Lin et al. *Research and Practice in Technology Enhanced Learning* (2025) 20:34
https://doi.org/10.58459/rptel.2025.20034

Research and Practice in Technology Enhanced Learning

RESEARCH **Free and Open Access**

Enhancing students' authentic mathematical problem-solving skills and confidence through error analysis of GPT-4 solutions

Yu-Feng Lin [1], Euphony Fu-Yu Yang [2*], Jeng-Shin Wu [3], Charles Y. C. Yeh [1], Chang-Yen Liao [1] and Tak-Wai Chan [1]

*Correspondence:
fyyang@mail.cgust.edu.tw
Clinical Competency Center,
Chang Gung University of Science and Technology,
No.261, Wenhua 1st Rd.,
Guishan Dist.,
Taoyuan City 33303, Taiwan
Full list of author information is available at the end of the article

Abstract

Authentic mathematical problems connect mathematics to real-life scenarios, making mathematics learning more meaningful. However, students often find it challenging to comprehend the complexity and extensive textual descriptions of authentic mathematical problems, resulting in a lack of mathematical confidence. This study aims to investigate whether error analysis learning activity of GPT-4 solutions can enhance the skill of fifth-grade students to solve authentic mathematical problems and foster their mathematical confidence. A quasi-experimental design was employed, involving 59 fifth-grade students from a primary school in northern Taiwan. The experimental group engaged in error analysis learning activity of GPT-4 solutions, while the control group received traditional instruction, with both groups using the same teaching materials. Quantitative assessments were conducted through tests on solving authentic mathematical problems and a mathematical confidence scale, complemented by qualitative data collected via semi-structured interviews. The results revealed that the experimental group showed significant improvement in solving authentic mathematical problems, both in pre- and post-test comparisons within the group and in post-test comparisons between groups. Furthermore, the low-achieving students in the experimental group showed a significant improvement in solving authentic mathematical problems compared to the control group. Additionally, the mathematical confidence of both high- and low-achieving students in the experimental group was significantly higher than that of the control group. This study confirms the effectiveness of GPT-4 in mathematics education, offering new teaching strategies and research directions for educators and researchers.

Keywords: GPT-4, Authentic mathematical problem-solving, Error analysis, Mathematical confidence

Introduction

Authentic mathematical problem-solving emphasizes integrating mathematics with real-life, making learning significantly more meaningful for students by illustrating the practical value of mathematics (Reinke et al., 2023; Van den Heuvel-Panhuizen & Drijvers, 2020). This method not only improves students' comprehension of mathematical concepts but also fosters essential 21st-century skills such as critical thinking and creativity (OECD, 2022). However, authentic mathematical problems are inherently complex and laden with extensive textual descriptions, making these problems challenging for students to comprehend and solve (Verschaffel et al., 2020). This is particularly problematic for low-achieving students, who often find these problems overwhelming and disheartening, which adversely affects their confidence in learning mathematics (Mullis et al., 2020). Consequently, educators need to shift from traditional methods of teaching mathematics to incorporating real-life contexts to effectively develop students' cognitive abilities and authentic problem-solving skills (Aulia & Prahmana, 2022; Laurens et al., 2017).

With the advancement of artificial intelligence (AI), numerous studies have begun exploring how to leverage these technologies to adjust teaching strategies and assist students in understanding and managing complex mathematical problems (Baidoo-Anu & Ansah, 2023; Rospigliosi, 2023). For example, utilizing GPT-4 to aid in learning mathematics has been an area of increasing exploration (Bubeck et al., 2023; OpenAI, 2023; Plevris et al., 2023). GPT-4 can provide detailed procedural steps and multiple problem-solving approaches, assisting elementary students in understanding the questions and fostering a variety of strategic skills (Frieder et al., 2023; Gattupalli et al., 2023). However, challenges remain as GPT-4 sometimes generates incomplete strategies or incorrect answers, potentially misguiding students and undermining their confidence (Plevris et al., 2023; Wardat et al., 2023). Overreliance on GPT-4 could also impede the development of critical thinking and reduce opportunities for peer collaboration in problem-solving (Lo et al., 2024).

Conversely, transforming these challenges into learning opportunities can harness the powerful capabilities of AI. For instance, using error analysis activities triggered by incomplete or incorrect responses from GPT-4 can engage students in discussions to analyze their problem-solving processes, identify errors, understand the reasons behind these mistakes, and formulate correct solutions. This is significant because, according to constructivist and metacognitive theories, error analysis can help students reflect on their thinking processes, identify areas for improvement, and deepen their conceptual understanding, critical thinking, and metacognitive skills (Alvidrez et al., 2024; Khasawneh et al., 2023; Kramarski & Zoldan, 2008; Toikka et al., 2024), especially for students with insufficient prior knowledge or lower academic achievements (Begolli et al., 2021; Kshetree et al., 2021).

While many studies have investigated the application of GPT-4 in mathematical error analysis, most have been simulations by researchers focused on assessing GPT-4's problem-solving capabilities and handling of mathematical problems (An et al., 2023; Frieder et al., 2023; Plevris et al., 2023; Shakarian et al., 2023; Supriyadi & Kuncoro, 2023; Zheng et al., 2023; Zong & Krishnamachari, 2023). Other studies have used surveys and interviews to understand the effects of GPT-4 in educational settings (Wardat et al., 2023; Zafrullah et al., 2023). However, direct applications of these methodologies in mathematics education settings are relatively rare, and there has been no prior research combining GPT-4's problem-solving examples with error analysis strategies to enhance students' skills in solving authentic mathematical problems and improving their confidence. Therefore, this study will integrate the advantages of GPT-4 with teaching strategies focused on error analysis (Hwang & Chen, 2023), attempting to use GPT-4 to decompose mathematical text problems, generate examples and explanations as starting points for mathematical discussions (Rong & Mononen, 2022), and guide students in small group interactions to analyze the problem-solving process, identify errors, and explain the reasons behind them, aiming to cultivate students' skills in solving authentic mathematical problems and to analyze the effectiveness of this teaching activity. Additionally, this study will further explore whether the error analysis learning activity using GPT-4 can enhance the problem-solving abilities and confidence of low-achieving students in authentic mathematical contexts.

Summarizing, the research questions of this study are as follows:

(1) Can error analysis learning activity of GPT-4 solutions effectively enhance students' authentic mathematical problem-solving skills?

(2) Can the error analysis learning activity of GPT-4 solutions enhance students' authentic mathematical problem-solving skills more in comparison with traditional instruction?

(3) Can error analysis learning activity of GPT-4 solutions strengthen students' mathematical confidence?

(4) Can the error analysis learning activity of GPT-4 solutions enhance students' mathematical confidence more in comparison with traditional instruction?

(5) What are students' perceived benefits and feelings after participating in the error analysis learning activity of GPT-4 solutions?

Literature review

Authentic mathematical problem-solving

Authentic mathematical problem-solving emphasizes the application of mathematical knowledge and skills in real-life contexts (OECD, 2022). Based on constructivist and situated learning theories, providing students with authentic mathematical problems can foster the development of their imagination and creative thinking (Samritin et al., 2023) and helps them establish rigorous mathematical concepts from everyday situations, thus sparking their interest in mathematics (Aulia & Prahmana, 2022; Niss & Jablonka, 2014). Therefore, authentic mathematical problem-solving holds significant value for mathematics education. For instance, it facilitates students in applying the mathematical knowledge and skills learned in the classroom to daily life scenarios, thereby understanding the connections between mathematics and social, cultural, or interpersonal interactions (Niss & Højgaard, 2011) and its practicality (Nuraina et al., 2021). Furthermore, authentic mathematical problem-solving nurtures critical 21st-century skills such as critical thinking, creativity, communication, and collaboration, which are essential in solving complex mathematical problems (OECD, 2022). Additionally, during the mathematical problem-solving process, students must effectively plan, monitor, and evaluate their problem-solving strategies, thus also serving as a method to develop their metacognitive skills (Roorda et al., 2024).

Despite its importance in mathematics education, implementing authentic mathematical problem-solving poses numerous challenges. For instance, the inherent complexity and extensive textual descriptions of authentic mathematical problems increase the difficulty for students to understand and resolve these problems (Verschaffel et al., 2020). Moreover, when students encounter procedural and conceptual misunderstandings while solving authentic mathematical problems, the lack of timely support or sufficient opportunities from teachers to express ideas or clarify concepts can limit their ability to solve authentic mathematical problems and their confidence in learning (Samritin et al., 2023; Santos-Trigo, 2024). Recent studies suggest that if teachers effectively utilize digital technology resources, they can aid students in understanding mathematical concepts, mastering problem-solving skills, and providing reflections and solutions for specific problems (Larrain & Kaiser, 2022; Santos-Trigo, 2024; Wildgans-Lang et al., 2020).

Applying GPT-4 in mathematics education

Since its release on November 30, 2022, ChatGPT has rapidly gained widespread adoption in the global education sector, highlighting the increasing importance and impact of artificial intelligence tools in education (İpek et al., 2023; Lo et al., 2024; OpenAI, 2022). ChatGPT, a large language model powered by artificial intelligence, offers various instructional applications in mathematics education, including providing personalized feedback, interactive dialogue, curriculum preparation, and assessments (Baidoo-Anu & Ansah, 2023; Sok & Heng, 2023). Studies indicate that these functionalities indeed enhance students' understanding of mathematical concepts and overall learning outcomes (Baidoo-Anu & Ansah, 2023; Rospigliosi, 2023). For instance, Pardos and Bhandari (2023) compared the effectiveness of algebraic prompt generation between ChatGPT and human tutors, and their findings suggest that both ChatGPT and human tutors can generate positive learning outcomes. Wu et al. (2023) discovered that using ChatGPT can alleviate students' fear associated with unsolvable problems, thereby reducing their anxiety. Therefore, effectively utilizing ChatGPT in mathematics instruction can aid students in understanding problem statements (İpek et al., 2023), reducing their fear of solving problems, and enhancing their mathematical performance and confidence.

