

Research paper

Guided AI Integration Model (GAIIM): Effects of Verification-Oriented Generative AI-Supported Learning on Undergraduate Students' Algebraic Problem-Solving

Isaac Davor * , Ernest Larbi , Francis Ohene Boateng , Michael Osei Tawiah 

University of Skills Training and Entrepreneurial Development (USTED), GHANA

*Corresponding Author: davorisaac70@gmail.com

Citation: Davor, I., Larbi, E., Boateng, F. O., & Tawiah, M. O. (2026). Guided AI integration model (GAIIM): Effects of verification-oriented generative AI-supported learning on undergraduate students' algebraic problem-solving. *Journal of Interdisciplinary Research in Artificial Intelligence and Society*, 2(1), Article 8. <https://doi.org/10.20897/jirais/19556>

Published: October 9, 2026

ABSTRACT

Artificial intelligence (AI) is increasingly being integrated into higher education; however, most instructional applications emphasize answer generation rather than mathematical reasoning and critical verification of AI-generated mathematical solutions. The current study examined the effectiveness of the Guided AI Integration Model (GAIIM), a verification-oriented instructional framework that utilizes AI for verification of students' algebraic problem-solving within a problem-solving context. An explanatory sequential mixed-methods design was employed. Quantitative data were collected from 86 undergraduate mathematics education students using a randomized pretest-posttest control group classroom design. During the 12-week intervention, students in the experimental group engaged in GAIIM-based instruction involving independent problem-solving, AI-generated solution comparison, error detection and justification, error resolution, and reflective verification, while the control group received conventional algebra instruction. Qualitative data were subsequently collected from 10 purposively selected students through semi-structured interviews and analyzed thematically to explain and elaborate on the quantitative findings. Students in the GAIIM group demonstrated significantly higher posttest performance in algebraic problem-solving, error detection, and error resolution after controlling for pretest scores. Qualitative findings further revealed stepwise verification, iterative engagement with AI feedback, conditional trust in AI-generated responses, and difficulties interpreting and correcting AI-generated mathematical reasoning. The findings suggest that structured AI-supported verification can support students' algebraic problem-solving and error-verification performance.

Keywords: algebraic problem-solving, artificial intelligence, error detection, error resolution, mixed methods, verification-oriented learning

Artificial intelligence (AI) is rapidly transforming higher education by changing how students learn, acquire knowledge, and complete academic tasks. However, the educational opportunities offered by AI are much broader than just creating academic material (Gimpel et al., 2023). University students increasingly use AI systems such as ChatGPT to support learning and complete academic tasks (Michel-Villarreal et al., 2023). Research indicates that the educational potential of AI is determined not merely by the availability of technology but by how it is integrated with instructional, learner, and contextual factors (Hornberger et al., 2023; Asamoah & Amarteifio, 2025; Xu & Ouyang, 2022). However, there is evidence from recent studies suggesting that the use of AI can foster successful

learning results if a proper instructional approach is used. Akendita et al. (2025) found that there exists a positive correlation between pre-service teachers of mathematics in Ghana's views on using AI and their performance. Boateng et al. (2026) reported positive relationships among AI-supported instructional practices, student engagement, and mathematical problem-solving. Similarly, Davor and Boateng (2026) found that AI-enhanced pedagogical practices were positively associated with students' mathematical language proficiency, with digital literacy and learning engagement contributing to these relationships. Kazadi Tshikolu et al. (2025) reported that AI-supported problem-based learning was associated with analytical and collaborative learning experiences among engineering students. Davor et al. (2026) reported associations among self-reported ChatGPT use, digital literacy, critical thinking, and mathematical problem-solving efficacy, suggesting that learner-related factors may shape students' experiences with AI-supported mathematical learning. Similarly, Yelamarthi et al. (2024) highlighted both the opportunities and challenges associated with generative AI in STEM-related higher education, emphasizing the importance of purposeful pedagogical integration.

Emerging evidence also suggests that the educational value of generative AI may depend on how its use is pedagogically structured, with scaffolded AI integration providing more deliberate support for learners than unscaffolded use (Boateng et al., 2026). Although students generally hold positive perceptions of AI in mathematics learning, researchers consistently emphasize that pedagogical guidance is essential if AI is to strengthen mathematical reasoning, metacognitive monitoring, and critical evaluation rather than encourage passive dependence on generated solutions (Contel & Cusi, 2025; Gqoli & Baidoo, 2025; Atuahene & Boateng, 2026; Joseph & Uzundu, 2024). Concerns about uncritical reliance on AI are also evident in AI-supported educational judgement, where questions of validity, reliability, fairness, and human oversight remain important. Experimental evidence in mathematics further shows that ChatGPT-generated support can improve learning while remaining vulnerable to quality-control problems; Pardos and Bhandari (2024) found significant learning gains from ChatGPT-generated mathematical help, although 32% of the generated hints initially failed quality checks. Zacharis and Papadakis (2025), for example, demonstrated limitations in the agreement between AI-generated and human educational assessments, emphasizing the continued importance of human oversight when AI outputs inform educational judgements. Such evidence reinforces the need to treat AI-generated outputs as resources requiring critical evaluation rather than as inherently authoritative responses.

However, there is a lack of effective instructional models that explicitly use AI to enhance students' ability to verify answers as part of solving algebraic problems. Although previous studies have examined students' perceptions of AI, AI-supported instruction, and AI-assisted problem solving, little empirical evidence exists regarding instructional models that deliberately position AI-generated solutions as resources for verification and comparison rather than as substitutes for students' independent problem-solving. The present study addresses this gap by evaluating the Guided AI Integration Model (GAIIIM), which structures students' interaction with AI around verification, comparison, and refinement of their own mathematical solutions.

Problem statement

Even though there is an increasing interest in the use of AI-based mathematics education, enhancing the capacity of students to solve algebraic problems stands out as one of the greatest problems in mathematics education. The ability to solve mathematical models and word problems is a major challenge among students due to the lack of understanding of algebraic relations (Jupri & Drijvers, 2016). One of the difficulties that the students experience while studying algebra is the meaningful application of symbols in various mathematical problems. Students often rely on symbols procedurally rather than relationally. As a result, there are many different situations in which students cannot use the mathematical concepts with which they have been working to complete problems within a variety of different contexts. Beyond procedural difficulties, students' engagement with algebraic problem-solving is also associated with learner-related factors. Julius (2022) reported significant relationships among students' self-concept, problem-solving skills, and attitudes toward solving algebraic problems.

Recently, Özcan (2024) demonstrated the value of structured worked-example instruction through the Worked-out Examples Involving Four Stages (WEI4S) approach, in which students analyzed worked solutions, explained solution procedures, examined incorrect solutions through error analysis, and applied their understanding to new algebraic problems. The findings showed that this structured engagement with worked examples improved students' algebraic problem-solving performance. This approach is particularly relevant to the present study because it demonstrates the instructional value of requiring learners to examine and reason about both correct and incorrect mathematical solutions rather than focusing solely on obtaining answers. However, Özcan's (2024) intervention was based on researcher-designed worked examples rather than AI-generated solutions, leaving unanswered how students can be systematically guided to evaluate, detect errors in, and resolve errors encountered in AI-generated algebraic solutions.

Besides the errors of a procedural nature, students encounter various other difficulties when solving mathematical problems (Cevikbas & Kaiser, 2026). These are of a cognitive as well as a motivational nature. For

example, in a study of 36 pre-service teachers, Kusuma et al. (2024) found that the higher the pre-service teachers' self-efficacy in solving mathematical problems, the more sophisticated their algebraic thinking was compared with that of their less confident peers. These concerns are also relevant at the university level, where mathematical problem-solving continues to receive considerable research attention. Santos-Trigo (2024) emphasized the continuing importance of problem solving as a central component of mathematics education, particularly in supporting learners' mathematical reasoning, exploration, and engagement with increasingly complex mathematical situations. This shows that mathematical problem-solving in the university not only supports students' academic achievement but also fosters higher-order mathematical thinking. The study conducted by Habtamu et al. (2022) found that the motivation for learning algebra was enhanced during the instructional phase. Also, Baumgartner et al. (2025) reported that students' approaches to problem solving changed after linear algebra instruction. However, the study focused on the effects of content instruction rather than the use of AI-supported verification to improve error detection and error resolution.

Most studies have focused on students' attitudes towards AI, teachers' awareness and issues faced when implementing AI, readiness of institutions, and the effectiveness of AI-assisted instruction (Akendita et al., 2025; Asamoah & Amarteifio, 2025; Atuahene & Boateng, 2026; Boateng et al., 2026; Gqoli & Baidoo, 2025). Though there is much promise in using AI to support mathematical verification, limited research has examined instructional models that systematically guide students in using AI to evaluate their solutions. Existing studies rarely distinguish between error detection and error resolution, and evidence from higher education, particularly in Ghana, remains limited. In order to bridge these gaps, this study examined the Guided AI Integration Model (GAIIM), a verification-oriented instructional model which helps students to compare, assess, and improve their algebraic solutions through AI-generated feedback. Employing an explanatory sequential mixed-methods research design, this study examined the effect of GAIIM on undergraduate students' algebraic problem-solving, error detection, and error resolution.

