MathEdu Journal of the Korean Society of Mathematical Education Series A

Research Article

The Mathematical Education 2024;63(2):139-163
https://doi.org/10.7468/mathedu.2024.63.2.139
pISSN 1225-1380, eISSN 2287-9633

Using ChatGPT as a proof assistant in a mathematics pathways course

Hyejin Park[1]*, Eric D. Manley[2]

[1]Assistant Professor of Mathematics, Drake University
[2]Associate Professor of Computer Science, Drake University

ABSTRACT

The purpose of this study is to examine the capabilities of ChatGPT as a tool for supporting students in generating mathematical arguments that can be considered proofs. To examine this, we engaged students enrolled in a mathematics pathways course in evaluating and revising their original arguments using ChatGPT feedback. Students attempted to find and prove a method for the area of a triangle given its side lengths. Instead of directly asking students to prove a formula, we asked them to explore a method to find the area of a triangle given the lengths of its sides and justify why their methods work. Students completed these ChatGPT-embedded proving activities as class homework. To investigate the capabilities of ChatGPT as a proof tutor, we used these student homework responses as data for this study. We analyzed and compared original and revised arguments students constructed with and without ChatGPT assistance. We also analyzed student-written responses about their perspectives on mathematical proof and proving and their thoughts on using ChatGPT as a proof assistant. Our analysis shows that our participants' approaches to constructing, evaluating, and revising their arguments aligned with their perspectives on proof and proving. They saw ChatGPT's evaluations of their arguments as similar to how they usually evaluate arguments of themselves and others. Mostly, they agreed with ChatGPT's suggestions to make their original arguments more proof-like. They, therefore, revised their original arguments following ChatGPT's suggestions, focusing on improving clarity, providing additional justifications, and showing the generality of their arguments. Further investigation is needed to explore how ChatGPT can be effectively used as a tool in teaching and learning mathematical proof and proof-writing.

Keywords ChatGPT, Proof construction, Proof evaluation, Proof assistant tool, Beliefs about proof, Large language model

Introduction

Learning to write mathematical proofs can be difficult for many learners (novice provers), and the ways in which Artificial Intelligence (AI) can be leveraged to assist these learners is an active research area. Research into automated theorem provers (ATPs), interactive theorem provers, and other proof assistants

Received April 8, 2024; **Revised** April 16, 2024; **Accepted** May 8, 2024

***Corresponding author** Hyejin Park
E-mail hyejin.park@drake.edu
2000 Mathematics Subject Classification 97C80

(e.g., Isabelle/HOL [Nipkow et al., 2002], Coq [Bertot & Castéran, 2013], Lean [de Moura et al., 2015]), have a history of harnessing AI algorithms to assist mathematicians with formal proofs of mathematical theorems. For example, a machine-generated proof of the Four-Color Theorem was first produced by Appel and Haken in 1977 (Appel & Haken, 1977) and then verified by Georges Gonthier in 2005 using Coq (Gonthier, 2008). These technologies utilize formal languages to express mathematical ideas and strategically apply rules of logic to formally prove inference. Large Language Models (LLMs) have recently been fine-tuned on proofs in these formal languages and have demonstrated significant improvements in proof generation on benchmark datasets. Examples of LLM-powered ATPs include Thor (Jiang et al., 2022), Baldur (First et al., 2023), and LeanDojo (Yang et al., 2023). While these tools were not primarily designed for mathematics education purposes, some of the techniques have been integrated into tools for mathematics education, like GeoGebra, to assist students in learning to prove through verification of proof correctness, generation and evaluation of conjectures, and proof generation (Botana et al., 2015).

Traditionally, Intelligent Tutoring Systems (ITSs) for mathematics include both interactive and automated theorem-proving components with language models, enabling communication with the system and mathematically sound inference. With recent advances in general-purpose LLMs like the Open AI GPT and Google Gemini families of models, the language components of ITSs are poised to improve remarkably. At the same time, consumer applications like ChatGPT, Microsoft Copilot, Google Gemini, and xAI Grok have made these LLMs widely available, and students are becoming increasingly familiar with them, using them for a variety of language-oriented tasks. While these LLMs alone do not explicitly incorporate proving components like specialized ITSs, they display procedural reasoning through their abilities to generate coherent program code and explain mathematical concepts. This raises the question of whether and how students can learn to write better natural-language-based proofs that can convince themselves or others with the assistance of general-purpose LLM-based applications. The effectiveness of using these LLM-based tools as proof assistants in developing students' abilities to read, understand, analyze, evaluate, and construct mathematical arguments (proofs) has not yet been explored.

Since ChatGPT was launched in November 2022, educators and related shareholders have hotly debated its potential use and misuse as a learning and teaching tool. In our study, we conjecture that ChatGPT can function as a proof assistant supporting students (novice provers) in their proving activities. To examine the capabilities of ChatGPT as a tool to assist students in generating arguments that can be considered proofs, we purposefully created a set of ChatGPT-embedded proving activities for learners. We explored how students used ChatGPT when asked to improve their arguments while interacting with ChatGPT after attempting to prove a mathematical claim by themselves. Our study context was the Mathematical Pathways course designed for college students to deeply explore mathematics concepts covered in school mathematics to build their conceptual understanding and reasoning skills, focusing on mathematical argumentation and proof. The following research questions guided our study:

RQ1) What conceptions do students possess about mathematical proof and proving?

RQ2) How do students initially construct arguments that can be considered proofs before revising them with ChatGPT assistance?

RQ3) What feedback does ChatGPT provide when students ask to evaluate whether their arguments look like mathematical proofs?

RQ4) What feedback from ChatGPT do students take into consideration when revising their arguments after attempting to construct the arguments themselves?

RQ5) How do students feel about using ChatGPT in the process of improving arguments?

Related Literature

1. Students' Proof Constructions

Proving is the heart of mathematical practice. The importance of learning and teaching proof and its multiple roles in mathematics, such as verification, explanation, discovery, communication, and systematization (de Villiers, 1990), have been spotlighted in mathematics education. Researchers emphasized its explanatory power and urged teachers to use proofs that explain to promote students' mathematical understanding (Hanna, 2000; Knuth, 2002). However, as a large body of research has shown (Coe & Ruthven, 1994; Harel & Sowder, 1998; Knapp, 2005; Moore, 1994; Selden & Selden, 1995; Selden, 2012), constructing mathematical proofs is challenging for novice provers. Moore (1994) reported that students' lack of understanding of mathematical concepts and methods of proof could lead them to fail to prove. Other difficulties students sometimes encounter when attempting to prove include not knowing what theorems or definitions to use (Weber, 2001) and sometimes even unsure how to start a proof (Moore, 1994). Prior studies also show that students' views of proof could influence their approaches to generating proofs (Healy & Hoyles, 2000). For instance, students may produce empirical arguments by relying on specific examples (or cases) when asked to prove a general statement. This is because they may see an empirical argument as an acceptable proof without understanding its limitations. Such an approach to producing arguments is the most common approach observed among many students across all grade levels based on results from prior studies. Weber and Alcock (2004) defined this proof production approach as semantic. According to their definition of this approach, it is "a proof of a statement in which the prover uses instantiation(s) of the mathematical object(s) to which the statement applies to suggest and guide the formal inferences that he or she draws" (p. 211). Alcock and Weber (2010) also identified students whose approach to proving is syntactic when asked to prove a mathematical statement. A syntactic proof production is one "which is written solely by manipulating correctly stated definitions and other relevant facts in a logically permissible way" (Weber & Alcock, 2004, p. 210). Exploring how to support students better to help them overcome their difficulties with proving, such as using logic, proof techniques, worked examples, proof assistants (e.g., GeoGebra automated reasoning tools), or a carefully designed instructional sequence with multiple tasks through which students can learn limitations of empirical arguments in terms of generality, has long been one of the researchers' interest investigation areas over the years (Botana et al., 2015; Epp, 2003; Knuth et al., 2009; Marty, 1986; Papadopoulos, 2016; Stylianides & Stylianides, 2009). This present study will contribute to the literature by examining the capabilities of ChatGPT as a proof assistant tool, exploring students' work with ChatGPT during their proof construction and evaluation activities.

2. Use of AI in Mathematics Classrooms

Van Vaerenbergh and Pérez-Suay (2022) give a taxonomy of Artificial Intelligence (AI) systems used in Intelligent Tutoring Systems (ITSs) such as Hypergraph Based Problem Solver (HBPS) (Arnau et al., 2013) and QED-Tutrix (Font et al., 2018, 2022) that can assist students in solving math problems as well as in learning proof and proving techniques. The different ITS components they describe include information extractors, reasoning engines, explainers, and data-driven modeling. Information extractors attempt to take the input produced by the student (e.g., their written work and drawings) and represent them internally so that reasoning engines (e.g., automated theorem provers [Botana et al., 2015; Fitting, 2012]) can determine sequences of inferences to generate a proof. The task of presenting information generated by the reasoning engines for human users is done by explainer systems. Data-driven modeling takes advantage of data on student work or interaction with the ITS to analyze the learning process or to create adaptive systems that adjust for particular student needs (Van Vaerenbergh & Pérez-Suay, 2022).

Recent advances in deep learning, particularly convolutional neural networks (Li et al., 2021) and transformers (Vaswani et al., 2017), have been used in breakthrough computer vision and natural language applications. These evolving technologies have the potential to improve each of the four components in Van Vaerenbergh and Perez-Suay's taxonomy, especially the information extractors and explainers. However, the nature of most current ITSs is special purpose. For example, HBPS works with word problems, and QED-Tutrix works with geometric proofs typical in secondary education settings. In their study, Arnau et al. (2013) show that HPBS can be used as an assistant tool to support students in solving algebraic word problems. They found that university students who used HPBS could eventually solve word problems that they initially could not and improve their algebraic translation competence significantly compared with the control group. However, these extant ITSs often require tedious manual labor to configure for a particular problem, along with special training needed to use the software (Font et al., 2018).

On the other hand, transformers, which underpin large language models (LLMs) like GPT-3 and GPT-4, on their own can offer a possible approach to assisting in proof construction. Unlike the specialized ITSs previously discussed, these models have been trained on extremely large datasets spanning numerous disciplines, giving them the ability to engage in diverse language tasks such as summarization, translation, and question-answering (Brown et al., 2020). Their ability to generate coherent explanations of mathematical concepts, act as a mathematics fact-querying engine (Frieder et al., 2024), and produce functional computer code (Chen et al., 2021; Finnie-Ansley et al., 2022) in response to natural language prompts demonstrates a form of procedural reasoning. They have demonstrated the ability to solve mathematical problems like symbolic integration and eigenvalue computation despite not explicitly engaging in mathematical reasoning (National Academies of Sciences, Engineering, and Medicine, 2023). All of this suggests that, while they do not possess an explicit reasoning engine like the ITSs discussed above, they still have potential as a proof tutor. This potential should be investigated as students become increasingly familiar with these LLMs through applications like ChatGPT. With this motivation, our study seeks to understand whether students can improve their proof construction and evaluation skills with the help of ChatGPT.

