Research Square

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

Middle School Students and Generative AI Solving Exponent Problems: A Comparative Analysis of Written Solutions and Solution Reliability

Sena Gencel
Istanbul Medeniyet University
İclal Sude Çetinkaya
Istanbul Medeniyet University
Fatmagül Ergün
fergun@29mayis.edu.tr

Istanbul 29 Mayıs University: Istanbul 29 Mayis Universitesi

Research Article

Keywords: generative artificial intelligence, mathematical problem solving, mathematical justification, error analysis, solution reliability

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11239535/v1

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

Additional Declarations: No competing interests reported.

Middle School Students and Generative AI Solving Exponent Problems: A Comparative Analysis of Written Solutions and Solution Reliability

Sena Gencel[a], İclal Sude Çetinkaya[a], and Fatmagül Ergün[b]*

[a] Mathematics Education Department, Istanbul Medeniyet University, Istanbul, Türkiye, Sena Gencel ORCID: 0009-0000-6562-5624, İclal Sude Çetinkaya ORCID: 0009-0003-3085-4322

[b] Mathematics Education Department, Istanbul 29 May University, Istanbul, Türkiye, Fatmagül Ergün ORCID: 0009-0002-4137-8184

* Corresponding author: Fatmagül Ergün; Email: fergun@29mayis.edu.tr

Abstract

This study compares the independent written solutions produced by 22 students aged 12–13 for 10 exponent problems with the first responses generated by ChatGPT, Claude, Gemini, and Grok for the same problems. Using a comparative case study design, the data were analyzed inductively in line with Braun and Clarke's (2006) six-phase thematic analysis approach. The solutions were examined in terms of problem representation and mathematization, strategy selection, mathematical execution, reasoning, justification, error patterns, and solution reliability. The findings showed that students' solutions were generally brief and symbolic, with intermediate relationships and justifications remaining implicit in many cases. Generative AI outputs made givens, intermediate steps, and justifications more explicit and more clearly sequenced. Student errors were concentrated primarily in procedural execution, exponent rules, decimal decomposition, and relationships between negative exponents and signs. Generative AI errors occurred particularly in tasks requiring interpretation of visual structures and mathematization of context. In some student and AI solutions, a process error co-occurred with a correct final answer, indicating that answer correctness alone does not establish solution reliability. Under the tasks, model versions, and first-response protocol examined in this study, no coded mathematical error was identified in the ChatGPT outputs. Errors were observed in specific visual and contextual tasks in the outputs of Claude, Gemini, and Grok. The findings indicate that generative AI solutions should be treated not as ready-made correct answers in mathematics education, but as digital mathematical products that require critical validation in terms of representation, strategy, execution, justification, and checking.

Keywords: generative artificial intelligence, mathematical problem solving, mathematical justification, error analysis, solution reliability

1. Introduction

Mathematical problem solving requires identifying relationships, selecting strategies, constructing representations, and justifying results. These processes draw on conceptual understanding, procedural fluency, strategic competence, and adaptive reasoning (National Research Council, 2001). Pólya (1945) identified understanding, planning, execution, and looking back as key stages of problem solving. Schoenfeld (1985, 1992) further highlighted decision making and monitoring. Consequently, solutions with identical final answers may differ in representation, strategy, execution, justification, and checking; these features must be considered when evaluating solution reliability.

The ability of generative artificial intelligence systems to produce mathematical explanations and step-by-step solutions through natural language has introduced a new dimension to mathematical problem solving. ChatGPT, Claude, Gemini, and Grok can organize information in a problem, propose solution strategies, carry out calculations, and justify results. Kasneci et al. (2023) discussed educational opportunities for large language models, Lo (2023) reviewed applications of ChatGPT in education, and Pepin et al. (2025) examined generative AI in mathematics education. Together, these discussions situate explanation, feedback, and solution support among the educational uses

of generative AI. Wardat et al. (2023) showed that ChatGPT can support mathematics teaching and learning. McGalliard and Otten (2025) found that generative AI can produce detailed responses to cognitively demanding secondary mathematics problems. Ergene and Ergene (2025) compared the performance of five AI systems on 390 mathematics problems; students and teachers identified detailed explanations, rapid access to solutions, and learning support among the educational affordances of these systems.

Fluent and well-sequenced solution texts generated by AI are not necessarily mathematically valid. Misinterpreting problem conditions, selecting unsuitable strategies, making procedural errors, and failing to verify a result adequately can undermine solution reliability. Wardat et al. (2023) reported that ChatGPT's mathematical performance varies according to problem and equation complexity, input data, and instructions. McGalliard and Otten (2025) identified substantial errors in responses to cognitively demanding problems. Korkmaz Guler et al. (2024) reported that ChatGPT-4 outperformed ChatGPT-3.5 and that both models produced mathematical errors. Wei (2024) showed that the performance of ChatGPT-4 and ChatGPT-4o on NAEP mathematics items varies by content area and item difficulty. Oh et al. (2024) identified differences in accuracy among ChatGPT-4, Claude 3 Opus, and Gemini Advanced, as well as difficulties with visual tasks. Spreitzer et al. (2024) reported that difficulties in interpreting context and mathematizing became more pronounced as problem complexity increased. These findings indicate that generative AI solutions require examination for mathematical accuracy and reliability before being used in education (UNESCO, 2023). The quality of generative AI solutions varies with the model, the task, and the prompt. Schorcht et al. (2024) showed that prompting techniques can affect process-related features such as strategy and representation more strongly than the content quality of solutions. Comparing four models, six problems, and four prompting techniques, Schorcht et al. (2026) found that content quality depended more on model and task characteristics, whereas process-related and pedagogical-contextual quality varied with prompt design. These findings highlight the value of comparing different AI tools on the same problems under the same prompting conditions and evaluating the visible features of their solutions using common criteria.

Solution reliability requires examining the method, mathematical grounds, and justifications together with the final result (Zhuang, 2026). Generative AI tools can present mathematically incorrect results with persuasive explanations (Frieder et al., 2023); users may accept incorrect ChatGPT solutions (Noster et al., 2024); and Bulut and Borromeo Ferri (2026) examined challenges associated with interpretation and validation in AI-supported mathematical modeling; Cevikbas and Kaiser (2026) also addressed issues of AI-supported mathematical modeling. Mathematical errors provide information about the conceptual and procedural structures used in a solution (Radatz, 1979) and constitute an analytical resource for examining mathematical thinking (Borasi, 1994). Errors should therefore be evaluated in relation to the stage at which they arise in the chain of problem representation, strategy, mathematical execution, reasoning, and justification, as well as their effects on subsequent steps.

Existing comparisons of human and generative AI contributions to mathematical problem solving have addressed different forms of support, participation, and performance. Jia et al. (2024) compared mathematical problem-solving hints written by teachers and generated by AI, finding that AI hints were perceived as useful but received lower ratings than teacher-written hints; evaluators could not consistently identify their source. This comparison concerned instructional hints rather than complete, independently produced student and AI solutions. Yoon et al. (2024) investigated three university students' interactions with generative AI during mathematical proving and emphasized critical evaluation of AI-generated arguments. Zhou et al. (2025), working with 29 university students, examined ChatGPT-supported scaffolding and found differences in how higher- and lower-performing students engaged with interpretation, evaluation, prompting, and procedural representations. These studies clarify how learners interact with AI support, but their focus does not establish how independently written student solutions compare with AI-generated solutions to identical problems. Getenet (2024) directly compared 58 preservice teachers and ChatGPT on the same word problem, reporting similarities in strategy use alongside frequent incorrect ChatGPT solutions; its focus on a single task and preservice teachers leaves room for examining multiple tasks and younger learners. Wei (2024) compared model accuracy on NAEP items with aggregated student performance, providing an outcome-level comparison rather than a paired analysis of written solution processes. Kortenkamp and Larkin (2026) studied student-like error patterns in repeated model responses to textbook problems and their reactions to instructional prompts,

without directly comparing those responses with independent written solutions from the same student group. Together, these studies establish the importance of both mathematical accuracy and observable solution processes while motivating a common, process-based comparison across human and multiple AI cases.

Exponent problems offer a focused yet mathematically varied context for such a comparison. Kieran (2004) emphasized the interpretation of symbolic relationships in algebra, while Usiskin (1988) discussed conceptual perspectives on school algebra relevant to exponent expressions. Scientific and decimal notation require attention to place value and numerical magnitude (Hiebert & Wearne, 1986); ratio tasks require multiplicative reasoning about quantities (Lamon, 2007); and area–perimeter and part–whole tasks involve mathematical visualization as discussed by Arcavi (2003) and conversion between representations as discussed by Duval (2006). Examining these demands within a common set of exponent problems makes it possible to identify where solutions succeed or break down in representation, strategy selection, execution, reasoning, and justification. In particular, tasks that require visual interpretation and mathematization provide opportunities to distinguish a correctly stated final answer from a mathematically reliable solution path.

The literature therefore identifies several complementary strands: Jia et al. (2024) studied AI-generated hints, Yoon et al. (2024) investigated learner–AI interactions during proving, and Zhou et al. (2025) examined ChatGPT-supported scaffolding. Getenet (2024) compared preservice teachers and ChatGPT on a word problem, whereas Wei (2024) compared model accuracy with aggregated student performance. Kortenkamp and Larkin (2026) examined model-generated error patterns and responses to instructional prompts. What remains less clearly documented is how independent middle school students' written solutions and the first responses of several generative AI systems to the same set of problems compare when evaluated through identical process-based criteria. In particular, the relationship among problem representation, strategy selection, mathematical execution, reasoning, justification, error propagation, and final-answer correctness requires closer examination across both student and AI products. Addressing this gap, the present study compares written solutions produced independently by students aged 12–13 with first responses from ChatGPT, Claude, Gemini, and Grok to the same ten exponent problems. Its purpose is to examine observable solution profiles and reliability, rather than to rank the tools solely by correct-answer rates.

In line with this purpose, the main research question is: What are the written solution profiles of students aged 12–13 and generative AI tools on the same mathematics problems, and how do these profiles differ in terms of representation, strategy, mathematical execution, reasoning, justification, error patterns, and solution reliability?

2. Theoretical Framework

The framework combines a process layer (problem representation, strategy, mathematical execution, and control) with a validation layer (epistemic, teleological, and communicative rationality) to evaluate observable solutions. The process layer draws on Pólya's (1945) stages of problem solving and Schoenfeld's (1985, 1992) account of strategic decisions and monitoring. Mathematical proficiency is informed by the National Research Council (2001), forms of understanding by Skemp (1976), and representation by Duval (2006). Stylianides (2007) provides a basis for evaluating mathematical justification, Toulmin (2003) for examining warrants in arguments, Radatz (1979) for classifying mathematical errors, and Borasi (1994) for interpreting errors as resources for investigating mathematical thinking. The validation layer focuses on the validity of a mathematical solution. Habermas (1998) provides the underlying account of rational action; Boero and Planas (2014) discuss rationality in mathematical activity. In the present framework, epistemic, teleological, and communicative rationality address, respectively, the validity of mathematical knowledge, the appropriateness of strategy for a mathematical goal, and the clarity of the solution account. Zhuang's (2026) validation approach likewise connects the result, method, mathematical grounds, and justifications within a single evaluation. This layer therefore provides a basis for examining the mathematical quality and reliability of the written solutions produced by students and generative AI.

