Undergraduate Students' Use of Generative Artificial Intelligence in Proving

Jihye Hwang
Arizona State University

Hyunkyoung Yoon
California State Polytechnic University, Pomona

Kyungwon Lee
Seoul National University

Oh Nam Kwon
Seoul National University

In this paper, we explore how undergraduate students use and evaluate generative artificial intelligence (genAI) in proving. We view proving as a human activity and proof as a production of proving, and hence, we believe that students need to be the ones who evaluate the genAI-generated arguments and write their proofs. In the initial phase of research on genAI in proving, we conducted interviews with three undergraduate students to examine how they use genAI in proving and how they write their proofs after their use of genAI. This study revealed that there is a flow in students' use of ChatGPT in proving, and there are factors that impact their use of genAI. We think that the flowchart presented here can serve as a guide for both teaching and research on the use of genAI in the context of proving.

Keywords: Artificial Intelligence (AI), Undergraduate Students, Proving, ChatGPT

Since the end of 2022, when a generative Artificial Intelligence (genAI) tool became public, ongoing discussions about the role of AI in education have accelerated. People started viewing genAI as a tool that can both benefit and hinder students' learning. As the prevalence of the use of genAI in education is inevitable, it is important to consider how educators can help and guide students' responsible use of genAI in education. Based on its affordances (e.g., instant feedback or personalized learning) and limitations (e.g., wrong information or ethical issues), we believe that students should be able to be aware of the pitfalls of AI tools and critically evaluate responses generated by AI tools.

We chose proving as our research context because proving is a context where students need to persuade themselves and convince others (Bieda & Lepak, 2014; Hanna, 2000; Hersh, 1993). Using the feature of genAI that provides instant conversation, it can be used as a good tool to practice convincing each other. Another benefit of using a genAI as a tool in proving can be that it does not necessarily involve negative human interactions with the user. For example, when students ask questions to their instructors or fellow students, power relationships exist between them. Then, some students may take mathematical statements and proofs from the authorities, including instructors, peers, or textbooks, without questioning them (Weber & Mejia-Ramos, 2014). On the contrary, using AI might help students feel more agency and authority by being free from those power dynamics because there is no human behind the tool and their (intentional or unintentional) judgment about the student. Thus, communicating with genAI can become a place where students can have authority in evaluating the argument if used appropriately.

As there have been only a few, if any, studies regarding the use of AI in undergraduate mathematics-level proving tasks, this study initiates exploring this area. We investigated the following research question: *How would undergraduate students use ChatGPT when they are proving a mathematical statement?* Through the study, we call for more conversation about the use of genAI tools in proving by proposing a set of aspects to be considered when using genAI in proving and how those aspects are connected and interact with each other, by revising the IDEE framework by Su and Yang (2023).

Backgrounds and Theoretical Perspectives

In this section, we review 1) the affordances and limitations of genAI, 2) our perspectives on proving, and 3) a framework for using a genAI in proving by modifying the IDEE framework (Su & Yang, 2023).

Affordances and Limitations of GenAI in Education

As emerging of genAI tools was recent, at the end of 2022, which is about a year ago, not much research exists in education in general, and even less in mathematics education. The most common form of current research articles is examining genAI as a tool for education to find affordances and limitations. Researchers viewed that genAI could be of benefit to students' learning, including but not limited to the following: 1) provide personalized and interactive learning environment tailored to individual students' preferences and levels (Baidoo-Anu & Owusu Ansah, 2023; Qadir, 2023; Su & Yang, 2023); 2) provide generate formative assessment tools (Baidoo-Anu & Owusu Ansah, 2023; Qadir, 2023); and 3) provide instant feedback (Baidoo-Anu & Owusu Ansah, 2023). GenAI tools, however, have some limitations and challenges as well. First, several researchers pointed out that genAI can provide wrong information depending on the data set that it was trained in (Baidoo-Anu & Owusu Ansah, 2023; Qadir, 2023; Su & Yang, 2023). Researchers claimed that the accuracy of genAI responses would be lower in complex tasks (Su & Yang, 2023). Second, ethical concerns exist regarding using genAI, such as plagiarism and overreliance (Qadir, 2023), bias in the data set (Baidoo-Anu & Owusu Ansah, 2023), users' privacy (Baidoo-Anu & Owusu Ansah, 2023; Qadir, 2023; Su & Yang, 2023), and equity issues (Cooper, 2023; Qadir, 2023).