THEORETICAL FRAMEWORK

Social constructivist theory, more precisely, the Zone of Proximal Development (ZPD), provides the foundation for this study to analyze the ways students utilize AI-generated feedback while solving algebra problems. The ZPD refers to the distance between what learners can accomplish independently and what they can accomplish with appropriate guidance or support from a more knowledgeable other (Vygotsky, 1978). Central to the concept are social interaction, guided assistance, and scaffolding, through which learners receive temporary support to perform tasks that exceed their current independent capabilities. Within AI-supported mathematics instruction, generative AI can similarly function as a cognitive scaffold when its use is structured to support rather than replace students' mathematical reasoning (Davor, 2026). As competence develops, such support is gradually reduced, enabling learners to assume greater responsibility and eventually perform the task independently. Applying the concept of the Zone of Proximal Development (ZPD) by Vygotsky (1978), this research considers error detection to be a metacognitive task, which could be performed via scaffolding assistance, while the solution of an error requires a learner to assimilate knowledge internally rather than externally. Social constructivism implies the notion of learning as an interactive process when a person actively constructs knowledge through interaction, self-reflection, and scaffolding assistance as compared to the passive way of acquiring information (Sivan, 1986; Shulman & Carey, 1984). Scaffolding, in turn, means the assistance provided by a teacher to a learner to move forward through the ZPD while solving problems (Wood et al., 1976).

Empirical research suggests that modern technologies such as AI can help the learner in solving a problem in terms of offering immediate feedback, guidance regarding reasoning, and metacognitive activities (Contel & Cusi, 2025; Kazadi Tshikolu et al., 2025). Scaffolding offered in a mathematics classroom through technology use improves learners' problem-solving skills and helps them to think about the process of solving the problem (Boateng et al., 2026). At the same time, students' productive engagement with AI-generated solutions may depend on structured instructional guidance that encourages metacognitive monitoring and critical evaluation. At the same time, students' productive engagement with AI-generated solutions may depend on structured instructional guidance that encourages metacognitive monitoring and critical evaluation, particularly because metacognitive self-regulation can play an important role in AI-supported higher-order learning (Davor et al., 2026). Within GAIIM, AI-generated solutions function as external comparison resources rather than as scaffolds in themselves. The structured comparison activities, instructor prompts, error-justification tasks, correction activities, and reflection provide the instructional scaffolding through which students engage with AI-generated solutions.

The process of verifying a solution includes metacognitive monitoring (error detection) and procedural and conceptual understanding of algebraic concepts (error resolution) (Kirsten & Greefrath, 2023; Sommerhoff et al., 2025). Therefore, error detection involves self-evaluation of one's solution in terms of its accuracy, while error resolution requires an understanding of relevant algebraic principles and operations. Based on these considerations, social constructivist theory, particularly the Zone of Proximal Development (ZPD), provides a basis for examining

how structured instructional support around AI-generated solutions contributes to students' algebraic problem-solving activities. In GAIIM, the AI-generated solution serves as an external reference, while the instructional activities surrounding its use provide the scaffolding.

Research questions

RQ1: What is the effect of GAIIM on undergraduate students' algebraic problem-solving performance?

RQ2a: What is the effect of GAIIM on students' ability to accurately detect errors in AI-generated algebraic solutions?

RQ2b: What is the effect of GAIIM on students' ability to correctly resolve errors in AI-generated algebraic solutions?

RQ3: How do undergraduate students engage with AI-generated feedback during the verification of their algebraic problem-solving processes?

RQ4: What challenges do students encounter when using AI as a verification tool rather than an answer-generating tool within the GAIIM framework?

METHODS

Research design

An explanatory sequential mixed-methods design was employed, in which quantitative data were first collected and analyzed, followed by qualitative data to explain and elaborate on the quantitative findings (Creswell & Creswell, 2018). The quantitative phase employed a randomized pretest-posttest control group classroom design. One intact Level 200 Mathematics Education class was initially selected through convenience sampling, after which the 86 participating students were randomly allocated at the individual level to the experimental and control conditions. The intervention was implemented through an active-learning approach in which students engaged in algebraic problem-solving using structured AI-generated feedback. The design allowed for an evaluation of the effects of GAIIM on students' algebraic problem-solving, error-detection, and error-resolution performance.

Participants and sampling

The target population for this study involved undergraduate students enrolled in the BSc Mathematics Education programme in one of the public universities in Ghana. A total of 86 Level 200 undergraduate students participated in the study. For the quantitative phase, convenience sampling was used to select one intact Level 200 Mathematics Education class. Before the intervention commenced, the 86 participating students were randomly allocated at the individual level to either the experimental group ($n = 43$) or the control group ($n = 43$). Each participant was assigned a unique identification number, after which a computer-generated randomization procedure was used to assign participants to the two instructional conditions in a 1:1 ratio. The allocation sequence was generated and implemented by an independent research assistant who was not involved in delivering the intervention. Thus, convenience sampling was used to select the participating class, whereas individual random allocation was used to assign participants to the experimental and control conditions. Both groups were taught by the same instructor, covered identical algebra topics, received the same instructional time throughout the 12-week intervention period, and followed the same assessment schedule and learning objectives.

For the qualitative phase, 10 students from the experimental group were purposively selected on the basis of their participation in the intervention, willingness to participate in follow-up interviews, and capacity to provide detailed accounts of their experiences with the GAIIM activities. Students were eligible for inclusion if they had participated in the GAIIM intervention, completed the quantitative phase, and consented to the follow-up interview. Students who did not complete the intervention or quantitative assessments or who declined participation in the interview were excluded. The qualitative sample was selected to provide explanatory accounts of students' experiences rather than according to predetermined quantitative performance profiles; consequently, integration was conducted at the outcome rather than participant level. The two instructional conditions were conducted separately during the intervention period to reduce direct exposure of control participants to the GAIIM activities. Nevertheless, because participants belonged to the same academic programme and could interact outside the scheduled instructional sessions, communication across conditions could not be completely controlled. [Table 1](#) presents the demographic characteristics of the 86 participants.

Table 1*Demographic characteristics of participants (N = 86)*

Variable	Category	Frequency (n)	Percentage (%)
Gender	Male	54	62.8
	Female	32	37.2
Age	18-20 years	21	24.4
	21-23 years	46	53.5
	24 years and above	19	22.1
Programme of Study	BSc Mathematics Education	86	100
Level of Study	Level 200	86	100

Instruments

The data collection procedure included two researcher-designed tools intended to measure the students' algebraic proficiency and verification skills. Algebraic Problem-Solving Test (APST) was composed of seven structured problems related to linear equations, quadratic equations, rational expressions, inequalities, logarithmic equations, and systems of equations. Each problem was scored using an analytic rubric that considered the accuracy of the solution process and the justification provided, resulting in a total score of 30 points. The second data collection tool, Algebraic Verification Task (AVT), aimed at measuring the students' skills in detecting and resolving errors in the AI-generated algebraic solutions. The AVT consisted of five verification tasks. For each task, students were required to determine whether the presented reasoning was mathematically valid and, where an error was identified, locate and justify the error using relevant algebraic principles and reconstruct the correct solution. Error-detection and error-resolution scores were derived separately from the same verification tasks using the scoring rubric presented in Appendix C. Error detection and error resolution were therefore treated as related dimensions of performance derived from the same verification tasks rather than as outcomes obtained from separate sets of tasks. The results for detecting and resolving errors were calculated separately. Two parallel versions of the instruments were developed for the pretest and posttest phases. The forms were mapped to the same algebraic content domains, intended learning outcomes, item formats, and scoring criteria. Numerical values and symbolic representations were varied between the forms to reduce direct recall and practice effects (Polotskaia et al., 2022).

The forms were reviewed by the expert panel to examine their alignment in content coverage, cognitive demand, and scoring expectations. Although the forms were matched through content mapping and expert review, they were not statistically equated; therefore, equivalence in empirical difficulty cannot be assumed. The AVT used deliberately constructed worked solutions that included both mathematically valid and erroneous solutions to provide standardized verification conditions across participants. Students were therefore required to determine whether an error was present before identifying, justifying, and resolving any error detected. These assessment tasks were designed to evaluate students' ability to identify, justify, and correct algebraic errors under common assessment conditions and were distinct from spontaneous inaccuracies that could arise during live interaction with ChatGPT.

Instructional procedure of the guided AI integration model (GAIIM)

The Guided AI Integration Model (GAIIM) was adopted as a structured instructional sequence to guide the students' use of artificial intelligence while solving algebra problems. The intervention was implemented over 12 weeks, from Week 2 to Week 13, with two instructional sessions per week, each lasting approximately 90 minutes. The instructional task included the process of algebra problem solving, such as linear equations, quadratic equations, inequalities, and rational expressions. Instructional material for both groups was kept constant to ensure equivalency. The instructional procedure in the Guided AI Integration Model (GAIIM) comprised six sequential phases. Although ChatGPT generated step-by-step solutions, these outputs were used pedagogically as reference solutions for verification and comparison rather than as substitutes for students' independent problem-solving.

Phase 1: Independent problem-solving attempt

Students solved algebra problems independently without using any AI tools. Their written solutions were collected and examined to establish their initial problem-solving approach before exposure to AI-generated solutions. Attention was given to the algebraic procedures, solution steps, calculations, and mathematical justifications evident in their written work.

Phase 2: AI-based solutions production

Students produced AI-based answers to algebra problems using the same standardized prompt throughout the intervention period. For each instructional session, all students in the experimental group worked on the same set of algebra problems and entered the same standardized prompt into ChatGPT for each problem. The prompt was: “Provide a step-by-step answer to the following algebra problem. Explain each math operation you performed, the algebraic rule used at each step, and show the solution in its entirety without skipping any calculations.”