Research on learning and teaching mathematics with ChatGPT assistance has been radically growing over the past few years since ChatGPT was launched; however, based on our literature review, *little* has been addressed yet about the potential use of ChatGPT as a proof assistant in the context of mathematics classrooms. Existing studies largely focus on investigating (both prospective and practicing) mathematics teachers' perspectives on using ChatGPT, such as in creating teaching materials (e.g., lesson plans and mathematical tasks) and teaching. For example, Wardat et al. (2023) studied ChatGPT user perspectives of mathematics teachers and other professionals through interviews and user experience testing. These users thought that ChatGPT could be used for interactive support of mathematics education and had the ability to solve many mathematical problems found in educational settings; however, they also noted that it had to be used with caution, given that it sometimes produces errors or inaccuracies. Gattupalli et al. (2023) studied pre-service teachers' preferences for mathematics materials (hints for solving mathematics problems) generated by both humans and GPT-4. They found that the participants preferred human-written material overall but appreciated the detail, clarity, and step-by-step instructions generated by the language model. Research has also shown the effectiveness of ChatGPT in promoting students' mathematics learning experience and their cognitive and affective development. For instance, Zafrullah et al. (2023) surveyed prospective mathematics teachers in Indonesia and found they had high interest in learning mathematics using ChatGPT. Similarly, Patero (2023) found that integrating ChatGPT into the high school mathematics curriculum increased students' interest, self-efficacy, and performance in mathematics. Butgereit and Martinus (2023) described the implementation of a project called Prof Pi, in which GPT-4 was integrated into Whatsapp and deployed for mathematics homework assistance in underserved areas, providing several examples of how it was able to help students with math problems in algebra, geometry, trigonometry, and calculus.

Methods

1. Study Context and Participants

To answer the research questions, we collected data from two sections of a Mathematics Pathways course during the Fall 2023 semester at a mid-sized university located in the Midwest United States. The Mathematics Pathways course is offered by the math department every semester for liberal arts or business majors. Students usually take this course either to fulfill the quantitative general education requirement or in preparation for precalculus to review mathematical concepts covered in school mathematics that they might have missed. Historically, students enrolled in this course lack confidence and interest in mathematics. So, the main goals of this course are to help students build conceptual understandings of mathematics using various tools (e.g., protractors, rulers, block-based and text-based programming with physical/virtual robots), help them develop their mathematical argumentation and proving skills, and help them see the utility of mathematics in everyday life or other disciplines (e.g., computer science, music, and art). The math content covered in this course includes fundamental topics in algebra, geometry, statistics, and probability (e.g., solving systems of equations, justifying area formulas, and identifying the relationship between experimental and theoretical probabilities). The first author of this paper was the instructor of this course. She taught two sections of the course in the same format. Twenty-nine and 30 students were enrolled in each section during the Fall 2023 semester. Among these students, 17 from the morning and 16 from the afternoon sections voluntarily participated in our study. Of the 33 students who agreed to participate, 29 completed the portions of the learning tasks and assignments included in our analysis.

2. Data

The data we focused on for this study included homework responses from 29 students who completed the homework after participating in the in-class ChatGPT-embedded proving activities on justifying the permutation formula. We engaged students in the in-class ChatGPT-embedded proving activities after discussing the meaning of permutation and providing examples of problems related to permutation. After this, students were assigned homework, which asked them to explore ChatGPT further as a proof assistant in a different problem context.

The format of the in-class activities and individual homework was the same except for problem statements and contexts. The in-class portion occurred during a learning activity on permutation. See Figure 1 for the in-class learning task. Before the day of the in-class ChatGPT activities, students were asked to prove a formula for permutation as homework and bring their argument to the next class. During the in-class ChatGPT activities, students were asked to pull up ChatGPT 3.5 (https://chat.openai.com/) and type in their argument in the prompt box, providing ChatGPT with what they tried to prove and asking ChatGPT to determine whether their argument looks like a proof. Based on ChatGPT feedback, students were asked to improve their arguments to be acceptable as proofs. During these activities, the instructor engaged students in a discussion about the limitations of ChatGPT and the possibility that it may sometimes hallucinate or generate erroneous information. Students evaluated ChatGPT responses and whether they agreed or disagreed with the feedback. Students also shared their proof-writing experiences with ChatGPT assistance with their classmates during a whole-class discussion.

After completing the in-class ChatGPT activities, we asked students to complete another set of proving

#1. Revisit the following statement:

> If n and r are integers with $1 \le r \le n$, then
> $${}_nP_r = \frac{n!}{(n-r)!}$$

How did you prove the statement? Do you think your argument can be acceptable as a mathematical proof? How would you evaluate your argument? What part(s) of the argument should be improved?

#2. Let's explore ChatGPT to see if we could use it as a tool to improve or revise an argument. Type in your argument from your HW in the prompt box in ChatGPT, asking whether your argument looks like a proof.

> **Example prompt**: I wrote an argument to describe why the permutation formula should be n!/(n-r)! Here's my argument: [*Type your argument*]. Do you think my argument is a proof? If not, do you have any suggestions to improve my argument?

a) Type the prompt that you used below (You may revise the example prompt above).

b) Take a screenshot of the response from ChatGPT and paste it in the box below.

c) Did ChatGPT make any suggestions that you think are wrong? What errors did ChatGPT make? Or do you think all of its suggestions were accurate? Explain.

d) Revise your argument based on feedback from ChatGPT. Write your revised argument in the given box below.

e) Type all the prompts that you used in this revision process.

f) Evaluate your revised argument. Do you think your argument is improved? Compare your original argument with the revised one.

g) Type in your revised argument (from part (d)) in the prompt box in ChatGPT, asking whether your argument can be considered a proof. Take a screenshot of the response from ChatGPT and paste it in the box below.

h) Do you agree with ChatGPT's evaluation of your argument? How did ChatGPT determine whether your argument is a proof? Was it similar or different to how you evaluated arguments? Describe what you noticed.

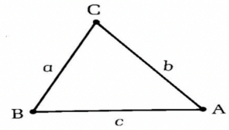
Figure 1. Task designed for the in-class ChatGPT-embedded proving activities.

activities as homework with and without ChatGPT assistance in a geometry problem context. Prior to the permutation lessons, students completed a geometry learning module in which they explored different ways to justify area formulas for 2D shapes and the Pythagorean theorem using definitions, formulas they already know, and/or moving and additivity principles. So, in class, students explored how to prove the Pythagorean theorem and the basic area formula for triangles ($^1/_2$ times *base* times *height*), considering all different types of triangles, which they are expected to possess as background knowledge for completing homework. The homework consists of two parts. The first part asked students to construct an argument describing a method to find an area of a triangle in terms of its side lengths and why that method works. Students were then asked to revise their arguments based on ChatGPT feedback as they interacted with it. After revising their arguments, they were also asked to answer a question about their experience using ChatGPT in constructing proofs. To capture students' work with ChatGPT, they were asked to include screenshots of ChatGPT responses when submitting their homework. See Figure 2 for homework instructions and prompts that we used to guide students' proving activities with ChatGPT assistance. Considering students' mathematical abilities, we first guided them to find the area of an equilateral triangle with each side length of *4 cm* as an example to ensure they understood what the problem was asking them to prove.

The second part asked students to reflect on their learning about proof and proving throughout the semester (Figure 3). In order to understand ways that students construct, evaluate, and revise arguments, as data for this study, we also included student responses to specific questions asking them to describe their views on mathematical proof and proving and their thoughts about good mathematical arguments

MathEdu

#1. Let's consider a triangle with side lengths a, b, and c. How can we calculate the area of this triangle? Find the area formula for a triangle with side lengths a, b, and c.



a) Make an argument supporting your claim (area formula) that can be accepted as a mathematical proof (*Hint*: How did you find the area of the equilateral triangle with each side measuring 4 cm?)

b) Do you think your argument is proof? Explain why or why not.

#2. Now, let's explore ChatGPT to improve or revise your argument. Type in your argument from part #1(a) in the prompt box in ChatGPT, asking whether your argument looks like a proof.

Example prompt: I wrote an argument to describe the area formula for a triangle with side lengths a, b, and c. Here's my argument: [*Type your argument*]. Do you think my argument is a proof? If not, do you have any suggestions to improve my argument?

a) Type the prompt that you used below (You may revise the example prompt).

b) Take a screenshot of the response from ChatGPT and paste it in the box below.

c) Did ChatGPT make any suggestions that you think are wrong? What errors did ChatGPT make? Or do you think all of its suggestions were accurate? Explain.

d) Revise your argument based on feedback from ChatGPT. Write your revised argument in the given box below.

e) Evaluate your argument. Do you think your argument is a mathematical proof? Explain why or why not.

f) Type in your revised argument (from part #2(d)) in the prompt box in ChatGPT, asking whether your argument looks like a proof.

g) Type the prompt that you used in this revision process.

h) Do you agree with ChatGPT's evaluation of your argument? How did ChatGPT determine whether your argument is a proof? Was it similar or different to how you evaluated arguments? Describe what you noticed.

i) How do you feel about using ChatGPT as your proof assistant to improve your argument (proof)? Reflect on your learning about proofwriting using ChatGPT.

Figure 2. Homework instructions and prompts used for guiding student proving activities with and without ChatGPT assistance.

Think about your journey of learning to prove and answer the following questions:

- What does it mean to prove something? What would you say if asked to give a definition for *proving* at this point?
- Have your conceptions of proof changed over this semester? If so, how?
- Why do we prove things in math? How important do you think proving is in mathematics?
- What makes a good mathematical argument that can be accepted as a mathematical proof?
- Do you think the class activities you engaged in helped you build your ability to construct mathematical arguments (proofs)? What activities were helpful for you to develop your proving skills? Share your experiences.
- One of the activities you did in class was programming to draw shapes. Using the Venn diagram, visualize the relationship between *proving* and *programming*. How are these two activities, proving and programming, similar or different? Clearly describe your current thoughts using the Venn Diagram you created.
- How do you feel about learning to prove using programming?

