covariates.

Students’ improvement was quantified using both raw gain and normalized gain (N-gain), which was categorized into low, medium, and high levels. Instrument reliability was assessed using Cronbach’s alpha, with values of $\alpha \geq 0.70$ considered acceptable [53]–[55]. In addition to expert-based content validation, validity evidence was examined by correlating pre-test and post-test scores (experimental group: $r = 0.545$; control group: $r = 0.965$), indicating acceptable score consistency across administrations.

4. RESULTS

4.1. Descriptive Analysis

Table 5 summarizes the pre-test and post-test performance of the experimental and control groups. The experimental group included 24 students. Pre-test scores ranged from 30 to 80 (M = 53.33, SD = 15.23). Following the intervention, post-test scores ranged from 40 to 90 (M = 68.33, SD = 12.74), indicating higher performance after treatment.

The control group included 23 students. Pre-test scores ranged from 20 to 70 (M = 49.13, SD = 12.40). Post-test scores ranged from 40 to 80 (M = 59.57, SD = 11.86). Although scores increased in the control group, the magnitude of improvement was smaller than that observed in the experimental group.

Table 5. Descriptive Analysis Results

Group	N	Minimum	Maximum	Mean	Std. Deviation
Experimental Pre-test	24	30	80	53.33	15.23
Experimental Post-test	24	40	90	68.33	12.74
Control Pre-test	23	20	70	49.13	12.40
Control Post-test	23	40	80	59.57	11.86

4.2. Normalized Gain (N-gain)

To further evaluate the effectiveness of ChatGPT-based adaptive learning, normalized gain (N-gain) was computed to quantify improvement from pre-test to post-test relative to the maximum possible gain. This metric facilitates a more comparable assessment of learning improvement across groups. Descriptive statistics for N-gain are presented in Table 6.

Table 6. N-gain Scores

Group	N	Mean N-Gain	Std. Deviation
Experimental	24	0.34	0.098
Control	23	0.21	0.072

The experimental group achieved a higher mean N-gain (0.34) than the control group (0.21), suggesting greater relative improvement under ChatGPT-based adaptive learning. Variability in N-gain was low in both groups, although the experimental group showed slightly greater dispersion (SD = 0.098 vs. 0.072).

Based on Hake’s [56] classification, the control group’s mean N-gain (0.21) falls in the low category, whereas the experimental group’s mean N-gain (0.34) falls in the medium category. This pattern indicates that the intervention was associated with stronger learning gains than conventional instruction.

4.3. Assumption Testing: Normality and Homogeneity of Variance

Prior to inferential analyses, normality and homogeneity assumptions were evaluated. Because each subgroup had fewer than 50 participants, normality was assessed using the Shapiro–Wilk test for each distribution (experimental/control × pre-test/post-test). Homogeneity of variance between groups was examined using Levene’s test. Results are shown in Tables 7 and 8.

Table 7. Tests of Normality (Shapiro–Wilk)

Group/Class	Statistic	df	Sig.	Interpretation
Pre-test Experimental	0.938	24	0.145	Normal ($p > 0.05$)
Post-test Experimental	0.940	24	0.167	Normal ($p > 0.05$)
Pre-test Control	0.932	23	0.121	Normal ($p > 0.05$)
Post-test Control	0.924	23	0.080	Normal ($p > 0.05$)

Table 8. Test of Homogeneity of Variance (Levene's Test)

Basis of Test	Levene Statistic	df1	df2	Sig.	Interpretation
Based on Mean	0.235	1	45	0.630	Homogeneous ($p > 0.05$)
Based on Median	0.155	1	45	0.696	Homogeneous ($p > 0.05$)
Based on Median (adj. df)	0.155	1	44.828	0.696	Homogeneous ($p > 0.05$)
Based on Trimmed Mean	0.206	1	45	0.652	Homogeneous ($p > 0.05$)

All Shapiro–Wilk tests were non-significant (0.080 – 0.167), indicating no evidence of meaningful deviation from normality (Table 7). Levene's test was also non-significant across estimators (0.652 – 0.696), supporting homogeneity of variance (Table 8). Therefore, parametric procedures were appropriate for subsequent analyses.

4.4. Within-Group Improvement Paired-Samples t-Tests

Within-group pre–post changes were examined using paired-samples t-tests (Table 9). Both groups showed statistically significant improvement in problem-solving scores from pre-test to post-test.

Table 9. Paired-Samples t-test

Group	Mean Difference	t-value	df	Sig.
Experimental	15.0	14.387	23	0.000
Control	10.435	13.651	22	0.000

The control group improved by 10.435 points, t-value (22) = 13.651, $p < 0.001$. The experimental group improved by 15.000 points, t-value (23) = 14.387, $p < 0.001$. The larger mean difference in the experimental group suggests greater improvement under the ChatGPT-supported condition.

4.5. Within-Between-Group Difference in Post-Test Scores: Independent-Samples t-test

To compare post-test performance between groups, an independent-samples t test was conducted (Table 10).

Table 10. Independent-Samples t Test (Post-test)

t-value	df	Sig.	Mean Diff.	95% CI Lower	95% CI Upper
2.439	45.0	0.019	8.768	1.528	16.008

Post-test scores differed significantly between groups, t-value (45) = 2.439, $p = 0.019$. The mean difference was 8.768 points (95% CI [1.528, 16.008]), indicating higher post-test performance in the experimental group than in the control group.