2.1. Mathematical Problem-Solving Process

Pólya (1945) describes understanding, planning, execution, and checking; Schoenfeld (1985, 1992) emphasizes strategic decisions and monitoring. These accounts support analysis of representation, strategy, execution, and control in the written solutions. Mathematical proficiency includes conceptual understanding, procedural fluency, strategic competence, and adaptive reasoning (National Research Council, 2001). Skemp (1976) distinguishes instrumental rule use from relational understanding of why rules work. Duval (2006) emphasizes conversion among representation systems and preservation of mathematical relationships. Visual misreading and incorrect part–whole or contextual translations can produce representation errors, including in AI-supported mathematical problem solving (Bulut & Borromeo Ferri, 2026). Stylianides (2007) links the acceptability of a mathematical claim to the validity of the mathematical grounds supporting that claim. In Toulmin's (2003) model of argumentation, the warrant establishes the connection between data and claim. Radatz (1979) treats mathematical errors as indicators of students' conceptual and procedural structures, while Borasi (1994) regards errors as an analytical resource for investigating mathematical thinking. These perspectives provide the basis for examining the grounds of a solution and the error patterns that emerge during the process. Taken together, these components of problem solving require the mathematical validity of a solution to be evaluated in relation to knowledge, strategy, and communication.

2.2. Validity of Mathematical Solutions

The validity of a mathematical solution is related to the correctness of the knowledge used, the appropriateness of the selected strategy for the goal, and the clarity with which the solution is expressed. The components of epistemic, teleological, and communicative rationality, adapted to mathematics education from Habermas's (1998) account of rational action, address these three dimensions. Boero (2006) discussed rationality in mathematical proving, Boero and Planas (2014) considered rationality in mathematical activity, and Goizueta et al. (2014) addressed related issues in mathematical argumentation. Urhan and Bülbül (2023) also examined mathematical problem-solving processes through these components of rationality. Epistemic rationality concerns the validity of the mathematical knowledge and claims used in a solution. Building on Habermas's (1998) account of rationality and its mathematical interpretation by Boero and Planas (2014), definitions, rules, theorems, and mathematical relationships are evaluated for both correctness and appropriate application to the problem. Zhuang (2026) examined AI solutions in terms of the correctness of rules, their appropriate application, and the validity of justifications. This component provides a theoretical basis for mathematical execution, reasoning, justification, and error patterns in the present study. Teleological rationality concerns the appropriateness of a selected strategy for the mathematical goal. Evaluation focuses on whether the solution path fits the problem and serves the intended mathematical purpose. Zhuang (2026) considered the logical use of a strategy, attainment of the mathematical goal, and efficiency of the method within this dimension. Teleological rationality therefore serves as a theoretical basis for explaining strategy selection and solution reliability in this study. Communicative rationality concerns expressing mathematical thinking clearly and in ways consistent with disciplinary conventions. Boero and Planas (2014) connect communicative rationality with mathematical discourse, while Zhuang (2026) examines the clarity and traceability of AI-generated mathematical solutions. In this study, communicative rationality provides a basis for evaluating the clarity of solution steps and justifications in the written product. Zhuang (2026) relates solution reliability to the joint evaluation of the result, method, mathematical grounds, and justifications. In the present study, the problem-solving process is considered together with epistemic, teleological, and communicative rationality, and solution profiles are identified through relationships among problem representation, strategy, mathematical execution, reasoning, justification, error patterns, and outcome. This integrated framework explains how student and generative AI solutions represent the problem, select and implement a strategy, structure reasoning and justification, and differ across stages of the written solution.

3. Method

3.1. Research Design

This study employed a comparative case study design to compare observable features in the written solutions produced by students aged 12–13 and generative AI systems for the same mathematics problems. A comparative case study allows bounded cases to be examined first within their own contexts and then compared across common criteria to identify similarities and differences (Yin, 2018). The study does not claim direct access to students' internal cognitive processes or to the internal operations of AI systems; the comparison is limited to representation, procedural work, relation building, justification, and error features visible in the written products.

The cases consisted of students' independent written solutions to 10 problems and the sets of first responses generated by ChatGPT, Claude, Gemini, and Grok for the same problems. The boundaries of the cases were defined by the common set of 10 problems, students' independent written solutions produced in a single session, and the first responses generated by the AI systems under the same standard prompt. Each case was first examined internally, after which cross-case comparisons were conducted in terms of problem representation and mathematization, strategy, mathematical execution, reasoning, justification, error patterns, and solution reliability. Accordingly, the terms "reasoning," "strategy," and "justification" in this article refer to mathematical features observable in the written solutions rather than to mental or system-internal processes.

3.2. Participants and Research Context

The human data source consisted of 22 students aged 12–13 attending a public school in Istanbul, Türkiye. Eligibility was based on having previously received instruction on exponents and volunteering to participate in the study. The 10 problems used in the study were administered to all students in a single session under the same conditions. Students were asked to solve the problems individually and, in addition to marking an answer choice, to show as much as possible of the calculations, representations, and explanations used in their solutions. The solution sheets were anonymized; students were coded Ö1–Ö22 and problems S1–S10.

3.3. Data Collection Instrument

The study used 10 exponent problems selected from mathematics questions published between 2018 and 2025 in the nationwide centralized examination administered in Türkiye to eighth-grade students for transition to upper-secondary education. Exponents were selected as a common mathematical content domain because they require coordination of symbolic rules, numerical magnitude, and structural relationships while also supporting a range of contextual, visual, and quantitative problem types. Focusing on one content domain made it possible to compare students' and generative AI systems' written solutions under shared mathematical conditions while examining differences across problem-solving demands. The selected problems involved evaluating exponential expressions, applying exponent rules, interpreting negative exponents and signs, decimal decomposition, scientific notation, prime-factor relationships, ratio and unit reasoning, area and perimeter, part–whole relationships, visual representation, and comparison of quantities. This range enabled examination of problem representation, strategy selection, mathematical execution, reasoning, justification, error patterns, and solution reliability. The core mathematical demands of the selected problems are presented in Table 1.

Table 1. Core mathematical demands of the problem set

Problem	Mathematical content	Mathematical task
S1	Geometric structure, area, and exponent relationships	Interpret the visual structure; relate area to exponential notation

Problem	Mathematical content	Mathematical task
S2	Exponential expressions, ratio, and unit	Relate the data; set up a ratio; calculate using units
S3	Negative exponents and signs	Interpret conditions; compare negative-exponent and sign structures
S4	Prime factors and exponential representations	Factor into primes; compare equivalent exponential expressions
S5	Decimal representation and comparison	Decompose decimal structure; compare quantities
S6	Decimal decomposition and condition	Relate place value to powers of 10; compare according to the given condition
S7	Visual part–whole relationships and area	Identify equal/identical parts; translate visual relationships into area calculations
S8	Algebraic modeling with exponential expressions	Model the total quantity; use variables and exponential expressions together
S9	Area–perimeter and part placement	Mathematize the visual arrangement; establish length/perimeter relationships
S10	Scientific notation	Express decimals using powers of 10 and convert to scientific notation

3.4. Data Collection

Data were collected in two stages from students' solution sheets and the written outputs generated by four generative AI systems for the same problem set. In the first stage, the 10 problems were administered to students in a single session under the same conditions. Students were asked to solve the problems individually and to show the calculations, representations, and explanations they used alongside their selected answers. At the end of the session, the solution sheets were collected, anonymized, and coded Ö1–Ö22.

In the second stage, the same 10 problems given to the students were presented to ChatGPT, Claude, Gemini, and Grok. To support comparability across models, the same standard prompt was used with all four tools: "Solve this problem step by step, explain your solution, and justify your calculations." Each problem was presented to each tool in a separate new chat, and the model's first response was recorded. The versions used and the dates of data generation are presented in Table 2. The possibility that this prompt made intermediate steps and justifications more explicit in the AI outputs was taken into account in the analysis. In the Gemini response to S7 and the Grok response to S8, two alternative solution paths produced within the same response differed in solution path, error pattern, and outcome; these paths were therefore retained as separate solution segments. When code frequencies were calculated, alternative solution paths belonging to the same mathematical task were counted once at the task level.

Table 2. Generative AI systems and data-generation dates

Tool	Version	Data-generation date	Application protocol
Gemini	Gemini 3.5 Flash	20 May 2026	Each problem in a separate new chat; first response used
Claude	Claude Sonnet 4.6	21 May 2026	Each problem in a separate new chat; first response used
Grok	Grok (fast version)	21 May 2026	Each problem in a separate new chat; first response used

Tool	Version	Data-generation date	Application protocol
ChatGPT	GPT-5.5 Thinking (ChatGPT Plus)	22 May 2026	Each problem in a separate new chat; first response used

3.5. Analytical Framework

The analytical framework was developed inductively from patterns observed in the student and generative AI solutions. No predetermined coding list was initially imposed on the data; instead, meaningful data units reflecting the solution process were identified and initial codes were generated to describe them. The codes were repeatedly reviewed against new data segments in terms of their definitions, scope, and distinguishing features. By grouping similar codes, the analytical themes of problem understanding and mathematization, strategy selection, mathematical execution, reasoning, justification, error patterns, and solution reliability were developed.

Once the themes and subcodes had stabilized in relation to the data, they were interpreted in connection with Pólya (1945), Schoenfeld (1985, 1992), the National Research Council (2001), Skemp (1976), Duval (2006), Stylianides (2007), Toulmin (2003), Radatz (1979), Borasi (1994), and the rationality framework (Habermas, 1998; Boero & Planas, 2014; Zhuang, 2026). Thus, the initial codes and themes were derived from the data; the theoretical sources were used not as predetermined coding categories but to explain and interpret the emergent analytical themes. Table 3 presents the data-based analytical dimensions, the features examined within each dimension, and the associated theoretical framework.

Table 3. Data-based analytical dimensions and associated theoretical framework

Analytical dimension	Associated theoretical framework
Problem understanding	Givens, conditions, goals, mathematization, and preservation of structural relationships (Pólya, 1945; Schoenfeld, 1985, 1992; Duval, 2006).
Strategy selection	Organization and selection of a goal-appropriate mathematical path; teleological suitability (Pólya, 1945; Schoenfeld, 1985, 1992; Habermas, 1998).
Mathematical execution	Accurate application of exponent rules, numerical conversions, and geometric procedures; procedural fluency and epistemic validity (National Research Council, 2001; Habermas, 1998; Zhuang, 2026).
Reasoning	Procedural and relational connections, structural preservation, and monitoring (Skemp, 1976; Duval, 2006; Schoenfeld, 1992; Habermas, 1998).
Justification	Explicit grounds linking mathematical claims to data, rules, or conditions; mathematical validity and communicative clarity (Toulmin, 2003; Stylianides, 2007; Habermas, 1998; Boero & Planas, 2014).
Error patterns	Calculation, symbolic-rule, representation, and mathematization breakdowns, including propagation to later steps (Radatz, 1979; Borasi, 1994; Duval, 2006).
Solution reliability	Consistency across representation, strategy, execution, justification, and outcome; teleological, epistemic, and communicative validity (Habermas, 1998; Boero & Planas, 2014; Zhuang, 2026).