In March 2023, OpenAI launched GPT-4, whose enhanced logical reasoning capabilities not only provide detailed solution steps and analyses but also more accurately solve complex mathematical problems. Gattupalli et al. (2023) found that GPT-4 can generate detailed steps and problem-solving processes, offering multiple strategies and different methods of approach. These features can provide teachers with clear step-by-step explanations useful in the classroom. Although GPT-4 has demonstrated convincing performance in understanding, explaining, and responding to queries, and can provide immediate problem-solving methods or strategies when students encounter difficulties (Wang et al., 2024), its limitations in reasoning and factual accuracy are noteworthy (Collins et al., 2024). For example, Frieder et al. (2023) noted that GPT-4's capabilities in handling algebra problems are significantly inferior to those of mathematics graduate students, often understanding the problem but failing to provide correct solutions for complex mathematical issues. Additionally, when dealing with complex mathematical word problems, GPT-4 frequently exhibits various errors including comprehension, factuality, specificity, and inference errors (Zheng et al., 2023), which can lead to incomplete problem-solving strategies or incorrect answers (Plevris et al, 2023; Wardat et al., 2023).

This aligns with the research question interpreting and critically assessing answers generated by AI tools correctly to ensure that students can accurately understand and apply mathematical concepts (Lan & Chen, 2024), fostering their mathematical understanding, critical thinking, and creative problem-solving abilities.

Mathematical error analysis

Constructivism emphasizes active participation and interactive learning, whereby learners construct knowledge through practice and reflection (Dillenbourg, 1999). Within this framework, errors are viewed as an integral part of learning; through error analysis, students can reflect on their thought processes, thereby enhancing their understanding of the problems and improving their problem-solving strategies (Rushton, 2018; Schneider & Artelt, 2010). Moreover, error analysis is a crucial teaching strategy for enhancing metacognitive abilities. Metacognition suggests that error analysis provides rich learning opportunities by triggering cognitive conflicts, thus fostering deep learning and comprehension (Khasawneh et al., 2023; Kramarski & Zoldan, 2008). Consequently, integrating constructivist and metacognitive learning theories with error analysis teaching strategies is essential for enhancing students' learning outcomes and problem-solving abilities.

In mathematics education, error analysis encompasses detecting computational errors, procedural mistakes, semantic context misunderstandings, and errors in extracting numerical information (Zhang, 2017). By analyzing and explaining errors, students can clarify what is correct and what is not, which not only promotes understanding but also enhances metacognitive abilities (Richey et al., 2019). From a constructivist learning perspective, learning is the process by which individuals actively construct knowledge, and error analysis activities help learners reconstruct knowledge by reflecting on and correcting their mistakes (Khasawneh et al., 2023). Many studies have highlighted the benefits of using error analysis to enhance students' mathematical problem-solving abilities. For instance, Rushton (2018) showed that analyzing and explaining correct and incorrect problem-solving examples can enhance students' understanding of mathematical concepts and the long-term retention of knowledge. Kshetree et al. (2021) found that addressing misconceptions and errors in the mathematics learning process can significantly enhance problem-solving capabilities on conceptual, procedural, and application levels. Begolli et al. (2021) also demonstrated that combining correct and incorrect examples could improve students' proportional reasoning and probability reasoning abilities, especially among students with insufficient prior knowledge or lower academic achievements.

Although error analysis learning activities have been shown to support deeper conceptual understanding, critical thinking, and metacognition (Alvidrez et al., 2024; Kramarski & Zoldan, 2008; Toikka et al., 2024), many teachers struggle to effectively use error analysis in the classroom. One main reason is that preparing error examples and analyzing mistakes can be too time-consuming (Rushton, 2018). Thus, introducing virtual student error responses or automatically generating them through AI could assist teachers in leading error analysis practices in the classroom (Khasawneh et al., 2023). Additionally, if error analysis is used too frequently in the classroom, it may increase anxiety and stress among students, especially those who already find academics challenging (Alvidrez et al., 2024). This study explores how to use GPT-4 to flip the errors it generates in authentic mathematical problems into materials for teaching error analysis and how to appropriately use error analysis learning activities to encourage students to actively share their understanding and questions through communication and interactive dialogues, thereby enhancing learning outcomes and reducing learning anxiety (Khasawneh et al., 2023; Yarman et al., 2020).

Methods

Participants

This study employed a quasi-experimental research design, involving two fifth-grade classes taught by the same mathematics teacher at a primary school in Northern Taiwan. Purposive sampling methods were used to select the participants. One class served as the experimental group, comprising 28 students (16 boys and 12 girls), who engaged in error analysis learning activity of GPT-4 solutions. The other class functioned as the control group, with 31 students (16 boys and 15 girls), employing traditional teaching methods. In total, 59 fifth-grade students participated in this study. Both classes used identical learning materials. To ensure homogeneity in mathematics academic performance between the two groups before the experiment, the Mann-Whitney U test was conducted by their previous midterm mathematics exams. Results showed no significant difference in the pre-experiment mathematics scores between the two groups (U=349.5, p=.97>.05). Furthermore, to explore the performance differences among students with varying levels of mathematical achievement within this study, students in both the experimental and control groups were divided into high-achievement and low-achievement groups based on the average scores of their midterm mathematics exams.

Materials

Teaching materials play a crucial role in aiding students learning and are an essential component in supporting the development of authentic mathematical problem-solving skills (Putri et al., 2020; Subekti & Prahmana, 2021). Based on the fifth-grade mathematics curriculum, this study selected problems set from authentic mathematical problem books (National Academy for Educational Research, 2020, 2021) covering various themes as the learning materials for both the experimental and control groups. These problems encompassed different life scenarios, such as sports and leisure, society and public, personal and life, health and hygiene, professions and science, and shopping and business activities, with a significant proportion of non-multiple-choice questions. Employing these materials for instruction allowed students to more flexibly apply their learned knowledge and skills and further integrate higher-order thinking skills such as critical thinking, analytical reasoning, systems thinking, as well as creativity and divergent thinking (Bingölbali & Bingölbali, 2021).

To ensure that GPT-4 could generate both correct and incorrect problem-solving processes during the experiment, we pre-tested ten selected authentic mathematical problems by entering them sequentially into different chat windows before the official experiment. The results showed an average correct response rate of 40%, and multiple queries did not significantly alter GPT-4’s outputs. This is consistent with the findings of

Plevris et al. (2023), indicating a certain level of inconsistency and error rate in GPT-4 when handling such mathematical problems.

Experimental procedure

The activity was conducted over ten weeks, with two sessions each week, each session lasting a total of 40 minutes. Before initiating the learning activities, students in the experimental group were divided into heterogeneous groups based on their midterm mathematics test scores, with each group consisting of three to four students. In the first week, the instructor introduced the detailed learning activity process to all students. Subsequently, all students participated in a pre-test, completing the Authentic Mathematical Problem-Solving Test (AMPS Test) and a confidence scale within 60 minutes. Starting from the second week, students were divided into the control group and the experimental group to engage in different learning activities. Students in the control group received traditional instruction: first, students solved problems independently; then the teacher encouraged students to share their solutions; next, the teacher explained the problem-solving strategies, and finally, they checked the answers. Students in the experimental group participated in the error analysis learning activity of GPT-4 solutions. Since these students had already learned how to use GPT-4 for inquiries in their computer science class, they were able to smoothly copy the problem content from the teacher-provided website and make inquiries.

To ensure the reliability and consistency of the experiment, the experimental group and the control group used the same teaching materials, instructional time, and instructor. This design aimed to eliminate other variables that might affect the experimental results, thereby more accurately assessing the impact of the error analysis learning activity of GPT-4 solutions on students' authentic mathematical problem-solving skills and learning confidence.

After completing the ten-week learning course, all students took the post-tests of the AMPS Test and confidence scale to evaluate the effectiveness of the experimental learning. Additionally, semi-structured interviews lasting 5 to 10 minutes each were conducted with students in the experimental group to gain deeper insights into their learning experiences and perceptions. Figure 1 illustrates the experimental procedures of this study.