4.6. Group Differences in Post-Test Scores and N-gain: One-Way ANOVA

To further evaluate group differences and report effect sizes, one-way ANOVAs were conducted for post-test scores and N-gain (Tables 11 and 12).

Table 11. One-Way ANOVA for Post-Test Scores by Group

Group/Class	Sum of Squared	df	Mean Square	F-value	p-value	η^2
Between Group	901.259	1	901.259	5.940	0.019	0.117
Within Group	6827.586	45	151.724			
Total	7728.845	46				

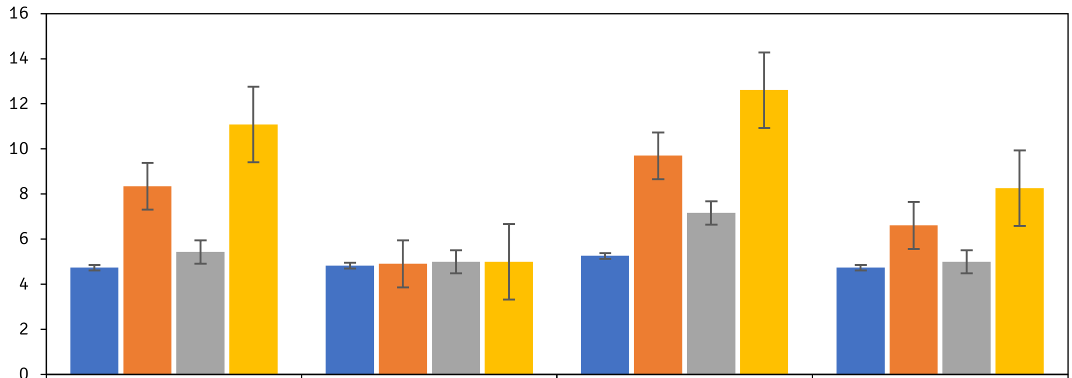
Table 12. One-Way ANOVA for N-gain by Group

Group/Class	Sum of Squared	df	Mean Square	F-value	p-value	η^2
Between Group	0.198	1	0.198	26.667	0.001	0.372
Within Group	0.335	45	0.007			
Total	0.533	46				

Post-test scores differed significantly by group, F-value (1, 45) = 5.94, p = 0.019, with a moderate effect size (η^2 = 0.117), indicating that 11.70% of the variance in post-test performance was attributable to group membership. Normalized gain also differed significantly by group, F-value (1,45) = 26.67, p = 0.001, with a large effect size (η^2 = 0.372), suggesting that 37.20% of the variance in learning gains was associated with the intervention. Overall, these findings indicate that the ChatGPT-supported condition was associated with higher post-test achievement and substantially greater relative learning improvement than conventional instruction.

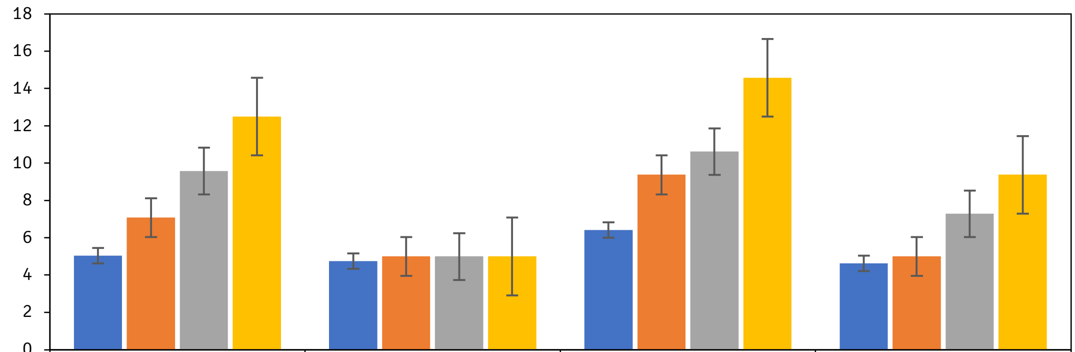
4.7. Polya's Problem-Solving Indicators

This study examined performance across Polya's four problem-solving stages—understanding the problem, devising a plan, carrying out the plan, and looking back—using pre-test and post-test assessments in both the control and experimental groups. Figures 3 and 4 summarize stage-level score patterns for the control group (conventional instruction) and the experimental group (ChatGPT-based adaptive learning), respectively.



	Understanding the Problem	Devising a Plan	Carrying Out the Plan	Rechecking/Looking Back
Pre-test (Q1)	4.74	4.83	5.26	4.74
Post-test (Q1)	8.35	4.91	9.7	6.61
Pre-test (Q2)	5.43	5	7.17	5
Post-test (Q2)	11.09	5	12.61	8.26

Figure 3. Score of Polya's Problem-Solving Indicators (Control Group)



	Understanding the Problem	Devising a Plan	Carrying Out the Plan	Rechecking/Looking Back
Pre-test (Q1)	5.04	4.75	6.42	4.63
Post-test (Q1)	7.08	5	9.38	5
Pre-test (Q2)	9.58	5	10.62	7.29
Post-test (Q2)	12.5	5	14.58	9.38

Figure 4. Score of Polya's Problem-Solving Indicators (Experimental Group)

The most pronounced gains were observed in understanding the problem and carrying out the plan (Control group). Average scores for understanding the problem increased from approximately 4.74–5.43 on the pre-test to 8.35–11.09 on the post-test, while scores for carrying out the plan rose from about 5.26–7.17 to 9.70–12.61. These improvements suggest that conventional instruction supported students' comprehension of problem statements and execution of solution procedures. In contrast, devising a plan remained largely unchanged at around 5, indicating persistent difficulty in strategy formulation. Looking back also improved (approximately 4.74–5.00 to 6.61–8.26), but gains were modest relative to those observed for understanding and execution.

