MATHEMATICS TEACHING RESEARCH JOURNAL
EARLY SPRING 2026
Vol 18 no 1

Comparative Analysis of Mathematical Arguments in Algebraic Proofs: Generative AI vs. University Students Approaches

Hendra Kartika[1], Lauren Jeneva Clark[2]

[1]Universitas Singaperbangsa Karawang, Karawang, Indonesia,

[2] University of Tennessee, Knoxville, USA

hendra.kartika@staff.unsika.ac.id, dr.jenevaclark@utk.edu

Abstract: The integration of artificial intelligence (AI) into mathematics education has attracted growing attention, particularly with the emergence of generative AI models capable of constructing mathematical arguments and proofs. This study compares the argumentation structures used by generative AI and undergraduate mathematics students when proving statements in number theory. Using a qualitative comparative approach, we analyzed responses generated by ChatGPT 3.5, YouChat, and 94 fifth-year undergraduate students at a state university in Indonesia. Participants were asked to justify a mathematical statement, and their responses were examined using the Claim–Evidence–Reasoning (CER) framework and argument mapping techniques. The results show that AI and most students reached similar conclusions. However, 9.6% of students produced different and more detailed responses, suggesting a higher level of sophistication in human mathematical reasoning. While AI systematically categorized cases based on integer values, some students extended their reasoning to non-integer domains, demonstrating greater interpretative flexibility. Moreover, AI predominantly relied on naïve empiricism, whereas students displayed varying levels of deductive reasoning. These findings underscore the need for further research to enhance AI's reasoning capabilities and to better integrate AI into pedagogical strategies that support deeper mathematical understanding beyond the current capacities of AI tools.

Keywords: Algebra, Argument Quality, Mathematical Argumentation

INTRODUCTION

The comparison between human and artificial intelligence reveals fundamental and consequential distinctions (Cope et al., 2020; Korteling et al., 2021a). Empirical evidence consistently demonstrates that humans outperform AI systems across a broad spectrum of cognitive and social tasks, particularly in novel and unpredictable contexts (Korteling et al., 2021b). Human superiority

is especially evident in social and psychosocial domains, where AI continues to encounter significant limitations. Specifically, interpreting human language and symbolism requires a nuanced and extensive contextual framework—an area in which current AI technologies remain markedly deficient (Korteling et al., 2021a). Nevertheless, AI systems exhibit a clear advantage in executing routine cognitive operations with high efficiency, particularly in domains where human performance is constrained (Markauskaite et al., 2022). These disparities underscore the necessity of a rigorous comparative analysis between AI-generated and student-generated mathematical arguments to clarify core differences in problem-solving strategies, logical structures, and conceptual depth. Whereas AI is driven by data-centric algorithms and pattern recognition, human cognition is characterized by the integration of intuition, creativity, and experiential knowledge.

Artificial intelligence (AI) has increasingly permeated various sectors of society, including the field of mathematics and its instructional methodologies (Engelbrecht & Borba, 2023). Mathematics is frequently identified as a particularly suitable domain for evaluating AI's reasoning capabilities, given its structured logic and definable problem sets (Glazer et al., 2024). Contemporary AI systems possess the capacity to analyze data, extract meaningful patterns, and generate informed inferences or predictions based on learned information (Celik et al., 2022). Recent advancements in generative AI have further demonstrated its proficiency in solving mathematical problems across a range of topics, including algebra, calculus, and geometry, thereby prompting intensified scholarly interest in its applicability to diverse mathematical disciplines (Oh, 2024). Moreover, empirical studies indicate that AI's functionality extends beyond mere computational efficiency; it also encompasses advanced competencies such as pattern recognition and logical reasoning, reinforcing its potential to engage with and solve complex mathematical problems (Lu et al., 2022; Xu et al., 2024).

In mathematics education, a growing body of research has explored the integration of artificial intelligence (AI) to support and enhance student learning. Graham (2023) identifies a range of innovative applications, including AI-generated visualizations and interactive simulations designed to facilitate the comprehension of complex mathematical concepts. AI also contributes to increased student engagement through the use of gamified learning experiences and personalized mathematical tasks.

Furthermore, AI systems offer targeted remediation for learners who encounter difficulties, thereby supporting differentiated instruction. Recent studies have focused specifically on the role of generative AI in mathematical reasoning. Park and Manley (2024) examine the use of ChatGPT in assisting students with the construction of mathematical arguments that qualify as formal proofs. Similarly, Dilling and Herrmann (2024) explore the utility of generative AI in supporting pre-service primary mathematics teachers in developing geometric proofs. Yoon et al. (2024) investigate how undergraduate students engage with generative AI when formulating mathematical proofs. Despite these promising developments, the integration of generative AI into

mathematics education presents significant challenges (Dwivedi et al., 2023; Wardat et al., 2023), particularly concerning the accuracy and reliability of AI-generated content. Moreover, the extent to which generative AI can meaningfully support the development, evaluation, and comparison of mathematical proofs alongside human-generated arguments remains unresolved. As such, a critical next step is to systematically assess the quality of AI-generated mathematical argumentation in relation to that produced by human learners. This evaluation is essential for advancing a deeper understanding of both the capabilities and limitations of AI within the context of mathematical reasoning and instruction.

Literature Review

Mathematical argumentation refers to the structured process through which reasoning is developed, articulated, and substantiated to analyze, explain, and solve mathematical problems (Kartika et al., 2024; Stylianides et al., 2013). It serves as a foundational precursor to formal proof, a core pillar of mathematical thinking and practice (Marks Krpan & Sahmbi, 2024). A mathematical argument consists of a coherent set of statements supported by evidence—such as data, warrants, and justifications—intended to establish the validity or refute the falsity of a specific claim or conclusion (Cardetti & LeMay, 2018; Zambak & Magiera, 2020). Within this framework, data supports the claim, and the logical connection between them must be explicitly validated through established mathematical principles, including axioms, definitions, theorems, or rules (Urhan & Zengin, 2023). Thus, a mathematical argument emerges as a reasoned sequence of propositions derived through rigorous argumentation (Kartika et al., 2023).