Figure 3. Prompts used for student reflection on their learning about proof and proving.

that can be considered proofs. Note that this assignment came near the end of the Mathematics Pathways course. Throughout the semester, students were exposed to thinking about the meanings of proof and its roles in mathematics as they participated in various proving activities involving constructing, evaluating, and revising arguments in algebra, geometry, and probability problem contexts, reading de Villiers's (1990) article illustrating different roles of proof in mathematics, watching two videos showing Andrew Wiles' journey of proving Fermat's Last Theorem, and writing a group discussion paper about proof. As part of in-class proving activities, students also created a proof evaluation rubric with their group members and evaluated given arguments using their rubric. After completing the proof evaluation work with their respective groups, students participated in a whole-class discussion. They compared the rubrics made by each group and discussed what constitutes a mathematically valid proof. Students also constructed arguments as they designed (block-based) programs to justify why they thought programs would run to draw target shapes (e.g., isosceles and acute triangles) with robots. After the drawing activities with robot programming, students were also asked to investigate how the process of programming and proving are connected. To capture their understanding of mathematical proof and proving at the end of the semester after completing all of these proving-related activities, we asked them to complete and submit their written reflections on proof based on their learning experience in the whole course, which was the second part of their individual ChatGPT-embedded proving activities homework assignment.

3. Data Analysis

Using open coding (Strauss & Corbin, 1990), we analyzed students' homework responses involving their original and revised arguments that they constructed with and without ChatGPT assistance, their evaluations of their original and revised arguments, and ChatGPT's feedback concerning proof, and their written reflections describing their perspectives on proof and proving, and on good mathematical arguments that can be viewed as proofs, and also their thoughts about the use of ChatGPT in proof construction. To explore what improvements students made when revising their original arguments based on ChatGPT feedback, we first analyzed and coded ChatGPT suggestions students had received while interacting with it in the revision process. We analyzed screenshots of the ChatGPT feedback students shared through their homework submissions. Our data analysis identified six themes across all the ChatGPT suggestions provided to students, including:

- **Clarity:** Suggesting students define symbols or mathematical terms being used or suggesting students clearly state what they are describing or what theorem is being used
- **Justification:** Suggesting students give reasons for statements in their proofs
- **Generalization:** Suggesting that students make statements that apply to all triangles
- **Error corrections:** Catching mistakes or errors that students may have made and suggesting how to fix them
- **Claim revisions:** Suggesting that students make a change to the claim
- **Questionable advice:** Suggesting inappropriate approaches that contain errors

Then, we analyze students' original and revised arguments, focusing on the parts in which they improved their original arguments using these six themes as codes, opening up new codes describing characteristics of students' revised arguments compared to their original arguments.

After analyzing their arguments and ChatGPT feedback, we examined their responses to evaluations of their original and revised arguments and of ChatGPT feedback on their arguments, their thoughts about the use of ChatGPT in generating arguments, and their perspectives on proof and proving and on good mathematical arguments that can be considered proofs. We first coded how they described proof and proving. We then coded their evaluations of their arguments and ChatGPT feedback, focusing on what features of mathematical proof they considered when determining whether their arguments were proofs and whether they agreed or disagreed with ChatGPT's feedback on improving their arguments to make them more proof-like. After coding these responses, we explored how their evaluations aligned with their perspective on proof and proving. Lastly, we coded students' responses about using ChatGPT as a proof assistance.

In the data analysis stage, to develop those codes, the authors of the paper first carefully read students' homework responses line-by-line and individually coded the responses using Excel spreadsheets after transferring student responses to Excel. During our weekly research meetings, to ensure the trustworthiness of our codes, we then shared the coding processes and compared codes each identified to ensure we coded them through the same line of interpretations. When our interpretations of student responses differed, we discussed them until we reached a consensus and re-coded some parts of the data based on definitions of the codes we made. After this initial coding process, we looked through all the codes made together, discussed what the codes told us, identified themes across the codes, and wrote analytic memos describing the characteristics of students' argument constructions, revisions, evaluations, characteristics of ChatGPT's feedback on students' arguments, students' views on proof and proving, and their thoughts about the use of ChatGPT in proof writing.

Results

Most students evaluated ChatGPT as an effective tool that can improve arguments based on their proof-writing experience with ChatGPT. They thought that ChatGPT evaluated their arguments similarly to how they usually evaluate arguments of themselves or others, which aligned with their perspectives on proof and proving. So, for the most part, they agreed with ChatGPT's evaluations of their arguments and revised parts of their original arguments as suggested by ChatGPT, focusing on improving clarity, providing additional justifications, and showing the generality of their arguments across all cases. However, when ChatGPT's suggestions did not make sense, they stuck with their original arguments without making any changes. Some students expressed that ChatGPT responses sometimes confused them, so their suggestions were not always helpful in the revision process of their arguments. In this results section, we will report our results in the following order: first, we will describe how they viewed proof and proving. we will then demonstrate how students constructed and evaluated their original arguments that they produced to show how to find the area of a triangle given its side lengths and why that method works, and whether their approaches to proof construction and evaluation align with their perspectives on proof and proving. We will then illustrate the types of ChatGPT feedback students received on their original arguments and describe the characteristics of the revised arguments students rewrote based on the ChatGPT feedback on their original arguments. Students' thoughts about using ChatGPT as a proof assistant will be shared at the end of this section.

1. RQ1: Students' Perspectives on Mathematical Proof and Proving

Of the 29 student participants, 28 described proving as explaining why a mathematical statement (claim) is true. For instance, when prompted to demonstrate what proving means to her, Amory[1)] wrote, "It means to come up with a thorough explanation for why something does something. I would find an explanation to prove why something is the way." Along the same line of thought, she described that a mathematical argument that can be accepted as a proof should involve "a thorough explanation that goes the steps and explanations [that are] correct and make sense." Describing such a view, Steven and Halley particularly emphasized the general aspect of mathematical proof. Halley stated that in proving, one should "provide concrete evidence about a claim" and "ensure it is applicable in any given situation." Steven also addressed that proving means showing that "It can be replicated with any relevant set of numbers to achieve an accurate result." Unlike other students, one student, Alma, exceptionally defined proving as verifying the truth of a statement, but she also mentioned that a mathematical proof should be logically constructed based on "axioms, definitions, and previously established theorems." So, for our participants, a justification(s) was an essential component that should be contained in a proof.

2. RQ2: Characteristics of Students' Original Argument Constructions and Evaluations

Based on our data analysis of students' original arguments about finding the area of a triangle using its side lengths, which they constructed by themselves before attempting to revise their arguments with ChatGPT assistance, and their responses about whether their arguments were proofs, we found that students' views on proof and proving influenced both their argument constructions and evaluations. Most students constructed their arguments by describing a method they came up with to find the area of a triangle, given its side lengths, step by step. Although their rationales for why they saw their arguments as proofs were slightly different, most of our student participants addressed that since they provided explanations of how and why their methods would work to find the area of any triangle in terms of three side lengths, they determined their arguments could be accepted as proofs (21 out of 29 students). Six students also mentioned reasons for using concrete examples in their arguments to provide clear explanations about their thinking about their claims about the method of finding the area of a triangle. Five other students commented that their arguments were mathematical proofs because they made generalizations to prove why their methods worked for all triangles. Some students evaluated their arguments as incomplete proofs because they did not thoroughly explain why. See Catherine's original argument below as an example:

> We can find the area of a triangle using the formula $\frac{1}{2}bh$. However, since the triangle that is given is not a right triangle, we must find the height. We can do this by drawing a line segment from base a to the hypotenuse. From here, we are doing the additive principle which means we are breaking down the shape into two different right triangles. Since b and c are congruent sides and we have created two right triangles, we can then perform the Pythagorean theorem, which states $a^2+b^2=c^2$. To make this easier, I provided an example to understand how this process works. Consider that b and c equal *12* and a equals 4. Once we draw a line from the hypotenuse, side length a has two different lengths that equal *2*. Then we can use the theorem, $2^2+12^2=h^2$. Therefore $4+144=h^2$. Then, we square root *148* to find h, which equals *12.17*. Since we have found the height, we can then put this back into the area formula, which is $\frac{1}{2}bh$. Therefore, $\frac{1}{2}(4)(12.17)=24.3$.

Catherine evaluated her original argument as a proof, indicating that "I am showing how each step occurs and why. I have used logic and reasoning to prove the area. From there, I also use an example to show how this process works." When evaluating, she focused on the extent of justification she provided about why and how her method should work, not catching errors she made in finding the height of a triangle using the Pythagorean theorem. Her way of constructing and evaluating the argument is aligned with her perspective on proof and proving. Catherine described that for her, proving means "showing something is true through logic and reasoning. You break down a concept step by step to show how each step contributes towards the product." She also elaborated that:

> A good mathematical argument [proof] involves breaking down each step of a process and explaining how it is true. You also need to make sure to define and show how mathematical terminology is used and how it works. I think that is also helps to add an example so that the audience can see how it works.

3. RQ3: Types of ChatGPT Feedback on Original Student Arguments

In this section, we describe the types of feedback that students received from ChatGPT when they asked

ChatGPT to evaluate whether their original arguments, which they constructed without help from others (including ChatGPT assistance), are mathematical proofs. They made arguments describing a method for finding the area of a triangle in terms of its three side lengths. ChatGPT gave students six different kinds of feedback to help improve their arguments: clarity, justification of steps, generalization, error corrections, claim revisions, and questionable advice. Figure 4 shows how often each of these argument improvement suggestions appeared. Note that some feedback exhibited multiple suggestions, so the total is greater than the number of students participating in the study.

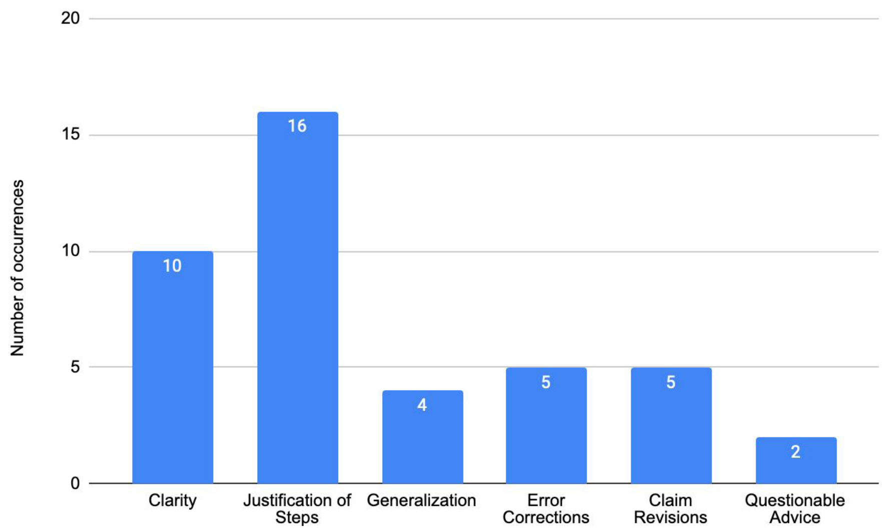


Figure 4. Types of proof feedback given by ChatGPT.