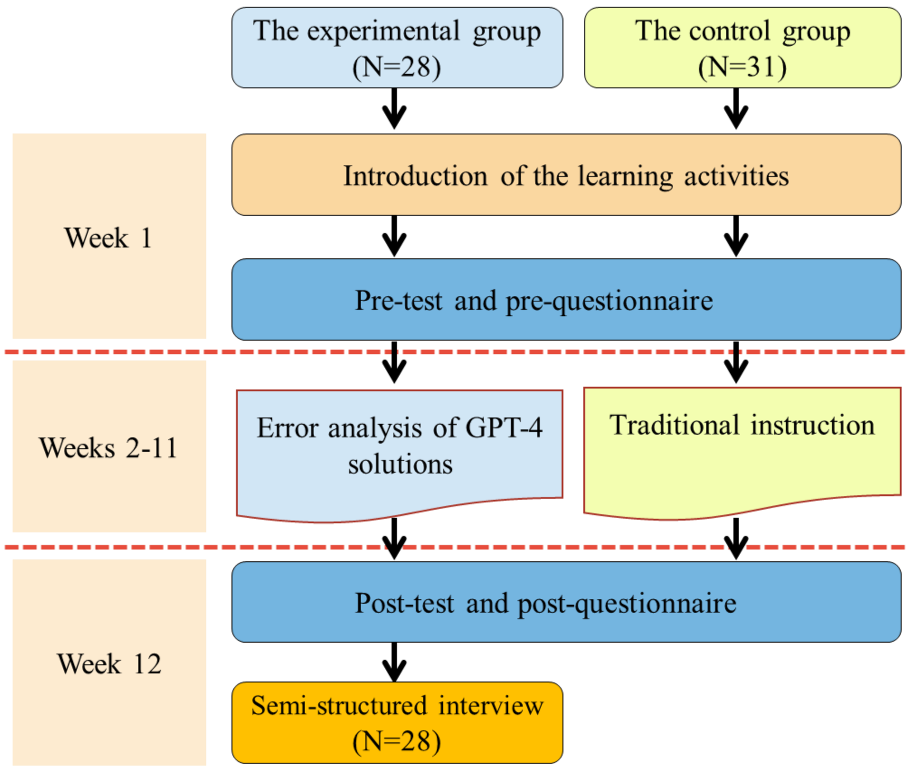

Fig. 1 Experimental procedures

Error analysis learning activity of GPT-4 solutions

To enhance students' skills in solving authentic mathematical problems, this study designed and utilized error analysis learning activity of GPT-4 solutions. Since GPT-4 is a subscription-based service with a dialogue limit of 40 per 3 hours, to maximize cost-effectiveness, we employed four GPT-4 accounts, with each account assigned to 8 students. Moreover, before initiating the learning activities, we pre-set each dialogue window according to the students' classroom seating arrangement, naming them in the format "class_group_seat number," as shown in Figure 2. This allowed students to easily identify and independently use their personal GPT-4 dialogue for solving authentic mathematical problems and conducting subsequent error analysis. Furthermore, GPT-4 provides only one answer each time to avoid interference from previous dialogues, ensuring that students can focus on solving and analyzing the current problem.

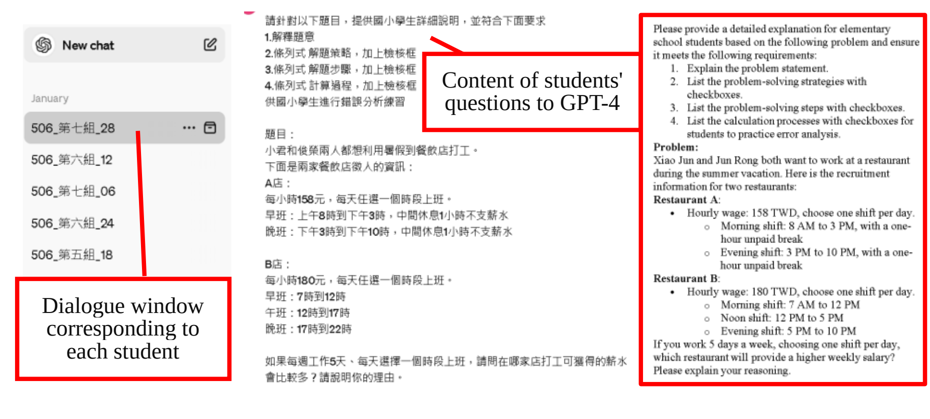

Fig. 2 Online group configuration and GPT-4 prompts diagram

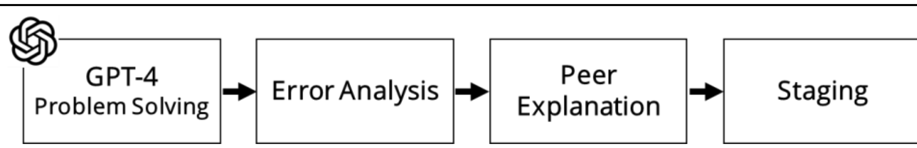

Fig. 3 The process of error analysis learning activity of GPT-4 solutions

The learning activity encompasses four sub-activities: GPT-4 Problem Solving, Error Analysis, Peer Explanation, and Staging. Figure 3 provides a detailed process of the learning activities. The content of the four sub-activities is described as follows:

GPT-4 problem-solving

This activity requires students to pose questions to GPT-4, which then generates explanations of the problem statement, problem-solving strategies, steps, and calculation processes based on authentic situational mathematics problems, aiding students in understanding the key points and processes of problem-solving. Figure 4 shows an example of a solution generated by GPT-4.

Error analysis

In this activity, students are required to carefully review the problem-solving process provided by GPT-4, compare their own explanations with those of GPT-4, and check each step for accuracy. If correct, they mark a check in the box; if incorrect, they annotate and correct the mistake. Any part that is not understood should be marked with a question mark. Figure 5 shows a record of a student's error analysis.

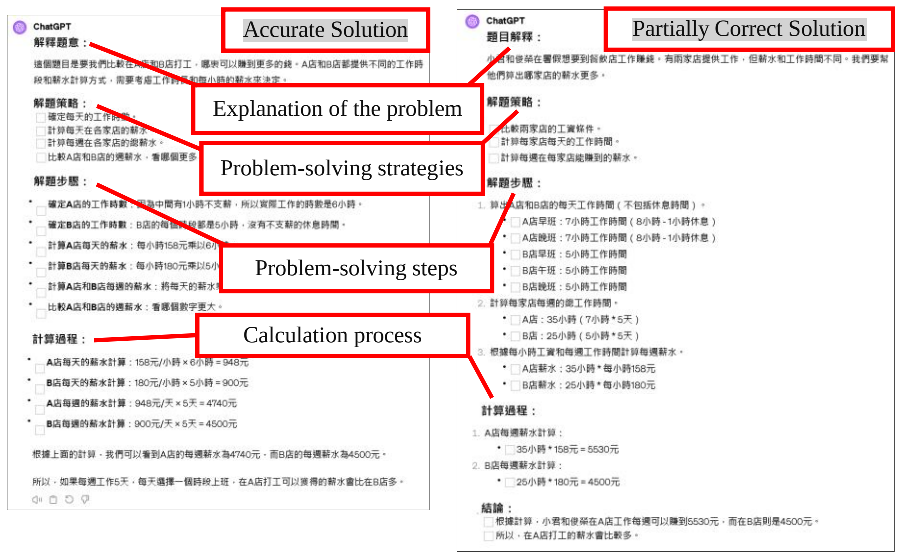

Fig. 4 Example of mathematical problem-solving explanations generated by GPT-4

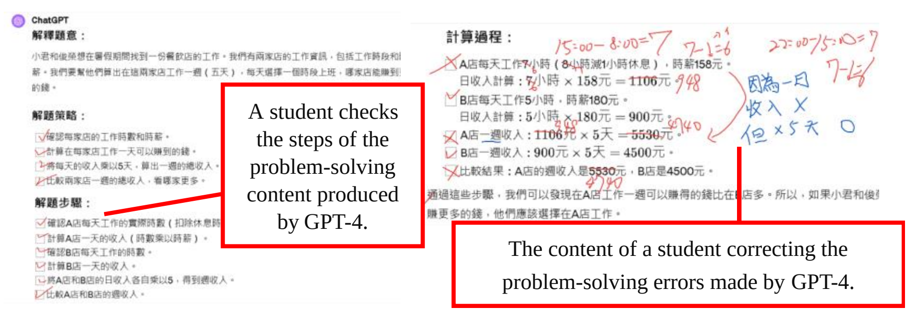

Fig. 5 A record of a student's error analysis

Peer explanation

Within students' groups, members are tasked with explaining to each other the meanings of authentic mathematical problems, GPT-4's problem-solving strategies and steps, as well as the outcomes of their error analysis. They must thoroughly explain the parts they believe are correct to ensure a deep understanding. For the parts identified as incorrect, they need to discuss the reasons behind the mistakes and how to correct them. Should there be any questions, these are to be raised and deliberated upon by group members, collaboratively determining the final results of their error analysis.

Staging

Finally, the teacher encourages each group of students to present their error analysis results to the entire class. They must explain the solutions derived from their error analysis to their classmates and then respond to any questions raised by their peers. Subsequently, the teacher uses examples generated by GPT-4 to demonstrate how to explain mathematical concepts and clarify some of the mistakes made by students to prevent similar errors in the future. Moreover, the teacher prompts students to consider if there are alternative methods to solve the problems, thereby fostering students' problem-solving thinking and facilitating the exchange of mathematical concepts and ideas among peers.

Evaluation tools

Authentic mathematical problem-solving assessment

To assess whether students' authentic mathematical problem-solving skills improved more in comparison with traditional instruction after participating in the error analysis learning activity of GPT-4 solutions, this study selected authentic mathematical problems from problem books published by the National Academy for Educational Research in Taiwan (2020, 2021) for both pre-test and post-test assessment of students' authentic mathematical problem-solving skills. These problems covered various themes, were reviewed by experts, and underwent pilot testing (Wu et al., 2022), demonstrating high discriminative power (average discrimination index of 0.44), making them highly suitable for assessing students' skills to solve authentic mathematical problems. The study utilized real-life scenario problems such as symmetric drawing, exchanging storage cans, making a six-sided die, Tangrams, and assessing the feasibility of triangles as assessment items for the pre-test and post-test, with these problems having an average pass rate of 0.44 and an average discrimination index of 0.42.