Based on this review, as mentioned previously, we view the critical use of genAI as a skill for students to take advantage of the tool, and hence, we aimed to provide a guide for the goal.

Proving and Proofs

In this study, we viewed proving as human activities with the goal of generating proofs. By proof, we adapt the definition of Stylianides (2007). Stylianides defined mathematical proof as

> … a mathematical argument, a connected sequence of assertions against a mathematical claim, with the following characteristics: 1. It uses statements accepted by the classroom community (*set of accepted statements*) that are true and available without further justification; 2. It employs forms of reasoning (*modes of argumentation*) that are valid and known to, or within the conceptual reach of, the classroom community; 3. It is communicated with forms of expression (*modes of argument representation*) that are appropriate and known to, or within the conceptual reach of, the classroom community. (p. 291).

The key point of this definition is that an argument can be referred to as proof when it is approved by the classroom community, which is their mathematical community. It aligns with other researchers' arguments that a mathematical argument becomes proof when a mathematical community agrees on it (Weber & Mejia-Ramos, 2014). From this perspective, what genAI produces is a mathematical argument, but it does not become proof until a mathematical community sanctions it. In other words, students who use genAI become the first person in the line who gets to decide if the generated argument can be qualified as a mathematical proof considering the three components mentioned in the definition of proof before they present this to other people. Thus, the subject who writes and presents mathematical proof should be and is a person, not genAI.

IDEE Framework

Su and Yang (2023) suggested a theoretical framework as a guide to using generative AI in education. The framework has four aspects:

Identify the Desired Outcomes. Identifying the objects of generative AI before using it.

Determine the Appropriate Level of Automation. Automating teaching or learning experiences using generative AI or using it as a supplement to traditional teaching methods, depending on the objectives.

Ensure Ethical Considerations. Considering ethical implications of using generative AI, including potential biases and their impact on teachers and students.

Evaluate the Effectiveness. Evaluating the effectiveness of generative AI in achieving the desired outcomes.

While the IDEE framework (Su & Yang, 2023) is for teaching and learning in general education, this study focuses on students' use of generative AI in the context of mathematical proof. We interpreted the framework by focusing on (1) mathematics education, more specifically for students' proving, and (2) students' use of generative AI for their learning. Thus, the following is our interpretation of the IDEE framework.

Identify the Desired Outcomes. Identifying the objects of generative AI before using it when proving mathematical statements

Determine the Appropriate Level of Automation. Deciding the ways to use generative AI and the extent to which they want to take from what generative AI generates depends on the objectives.

Ensure Ethical Considerations. Students' recognition of potential ethical issues, such as determining whether using generative AI for homework aligns with academic integrity.

Evaluate the Effectiveness. Evaluating the effectiveness of generative AI in achieving the desired outcomes when proving mathematical statements.

Methods

Data Collection

Participants. In July 2023, we conducted individual interviews with three participants, Yuna, Jiho, and Kitae (all pseudonyms), from two different universities in Korea. All of them were freshmen mathematics education majors. Yuna was a woman, and Jiho and Kitae were men. Yuna took a Linear Algebra course, which she described as a proof-based course, as well as Calculus I, and the other two took only Calculus I.

GenAI. We used ChatGPT 3.5 as a genAI in this study because it was the most commonly used genAI in Korea at the time of data collection, July 2023.