Oftentimes, feedback that ChatGPT provided students centered on clarity, emphasizing forms of formal proof and the importance of readability of proof that can help communicate ideas clearly to proof readers. These ChatGPT responses mainly dealt with suggestions for defining mathematical variables that students used for unknown quantities (e.g., unknown side lengths of a triangle) in their arguments. For instance, when Kyle asked ChatGPT whether her argument looked like a proof (see Kyle's argument below):

> Find the height of the triangle by creating two right triangles. Find the height by using the Pythagorean Theorem. Use the area formula of $\frac{1}{2}\times b\times h$ to find the area of both right triangles. Add both of the areas to find the total area of the scalene triangle. This finds the solution because it takes into account the unknown of the sides of the whole triangle by splitting it up into two and adding them together to make the whole triangle again.

ChatGPT suggested improving her argument for clarity, pointing out, "At the beginning, explicitly state that *a, b*, and *c* are the side lengths of the triangle and *h* is the height." It also commented that "clearly state that you are using the Pythagorean Theorem to find the height of the triangle. For example, by applying the Pythagorean Theorem to the right triangles." Shelby also received similar feedback from ChatGPT on her argument:

> The area formula for a triangle is *(1/2)(base)(height)=area*. The triangle above gives us side lengths of *a, b*, and *c*. While any of these side lengths could be the base, I am going to call *c* the side base. With *c* as the base, the height would be a line perpendicular to the base that runs up to angle *C*. Then, using the Pythagorean theorem, the height of the triangle can be found. $a^2+b^2=c^2$, with the height being the variable that you are solving for. That value can then be plugged into the height in the area formula.

Highlighting the importance of making arguments clear for proof readers, ChatGPT provided suggestions that

> define the variables you are using (*a, b, c, A*, *base, height*) before diving into the explanation. *This helps the reader follow your argument more easily*. When you mention 'angle *C*,' explicitly state that it is the angle opposite side *c.* This helps to avoid confusion, especially if your audience is not familiar with the standard notation.

Other suggestions included specifying what side of a triangle students picked as a base and how students identified the height for the selected base.

Another area where student-written arguments were lacking, and ChatGPT provided feedback was the area of justification, encouraging students to give reasons for their statements explicitly. For example, when Adrian asked ChatGPT whether her argument "I may get the area of the triangle using the formula (*a*d*)/*2* by drawing a line that splits the triangles evenly in half and identifying that line as "*d*," where *d* is the triangle's height," is a mathematical proof, not only ChatGPT pointed out that Adrian should "clearly define the terms" she is using in her argument, but also it suggested to provide "a brief justification for why this formula [(*a*d*)/*2*] works" to find the area of a triangle to make her argument "complete." Similarly, for Amory's argument,

> To find the area of the triangle, multiply the base by the height using the area formula. Since sides *b* and *c* are equal, indicating congruence, set the new base length to *6*. Determine the height using the Pythagorean theorem, take the square root of the height, and finally, calculate the triangle's area using the relevant formula.

ChatGPT suggested "adding more explicit steps and ensuring a clear, logical flow" to make a more "solid," "convincing," and "rigorous" argument. ChatGPT presented additional details, such as how to use the Pythagorean Theorem to find the height of the given triangle that Amory should consider in revising her original argument.

In some cases where students attempted to find a triangle area using specific numbers, ChatGPT tried to direct them to generalize their arguments for all triangles. Sometimes, it did not explicitly point out that specific examples are not general proofs. However, it included statements like "In your Pythagorean theorem calculation, it would be clearer to use variables instead of specific numbers" (ChatGPT response given to Hope's argument) to encourage generalization for improving example-based arguments that students produced towards general arguments. ChatGPT also provided suggestions encouraging students to write their arguments in more general ways. For instance, for Adrian's argument described above, where she did not consider all different types of triangles, ChatGPT commented, "If applicable, mention that this formula [$\frac{1}{2}$× base×height] is not limited to a specific type of triangle (e.g., right-angled triangle) and can be applied more broadly." Scout also received similar ChatGPT feedback on his argument. It said, "Your argument seems to heavily rely on the right-angled triangle. The area formula applies to all triangles, regardless of whether they're acute, obtuse, or right-angles."

Another type of ChatGPT feedback dealt with identifying and correcting errors students made in their arguments. Most students made errors by making incorrect assumptions (e.g., assuming a given shape is one type of triangle such as isosceles, equilateral, or scalene triangle) or misapplying a theorem/formula they already know. In cases of incorrect assumptions, ChatGPT usually went along with the assumption students made and provided suggestions that fit within the same assumptions. However, ChatGPT was able to catch some errors in their applications of formulas. As an example, Caroline received the following suggestion from ChatGPT, which identified a typo she made in her argument:

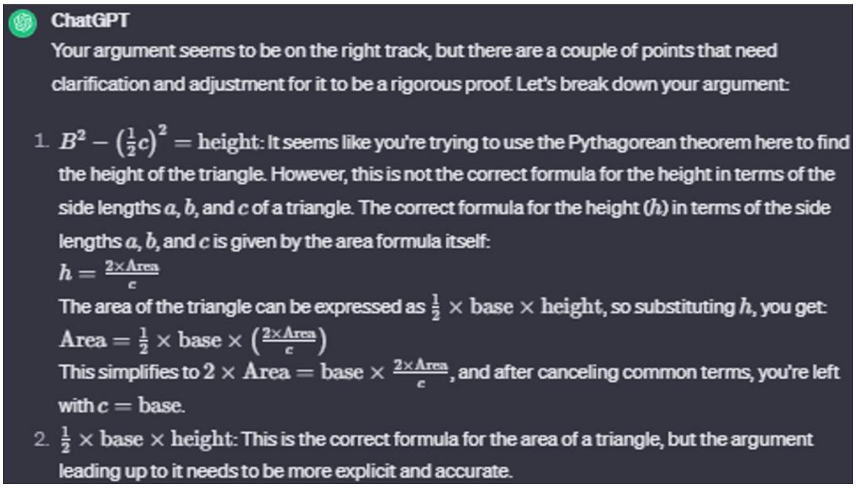
> The Pythagorean theorem typically states as $c^2=a^2+b^2$, where *c* is the hypotenuse and *a* and *b* are the legs of a right triangle. So, your formula, $A^2+B^2=H$, might be a typo or misrepresentation of the Pythagorean theorem.

In this ChatGPT-embedded proving assignment, we did not explicitly tell students that they could find the area of a triangle using its side lengths by applying an existing formula known as Heron's formula. However, ChatGPT sometimes understood the goal of the task well enough based on the prompt that students provided when interacting with ChatGPT and suggested using Heron's formula to figure out the task. In most of these cases, it suggested reformulating their claim as Heron's formula. For example, in her argument, Halley described why a standard formula for the area of a triangle works using the area formula for a rectangle. See below for Halley's argument:

> A rectangle which has been divided into two makes two congruent triangles. To find the area of the rectangle, you multiply the base by the height, and since the triangles, specifically the right-angle triangle is formed from the rectangle and is half of the shape [rectangle], finding its area means finding half of the triangle, hence the formula will be $^1/_2 \times b \times h$.

ChatGPT called attention to the fact that her explanation "lacks the connection to the side lengths *a, b*, and *c* of the triangle, and it does not directly derive the formula for the area of a triangle using these side lengths" and proposed to "provide a more comprehensive proof for the formula of the area of a triangle using side lengths *a, b*, and *c*" introducing Heron's formula, which can calculate the area of a triangle given its side lengths, $A = \sqrt{s(s-a)(s-b)(s-c)}$, where *s* is the semi-perimeter of the triangle ($s = \frac{a+b+c}{2}$). Four other students received feedback similar to this.

The last type of ChatGPT feedback involves undesirable and questionable advice, which can either trivialize the problem or confuse the student. This undesirable feedback comes from the fact that large language models like ChatGPT are prone to hallucinate or generate nonsensical or non-factual content. Our data analysis shows that this occurred in a few cases. Look at Figure 5 presenting the ChatGPT feedback that Steven received on his original argument "$B^2 - \left(\frac{1}{2} \cdot c\right)^2$ = height, then, $\frac{1}{2}$ • base • height" as an example. Although it identified an error Steven made in finding the height for a triangle using the Pythagorean theorem, it suggested a revision with circular reasoning to prove why a standard area formula for a triangle works, which was not a goal of this task.



ChatGPT
Your argument seems to be on the right track, but there are a couple of points that need clarification and adjustment for it to be a rigorous proof. Let's break down your argument:

1. $B^2 - \left(\frac{1}{2}c\right)^2 = \text{height}$: It seems like you're trying to use the Pythagorean theorem here to find the height of the triangle. However, this is not the correct formula for the height in terms of the side lengths a, b, and c of a triangle. The correct formula for the height (h) in terms of the side lengths a, b, and c is given by the area formula itself:
$h = \frac{2 \times \text{Area}}{c}$
The area of the triangle can be expressed as $\frac{1}{2} \times \text{base} \times \text{height}$, so substituting h, you get:
$\text{Area} = \frac{1}{2} \times \text{base} \times \left(\frac{2 \times \text{Area}}{c}\right)$
This simplifies to $2 \times \text{Area} = \text{base} \times \frac{2 \times \text{Area}}{c}$, and after canceling common terms, you're left with $c = \text{base}$.
2. $\frac{1}{2} \times \text{base} \times \text{height}$: This is the correct formula for the area of a triangle, but the argument leading up to it needs to be more explicit and accurate.

Figure 5. ChatGPT feedback on Steven's argument.

4. RQ4: Students' Use of ChatGPT Feedback in Revising Their Original Arguments

Twenty students (out of 29) improved their original arguments, while six made no substantial change to their arguments. Three students changed their arguments, but the result was no better than the original. In this section, we describe what improvement(s) students made in revising their original arguments to make them more proof-like. Note that most students (25 out of 29) thought their original arguments were proofs but could be improved to make them more formal proof-like. Two other students, Jerry and Alba, did not think of their original arguments as proofs because of the lack of information, such as numbers given in the problem task, even though their arguments felt somewhat sound. They thought that specific numbers should be given for side lengths or angles to find the area of a triangle. There were no answers from the remaining two students about their determinations of whether their original arguments were proofs.