While extensive research has investigated students' abilities to construct mathematical arguments, comparatively less attention has been paid to how these arguments align or differ from those generated by artificial intelligence. This research gap is particularly salient given the growing role of AI tools in mathematical problem-solving, necessitating a deeper examination of how mathematical reasoning manifests in both human and machine-generated contexts.

The integration of generative AI technologies—such as ChatGPT and adaptive learning platforms—has introduced significant shifts within mathematics education. These technologies not only influence instructional practices but also provoke critical theoretical discussions about the nature of mathematical knowledge and the evolving roles of agents within human-technology collectives (Engelbrecht & Borba, 2023). Although generative AI models exhibit impressive capabilities in generating coherent, human-like text (Rane, 2023), they remain susceptible to producing contextually irrelevant, inaccurate, or misleading outputs (Wardat et al., 2023; Biton & Segal, 2025). Their operation as opaque "black box" systems further complicate this issue, as they can yield plausible yet erroneous responses that may even mislead subject-matter experts (Mishra et al., 2023; Fütterer et al., 2023). Instances of fabricated content highlight the risk of undermining

user confidence and distorting perceptions of one's own mathematical competence (Emsley, 2023).

It is therefore imperative to view generative AI as a supplementary educational tool rather than a replacement for human educators (Tenhundfeld, 2023). AI-generated content must be subject to critical scrutiny by individuals with the requisite disciplinary expertise to assess its validity (Sabzalieva & Valentini, 2023). The ability to detect and correct mathematical errors in AI outputs is essential for their responsible use, particularly in educational contexts where such tools may be employed to construct examples or instructional materials (Noster et al., 2024; Schorcht et al., 2024). To harness the full potential of generative AI in mathematics education, it is essential to integrate its capabilities with human critical thinking and disciplinary knowledge.

Beyond verifying accuracy, a deeper and more pressing concern lies in evaluating the nature and quality of AI-generated mathematical reasoning. In particular, comparative analyses between AI- and human-generated arguments are essential to understanding the degree to which AI systems can replicate—or diverge from—the logical coherence, conceptual depth, and justification that characterize rigorous mathematical argumentation.

METHOD

This study adopted a qualitative research design, utilizing document analysis as the primary methodological approach. Document analysis entails the systematic examination and interpretation of written texts—both handwritten and digitally generated—to extract meaningful patterns and insights (Bowen, 2009). In this context, the analysis focused on identifying similarities and differences in the argumentative strategies employed by generative AI and student participants. The objective was to develop a nuanced understanding of the distinctive reasoning methods and epistemic contributions each source brings to the domain of mathematical argumentation.

Participants

The generative AI models utilized in this study were ChatGPT 3.5 and YouChat, both of which are publicly accessible and widely recognized for their capabilities in natural language processing and mathematical problem-solving. The human participants comprised 94 fifth-year undergraduate mathematics education students, aged 19 to 21, from Singaperbangsa Karawang University in West Java, Indonesia. All participants had previously completed a course in Number Theory during their first semester, thereby ensuring a foundational competence relevant to the mathematical statement under investigation. Participation was voluntary, and students were informed that all data would be anonymized and used solely for research purposes, with no implications for their academic standing.

The selection of both advanced undergraduate students and specific generative AI models was deliberate. Fifth-year students were chosen due to their extensive mathematical training, including exposure to proof-based reasoning and formal argumentation, rendering them well-positioned to critically assess and construct mathematical statements. Their academic maturity enabled a meaningful exploration of human mathematical reasoning at a level characterized by abstract and deductive thought.

ChatGPT 3.5 and YouChat were selected as AI comparators based on their accessibility and their documented ability to generate coherent mathematical explanations and structured arguments. Both models, trained on large-scale datasets, are capable of producing algorithmically organized responses that simulate human-like reasoning. By comparing the outputs of these AI systems with those of advanced mathematics students, the study aimed to evaluate the reasoning capabilities of generative AI in relation to human cognition. This comparative analysis also illuminated key distinctions in argument construction, treatment of implicit assumptions, and the use of counterexamples—critical components of rigorous mathematical argumentation.

Data Collection Process

In this study, participants were given 30–40 minutes to justify a mathematical statement in number theory, adapted from Lee (2016). They were required to construct their arguments independently, without assistance from online sources. To address any potential misunderstandings related to terminology or definitions, participants were permitted to seek clarification from the researcher.

The task was intentionally designed to align with the participants' prior coursework, integrating both fundamental and advanced elements of number theory to assess a broad spectrum of mathematical reasoning competencies. Central to the task was the construction of a deductive proof, requiring students to engage in formal logical reasoning and structured argumentation. While some participants responded with generalized explanations—often relying on mathematical language rather than specific examples (e.g., “The sum is even because odd plus odd is even”)—a rigorous refutation of the given statement required the identification of a specific counterexample and a logical demonstration that the implication was false (Lee, 2016).

To ensure the task's appropriateness and efficacy, a pilot study was conducted involving 31 fifth-year mathematics education students from the same institution. The final version of the task was selected based on criteria including clarity, the potential to generate meaningful engagement, and the capacity to elicit nontrivial and analytically rich responses. This validation process confirmed the task’s effectiveness in prompting essential components of mathematical reasoning, such as the construction of formal justifications, identification of logical inconsistencies, and formulation of counterexamples.

Handwritten student responses were collected and subsequently categorized by thematic content. The same task prompt was also submitted to ChatGPT 3.5 and YouChat, with their outputs systematically recorded for comparative analysis. As illustrated in Figure 1, the task did not explicitly define the variable x as belonging to a specific numerical set. This omission proved to be a critical factor in evaluating both student and AI-generated arguments, as it directly impacted the interpretation and validity of the logical structure underlying each response.

Benar atau salah pernyataan ini: "Jika $3x$ dikalikan dengan $(1 + x^2)$, maka hasilnya adalah bilangan genap."
(Translation) Is this statement true or false? "If $3x$ is multiplied by $(1 + x^2)$, then the product is an even number."