To evaluate the effectiveness of the intervention, a Mann-Whitney U test was used to examine the students' performance in authentic mathematical problem-solving. This statistical test was chosen because it is a non-parametric test that does not assume normal distribution of the data, making it suitable for small sample sizes and ordinal data. It helps determine whether there is a significant difference in the improvement of problem-solving skills between students who participated in the error analysis learning activity and those who received traditional instruction.

Mathematical confidence scale

For assessing whether students' mathematical confidence increased after participating in the learning activities, this study adopted the Mathematical Confidence Scale developed by TIMSS 2019 (Martin et al., 2020), the reliability of which was sound (Cronbach's alpha=0.85). The questionnaire consists of nine items, including five reverse-scored items and four positively worded items. The questionnaire of confidences was used to assess students' perceived mathematics abilities. A sample item was "I am good at working out difficult mathematics problems". These items were assessed according to a 4-point Likert scale, ranging from "strongly agree (4)", "agree (3)", "disagree (2)", and "strongly disagree (1)" in this questionnaire.

The items on the scale are closely linked to constructivist and metacognitive learning theories. For instance, the item "I am good at solving difficult math problems" reflects, from a constructivist perspective, the student's successful experiences in solving math problems, thereby constructing their self-perception of ability. Additionally, from a metacognitive perspective, this item prompts students to reflect on and evaluate the strategies and methods they use in problem-solving, further understanding their thinking patterns and problem-solving strategies. Such questionnaire items help us better understand students' self-assessments, situational responses, and metacognitive strategies during the learning process, thereby effectively assessing their learning confidence.

Interview transcription

To gain a deeper understanding of students' learning experiences and perceptions regarding the error analysis learning activity of GPT-4 solutions, this study conducted semi-structured interviews with all participants in the experimental group, totaling 28 students, after the teaching activities concluded. The interview questions were designed based on constructivist learning theory and metacognitive theory to guide students in reflecting on their experiences using GPT-4 and the key skills they utilized to facilitate error analysis. These skills included collaborative problem-solving, critical thinking, mathematical creativity, and metacognition (Almulla, 2023; Fitzsimons & Fhloinn, 2023; Laurens et al., 2017; Nilimaa, 2023; Tay et al., 2024), with each skill corresponding to specific interview questions. For details on the key skills, skill descriptions, and corresponding interview questions, please see Table 1.

(1) Interview methodology: Adhering to the principles of qualitative research interviews, the interviews were conducted one-on-one in a secure environment by an interviewer who was not the students' regular teacher. Students were informed that the interview content would not affect their academic grades, encouraging them to express their thoughts and experiences genuinely. Each student interview was limited to 10 minutes, and care was taken to avoid leading questions and a questioning tone to ensure the accuracy and credibility of the interview data.

(2) Evaluation method: The interview results were assessed by three mathematics teachers with over 15 years of teaching experience, using the Fleiss' kappa statistical method (Fleiss & Cohen, 1973) to evaluate the consistency among raters, to ensure the reliability and validity of the categorization results.

(3) Result analysis: The results were analyzed and interpreted using constructivist learning theory and metacognitive theory to ensure that the findings were grounded in theoretical principles and practical significance.

Table 1 The key skills of AMPS with definitions and interview questions

Key skills and Definition	Interview questions
Understanding Identifying problems and understanding the relationship between conditions and objectives.	How has using GPT-4 changed your understanding of mathematical problems and your ability to grasp key points? What impact has this had on your problem-solving abilities?
Collaborative problem-solving To seek consensus, dialectically discerning and clarifying one's viewpoints.	When solving problems with GPT-4, have you noticed any errors while sharing or demonstrating your solution process with classmates? How did you address these errors?
Critical thinking Identifying errors, analyzing causes, evaluating solutions, confirming outcomes.	What is your opinion on the errors made by GPT-4 during problem-solving? How do you handle and clarify these mistakes?
Mathematical creativity Conceiving and creating new methods to solve problems.	While using GPT-4 for problem-solving, have you attempted any new methods to solve mathematical problems? Can you share your experiences?
Declarative Knowledge Understanding of knowledge and comprehension of concepts or facts.	How does using GPT-4 for problem-solving compare to listening to a teacher explain mathematical concepts? Which method helps you better understand these concepts?
Procedural knowledge Demonstrates how students approach problem-solving and understand solution steps.	When solving problems with GPT-4 and listening to teacher explanations, how do you understand and verify the solution steps?
Conditional knowledge Knowing when to use strategies, select methods, and seek assistance.	When you encounter difficulties, do you prefer to use GPT-4's solutions or seek help from others?
Mathematical confidence Persistence, perseverance, and focus in solving problems without anxiety.	After using GPT-4 for problem-solving, have you become more patient when facing challenging mathematical problems? Has your anxiety decreased, and has your confidence in solving problems increased?

Results

Authentic mathematical problem-solving performance

To ensure homogeneity between the experimental and control groups in terms of authentic mathematical problem-solving skills before the experiment began, the study utilized the Mann-Whitney U test to compare the pre-test scores of both groups in solving authentic mathematical problems. The results, presented in Table 2, showed no significant difference in the pre-test scores for authentic mathematical problem-solving between the experimental and control groups (U=376, p=.367). Additionally, there was no significant difference in the pre-test scores between high-achieving (U=107, p=.828) and low-achieving students (U=70, p=.109) in both the experimental and control groups.

This study's first research question focused on assessing whether participation in the error analysis learning activity of GPT-4 solutions enhanced the authentic mathematical problem-solving skills of students in the experimental group. A Wilcoxon Signed-Rank test analysis was conducted. The results show that the overall scores of the experimental group significantly improved after the intervention (z=−4.318, p=0.000 < .05). Results shown in Table 3 indicate that high-achieving students significantly improved after the experimental intervention (z=−3.203, p=.001 < .05). Similarly, results presented in Table 4 demonstrate that low-achieving students also made significant progress in the post-test (z=−2.956, p=.003 < .05). These results suggest that students of varying achievement levels participating in the error analysis learning activity of GPT-4 solutions can significantly enhance their authentic mathematical problem-solving skills.

To assess the second research question, whether students' participation in the error analysis learning activity of GPT-4 solutions enhances their authentic mathematical problem-solving skills more in comparison with traditional instruction after the experiment. The Mann-Whitney U test was conducted to compare the post-test scores of both groups in solving authentic mathematical problems. In the post-test, the experimental group scored higher than the control group, but the difference was not statistically significant (U=316, p=.069). Further comparison between high-achieving and low-achieving students within

Table 2 Mann-Whitney U test results for pre- and post-test scores in AMPS

Students	Groups	N	Pre-test				Post-test			
			M	*SD*	*U*	*p*	*M*	*SD*	*U*	*p*
WC	EG	28	3.61	1.55	376	0.367	5.71	2.21	316	0.069
	CG	31	3.23	1.50			4.58	1.79		
HA	EG	14	4.14	1.79	107	0.828	6.57	2.50	98	0.541
	CG	16	3.94	1.39			5.69	1.49		
LA	EG	14	3.07	1.07	70	0.109	4.86	1.51	48	0.010*
	CG	15	2.47	1.25			3.40	1.24		

Note. * $p < .05$.

WC: Whole Class; HA: High-achieving; LA: Low-achieving; EG: Experimental Group; CG: Control Group

Table 3 Wilcoxon Signed-Rank test results for HA students' pre- and post-test scores in AMPS

Pre- and post-test measurements	N	Mean rank	Rank sum	*z*	*p*
Negative rank	0	0	0	-3.203	0.001*
Positive rank	13	7	91		
Tie	1				

Note. * $p < .05$.

Table 4 Wilcoxon Signed-Rank test results for LA students' pre- and post-test scores in AMPS

Pre- and post-test measurements	N	Mean rank	Rank sum	*z*	*p*
Negative rank	0	0	0	-2.956	0.003*
Positive rank	11	6	66		
Tie	3				

Note. * $p < .05$.

the groups revealed that high-achieving students in the experimental group scored higher than those in the control group, yet the difference was not significant (*U*=98, *p*=.541). However, low-achieving students in the experimental group significantly outperformed those in the control group (*U*=48, *p*=.01 < .05). This result may imply that, compared to traditional teaching methods, low-achieving students participating in error analysis learning activity of GPT-4 solutions could better enhance their authentic mathematical problem-solving skills.

Mathematical confidence

To ensure homogeneity between the experimental and control groups in terms of mathematical confidence before the experiment began, the study used the Mann-Whitney U test to compare the pre-test scores of mathematical confidence between the two groups. The results, presented in Table 5, showed no significant difference in the pre-test scores of mathematical confidence between the experimental and control groups (*U*=418, *p*=.808). Additionally, there was no significant difference in the pre-test scores between high-achieving (*U*=89, *p*=.338) and low-achieving students (*U*=67, *p*=.094) in both the experimental and control groups.

This study's third research question focused on assessing whether participation in the error analysis learning activity of GPT-4 solutions enhanced the mathematical confidence in the experimental group. A Wilcoxon Signed-Rank test analysis was conducted. The results show that the overall scores of the experimental group significantly improved after the intervention (z=−3.881, *p*=0.000 < .05). Table 6 revealed that the scores of high-achieving students after the intervention were significantly higher than their scores before the intervention (z=−3.184, *p*=.001 < .05). Similarly, the results presented in Table 7 indicated that the scores of low-achieving students after the intervention were significantly higher than their scores before the intervention (z=−2.103, *p*=.035 < .05). These results indicate that participating in the error analysis learning activity of GPT-4 solutions significantly enhances the mathematical confidence of students with different achievement levels, particularly showing more significant effects for high-achieving students.