3.6. Data Analysis

Data analysis followed Braun and Clarke's (2006) six-phase approach to thematic analysis. In the first phase, the student and generative AI solutions were read repeatedly to become familiar with the dataset. In the second phase, each written solution was treated as the unit of analysis, and meaningful data units reflecting the solution process were coded inductively. In the third phase, similar codes were grouped to generate candidate analytical themes. In the fourth phase, these themes were reviewed against both the relevant coded data segments and the dataset as a whole. In the fifth phase, the scope and distinguishing features of the themes were clarified, the themes were named, and a codebook

containing inclusion and exclusion criteria and example data segments for each code was finalized. In the final phase, the themes were reported together with solution examples and code frequencies that supported cross-case comparison. In line with Braun and Clarke's (2006) emphasis, these phases were not treated as a rigid linear sequence; instead, analysis involved iterative movement among the data, codes, and themes. Because a single solution could contain more than one mathematical process feature, multiple codes were permitted within the same solution.

The analysis was supported by review from multiple researchers. The first two researchers examined the dataset together and developed the initial codes, candidate themes, and codebook. After the codebook had been developed, the third researcher re-examined the dataset using the same definitions and criteria. Data segments that received different interpretations were compared by all three researchers; the scope of the code, the meaning of the data segment, and its fit with the theme were discussed, and the final coding was determined by consensus. This process supported consistent application of the code and theme definitions across student and generative AI solutions.

After coding, the data were organized in matrices crossing the student, ChatGPT, Claude, Gemini, and Grok cases with problem numbers. Solution patterns within each case were examined first, followed by comparisons of similarities and differences across cases. The student dataset comprised 220 written solutions from 22 students across 10 tasks. Each AI case comprised the first response to 10 tasks. In the findings tables, student codes are reported as n/220 and percentages, whereas AI codes are reported as the number of tasks out of 10 in which the code appeared. These indicators were not treated as directly comparable equivalent proportions; they were used only to support the visibility of qualitative patterns within each case. Alternative solution paths in Gemini S7 and Grok S8 were retained as separate solution segments, but the same code was counted only once in task-level counts.

4. Findings

Findings are organized into four axes: representation and mathematization; strategy and execution; reasoning and justification; and error patterns and reliability. Codes describe observable features of written solutions, with examples showing where mathematical breakdowns occurred.

The analysis covered 220 student solutions and the first response of each AI system to 10 tasks. Multiple codes could occur in one solution. Student indicators are n/220 (%), whereas AI indicators are tasks/10; these denominators are not statistically equivalent. For AI, terms such as reasoning and justification refer only to observable output.

4.1. Problem Representation and Mathematization

Table 4 summarizes codes, frequencies, and examples for problem representation and mathematization.

Table 4. Codes, code frequencies, and example solutions for problem representation and mathematization

Theme	Code	Code visibility	Student example solution	AI example solutions
Problem representation	Mathematizing the context	Students: 98/220 (44.5%) ChatGPT: 10/10 Claude: 10/10 Gemini: 10/10 Grok: 10/10	Ö1–S2, Ö19–S7	ChatGPT–S2, Claude–S2, Gemini–S2, Grok–S2
Problem representation	Noticing the condition	Students: 79/220 (35.9%) ChatGPT: 4/10 Claude: 4/10 Gemini: 4/10 Grok: 4/10	Ö1–S3/S4	ChatGPT–S3–S6, Claude–S3–S6, Gemini–S3–S6, Grok–S3–S6
Problem representation	Identifying givens	Students: 40/220 (18.2%) ChatGPT: 9/10 Claude: 10/10 Gemini: 10/10 Grok: 10/10	Ö1–S2, Ö19–S7	ChatGPT–S1/S2, Claude–S1/S2, Gemini–S1/S2, Grok–S1/S2

Theme	Code	Code visibility	Student example solution	AI example solutions
Problem representation	Identifying the quantity to be compared	Students: 41/220 (18.6%) ChatGPT: 2/10 Claude: 2/10 Gemini: 2/10 Grok: 2/10	Ö1–S4, Ö3–S5	ChatGPT–S5/S6; Claude–S5/S6; Gemini–S5/S6; Grok–S5/S6
Problem representation	Unit awareness	Students: 16/220 (7.3%) ChatGPT: 1/10 Claude: 1/10 Gemini: 1/10 Grok: 1/10	Ö1–S2, Ö21–S2	ChatGPT–S2; Claude–S2; Gemini–S2; Grok–S2
Representation relationships	Establishing a part–whole relationship	Students: 18/220 (8.2%) ChatGPT: 3/10 Claude: 3/10 Gemini: 3/10 Grok: 3/10	Ö19–S7	ChatGPT–S1/S7/S9; Claude–S1/S7/S9; Gemini–S1/S7/S9; Grok–S1/S7/S9
Representation relationships	Recognizing equal/identical part relationships	Students: 11/220 (5%) ChatGPT: 2/10 Claude: 2/10 Gemini: 2/10 Grok: 2/10	Ö1–S7, Ö2–S7	ChatGPT–S7/S9; Claude–S7/S9; Gemini–S7/S9; Grok–S7/S9
Representation relationships	Visualizing the context	Students: 4/220 (1.8%) ChatGPT: 0/10 Claude: 0/10 Gemini: 0/10 Grok: 0/10	Ö1–S1, Ö18–S1	no clear example

Figure 1 compares student and AI solutions to the visual part–whole and area task S7.

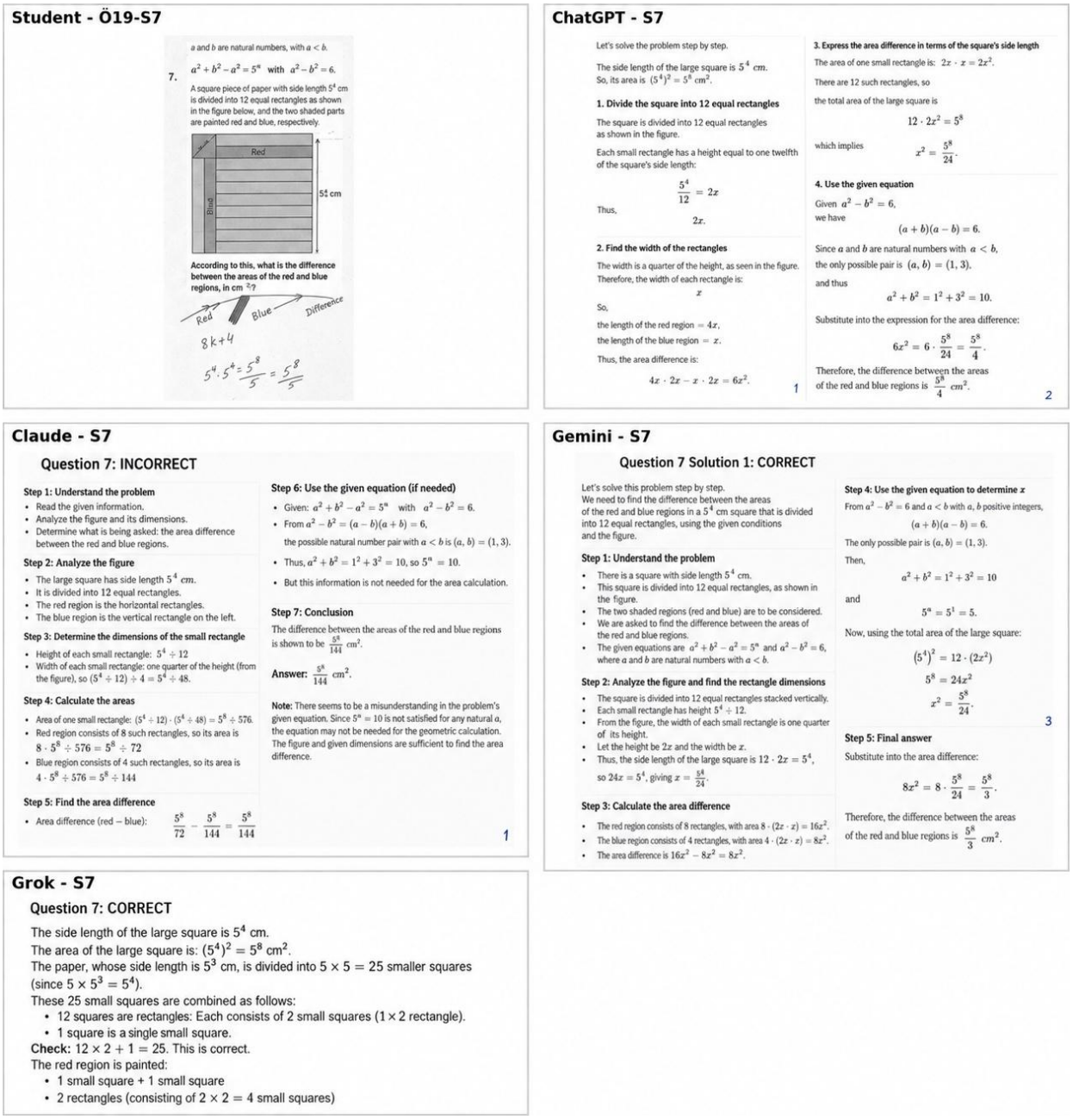


Figure 1. Comparative example solutions for Problem S7: (a) Ö19–S7, (b) ChatGPT–S7, (c) Claude–S7, (d) Gemini–S7, (e) Grok–S7

These codes represent the initial chain of the solution: the solver first selects which information in the problem is mathematically functional and then transforms that information into a mathematical structure expressed as quantities, relationships, units, figures, or symbols. Pólya's (1945) problem-understanding phase was used to explain why givens, goals, and conditions were examined separately at the beginning of a solution. Schoenfeld's (1985) account of problem representation provides a theoretical basis for explaining how the representation constructed at the outset can influence subsequent strategic decisions. Accordingly, the codes in this section describe the quality of the mathematical model constructed before later procedural steps.

In student solutions, problem information was processed primarily by translating the context into a mathematical relationship and recognizing the condition of the problem. Visualizing the context was less common. In those instances, students reconstructed the relationship in the problem through diagrams, partitions, or markings. Arcavi (2003) describes visual representations as tools for noticing and organizing mathematical relationships. Accordingly, the visual structures in students' solutions were evaluated in terms of the extent to which they preserved the relevant mathematical relationships.

Problem representation in the AI tools was more explicit and more clearly sequenced. In tasks requiring part–whole or equal-part relationships, parts of the figure were counted, quantities were described verbally, and this information was converted into formulas or equations. From the perspective of communicative rationality, an explicit and traceable representation makes a solution easier to follow. Mathematical validity, however, depends on whether the constructed representation is consistent with the structure of the problem (Habermas, 1998).

Across cases, the decisive issue was whether the representation preserved the original quantitative and structural relationships (Duval, 2006). Misreading a figure could carry into subsequent calculations.

The visual part–whole and area problem in Figure 1 illustrates how the same mathematical structure was represented by a student and four generative AI tools. In the student solution, the large square was reconstructed through visual partitions and markings; the part–whole relationship was established directly on the figure, and the area relationship was expressed through a short solution chain. ChatGPT and Grok described the number and dimensions of the parts through verbal-symbolic steps and explicitly connected the total number of parts to the area. Claude initially misinterpreted the visual structure, and this interpretation carried into subsequent mathematization. Gemini's alternative solution paths contained errors involving the number of parts, the part–whole relationship, and counting.

These cases show that explicitness alone cannot establish validity: representation errors can propagate through a fluent solution (Duval, 2006; Zhuang, 2026).

4.2. Strategy Development and Mathematical Execution

Table 5 summarizes strategy-development codes and examples.