Interviews. At the beginning of the interview, students were given six mathematical statements, mostly from number theory and one from advanced calculus. The statements were written both in Korean and English. Students were asked to choose one or two statements they wanted to prove or disprove and allowed to use any resources. We did not ask them to use ChatGPT for this part to see if it occurs naturally for them. However, none of them used the ChatGPT in the first part, so we asked them if they were aware of the tool and had used it. All of them said they had used ChatGPT before, so we asked them to use ChatGPT for proving tasks, observed how they used ChatGPT, and evaluated the argument generated by ChatGPT. During the interviews, a main interviewer led the interview, and a secondary interviewer took notes and asked supplementary questions if needed. Each interview lasted 80 to 100 minutes and was transcribed into Korean after the collection.

Data Analysis

For data analysis, we used the modified IDEE framework. We first coded the transcripts using the four categories: 1) desired outcomes, 2) level of automation, 3) ethical consideration, and 4) effectiveness. In applying the initial categories, we viewed the level of automation as students' decision process to get an argument and effectiveness as students' evaluation of the argument generated by ChatGPT. If students' use of ChatGPT in proving does not fit into one of the four categories, we coded those as others. First, each participant's transcripts were coded by one of the team members; then, different team members did the second round for cross-checking purposes. When the team members detected differences, it was discussed until the team agreed on specific coding. Through this repetitive process, we revised and specified the codes (Table 1).

Table 1 Code Descriptions

Codes	Revised Codes
Desired outcomes	Objectives of using genAI.
Level of automation	Decisions on prompts to ask (differently) to genAI.
	Decisions on continuing conversation with genAI.
	Decisions on how much they want to take from the generated arguments by AI.
Ethical consideration	Opinions regarding ethical issues
Evaluation of effectiveness	Evaluations of genAI's argument
Others	GenAI belief
	Proving belief
	Proving knowledge

Findings

In this section, we present how Yuna used and understood ChatGPT as a learning resource and how she used ChatGPT in a proving task. We only present Yuna's case here because of the page constraints.

Yuna's Previous Experiences in Using ChatGPT

Yuna was the participant who often used ChatGPT for her undergraduate education in general. She used ChatGPT in mathematics when she had a problem that she did not know how to solve or did not have enough time to complete assignments. Yuna generally noted that ChatGPT was a useful tool that expanded her draft and met the quantity requirement for writing assignments in other contexts. She said ChatGPT was not that useful in studying mathematics because 1) it is not easy to type mathematical symbols into it, and 2) sometimes it generates different answers even if the same problem was typed in. Because of these reasons, she preferred searching online (e.g., Google or MathStackExchange) to ChatGPT for proving tasks. Lastly, despite her negative evaluations for ChatGPT in proving, that she thought using ChatGPT was not cheating and did not have an ethical problem. She thought using ChatGPT did not matter what she used as resources for assignments and said maybe it was even better than copying peers' assignments.

Yuna's Use of ChatGPT in a Proving Task

As she did not use ChatGPT without the interviewer's direction, we asked her to use ChatGPT. In Yuna's use of ChatGPT for a proving task, we found that she had a conversation with ChatGPT by entering prompts several times. We found a repetitive pattern of 1) making decisions on prompts, 2) executing the prompts in ChatGPT, and 3) evaluating the responses until Yuna felt that she was ready to write her proof, which is detailed below (Yuna's pattern of using ChatGPT in proving tasks in Figure 1).