Table 5 Mann-Whitney U test results for pre- and post-test scores in mathematical confidence

Students	Groups	N	Pre-test				Post-test			
			M	*SD*	*U*	*p*	*M*	*SD*	*U*	*p*
WC	EG	28	23.14	6.81	418	0.808	26.89	5.81	215	0.001*
	CG	31	22.29	5.36			21.16	6.26		
HA	EG	14	22.86	7.77	89	0.338	28.43	5.23	64	0.045*
	CG	16	24.75	4.84			23.19	6.43		
LA	EG	14	23.43	5.97	67	0.094	25.36	6.13	43	0.007*
	CG	15	19.67	4.72			19.00	5.48		

Note. * $p < .05$.
WC: Whole Class; HA: High-achieving; LA: Low-achieving; EG: Experimental Group; CG: Control Group

Table 6 Wilcoxon Signed-Rank test results for high-achieving students' pre- and post-test scores in mathematical confidence

Pre- and post-test measurements	N	Mean rank	Rank sum	*z*	*p*
Negative rank	0	0	0	-3.184	0.001*
Positive rank	13	7	91		
Tie	1				

Note. * $p < .05$.

Table 7 Wilcoxon Signed-Rank test results for low-achieving students' pre- and post-test scores in mathematical confidence

Pre- and post-test measurements	N	Mean rank	Rank sum	*z*	*p*
Negative rank	2	1.5	3	-2.103	0.035*
Positive rank	6	5.5	33		
Tie	6				

Note. * $p < .05$.

The fourth research question focused on assessing whether students' participation in the error analysis learning activity of GPT-4 solutions enhanced their mathematical confidence more in comparison with traditional instruction. The study utilized the Mann-Whitney U test to compare the post-test scores of mathematical confidence between the two groups. The results, presented in Table 5, indicated that the post-test scores of students in the experimental group were significantly higher than those in the control group (U=215, p=.001 < .05).

Upon further observation of the performance of students with different achievement levels, it was found that the post-test scores of high-achieving students in the experimental group were significantly higher than those of high-achieving students in the control group (U=64, p=.045 < .05), and the post-test scores of low-achieving students in the experimental group were also significantly higher than those of low-achieving students in the control group (U=43, p=.007 < .05). This result suggests that participating in error analysis learning activity of GPT-4 solutions is very effective in enhancing the mathematical confidence of both high-achieving and low-achieving students, compared to traditional instruction.

Interview results

Table 8 presents the interview results on the impact of participating in the error analysis learning activity of GPT-4 solutions on solving authentic mathematical problems and mathematical confidence. The key skills improved in solving authentic mathematical problems include understanding skills, collaborative problem-solving, critical thinking, mathematical creativity, and metacognitive knowledge, while mathematical confidence encompasses attitudes towards problem-solving.

Table 8 The key skills of authentic mathematical problem-solving and interview examples

Key skills	Interview examples
Understanding	• *Initially, I couldn't understand the questions or grasp the key points. Through analyzing the solutions provided by GPT-4, I was able to better comprehend the meaning of the questions.* • *GPT-4's explanations help me understand each condition and the problem requirements, enabling more accurate problem-solving.*
Collaborative problem-solving	• *When I don't understand GPT-4's explanations, I discuss problem-solving methods with my classmates to clarify my confusion.* • *When sharing with classmates or presenting the problem-solving process of GPT-4, it was only after a classmate pointed it out that I realized there were errors in GPT-4's solutions.* • *When we have different views on GPT-4's solutions, I first discuss it with classmates and share my understanding.*
Critical thinking	• *Analyzing GPT-4's mistakes taught me that making errors is okay; what's more important is to learn how to identify and correct them.* • *While analyzing GPT-4's problem-solving process, when my thoughts differ from my classmates', I actively listen to others' opinions to understand whether my ideas are correct.* • *When a classmate presents, I point out any inaccuracies or unclear descriptions and share my own perspective.*
Mathematical creativity	• *GPT-4 has shown a variety of problem-solving methods, which allows me to use diverse strategies more confidently when facing different types of mathematical problems.* • *Using GPT-4, I discovered that combining different methods can effectively solve authentic mathematical problems.*
Declarative Knowledge	• *Teachers might not elaborate on simple mathematical concepts, but GPT-4 provides comprehensive explanations.* • *Analyzing GPT-4's solutions has made it clearer for me how to explain the meaning of the questions and the significance of the formulas to my classmates.* • *Through the teacher's detailed explanations, I gained a clearer understanding of why these areas are prone to errors and the correct methods to solve them.*
Procedural knowledge	• *I use GPT-4's solutions to verify if my understanding of the problem is correct and to see if there are different perspectives to consider.* • *When the teacher explains too quickly, I can use GPT-4's solutions to understand at my own pace.* • *GPT-4 offered multiple solutions and detailed steps, helping me verify whether my thought process and calculations were correct, and also expanded my methods of solving problems.*
Conditional knowledge	• *Sometimes GPT-4's solutions are not entirely correct, but through the process of analyzing mistakes and thinking, I try to use the knowledge I have learned to examine and find solutions to the problem.* • *I prefer using GPT-4's explanations because I'm too shy to ask questions.* • *Although GPT-4 provides detailed explanations, I wish it could offer easier-to-understand methods, such as drawing or verbal explanations.*
Mathematical confidence	• *Facing authentic mathematical problems, I've become more patient. I no longer give up because a problem seems complex; instead, I try to understand it gradually.* • *By analyzing the answers provided by GPT-4, I've become more focused on reading and understanding questions, which makes it easier for me to identify errors in the problem-solving process.* • *When faced with challenging questions I don't know how to solve, the explanations of formulas and problem-solving processes by GPT-4 effectively reduce my anxiety. Now, even when facing difficult problems, I have more confidence to tackle them step by step.*

Moreover, the consistency of the student interview sample classification was evaluated by three experts, with a Fleiss' Kappa value of 0.934 (where a small effect=0.20; medium effect=0.50; large effect=0.80; Fleiss & Cohen, 1973). This indicates that the evaluators achieved almost perfect agreement, demonstrating high consistency in classifying the interview samples, thus ensuring the reliability and validity of the classification results.

According to the interview results, 93% of students felt that detailed explanations and analyses provided by GPT-4 on the purpose, strategy, steps, and calculation processes of authentic mathematical problems helped enhance their understanding. When students explained and analyzed GPT-4's solutions among peers and during presentations, they could clarify doubts, making 79% of students feel it fostered collaborative problem-solving among peers. Furthermore, 43% felt it aided in developing their critical thinking; due to GPT-4 showcasing various problem-solving methods that encouraged different ways of thinking, 43% believed they had enhanced their creative problem-solving diversity. Moreover, 86% of students demonstrated how they improved their metacognitive skills from aspects such as declarative, procedural, and conditional knowledge. Additionally, 86% of students thought this teaching activity enhanced their patience, perseverance, and concentration when solving authentic mathematical problems, aligning with mathematical confidence-related questions in the TIMSS questionnaire.

In summary, through GPT-4's explanations of authentic mathematical problems and collaborative error analysis learning activities, students were able to deeply understand the mathematical problems. During discussing and sharing, not only effectively trained students' problem-solving methods deconstruct and analyze information from classmates and GPT-4, identifying potential problem-solving errors but also learned how to correct them and explain to classmates, thereby developing critical thinking. Through continuous group and class dialogues, students practiced divergent and convergent thinking, tested knowledge learned previously, and sought diverse problem-solving methods, which later enabled them to generate innovative solutions when faced with different types of mathematical problems. The interactions among students created a positive learning atmosphere, fostering an active learning motivation and confidence. The interview results provide evidence of the effective enhancement of key skills for authentic mathematical problem-solving and mathematical confidence among students.

Discussion

Previous studies have examined AI in mathematical problem-solving (Gattupalli et al., 2023), the impact of error analysis learning activities on mathematical problem-solving skills (Alvidrez et al., 2024; Kramarski & Zoldan, 2008; Toikka et al.,2024), and the significance of authentic mathematical problem-solving on mathematical confidence (Samritin et al., 2023; Santos-Trigo, 2024). However, studies on the actual application of

GPT-4 in the mathematics education environment are relatively scarce, and there is a lack of research on combining GPT-4 solution examples and error analysis strategies to enhance students' skills in solving authentic mathematical problems and their mathematical confidence. Thus, this study provides new insights into how these variables interact.

Our findings indicate that, compared to traditional teaching methods, integrating GPT-4 solution examples and error analysis strategies can better enhance students' skills in solving authentic mathematical problems and their mathematical confidence, especially for low-achieving students. This result contrasts with previous studies, which tended to find that authentic mathematical problems are more complex and challenging, often leading to low-achieving students feeling overwhelmed and lacking in mathematical confidence (Mullis et al., 2020).

However, encouraging students to actively practice solving authentic mathematical problems and engage in logical reasoning can help improve their problem-solving capabilities, consistent with the views of Zainiyah and Marsigit (2018). When students design and create new methods to solve authentic mathematical problems, it helps enhance creativity and develop diverse problem-solving strategies, thereby achieving better mathematics learning outcomes (Khalid, 2020; Schunk, 2012).

The study also found that students had to apply their conditional knowledge to solve problems while using GPT-4 for problem-solving and error analysis. Conditional knowledge refers to an individual's understanding of when and why to use specific cognitive strategies, including knowledge of the learning situation and contextual factors such as whether a task is difficult or resources are available. This triggered metacognitive monitoring of their own capabilities (Khasawneh et al., 2023; Kshetree et al., 2021; Toikka et al., 2024), making students aware of potential knowledge gaps.