Students who received ChatGPT-based adaptive scaffolding showed larger and more consistent improvements across indicators (Experimental group). The strongest gains again occurred in understanding the problem and carrying out the plan. Scores for understanding the problem increased from roughly 5.04–9.58 on the pre-test to 7.08–12.50 on the post-test. Similarly, carrying out the plan improved from approximately 6.42–10.62 to 9.38–14.58. These patterns suggest that ChatGPT's interactive prompts and adaptive feedback effectively supported both comprehension and procedural execution. Notably, looking back increased more substantially in the experimental group than in the control group, indicating that access to ChatGPT may have encouraged students to reflect on and verify their solutions more consistently. However, as in the control group, devising a plan remained the weakest stage, showing only minimal change from pre-test to post-test.

5. DISCUSSION

This study indicates that ChatGPT-assisted adaptive learning can substantially enhance students' mathematical problem-solving performance relative to conventional instruction. The findings are consistent with Polya's [48] framework, which conceptualizes problem solving as a cycle comprising four interrelated phases: understanding the problem, devising a plan, carrying out the plan, and looking back.

At the understanding stage, students in the experimental group demonstrated more consistent improvement than those in the control group. ChatGPT appeared to support this phase by reformulating problems in simpler language, highlighting critical information, and connecting tasks to students' prior knowledge. This pattern is consistent with Fitriani et al. [49], who emphasize the importance of early problem clarification for strengthening mathematical reasoning. In this role, ChatGPT may function as an accessible mediator that reduces barriers to comprehension by translating complex statements into manageable components [57]. By contrast, students in the control group relied primarily on teacher explanations and peer discussion, which may explain their comparatively smaller gains.

At the planning stage, performance remained relatively low in both groups. Although ChatGPT provided alternative strategies, many students appeared insufficiently prepared to evaluate, compare, and select the most appropriate approach. This finding aligns with Noster, Gerber, and Siller [26], who reported that students and pre-service teachers may accept ChatGPT-generated outputs without adequately scrutinizing their validity. Similarly, Busuttil and Calleja [28] argue that effective classroom use of AI requires deliberate pedagogical supports that cultivate metacognitive competencies, including anticipating outcomes, weighing alternatives, and justifying a strategy before executing it. Related concerns have also been raised by Zhai, Wibowo, and Li [58], who found that overreliance on AI dialogue systems can inhibit the development of independent reasoning.

At the implementation stage, the experimental group exhibited greater progress than the control group. ChatGPT's step-by-step explanations likely served as digital scaffolding that reduced cognitive load and supported more systematic execution [59], [60]. At the same time, this affordance may introduce a risk of overdependence. Some students appeared to replicate AI-generated procedures without fully internalizing the underlying mathematical logic [27]. Without explicit guidance to evaluate steps critically, reliance on ChatGPT may foster passive learning and reduce autonomy and creativity. Evidence that short AI-literacy interventions do not reliably prevent uncritical adoption of incorrect recommendations further underscores the importance of structured, guided reflection [61].

At the looking back stage, ChatGPT appeared to prompt students to verify solutions (e.g., via substitution) and interpret results within context, thereby strengthening reflective engagement in the experimental group. In contrast, students in the control group often omitted this phase in the absence of explicit prompts. This pattern aligns with Kapur's [62] concept of productive failure, which emphasizes deeper learning through testing, critiquing, and revising solutions. It is also consistent with Hyeon Jo et al. [63], who suggest that AI-supported reflective engagement can improve error detection and conceptual understanding.

The N-gain results reinforce the stage-level patterns: the experimental group achieved medium improvement (0.34), whereas the control group remained in the low range (0.21). Independent-samples t tests further indicated a statistically significant between-group difference ($p < 0.05$), suggesting that the observed improvements were attributable to the AI-supported intervention rather than incidental variation. This finding is consistent with Huang and Tu [64], who reported moderate but reliable positive effects of AI integration on students' mathematical reasoning in a systematic review.

From a theoretical perspective, the results contribute to the Knowledge for Teaching Mathematics with Technology (KTMT) framework by underscoring that educational technology should function not as an answer generator but as a cognitive scaffold that promotes critical, reflective, and adaptive thinking. Practically, the findings suggest that ChatGPT can serve as a supplementary learning tool for both independent study and classroom instruction. However, teacher mediation remains essential: teachers must ensure that students move beyond obtaining answers toward understanding and justifying the underlying mathematical reasoning [65], [66]. As Kanbul, Adamu, and Mohammed [67] emphasize, teacher preparedness is a key determinant of whether AI strengthens or undermines educational quality.