Most students attempted to revise their original arguments to make them more proof-like, considering ChatGPT feedback by improving clarity, justifying each step, proving generalizations, correcting errors, or revising their claims (Figure 6). They did so because ChatGPT's suggestions made sense to them and aligned with their perspectives on proof and proving. Their approach to proving is similar to the syntactic proof production style (Weber & Alcock, 2004) in that they tried to make arguments using the theorem or formula they already knew, not relying on a few examples.

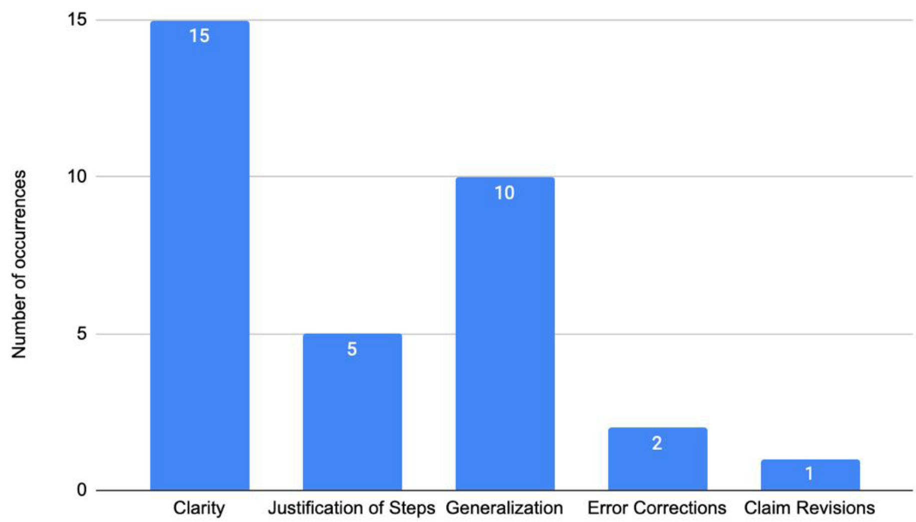


Figure 6. Types of improvements in students' revised arguments.

Of 29 students, 15 improved their original arguments for clarity by clearly describing the symbols they used. For example, Kyle began her original argument with the following sentences, "Find the height of the triangle by creating two right triangles. Find the height by using the Pythagorean Theorem. Use the area formula of $\frac{1}{2} \times b \times h$ to find the area of both right triangles." Note that variables b and h are used in Kyle's original argument without defining what each variable represents in this problem context. As described in the previous section, suggestions Kyle received from ChatGPT regarding improving this intro part of her argument were, "At the beginning, explicitly state that a, b, and c are the side lengths of the triangle and h is the height." Kyle thought its suggestions were "correct as I did not define the variables well in my first argument, and I could be more specific in what the steps are to find the answer." Kyle then revised her argument's intro: "Find the height of the triangle by creating two right triangles and label the altitude [height] h. Label the three sides a, b, and c." Another example of students' argument improvement in clarity came from Alfie, who originally used inconsistent mathematical notation in his argument:

> I think we can still apply the *base×height / 2* formula here with a few notes to fully explain the process. Base=any length side (a, b, c). Height would now=a line perpendicular to the base running from the base to the opposite point, e.g., $a \rightarrow A$, $b \rightarrow B$, and c=C. You can now work out [the] area from this information

using the *base×height* / *2*. Therefore, for example, the area could be $c \times c \rightarrow C$ / 2.

Accepting ChatGPT suggestions on his original argument for more precise language and notation use (see Figure 7 for the feedback that Alfie received from ChatGPT), although he thought his original argument was a mathematical proof, Alfie rewrote it as follows to enhance clarity:

> I propose using the formula for the area of a triangle, which is given by the expression $^1/_2$ **base*height*. Let's consider side '*c*' as the base of the triangle with its opposite vertex '*C*.' The height, denoted as '*h*,' is defined as the perpendicular distance from side '*c*' to Vertex '*C*'. To find the area, we apply the formula: $area={}^1/_2*c*h$.

1. **Use more precise language:**
   Instead of saying "any length side (a, b, c)," specify that the base is the side opposite the height you're considering. For example, you could say, "Let's consider side a as the base, and the height will be the perpendicular from A to side a."
2. **Be specific about the perpendicular distance:**
   State explicitly that the height is the perpendicular distance from the vertex to the line containing the base. You can mention that this distance is measured along the line containing the base.
3. **Avoid using the same symbol for different things:**
   In your example, you mentioned "c x c -> C / 2," which might be confusing because it uses the same symbol 'c' for the side length and the perpendicular distance. Consider using a different symbol for the height, perhaps 'h,' to avoid confusion.
4. **Use proper mathematical notation:**
   Instead of writing "base x height / 2," you might want to use mathematical symbols and notation. You can write it as "1/2 * base * height."

Figure 7. ChatGPT's feedback on Alfie's original argument.

Five students included additional justification(s) to their original arguments while revising them. These students viewed their original arguments as mathematical proofs because they felt they provided good enough explanations using valid reasoning in their arguments. However, they agreed with ChatGPT's suggestions to add more justifications to make their original arguments more complete. One example is from Billy, whose first-draft argument was,

> To find the area of this triangle, you will need to split the triangle into *2* at the altitude [height], and that line is your height. Then, you need to find the height of the triangle, you can do this by using a side length and an angle. For example, $b \times \cos A$. Then we can just plug that back into the original formula and get $a=\frac{1}{2} \times a \times b \times \cos A$.

In the revision, Billy explicitly invoked the definition of cosine to justify his height calculation for the triangle, writing, "[a] side *b* and [an] angle *A*, according to the definition of cosine, $b \times \cos A$ represents the length of the adjacent side in a right triangle formed by the altitude [height]. This equation shows the height of the triangle." Regarding ChatGPT suggestions about adding more details, Billy felt "it is more knit picky stuff," but he evaluated that his revised argument is "definitely better" than his original one. Another example comes from Kelly, whose original argument was:

> Pythagorean's theorem can be used to find the missing height of a scalene triangle to then find the area. To start, draw a perpendicular line from one vertex to the base, creating two triangles with a right angle. You can then use Pythagorean's theorem, $A^2+B^2=H$, to find the height. This can then be plugged into the

triangle area formula $\frac{1}{2}$×b×h to solve for the area.

She then revised this to:

> To find the [area] formula of a scalene triangle, the side lengths *a, b*, and *c* can be used to find the height by using Pythagorean's Theorem. To start, draw a perpendicular line from one vertex to the opposite side, creating two right triangles. *A* and *B* are the two opposing sides, whereas *C* is where the base length where the perpendicular line is dropped on to. Pythagorean's theorem can then be used with *A* and *B* representing the segments of side *C* divided by the perpendicular line. The length found using the theorem is then the height (*H*) of the triangle. After finding the height using Pythagorean's Theorem, the height can be applied to the formula for the area of a triangle $\frac{1}{2}$*b*h. Then, calculate using the previously found variables to find the area of the scalene triangle.

1. **Establishing the approach:**
   Your initial statement about using Pythagoras' theorem to find the missing height of a scalene triangle and then using the area formula is a valid approach.
2. **Describing the process:**
   - Start by drawing a perpendicular line from one vertex to the opposite side, creating two right triangles.
   - You can denote the sides of the triangle as a, b, and c, where 'c' is the side on which the perpendicular is dropped.
   - Use Pythagoras' theorem to find the length of the perpendicular (height).
     - The theorem states: $a^2 + b^2 = c^2$, where 'a' and 'b' are the legs of the right triangle and 'c' is the hypotenuse.
     - In this case, 'a' and 'b' represent the segments of the side 'c' divided by the perpendicular line.
     - The length found using Pythagoras' theorem is the height (h) of the triangle.
3. **Applying the area formula:**
   - Once you've found the height using Pythagoras' theorem, apply it to the formula for the area of a triangle: $\frac{1}{2} \times \mathrm{base} \times \mathrm{height}$.
   - Plug in the known values for the base and height (which you found using Pythagoras' theorem).
   - Calculate to find the area of the scalene triangle.

To improve the argument:

- Clarify that Pythagoras' theorem is being applied to one of the right-angled triangles formed by dropping the perpendicular, and the hypotenuse 'c' in this case represents the side of the triangle on which the perpendicular is dropped.
- Explicitly state that the obtained length from Pythagoras' theorem is the height of the triangle.
- Ensure the steps are clearly explained and connected, emphasizing how the calculated height is used to find the area using the triangle area formula.

By refining and clearly delineating each step in the process, you can provide a more comprehensive and structured argument.

Figure 8. ChatGPT's suggestions on Kelly's original argument.

Kelly viewed both her original and revised arguments as proofs. However, she evaluated that the revised one was "much more clarifying than previously" since she broke down "the steps more thoroughly, making it easier to comprehend and understand how to solve." Look at Figure 8 above for ChatGPT's suggestions Kelly received regarding her original argument. Kelly felt that "overall, ChatGPT helped guide me in making a stronger proof, and showing me what parts needed more clarifying and, overall, how to strengthen my argument and make it proof."

In ten cases, we observed that students attempted to improve their original arguments to make them more general, showing the generality of their arguments across all triangles. Initially, they thought their original arguments were proofs without recognizing the lack of generality in their arguments until ChatGPT pointed it out. Students receiving ChatGPT's feedback on the lack of their original arguments in terms of generalization added wording like "This method is applicable to any triangle" (George) or "This makes it repeatable and a generalized formula" (Billy) in their revised arguments. The following argument is Scout's first draft of the argument, including the diagram showing the dotted line segment Scout drew to indicate the height of a triangle for the chosen base (see Figure 9):

> The area formula for triangles is $base \times height \times \frac{1}{2}$. In order to find the area, you must use the Pythagorean theorem to find the hypotenuse. From there, you must divide the base by 2 and use the Pythagorean theorem again to find the height (dotted line on the diagram below). Once you have both the base value (c) and height, you must follow the area formula, $base \times height \times \frac{1}{2}$.