Through peer explanations and presentations, students had more opportunities to understand situational problems (Rushton, 2018; Yarman et al., 2020), discuss issues, and find solutions, promoting active participation in the activities. This transformed their mathematical knowledge from vague concepts to concrete content, and many reasoning and thinking blind spots were clarified during the checking of GPT-4's solutions or peer explanations (Fuchs et al., 2020), aligning with the initial survey results of GPT-4's application in mathematics education (Patero, 2023; Wardat et al., 2023).

Furthermore, students' mathematical confidence significantly increased after participating in the GPT-4 solution error analysis learning activity, compared to traditional teaching methods. This aligns with previous research findings, which emphasize the importance of providing a supportive learning environment, positive teacher-student interactions, peer support relationships, and self-regulated learning strategies in enhancing students' mathematical confidence (Eccles & Wigfield, 2020; Ryan & Deci, 2020; Schunk & DiBenedetto, 2020).

In this study, we used the Mathematical Confidence Scale developed by TIMSS 2019 (Martin et al., 2020). The Cronbach's alpha value of this scale was 0.85, indicating good internal consistency and ensuring the reliability of the scale in measuring students' mathematical confidence.

To verify the reliability and validity of our study outcomes, we invited three experts to evaluate the consistency of the student interview sample classification. The results showed a Fleiss' Kappa value of 0.934, indicating almost perfect agreement among the evaluators, thus confirming the high consistency and reliability of the classification results.

By creating a classroom environment that allows for mathematical errors, students were able to express their ideas boldly and accept feedback during peer explanations and presentations, effectively reducing their mathematical anxiety (AlAli & Al-Barakat, 2022; Khasawneh et al., 2023).

Additionally, through GPT-4's explanations, problem-solving strategies, steps, and calculation processes, students were helped to understand the key points of problem-solving, motivating them to actively explore and patiently seek errors in the problem-solving process (Sánchez-Ruiz et al., 2023), thereby enhancing their problem-solving confidence (Patero, 2023; Surya & Putri, 2017).

However, the reasons for the increase in mathematical confidence were not the same for high-achieving and low-achieving students in the experimental group. According to interview results, high-achieving students had more opportunities to use mathematical representations to communicate ideas and solve problems with others when sharing different problem-solving strategies from GPT-4 within their groups, ensuring the correctness of their problem-solving processes, fostering higher-level reasoning (Laurens et al., 2017), and inspiring more creative problem-solving methods, thus having higher mathematical confidence.

Conversely, low-achieving students lacked foundational knowledge and problem-solving application skills and still relied heavily on teacher explanations and peer interactions even after reviewing GPT-4's problem-solving explanations. Thus, GPT-4 cannot completely replace the role of the teacher, as teachers can more quickly and accurately identify students' learning gaps, providing personalized support and instructional adjustments to accommodate different learning styles (Supriyadi & Kuncoro, 2023).

Despite this, low-achieving students were still able to overcome obstacles in understanding authentic mathematical problem-solving processes and comprehend potential errors and their reasons in GPT-4's problem-solving explanations through GPT-4's explanations and discussions with peers, thereby enhancing their mathematical confidence (Zafrullah et al., 2023).

Limitations of the study

This study has three limitations. First, it employed purposive sampling, and while the schools are located in a large city, they are positioned in the suburbs. To generalize the results nationally, more extensive sampling and educational interventions would be required to confirm if the educational outcomes remain consistent. Second, the school's provision of relevant information technology courses means students are proficient in using tablets and operating GPT-4 in the classroom. Schools without such resources or students with poorer IT skills might experience difficulties that could affect the course process and the effectiveness of the experimental curriculum. Third, the high cost of the paid version of GPT-4 may pose a financial burden on educational institutions and individual users, which could prevent the widespread adoption of this experimental curriculum. However, with ongoing changes in technology, future costs may decrease, potentially enabling more schools, teachers, and students to benefit from this technology.

Conclusions and future works

This study aimed to explore the impact of error analysis learning activities using GPT-4 solutions on enhancing fifth-grade students' authentic mathematical problem-solving skills and confidence. A quasi-experimental design was employed, involving 59 fifth-grade students from an elementary school in northern Taiwan. The experimental group participated in error analysis learning activities using GPT-4 solutions, while the control group received traditional instruction, both using the same materials. Quantitative assessments were conducted through authentic mathematical problem-solving tests and a mathematical confidence scale, supplemented by semi-structured interviews to collect qualitative data.

The results indicated that the experimental group showed a significant improvement in their ability to solve authentic mathematical problems, outperforming the control group in both within-group pre- and post-test comparisons and between-group post-test comparisons. Furthermore, low-achieving students in the experimental group demonstrated a significantly greater improvement in solving authentic mathematical problems compared to the control group. These findings confirm the effectiveness of error analysis learning activities, particularly for low-achieving students. Additionally, students in the experimental group exhibited significantly higher mathematical confidence than those in the control group, with both high- and low-achieving students showing increased confidence in mathematics. This suggests that error analysis learning activities using GPT-4 not only enhance students' problem-solving skills but also boost their confidence in learning mathematics.

This study makes a significant contribution to the existing literature, which has primarily focused on evaluating the problem-solving capabilities of GPT-4 and its application in mathematical problem-solving (An et al., 2023; Frieder et al., 2023; Plevris et al., 2023), or on understanding the impact of GPT-4 in educational settings through questionnaires and interviews (Wardat et al., 2023; Zafrullah et al., 2023). However, combining GPT-4 problem-solving examples with error analysis strategies and applying them to improve students' authentic mathematical problem-solving skills and confidence is relatively rare. Therefore, this study integrates GPT-4 solutions into error analysis learning activities and investigates their impact on authentic mathematical problem-solving skills and confidence. The findings suggest that this activity not only promotes a deeper understanding of the authentic mathematical problem-solving process but also significantly enhances students' mathematical confidence, especially for low-achieving students (Begolli et al., 2021; Kshetree et al., 2021).

The study also found that students need to apply their conditional knowledge when using GPT-4 for problem-solving and error analysis, which promotes metacognitive monitoring and helps students become aware of knowledge gaps. Through peer explanations and staging, students have more opportunities to understand contextual problems, discuss issues, and find solutions, fostering active participation. These findings are consistent with previous research, highlighting the importance of a supportive learning environment and positive student interactions in enhancing students' mathematical confidence (Eccles & Wigfield, 2020; Ryan & Deci, 2020; Schunk & DiBenedetto, 2020).

In conclusion, this study provides new empirical evidence confirming the effectiveness of error analysis learning activities using GPT-4 solutions and offers valuable references for future teaching practices and research.

This research suggests that teachers should encourage students to use GPT-4 with a critical mindset and learning attitude, viewing it as an opportunity for learning challenges to enhance their mathematical problem-solving and comprehension skills. Additionally, considering the significant effect of peer interaction and error analysis of GPT-4's problem-solving explanations on improving the learning outcomes of low-achieving students, it is recommended that teachers fully utilize GPT-4 to enhance discussions among students, thereby promoting cooperative learning and diversified problem-solving thinking.

Currently, there are various AI tools available; this study specifically used the paid version of GPT-4. Future research could further explore GPT-4o and GPTs and compare different AI tools in generating solutions to authentic mathematical problems, to determine if errors still occur, if there are different types of errors, and whether the content generated by these tools can serve as educational material for error analysis. Additionally, educators integrating AI tools into their teaching should consider implementing comprehensive training programs that include tablet operation, AI tool application, information literacy, and ethics education. These programs are crucial to ensure that students can effectively and accurately utilize these technologies, thereby enhancing their academic competencies.

Abbreviations

AMPS: Authentic Mathematical Problem-Solving; GPT-4: Generative Pre-trained Transformer 4; AI: Artificial Intelligence; TIMSS: Trends in International Mathematics and Science Study

Acknowledgements

We extend our sincere gratitude to the class teachers and students at Renhe Elementary School in Taoyuan, Taiwan, for their invaluable participation in our study. Their cooperation greatly facilitated the research process.

Authors' contributions

Yu-Feng Lin was responsible for instruction, conceptualization, project administration, methodology, investigation, data curation, formal analysis, interpretation of data, writing—original draft, review of the manuscript, and contributed to editing of the manuscript. Euphony F. Y. Yang was responsible for conceptualization, methodology, formal analysis, interpretation of data, writing—original draft, and review of the manuscript, and contributed to editing of the manuscript, and supervision. Jeng-Shin Wu was responsible for formal analysis, interpretation of data and review of the manuscript, contributed to editing of the manuscript. Charles Y. C. Yeh was responsible for review of the manuscript. Chang-Yen Liao contributed significantly to the design and analysis framework of the error analysis learning activities in the study, which greatly enhanced the methodological rigor and strengthened the results of the research. Tak-Wai Chan was responsible for review of the manuscript, and supervision. All authors read and approved the final manuscript.

Authors' information

Yu-Feng Lin is a Ph.D. student at Graduate Institute of Network Learning Technology, National Central University in Taiwan. His research interests include computer-supported learning in mathematics, game-based learning, and self-explanation on learning mathematics.

Euphony Fu-Yu Yang is an Assistant Research Fellow at the Clinical Competence Center of Chang Gung University of Science and Technology in Taiwan. She has extensive experience in artificial intelligence in education, computer-supported collaborative learning, mobile and ubiquitous learning, and digital game-based learning, contributing to the development of innovative educational technologies and research.