Phase 3: Step-wise comparison

Students compared their independent answers and AI-based solutions in a step-by-step manner using the worksheet for step-by-step comparison. Specific steps at which the students’ solutions differed from the AI-generated solutions were identified.

Phase 4: Error detection and justification

Students identified the error and provided written explanations of why the identified steps were mathematically incorrect. The instructor provided prompts such as “What wrong assumption did you make?” and “What algebraic rule was broken?”

Phase 5: Error resolution

Students corrected the identified error and reconstructed the solution. The correction was justified using relevant algebraic rules and principles.

Phase 6: Reflective verification

Students completed a short reflective task in which they described how the error occurred, how it was detected, and how it was corrected. Fidelity was implemented through a standardized checklist that was filled out after every lesson. The checklist included the implementation of all GAIIM phases, use of standardized prompts, time taken to complete the lesson, and verification activities. The students were not allowed to submit AI-generated solutions directly; the students had to write down their own answers, justifying them. In this way, the pedagogical purpose of the AI-generated solutions was verification and comparison rather than direct answer submission. Each instructional period consisted of four algebraic verification activities. The tasks involved solving the problems independently, generating the AI answer using the standardized prompt, comparing the two answers using the verification worksheet, noting any differences, justifying the errors found, generating the correct answer, and carrying out a reflective verification activity.

Control group instruction

The participants of the control group were taught the conventional way of algebraic instruction that comprised explanation by the teacher, examples, and exercises in the class, as well as practice by themselves. Similar algebraic instruction topics, time duration, worksheets, and testing schedule followed in the experimental group were adhered to maintain instructional equivalence. However, students did not use ChatGPT or engage in AI-based verification activities during the instructional period. Accordingly, the comparison examined the effect of the GAIIM instructional package relative to conventional instruction rather than isolating the independent contribution of AI-generated verification from the other structured activities incorporated within GAIIM.

Validity and reliability of instruments

Content validity was established through review by three experts with expertise in mathematics education, algebra instruction, and educational measurement. The experts evaluated the items for relevance, clarity, representativeness of the intended algebraic domains, appropriateness of difficulty, and alignment with the study objectives and scoring criteria. Their comments informed revisions to item wording, mathematical structure, and scoring specifications before pilot testing (Teresi et al., 2022). A pilot study involving 25 undergraduate mathematics education students produced reliability coefficients of .86 for the APST, .83 for error detection, and .81 for error resolution. Two raters independently scored 30% of the pilot responses, yielding strong inter-rater reliability ($ICC = .91$). Scoring disagreements were resolved through discussion and consensus.

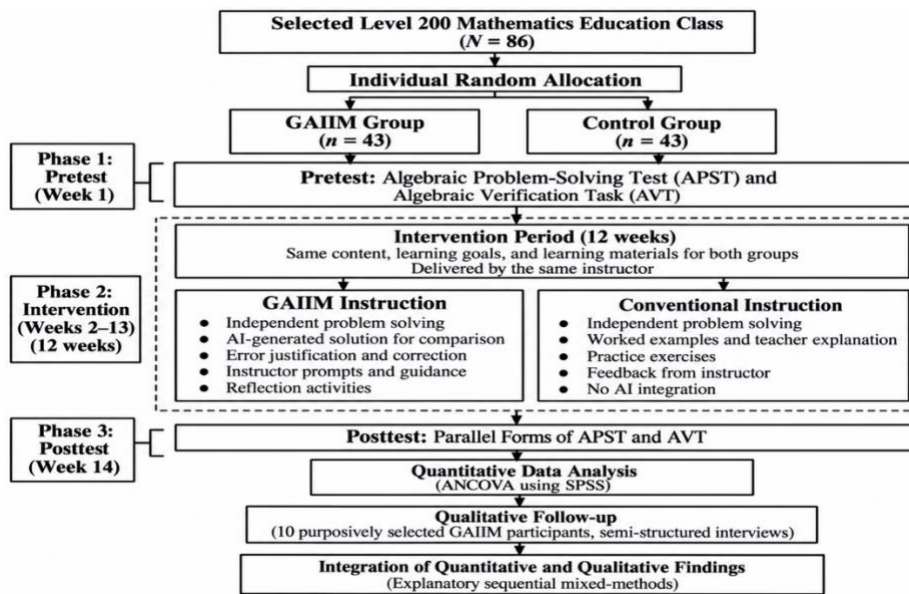
Procedure

The study followed three main stages: pretest, intervention, and posttest, and was conducted over 14 weeks. The pretest was administered in Week 1, the intervention was implemented from Weeks 2 to 13, and the posttest was administered in Week 14. Both groups completed the Algebraic Problem-Solving Test and the five-task

Algebraic Verification Task during the pretest stage. During the 12-week intervention period, both groups were taught by the same instructor and followed the same algebraic content, learning goals, instructional schedule, and core learning materials, while differing in the instructional approach used (see Figure 1). Both groups completed the parallel posttest forms of the APST and AVT after the intervention. The pretest and posttest assessments were administered under the same standardized testing conditions for both instructional groups. Responses to the APST and AVT were independently scored by two raters using the predetermined analytic scoring rubrics. To reduce the possibility of scoring bias, participants' scripts were identified using participant codes rather than instructional-group labels during scoring. Consequently, the scorers were not informed of participants' group membership while evaluating the responses.

Figure 1

Study design and procedural sequence



ChatGPT implementation

ChatGPT-4o was accessed through the OpenAI web interface using free accounts during the 12-week intervention. All interactions were conducted in English, and students generated AI responses live during the instructional sessions using the standardized prompt provided above. A new chat was initiated for each algebraic problem to minimize the influence of previous interactions on subsequent responses. Because ChatGPT-4o could produce variations in explanations even when the same prompt was used, students worked with the first complete response generated for each task rather than repeatedly prompting the system to obtain a preferred response. Where an AI-generated solution contained an error or differed from the student's solution, the response was retained for comparison, error detection, justification, and correction rather than being replaced. Because responses could vary across students even under the standardized prompt, the intervention standardized the procedure for engaging with AI rather than guaranteeing identical AI-generated content for every participant.

Data analysis

The quantitative data analysis was carried out by means of ANCOVA using SPSS Version 23 to determine the effects of the interventions after controlling for the pretests. Homogeneity of variance was assessed using Levene's test, while homogeneity of regression slopes was examined by testing the interaction between instructional group and the corresponding pretest covariate for each outcome. Estimated marginal means, standard errors, adjusted group comparisons, and 95% confidence intervals were computed to facilitate interpretation of the adjusted group differences. The magnitude of effect was determined with partial eta squared (η^2), which represents .01, .06, and .14 as small, medium, and large effects, respectively. The qualitative data were analyzed thematically following Braun and Clarke's (2006) six-phase procedure. The interview recordings were transcribed verbatim and read repeatedly to support familiarization with the data. Initial codes were generated from participants' accounts and organized around patterns related to engagement with AI-generated feedback and challenges encountered during verification. The coding was primarily inductive, while the research questions guided the organization of the resulting themes. Candidate themes were reviewed against the coded extracts and the full dataset before being refined and named. The researchers discussed the coding and theme structure and resolved differences through consensus.

RESULTS

Assumption testing

The assessed ANCOVA assumptions were adequately satisfied. An initial assessment of variance for each of the dependent variables by group utilizing Levene's test did not reveal any significant differences among groups (p -values $> .05$; see [Tables 2](#) and [3](#)). Furthermore, an assessment of the homogeneity of regression slopes for each of the dependent variables revealed no significant interaction between group and pretest measures (all p -values $> .05$). Additionally, a review of the data utilizing the standardized residuals did not reveal any extremely large or small values that could affect the results of the ANCOVA. Thus, the use of ANCOVA for testing the current set of hypotheses is appropriate.

Table 2

Descriptive statistics for pretest and posttest scores by group

Outcome	Test	Group	N	M	SD	95% CI
Error Detection	Pretest	Control	43	4.98	1.61	[4.48, 5.47]
		GAIIM	43	4.91	1.60	[4.41, 5.40]
	Posttest	Control	43	5.28	1.88	[4.70, 5.86]
		GAIIM	43	7.28	1.96	[6.68, 7.88]
Error Resolution	Pretest	Control	43	5.05	0.93	[4.76, 5.33]
		GAIIM	43	4.93	0.86	[4.67, 5.19]
	Posttest	Control	43	6.42	1.18	[6.06, 6.78]
		GAIIM	43	7.05	1.17	[6.69, 7.41]
Algebraic Problem-Solving	Pretest	Control	43	19.86	3.71	[18.72, 21.00]
		GAIIM	43	20.72	3.44	[19.66, 21.78]
	Posttest	Control	43	22.51	4.26	[21.20, 23.82]
		GAIIM	43	27.60	4.34	[26.27, 28.94]

Specifically, for the measure of error detection, the control group achieved a mean of 4.98 with a standard deviation of 1.61 on the pretest. In comparison, the GAIIM group achieved a mean of 4.91 with a standard deviation of 1.60 on the pretest measure of error detection. On the posttest measure of error detection, the GAIIM group achieved a mean of 7.28 with a standard deviation of 1.96, whereas the control group achieved a mean of 5.28 with a standard deviation of 1.88 on the posttest measure. From the pretest data for error resolution, the control group scored a mean of 5.05 ($SD = 0.93$), whilst the GAIIM group scored a mean of 4.93 ($SD = 0.86$). The scores for the two groups on the posttest for error resolution were 7.05 ($SD = 1.17$) for the GAIIM group and 6.42 ($SD = 1.18$) for the control group. For algebraic problem solving, the two groups of participants were found to be similar at pretest (Control $M = 19.86$, $SD = 3.71$; GAIIM $M = 20.72$, $SD = 3.44$). However, for posttest, the GAIIM group achieved substantially higher scores for algebraic problem solving ($M = 27.60$, $SD = 4.34$) than did the Control group ($M = 22.51$, $SD = 4.26$).