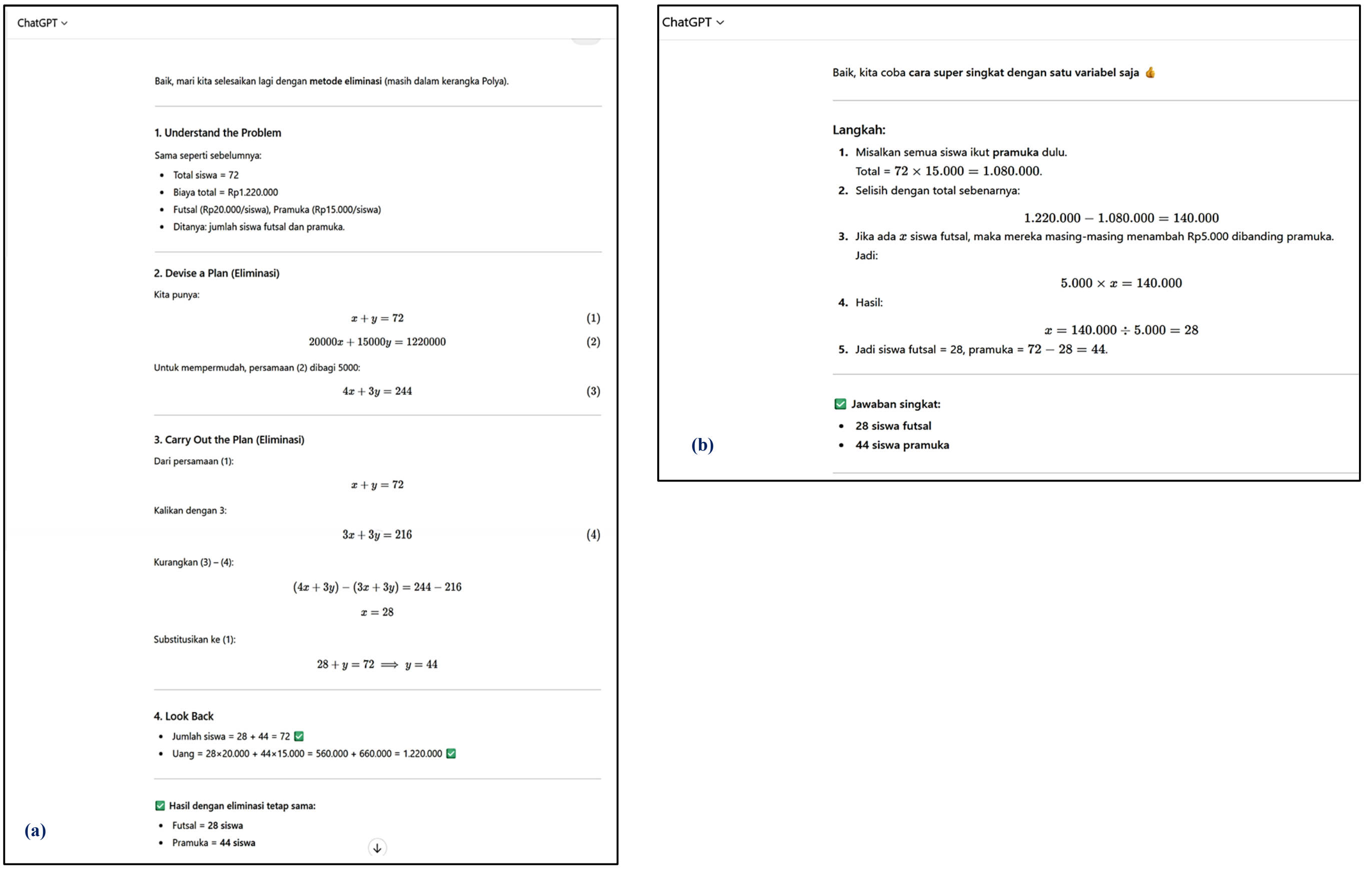


Figure 5. Question 1 with ChatGPT in Mathematics Learning. (a) Polya's Four Problem-Solving Indicators with the elimination method; (b) Alternative Short Solution with One Variable

Building on these implications, Figure 4 illustrates how ChatGPT can support students in solving SPLDV problems through two approaches. In Figure 4a, the solution process is explicitly structured according to Polya's four indicators. ChatGPT guides students to define variables from the given context, proposes an elimination-based plan, supports systematic execution, and then prompts verification through substitution. This approach reinforces a transparent reasoning sequence that can help students internalize the logic of problem solving rather than focusing solely on the final answer.

Figure 4b presents an alternative, more concise solution using a one-variable substitution approach. Although this strategy may be more intuitive and efficient for some learners, its brevity may also reduce opportunities to practice the systematic reasoning emphasized in Polya's framework. Thus, while shortcut methods can expedite computation, they may not consistently support the development of reflective and transferable problem-solving competence.

Providing multiple valid solution pathways can promote strategic flexibility and accommodate differences in students' cognitive preferences. This interpretation aligns with Szabo et al. [68] and Menezes et al. [52], who argue that problem-solving instruction should allow for varied approaches because diversity can enrich learning experiences. Nonetheless, teachers should emphasize that while concise solutions are mathematically valid, mastery of systematic problem-solving processes remains crucial for tackling nonroutine and complex tasks [50].

6. IMPLICATION AND LIMITATIONS

Overall, the findings align with prior research showing that AI integration can increase engagement, support personalization, and provide timely feedback [69], [70]. Related studies also highlight the value of problem-based learning for strengthening conceptual understanding and higher-order thinking [71], [72]. In this study, ChatGPT appeared to function as adaptive scaffolding by supporting procedural execution while also encouraging students to compare strategies, check correctness, and develop cognitive flexibility [73], [74]. However, realizing these benefits requires teacher-guided routines that prompt students to justify decisions and evaluate outputs critically.