Figure 1: Argumentative Tasks Frame

Before analyzing the data, the responses obtained by the students were compared to determine similarities and differences. Students who had given similar answers were grouped and selected one data from each group for in-depth analysis. In addition, we compared the responses obtained from ChatGPT and YouChat to assess consistency between the two sources. For instance, ChatGPT and YouChat evaluate the statement under two conditions, namely x is even and x is odd. Then ChatGPT and YouChat conclude that this statement is true. This indicates that the responses from both sources are consistent and, although mathematically flawed, this establishes the credibility of this data for further analysis.

Data Analysis Process

All arguments were qualitatively analyzed through close reading based on the dimensions of structure, content, and recipient orientation (Meyer & Schnell, 2020). To develop coding classifications for analyzing participants' argument structures in Deductive-Proof Construction, this research utilized the CER (Claim, Evidence, Reasoning) model (McNeill & Krajcik, 2008), which is a three-step process: making statements, providing supporting evidence, and explaining the logic behind their statement. Using the CER model, the researchers focused on identifying the conclusions or claims made by each participant. In particular, a claim refers to the participant's response to the question, *"Is this statement true or false?"*

In addition to claims, the researchers examined the evidence and reasoning provided by the participants (Kartika et al., 2023). Evidence refers to the data or information gathered to back up the claim, while reasoning involves the mathematical rule or principle that connects the evidence to the claim. Meyer and Schnell (2020) suggested several guiding questions to identify the structure of an argument. For example: Does the claim answer the question? Is there sufficient

evidence to support the claim? Is the reasoning adequate to explain how the evidence supports the claim? These types of questions assist in evaluating the strength and validity of participants' arguments.

Furthermore, the researchers constructed argument maps by synthesizing all arguments formulated by the generative AI models and student participants. Each map illustrated the relationships among the claims, evidence, and reasoning presented in the respective solution narratives. Following the construction of these maps, the researchers analyzed the data to identify recurring patterns and draw conclusions regarding the quality of the arguments. Argument quality was evaluated using the framework developed by McNeill and Krajcik (2008), as adapted by Kartika et al. (2023) and presented in Table 1.

Component of Argument	Quality	Description	Code
Claim	Low	Not making any claims, or making inaccurate or false claims.	C_L
	Moderate	Making claims that are accurate but incomplete	C_M
	High	Making a claim that is both accurate and complete.	C_H
Evidence	Low	Not providing evidence, or only providing inappropriate evidence (evidence that does not support the claim).	E_L
	Moderate	Providing precise but insufficient evidence to support a claim. This may include some incongruous pieces of evidence.	E_M
	High	Providing precise and sufficient evidence to support a claim.	E_H
Reasoning	Low	Failure to provide reasoning, or only providing reasoning that does not connect evidence to claims.	R_L
	Moderate	Providing reasoning that connects claims and evidence, reiterating evidence and/or incorporating some, but insufficient scientific principles.	R_M
	High	Providing reasoning that connects evidence with claims, encompassing appropriate and adequate scientific principles.	R_H

Table 1: Guidelines for Assessing Argument Quality

We also identified the levels of proof formulated by Generative AI and the students. According to Balacheff (1988), there are four levels of proof.

1. Naïve Empiricism: This level involves testing a mathematical statement by verifying a few specific cases.

2. Crucial Experiment: At this level, a special or extreme case is used to prove the truth of a statement. Students at this stage recognize the limitations of generalizing from a few examples and begin to address the problem of insufficient verification.
3. Generic Example: This level involves demonstrating the truth of a mathematical statement through a representative operation, which serves as a basis for generalization.
4. Thought Experiment: The highest level, this involves investigating the properties of operations to justify the truth of mathematical statements. A thought experiment eliminates the reliance on specific examples by building on the concept of a generic example.

Stylianides (2019) categorized Naïve Empiricism and Crucial Experiments as weak forms of argumentation, whereas Generic Examples and Thought Experiments represent strong and more rigorous approaches to proof.

The results of the analysis are presented clearly and concisely, supported by tables, graphics, generative AI outputs, and student artifacts to highlight the main findings. The coding process, which involved iterative discussions and collaboration, ensured the reliability and validity of the findings. Specifically, the first author conducted individual coding, followed by discussions to review and refine the proposed categorizations. A final agreement on the categorization was reached through this collaborative approach.

RESULTS AND DISCUSSIONS

In this study, participants were tasked with evaluating a mathematical statement in number theory: *"If 3x is multiplied by* $(1+x^2)$*, then the product is an even number."* The findings related to the question *"Is this statement true or false?"* are summarized in Table 2. This table presents a comparative overview of responses from two participant groups: generative AI models and undergraduate students. Each response was classified as either "True" or "False," with the corresponding frequency and percentage reported for each category, thereby facilitating a clear comparison of the interpretative tendencies across human and AI-generated reasoning.

Responses	Participants	
	Generative AI (n=2)	Students (n=94)
True	2 (100%)	85 (90.4%)
False	0 (0.0%)	9 (9.6%)

Table 2: Frequency of Responses between Generative AI and Students

Table 2 reveals that the generative AI systems uniformly classified the statement as "True," with both responses (100%) indicating full agreement. Likewise, the majority of student participants (n = 94) concurred with this assessment, with 85 students (90.4%) selecting "True." However, a

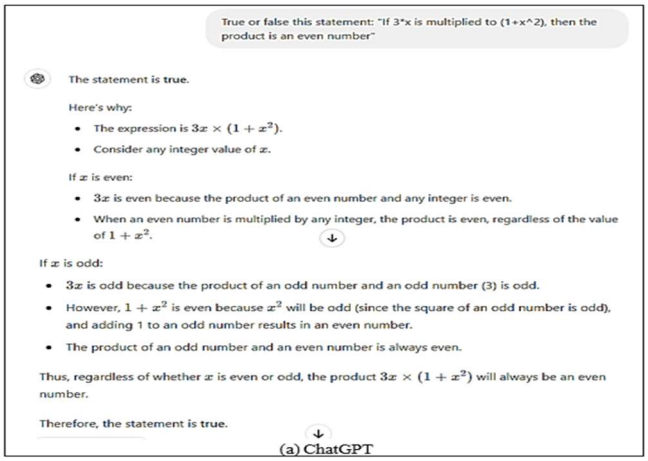
minority of students—9 individuals (9.6%)—responded with "False," indicating a degree of divergence in reasoning between subsets of the human participants and the AI models.