Jeng-Shin Wu is an Associate Researcher at the Research Center for Curriculum and Instruction at the National Academy for Educational Research (NAER) in Taiwan. His research interests encompass educational testing, large-scale educational surveys, and mathematics literacy assessment.

Charles Y. C. Yeh is currently a Postdoctoral Researcher in Graduate Institute of Network Learning Technology at National Central University in Taiwan. His research interests include mathematics and digital game-based learning.

Chang-Yen Liao is an Assistant Professor of Department of Hakka Language and Social Sciences at National Central University in Taiwan. He specializes in technology-enhanced learning and artificial intelligence applications in education.

Tak-Wai Chan is a Chair Professor of the Graduate Institute of Network Learning Technology at National Central University in Taiwan. He has worked on various areas of digital technology-supported learning, including artificial intelligence in education, computer-supported collaborative learning, digital classrooms, online learning communities, mobile and ubiquitous learning, digital game-based learning, and, most recently, technology-supported mathematics and language arts learning.

Funding

Not applicable.

Availability of data and materials

As a series of subsequent research papers are still in progress, for now, it is temporarily impossible to share research data sets.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1] National Central University, No. 300, Zhongda Rd., Zhongli District, Taoyuan City 32001, Taiwan.

[2] Chang Gung University of Science and Technology, No.261, Wenhua 1st Rd., Guishan Dist., Taoyuan City 33303, Taiwan.

[3] National Academy for Educational Research, No.2, Sanshu Rd., Sanxia Dist., New Taipei City 237201, Taiwan.

Received: 9 April 2024 Accepted: 21 August 2024

Published online: 1 January 2025 (Online First: 18 November 2024)

References

AlAli, R., & Al-Barakat, A. (2022). Using structural equation modeling to assess a model for measuring creative teaching perceptions and practices in higher education. *Education Sciences*, *12*(10), 690. https://doi.org/10.3390/educsci12100690

Almulla, M. A. (2023). Constructivism learning theory: A paradigm for students' critical thinking, creativity, and problem solving to affect academic performance in higher education. *Cogent Education*, *10*(1), 2172929. https://doi.org/10.1080/2331186X.2023.2172929

Alvidrez, M., Louie, N., & Tchoshanov, M. (2024). From mistakes, we learn? Mathematics teachers' epistemological and positional framing of mistakes. *Journal of Mathematics Teacher Education*, *27*(1), 111–136. https://doi.org/10.1007/s10857-022-09553-4

An, S., Ma, Z., Lin, Z., Zheng, N., Lou, J. G., & Chen, W. (2023). *Learning from mistakes makes LLM better reasoner*. https://doi.org/10.48550/arXiv.2310.20689

Aulia, E. T., & Prahmana, R. C. I. (2022). Developing interactive e-module based on realistic mathematics education approach and mathematical literacy ability. *Jurnal Elemen*, *8*(1), 231–249. https://doi.org/10.29408/jel.v8i1.4569

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52–62. https://dx.doi.org/10.2139/ssrn.4337484

Begolli, K. N., Dai, T., McGinn, K. M., & Booth, J. L. (2021). Could probability be out of proportion? Self-explanation and example-based practice help students with lower proportional reasoning skills learn probability. *Instructional Science*, *49*, 441–473. https://doi.org/10.1007/s11251-021-09550-9

Bingölbali, E., & Bingölbali, F. (2021). An examination of open-ended mathematics questions' affordances. *International Journal of Progressive Education*, *17*(4), 1–16. https://doi.org/10.29329/ijpe.2021.366.1

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. https://doi.org/10.48550/arXiv.2303.12712

Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J. B., Hart, W., Gowers, T., Li, W., Weller, A., & Jamnik, M. (2024). Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, *121*(24), e2318124121. https://doi.org/10.1073/pnas.2318124121

Dillenbourg, P. (1999). What do you mean by collaborative learning?. In P. Dillenbourg (Ed.), *Collaborative-learning: Cognitive and computational approaches* (pp. 1–19). Elsevier.

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, *61*, 101859. https://doi.org/10.1016/j.cedpsych.2020.101859

Fitzsimons, A., & Ní Fhloinn, E. (2023). The CoPs model for collaborative problem-solving in mathematics. *Irish Educational Studies*, 1–18. https://doi.org/10.1080/03323315.2023.2189137

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*(3), 613–619. https://doi.org/10.1177/001316447303300309

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2023). Mathematical capabilities of ChatGPT. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt & S. Levine (Eds.), *Advances in Neural Information Processing Systems*, *36*, *Proceedings of the 37th International Conference on Neural Information Processing Systems* (pp. 27699–27744). ACM. https://doi.org/10.48550/arXiv.2302.08760

Fuchs, D., Fuchs, L. S., & Abramson, R. (2020). Peer-Assisted Learning Strategies (PALS): A validated classwide program for improving reading and mathematics performance. In A. L. Reschly, A. J. Pohl & S. L. Christenson (Eds.), *Student engagement: Effective academic, behavioral, cognitive, and affective interventions at school* (pp. 109–120). Springer. https://doi.org/10.1007/978-3-030-37285-9_6

Gattupalli, S., Maloy, R. W., & Edwards, S. (2023). *Comparing teacher-written and AI-generated math problem solving strategies for elementary school students: Implications for classroom learning*. https://doi.org/10.7275/8sgx-xj08

Hwang, G. J., & Chen, N. S. (2023). Exploring the potential of generative artificial intelligence in education: Applications, challenges, and future research directions. *Journal of Educational Technology & Society*, *26*(2). https://doi.org/10.30191/ETS.202304_26(2).0014

İpek, Z. H., Gözüm, A. I. C., Papadakis, S., & Kallogiannakis, M. (2023). Educational applications of the ChatGPT AI system: A systematic review research. *Educational Process: International Journal*, *12*(3), 26–55. https://doi.org/10.22521/edupij.2023.123.2

Khalid, M., Saad, S., Hamid, S. R. A., Abdullah, M. R., Ibrahim, H., & Shahrill, M. (2020). Enhancing creativity and problem solving skills through creative problem solving in teaching mathematics. *Creativity Studies*, *13*(2), 270–291. https://doi.org/10.3846/cs.2020.11027

Khasawneh, A. A., Al-Barakat, A. A., & Almahmoud, S. A. (2023). The impact of mathematics learning environment supported by error-analysis activities on classroom interaction. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*(2), em2227. https://doi.org/10.29333/ejmste/12951

Kramarski, B., & Zoldan, S. (2008). Using errors as springboards for enhancing mathematical reasoning with three metacognitive approaches. *The Journal of Educational Research*, *102*(2), 137–151. https://doi.org/10.3200/JOER.102.2.137-151

Kshetree, M. P., Acharya, B. R., Khanal, B., Panthi, R. K., & Belbase, S. (2021). Eighth grade students' misconceptions and errors in mathematics learning in Nepal. *European Journal of Educational Research*, *10*(3), 1101–1121. https://doi.org/10.12973/eu-jer.10.3.1101

Lan, Y. J., & Chen, N. S. (2024). Teachers' agency in the era of LLM and generative AI. *Educational Technology & Society*, *27*(1), I-XVIII. https://www.jstor.org/stable/48754837

Larrain, M., & Kaiser, G. (2022). Interpretation of students' errors as part of the diagnostic competence of pre-service primary school Teachers. *Journal for Mathematics Didactics*, *43,* 39–66. https://doi.org/10.1007/s13138-022-00198-7

Laurens, T., Batlolona, F. A., Batlolona, J. R., & Leasa, M. (2017). How does realistic mathematics education (RME) improve students' mathematics cognitive achievement? *Eurasia Journal of Mathematics, Science and Technology Education*, *14*(2), 569–578. https://doi.org/10.12973/ejmste/76959

Lo, C. K., Hew, K. F., & Jong, M. S. Y. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education*, *219*, 105100. https://doi.org/10.1016/j.compedu.2024.105100

Martin, M. O., von Davier, M., & Mullis, I. V. (2020). *Methods and procedures: TIMSS 2019 Technical Report*. International Association for the Evaluation of Educational Achievement.

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. TIMSS & PIRLS International Study Center, Boston College.

National Academy for Educational Research. (2020). *Literacy-oriented assessment items research and development for teachers' professional development (first year): Mathematics items*. National Academy for Educational Research. https://ptips.ntct.edu.tw/var/file/69/1069/img/173342938.pdf (in Chinese)

National Academy for Educational Research. (2021). *Literacy-oriented assessment items research and development for teachers' professional development (second year): Mathematics items*. National Academy for Educational Research. https://ptips.ntct.edu.tw/var/file/69/1069/img/376480000A_1100252353_ATTACH1.pdf (in Chinese)

Nilimaa, J. (2023). New examination approach for real-world creativity and problem-solving skills in mathematics. *Trends in Higher Education*, *2*(3), 477–495. https://doi.org/10.3390/higheredu2030028

Niss, M., & Jablonka, E. (2014). Mathematical literacy. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 391–396). Springer Netherlands. https://doi.org/10.1007/978-94-007-4978-8_100

Niss, M., & Højgaard, T. (Eds.). (2011). *Competencies and mathematical learning: Ideas and inspiration for the development of mathematics teaching and learning in Denmark*. Roskilde University.