Table 5. Codes, code frequencies, and example solutions for strategy development

Theme	Code	Code visibility	Student example solution	AI example solutions
Strategy development	Organizing data	Students: 119/220 (54.1%) ChatGPT: 10/10 Claude: 10/10 Gemini: 10/10 Grok: 10/10	Ö1–S2/S4, Ö3–S5	ChatGPT–S1–S10; Claude–S1–S10; Gemini–S1–S10; Grok–S1–S10
Strategy development	Selecting an appropriate rule/formula	Students: 84/220 (38.2%) ChatGPT: 8/10 Claude: 8/10 Gemini: 8/10 Grok: 8/10	Ö1–S3/S4, Ö17–S10	ChatGPT–S1–S4/S7–S10; Claude–S1–S4/S7–S10; Gemini–S1–S4/S7–S10; Grok–S1–S4/S7–S10
Strategy development	Decomposing the problem into subproblems	Students: 2/220 (0.9%) ChatGPT: 7/10 Claude: 8/10 Gemini: 8/10 Grok: 6/10	Ö1–S2, Ö17–S7	ChatGPT–S2/S4/S5/S7–S10; Claude: 8 tasks; Gemini: 8 tasks; Grok: 6 tasks
Strategy development	Making comparisons	Students: 51/220 (23.2%) ChatGPT: 4/10 Claude: 4/10 Gemini: 4/10 Grok: 4/10	Ö3–S5, Ö5–S6	ChatGPT–S3–S6; Claude–S3–S6; Gemini–S3–S6; Grok–S3–S6

Theme	Code	Code visibility	Student example solution	AI example solutions
Strategy development	Setting up a ratio	Students: 7/220 (3.2%) ChatGPT: 1/10 Claude: 0/10 Gemini: 2/10 Grok: 0/10	Ö21–S7	ChatGPT–S7 Gemini–S2/S7
Strategy diversity	Constructing a table or list	Students: 0/220 (0%) ChatGPT: 1/10 Claude: 1/10 Gemini: 1/10 Grok: 1/10	no clear example	ChatGPT–S3; Claude–S3; Gemini–S3; Grok–S3
Strategy diversity	Working backward	Students: 0/220 (0%) ChatGPT: 1/10 Claude: 0/10 Gemini: 0/10 Grok: 1/10	no clear example in the coding matrix	ChatGPT–S5, Grok–S5
Strategy diversity	Trial and error	Students: 0/220 (0%) ChatGPT: 0/10 Claude: 1/10 Gemini: 1/10 Grok: 0/10	no clear example	Claude–S5, Gemini–S5

Table 6 summarizes mathematical-execution codes and examples.

Table 6. Codes, code frequencies, and example solutions for mathematical execution

Theme	Code	Code visibility	Student example solution	AI example solutions
Mathematical execution	Applying exponent rules	Students: 65/220 (29.5%) ChatGPT: 5/10 Claude: 5/10 Gemini: 6/10 Grok: 5/10	Ö1–S1/S2/S3	ChatGPT–S1/S3/S7–S9; Claude–S1/S3/S7–S9; Gemini–S1–S3/S7–S9; Grok: 5 tasks
Mathematical execution	Calculating the value of an exponential expression	Students: 45/220 (20.5%) ChatGPT: 6/10 Claude: 6/10 Gemini: 5/10 Grok: 6/10	Ö1–S1/S3/S4	ChatGPT–S1–S4/S8/S9; Claude: 6 tasks; Gemini: 5 tasks; Grok: 6 tasks
Mathematical execution	Comparing quantities	Students: 51/220 (23.2%) ChatGPT: 4/10 Claude: 4/10 Gemini: 4/10 Grok: 4/10	Ö1–S4/S5, Ö3–S5	ChatGPT–S3–S6; Claude–S3–S6; Gemini–S3–S6; Grok–S3–S6
Mathematical execution	Converting to decimal form	Students: 34/220 (15.5%) ChatGPT: 2/10 Claude: 2/10 Gemini: 2/10 Grok: 2/10	Ö1–S5/S6, Ö4–S5	ChatGPT–S5/S6; Claude–S5/S6; Gemini–S5/S6; Grok–S5/S6
Mathematical execution	Proportional calculation	Students: 13/220 (5.9%) ChatGPT: 1/10 Claude: 1/10 Gemini: 2/10 Grok: 1/10	Ö1–S2, Ö2–S7	ChatGPT–S2; Claude–S2; Gemini–S2/S7; Grok–S2
Mathematical execution	Calculating area	Students: 10/220 (4.5%) ChatGPT: 1/10 Claude: 1/10 Gemini: 1/10 Grok: 1/10	Ö4–S7, Ö19–S7	ChatGPT–S7; Claude–S7; Gemini–S7; Grok–S7

Theme	Code	Code visibility	Student example solution	AI example solutions
Mathematical execution	Calculating perimeter	Students: 9/220 (4.1%) ChatGPT: 1/10 Claude: 1/10 Gemini: 1/10 Grok: 1/10	Ö6–S9	ChatGPT–S9; Claude–S9; Gemini–S9; Grok–S9
Mathematical execution	Converting to scientific notation	Students: 11/220 (5%) ChatGPT: 1/10 Claude: 1/10 Gemini: 1/10 Grok: 1/10	Ö6–S10	ChatGPT–S10; Claude–S10; Gemini–S10; Grok–S10

Figure 2 compares solutions to the decimal representation and comparison task S5.

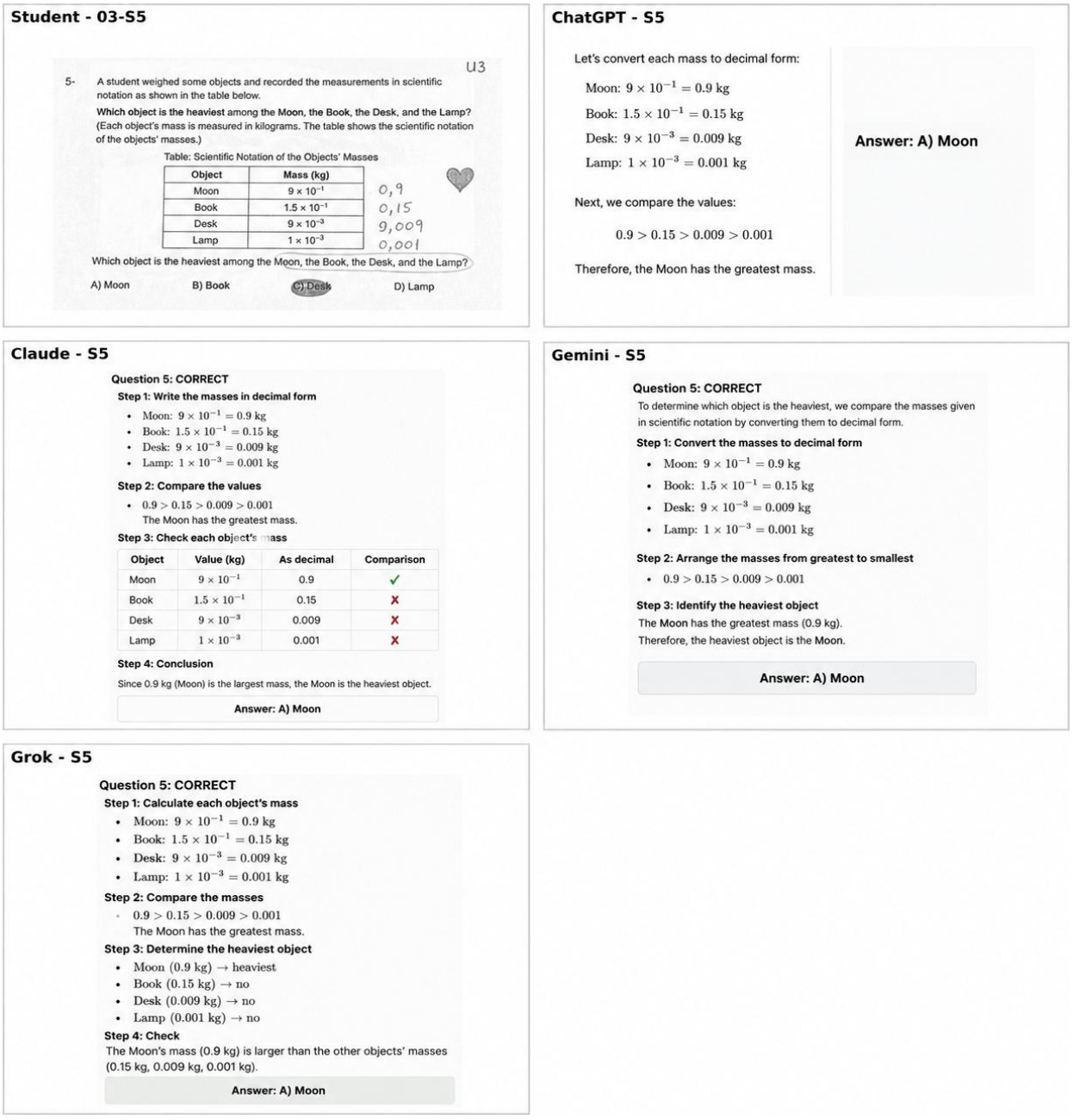


Figure 2. Comparative example solutions for Problem S5: (a) Ö3–S5, (b) ChatGPT–S5, (c) Claude–S5, (d) Gemini–S5, (e) Grok–S5

The mathematical-execution dimension captures how the selected path is carried out through operations such as applying exponent rules, calculating exponential expressions, decimal conversion, proportional calculation, area–perimeter calculation, quantity comparison, and scientific notation. This section therefore focuses on two successive questions: Was the solution path selected appropriately for the mathematical goal, and was that path executed correctly? Pólya's (1945) separation of devising a plan from carrying out the plan provides a theoretical basis for treating strategy selection and mathematical execution as distinct. Schoenfeld (1985) likewise distinguishes strategic decisions from the monitoring and execution of those decisions as different problem-solving functions.

In student solutions, strategy development was most visible in organizing data, selecting an appropriate rule or formula, and comparing relevant quantities. In the problem involving decimal representation and quantity comparison, the required conversions and comparisons were combined into a short procedural chain. In one student solution comparing prime factors and equivalent exponential expressions, the problem condition was identified correctly and an appropriate path was selected, but an error in applying the exponent rule changed the outcome. These examples show that students generally constructed the solution path within a compressed sequence of procedures and that strategic decisions and mathematical execution served distinct functions within the solution.

In the AI tools, strategy development was more explicit through the reorganization of data and decomposition of the problem into intermediate goals. In most tasks, the tools named an appropriate rule or formula, identified the comparison criterion, and divided the solution path into substeps. Depending on the task, strategies such as constructing a table or list, working backward, and trial and error were also expressed explicitly. The code frequencies in the table support that AI outputs verbalized intermediate goals, options, and groups of data more often than student solutions. This structure made the solution path easier to trace. From the perspective of teleological rationality, the primary criterion is whether the selected strategy is appropriate for the mathematical goal (Habermas, 1998). Some solutions that presented an explicit and detailed strategy still contained mathematical inconsistencies. Strategy appropriateness and efficiency were therefore evaluated together with subsequent solution steps.