Several limitations should be noted. First, devising a plan remained the weakest stage, suggesting that ChatGPT did not sufficiently strengthen students' strategic planning. Second, occasional inaccuracies or contextually inappropriate responses may mislead students who lack robust verification skills. Third, some students demonstrated dependency, consulting ChatGPT prematurely rather than attempting independent reasoning.

Accordingly, integrating ChatGPT into mathematics learning should be situated within a blended learning design that combines AI support with structured human mediation. Complementary strategies—such as problem posing, peer discussion, and guided reflection—are likely necessary to ensure that adaptive learning extends beyond procedural proficiency to foster creativity, autonomy, and higher-order reasoning.

7. CONCLUSION

This study shows that adaptive mathematics instruction supported by ChatGPT can enhance students' problem-solving performance relative to conventional teaching. Significant differences between the experimental and control groups were observed across both normalized learning gains (N-gain) and inferential comparisons (t-test results). At the process level, the strongest improvements occurred in Polya's indicators of understanding the problem and carrying out the plan. However, consistently weaker performance in devising a plan across both groups highlights the need for instructional designs that more deliberately cultivate strategic planning and other higher-order cognitive skills.

Theoretically, the findings suggest that LLM-based tutoring may preferentially strengthen procedural components of problem solving particularly comprehension and execution while providing more limited support for strategic planning. This distinction extends technology-integration frameworks by clarifying which cognitive subprocesses appear most responsive to generative-AI scaffolding. Practically, teachers should embed ChatGPT use within a structured verify-and-explain routine in which students (a) justify their chosen strategy, (b) evaluate AI outputs against mathematical principles, and (c) produce a brief written reflection before receiving confirmation or moving forward.

ChatGPT-supported instruction appears most beneficial for supporting students' progression through problem comprehension and execution, whereas strategic planning remains a persistent bottleneck. Accordingly, generative AI tools should be implemented within teacher-guided routines that explicitly develop planning heuristics and metacognitive monitoring. Future research should replicate these findings using larger samples, longer intervention periods, and implementation fidelity measures. Although the present results indicate meaningful improvements under structured ChatGPT scaffolding, the conclusions are limited by the quasi-experimental design, small sample size, and single-school setting. Replications using randomized or well-matched designs, extended durations, and systematic fidelity checks are needed to strengthen causal inference and generalizability.

ACKNOWLEDGMENTS

The authors would like to express their deepest gratitude to the University, participants and colleagues for their exceptional support and resources, which were critical in facilitating this research.

CONFLICTS OF INTEREST

The authors declare that no conflicts of interest are associated with this study. All aspects of the research were conducted with the utmost integrity and transparency.

DATA AVAILABILITY

The datasets utilized and analyzed during this research are available from the corresponding author upon reasonable request.

ETHICAL STATEMENTS

The authors confirm that the study complied with all applicable local laws, ethical standards, and institutional guidelines, including obtaining approval from relevant ethics committees.

FUNDING

This research was conducted without financial support. The authors confirm that no funding was received for this study's research, analysis, or publication.

REFERENCES

[1] M. Jurdak, "Learning and Teaching Real World Problem Solving in School Mathematics," Learn. Teach. Real World Probl. Solving Sch. Math., 2016, https://doi.org/10.1007/978-3-319-08204-2

[2] J. Nilimaa, "New Examination Approach for Real-World Creativity and Problem-Solving Skills in Mathematics," Trends High. Educ., vol. 2, no. 3, pp. 477–495, 2023, https://doi.org/10.3390/higheredu2030028

[3] B. Trilling and C. Fadel, 21St Century Skills: Learning for Life in Our Times, vol. 47, no. 10. John Wiley & Sons, 2010. https://doi.org/10.5860/choice.47-5788

[4] OECD, "An OECD Learning Framework 2030," The Future of Education and Labor. Springer, pp. 23–35, 2019. https://doi.org/10.1007/978-3-030-26068-2_3

[5] L. W. Anderson, D. R. Krathwohl, and B. S. Bloom, A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. Addison Wesley Longman, Inc., 2001.

[6] S. M. Brookhart, How to assess higher-order thinking skills in your classroom. Ascd, 2010.

[7] J. Cai, "What Research Tells Us About Teaching Mathematics Through Problem Solving," IEEE Sens. J., vol. 5, no. 2, pp. 273–276, 2005.

[8] E. G. Harskamp and C. J. M. Suhre, "Improving mathematical problem solving: A computerized approach," Comput. Human Behav., vol. 22, no. 5, pp. 801–815, 2006, https://doi.org/10.1016/j.chb.2004.03.023

[9] H. J. Pol, E. G. Harskamp, C. J. M. Suhre, and M. J. Goedhart, "How indirect supportive digital help during and after solving physics problems can improve problem-solving abilities," Comput. Educ., vol. 53, no. 1, pp. 34–50, 2009, https://doi.org/10.1016/j.compedu.2008.12.015

[10] F. K. Lester, "Thoughts about research on mathematical problem- solving instruction," Math. Enthus., vol. 10, no. 1–2, pp. 245–278, 2013, https://doi.org/10.54870/1551-3440.1267