The comparative analysis of responses from the two groups reveals noteworthy patterns. The high proportion of "True" responses from both generative AI (100%) and student participants (90.4%) suggests a significant convergence between AI-generated reasoning and prevailing human interpretations. This alignment may be interpreted as an indicator of the apparent validity of the mathematical statement as perceived by the majority of respondents. However, the presence of a 9.6% deviation among students warrants closer examination, as it highlights the potential influence of alternative reasoning pathways in human cognition.

This divergence may stem from various underlying factors. Differences in conceptual versus procedural understanding likely play a role in how individuals evaluate mathematical statements. Students with greater exposure to formal proof techniques or training in identifying counterexamples may be more predisposed to scrutinize assertions and challenge superficially plausible conclusions. Moreover, cultural and pedagogical contexts—including norms that emphasize critical thinking, questioning of assumptions, and diverse problem-solving strategies—may further contribute to these differing interpretations. These findings suggest that while generative AI can replicate dominant reasoning patterns, it remains limited in capturing the nuanced, contextually informed, and often experience-driven dimensions of human mathematical thought.

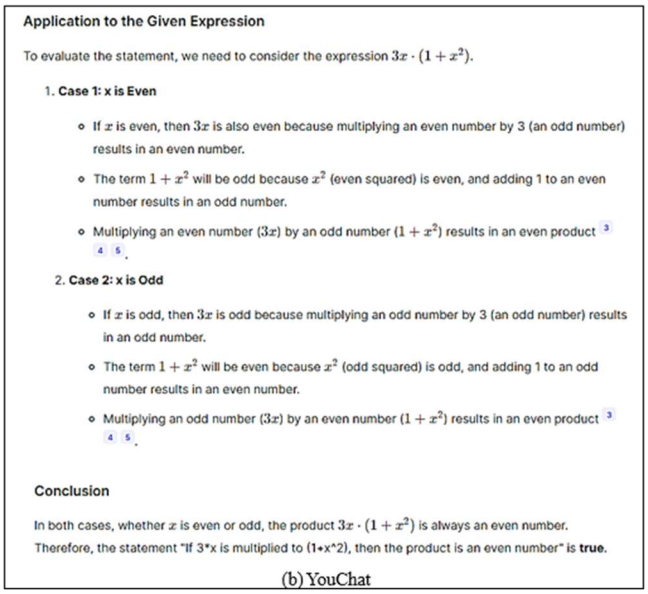
Argumentative Structure Generated by Artificial Intelligence

Figure 2 and 3 presents the response generated by the artificial intelligence model in relation to the given mathematical task. This figure illustrates the AI's approach to constructing a mathematical argument, highlighting its reasoning process, logical structure, and the methods employed to assess the validity of the statement.



(a) ChatGPT

Figure 2: ChatGPT Responses to the Mathematical Task



Application to the Given Expression

To evaluate the statement, we need to consider the expression $3x \cdot (1 + x^2)$.

1. **Case 1: x is Even**
   - If x is even, then $3x$ is also even because multiplying an even number by 3 (an odd number) results in an even number.
   - The term $1 + x^2$ will be odd because x^2 (even squared) is even, and adding 1 to an even number results in an odd number.
   - Multiplying an even number ($3x$) by an odd number ($1 + x^2$) results in an even product [3] [4] [5].
2. **Case 2: x is Odd**
   - If x is odd, then $3x$ is odd because multiplying an odd number by 3 (an odd number) results in an odd number.
   - The term $1 + x^2$ will be even because x^2 (odd squared) is odd, and adding 1 to an odd number results in an even number.
   - Multiplying an odd number ($3x$) by an even number ($1 + x^2$) results in an even product [3] [4] [5].

Conclusion

In both cases, whether x is even or odd, the product $3x \cdot (1 + x^2)$ is always an even number.
Therefore, the statement "If 3*x is multiplied to (1+x^2), then the product is an even number" is **true**.

(b) YouChat

Figure 3: YouChat Responses to the Mathematical Task

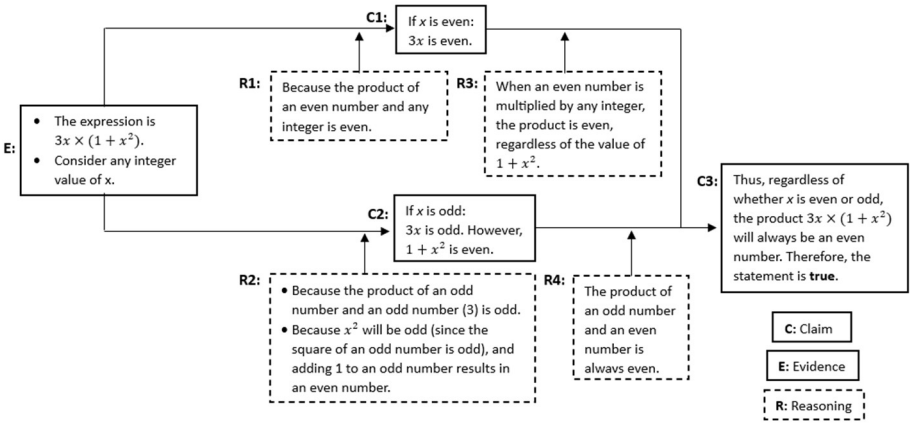


Figure 4: Argument Map Representing the Reasoning of the Generative AI

As illustrated in Figure 4, the generative AI tools—specifically ChatGPT and YouChat—began their analysis by restating the algebraic expression $3x \times (1+x^2)$ and proceeded under the assumption that the variable x is an integer. The AI then employed a case-based approach, identifying two exhaustive categories for x: namely, that x is either an even or an odd integer.

In the first case, where x is even, the AI reasoned that $3x$ must also be even, based on the mathematical principle that the product of an even number and any integer is always even. Consequently, multiplying this even result by any value—including the expression $1+x^2$—will yield an even product, in accordance with established properties of integer multiplication.