Applying exponent rules, calculating the value of an exponential expression, comparing quantities, and converting to decimal form appeared in both student and AI solutions depending on the mathematical demand of the task. The numerical indicators in the table show that AI tools more often wrote these procedures as separate steps, whereas students presented the same procedures in shorter, more integrated chains. Proportional calculation, area–perimeter calculation, and scientific notation also appeared according to the structure of the relevant problems. The National Research Council (2001) conceptualizes procedural fluency as the capacity to carry out appropriate procedures accurately, flexibly, and efficiently. Within this framework, the quality of execution was evaluated less by the amount of written detail than by the correctness of the mathematical knowledge used and its consistency with the structure of the problem.

Tasks involving negative exponents and signs required accurate reading of symbolic structure; decimal notation required preservation of the relationship between place value and powers of 10; and ratio problems required establishing the multiplicative relationship between quantities. Hiebert and Wearne (1986) showed that procedural knowledge and conceptual structure can diverge in decimal-number knowledge. Lamon (2007) emphasizes that proportional reasoning requires establishing a multiplicative relationship between two quantities. Mathematical-execution codes therefore capture more than whether students and AI tools completed a procedure; they also indicate whether the procedure was consistent with the underlying mathematical structure.

The decimal representation and quantity-comparison problem in Figure 2 illustrates how the written visibility of strategy and mathematical execution differed. The student carried out the necessary mass conversions, identified the quantities to be compared, and reached the correct answer. Some intermediate decisions were expressed only in short lines of calculation. ChatGPT, Claude, Gemini, and Grok calculated the quantities in separate steps, stated the comparison condition explicitly, and evaluated the options in sequence. Strategy and execution appeared in a more compressed form in the student solution, whereas intermediate steps were written in greater detail in the generative

AI outputs. In this problem, the key determinant of mathematical quality was whether decimal conversions and quantity comparisons were carried out consistently with the problem condition.

Together, the findings distinguish strategic appropriateness (teleological validity) from correct mathematical application (epistemic validity), while written clarity concerns communicative rationality (Habermas, 1998; Zhuang, 2026).

4.3. Reasoning and Justification

Table 7 summarizes reasoning codes and examples.

Table 7. Codes, code frequencies, and example solutions for reasoning

Theme	Code	Code visibility	Student example solution	AI example solutions
Reasoning	Procedural reasoning	Students: 80/220 (36.4%) ChatGPT: 3/10 Claude: 3/10 Gemini: 2/10 Grok: 3/10	Ö1–S2/S3/S4, Ö12–S8	ChatGPT–S2/S4/S8, Claude–S2/S4/S8, Gemini–S2/S8, Grok–S2/S4/S8
Reasoning	Mental/implicit reasoning	Students: 63/220 (28.6%) ChatGPT: 0/10 Claude: 0/10 Gemini: 0/10 Grok: 0/10	Ö3–S4/S5, Ö5–S1	no clear example
Reasoning	Contextual reasoning	Students: 47/220 (21.4%) ChatGPT: 6/10 Claude: 6/10 Gemini: 6/10 Grok: 6/10	Ö1–S2/S5, Ö19–S6	ChatGPT–S1/S2/S7–S10; Claude–S1/S2/S7–S10; Gemini–S1/S2/S7–S10; Grok–S1/S2/S7–S10
Reasoning	Comparative reasoning	Students: 27/220 (12.3%) ChatGPT: 4/10 Claude: 4/10 Gemini: 3/10 Grok: 4/10	Ö1–S4, Ö12–S4	ChatGPT–S3–S6; Claude–S3–S6; Gemini–S3–S6; Grok–S3–S6
Reasoning	Relational reasoning	Students: 20/220 (9.1%) ChatGPT: 0/10 Claude: 0/10 Gemini: 0/10 Grok: 0/10	Ö19–S7, Ö20–S10	no clear example
Reasoning	Visual reasoning	Students: 16/220 (7.3%) ChatGPT: 1/10 Claude: 1/10 Gemini: 1/10 Grok: 1/10	Ö12–S9, Ö19–S7	ChatGPT–S7; Claude–S7; Gemini–S7; Grok–S7
Reasoning	Proportional reasoning	Students: 11/220 (5%) ChatGPT: 2/10 Claude: 1/10 Gemini: 2/10 Grok: 1/10	Ö17–S2, Ö2–S7	ChatGPT–S2/S7 Claude–S2 Gemini–S2/S7 Grok–S2
Reasoning	Validating reasoning	Students: 0/220 (0%) ChatGPT: 1/10 Claude: 1/10 Gemini: 1/10 Grok: 0/10	no clear example	ChatGPT–S1, Claude–S5, Gemini–S5

Table 8 summarizes justification codes and examples.

Table 8. Codes, code frequencies, and example solutions for justification

Theme	Code	Code visibility	Student example solution	AI example solutions
Justification	Procedural justification	Students: 86/220 (39.1%) ChatGPT: 1/10 Claude: 2/10 Gemini: 2/10 Grok: 2/10	Ö1–S1/S2/S3	ChatGPT–S2 Claude–S1/S2 Gemini–S1/S2 Grok–S1/S2
Justification	Context-based justification	Students: 11/220 (5%) ChatGPT: 6/10 Claude: 5/10 Gemini: 6/10 Grok: 5/10	Ö2–S1, Ö19–S7	ChatGPT–S2/S5–S9; Claude: 5 tasks; Gemini: 6 tasks; Grok: 5 tasks
Justification	Visual justification	Students: 7/220 (3.2%) ChatGPT: 0/10 Claude: 0/10 Gemini: 0/10 Grok: 0/10	Ö1–S1, Ö19–S7	no clear example
Justification	Comparison-based justification	Students: 5/220 (2.3%) ChatGPT: 2/10 Claude: 2/10 Gemini: 2/10 Grok: 2/10	Ö1–S4, Ö2–S4	ChatGPT–S3/S4; Claude–S3/S4; Gemini–S3/S4; Grok–S3/S4
Justification	No justification	Students: 7/220 (3.2%) ChatGPT: 0/10 Claude: 0/10 Gemini: 0/10 Grok: 0/10	Ö5–S1 and similar brief/implicit solutions	no clear example
Justification	Explaining why the rule works	Students: 0/220 (0%) ChatGPT: 2/10 Claude: 2/10 Gemini: 5/10 Grok: 0/10	no clear example	ChatGPT–S3/S6 Claude–S3/S5 Gemini–S5/S6/S8–S10
Justification	Justification by naming the rule	Students: 0/220 (0%) ChatGPT: 0/10 Claude: 0/10 Gemini: 4/10 Grok: 1/10	no clear example	Gemini–S1/S2/S3/S10 Grok–S3
Justification	Table-based justification	Students: 0/220 (0%) ChatGPT: 1/10 Claude: 1/10 Gemini: 1/10 Grok: 1/10	no clear example	ChatGPT–S3; Claude–S3; Gemini–S3; Grok–S3
Justification	Condition-based justification	Students: 0/220 (0%) ChatGPT: 3/10 Claude: 3/10 Gemini: 3/10 Grok: 3/10	no clear example	ChatGPT–S5/S6/S10; Claude–S5/S6/S10; Gemini–S5/S6/S10; Grok–S5/S6/S10
Justification	Justification by checking	Students: 0/220 (0%) ChatGPT: 1/10 Claude: 1/10 Gemini: 0/10 Grok: 0/10	no clear example	ChatGPT–S1, Claude–S7

Figure 3 compares solutions to a decimal-place-value and maximum-limit task.

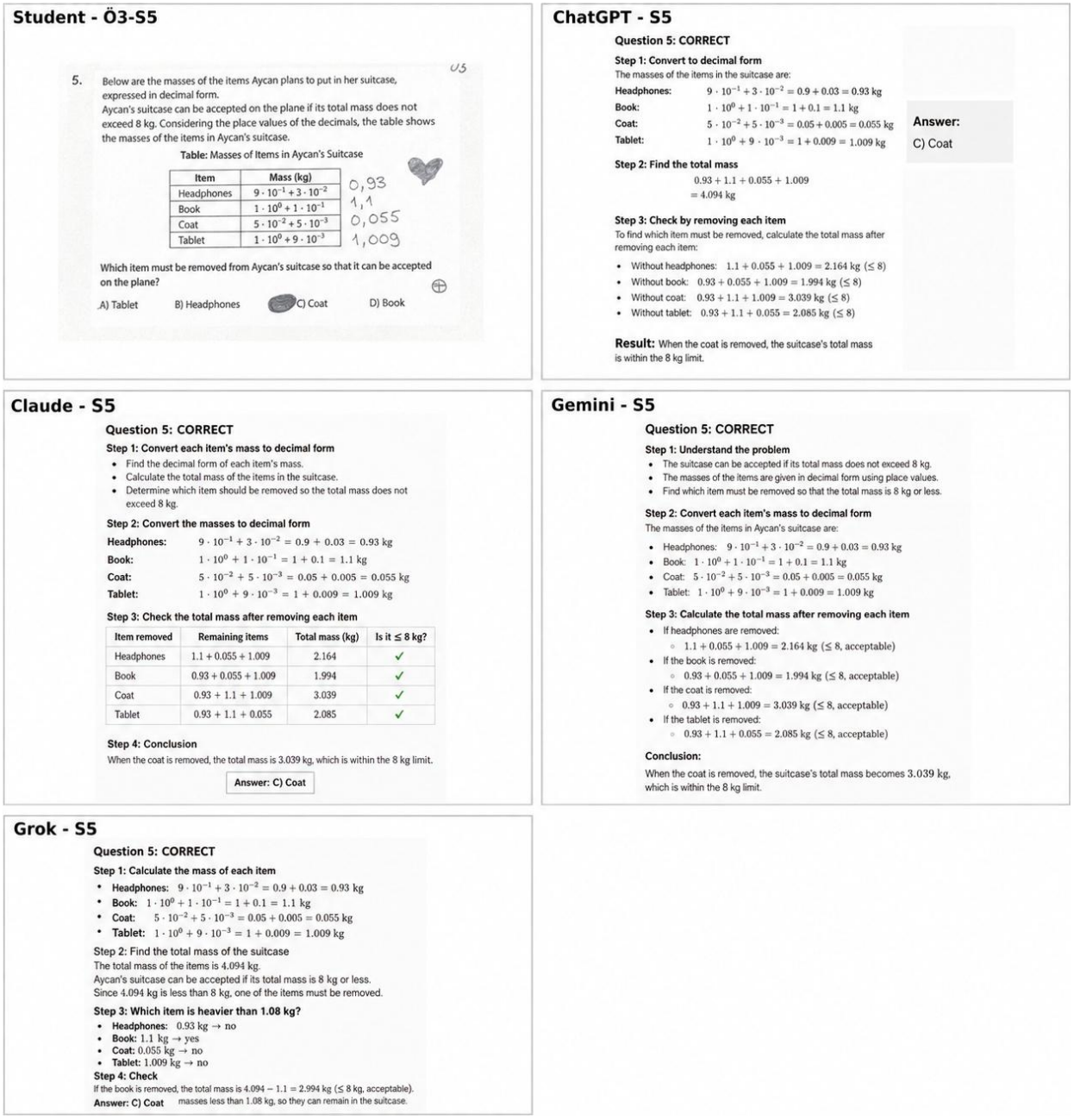


Figure 3. Comparative example solutions for a suitcase-mass problem involving decimal notation: (a) student solution, (b) ChatGPT, (c) Claude, (d) Gemini, (e) Grok

These codes indicate which mathematical relationships became visible in the written solution. The justification dimension examines the mathematical basis used to support a procedure, rule, solution path, or result through codes for justification by procedure, context, visual representation, comparison, naming a rule, explaining why a rule works, using a table, invoking a condition, or checking. Skemp (1976) describes instrumental understanding as applying a rule and relational understanding as understanding why the rule is valid and how concepts are connected. This distinction provides a basis for interpreting reasoning patterns in terms of the quality of the mathematical connections used. Stylianides (2007) links the acceptability of a mathematical claim to the validity of the mathematical foundation supporting it.