[11] K. Stacey, "The place of problem solving in contemporary mathematics curriculum documents," J. Math. Behav., vol. 24, no. 3–4, pp. 341–350, 2005, https://doi.org/10.1016/j.jmathb.2005.09.004
[12] G. Polya, How to Solve It: A New Aspect of Mathematical Method. Princeton University Press, 2004.
[13] D. Jonassen, "Supporting Problem Solving in PBL," Interdiscip. J. Probl. Learn., vol. 5, no. 2, pp. 95–119, 2011, https://doi.org/10.7771/1541-5015.1256
[14] K. Manasa Lakshmi, K. Michael Angelo, S. Naga Lakshmi, and P. Ramu, "Smart approach for e-health monitoring system," in AIP Conference Proceedings, 2025, vol. 3162, no. 1, p. 20086. https://doi.org/10.1063/5.0249448
[15] I. Campo-Ruiz, "Artificial intelligence may affect diversity: architecture and cultural context reflected through ChatGPT, Midjourney, and Google Maps," Humanit. Soc. Sci. Commun., vol. 12, no. 1, pp. 1–13, 2025, https://doi.org/10.1057/s41599-024-03968-5
[16] Q. Wang, T. Sun, and R. Li, "Does Artificial Intelligence (AI) enhance green economy efficiency? The role of green finance, trade openness, and R&D investment," Humanit. Soc. Sci. Commun., vol. 12, no. 1, pp. 1–22, 2025, https://doi.org/10.1057/s41599-024-04319-0
[17] I. H. Cho and A. Chapagain, "Self-evolving artificial intelligence framework to better decipher short-term large earthquakes," Sci. Rep., vol. 14, no. 1, p. 21934, 2024, https://doi.org/10.1038/s41598-024-72667-7
[18] G. Zhang, A. Raina, J. Cagan, and C. McComb, "A cautionary tale about the impact of AI on human design teams," Des. Stud., vol. 72, p. 100990, 2021.
[19] S. Morandini, F. Fraboni, M. De Angelis, G. Puzzo, D. Giusino, and L. Pietrantoni, "the Impact of Artificial Intelligence on Workers' Skills: Upskilling and Reskilling in Organisations," Informing Sci., vol. 26, pp. 39–68, 2023, https://doi.org/10.28945/5078
[20] H. Ackermann et al., "Physical embodiment and anthropomorphism of AI tutors and their role in student enjoyment and performance," npj Sci. Learn., vol. 10, no. 1, p. 1, 2025, https://doi.org/10.1038/s41539-024-00293-z
[21] N. Bienefeld, E. Keller, and G. Grote, "AI Interventions to Alleviate Healthcare Shortages and Enhance Work Conditions in Critical Care: Qualitative Analysis," J. Med. Internet Res., vol. 27, p. e50852, 2025, https://doi.org/10.2196/50852
[22] I. Roll and R. Wylie, "Evolution and Revolution in Artificial Intelligence in Education," Int. J. Artif. Intell. Educ., vol. 26, no. 2, pp. 582–599, 2016, https://doi.org/10.1007/s40593-016-0110-3
[23] W. Holmes and I. Tuomi, "State of the art and practice in AI in education," Eur. J. Educ., vol. 57, no. 4, pp. 542–570, 2022, https://doi.org/10.1111/ejed.12533
[24] T. Shahzad, T. Mazhar, M. U. Tariq, W. Ahmad, K. Ouahada, and H. Hamam, "A comprehensive review of large language models: issues and solutions in learning environments," Discov. Sustain., vol. 6, no. 1, p. 27, 2025, https://doi.org/10.1007/s43621-025-00815-8
[25] T. Brown et al., "Language models are few-shot learners," Adv. Neural Inf. Process. Syst., vol. 33, pp. 1877–1901, 2020.
[26] N. Noster, S. Gerber, and H.-S. Siller, "Pre-Service Teachers' Approaches in Solving Mathematics Tasks with ChatGPT," Digit. Exp. Math. Educ., vol. 10, no. 3, pp. 543–567, 2024, https://doi.org/10.1007/s40751-024-00155-8
[27] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," Learn. Individ. Differ., vol. 103, p. 102274, 2023, https://doi.org/10.1016/j.lindif.2023.102274
[28] L. Busuttil and J. Calleja, "Teachers' Beliefs and Practices About the Potential of ChatGPT in Teaching Mathematics in Secondary Schools," Digit. Exp. Math. Educ., vol. 11, no. 1, pp. 140–166, 2025, https://doi.org/10.1007/s40751-024-00168-3
[29] B. Pepin, N. Buchholtz, and U. Salinas-Hernández, "A Scoping Survey of ChatGPT in Mathematics Education," Digit. Exp. Math. Educ., vol. 11, no. 1, pp. 9–41, 2025, https://doi.org/10.1007/s40751-025-00172-1
[30] K. Łodzikowski, P. W. Foltz, and J. T. Behrens, "Generative AI and Its Educational Implications," in Postdigital Science and Education (Netherlands), vol. Part F3835, Springer, 2024, pp. 35–57. https://doi.org/10.1007/978-3-031-64487-0_2
[31] R. Zhou et al., "Exploring ChatGPT-Facilitated Scaffolding in Undergraduates' Mathematical Problem Solving," J. Comput. Assist. Learn., vol. 41, no. 4, p. e70077, 2025, https://doi.org/10.1111/jcal.70077
[32] B. A. Becker et al., "Generative AI in Introductory Programming," Comput. Sci. Curricula 2023, Curric. Pract. Vol., vol. 1, p. 25, 2023.
[33] C. Fadel, W. Holmes, and M. Bialik, Artificial intelligence in education: Promises and implications for teaching and learning. The Center for Curriculum Redesign, Boston, MA, vol. 14, no. 4. Center for Curriculum Redesign, 2019.
[34] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education – where are the educators?," Int. J. Educ. Technol. High. Educ., vol. 16, no. 1, pp. 1–27, 2019, https://doi.org/10.1186/s41239-019-0171-0
[35] D. Loewenberg Ball, M. H. Thames, and G. Phelps, "Content knowledge for teaching: What makes it special?," J. Teach. Educ., vol. 59, no. 5, pp. 389–407, 2008, https://doi.org/10.1177/0022487108324554
[36] J. van de Pol, M. Volman, and J. Beishuizen, "Scaffolding in teacher-student interaction: A decade of research," Educ. Psychol. Rev., vol. 22, no. 3, pp. 271–296, 2010, https://doi.org/10.1007/s10648-010-9127-6
[37] D. Holton and D. Clarke, "Scaffolding and metacognition," Int. J. Math. Educ. Sci. Technol., vol. 37, no. 2, pp. 127–143, 2006, https://doi.org/10.1080/00207390500285818
[38] E. A. Davis and N. Miyake, "Explorations of Scaffolding in Complex Classroom Systems," in Scaffolding: A special issue of the journal of the learning sciences, Psychology Press, 2018, pp. 265–272. https://doi.org/10.4324/9780203764411-1
[39] J. Hattie, M. Gan, and C. Brooks, "Instruction Based on Feedback," in Handbook of Research on Learning and Instruction, Second edition, Routledge, 2016, pp. 290–324. https://doi.org/10.4324/9781315736419-23
[40] V. J. Shute, "Focus on Formative Feedback," ETS Res. Rep. Ser., vol. 2007, no. 1, pp. i–47, 2007, https://doi.org/10.1002/j.2333-8504.2007.tb02053.x
[41] B. Wisniewski, K. Zierer, and J. Hattie, "The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research," Front. Psychol., vol. 10, p. 487662, 2020, https://doi.org/10.3389/fpsyg.2019.03087
[42] F. M. Van der Kleij, R. C. W. Feskens, and T. J. H. M. Eggen, "Effects of Feedback in a Computer-Based Learning