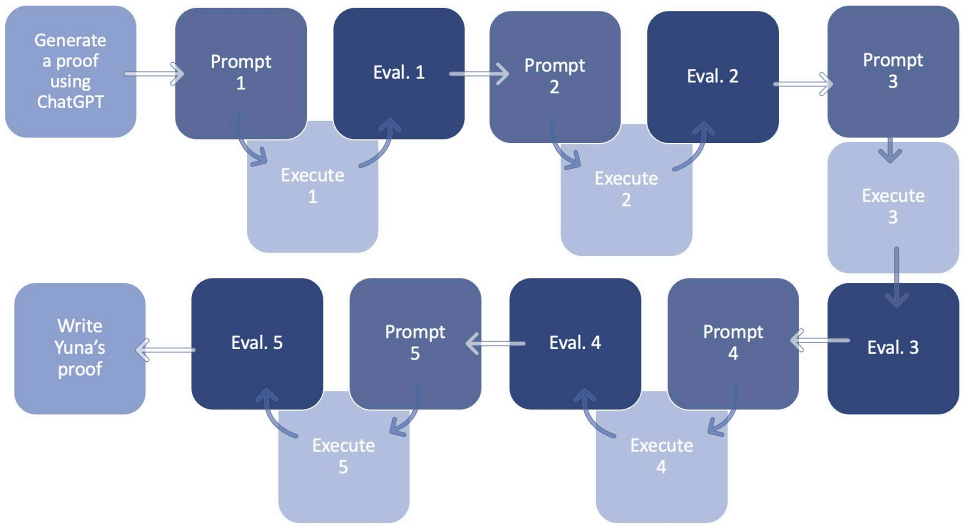

Figure 1 Yuna's pattern of using ChatGPT in proving tasks

Given a set of questions, Yuna selected the following statement to use ChatGPT for proving: *For all integers m and n, if the product mn is odd, then m is odd and n is odd.* As soon as she chose the problem, she copied and pasted the statement directly to prompt ChatGPT (Prompt 1 in Figure 2). ChatGPT provided an argument claiming that the statement was false. When Yuna read the ChatGPT-generated argument, she found the argument was conflicting with her conjecture that the statement was true. Then, she revised the prompt by adding "can you prove this statement:" in front of the previous prompt and typing "m*n" instead of "mn" to clarify that it is about multiplication (Prompt 2 in Figure 2). With Prompt 2, ChatGPT generated an argument using proof by contradiction that discussed the other three cases, except the case when both *m* and *n* are odd. Yuna initially felt that the argument seemed correct but soon questioned:

> It seems like it is saying the right thing… There exist four cases for *m* and *n,* and it used all the cases except when both *m* and *n* are odds, and claiming these cases led to a conclusion that *mn* is even, but I don't know where the remaining case is. [Translated from Korean]

Then, we asked if she knew what the word "proof by contradiction" meant and translated it into Korean. Then she typed Prompt 3 in Figure 2 in Korean, saying: "There are too many English words than mathematical expressions compared to what I am used to, so it is hard to recognize the argument in skimming." When ChatGPT produced the response for Prompt 3, she was not satisfied again and prompted Prompt 4 in Figure 2. Yuna noted that both responses did not include many mathematical expressions, and she did not feel ChatGPT's response was rigorous enough to be considered as proof.

When the interviewer asked if she wanted to prompt further, she typed Prompt 5 (Figure 2), adding "considering cases for m and n," into the prompt because she did not understand why the argument generated by ChatGPT used proof by contradiction. She felt this response was similar to what she was thinking and decided she would not use ChatGPT anymore as she knew what to do. Yuna noted that she probably would not use proof by contradiction to write her final answer. Even though Yuna said that she would not use "proof by contradiction" in her final answer, she used the format of proof by contradiction without the word “proof by contradiction” in her answer.

Prompt	Yuna’s Prompts	Translation
1	For all integers m and n, if the product mn is odd, then m is odd and n is odd.	N/A
2	can you prove this statement: For all integers m and n, if the product m*n is odd, then m is odd and n is odd.	N/A
3	모든 정수 m 과 n 에 대하여, 만약 그 곱 mn 이 홀수이면, m 도 홀수이고 n 도 홀수이다.	For all integers m and n, if the product mn is odd, then m is odd and n is odd.
4	귀류법을 사용해서 모든 정수 m 과 n 에 대하여, 만약 그 곱 *mn* 이 홀수이면, *m* 도 홀수이고 *n* 도 홀수이다. 를 증명해봐.	By using proof by contradiction, prove. For all integers m and n, if the product mn is odd, then m is odd and n is odd.
5	귀류법을 사용해서 m 과 n 을 경우를 나누어서 아까의 명제를 증명해봐.	By using proof by contradiction and considering cases for m and n, prove the previous statement

Figure 2 Yuna’s prompts used in ChatGPT

Discussions

By combining the analysis of Yuna’s and the other students’ use of ChatGPT in their proving tasks, we present a flow chart (Figure 3) to illustrate students’ process of using ChatGPT in proving by revising IDEE framework (Su & Yang, 2023).