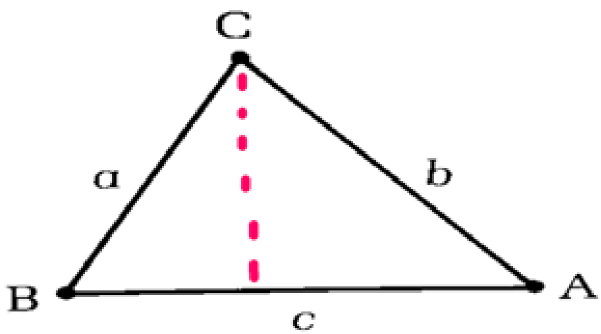
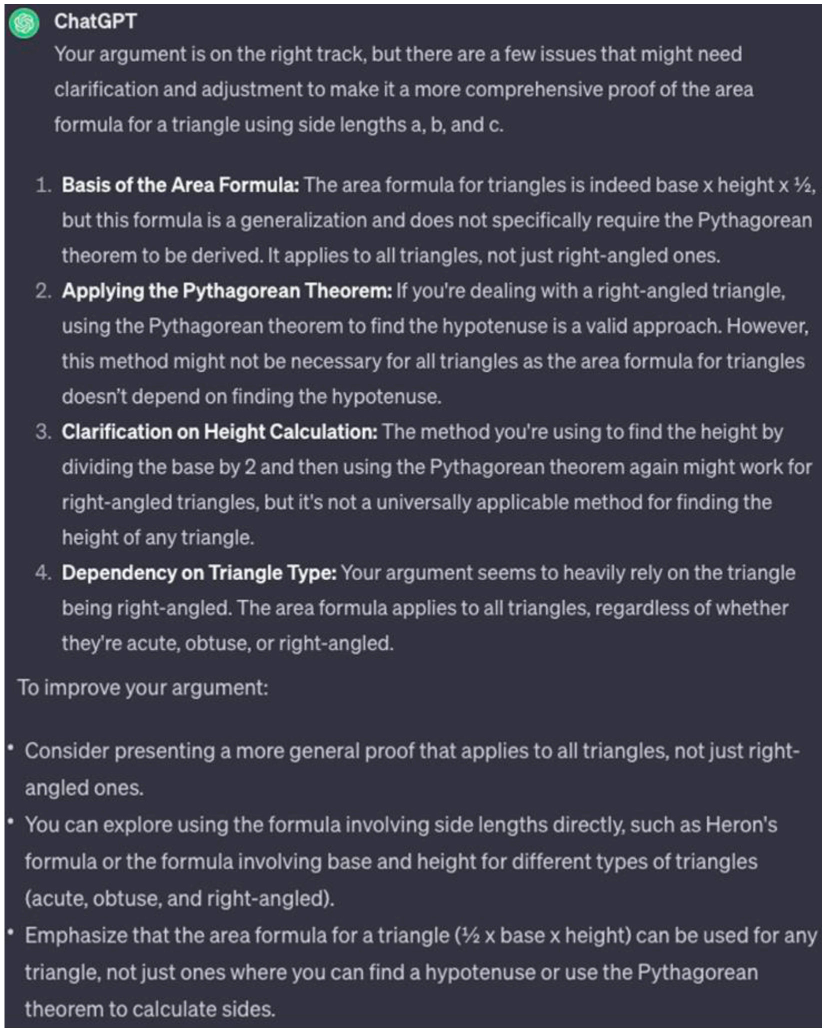


Figure 9. Scout's drawing of the dotted line segment indicating the height of a triangle.

In the revision, he focused on describing how his method of finding an area of a triangle using its side lengths would work for any triangle, highlighting that the standard area formula (i.e., $\frac{1}{2} \times base \times height$) can be applicable for all different types of triangles to find areas of those shapes:

> The area of a triangle [is] given by the formula: $area = \frac{1}{2} \times base \times height$. This formula applies universally to all triangles, regardless of their type (acute, obtuse, or right-angled). To find the height of a triangle, several methods are employed depending on the information given. 1. When a perpendicular from a vertex to the base is known, the length of this perpendicular directly represents the height. 2. In cases where side lengths are given but the perpendicular is not known, trigonometry or other geometric relationships can be used to find the height. Once the base and height are found, the area of the triangle can be found using the formula $base \times height \times \frac{1}{2}$. Thus, the area formula $area = \frac{1}{2} \times base \times height$ is a comprehensive method to determine the area of a triangle.

Figure 10 below includes the suggestions that Scout received from ChatGPT for the improvement of his original argument.


ChatGPT

Your argument is on the right track, but there are a few issues that might need clarification and adjustment to make it a more comprehensive proof of the area formula for a triangle using side lengths a, b, and c.

1. **Basis of the Area Formula:** The area formula for triangles is indeed base x height x ½, but this formula is a generalization and does not specifically require the Pythagorean theorem to be derived. It applies to all triangles, not just right-angled ones.
2. **Applying the Pythagorean Theorem:** If you're dealing with a right-angled triangle, using the Pythagorean theorem to find the hypotenuse is a valid approach. However, this method might not be necessary for all triangles as the area formula for triangles doesn't depend on finding the hypotenuse.
3. **Clarification on Height Calculation:** The method you're using to find the height by dividing the base by 2 and then using the Pythagorean theorem again might work for right-angled triangles, but it's not a universally applicable method for finding the height of any triangle.
4. **Dependency on Triangle Type:** Your argument seems to heavily rely on the triangle being right-angled. The area formula applies to all triangles, regardless of whether they're acute, obtuse, or right-angled.

To improve your argument:

- Consider presenting a more general proof that applies to all triangles, not just right-angled ones.
- You can explore using the formula involving side lengths directly, such as Heron's formula or the formula involving base and height for different types of triangles (acute, obtuse, and right-angled).
- Emphasize that the area formula for a triangle (½ x base x height) can be used for any triangle, not just ones where you can find a hypotenuse or use the Pythagorean theorem to calculate sides.

Figure 10. ChatGPT's suggestions on Scout's original argument.

Two students made slight changes in their original arguments by correcting errors ChatGPT had their attention to while interacting with it in their argument revision process. For instance, as discussed in the previous section about types of proof feedback given by ChatGPT, ChatGPT was able to catch a typo Caroline made in writing the Pythagorean theorem equation. Caroline wrote the equation correctly in her revised argument. When asked to evaluate ChatGPT's feedback in terms of correctness, she addressed that ChatGPT did a good job catching the error(s) she made in her homework responses. Only one student, Clay, changed his claim as suggested by ChatGPT. His original claim was that the area formula for a triangle in terms of its side lengths is "$\frac{1}{2}ab$." ChatGPT indicated the problem with this claim that it is "not a correct formula for finding the area of a triangle with side lengths *a, b,* and *c*" and introduced Heron's formula as the "correct" formula to use in this problem situation. Based on ChatGPT, Clay changed his clam to "$Area=\sqrt{s\cdot(s-a)\cdot(s-b)\cdot(s-c)}$" without providing any justifications to support his new claim. He evaluated that his revised argument was "kind of a proof but not complete." He thought that to be complete, he should have to "show more steps" in this argument.

5. RQ5: Students' Thoughts about Using ChatGPT as a Proof Assistant

Students' reactions to using ChatGPT as a proof assistant to improve their arguments were mostly positive, with 17 of 29 students saying only positive things about their proof writing experiences with ChatGPT (see Figure 11). Among 17 students, six of them remarked on the ability of ChatGPT to help them improve their proof-writing clarity. Alba noted that "it [ChatGPT] cleared my proofs up and made them more mathematically centered and direct regarding what I am proving." Kelly also mentioned, "It helps provide more guidance and clear up what I was trying to stay." Four students mentioned ChatGPT was a useful stand-in for another person's perspective. For example, Morgan mentioned, "If you don't have someone very knowledgeable about math around, you can use ChatGPT and get some good feedback on your proof." Carmel noted it "is a great replacement for a writing partner or similar coach or peer." Montana also remarked that it is a good tool because it allows her to "get third-person feedback" on her writing. Eight students commented that ChatGPT improved their arguments by identifying and correcting errors or pointing out things that were otherwise missing from their arguments. Bill wrote, "It shines [a] light on ideas and flaws in my proofs that I would have never found on my own." George also mentioned:

> I really like the use of ChatGPT as my proof assistant. I love the depth of feedback it gives, and its brutal, honestly. If my proof is way off the mark, it will let me know. It also shows multiple areas on how to refine and improve my arguments. Overall, it helps me see where I'm lacking in my proofs and how to fix them. I'm sure if I used this program constantly, my proof-writing skills would increase drastically.

As seen in Figure 11, six other students were mixed, saying some positive things but with caution about some aspects of ChatGPT, while six students had a primarily negative reaction. Among the negative aspects of their interaction with ChatGPT as a proof assistant, one common complaint (4 out of 12 students) was difficulty communicating with the language model. Students admitted that they had a hard time understanding ChatGPT's suggestions: "It was harder for me to learn it like that because it wasn't clear all the time" (Adrian) and "I think my proof-writing has gotten better, but learning from ChatGPT can be difficult if it doesn't explain it well" (Alma). Others felt that ChatGPT did not understand what they meant: "Sometimes the things that we are trying to communicate do not necessarily come across clearly to ChatGPT, which can make you confused" (Catherine). The other negative comments focused mainly on limitations in the language model's factual reliability and the students' inability to verify the content it generated (6 students). In one example, Alfie says, "I think it has definitely made my work more concise, but I'd definitely want a mathematical professional to validate the claims made by ChatGPT." Scout also did not trust his ability to know when ChatGPT was wrong:

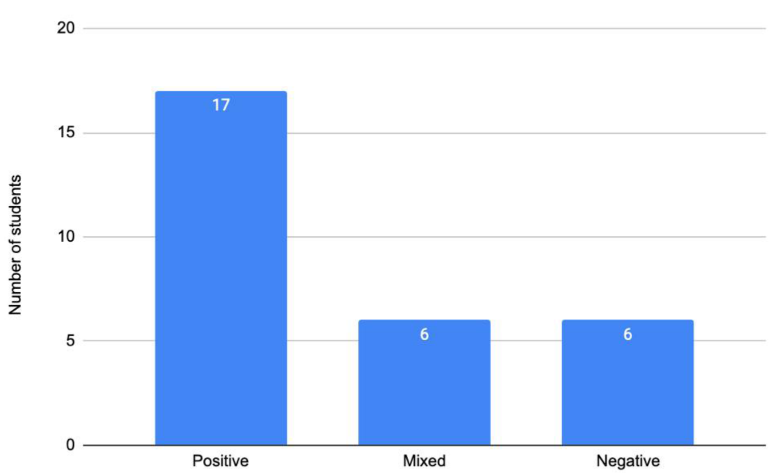


Figure 11. Students' thoughts on using ChatGPT as a proof assistant.

"I think, in math, it is useful to other people, but I have a limited knowledge of formulas and vernacular that would be crucial for me to find any inaccuracies and skips of logic."