Table 3

Independent-samples t -tests for baseline equivalence

Variable	Control M (SD)	GAIIM M (SD)	Levene's F	P-value	t(df)	P-value
Error Detection	4.98 (1.61)	4.91 (1.60)	0.046	.831	0.201 (84)	.841
Error Resolution	5.05 (0.93)	4.93 (0.86)	0.142	.707	0.605 (84)	.547
Algebraic Problem-Solving	19.86 (3.71)	20.72 (3.44)	0.397	.530	-1.115 (84)	.268

Again, an independent samples t -test was employed to evaluate the equivalence of the control and GAIIM group at the baseline before the implementation of the intervention. The results from Levene's test in [Table 3](#) shows that the assumption of homogeneity of variances was met for error detection, $F(1, 84) = 0.046$, $p = .831$, error resolution, $F(1, 84) = 0.142$, $p = .707$, and algebraic problem solving, $F(1, 84) = 0.397$, $p = .530$. Results from independent-samples t -test showed that there were no statistically significant differences between the groups in any of the pretest scores.

Table 4*Tests of the homogeneity of regression slopes assumption*

Outcome Variable	Interaction Term	F (1, 82)	P-value	Decision
Algebraic Problem-Solving	Group x Pre APS	1.358	.247	Assumption satisfied
Error Detection	Group x Pre Detection	0.022	.881	Assumption satisfied
Error Resolution	Group x pre-Resolution	0.872	.353	Assumption satisfied

The homogeneity-of-regression-slopes assumption was investigated by testing the interaction between instructional group and the respective pretest covariates for each dependent variable. As presented in **Table 4**, there was no significant interaction between instructional group and the pretest scores on algebraic problem solving, $F(1, 82) = 1.358$, $p = .247$. Likewise, there was no significant interaction between instructional group and the pretest scores on error detection, $F(1, 82) = 0.022$, $p = .881$, and there was no significant interaction between instructional group and pretest scores on error resolution, $F(1, 82) = 0.872$, $p = .353$. From these results, the researchers infer that the regression slope for each covariate and its respective posttest score was the same across instructional groups.

RQ1: What is the effect of GAIIM on undergraduate students' algebraic problem-solving performance?

To examine the effect of the Guided AI Integration Model (GAIIM) on students' posttest algebraic problem-solving performance, a one-way analysis of covariance (ANCOVA) was conducted using posttest algebraic problem-solving scores as the dependent variable while controlling for pretest algebraic problem-solving scores. From **Table 5**, the overall ANCOVA model was statistically significant, $F(2, 83) = 152.470$, $p < .001$, accounting for 78.6% of the variance in posttest algebraic problem-solving scores, $R^2 = .786$, adjusted $R^2 = .781$. The pretest score was a significant covariate, $F(1, 83) = 202.536$, $p < .001$, partial $\eta^2 = .709$.

Table 5*Analysis of covariance (ANCOVA) for posttest algebraic problem-solving scores*

Source	SS	df	MS	F	P-value	Partial η^2
Pretest APS	1103.007	1	1103.007	202.536	< .001	.709
Group	377.611	1	377.611	69.338	< .001	.455
Error	452.016	83	5.446			
Corrected Model	1660.693	2	830.346	152.470	< .001	.786

Model fit: $R^2 = .786$; Adjusted $R^2 = .781$.

Table 6*Adjusted posttest means and group comparison for algebraic problem-solving*

Group	Adjusted Mean	SE	95% CI	
Control	22.95	0.36	[22.24, 23.66]	
GAIIM	27.17	0.36	[26.46, 27.88]	
Group Comparison	Mean Difference	SE	P-value	95% CI
Control – GAIIM	-4.222	0.507	< .001	[-5.230, -3.213]

Most importantly, the instructional group had a statistically significant effect on posttest algebraic problem-solving after controlling for pretest differences, $F(1, 83) = 69.338$, $p < .001$, partial $\eta^2 = .455$. The GAIIM group obtained a significantly higher adjusted posttest mean (adjusted $M = 27.17$, $SE = 0.36$) than the control group (adjusted $M = 22.95$, $SE = 0.36$) (**Table 6**). The adjusted group comparison showed a statistically significant mean difference of 4.22 points between the two groups ($p < .001$, 95% CI [-5.230, -3.213]).

RQ2a: What is the effect of GAIIM on students' ability to accurately detect errors in AI-generated algebraic solutions?

An ANCOVA was also used to determine the effect of GAIIM on the posttest error-detection measures of students. The pretest error-detection measures of students were used as covariates for this analysis. The overall ANCOVA model in **Table 7** was significant, $F(2, 83) = 156.127$, $p < .001$. This model explained 79.0% of the variance in posttest error-detection of students' measures ($R^2 = .790$, adjusted $R^2 = .785$). The analysis for the covariates indicated that the pretest error-detection measures of students significantly predicted their posttest error-detection measures, $F(1, 83) = 226.264$, $p < .001$, partial $\eta^2 = .732$.

Table 7*Analysis of covariance (ANCOVA) for posttest error detection scores*

Source	SS	df	MS	F	P-value	Partial η^2
Pretest Error Detection	226.292	1	226.292	226.264	< .001	.732
Group	92.197	1	92.197	92.186	< .001	.526
Error	83.010	83	1.000			
Corrected Model	312.292	2	156.146	156.127	< .001	.790

Model fit: $R^2 = .790$; Adjusted $R^2 = .785$.**Table 8***Adjusted posttest means and group comparison for error detection*

Group	Adjusted Mean	SE	95% CI	
Control	5.24	0.15	[4.94, 5.55]	
GAIIM	7.32	0.15	[7.01, 7.62]	
Group Comparison	Mean Difference	SE	P-value	95% CI
Control--GAIIM	-2.071	0.216	< .001	[-2.500, -1.642]

After controlling for pretest error-detection scores, the instructional group had a statistically significant effect on posttest error-detection scores, $F(1, 83) = 92.186$, $p < .001$, partial $\eta^2 = .526$. Students in the GAIIM group had an adjusted posttest mean of 7.32 ($SE = .15$) for error-detection of algebraic errors, which was significantly higher than the adjusted posttest mean of 5.24 ($SE = .15$) for students in the control group (see [Table 8](#)). The adjusted group comparison showed a statistically significant mean difference of 2.07 points between the groups, $p < .001$, 95% CI [-2.500, -1.642].

RQ2b: What is the effect of GAIIM on students' ability to correctly resolve errors in AI-generated algebraic solutions?

To compare posttest error-resolution scores between the GAIIM and control groups while controlling for pretest error-resolution scores, a one-way analysis of covariance (ANCOVA) was conducted. From [Table 9](#), the overall ANCOVA model was found to be significant, $F(2, 83) = 12.748$, $p < .001$. This model accounted for 23.5% of the variance in the posttest error-resolution scores of the students in the study, with $R^2 = .235$ and adjusted $R^2 = .217$. The pretest error-resolution score was found to be a significant covariate, $F(1, 83) = 18.130$, $p < .001$, partial $\eta^2 = .179$.

Table 9*Analysis of covariance (ANCOVA) for posttest error resolution scores*

Source	SS	df	MS	F	P-value	Partial η^2
Pretest Error Resolution	20.862	1	20.862	18.130	< .001	.179
Group	10.279	1	10.279	8.932	.004	.097
Error	95.510	83	1.151			
Corrected Model	29.339	2	14.669	12.748	< .001	.235

Model fit: $R^2 = .235$; Adjusted $R^2 = .217$.**Table 10***Adjusted posttest means and group comparison for error resolution*

Group	Adjusted Mean	SE	95% CI	
Control	6.39	0.16	[6.06, 6.71]	
GAIIM	7.08	0.16	[6.75, 7.41]	
Group Comparison	Mean Difference	SE	P-value	95% CI
Control – GAIIM	-0.693	0.232	.004	[-1.154, -0.232]

The analysis of error-resolution ability, after controlling for pretest scores, revealed that the instructional group had a statistically significant effect on posttest error-resolution scores, $F(1, 83) = 8.932$, $p = .004$, partial $\eta^2 = .097$. The posttest scores of error-resolution abilities for the students in the GAIIM instructional group had an adjusted posttest mean of 7.08 ($SE = .16$) and were significantly higher than that of the control group students with an adjusted posttest mean of 6.39 ($SE = .16$) ([Table 10](#)). The adjusted group comparison showed a statistically significant mean difference of 0.69 points between the groups, $p = .004$, 95% CI [-1.154, -0.232]. This finding indicated a medium effect for the GAIIM intervention, partial $\eta^2 = .097$.