Environment on Students' Learning Outcomes: A Meta-Analysis," Rev. Educ. Res., vol. 85, no. 4, pp. 475–511, 2015, https://doi.org/10.3102/0034654314564881
[43] Z. A. Pardos and S. Bhandari, "ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills," PLoS One, vol. 19, no. 5 May, p. e0304013, 2024, https://doi.org/10.1371/journal.pone.0304013
[44] A. Gupta, J. Reddig, T. Calò, D. Weitekamp, and C. J. MacLellan, "Beyond Final Answers: Evaluating Large Language Models for Math Tutoring," in Lecture Notes in Computer Science, 2025, vol. 15877 LNAI, pp. 323–337. https://doi.org/10.1007/978-3-031-98414-3_23
[45] M. Turmuzi, S. Azmi, and N. M. I. Kertiyani, "ChatGPT in school mathematics education: A systematic review of opportunities, challenges, and pedagogical implications," Teach. Teach. Educ., vol. 170, p. 105286, 2026, https://doi.org/10.1016/j.tate.2025.105286
[46] T. Jiang, S. Huang, Z. Luan, D. Wang, and F. Zhuang, "Scaling Sentence Embeddings with Large Language Models," in EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024, 2024, pp. 3182–3196. https://doi.org/10.18653/v1/2024.findings-emnlp.181
[47] G. Polya, How to solve it: A new aspect of mathematical method. Princeton university press, 2014. https://doi.org/10.2307/2306109
[48] G. Polya, How to Solve It. Princeton Press, 2019. https://doi.org/10.2307/j.ctvc773pk
[49] F. Fitriani, R. Hayati, S. Sugeng, S. Srimuliati, and T. Herman, "Students' Ability To Solve Mathematical Problems Through Polya Steps," J. Eng. Sci. Technol., vol. 17, pp. 25–32, 2022.
[50] Z. Aba-Oli, K. Koyas, and A. Husen, "Higher-order thinking skills-oriented problem-based learning interventions in mathematics: A systematic literature review," Sch. Sci. Math., vol. 125, no. 3, pp. 214–231, 2025, https://doi.org/10.1111/ssm.12676
[51] M. Sadak, L. Incikabi, F. Ulusoy, and M. Pektas, "Investigating mathematical creativity through the connection between creative abilities in problem posing and problem solving," Think. Ski. Creat., vol. 45, p. 101108, 2022, https://doi.org/10.1016/j.tsc.2022.101108
[52] L. Menezes, H. Oliveira, and A. P. Canavarro, "Inquiry-based mathematics teaching: The case of Célia," in Educational Paths to Mathematics: A CIEAEM Sourcebook, Springer, 2015, pp. 305–321.
[53] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, and R. L. Tatham, Multivariate Data Analysis, 8th ed. Edinburgh Gate, Harlow: Pearson Education Limited, 2019.
[54] B. Tabachnick and L. Fidell, Using Multivariate Statistics Title: Using multivariate statistics, vol. 5, no. 7th. Boston, Massachusetts, United State: Allyn & Bacon/Pearson Education, 2019.
[55] K. S. Taber, "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education," Res. Sci. Educ., vol. 48, no. 6, pp. 1273–1296, 2018, https://doi.org/10.1007/s11165-016-9602-2
[56] R. R. Hake, "Analyzing change/gain scores." 1999.
[57] F. Fui-Hoon Nah, R. Zheng, J. Cai, K. Siau, and L. Chen, "Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration," Journal of information technology case and application research, vol. 25, no. 3. Taylor & Francis, pp. 277–304, 2023. https://doi.org/10.1080/15228053.2023.2233814
[58] C. Zhai, S. Wibowo, and L. D. Li, "The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review," Smart Learn. Environ., vol. 11, no. 1, p. 28, 2024, https://doi.org/10.1186/s40561-024-00316-7
[59] M. Doorman, P. Drijvers, T. Dekker, M. van den Heuvel-Panhuizen, J. de Lange, and M. Wijers, "Problem solving as a challenge for mathematics education in The Netherlands," ZDM Math. Educ., vol. 39, no. 5, pp. 405–418, 2007, https://doi.org/10.1007/s11858-007-0043-2