Conclusion and Future Directions

Our study illustrates six different kinds of feedback ChatGPT provided students about students' self-constructed arguments. In most cases, ChatGPT's feedback started with praise, such as "Your argument is a good start," to encourage and list parts students should consider revising to improve their arguments. Many of our student participants found ChatGPT's feedback helpful in improving arguments, making them clearer, more complete, and more general. Most students, except a couple, saw their original arguments as mathematical proofs, believing that they provided enough justifications for the claims they made in their arguments, which aligned with their perspectives on proof and proving. When asked to improve their arguments using ChatGPT's suggestions, they attempted to make their original arguments more proof-like, such as by clearly defining terms and stating theorems used, adding more justifications to statements they made in their arguments, showing the generality of their arguments across all triangles, and correcting errors. Students saw their revised arguments as proofs but refined ones. Excluding only a few students who produced arguments using specific examples, most students tried to describe and verify their methods of finding the area of a triangle in terms of its three side lengths using the Pythagorean theorem and the basic triangle area formula they already know. Their revised arguments were still incomplete and/or invalid based on our evaluations. However, the way our student participants produced their arguments for proving was similar to the syntactic approach (Weber & Alcock, 2004).

While interacting with ChatGPT, students got quick feedback from ChatGPT, which helped them easily identify what part(s) of their arguments should be improved. For instance, as Chazan (1993) described in his paper, some of our participants also saw their original arguments that they had constructed only considering one type of triangle as proofs. They did so by relying entirely on a shape pictured in the diagram. Not recognizing the general aspect of a given shape, they simply interpreted it as one type of triangle, such as a scalene triangle, based on how they could see it. Thus, they made their original arguments by considering one type of triangle. However, when ChatGPT pointed out that they should consider all types of triangles to make valid proofs, students caught the limitations of their arguments. So, in their revision processes with ChatGPT's assistant, students tried to revise their arguments considering all types to make them general arguments. ChatGPT also helped students correct their errors (e.g., inappropriately applying the Pythagorean theorem in finding the height of a triangle for the chosen base). Overall, students thought that ChatGPT's line-by-line evaluations of their arguments were similar to the way that they usually evaluate the arguments of others and themselves. They also agreed with ChatGPT's suggestions to make their arguments more proof-like for the most part since what it suggested was also aligned with their perspectives on proof and proving. However, when students encountered ChatGPT's nonsensical or difficult-to-understand suggestions during the argument revision process, they chose to ignore the suggestions. Like previous studies pointed out about ChatGPT's limitations in terms of its mathematical capacities and accuracies (Frieder et al., 2024; Kang, Y., 2024; Wardat et al., 2023), ChatGPT also sometimes made errors in responses to students' questions in our study. Their errors were mainly due to misunderstandings or misinterpretations of prompts given by students. Our student participants exercised caution in utilizing ChatGPT-generated information, perhaps due to the in-class discussion on the limits of ChatGPT. When ChatGPT suggested a new idea, such as using Heron's formula for finding the triangle area in terms of its side lengths, most students, excluding one student, ignored its suggestions. *No* students asked ChatGPT follow-up questions about Heron's formula, how to deduce it, or whether and how they could derive it by building on their original arguments. Students may believe that their arguments looked good at the moment when revising arguments, so they did not think they needed to attempt to prove using the new idea that ChatGPT suggested. However, some students admitted that they were unfamiliar with Heron's formula, so they were unsure what to do with it in their revision processes and ended up not using it in their arguments. So, from our study, we see that students' views on proof and proving and their abilities to understand and interpret ChatGPT's feedback, which is also related to their knowledge about mathematics and geometry associated with the given proving task, affected their work of constructing arguments with ChatGPT's assistance.

We think there are a number of things that could lead to better outcomes in using large language models (LLMs) like ChatGPT as a proof assistant in a mathematics classroom:

1. Spend more time training students to use and critique ChatGPT's proof feedback: As described in the methods section, students in this study used ChatGPT in in-class activities before working on the ChatGPT-embedded individual homework activities that we used as data for this study. We observed that some students did not approach the homework activities in the same way as they were shown in the in-class activities. More explicit scaffolding is needed for students to have them carefully read and evaluate ChatGPT's feedback and to guide them to arrive at an (expected) final claim in generating arguments with ChatGPT's assistance. "Teaching students to *think critically* is essential when utilizing AI systems" (Frick, 2024, p. 22). Students should build abilities to "evaluate, verify, question AI outputs, detect hallucinations, biases, inaccuracies" (Cain, 2024, p. 50). Additional exercises on using ChatGPT as a proof assistant also might help students perform better on homework.
2. Use more advanced models: In this study, students used ChatGPT-3.5 because of its free availability at the time of the class. However, newer models, such as GPT-4.0, have the potential to provide better feedback to users. Koubaa (2023) compared the capabilities of GPT-3.5 and GPT-4.0 on several benchmark data sets, including the Math section of the SAT exam, the Quantitative section of the Graduate Record Examination (GRE), and the Advanced Placement (AP) Calculus BC exam. On each of the exams, GPT-4.0 scored better than GPT3.5: 89th percentile vs. 70th percentile on SAT Math, 80th percentile vs. 25th percentile on GRE Quantitative, and 59th percentile vs. 7th percentile on AP Calculus BC. The improved performance on math-related benchmarks may indicate an ability to provide better feedback on mathematical proofs, and further studies should be conducted to establish this. We have anecdotally observed better proof feedback from GPT-4.0, which has since become freely available through the Microsoft Copilot application. This is another avenue to explore.
3. Prompt engineering: This study focused on using simple prompts similar to what a novice user might be able to come up with on their own. However, it may be that experimenting with different kinds of prompts might help the models orient the feedback more appropriately for the audience. As noted by Bozkurt and Sharma (2023), "By approaching a conversational generative AI strategically, with clear purpose, tone, role, and context, a prompt-based conversational pedagogy can be established, enabling communication and interaction that facilitate teaching and learning effectively." They suggest creating prompts that are concise, clearly define the objective, and include all appropriate context and detail while emphasizing testing and iteration to ensure they produce the best feedback for the task. Further experimentation may reveal the best kinds of prompts to help novice proof-writers improve their proof-writing skills.

Better outcomes may also be realized by creating new LLM-based proof assistant tools for novice provers. The results of this research and future research along these lines should inform the design of such systems. Developers should consider how learning objectives of particular learning activities can be infused into special-purpose LLM-based applications, whether through fine-tuning of the model, prompting, or other means of adaptation. Some of the difficulties and inconsistencies experienced by students working with a general-purpose tool might be mitigated if an educator could specify information relevant to the learning activity, like the problem statement, evaluation criteria, and/or instructions on the kinds of feedback that students should receive.

Hanna et al. (2019) points out that "we already have some sense that proof assistants greatly diminish the need for verification and justification, but we know almost nothing of their potential contribution to other roles of proof, such as explanation, communication, discovery, and systematization, or how they now may become more relevant as pedagogical motivation for the learning of proof in the classroom" (p. 9). While working on the ChatGPT-embedded proving activities we purposefully designed for this study, communicating with ChatGPT through argumentation about a method for finding the area of a triangle in terms of its side lengths, our student participants engaged in verifying, justifying, and explaining their methods as well as evaluating sample arguments suggested by ChatGPT. In this interaction, ChatGPT often asked students to increase the readability of proofs for the audience by clarifying symbols they used in their arguments and justifying each step. So, we believe that ChatGPT has the potential to contribute to helping students learn proof as "a form of discourse" (Volmink, 1990), a vehicle to interchange ideas among mathematics communities based on shared meanings of mathematics concepts and proof. Further investigations are needed on how LLMs like ChatGPT could contribute to other roles of proof in the context of mathematics classrooms. Students still learn existing knowledge in school mathematics, even at a college level. Teaching existing knowledge is vital because it is the foundation for building new knowledge. However, as new technologies are invented, we need to think about new pedagogical approaches to teaching proof. Currently, most proof-related practices focus on engaging students in generating, analyzing, evaluating, justifying, or communicating arguments (proofs) about mathematical claims, already known true or false statements by the mathematics community. However, with new proof technologies, like mathematicians, students may be able to have experience discovering new ideas during their school mathematics experience. "Students and educators transition from passive recipients to active co-creators of their learning experiences" (Cain, 2024, p. 47). Thus, further investigations are needed to explore whether ChatGPT has the capabilities to engage students in discovering new ideas in mathematics.

Our study only began to address some potential capabilities of ChatGPT as a tool for supporting students in generating arguments that can be accepted as proofs. This study examined student work with ChatGPT assistance, focusing on one geometry-proving task. Our study context was also a mathematics pathways course designed to review fundamental concepts covered in school mathematics for college students. So, this study context might be considered a high school mathematics classroom. However, considering other groups of student populations (e.g., elementary, middle, and high school students) and using different levels of proving tasks, further studies are also needed to explore the capabilities of ChatGPT as a proof assistant in various mathematics classrooms. Comparing the capabilities of ChatGPT with the capabilities of available proof technologies such as GeoGebra's automated proving tools (Hohenwarter et al., 2019), Lean theorem provers (Avigad, 2019), and LLM-based proof generators (First et al., 2023) and examining how each of these tools can support students in proof-based activities should be investigated. Such a comparative analysis of proof technologies will help both schoolteachers and college instructors who teach proof in their instructional decision-making about what proving technology would be appropriate to use in what situations.

Endnote

[1] All student names are pseudonyms.

ORCID

Hyejin Park: https://orcid.org/0000-0002-0451-5728
Eric D. Manley: https://orcid.org/0009-0008-9372-5567

Conflict of Interest

The authors declare that they have no competing interests.

References

Alcock, L., & Weber, K. (2010). Referential and syntactic approaches to proving: Case studies from a transition-to-proof course. In F. Hitt, D. Holton, & P. Thompson (Eds.), *Research in collegiate mathematics education VII* (pp. 93-114). American Mathematical Society.

Appel, K., & Haken, W. (1977). The solution of the four-color-map problem. *Scientific American, 237*(4), 108-121. https://www.jstor.org/stable/24953967

Arnau, D., Arevalillo-Herráez, M., Puig, L., & González-Calero, J. A. (2013). Fundamentals of the design and the operation of an intelligent tutoring system for the learning of the arithmetical and algebraic way of solving word problems. *Computers & Education, 63*, 119-130. https://doi.org/10.1016/j.compedu.2012.11.020

Avigad, J. (2019). Learning logic and proof with an interactive theorem prover. In G. Hanna, M. de Villiers, & D. Reid (Eds.), *Proof technology in mathematics research and teaching, Series: Mathematics education in the digital era*, (Vol. 14, pp. 277-290). Springer.