Contextual reasoning appeared in solutions that connected quantities to their meaning in the problem. Comparative and relational reasoning appeared in examples that directly connected quantities, representations, or rules. This pattern suggests that some students embedded justification within the procedural sequence rather than expressing it as a separate explanation. The code frequencies in the table likewise indicate that justification through procedures was the dominant form in student solutions, while justifications based on context, visual representation, or comparison appeared in fewer cases. The mental/implicit code refers to solutions in which intermediate mathematical connections remained unstated in the written product.

ChatGPT, Claude, and Grok explicitly stated comparison criteria in many solutions; similar examples appeared more selectively in Gemini. Validating reasoning occurred in specific tasks in ChatGPT, Claude, and Gemini when a result or intermediate calculation was checked again. Mental/implicit and relational reasoning codes appeared infrequently in AI outputs. This pattern indicates that AI outputs more extensively verbalized mathematical relationships through explanations, comparisons, and checking statements. The mathematical quality of reasoning was evaluated in terms of the fit between the relationship used and the structure of the problem rather than the length of the explanation.

The tools more often grounded procedures in the problem context or problem condition; in some solutions they explained why a rule was valid, named the rule, used a table, or justified the result by checking it. Context-based justifications were more visible in ChatGPT and Gemini. Gemini more frequently used statements that explained why a rule worked and named the rule. Justifications through checking appeared in ChatGPT and Claude. The numerical indicators in the table support that AI outputs more often made the connection between a procedure and its justification explicit than did student solutions. In Toulmin's (2003) model of argumentation, a warrant links data to a claim. In this study, the visibility of a justification and the validity of the mathematical grounds used were examined separately.

The suitcase-mass problem in Figure 3 shows how reasoning and justification were expressed in a task requiring decimal interpretation, total-mass calculation, and evaluation of a limiting condition. In the student solution, the masses were converted and compared through a short sequence of calculations, and the answer was reached with minimal written explanation. ChatGPT, Claude, and Gemini rewrote the item masses in decimal form, calculated the total mass, and justified the answer by checking whether the suitcase would remain within the 8 kg limit after removing each option. Grok also organized the solution step by step, but it focused on identifying an item above a threshold rather than consistently applying the total-mass condition, which produced a different process–outcome pattern. This example shows that reasoning quality in quantity problems depends not only on computing correct values but also on maintaining the governing condition of the task and using that condition to justify the final choice.

Thus, solution clarity and justification validity were evaluated separately; the same prompt explicitly requested explanations from all AI systems.

4.4. Error Patterns and Solution Reliability

Table 9 summarizes error codes, frequencies, and examples.

Table 9. Codes, code frequencies, and example solutions for error patterns

Theme	Code	Code visibility	Student example solution	AI example solutions
Procedural/conceptual error	Calculation error	Students: 11/220 (5%) ChatGPT: 0/10 Claude: 0/10 Gemini: 0/10 Grok: 0/10	Ö1–S4, Ö6–S7, Ö13–S2	no clear example
Procedural/conceptual error	Incorrect use of an exponent rule	Students: 9/220 (4.1%) ChatGPT: 0/10 Claude: 0/10 Gemini: 0/10 Grok: 0/10	Ö1–S3/S4, Ö14–S2	no clear example

Theme	Code	Code visibility	Student example solution	AI example solutions
Procedural/conceptual error	Decimal-decomposition error	Students: 9/220 (4.1%) ChatGPT: 0/10 Claude: 0/10 Gemini: 0/10 Grok: 0/10	Ö6–S5/S6, Ö7–S5	no clear example
Procedural/conceptual error	Negative-exponent and sign error	Students: 4/220 (1.8%) ChatGPT: 0/10 Claude: 0/10 Gemini: 0/10 Grok: 0/10	Ö18–S3, Ö20–S3, Ö21–S3	no clear example
Representational error	Visual/figure-reading error	Students: 1/220 (0.5%) ChatGPT: 0/10 Claude: 1/10 Gemini: 1/10 Grok: 0/10	Ö1–S1	Claude–S7 Gemini–S7.1/S7.2
Representational error	Part–whole relationship error	Students: 0/220 (0%) ChatGPT: 0/10 Claude: 0/10 Gemini: 1/10 Grok: 0/10	no clear example	Gemini–S7.1/S7.2
Representational error	Counting error	Students: 0/220 (0%) ChatGPT: 0/10 Claude: 0/10 Gemini: 1/10 Grok: 0/10	no clear example	Gemini–S7.1/S7.2
Mathematization error	Incorrect mathematization of context	Students: 1/220 (0.5%) ChatGPT: 0/10 Claude: 2/10 Gemini: 2/10 Grok: 2/10	Ö1–S2	Claude–S7/S9 Gemini–S7/S9 Grok–S8/S9

Figure 4 compares solutions to the area–perimeter and part-placement task S9.

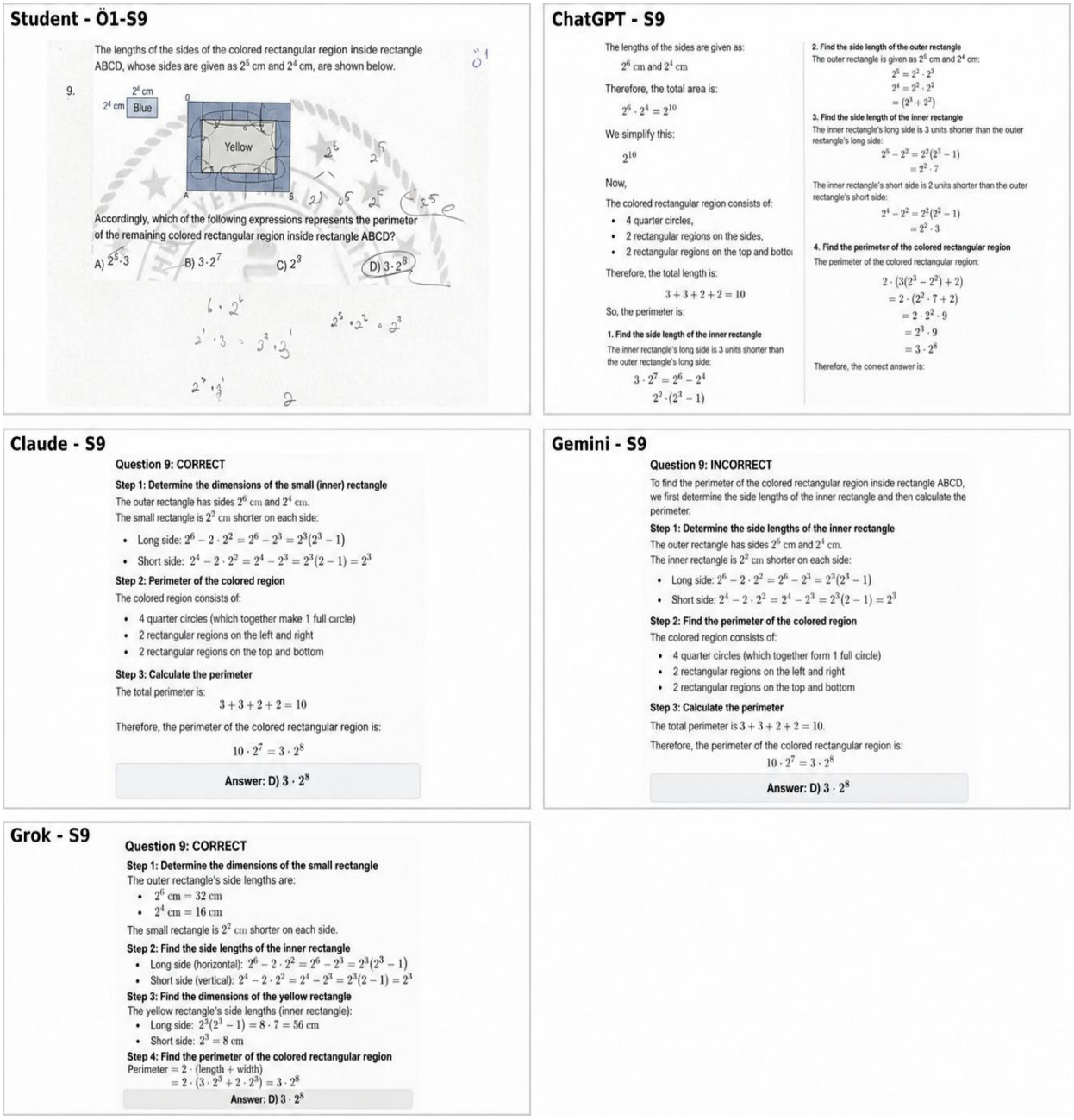


Figure 4. Comparative example solutions for Problem S9: (a) Ö1–S9, (b) ChatGPT–S9, (c) Claude–S9, (d) Gemini–S9, (e) Grok–S9

Error codes distinguish calculation, exponent-rule, decimal-decomposition, negative-exponent/sign, area–perimeter, counting, and mathematization errors. Reliability was judged by considering the solution path, mathematical grounds, justification, and final answer together (Radatz, 1979; Borasi, 1994; Zhuang, 2026).

Calculation errors, incorrect use of exponent rules, errors in decimal decomposition, and incorrect treatment of the relationship among negative exponents and signs were the most prominent forms in the student data. Visual or figure-reading errors and incorrect mathematization of context occurred in fewer cases. Problems involving negative exponents and signs, exponent rules, equivalent exponential expressions based on prime factors, and decimal decomposition made symbolic and procedural breakdowns particularly visible. In some solutions, the problem was

represented correctly and an appropriate strategy was selected, but an error during execution changed the outcome. Some solutions containing an error code nevertheless ended with the correct answer. This pattern shows that answer correctness and the mathematical coherence of the process can diverge. Borasi (1994) argues that errors can be examined as sources of data that reveal the structure of mathematical thinking. The present finding therefore underscores the importance of examining the procedural or conceptual chain through which a correct answer is reached.

ChatGPT's solutions to the tasks examined were mathematically coherent. In some Claude, Gemini, and Grok solutions, the context was mathematized incorrectly. Visual or figure-reading errors occurred in Claude and Gemini. Errors involving part–whole relationships and counting appeared in some of Gemini's solution paths. These forms of error were made visible by the visual part–whole and area problem, the problem combining area–perimeter relationships with part placement, and the algebraic-modeling problem involving exponential expressions. Duval (2006) regards preservation of mathematical structure during conversions between representation systems as fundamental to understanding. In the visual part–whole problem, representation-level errors in some Claude and Gemini solutions carried over into subsequent area or part calculations. The calculations followed an orderly sequence, but the mathematical model remained incorrect. This pattern indicates that, for some problem types, breakdowns in generative AI occurred before calculation, when the visual relationship was read and the context was transformed into a mathematical model.