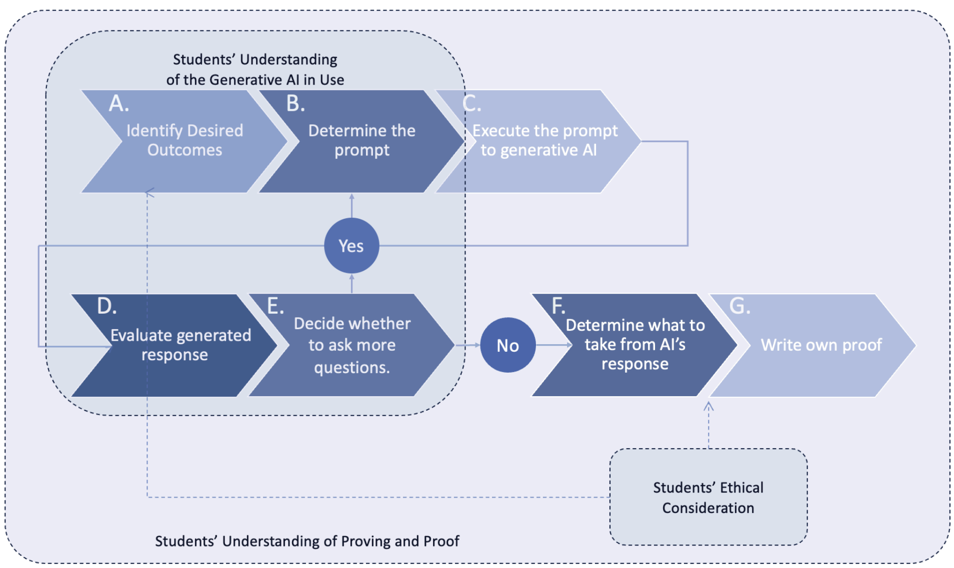

Figure 3 Flowchart of students’ use of ChatGPT in proving

A. **Identify desired outcomes.** Students had their objectives by using genAI in proving. The outcomes could be determined by students themselves or given by other people such as an instructor or researcher.
B. **Determine the prompt.** Once students decided on their objectives for using genAI, they decided how they wanted to ask such questions. Yuna’s various prompts are in Figure 2.
C. **Execute the prompt to genAI.** As students entered their determined prompts, they received the answers from the ChatGPT immediately.
D. **Evaluate the generated response.** When students evaluated generated responses, their evaluation varied based on the objectives identified in the first stage. As mentioned in the

example, Yuna made several evaluations of the arguments based on her understanding of proofs and mathematical knowledge.

E. **Decide whether to ask more questions.** At this point, students decided if they wanted to use genAI further to get another response based on their evaluation at D. Based on the answer, students could repeat the cycle between B and E as many as they wanted to. In Yuna's case, she repeated the cycle five times until she decided to ask no further questions.
F. **Determine what to take from AI's response.** When they decided to stop using genAI, students made a decision on which part of genAI's responses they would like to take and discard for their proof. It could include a set of accepted arguments, modes of argumentation, and modes of argument representations. For example, Yuna hesitated to use proof by contradiction in her proof because that was not the mode of argumentation she had seen in her mathematics classrooms.
G. **Write their own proof.** As a last step, students wrote their proof as the author of their proof. This stage emphasizes proof is human activity, and what counts as a proof is determined by humans, and in this case, the user of the genAI.