In the second case, where x is odd, the AI asserted that $3x$ is odd, as the product of two odd numbers remains odd. It further noted that x^2 would also be odd (given that the square of an odd number is odd), and therefore, $1+x^2$ must be even. Since the product of an odd number and an even number is always even, the entire expression $3x \times (1+x^2)$ would again yield an even result.

Based on these two scenarios, both of which result in an even product, the AI concluded that the expression produces an even number for all integer values of x. Accordingly, the AI affirmed the truth of the given mathematical statement, asserting its universal validity within the set of integers.

Argumentative Structure Demonstrated by a Student

The mathematical reasoning constructed by a student participant is presented in Figure 5. The figure illustrates the student's reasoning process, including the formulation of claims, use of mathematical principles, and the logical progression employed to justify or refute the validity of the statement.

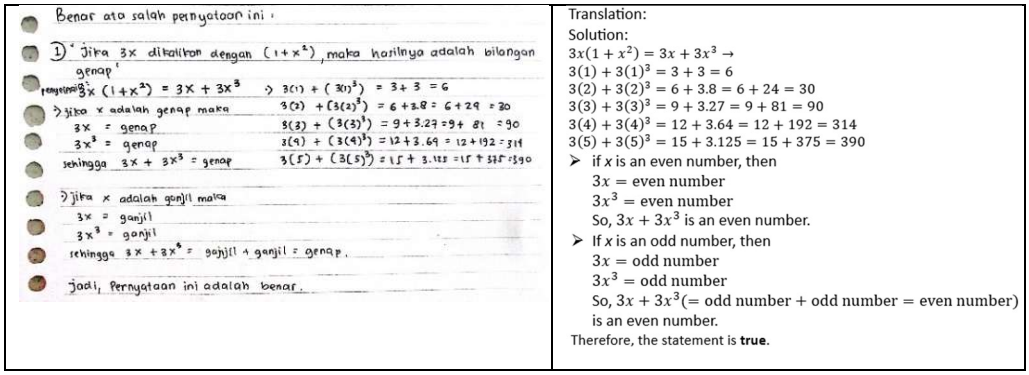


Figure 5: Responses of Students to the Mathematical Task

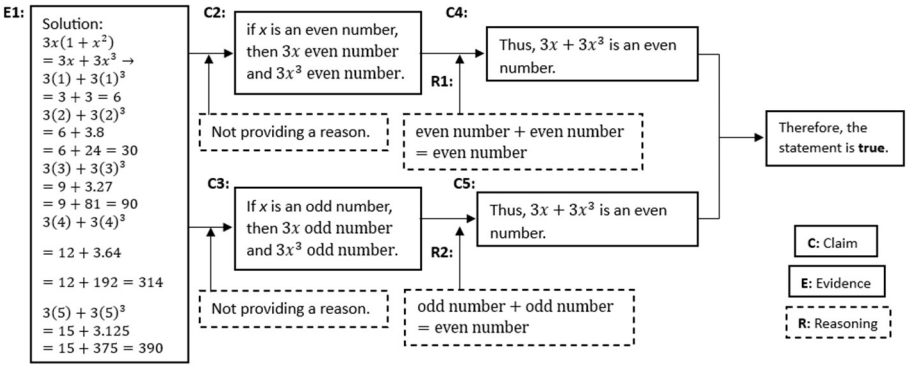


Figure 6: Structure of the Student's Argument

Figure 6 illustrates the structure of a student's argument when evaluating the given mathematical statement. The student initially approached the problem by rewriting the expression $3x \times (1+x^2)$ and expanding it to $3x+3x^2$. Following this, the student substituted several integer values into the expanded equation. For example, when x=1, the calculation $3x+3x^2=3(1)+3(1)^2=3+3=6$, and for x=2, the equation becomes $3(2)+3(2)^3=6+3.8=6+24=30$. Based on these computations, the student identified two key conditions:

1. If x is an even number, then both $3x$ and $3x^2$ are even, and therefore, $3x+3x^3$ is even.
2. If x is an odd number, then both $3x$ and $3x^2$ are odd, and since the sum of two odd numbers is even, $3x+3x^3$ is even.

   In conclusion, the student determined that the mathematical statement was true.

The Claim-Evidence-Reasoning (CER) framework was employed to analyze the structure, content, and recipient orientation of the mathematical arguments produced by both Generative AI and university students when evaluating number theory statements. The analysis revealed that both Generative AI and the majority of students reached the same conclusion, affirming the truth of the statement under consideration. However, a subset of students diverged, concluding that the statement was false, as demonstrated in Figure 7.