Bertot, Y., & Castéran, P. (2013). *Interactive theorem proving and program development: Coq'Art: The calculus of inductive constructions.* Springer Science & Business Media. https://doi.org/10.1007/978-3-662-07964-5

Botana, F., Hohenwarter, M., Janičić, P., Kovács, Z., Petrović, I., Recio, T., & Weitzhofer, S. (2015). Automated theorem proving in GeoGebra: Current achievements. *Journal of Automated Reasoning, 55*(1), 39-59. https://doi.org/10.1007/s10817-015-9326-4

Bozkurt, A., & Sharma, R. C. (2023). Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world. *Asian Journal of Distance Education, 18*(2), i–vii. http://www.asianjde.com/ojs/index.php/AsianJDE/article/view/749

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)* (pp. 1877-1901). Curran Associates Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Butgereit, L., & Martinus, H. (2023). Prof Pi: Using WhatsApp bots and GPT-4 for tutoring mathematics in underserved areas. In A. Seeam, V. Ramsurrun, S. Juddoo, & A. Phokeer (Eds.), *Proceedings of the International Conference on Innovations and Interdisciplinary Solutions for Underserved Areas* (pp. 278-289). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-51849-2_19

Cain, W. (2024). Prompting change: Exploring prompt engineering in large language model AI and its potential to transform education. *TechTrends, 68*(1), 47-57. https://doi.org/10.1007/s11528-023-00896-0

Chazan, D. (1993). High school geometry students' justification for their views of empirical evidence and mathematical proof. *Educational Studies in Mathematics, 24*(4), 359-387. https://doi.org/10.1007/BF01273371

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., ... & Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*. https://doi.org/10.48550/arXiv.2107.03374

Coe, R., & Ruthven, K. (1994). Proof practices and constructs of advanced mathematics students. *British Educational Research Journal, 20*(1), 41-53. https://doi.org/10.1080/0141192940200105

de Moura, L., Kong, S., Avigad, J., Van Doorn, F., & von Raumer, J. (2015). The Lean theorem prover (system description). In A. Felty, & A. Middeldorp (Eds.), *Automated Deduction-CADE-25* (pp. 378-388). Springer International Publishing. https://doi.org/10.1007/978-3-319-21401-6_26

de Villiers, M. (1990). The role and function of proof in mathematics. *Pythagoras, 24*(24), 17-24.

Epp, S. S. (2003). The role of logic in teaching proof. *The American Mathematical Monthly, 110*(10), 886-899. https://doi.org/10.1080/00029890.2003.11920029

Finnie-Ansley, J., Denny, P., Becker, B. A., Luxton-Reilly, A., & Prather, J. (2022). The robots are coming: Exploring the implications of OpenAI Codex on introductory programming. In J. Sheard, & P. Denny (Eds.), *Proceedings of the*

24th Australasian Computing Education Conference (pp. 10-19). Association for Computing Machinery. https://doi.org/10.1145/3511861.3511863

First, E., Rabe, M. N., Ringer, T., & Brun, Y. (2023). Baldur: Whole-proof generation and repair with large language models. *In Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (pp. 1229-1241). Association for Computing Machinery. https://doi.org/10.1145/3611643.3616243

Fitting, M. (2012). *First-order logic and automated theorem proving.* Springer Science & Business Media.

Font, L., Gagnon, M., Leduc, N., & Richard, P. R. (2022). Intelligence in QED-Tutrix: Balancing the interactions between the natural intelligence of the user and the artificial intelligence of the tutor software. In P. R. Richard, M. P. Vélez, & S. Van Vaerenbergh (Eds.), *Mathematics Education in the Age of Artificial Intelligence: How Artificial Intelligence Can Serve Mathematical Human Learning* (pp. 45-76). Springer International Publishing. https://doi.org/10.1007/978-3-030-86909-0_3

Font, L., Richard, P. R., & Gagnon, M. (2018). Improving QED-Tutrix by automating the generation of proofs. *arXiv preprint arXiv:1803.01468.* https://doi.org/10.48550/arXiv.1803.01468

Frick, T. (2024). Are we dupes? Limitations of AI systems: What should educators do with them? *TechTrends, 68*(1), 14-26. https://doi.org/10.1007/s11528-023-00893-3

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2024). Mathematical capabilities of ChatGPT. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Track on Datasets and Benchmarks*. https://proceedings.neurips.cc/paper_files/paper/2023/file/58168e8a92994655d6da3939e7cc0918-Paper-Datasets_and_Benchmarks.pdf

Gattupalli, S., Lee, W., Allessio, D., Crabtree, D., Arroyo, I., & Woolf, B. (2023, July 7). *Exploring pre-service teachers' perceptions of large language models-generated hints in online mathematics learning* [Virtual Presentation]. AIED2023 Empowering Education with LLMs - the Next-Gen Interface and Content Generation, Tokyo, Japan.

Gonthier, G. (2008). Formal proof-the four-color theorem. *Notices of the AMS, 55*(11), 1382-1393.

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

Hanna, G., Reid, D., & de Villiers, M. (2019). Proof technology: Implications for teaching. In G. Hanna, D. A. Reid, & M. de Villiers (Eds.), *Proof Technology in Mathematics Research and Teaching, Series: Mathematics Education in the Digital Era* (Vol. 14, pp. 3-9). Springer. https://doi.org/10.1007/978-3-030-28483-1_1

Harel, G., & Sowder, L. (1998). Students' proof schemes: Results from exploratory studies. In A. H. Schoenfeld, J. Kaput, & E. Dubinsky (Eds.), *Research in collegiate mathematics education III* (pp. 234-283). American Mathematical Society.

Healy, C., & Hoyles, L. (2000). A study of proof conceptions in algebra. *Journal for Research in Mathematics Education, 31*(4), 396-428. https://doi.org/10.2307/749651

Hohenwarter, M., Kovács, Z., & Recio, T. (2019). Using GeoGebra automated reasoning tools to explore geometric statements and conjectures. In G. Hanna, M. de Villiers, & D. Reid (Eds.), *Proof technology in mathematics research and teaching, Series: Mathematics Education in the Digital Era* (Vol. 14, pp. 215-236). Springer.

Jiang, A. Q., Li, W., Tworkowski, S., Czechowski, K., Odrzygóźdź, T., Miłoś, P., ... & Jamnik, M. (2022). Thor: Wielding hammers to integrate language models and automated theorem provers. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho & A. Oh (Eds.), *Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022)* (pp. 8360-8373). https://proceedings.neurips.cc/paper_files/paper/2022/file/377c25312668e48f2e531e2f2c422483-Paper-Conference.pdf

Kang, Y. (2024). A study on the didactical application of ChatGPT for mathematical word problem solving. *Communications of Mathematical Education, 38*(1), 49-67. https://doi.org/10.7468/jksmee.2024.38.1.49

Knapp, J. (2005). *Learning to prove in order to prove to learn.* https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ed7046da768e6b39c38f7db472a0ff158230b075

Knuth, E. J. (2002). Proof as a tool for learning mathematics. *Mathematics Teacher, 95*(7), 486-490. https://doi.org/10.5951/MT.95.7.0486

Knuth, E. J., Choppin, J. M., & Bieda, K. N. (2009). Proof: Examples and beyond. *Mathematics Teaching in the Middle School, 15*(4), 206-211. https://doi.org/10.5951/MTMS.15.4.0206

Koubaa, A. (2023). GPT-4 vs. GPT-3.5: A concise showdown. https://doi.org/10.20944/preprints202303.0422.v1

Kovács, Z. (2015). *Computer based conjectures and proofs in teaching euclidean geometry* [Doctoral dissertation, Johannes Kepler University].

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Transactions on Nneural Networks and Learning Systems, 33*(12), 6999-7019. https://doi.org/10.1109/TNNLS.2021.3084827

Marty, R. H. (1986). Teaching proof techniques. *Mathematics in College* (Spring/Summer), 46-53.

Moore, R. C. (1994). Making the transition to formal proof. *Educational Studies in Mathematics, 27*, 249-266. https://doi.org/10.1007/BF01273731

National Academies of Sciences, Engineering, and Medicine. (2023). *Artificial intelligence to assist mathematical reasoning: Proceedings of a workshop*. The National Academies Press. https://doi.org/10.17226/27241
Nipkow, T., Wenzel, M., & Paulson, L. C. (2002). *Isabelle/HOL: A proof assistant for higher-order logic.* Springer Berlin Heidelberg.
Papadopoulos, D. (2016). *Transitioning to proof with worked examples*. Drexel University.
Patero, J. L. (2023). Revolutionizing Math Education: Harnessing ChatGPT for student success. *International Journal of Advanced Research in Science, Communication and Technology, 3*(1). 807-813. https://doi.org/10.48175/IJARSCT-12375
Selden, A. (2012). Transitions and proof and proving at tertiary level. In G. Hanna, & M. de Villiers (Eds.), *Proof and proving in mathematics education* (pp. 391-414). Springer. https://doi.org/10.1007/978-94-007-2129-6_17
Selden, J., & Selden, A. (1995). Unpacking the logic of mathematical statements. *Educational Studies in Mathematics, 29*(2), 123-151. https://doi.org/10.1007/BF01274210
Strauss, A. L., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Sage.
Stylianides, G. J., & Stylianides, A. J. (2009). Facilitating the transition from empirical argument to proof. *Journal for Research in Mathematics Education, 40*(3), 314-352. https://doi.org/10.5951/jresematheduc.40.3.0314
Van Vaerenbergh, S., & Pérez-Suay, A. (2022). A classification of artificial intelligence systems for mathematics education. In P. R. Richard, M. P. Vélez, & S. Van Vaerenbergh (Eds.), *Mathematics education in the age of artificial intelligence: How artificial intelligence can serve mathematical human learning* (pp. 89-106). Springer International Publishing. https://doi.org/10.1007/978-3-030-86909-0_5
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In U. von Luxburg, & I. Guyon (Eds.), *Proceedings of the Advances in Neural Information Processing Systems 30* (NeurIPS 2017) (pp. 6000–6010). Curran Associates Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
Volmink, J. D. (1990). The nature and role of proof in mathematics education. *Pythagoras, 23*, 7–10.
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
Weber, K. (2001). Student difficulty in constructing proofs: The need for strategic knowledge. *Educational Studies in Mathematics, 48*(1), 101-119. https://doi.org/10.1023/A:1015535614355
Weber, K., & Alcock, L. (2004). Semantic and syntactic proof productions. *Educational Studies in Mathematics, 56*, 209-234. https://doi.org/10.1023/B:EDUC.0000040410.57253.a1
Yang, K., Swope, A. M., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R., & Anandkumar, A. (2023). LeanDojo: Theorem proving with retrieval-augmented language models. *In Proceedings of the 37th Conference on Neural Information Processing Systems (NIPS '23)*. https://doi.org/10.48550/arXiv.2306.15626
Zafrullah, Z., Hakim, M. L., & Angga, M. (2023). ChatGPT Open AI: Analysis of mathematics education students learning interest. *Journal of Technology Global, 1*(1), 1-10. https://penaeducentre.com/index.php/JTeG/article/view/35/33