Observed effect-size patterns for detection and resolution skills

The estimated group effect for error detection was partial $\eta^2 = .526$, whereas the estimated group effect for error resolution was partial $\eta^2 = .097$. Pearson's correlation showed a weak, nonsignificant relationship between posttest error detection and error resolution, $r = .083$, $p = .448$, indicating that better error detection was not necessarily associated with better error resolution. Numerically, the estimated effect was larger for error detection. However, because the effect sizes were obtained from separate outcome models, their numerical difference should not be interpreted as evidence that the effect of GAIIM on error detection was statistically greater than its effect on error resolution. The observed pattern was consistent with stronger performance in error detection, while successful error correction also required procedural and conceptual knowledge. The qualitative findings provided further context for this pattern, as several interviewed reported being able to identify discrepancies while experiencing difficulty reconstructing correct solutions.

QUALITATIVE FINDINGS

Thematic analysis was applied to the collected qualitative data in accordance with the six phases proposed by Braun and Clarke (2006). This method allowed the identification of patterns in how students described their engagement with AI-generated feedback during algebraic problem-solving. Coding was initially aimed at detecting interactions with AI-generated feedback and difficulties related to the verification of students' solutions. Initial codes included stepwise comparison, discrepancy detection, iterative revision, explanation complexity, automation dependency, and correction difficulty. The analysis was organized according to the two qualitative research questions. Three themes were developed in relation to students' engagement with AI-generated feedback during verification (RQ3), while four themes captured the challenges students encountered when using AI as a verification rather than an answer-oriented tool (RQ4).

RQ3: How do undergraduate students engage with AI-generated feedback during the verification of their algebraic problem-solving processes?

Regarding the third research question, the themes that were generated were refined and explained to focus on how students interacted with feedback from the artificial intelligence system while doing verification activities. It is clear from the findings that the interaction between the two was a process that was characterized by evaluation, development, and analysis of feedback from the AI system. The three main themes identified are: (i) stepwise verification of solutions, (ii) iterative engagement with AI feedback, and (iii) conditional trust in AI-generated responses.

Theme 1: Stepwise verification of solutions

From the study, verification takes place systematically and in stages as part of the algebra problem-solving process. Instead of verifying based on the answer alone, students compare their work with that of the AI until they notice a difference between the two. In this way, problem-solving becomes more focused on the process rather than the result.

For example, one participant noted:

"Instead of looking only at the final answer, I compare every step of my solution with the AI's explanation. When I notice that one of my steps is different, I stop and try to understand why before I continue. That helps me identify exactly where I made the mistake." (P3)

This response illustrates a deliberate verification strategy in which the student decomposes the solution into discrete steps and evaluates each component independently. This behaviour is consistent with a monitoring process in which the student checks the consistency of their reasoning against an external reference.

Similarly, another student noted:

"Sir, I do not look at the final answer first because I want to know where my own reasoning starts to differ from the AI. I compare each step carefully, and when I find a difference, I go back to understanding why my approach is different before I continue." (P7)

This suggests that the student focuses on comparing the reasoning process rather than considering only the final answer. The response shows that the participants attended to differences within individual solution steps rather than focusing only on the final answer.

In addition, one participant stated:

"Sir, when my solution matches the AI at each step, I become more confident that I am using the correct method. It reassures me that my reasoning is on the right track before I even reach the final answer." (P2)

This response shows that the participants described checking the solution continuously during problem-solving rather than only at the end. The response illustrates continuous checking of the solution process rather than

verification only at the final-answer stage. The theme shows how students used AI-generated solutions as reference points for stepwise comparison within the GAIIM activities.

Theme 2: Iterative Engagement with AI Feedback

The findings indicate that students engage with AI feedback cyclically and recursively, repeatedly revising their solutions in response to discrepancies identified during verification. AI is not used as a static reference point but as part of an ongoing process of refinement.

For example:

“Whenever my answer does not match the AI’s response, I go back through each step instead of starting again. Sometimes I discover that one mistake affects the next step, so I keep checking until I understand where the error began.” (P6)

This response reflects a recursive verification process, where the student re-engages with earlier steps to identify and correct inconsistencies.

Another participant noted:

“After correcting one step, I compare my solution with the AI again because one mistake can affect the rest of the solution. I keep checking until every step follows the correct mathematical process.” (P5)

This illustrates a layered verification strategy, where each correction triggers further evaluation. The student does not assume that fixing one error resolves the entire solution, demonstrating awareness of the interconnected nature of algebraic reasoning.

Similarly:

“I often go back more than once because one mistake can affect the rest of the solution. As I correct each step, I discover other errors, so I continue checking until the whole solution makes sense.” (P9)

This highlights the iterative nature of problem-solving, where correcting one error may lead the student to identify additional problems in the solution. This response illustrates repeated checking and revision following interaction with the feedback. Taken together, these responses describe a repeated process of checking and revising solutions following interaction with AI-generated feedback.

Theme 3: Conditional trust in AI-generated responses

The findings reveal that students do not uniformly accept AI-generated outputs as authoritative. Instead, they demonstrate conditional trust, where AI responses are evaluated against their own reasoning and prior knowledge. For example:

“Even if the AI provides an answer immediately, I don’t accept it automatically. I first compare it with my own reasoning to see whether the explanation makes mathematical sense before deciding whether to change my solution.” (P4)

This suggests deliberate evaluation of the AI-generated response rather than its automatic acceptance. The emphasis on “making sense” shows that the participant considered the reasoning before accepting the AI-generated response.

Another participant stated:

“Sometimes the AI gives a solution that is different from mine, so I do not accept it immediately. I first compare its explanation with my own reasoning to understand why the difference occurred before deciding which approach is correct.” (P8)

This suggests that discrepancies between the students’ reasoning and AI output prompted further comparison of the two solutions. Rather than rejecting or accepting the AI response outright, the student described comparing the two solutions before deciding which approach was appropriate.

In addition, another participant noted:

“I trust the AI because it usually provides helpful explanations, but I still compare its solution with my own. I know that AI can also make mistakes, so I only accept its answer after I am convinced that the mathematical reasoning is correct.” (P1)

This reflects a balanced stance toward AI, where trust is accompanied by verification. The recognition that AI can produce errors shows that the participant did not always accept AI-generated responses uncritically. These responses show that some participants did not accept AI-generated solutions automatically and instead compared them with their own reasoning before deciding whether to accept or revise a solution.

RQ4: What challenges do students encounter when using AI as a verification tool rather than an answer-generating tool within the GAIIM framework?

Theme 1: Difficulty interpreting AI-generated explanations

A major challenge identified was students’ difficulty in interpreting AI-generated explanations, particularly when reasoning involved multiple symbolic transformations.

For example:

“Sometimes the AI explains the solution using too many mathematical steps or symbols, and I struggle to understand how it arrived at the answer. When that happens, it becomes difficult for me to verify whether my own reasoning is correct.” (P2)

This indicates that even when an AI-generated answer appears correct, difficulty following its intermediate steps can make verification challenging for students.

Another participant explained:

“Although the AI explains the solution step by step, some of the transitions are not clear to me. I sometimes know what the next step is, but I do not understand why it was taken, which makes verification difficult.” (P10)

This suggests that difficulty following intermediate reasoning may limit the depth of verification students can undertake.

Similarly:

“The answer may be correct, but I sometimes find the explanation difficult to follow. When I cannot understand the reasoning behind the steps, it becomes harder for me to use feedback to improve my own solution.” (P6)

This suggests that correctness alone is insufficient for learning if the reasoning is not accessible. Difficulty interpreting the explanations may limit students’ ability to use AI-generated solutions effectively for verification.

Theme 2: Procedural difficulty in error correction

Although several participants reported being able to detect errors, they also described difficulty correcting them, particularly when reconstructing the appropriate mathematical procedure.

For example, a participant noted:

“Sometimes I can clearly see that my solution is wrong after comparing it with the AI, but I still struggle to know which mathematical idea I misunderstood or how to correct the step.” (P5)

This highlights a gap between recognition and resolution, suggesting that error detection does not automatically translate into correction.

Another participant stated:

“I can tell that I have made a mistake after comparing my solution with the AI, but I often struggle to identify exactly which step I should change or how to correct it without creating another error.” (P3)

This reflects uncertainty in identifying the exact point of intervention within the solution process. The response suggests difficulty identifying the procedural knowledge required to address the inconsistency.

In addition, a participant noted:

“The AI helps me see the correct way to solve the problem, but I still have difficulty transferring that approach to my own solution. I often need more time to understand how the explanation applies to the mistake I made.” (P7)

This indicates difficulty in transferring observed solutions into personal problem-solving strategies. The response shows that identifying a discrepancy did not necessarily enable the student to reconstruct a correct solution. The participants’ accounts suggest that successful correction required additional procedural and conceptual knowledge. The apparently contrasting accounts reflected variability in participants’ reported AI-use behaviours rather than mutually exclusive positions. Some participants described verification-oriented practices while also acknowledging occasions of greater reliance on AI. For example, P1 explicitly described initially accepting AI-generated solutions without verification but later recognizing the need to compare the AI’s reasoning with personal solutions. Similarly, P4 and P8 reported evaluating AI-generated responses in some situations while also acknowledging occasions when speed or confidence in the AI response reduced the extent of verification. These accounts suggest that participants’ engagement with AI varied across situations and, for some participants, across stages of their experience with GAIIM.

Theme 3: Automation reliance

The findings reveal that some students continue to rely on AI-generated solutions without engaging in verification, reflecting the persistence of answer-oriented learning behaviors.