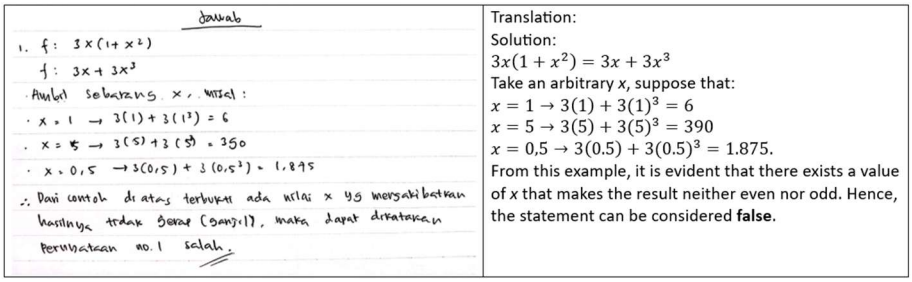


Figure 7: Responses to the Mathematical Task from an Alternative Student Group

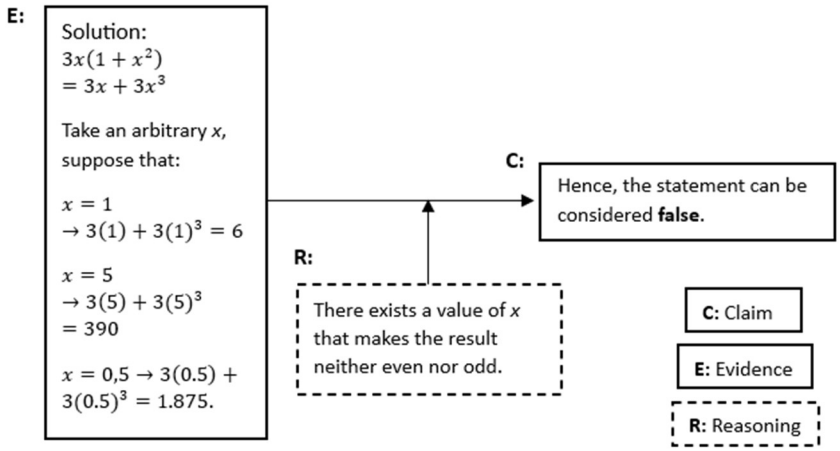


Figure 8: Structure of an Argument Involving a Counterexample

As shown in Figure 8, this alternative group of students employed counterexamples in their reasoning. Rather than presuming that x was restricted to integer values, they systematically explored non-integer inputs. For example, substituting $x = 0.5$ into the equation $3(0.5)+3(0.5)^3=1.875$ yielded a result that contradicted earlier conclusions. This divergence arose from the fact that the problem did not explicitly define the domain of x. While the Generative AI consistently assumed x to be an integer, the students extended their reasoning to include non-integer values, thereby revealing a fundamental conceptual difference in interpretative strategy. This discrepancy highlights the critical role of counterexamples in mathematical reasoning, as they challenge implicit assumptions and contribute to the refinement and validation of mathematical arguments.

Generative AI models, such as ChatGPT and YouChat, frequently assume that x is an integer without explicitly defining this constraint within the problem context—an approach that aligns with what may be termed naïve empiricism (Stylianides, 2019). These systems predominantly rely on pattern recognition rather than formal deductive reasoning, often extrapolating from prototypical cases without systematically accounting for the complete domain of possible values for x. This methodological shortcoming compromises the validity of their reasoning, as it neglects edge cases and alternative numerical domains, thereby exposing a fundamental limitation in AI-generated mathematical argumentation.

In contrast, several students demonstrated a more analytical approach—characterized by the use of thought experiments (Stylianides, 2019)—through the systematic interrogation of implicit assumptions and the rigorous testing of a range of values for x, thereby engaging in fundamental practices of mathematical reasoning. Unlike generative AI systems, which tend to draw conclusions based on unexamined assumptions, these students employed formal justifications grounded in the use of counterexamples and logical constraints—elements that the AI models failed to incorporate. This distinction aligns with the findings of Pielsticker et al. (2024), who reported that ChatGPT exhibits difficulty maintaining logical consistency and rigor in the construction of mathematical proofs. Similarly, Frieder et al. (2023) concluded that although ChatGPT is occasionally capable of generating insightful arguments, its overall proficiency in mathematical proof remains limited due to its reliance on heuristic-based reasoning rather than structured formal logic.

The comparison between AI-generated and student-generated mathematical reasoning reveals a significant pedagogical opportunity: the strategic utilization of AI not merely as a computational tool, but as an instructional catalyst to enhance students' engagement with mathematical argumentation. These findings underscore the imperative for educators to transition from a passive reception of AI outputs to a critical, inquiry-driven approach—one that actively interrogates the logical structure, assumptions, and limitations inherent in AI-generated reasoning. This shift can be operationalized through the deliberate implementation of targeted instructional strategies.

The findings also point to the necessity of explicitly addressing the epistemological limitations of AI-based tools within the mathematics classroom. It is essential that students are not led to accept AI-generated arguments uncritically or regard them as definitive sources of mathematical truth. Rather, educators should frame AI as a heuristic device—useful for exploration and conjecture, but ultimately subordinate to disciplined reasoning and formal justification.

CONCLUSION

The study found that Generative AI, exemplified by ChatGPT and YouChat, relies on algorithmic reasoning based on specific assumptions, such as treating x as an integer. This methodology, classified as naïve empiricism, simplifies the argumentation process but limits its adaptability to broader mathematical contexts. In contrast, students demonstrated greater flexibility by employing algebraic manipulation and thought experimentation, enabling them to consider possibilities beyond integer values. These findings underscore both the potential and limitations of Generative AI in mathematical reasoning. While AI tools can generate structured and accurate responses in well-defined scenarios, their dependence on predefined assumptions and occasional struggles with contextual nuances highlight the necessity of human oversight in mathematical problem-solving. The study reinforces the role of AI as a complementary tool in education, enhancing learning and fostering critical thinking rather than replacing human educators. By integrating Generative AI with the reasoning skills of educators and students, mathematics education can adopt a more dynamic and robust framework that encourages analytical thinking and deeper conceptual understanding.

While the present study offers meaningful insights into the comparative argumentation patterns of generative AI and undergraduate students, several limitations merit critical consideration. First, the sample of generative AI systems was limited to two models—ChatGPT and YouChat—which, although representative of current state-of-the-art tools, do not encompass the full range of available AI technologies. Consequently, the findings may not generalize to other AI systems with different training data, architectures, or reasoning mechanisms. Second, the student sample was drawn from a single population of undergraduate participants, whose prior exposure to mathematical proof techniques, familiarity with number theory, or educational background may not reflect broader or more diverse student cohorts. Such homogeneity in the student group introduces the potential for selection bias, limiting the external validity of the results.

Moreover, the study's focus on a single mathematical statement, while analytically productive, constrains the scope of inference regarding AI and human reasoning across other mathematical domains or problem types. Finally, although the Claim-Evidence-Reasoning (CER) framework facilitated a structured analysis of argumentation, its application may have introduced interpretive subjectivity, particularly in categorizing the quality and nature of students' reasoning. Future research should consider expanding the range of mathematical tasks, incorporating a more diverse

set of AI models and student populations, and employing complementary analytical frameworks to enhance generalizability and interpretive rigor.

Future research should examine how generative AI systems handle mathematical problems characterized by undefined variables or ambiguous conditions—scenarios that require inference beyond explicitly stated assumptions. Investigating AI reasoning across a broader range of mathematical domains, including non-integer, irrational, and complex numbers, may further illuminate the strengths and limitations of AI-generated argumentation. Additionally, studies exploring the capacity of AI to construct advanced mathematical proofs and provide step-by-step justifications in higher-level mathematics could yield critical insights into areas for algorithmic refinement. Research might also focus on how AI engages with students in open-ended mathematical discussions and adapts to diverse pedagogical strategies, thereby offering a more nuanced understanding of its potential role in educational settings. Furthermore, longitudinal investigations into the sustained impact of AI integration on students' mathematical reasoning could reveal how such tools influence the development of critical thinking over time.