Alternative paths in Gemini S7 and Grok S8 demonstrated that one first response could contain different strategies, errors, or outcomes. Such paths were retained analytically but counted once per code at task level.

The area–perimeter and part-placement problem in Figure 4 illustrates the relationship between process accuracy and outcome within the same mathematical content. The student identified the givens and the part–whole relationship, selected an appropriate formula, and reached an incorrect result because of an exponent-rule and calculation error. ChatGPT decomposed the problem into subproblems, established the relationship between perimeter and number of parts coherently, and reached the correct answer. In the Claude and Grok solutions, an incorrect-mathematization code co-occurred with the correct final answer. In the Gemini solution, incorrect mathematization also produced an incorrect result. The fact that the same mathematical content produced different process–outcome patterns across the five solution sources shows that solution reliability should be evaluated by considering problem representation, execution, and outcome together.

Within a single solution, teleological, epistemic, and communicative qualities could occur at different levels. Teleological rationality concerns whether the selected path is appropriate for the mathematical goal. In the student solution to the area–perimeter and part-placement problem, the strategy focusing on the part–whole and perimeter relationships provided an appropriate starting point, but the subsequent exponent calculation was carried out incorrectly. Epistemic rationality concerns the validity of the mathematical knowledge, rules, and claims used; in this example, the stage that undermined reliability was mathematical execution. Communicative rationality concerns whether the solution is expressed clearly and traceably. For the same problem, Claude and Grok produced orderly solution texts and reached the correct option, yet the coded incorrect mathematization weakened the epistemic coherence of the process. In Gemini, the representation error was visible in both the process and the outcome. Zhuang (2026) recommends that validation of AI solutions separately examine the appropriateness of the method for the goal, the validity of the mathematical knowledge used, and the adequacy of the justifications. The findings indicate that this multidimensional approach to validation is also necessary for student solutions.

Errors differed by mathematical demand: student work most visibly involved symbolic procedures, while some AI outputs failed in visual interpretation and contextual mathematization. The location of the error mattered for evaluating reliability.

5. Discussion and Conclusion

Students generally represented problem information through brief symbolic expressions, calculations, or visual markings, moved directly from the solution path to execution, and embedded their justifications within procedural

sequences. Generative AI tools, influenced in part by the standard prompt, produced more detailed textual solutions that named the givens, decomposed the problem into substeps, and made justifications visible in separate statements. Daher and Gierdien (2024) found that ChatGPT can produce mathematically incorrect results within well-sequenced and linguistically organized solution texts. Jia et al. (2024) showed that AI-generated problem-solving hints were rated lower than hints produced by human teachers and that evaluators could not consistently identify the source of the hints. These findings support evaluating linguistic clarity separately from mathematical and pedagogical quality.

The common prompt asked AI systems to solve step by step, explain, and justify; students received different instructions. Accordingly, differences in the amount of visible explanation cannot be interpreted as intrinsic reasoning superiority.

The findings on problem representation and mathematization show that how problem information is carried into a solution shapes subsequent stages. According to Duval (2006), mathematical understanding depends on preserving mathematical structure across transformations between representations. Spreitzer et al. (2024) reported that ChatGPT's performance declines as contextual and mathematical complexity increases. Cevikbas and Kaiser (2026) found that large language models can experience difficulty processing complex problem situations and visual information. Similar difficulties have been reported by Bulut and Borromeo Ferri (2026) in visually intensive modeling tasks and by Wardat et al. (2023) in interpreting relationships in geometry problems. Busuttil and Calleja (2025) reported that secondary mathematics teachers viewed ChatGPT's capacity to interpret visual mathematical representations as an important limitation. Ergene and Ergene (2025) showed that performance varies by model and problem complexity. The present findings indicate that representation form, mathematical content, and model characteristics jointly shape solution reliability in visual and contextual problems.

A second central finding is the distinction between strategy selection and mathematical execution. In some generative AI solutions, the overall solution path was appropriate for the mathematical goal, but an error in the initial mathematization undermined the validity of the solution. Pólya (1945) treats devising a plan and carrying out the plan as separate stages. Schoenfeld (1985, 1992) describes selecting, implementing, and monitoring a strategy as distinct problem-solving functions. The National Research Council (2001) likewise defines strategic competence and procedural fluency as separate components. Getenet (2024) showed that preservice teachers and ChatGPT can use similar strategies and that many of ChatGPT's solutions generated through those strategies were incorrect. The findings therefore support examining the suitability of a solution path separately from the mathematical correctness with which it is carried out.

Reasoning and justification became visible in different ways in student and generative AI solutions. In student solutions, intermediate connections often remained implicit within procedural sequences. Skemp's (1976) distinction between instrumental and relational understanding helps explain the mathematical quality of these connections. Stylianides (2007) emphasizes that a claim must be supported by valid mathematical grounds, and Toulmin (2003) identifies the warrant as the link between data and claim. Schorcht et al. (2024) found that prompts requesting step-by-step solutions made strategies more visible while retrospective evaluation remained infrequent. Daher and Gierdien (2024) showed that an explanatory solution text can still contain an incorrect mathematical relationship. Accordingly, the written clarity of a justification and its mathematical validity were treated as separate analytical features.

Error patterns show that the stage at which a solution breaks down differs between the student and generative AI data. Radatz (1979) and Borasi (1994) argue that mathematical errors can help reveal the conceptual and procedural structures used by the solver. Kortenkamp and Larkin (2026) showed that large language model outputs can produce persuasive errors resembling student responses. Sapkota and Bondurant (2024) found that some mathematical tasks generated by ChatGPT were conceptually or procedurally incorrect and ambiguous. McGalliard and Otten (2025) showed that detailed responses can coexist with substantial mathematical errors. The present comparison indicates that error sources need to be distinguished between representation and mathematization at the beginning of a solution and later stages of mathematical execution.

An important implication of the error findings is that a correct answer does not always coincide with a reliable process. Noster et al. (2024) found that preservice teachers can accept incorrect ChatGPT responses. In adaptations of

Habermas's (1998) framework to mathematics education, epistemic rationality concerns the validity of knowledge and claims, teleological rationality concerns the appropriateness of the strategy for the goal, and communicative rationality concerns whether the solution is expressed clearly and traceably (Boero & Planas, 2014). Urhan et al. (2026) identified problems of epistemic rationality in mathematical reasoning with ChatGPT. Zhuang (2026) recommends questioning the method, mathematical grounds, and justifications together when evaluating AI solutions. In the present study, solution reliability was determined through the mathematical validity of each stage in the solution chain and the coherence among those stages; the final result was treated as one component of that structure.

Differences among models indicate that generative AI outputs should be interpreted in relation to the system used, task type, prompt, and mathematical content. Oh et al. (2024) showed that the performance of ChatGPT, Claude, and Gemini varies by task type. Schorcht et al. (2026) found differences among models in content, process-related, and pedagogical-contextual quality. Ergene and Ergene (2025) showed that the performance of different AI systems is related to problem complexity. Korkmaz Guler et al. (2024) identified variation in achievement and error patterns across different ChatGPT versions. The model differences reported here should therefore be understood as time-bound output characteristics specific to the tasks and data-generation conditions examined.

The main educational implication is that solutions generated by AI can be used not as ready-made sources of "correct solutions" for students, but as digital solution products that require mathematical validation. This approach is consistent with studies that recommend critical evaluation of generative AI solutions and explicit questioning of their mathematical accuracy (Gurl et al., 2025; McGalliard & Otten, 2025; Noster et al., 2024). The analytical dimensions used in this study can be converted into a sequence of classroom validation questions: Are the problem and givens represented correctly? Is the selected strategy appropriate for the problem conditions and mathematical goal? Yoon et al. (2024) found that evaluating mathematical arguments generated by AI can activate students' evaluative behaviors. Zhou et al. (2025) showed that higher-performing students more frequently engaged in interpretation and evaluation during AI-supported problem solving. From this perspective, generative AI outputs can be incorporated into digital mathematics experiences designed to have students compare representation, strategy, execution, and justification.

Overall, student solutions tended to be brief and symbolic, whereas AI responses presented more explicit intermediate steps. Students' errors were frequently procedural, and some AI errors arose during visual interpretation or mathematization. Reliable evaluation must inspect representations, methods, grounds, and outcomes together, since correct answers can coexist with faulty processes.

5.1. Limitations and Recommendations

This study is limited to 22 students aged 12–13 attending a public school in Istanbul and to 10 exponent problems. The student dataset consists of 220 written solutions, whereas each generative AI case consists of 10 first responses. The tables were therefore used to show the visibility of qualitative patterns within each case, not to make statistical comparisons of proportions across groups. The codes "reasoning," "strategy," and "justification" do not directly represent the full cognitive processes of the students or the internal operations of the AI systems.

The generative AI data were produced with the versions of ChatGPT, Claude, Gemini, and Grok available at the time of the study. The same standard prompt was given to all four tools. Because the prompt explicitly requested a step-by-step explanation and justification of calculations, it may have increased the written visibility of intermediate steps and justifications in the AI outputs (Schorcht et al., 2024, 2026). Students were not given an equally structured step-by-step prompt. Differences between students and AI in explanation length or code frequency were therefore not interpreted as evidence of cognitive superiority. Alternative solution paths within the same response in Gemini S7 and Grok S8 were retained in the qualitative analysis, while the same code was counted once in task-level counts.

Future research could include larger student groups from different grade levels and school contexts and extend the problem set to domains such as algebra, geometry, ratio and proportion, probability, and mathematical modeling. In addition to multiple-choice tasks, open-ended problems and tasks requiring multiple representations could support more detailed examination of cases in which a correct answer co-occurs with errors in the solution process.

Experimental conditions using more equivalent explanation prompts for humans and AI could help determine more clearly whether differences in written visibility arise from solver type or from prompt structure.

In studies of generative AI, examining the same problems across different model versions, prompting techniques, and repeated independent generations is important for determining the stability of solution patterns. Comparing prompts that request a short answer, a step-by-step solution, or explicit justification under separate conditions could reveal the extent to which the visibility of process codes is shaped by prompt design (Schorcht et al., 2024, 2026). Given the mathematization errors identified in tasks involving visual and part–whole relationships, multimodal problems should also be examined with larger sets of examples; testing the same task under text-only, image-only, and combined presentation conditions could help disentangle the sources of representation-related difficulties.

For mathematics education practice, activities can be designed in which student and AI solutions are compared in terms of problem representation, strategy, mathematical execution, justification, checking, and error patterns. Such critical-evaluation activities are consistent with recent studies that recommend incorporating generative AI solutions into students' mathematical thinking and validation processes (Gurl et al., 2025; McGalliard & Otten, 2025; Zhuang, 2026). Experimental or design-based studies of these digital solution-validation activities could help explain how students develop not only the habit of obtaining solutions from AI but also the ability to question AI-generated mathematical information, identify errors, and construct justifications.

Statements and Declarations

Funding: No funds, grants, or other support was received.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

Ethics Approval: The study received ethics approval from the Scientific Research and Publication Ethics Committee of the authors' institution before data collection. Identifying committee details and the approval number are withheld for blinded review and will be provided in the unblinded version.

Consent to Participate: Written informed consent was obtained from the participants' parents or legal guardians, and age-appropriate assent was obtained from the participating students before data collection.

Consent for Publication: Not applicable. No identifiable participant information is presented in the manuscript.