For example:

“At the beginning, I relied on the AI because I believed its answers were always correct. I would copy the solution without checking it, but later I realised that I needed to compare its reasoning with my own to understand the mathematics” (P1)

This response illustrates an initially answer-oriented use of AI in which the participant reported copying the generated solution without verification.

Another participant noted:

“If the AI provides the answer immediately, I do not always take the time to compare it with my own work. I tend to trust the response because it is fast and seems reliable, even though I know I should verify it.” (P4)

This response shows that the participant sometimes prioritized obtaining a quick answer over comparing the AI-generated solution with their own work.

Similarly:

“There are times when I trust the AI more than myself. If its answer is different from mine, I usually think I made the mistake, so I do not spend much time checking whether my reasoning was actually correct.” (P8)

This response shows that the participant sometimes gave greater weight to the AI-generated solution than to their own reasoning, which could limit careful verification.

Theme 4: Difficulty transitioning to verification-based use

Students also reported difficulty shifting from using AI primarily for obtaining answers to using it for verification and comparison.

For example:

“Before using the GAIIM approach, I mostly used the AI to get answers. Learning to use it for verification was more challenging because I had to focus on understanding the reasoning behind each step instead of just looking at the final solution.” (P6)

This shows that the participant experienced verification-based use as more demanding than obtaining a final answer directly.

Another participant explained:

“I was used to asking the AI for the final answer, so changing to checking every step took much more effort. I had to think carefully about my own reasoning and compare it with the AI instead of simply accepting the solution.” (P9)

This suggests that verification required sustained attention and repeated comparison, which some students found demanding.

Similarly, a participant stated that:

“Getting the answer from the AI is quick, but verifying the solution requires more effort. I have to think through every step instead of simply accepting the final answer, which is more challenging but helps me understand the mathematics better.” (P2)

This highlights the contrast between obtaining an answer directly and engaging in the more effortful process of checking each step. Collectively, these responses indicate that adopting a verification-oriented approach requires a shift in learning behavior, supported by structured instructional guidance such as the GAIIM framework.

Integration of quantitative and qualitative findings

The qualitative findings helped explain the quantitative patterns rather than merely confirm them. Quantitatively, GAIIM students demonstrated higher adjusted algebraic problem-solving, error-detection, and error-resolution performance. Qualitatively, participants described comparing solution steps, identifying discrepancies, and revising their reasoning, providing a plausible explanation for the stronger performance observed in the GAIIM group. However, these accounts represent students' reported experiences and should not be interpreted as direct evidence of underlying cognitive change.

Integration was particularly informative for error detection and resolution. Although GAIIM significantly improved both outcomes, the estimated effect was larger for detection (partial $\eta^2 = .526$) than resolution (partial $\eta^2 = .097$), while their posttest correlation was weak and nonsignificant, $r = .083$, $p = .448$. Participants similarly reported that recognising a discrepancy did not always mean knowing how to correct it. Thus, the qualitative evidence explains why detection and resolution may represent related but distinguishable verification processes: comparison can reveal an error, whereas correction additionally requires sufficient mathematical knowledge to reconstruct a valid solution.

The strands also revealed qualification rather than complete convergence. While some participants described deliberate verification of AI responses, others reported occasions of automation reliance or greater confidence in AI-generated solutions than in their own reasoning. GAIIM therefore did not produce uniformly critical AI use across all situations. Because interviewees were not selected according to individual quantitative performance profiles, these integrated interpretations apply at the outcome level rather than as participant-level explanations.

DISCUSSION

After controlling for pretest scores, students in the GAIIM group demonstrated significantly higher posttest algebraic problem-solving performance than students in the control group. This finding is consistent with the positive relationships between AI-related factors and mathematical learning reported by Akendita et al. (2025) and Boateng et al. (2026). However, because those studies employed non-experimental designs, their findings provide associative rather than causal evidence. Previous research has examined patterns of error identification and correction within mathematical problem-solving (Ahzan et al., 2022; Campanilla & Mendoza, 2024). The present study extends this line of inquiry by examining error detection and error resolution separately within an AI-supported verification context. Although the estimated effect size was numerically larger for error detection than for error resolution, the present analyses did not directly test whether the two effects differed statistically. This pattern is consistent with previous research concerning monitoring processes in AI-assisted learning environments

(Contel & Cusi, 2025). The findings further suggest that successful error detection did not necessarily correspond with successful error resolution.

The qualitative findings provide further insight into how students engaged with AI-generated feedback during verification tasks. Stepwise comparison allowed students to detect discrepancies in their calculations. This pattern is consistent with Contel and Cusi's (2025) discussion of the role of AI-generated feedback in supporting monitoring during mathematical problem-solving. However, the qualitative findings showed that several participants experienced difficulty correcting errors even after identifying discrepancies in their solutions. These findings suggest that participants experienced different levels of difficulty in identifying and correcting errors, with the qualitative accounts indicating that correction often required additional procedural and conceptual knowledge.

This interpretation is consistent with studies that distinguish between error identification and error correction as related but different aspects of mathematical verification (Kirsten & Greefrath, 2023; Sommerhoff et al., 2025). In the present study, the quantitative and qualitative findings provide insight into how students experienced error detection and error resolution within an AI-assisted verification environment. More specifically, the findings suggest that structured verification can support students in identifying errors, whereas successful correction also depends on the procedural and conceptual knowledge required to reconstruct a valid solution.

From a social constructivist perspective, the structured GAIIM activities surrounding the AI-generated solutions function as instructional scaffolds by guiding students through comparison, justification, correction, and reflection. The AI-generated solution itself primarily serves as an external reference against which students evaluate their own reasoning. The qualitative findings further support previous evidence that students require pedagogical guidance to engage critically with AI rather than relying on AI-generated answers without verification (Asamoah & Amarteifio, 2025; Atuahene & Boateng, 2026; Zacharis & Papadakis, 2025). The qualitative findings indicate that some students initially approached AI primarily as an answer-generation tool rather than as a verification tool, highlighting the importance of explicit pedagogical guidance during AI-supported learning. In addition to the performance outcomes, the qualitative findings provide insight into how participants described their engagement with verification activities within GAIIM. These findings suggest that error correction may require additional instructional support beyond identifying discrepancies in AI-generated and student-generated solutions. However, the qualitative findings also indicate that GAIIM did not result in uniformly verification-oriented AI use, as some participants continued to report difficulty evaluating AI explanations or showed greater reliance on AI-generated responses in some situations. This suggests that structured AI integration may support verification practices without necessarily eliminating students' dependence on AI or difficulties in evaluating its reasoning.

CONCLUSION

The present study examined how verification-focused AI can be integrated into undergraduate students' algebraic problem-solving. Specifically, the study provides evidence that error detection and error resolution can be examined as related dimensions of students' algebraic verification performance. In the context of this study, the estimated group effect was numerically larger for error-detection performance than for error-resolution performance. The study provides evidence concerning the use of GAIIM as a structured approach to AI-supported verification in undergraduate algebraic problem-solving within the context investigated. The findings further suggest that successful verification involves both recognising problematic reasoning and being able to generate an appropriate mathematical correction. Consequently, the study provides insight into how structured AI-supported verification may contribute to undergraduate students' algebraic problem-solving beyond the production of final answers.

Implications

The findings from this study also have implications for the pedagogical use of generative AI in higher education. First, educators may design instructional activities in which AI-generated solutions are used primarily for verification and comparison rather than as substitutes for students' independent problem-solving. To illustrate, in the above example, the student would first solve the problem on his or her own, then compare it to the AI-generated solution for the same problem. The student would then identify any discrepancies between the two solutions and justify any resulting modifications to his or her original solution. Such activities may encourage students to evaluate AI-generated solutions rather than accept them automatically. Second, the observed patterns for error detection and error resolution suggest that both skills require explicit instructional attention. Structured comparison with AI-generated solutions may support students in recognizing errors, while additional instructional support may be needed to strengthen the procedural and conceptual knowledge required for correcting them. Mathematics instructors may therefore consider incorporating structured verification tasks that explicitly distinguish error identification from error correction. Findings from this study may inform mathematics educators interested in incorporating structured verification activities into undergraduate students' use of generative AI.

Further research is required before the findings are extended to other STEM disciplines, educational levels, or learning contexts.

Limitations and future research

The research was carried out using a sample of mathematics education undergraduate students from one public university. In addition, the experimental design used in this study was based on normal classroom environments. Although the participants were randomly allocated to the experimental and control conditions, the study involved students drawn through convenience sampling from a single Mathematics Education class at one public university. Consequently, the findings should be interpreted within this institutional and disciplinary context and should not be generalized uncritically to other populations or educational settings. Because participants belonged to the same academic programme, interaction between students across the two instructional conditions outside the scheduled sessions could not be completely controlled. Consequently, possible contamination between conditions cannot be entirely excluded. Students' independent use of generative AI outside the scheduled instructional sessions was also not directly monitored; therefore, possible exposure to AI-supported learning beyond the planned intervention conditions cannot be completely ruled out. Moreover, GAIIM combined AI-generated comparison with structured justification, correction, instructor prompting, and reflection. The present design therefore does not permit the observed effects to be attributed exclusively to the AI component of the intervention. Finally, the qualitative phase involved a relatively small number of participants and was intended to provide explanatory depth rather than statistical generalization. The parallel forms were not statistically equated, and qualitative participants were not selected by quantitative performance profiles, limiting form-equivalence claims and participant-level mixed-methods integration. Future research should examine the Guided AI Integration Model across different STEM disciplines, educational levels, and learning contexts. Longitudinal studies are also needed to determine whether extended exposure to verification-oriented AI instruction produces sustained improvements in procedural error-resolution skills. Future studies should further investigate how learner characteristics, such as AI literacy, prior mathematical knowledge, and metacognitive skills, influence the effectiveness of verification-oriented AI learning environments.