REFERENCES

[1] Balacheff, N. (1988). *A Study of Students' Proving Processes at The Junior High School Level,* The National Council of Teachers of Mathematics

[2] Biton, Y., & Segal, R. (2025). Learning to Craft and Critically Evaluate Prompts: The Role of Generative AI (ChatGPT) in Enhancing Pre-Service Mathematics Teachers' TPACK and Problem-Posing Skills, *International Journal of Education in Mathematics, Science, and Technology (IJEMST)* **13** (1), (pages 202-223) doi:10.46328/ijemst.4654

[3] Bowen, G. A. (2009). Document Analysis as A Qualitative Research Method, *Qualitative Research Journal* **9** (2), (pages 27–40) doi:10.3316/qrj0902027

[4] Cardetti, F., & LeMay, S. (2018). Argumentation: Building Students' Capacity for Reasoning Essential to Learning Mathematics and Sciences, *PRIMUS* **29** (8), (pages 775–798) doi:10.1080/10511970.2018.1482581

[5] Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The Promises and Challenges of Artificial Intelligence for Teachers: A Systematic Review of Research, *TechTrends* **66** (4), (pages 616–630) doi:10.1007/s11528-022-00715-y

[6] Cope, B., Kalantzis, M., & Searsmith, D. (2020). Artificial Intelligence for Education: Knowledge and Its Assessment in AI-Enabled Learning Ecologies, *Educational Philosophy and Theory*, **53** (12), (pages 1229–1245) doi:10.1080/00131857.2020.1728732

[7] Dilling, F., & Herrmann, M. (2024). Using Large Language Models to Support Pre-Service Teachers Mathematical Reasoning—An Exploratory Study on ChatGPT as An Instrument

for Creating Mathematical Proofs in Geometry, *Frontiers in Artificial Intelligence* **7**, (pages 1-14) doi:10.3389/frai.2024.1460337

[8] Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). "So What if ChatGPT Wrote it?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy, *International Journal of Information Management* **71**, (pages 1-63) doi:10.1016/j.ijinfomgt.2023.102642

[9] Emsley, R. (2023). ChatGPT: These are Not Hallucinations – They're Fabrications and Falsifications, *Schizophrenia* **9**, (pages 1-2) doi:10.1038/s41537-023-00379-4

[10] Engelbrecht, J., & Borba, M. C. (2023). Recent Developments in Using Digital Technology in Mathematics Education, *ZDM* **56** (2), (pages 281–292) doi:10.1007/s11858-023-01530-2

[11] Frieder, S., Pinchetti, L., Griffiths, R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., & Berner, J. (2023). Mathematical Capabilities of ChatGPT, *arXiv (Cornell University)* (pages 1-37) doi:10.48550/arXiv.2301.13867

[12] Fütterer, T., Fischer, C., Alekseeva, A., Chen, X., Tate, T., Warschauer, M., & Gerjets, P. (2023). ChatGPT in Education: Global Reactions to AI Innovations, *Scientifc Reports* **13** (1), (pages 1-14) doi:10.1038/s41598-023-42227-6

[13] Glazer, E., Erdil, E., Besiroglu, T., Chicharro, D., Chen, E., Gunning, A., Olsson, C. F., Denain, J., Ho, A., De Oliveira Santos, E., Järviniemi, O., Barnett, M., Sandler, R., Sevilla, J., Ren, Q., Pratt, E., Levine, L., Barkley, G., Stewart, N., . . . Enugandla, S. V. (2024). FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI, *arXiv (Cornell University)* (pages 1-26) doi:10.48550/arxiv.2411.04872

[14] Gong, C., Jing, C., Chen, X., Pun, C. M., Huang, G., Saha, A., Nieuwoudt, M., Li, H., Hu, Y., & Wang, S. (2023). Generative AI for Brain Image Computing and Brain Network Computing: A Review, *Frontiers in Neuroscience*, **17** (pages 1-24) doi:10.3389/fnins.2023.1203104

[15] Graham, G. (2023). *AI and Math Education: Exploring New Dimensions of Learning*. Retrieved from: https://www.futureschoolai.com/blog/ai-and-math-education-exploring-new-dimensions of-learning.

[16] Hu, J., & Sui, G. (2024). Application of Generative Artificial Intelligence in Linear Algebra Teaching, In *Advances in Social Science, Education and Humanities Research/Advances in social science, education and humanities research* (pages 124–129) doi:10.2991/978-2-38476-297-2_16

[17] Kartika, H., Budiarto, M. T., Fuad, Y., Clark, L. J., & Jeonghyeon, K. (2023) Comparative Analysis of Students' Argumentation Patterns in the Context of Algebraic Problems, *Mathematics Teaching-Research Journal Online,* Vol **15** N-5

[18] Kartika, H., Warmi, A., Urayama, D., & Suprihatiningsih, S. (2024) Mathematical Argumentation in Higher Education: A Systematic Literature Review, *Journal of University Teaching and Learning Practice*, **21** (7) (pages 1-24)

[19] Korteling, J. E., Van De Boer-Visschedijk, G. C., Blankendaal, R. a. M., Boonekamp, R. C., & Eikelboom, A. R. (2021a). Human- versus Artificial Intelligence, *Frontiers in Artificial Intelligence*, **4** doi:10.3389/frai.2021.622364

[20] Korteling, J. E., Gerritsma, J., and Toet, A. (2021b). Retention and Transfer of Cognitive Bias Mitigation Interventions: A Systematic Literature Study, *Front. Psychol,* **12** (pages 1–20) doi:10.3389/fpsyg.2021.629354

[21] Lee, K. (2016). Students' Proof Schemes for Mathematical Proving and Disproving of Propositions, *The Journal of Mathematical Behavior*, **41** (pages 26–44) doi:10.1016/j.jmathb.2015.11.005