Data Availability: The data supporting the findings of this study are available from the corresponding author upon reasonable request. Because the dataset includes minors' written work, data sharing is subject to the conditions of the ethics approval and the consent obtained from participants and their parents or legal guardians.

Author Contributions: Author 1: investigation, formal analysis, and writing – review and editing. Author 2: investigation, formal analysis, and writing – review and editing. Author 3: conceptualization, methodology, investigation, supervision, formal analysis, writing – original draft, and writing – review and editing. All authors read and approved the final manuscript.

References

Arcavi, A. (2003). The role of visual representations in the learning of mathematics. *Educational Studies in Mathematics, 52*, 215–241.

Boero, P. (2006). Habermas' theory of rationality as a comprehensive frame for conjecturing and proving in school. In J. Novotná, H. Moraová, M. Krátká, & N. Stehlíková (Eds.), *Proceedings of the 30th conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 185–192). PME.

Boero, P., & Planas, N. (2014). Habermas' construct of rational behavior in mathematics education: New advances and research questions. In P. Liljedahl, C. Nicol, S. Oesterle, & D. Allan (Eds.), *Proceedings of the Joint Meeting of PME 38 and PME-NA 36* (Vol. 1, pp. 205–235). PME.

Borasi, R. (1994). Capitalizing on errors as "springboards for inquiry": A teaching experiment. *Journal for Research in Mathematics Education, 25*(2), 166–208. https://doi.org/10.5951/jresematheduc.25.2.0166

Brasken, M., Berts, K.-E., Wallin, S., & Frilund, S. (2025). Students' use of ChatGPT in an algebra class: A case study of prompts and attitudes. *LUMAT: International Journal on Math, Science and Technology Education, 13*(1), 16. https://doi.org/10.31129/LUMAT.13.1.2877

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

Bulut, N., & Borromeo Ferri, R. (2026). AI-powered learning: Exploring secondary school students' use of ChatGPT for solving mathematical modelling problems. *ZDM–Mathematics Education*. https://doi.org/10.1007/s11858-026-01803-6

Busuttil, L., & Calleja, J. (2025). Teachers' beliefs and practices about the potential of ChatGPT in teaching mathematics in secondary schools. Digital Experiences in Mathematics Education, 11, 140–166. https://doi.org/10.1007/s40751-024-00168-3

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*. https://doi.org/10.1007/s11858-026-01786-4

Çelik, H., & Ay, Z. (2025). Evaluation of the mathematics questions on the Secondary Education Transition Examination (LGS) according to process standards. *Education and Science, 50*(Supplement 1), 147–171. https://doi.org/10.15390/EB.2025.14103

Daher, W., & Gierdien, F. (2024). Use of language by generative AI tools in mathematical problem solving: The case of ChatGPT. African Journal of Research in Mathematics, Science and Technology Education, 28(2), 222–235. https://doi.org/10.1080/18117295.2024.2384676

Duran, A. A. (2024). Liselere Geçiş Sistemi matematik sorularının bilişsel istem düzeyleri bağlamında değerlendirilmesi [Evaluation of High School Entrance System (LGS) mathematics questions in terms of cognitive demand levels]. *Necmettin Erbakan Üniversitesi Ereğli Eğitim Fakültesi Dergisi, 6*(1), 65–85. https://doi.org/10.51119/ereegf.2024.72

Duval, R. (2006). A cognitive analysis of problems of comprehension in a learning of mathematics. *Educational Studies in Mathematics, 61*, 103–131.

Ergene, O., & Ergene, B. C. (2025). AI ChatBots' solutions to mathematical problems in interactive e-textbooks: Affordances and constraints from the eyes of students and teachers. Education and Information Technologies, 30, 509–545. https://doi.org/10.1007/s10639-024-13121-z

Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., & Berner, J. (2023). Mathematical capabilities of ChatGPT. *Advances in Neural Information Processing Systems, 36*, 27699–27744.

Getenet, S. (2024). Pre-service teachers and ChatGPT in multistrategy problem-solving: Implications for mathematics teaching in primary schools. *International Electronic Journal of Mathematics Education, 19*(1), em0766. https://doi.org/10.29333/iejme/14141

Goizueta, M., Mariotti, M. A., & Planas, N. (2014). Validating in the mathematics classroom. In S. Oesterle, P. Liljedahl, C. Nicol, & D. Allan (Eds.), *Proceedings of the Joint Meeting of PME 38 and PME-NA 36* (Vol. 3, pp. 169–176). PME.

Gurl, T. J., Markinson, M. P., & Artzt, A. F. (2025). Using ChatGPT as a lesson planning assistant with preservice secondary mathematics teachers. Digital Experiences in Mathematics Education, 11, 114–139. https://doi.org/10.1007/s40751-024-00162-9

Habermas, J. (1998). *On the pragmatics of communication* (M. Cooke, Ed.). MIT Press.

Hiebert, J., & Wearne, D. (1986). Procedures over concepts: The acquisition of decimal number knowledge. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 199–223). Lawrence Erlbaum Associates.

Jia, J., Wang, T., Zhang, Y., & Wang, G. (2024). The comparison of general tips for mathematical problem solving generated by generative AI with those generated by human teachers. Asia Pacific Journal of Education, 44(1), 8–28. https://doi.org/10.1080/02188791.2023.2286920

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kieran, C. (2004). Algebraic thinking in the early grades: What is it? *The Mathematics Educator, 8*(1), 139–151.

Korkmaz Guler, N., Dertli, Z. G., Boran, E., & Yildiz, B. (2024). An artificial intelligence application in mathematics education: Evaluating ChatGPT's academic achievement in a mathematics exam. *Pedagogical Research, 9*(2), em0188. https://doi.org/10.29333/pr/14145

Kortenkamp, U., & Larkin, K. (2026). LLMs as useful training tools for mathematics teachers: Opportunities or hallucinations? *ZDM–Mathematics Education*. https://doi.org/10.1007/s11858-026-01813-4

Lamon, S. J. (2007). Rational numbers and proportional reasoning: Toward a theoretical framework. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 629–667). Information Age Publishing.

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

McGalliard, W., & Otten, S. (2025). AI responses to challenging problems and educator responses to AI availability. Digital Experiences in Mathematics Education, 11, 319–332. https://doi.org/10.1007/s40751-024-00167-4

Millî Eğitim Bakanlığı. (2023, 7 Eylül). *Liselere Geçiş Sistemi [High School Transition System]*. Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü. https://odsgm.meb.gov.tr/www/liselere-gecis-sistemi/icerik/1012

National Research Council. (2001). *Adding it up: Helping children learn mathematics* (J. Kilpatrick, J. Swafford, & B. Findell, Eds.). National Academies Press. https://doi.org/10.17226/9822

Noster, N., Gerber, S., & Siller, H.-S. (2024). Pre-service teachers' approaches in solving mathematics tasks with ChatGPT. *Digital Experiences in Mathematics Education, 10*, 543–567. https://doi.org/10.1007/s40751-024-00155-8

Oh, S., Yoon, J., Chung, Y., Cho, Y., Shim, H., & Kwon, O. N. (2024). Analysis of generative AI's mathematical problem-solving performance: Focusing on ChatGPT 4, Claude 3 Opus, and Gemini Advanced. *The Mathematical Education, 63*(3), 549–571. https://doi.org/10.7468/mathedu.2024.63.3.549

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

Pepin, B., Buchholtz, N., & Salinas-Hernández, U. (2025). A scoping survey of ChatGPT in mathematics education. *Digital Experiences in Mathematics Education, 11*, 9–41. https://doi.org/10.1007/s40751-025-00172-1

Pólya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton University Press.

Radatz, H. (1979). Error analysis in mathematics education. *Journal for Research in Mathematics Education, 10*(3), 163–172. https://doi.org/10.5951/jresematheduc.10.3.0163

Sapkota, B., & Bondurant, L. (2024). Assessing concepts, procedures, and cognitive demand of ChatGPT-generated mathematical tasks. International Journal of Technology in Education, 7(2), 218–238. https://doi.org/10.46328/ijte.677

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Academic Press.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334–370). Macmillan.

Schorcht, S., Buchholtz, N., & Baumanns, L. (2024). Prompt the problem—Investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. *Frontiers in Education, 9*, 1386075. https://doi.org/10.3389/feduc.2024.1386075

Schorcht, S., Müller, F. A., & Buchholtz, N. (2026). No one-size-fits-all: A study of prompt techniques and large language models to enhance AI's mathematics educational quality. *ZDM–Mathematics Education*. https://doi.org/10.1007/s11858-026-01784-6

Skemp, R. R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching, 77*, 20–26.

Spreitzer, C., Straser, O., Zehetmeier, S., & Maaß, K. (2024). Mathematical modelling abilities of artificial intelligence tools: The case of ChatGPT. *Education Sciences, 14*(7), 698. https://doi.org/10.3390/educsci14070698

Stylianides, A. J. (2007). Proof and proving in school mathematics. *Journal for Research in Mathematics Education, 38*(3), 289–321.

Taani, O., & Alabidi, S. (2025). ChatGPT in education: Benefits and challenges of ChatGPT for mathematics and science teaching practices. *International Journal of Mathematical Education in Science and Technology, 56*(9), 1748–1777. https://doi.org/10.1080/0020739X.2024.2357341

Toulmin, S. E. (2003). *The uses of argument* (Updated ed.). Cambridge University Press. (Original work published 1958)

Turmuzi, M., Azmi, S., & Kertiyani, N. M. I. (2026). ChatGPT in school mathematics education: A systematic review of opportunities, challenges, and pedagogical implications. *Teaching and Teacher Education, 170*, 105286. https://doi.org/10.1016/j.tate.2025.105286

UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO.

Urhan, S., & Bülbül, A. (2023). Habermas' construct of rationality in the analysis of the mathematical problem-solving process. *Educational Studies in Mathematics, 112*(1), 175–197. https://doi.org/10.1007/s10649-022-10188-8

Urhan, S., Gençaslan, O., & Dost, Ş. (2026). Is ChatGPT a rational assistant for university students during mathematical reasoning? *Science & Education*. https://doi.org/10.1007/s11191-026-00725-9

Usiskin, Z. (1988). Conceptions of school algebra and uses of variables. In A. F. Coxford & A. P. Shulte (Eds.), *The ideas of algebra, K–12* (pp. 8–19). National Council of Teachers of Mathematics.

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

Wei, X. (2024). Evaluating ChatGPT-4 and ChatGPT-4o: Performance insights from NAEP mathematics problem solving. *Frontiers in Education, 9*, 1452570. https://doi.org/10.3389/feduc.2024.1452570

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE.

Yoon, H., Hwang, J., Lee, K., Roh, K. H., & Kwon, O. N. (2024). Students' use of generative artificial intelligence for proving mathematical statements. *ZDM–Mathematics Education, 56*, 1531–1551. https://doi.org/10.1007/s11858-024-01629-0

Zhou, R., He, X., Fan, Q., Li, Y., Li, Y., Xiao, X., & Fang, J. (2025). Exploring ChatGPT-facilitated scaffolding in undergraduates' mathematical problem solving. Journal of Computer Assisted Learning, 41(4), e70077. https://doi.org/10.1111/jcal.70077

Zhuang, Y. (2026). Empowering students to critically validate AI-generated mathematical solutions through the rational questioning approach. *Educational Studies in Mathematics*. https://doi.org/10.1007/s10649-026-10528-y