Acknowledgement

The authors gratefully acknowledge all interview and focus group participants for their valuable time, insights, and contributions to this study.

Findings

The authors received no financial support for the research and/or authorship of this article.

Ethical statement

This study received ethics exemption from the relevant institutional research ethics committee, consistent with institutional guidelines for qualitative research involving voluntary adult participants in non-sensitive educational contexts. Written informed consent was obtained from all participants prior to data collection. Participants were fully informed of the study's objectives, data handling procedures, and their right to withdraw at any time without consequence. All data were anonymized, and participants are represented in the findings without personal or institutional identifiers.

Competing interests

The authors declare no competing interests.

Author contributions

I.D: Developed conceptualization, introduction, methodology development, data collection, analysis, interpretation, and manuscript writing. E.L.: Contributed to conceptualization, data analysis, and interpretation. FOB: Contributed to data analysis, interpretation, and methodology. M.O.T: Developed conceptualization, introduction, methodology development, and data collection.

Data availability

The qualitative data supporting the findings of this study are not publicly available in order to protect participant confidentiality. Reasonable requests for further information may be directed to the corresponding author, subject to ethical considerations.

AI disclosure

No generative AI tools were used in writing this manuscript. Conventional tools, including spell-check, grammar software, and Mendeley for reference management, were used for editing and formatting only. All analysis, interpretation, and writing were completed solely by the authors.

Biographical sketch

Isaac Davor is an M.Phil. student in Mathematics Education at the University of Skills Training and Entrepreneurial Development (USTED), Kumasi, Ghana. His research focuses on mathematics education, self-regulated learning, metacognition, and technology-enhanced learning. He is particularly interested in examining how students' cognitive and metacognitive processes influence mathematical problem-solving and achievement in technology-supported learning environments.

Enerst Larbi is the Head of the Department of Mathematics at the University of Skills Training and Entrepreneurial Development (USTED), Kumasi, Ghana. His academic and research interests include mathematics education, mathematical problem-solving, mathematics teaching and learning, and curriculum development.

Francis Obene Boateng is an Associate Professor of Applied Mathematics at the University of Skills Training and Entrepreneurial Development (USTED), Kumasi, Ghana, where he serves as Dean of Students Affairs. He holds a Ph.D. in Applied Mathematics. His research interests include computational mathematics, mathematical modelling, and the integration of information and communication technology in mathematics education.

Michael Osei Tawiah holds an M.Phil. in Mathematics from the University of Skills Training and Entrepreneurial Development (USTED), Ghana. His research interests include mathematical modelling, computational mathematics, mathematics education, and mathematical problem-solving.

Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and do not necessarily reflect the views of Lectito Publications and/or the editor(s). Lectito Publications and/or the editor(s) disclaim responsibility for any injury to persons or property resulting from any ideas, methods, instructions, or products referred to in the content.

REFERENCES

- Ahzan, Z. N., Simarmata, J. E., & Mone, F. (2022). Using Newman error analysis to detect students' error in solving junior high school mathematics problem. *Jurnal Pendidikan MIPA*, 23(2), 459-473. <https://doi.org/10.23960/jpmipa/v23i2.pp459-473>
- Akendita, P. A., Boateng, F. O., Arthur, Y. D., & Banson, G. M. (2025). Investigating the effect of perception of artificial intelligence on pre-service teachers' mathematics performance in Ghana: the moderating role of interest. *Artificial Intelligence in Education*, 1(1), 93-113. <https://doi.org/10.1108/AIIE-09-2024-0004>
- Asamoah, M. K., & Amarteifio, J. (2025). Overcoming barriers and ethical challenges in implementing intelligent tutoring systems in Ghanaian higher education. *Discover Education*, 4(1), 179. <https://doi.org/10.1007/s44217-025-00593-5>
- Atuahene, E., & Boateng, F. O. (2026). Mathematics teachers' awareness, perceptions, and challenges in using ChatGPT. *Discover Education*, 5(1), 88. <https://doi.org/10.1007/s44217-025-01083-4>
- Baumgartner, V., Dagwati, S., Trninic, D., Akveld, M., Caspar, A., Hungerbühler, N., & Kapur, M. (2025). Problem-solving before instruction for learning linear algebra in university mathematics. *Instructional Science*, 53(6), 1573-1602. <https://doi.org/10.1007/s11251-025-09709-8>
- Boateng, O. F., Davor, I., & Appiah Manu, C. (2026). Innovative teaching methods supported by artificial intelligence and students' mathematical problem-solving: the mediating role of student engagement. *EIKI Journal of Effective Teaching Methods*, 4(1). <https://doi.org/10.59652/7tsj5730>
- Boateng, F. O., Davor, I., Larbi, E., & Boateng, E. A. (2026). *Scaffolded and unscaffolded generative AI use in mathematics teacher education: Effects on pre-service teachers' instructional design competence* [Preprint]. Research Square. <https://doi.org/10.21203/rs.3.rs-10778501/v1>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. <https://doi.org/10.1191/1478088706qp063oa>
- Campanilla, N. S., & Mendoza, C. R. (2024). Error pattern analysis of the mathematics problem solving of grade 10 learners. *Journal of Social, Humanity, and Education*, 4(4), 245-262. <https://doi.org/10.35912/jshe.v4i4.2125>
- Cevikbas, M., & Kaiser, G. (2026). The use of large language models to solve mathematical modeling problems: Preservice mathematics teachers' use practices, perceived affordances and challenges, and trustworthiness

- judgments of AI-generated outputs. *ZDM—Mathematics Education*, 1-17. <https://doi.org/10.1007/s11858-026-01786-4>
- Contel, F., & Cusi, A. (2025). Investigating the role of ChatGPT in supporting metacognitive processes during problem-solving activities. *Digital Experiences in Mathematics Education*, 11(1), 167–191. <https://doi.org/10.1007/s40751-024-00164-7>
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE.
- Davor, I. (2026). Generative AI-supported instruction as a cognitive scaffold: Effects on senior high school students' geometric reasoning and proof construction. *Frontiers in Education*, 11, 1869961. <https://doi.org/10.3389/feduc.2026.1869961>
- Davor, I., & Boateng, F. O. (2026). AI-enhanced pedagogical practices and mathematical language proficiency in STEM education. *American Journal of STEM Education*, 22, 237–262. <https://doi.org/10.32674/6h0e3376>
- Davor, I., Larbi, E., & Boateng, F. O. (2026). Artificial intelligence-supported learning and higher-order cognitive outcomes: The mediating role of metacognitive self-regulation. *Frontiers in Education*, 11, 1855869. <https://doi.org/10.3389/feduc.2026.1855869>
- Davor, I., Boateng, F. O., Osei-Tutu, P., Osei-Bosompem, A., Patience, A. S., & Adusei, L. A. (2026). Examining the mediating roles of digital literacy and critical thinking in the relationship between ChatGPT use and university students' mathematical problem-solving efficacy. *Journal of Mathematics, Science and Technology Education*, 3(1), 1–15. <https://doi.org/10.12973/jmste.3.1.1>
- Gimpel, H., Hall, K., Decker, S., Eymann, T., Lämmermann, L., Mädche, A., Röglinger, M., Ruiner, C., Schoch, M., Schoop, M., Urbach, N., & Vandirck, S. (2023). *Unlocking the power of generative AI models and systems such as GPT-4 and ChatGPT for higher education: A guide for students and lecturers* (Discussion Paper No. 02-2023). University of Hohenheim.
- Gqoli, N., & Baidoo, J. (2025). The integration of artificial intelligence in mathematics education: towards developing a conceptual framework for understanding student perspectives in higher education. *E-Journal of Humanities, Arts and Social Sciences*, 2660–2674. <https://doi.org/10.38159/ehass.202561031>
- Habtmu, S. B., Mulugeta, A. A., & Mulugeta, W. G. (2022). The effect of cooperative problem-solving method on students' motivation towards learning algebra. *Pedagogical Research*, 7(2), em0123. <https://doi.org/10.29333/pr/11906>
- Hornberger, M., Bewersdorff, A., & Nerdel, C. (2023). What do university students know about Artificial Intelligence? Development and validation of an AI literacy test. *Computers and Education: Artificial Intelligence*, 5, 100165. <https://doi.org/10.1016/j.caeai.2023.100165>
- Joseph, O. B., & Uzondur, N. C. (2024). Integrating AI and machine learning in STEM education: Challenges and opportunities. *Computer Science & IT Research Journal*, 5(8), 1732–1750. <https://doi.org/10.51594/csitrj.v5i8.1379>
- Julius, E. (2022). The relationship between self-concept and problem-solving skills on students' attitude towards solving algebraic problems. *Contemporary Mathematics and Science Education*, 3(2), ep22020. <https://doi.org/10.30935/conmaths/12509>
- Jupri, A., & Drijvers, P. H. M. (2016). Student difficulties in mathematizing word problems in algebra. *Eurasia Journal of Mathematics, Science and Technology Education*, 12(9), 2481–2502. <https://doi.org/10.12973/eurasia.2016.1299a>
- Kazadi Tshikolu, R., Kule Mukuhi, D., Nzalalemba Kabwangala, T., Ntiaka Muzakwene, J., & Sunda-Meya, A. (2025). AI-enhanced problem-based learning for sustainable engineering education: The AIPL framework for developing countries. *Sustainability*, 17(20), 9038. <https://doi.org/10.3390/su17209038>
- Kirsten, K., & Greefrath, G. (2023). Proof construction and in-process validation – Validation activities of undergraduates in constructing mathematical proofs. *The Journal of Mathematical Behavior*, 70, 101064. <https://doi.org/10.1016/j.jmathb.2023.101064>
- Kusuma, A. P., Waluya, St. B., Rochmad, R., & Mariani, S. (2024). Algebraic thinking profile of pre-service teachers in solving mathematical problems in relation to their self-efficacy. *Eurasia Journal of Mathematics, Science and Technology Education*, 20(11), em2532. <https://doi.org/10.29333/ejmste/15580>
- Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R., & Gerardou, F. S. (2023). Challenges and opportunities of generative ai for higher education as explained by ChatGPT. *Education Sciences*, 13(9), 856. <https://doi.org/10.3390/educsci13090856>
- Özcan, Z. Ç. (2024). Evaluating the impact of the WEI4s instructional approach on middle school students' algebraic problem-solving skills. *Education Sciences*, 14(1), 102. <https://doi.org/10.3390/educsci14010102>
- Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLOS ONE*, 19(5), e0304013. <https://doi.org/10.1371/journal.pone.0304013>