This process underscores that while automation tools like genAI are used, proving remains fundamentally a human activity. In her interactions with genAI, Yuna maintained the authority to critique genAI's arguments and determine their incorporation in her proofs. This finding resonates with the significance of students' agency and autonomous actions in proving (Castle et al., 2022) could critically engage with genAI (Cooper, 2023) by resisting passively accepting arguments from external resources, including gen AI, by giving them the authority (Castle et al., 2022).

In addition to the processes, we identified three other aspects (highlighted in shaded areas of proving Figure 3) considered in their use of genAI for proving:

1. **Students' understanding of proving and proof:** Influences the entire process, including evaluating genAI's suggestions and writing students' final proof. For instance, Yuna believed a mathematical proof should be more equation-heavy and less verbose. She thus modified ChatGPT's verbose argument to a more symbolic form in her proof.
2. **Students' understanding of genAI in use:** Concerns how students perceive the genAI tool they employ. This might include their knowledge about genAI's capabilities based on prior experiences or beliefs. Yuna, believing it cumbersome to type mathematical notations, often refrained from doing so.
3. **Students' ethical considerations:** Relates to students' awareness of potential ethical dilemmas in using genAI. This might be influenced by class norms or personal beliefs about leveraging resources. Yuna, for example, felt using ChatGPT was "better" than copying others' work as at least one tried to type questions and checked the argument by ChatGPT.

We believe these three factors would need to be considered in designing and implementing tasks using genAI. Also, the presented flowchart in this paper could provide a guideline to mathematics instructors to devise their proving activities using genAI in their teaching by suggesting some sequences that should be considered. Also, it means that instructors need to know their students' understanding of proving and proof and genAI and should develop and negotiate the classroom norm with their students. This flow chart could also serve as a tool for analyzing how students use genAI in their proving tasks and the extent to which students critically think about the responses generated by genAI.

References

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative Artificial Intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. SSRN Electronic Journal, 1–22. https://doi.org/10.2139/ssrn.4337484

Bieda, K. N., & Lepak, J. (2014). Are you convinced? Middle-grade students' evaluations of mathematical arguments. School Science and Mathematics, 114(4), 166–177. https://doi.org/10.1111/ssm.12066

Castle, S. D., Smith III, J. P., Levin, M., Hwang, J., Karunakaran, S. S., Küchle, V. A. B., & Elmore, R. (2022). Shifts in external authority and resources for sense-making in the transition to proof-intensive mathematics: the case of Amelia. In S. S. Karunakaran & A. Higgins (Eds.), Proceedings of the 24th Annual Conference on Research in Undergraduate Mathematics Education (pp. 100–107).

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative Artificial Intelligence. Journal of Science Education and Technology, 32(3), 444–452. https://doi.org/10.1007/s10956-023-10039-y

Hanna, G. (2000). Proof, explanation and exploration: An overview. Educational Studies in Mathematics, 44(1–3), 5–23. https://doi.org/10.1023/A:1012737223465

Hersh, R. (1993). Proving is convincing and explaining. Educational Studies in Mathematics, 24(4), 389–399. https://doi.org/10.1007/BF01273372

Qadir, J. (2023). Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. In IEEE (Ed.), 2023 IEEE Global Engineering Education Conference (pp. 1–9). IEEE.

Stylianides, A. J. (2007). Proof and proving in mathematics education. Journal for Research in Mathematics Education Journal for Research in Mathematics Education, 38(3), 289–321. https://doi.org/10.1007/978-94-007-2129-6

Su, J., & Yang, W. (2023). Unlocking the power of ChatGPT: a framework for applying generative AI in education. ECNU Review of Education, 1–12. https://doi.org/10.1177/20965311231168423

Weber, K., & Mejia-ramos, J. P. (2014). Mathematics majors' beliefs about proof reading. International Journal of Mathematical Education in Science and Technology, 45(1), 89–103. https://doi.org/10.1080/0020739X.2013.790514