[22] Lu, P., Qiu, L., Yu, W., Welleck, S., & Chang, K. (2022). A Survey of Deep Learning for Mathematical Reasoning, *arXiv (Cornell University),* doi:10.48550/arxiv.2212.10535

[23] Markauskaite, L., Marrone, R., Poquet, O., Knight, S., Martinez-Maldonado, R., Howard, S., Tondeur, J., De Laat, M., Shum, S. B., Gašević, D., & Siemens, G. (2022). Rethinking The Entwinement Between Artificial Intelligence and Human Learning: What Capabilities Do Learners Need for A World With AI? *Computers and Education Artificial Intelligence*, **3** (100056) doi:10.1016/j.caeai.2022.100056

[24] Marks Krpan, C., & Sahmbi, G. (2024). Arguing for Access: Teachers' Perspectives on The Use of Argumentation in Elementary Mathematics, *International Journal of Education in Mathematics, Science, and Technology (IJEMST)* **12** (5) (pages 1320-1339) doi:10.46328/ijemst.4385

[25] McNeill, K. L., & Krajcik, J. (2008). Inquiry and Scientific Explanations: Helping Students Use Evidence and Reasoning, In J. Luft, R. L. Bell, & J. Gess-Newsome (Eds.), *Science as Inquiry in the Secondary Setting* (pages 121-134), Arlington, VA: National Science Teacher Association

[26] Meyer, M., & Schnell, S. (2020). What Counts as a "good" Argument in School? -How Teachers Grade Students' Mathematical Arguments, *Educational Studies in Mathematics* **105** (1), (pages 35-51) doi:10.1007/s10649-020-09974-z

[27] Mishra, P., Warr, M., & Islam, R. (2023). TPACK in the Age of ChatGPT and Generative AI, *Journal of Digital Learning in Teacher Education*, **39** (4), (pages 235–251) doi:10.1080/21532974.2023. 2247480

[28] Noster, N., Gerber, S., & Siller, S. (2024). Pre-Service Teachers' Approaches in Solving Mathematics Tasks with ChatGPT, *Digital Experiences in Mathematics Education*, **10**, (pages543-567) doi:10.1007/s40751-024-00155-8

[29] Oh, S. (2024). Evaluating Mathematical Problem-Solving Abilities of Generative AI models: Performance Analysis of ChatGPT 4o1 and ChatGPT 4O using the Korean College Scholastic Ability Test. *IEEE Access*, **1** doi:10.1109/access.2024.3523703

[30] Park, H., & Manley, E. D. (2024). Using ChatGPT as A Proof Assistant in A Mathematics Pathways Course, *The Mathematics Education,* **63** (2), (pages 139-163) doi:10.7468/mathedu.2024.63.2.139

[31] Pielsticker, F., Holten, K., Dilling, F., & Witzke, I. (2024). Promoting Negotiation and Argumentation Processes Through the Use of Generative AI Language Models in School Mathematics Learning? Initial Insights and Perspectives from Empirical Research, *Newsletter of the GDM*, **116**, (pages 14-22)

[32] Rane, N. (2023). Enhancing Mathematical Capabilities Through ChatGPT and Similar Generative Artificial Intelligence: Roles and Challenges in Solving Mathematical Problems, *SSRN Electronic Journal,* (pages 1-9) doi:10.2139/ssrn.4603237

[33] Relmasira, S. C., Lai, Y. C., & Donaldson, J. P. (2023). Fostering AI Literacy in Elementary Science, Technology, Engineering, Art, and Mathematics (STEAM) Education in The Age of Generative AI, *Sustainability*, **15** (18), 13595 doi:10.3390/su151813595

[34] Sabzalieva, E., & Valentini, A. (2023). *ChatGPT and Artificial Intelligence in Higher Education: Quick Start Guide,* UNESCO. Retrieved from: https://www.iesalc.unesco.org/wpcontent/uploads/2023/04/ChatGPT-and-Artificial-Intelligence-in-higher-education-Quick-Start-guide_EN_FINAL.pdf

[35] 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** doi:10.3389/feduc.2024.1386075

[36] Stylianides, A. J. (2019). Secondary Students' Proof Constructions in Mathematics: The Role of Written Versus Oral Mode of Argument Representation, *Review of Education*, **7** (1), (pages 156-182)

[37] Stylianides, G. J., Stylianides, A. J., & Shilling-Traina, L. N. (2013). Prospective Teachers' Challenges in Reasoning-and-Proving, *International Journal of Science and Mathematics Education*, **11**, (pages 1463-1490)

[38] Tenhundfeld, N. L. (2023). Two Birds with One Stone: Writing a Paper Entitled "ChatGPT as a Tool for Studying Human-AI Interaction in the Wild" with ChatGPT, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, **67** (1), (pages 2007–2012) doi:10.1177/21695067231192916

[39] Urhan, S., & Zengin, Y. (2023). Investigating University Students' Argumentations and Proofs Using Dynamic Mathematics Software in Collaborative Learning, Debate, and Self-Reflection Stages, *International Journal of Research in Undergraduate Mathematics Education,* **10**, (pages 380-407) doi:10.1007/s40753-022-00207-7

[40] 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), (pages 1-18) doi:10.29333/ejmste/13272

[41] Xu, Y., Liu, X., Liu, X., Hou, Z., Li, Y., Zhang, X., Wang, Z., Zeng, A., Du, Z., Zhao, W., Tang, J., & Dong, Y. (2024). ChatGLM-Math: Improving Math Problem-Solving in Large Language Models with a Self-Critique Pipeline, *arXiv (Cornell University),* doi:10.48550/arxiv.2404.02893

[42] Yoon, H., Hwang, J., Lee, K., Roh, K. H., & Kwon, O. N. (2024). Students' Use of Generative Artificial Intelligence for Proving Mathematical Statements, *ZDM*, **56** (7), (pages 1531–1551) doi:10.1007/s11858-024-01629-0

[43] Zambak, V. S., & Magiera, M. T. (2020). Supporting Grades 1–8 Preservice Teachers' Argumentation Skills: Constructing Mathematical Arguments in Situations That Facilitate Analyzing Cases, *International Journal of Mathematical Education in Science and Technology*, **51** (8), (pages 1196-1223) doi:10.1080/0020739X.2020.1762938