- Polotskaia, E., Fellus, O. O., Cavalcante, A., & Savard, A. (2022). Students' problem solving and transitioning from numerical to relational thinking. *Canadian Journal of Science, Mathematics and Technology Education*, 22(2), 341–364. <https://doi.org/10.1007/s42330-022-00218-1>
- Santos-Trigo, M. (2024). Problem solving in mathematics education: Tracing its foundations and current research-practice trends. *ZDM–Mathematics Education*, 56, 211–222. <https://doi.org/10.1007/s11858-024-01578-8>
- Shulman, L. S., & Carey, N. B. (1984). Psychology and the limitations of individual rationality: Implications for the study of reasoning and learning. *Review of Educational Research*, 54(4), 503–530. <https://doi.org/10.3102/00346543054004503>
- Sivan, E. (1986). Motivation in social constructivist theory. *Educational Psychologist*, 21(3), 209–233. https://doi.org/10.1207/s15326985ep2103_4
- Sommerhoff, D., Kollar, I., & Ufer, S. (2025). *Undergraduate Students' Mathematical Proof Skills: Examining the Impact of Six Cognitive Resources* (pp. 45–71). https://doi.org/10.1007/978-3-031-85004-2_3
- Teresi, J. A., Yu, X., Stewart, A. L., & Hays, R. D. (2022). Guidelines for designing and evaluating feasibility pilot studies. *Medical care*, 60(1), 95–103. <https://doi.org/10.1097/MLR.0000000000001664>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2), 89–100. <https://doi.org/10.1111/j.1469-7610.1976.tb00381.x>
- Xu, W., & Ouyang, F. (2022). The application of AI technologies in STEM education: A systematic review from 2011 to 2021. *International Journal of STEM Education*, 9, Article 59. <https://doi.org/10.1186/s40594-022-00377-5>
- Yelamarthi, K., Dandu, R., Rao, M., Yanambaka, V. P., & Mahajan, S. (2024). Exploring the potential of generative AI in shaping engineering education: Opportunities and challenges. *Journal of Engineering Education Transformations*, 37(Special Issue 2), 439–445. <https://doi.org/10.16920/jeet/2024/v37is2/24072>
- Zacharis, G., & Papadakis, S. (2025). Can AI grade like a human? Validity, reliability, and fairness in university coursework assessment. *Educational Process: International Journal*, 19, e2025591. <https://doi.org/10.22521/edupij.2025.19.591>

APPENDICES

Appendix A

Algebraic Problem-Solving Test (Pretest- Form A)

1. Solve: $\frac{2x}{x-1} - \frac{x}{x+1} = \frac{3}{x^2-1}$
2. Solve: $\ln(x^2 - 5x + 6) = \ln(2x - 3)$
3. Solve: $\begin{cases} y = x^2 - 2x + 3 \\ y = \frac{1}{x-1} \end{cases}$
4. Determine values of k such that: $x^2 + (k - 3)x + k = 0$ has two distinct real roots.
5. Solve: $2^{x+2} = 8^{2x-1}$
6. Solve: $\frac{x^2-4}{x^2-x-6} \leq 0$
7. Solve: $\sqrt{x+3} + x = 3$

Algebraic Problem-Solving Test (Posttest -Form B)

1. Solve: $\frac{3x}{x-2} - \frac{2x}{x+2} = \frac{4}{x^2-4}$
2. Solve: $\ln(x^2 - 7x + 10) = \ln(3x - 4)$
3. Solve: $\begin{cases} y = x^2 - 3x + 2 \\ y = \frac{2}{x-1} \end{cases}$
4. Determine values of k such that: $x^2 + (k - 4)x + k = 0$ has equal roots.
5. Solve: $3^{x+1} = 9^{2x-1}$
6. Solve: $\frac{x^2-9}{x^2-x-12} \geq 0$
7. Solve: $\sqrt{x+5} - x = 1$

Appendix B: Algebraic Verification Tasks (AVT)

Each verification task required students to (a) solve the algebra problem independently, (b) compare their solution with the AI-generated solution, (c) determine whether the AI-generated reasoning was mathematically valid, (d) identify and justify any error detected using relevant algebraic principles, and (e) reconstruct the solution where correction was required.

Task 1: Rational Equation

Solve:

$$\frac{2}{x-1} + \frac{3}{x+1} = \frac{5x}{x^2-1}$$

AI-Generated Solution:

$$\begin{aligned} \frac{2(x+1) + 3(x-1)}{(x-1)(x+1)} &= \frac{5x}{(x-1)(x+1)} \\ \frac{2x+2+3x-3}{(x-1)(x+1)} &= \frac{5x}{(x-1)(x+1)} \\ \frac{5x-1}{(x-1)(x+1)} &= \frac{5x}{(x-1)(x+1)} \\ 5x-1 &= 5x \Rightarrow -1 = 0 \end{aligned}$$

Conclusion (AI): No solution exists.

Task 2: Logarithmic Equation

Solve:

$$\ln(x-2) + \ln(x-4) = \ln(8)$$

AI Solution

$$\ln[(x-2)(x-4)] = \ln(8)$$

$$\begin{aligned} (x-2)(x-4) &= 8 \\ x^2 - 6x + 8 &= 8 \\ x^2 - 6x &= 0 \\ x(x-6) &= 0 \Rightarrow x = 0 \text{ or } x = 6 \end{aligned}$$

Conclusion (AI): $x = 0, 6$

Task 3: Parametric Quadratic

Find the values of k such that: $x^2 - 4x + k = 0$ has exactly one real solution.

AI-Generated Solution:

$$\begin{aligned} b^2 - 4ac &= 0 \\ (-4)^2 - 4(1)(k) &= 0 \\ 16 - 4k &= 0 \Rightarrow k = 4 \end{aligned}$$

Conclusion (AI): The equation has one solution when $k = 4$

Task 4: System of Equations

Solve:

$$\begin{cases} y = x^2 - 1 \\ y = 2x + 3 \end{cases}$$

AI-Generated Solution:

$$\begin{aligned} x^2 - 1 &= 2x + 3 \\ x^2 - 2x - 4 &= 0 \\ (x-2)^2 &= 8 \\ x &= 2 \pm 2\sqrt{2} \end{aligned}$$

Conclusion (AI): Solutions are $x = 2 \pm 2\sqrt{2}$

Task 5: Algebraic Simplification

Simplify:

$$\frac{x^3 - x}{x}$$

$$\text{AI-Generated Solution: } = \frac{x(x^2-1)}{x}$$

$$= x^2 - 1$$

Conclusion (AI): $x^2 - 1$

Appendix C: Scoring Rubric for Verification Tasks

Error Detection (RQ2a)

Score	Criteria	Description
0	No valid verification	Incorrectly accepts an erroneous solution or incorrectly identifies a valid solution as erroneous
1	Partial verification	Correctly judges whether an error is present but provides an incomplete or unclear justification
2	Accurate verification	Correctly determines whether the solution is valid and, where an error exists, precisely identifies and justifies the error

Error Resolution (RQ2b)

Score	Criteria	Description
0	No valid resolution	No correction or incorrect correction
1	Partial resolution	Correct approach but incomplete or computational error
2	Successful resolution	Correct correction leading to valid solution

Note: Error-resolution scoring was applied to tasks requiring mathematical correction. Where the presented solution was mathematically valid, no corrective reconstruction was required.

Appendix D: Semi-Structured Interview Guide

1. How did you use AI-generated feedback when solving algebraic problems?
2. Can you describe how you compared your solution with the AI-generated solution?
3. What steps did you take when your answer differed from the AI-generated response?
4. How did AI help you identify errors in your solution?
5. Were you always able to correct the errors you identified?
6. What difficulties did you experience when trying to understand AI-generated explanations?
7. Did you ever rely on AI-generated solutions without checking them? Why?
8. How did your use of AI change during the learning process?
9. Which aspects of using AI helped you learn the most?
10. What challenges do you face when using AI as a verification tool instead of just getting answers?