Nuraina, N., Fauzi, K. M. A., & Simbolon, N. (2021). The effect of realistic mathematics educations (RME) approach based on ethnomatics on the improvement of concept understanding ability and students' learning motivation in elementary school AlKausar City of Langsa. *Budapest International Research and Critics in Linguistics and Education (BirLE) Journal*, *4*(1), 543–554. https://doi.org/10.33258/birle.v4i1.1707

OECD. (2022). *PISA 2022 Mathematics Framework (Draft)*. https://pisa2022-maths.oecd.org/

OpenAI. (2022). *ChatGPT: Optimizing language models for dialogue*. https://openai.com/blog/chatgpt/

OpenAI. (2023). GPT-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf

Pardos, Z. A., & Bhandari, S. (2023). *Learning gain differences between ChatGPT and human tutor generated algebra hints*. https://doi.org/10.48550/arXiv.2302.06871

Patero, J. L. (2023). Revolutionizing math education: Harnessing ChatGPT for student success. *Mathematics Education*, *3*(1), 807–813. https://doi.org/10.48175/ijarsct-12375

Plevris, V., Papazafeiropoulos, G., & Jiménez Rios, A. (2023). Chatbots put to the test in math and logic problems: A comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *AI*, *4*(4), 949–969. https://doi.org/10.3390/ai4040048

Putri, D. A., Susanti, V. D., & Apriandi, D. (2020). Pengembangan Modul Berbasis RME Untuk Meningkatkan Kemampuan Literasi Matematika Siswa Kelas Xi SMK. *Prima Magistra: Jurnal Ilmiah Kependidikan*, *1*(2), 138–146. https://doi.org/10.37478/jpm.v1i2.470

Reinke, L. T., Stephan, M., Ayan-Civak, R., & Casto, A. R. (2023). Teachers' press for contextualization to ground students' mathematical understanding of ratio. *Journal of Mathematics Teacher Education*, *26*(3), 335–361. https://doi.org/10.1007/s10857-022-09531-w

Richey, J. E., Andres-Bray, J. M. L., Mogessie, M., Scruggs, R., Andres, J. M. A. L., Star, J. R., Baker, R. S., & McLaren, B. M. (2019). More confusion and frustration, better learning: The impact of erroneous examples. *Computers & Education*, *139*, 173–190. https://doi.org/10.1016/j.compedu.2019.05.012

Rong, L., & Mononen, R. (2022). Error-analysis of students with mathematics learning difficulties in Tibet. *Asian Journal for Mathematics Education*, *1*(1), 52–65. https://doi.org/10.1177/27527263221089357

Roorda, G., de Vries, S., & Smale-Jacobse, A. E. (2024). Using lesson study to help mathematics teachers enhance students' problem-solving skills with teaching through problem solving. *Frontiers in Education*, *9*, 1331674. https://doi.org/10.3389/feduc.2024.1331674

Rospigliosi, P. A. (2023). Artificial intelligence in teaching and learning: What questions should we ask of ChatGPT? *Interactive Learning Environments*, *31*(1), 1–3. http://dx.doi.org/10.1080/10494820.2023.2180191

Rushton, S. J. (2018). Teaching and learning mathematics through error analysis. *Fields Mathematics Education Journal*, *3*(1), 1–12. https://doi.org/10.1186/s40928-018-0009-y

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, *61*, 101860. https://doi.org/10.1016/j.cedpsych.2020.101860

Samritin, S., Natsir, S. R., Manaf, A., & Sari, E. R. (2023). The effect of realistic mathematics education implementation in mathematics learning in elementary school. *Formatif: Jurnal Ilmiah Pendidikan MIPA*, *13*(1). http://dx.doi.org/10.30998/formatif.v13i1.16522

Sánchez-Ruiz, L. M., Moll-López, S., Nuñez-Pérez, A., Moraño-Fernández, J. A., & Vega-Fleitas, E. (2023). ChatGPT challenges blended learning methodologies in engineering education: A case study in mathematics. *Applied Sciences*, *13*(10), 6039. https://doi.org/10.3390/app13106039

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

Schneider, W., & Artelt, C. (2010). Metacognition and mathematics education. *ZDM–Mathematics Education*, *42*, 149–161. https://doi.org/10.1007/s11858-010-0240-2

Schunk, D. H. (2012). *Learning theories an educational perspective*. Pearson Education, Inc.

Schunk, D. H., & DiBenedetto, M. K. (2020). Motivation and social cognitive theory. *Contemporary Educational Psychology*, *60*, 101832. https://doi.org/10.1016/j.cedpsych.2019.101832

Shakarian, P., Koyyalamudi, A., Ngu, N., & Mareedu, L. (2023). *An independent evaluation of ChatGPT on mathematical word problems (MWP)*. https://doi.org/10.48550/arXiv.2302.13814

Sok, S., & Heng, K. (2023). ChatGPT for education and research: A review of benefits and risks. *Cambodian Journal of Educational Research*, *3*(1), 110–121. https://doi.org/10.62037/cjer.2023.03.01.06

Subekti, M. A. S., & Prahmana, R. C. I. (2021). Developing interactive electronic student worksheets through discovery learning and critical thinking skills during pandemic era. *Mathematics Teaching-Research Journal*, *13*(2), 137–176.

Supriyadi, E., & Kuncoro, K. S. (2023). Exploring the future of mathematics teaching: Insight with ChatGPT. *Union: Jurnal Ilmiah Pendidikan Matematika*, *11*(2), 305–316. https://doi.org/10.30738/union.v11i2.14898

Surya, E., & Putri, F. A. (2017). Improving mathematical problem-solving ability and self-confidence of high school students through contextual learning model. *Journal on Mathematics Education*, *8*(1), 85–94. https://doi.org/10.22342/jme.8.1.3324.85-94

Tay, L. Y., Chan, M., Chong, S. K., Tan, J. Y., & Aiyoob, T. B. (2024). Learning of mathematics: A metacognitive experiences perspective. *International Journal of Science and Mathematics Education*, *22*(3), 561–583. https://doi.org/10.1007/s10763-023-10385-8

Toikka, S., Eronen, L., Atjonen, P., & Havu-Nuutinen, S. (2024). Combined conceptualisations of metacognitive knowledge to understand students' mathematical problem-solving. *Cogent Education*, *11*(1), 2357901. https://doi.org/10.1080/2331186X.2024.2357901

Van den Heuvel-Panhuizen, M., & Drijvers, P. (2020). Realistic mathematics education. In S. Lerman (Eds.), *Encyclopedia of Mathematics Education* (pp. 713–717). Springer, Cham. https://doi.org/10.1007/978-3-030-15789-0_170

Verschaffel, L., Schukajlow, S., Star, J., & Van Dooren, W. (2020). Word problems in mathematics education: A survey. *ZDM–Mathematics Education*, *52*, 1–16. https://doi.org/10.1007/s11858-020-01130-4

Wang, A., Prihar, E., Haim, A., Heffernan, N. (2024). Can large language models generate middle school mathematics explanations better than human teachers?. In A. M. Olney, I. A. Chounta, Z. Liu, O. C. Santos & I. I. Bittencourt (Eds.), *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky. AIED 2024. Communications in Computer and Information Science, vol 2151* (pp. 242–250). Springer, Cham. https://doi.org/10.1007/978-3-031-64312-5_29

Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*(7), em2286. https://doi.org/10.29333/ejmste/13272

Wildgans-Lang, A., Scheuerer, S., Obersteiner, A., Fischer, F., & Reiss, K. (2020). Analyzing prospective mathematics teachers' diagnostic processes in a simulated environment. *ZDM—Mathematics Education*, *52*(2), 241–254. https://doi.org/10.1007/s11858-020-01139-9

Wu, J. S., Hsieh, C. J., & Huang, Y. K. (2022). An investigation into the performances of sixth graders in answering literacy-oriented assessments in mathematics. *Taiwan Journal of Mathematics Education*, *9*(2), 87–111. https://doi.org/10.6278/tjme.202210_9(2).004

Wu, Y., Jia, F., Zhang, S., Li, H., Zhu, E., Wang, Y., Lee, Y. T., Peng, R., Wu, Q., & Wang, C. (2023). *An empirical study on challenging math problem solving with GPT-4*. https://doi.org/10.48550/arXiv.2306.01337

Yarman, Y., Fauzan, A., Armiati, A., & Lufri, L. (2020). Analysis of student errors in solving ordinary differential equations based on SOLO taxonomy. *Journal of Physics: Conference Series*, *1554*, 012065. https://doi.org/10.1088/1742-6596/1554/1/012065

Zafrullah, Z., Hakim, M. L., & Angga, M. (2023). ChatGPT open AI: Analysis of mathematics education students learning interest. *Journal of Technology Global*, *1*(01), 1–10.

Zainiyah, U., & Marsigit. (2018). Mathematical literacy: How if viewed from mathematics problem solving ability of high-grade's elementary school students? *Jurnal Riset Pendidikan Matematika*, *4*(1), 5–14. https://doi.org/10.48550/arXiv.2304.10513

Zhang, D. (2017). Effects of visual working memory training and direct instruction on geometry problem solving in students with geometry difficulties. *Learning Disabilities-A Contemporary Journal*, *15*(1), 117–138.

Zheng, S., Huang, J., & Chang, K. C. C. (2023). *Why does ChatGPT fall short in answering questions faithfully?* https://doi.org/10.48550/arXiv.2304.10513

Zong, M., & Krishnamachari, B. (2023). Solving math word problems concerning systems of equations with GPT models. *Machine Learning with Applications*, *14*, 100506. https://doi.org/10.1609/aaai.v37i13.26896

Publisher's Note

The Asia-Pacific Society for Computers in Education (APSCE) remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.