[60] K. Reiss and G. Törner, "Problem solving in the mathematics classroom: The German perspective," ZDM - Int. J. Math. Educ., vol. 39, no. 5–6, pp. 431–441, 2007, https://doi.org/10.1007/s11858-007-0040-5
[61] C. Mah, H. Walker, L. Phalen, S. Levine, S. W. Beck, and J. Pittman, "Beyond CheatBots: Examining Tensions in Teachers' and Students' Perceptions of Cheating and Learning with ChatGPT," Educ. Sci., vol. 14, no. 5, p. 500, 2024, https://doi.org/10.3390/educsci14050500
[62] M. Kapur, "Examining Productive Failure, Productive Success, Unproductive Failure, and Unproductive Success in Learning," Educ. Psychol., vol. 51, no. 2, pp. 289–299, 2016, https://doi.org/10.1080/00461520.2016.1155457
[63] H. Jo, "From concerns to benefits: a comprehensive study of ChatGPT usage in education," Int. J. Educ. Technol. High. Educ., vol. 21, no. 1, p. 35, 2024, https://doi.org/10.1186/s41239-024-00471-4
[64] G.-J. Hwang and Y.-F. Tu, "Roles and research trends of artificial intelligence in mathematics education: A bibliometric mapping analysis and systematic review," Mathematics, vol. 9, no. 6, p. 584, 2021, https://doi.org/10.3390/su142013185
[65] O. Taani and S. Alabidi, "ChatGPT in education: benefits and challenges of ChatGPT for mathematics and science teaching practices," Int. J. Math. Educ. Sci. Technol., pp. 1–30, 2024, https://doi.org/10.1080/0020739X.2024.2357341
[66] S. A. Costa, "Reasoning, Argumentation, and Mathematical Learning: How ChatGPT and AI Can Knowledge Build," in Mathematics and Education in an AI Era: Cognitive Science, Technological, and Semiotic Perspectives, Springer, 2025, pp. 89–115. https://doi.org/10.1007/978-3-031-80696-4_6
[67] S. Kanbul, I. Adamu, and Y. B. Mohammed, "A Global Outlook on AI-Predicted Impacts of ChatGPT on Contemporary Education," SAGE Open, vol. 14, no. 3, p. 21582440241266370, 2024, https://doi.org/10.1177/21582440241266370
[68] Z. K. Szabo, P. Körtesi, J. Guncaga, D. Szabo, and R. Neag, "Examples of problem-solving strategies in mathematics education supporting the sustainability of 21st-century skills," Sustain., vol. 12, no. 23, pp. 1–28, 2020, https://doi.org/10.3390/su122310113
[69] A. Jančařík, J. Michal, and J. Novotná, "Using AI Chatbot for Math Tutoring," J. Educ. Cult. Soc., vol. 14, no. 2, pp. 285–296, 2023, https://doi.org/10.15503/jecs2023.2.285.296
[70] D. Lee and S. Yeo, "Developing an AI-based chatbot for practicing responsive teaching in mathematics," Comput. Educ., vol. 191, p. 104646, 2022, https://doi.org/10.1016/j.compedu.2022.104646
[71] I. Sasson, I. Yehuda, and N. Malkinson, "Fostering the skills of critical thinking and question-posing in a project-based

learning environment," Think. Ski. Creat., vol. 29, pp. 203–212, 2018, https://doi.org/10.1016/j.tsc.2018.08.001

[72] P. Kwangmuang, S. Jarutkamolpong, W. Sangboonraung, and S. Daungtod, "The development of learning innovation to enhance higher order thinking skills for students in Thailand junior high schools," Heliyon, vol. 7, no. 6, 2021, https://doi.org/10.1016/j.heliyon.2021.e07309

[73] M. V. Pop, G. Tonț, F.-V. Flonta, and M. Flore, "Agentic AI in STEM Education: Enhancing Cognitive Flexibility and Workforce Readiness," BRAIN. Broad Res. Artif. Intell. Neurosci., vol. 16, no. 1 Sup1, p. 239, 2025, https://doi.org/10.70594/brain/16.s1/20

[74] L. Y. Tan, S. Hu, D. J. Yeo, and K. H. Cheong, "Artificial intelligence-enabled adaptive learning platforms: A review," Comput. Educ. Artif. Intell., vol. 9, p. 100429, 2025, https://doi.org/10.1016/j.caeai.2025.100429