Can AI Grade Mathematical Proofs?

Nathan Carter, Gabi Friedman, and Omeed Tavakoli, Bentley University

Abstract

Many projects are attempting to create software applications that can grade students' mathematical proofs, giving feedback on their reasoning, including where it is correct and where it is not. In 2025, it is necessary to ask how well generative AI can perform at this task. Can LLMs give accurate and helpful feedback on the highly structured and specific type of reasoning done in mathematics?

Our study answers this question in two ways. First, if we give a carefully crafted prompt to an LLM designed for mathematical work, how reliably and helpfully can it grade proofs written by undergraduate students in their first proof-based mathematics course? Second, if we give a popular LLM chat bot a prompt that a student might, do the results differ?

We gathered sample student proofs from mathematics courses at several different institutions and asked both human experts and generative AI to give feedback on those proofs, in the same format. We find that AI agrees with human experts less than 75% of the time even on the simplest measures of proof correctness, meaning that it is not yet ready to use for giving students reliable feedback.

Background and Significance

Since at least the 1990s, researchers have tried to create software for real-time feedback on student proof writing. In 2025, could we build such software on top of AI, making LLM math 'tutoring' accessible and reliable? We therefore ask: **Can AI grade mathematical proofs?**

AI coding assistants are continually improving, and proof-writing shares similarities with coding (structure, technicalities, precise meaning). But researchers at Purdue found that 52% of ChatGPT answers to programming questions contained incorrect information. Will AI fail to provide accurate feedback on proof-writing as well?

Two Research Questions

1. *Embedded use case:* Should software developers who are building a proof-checking app use AI for grading? We chose AIs with proven math credentials and carefully engineered a prompt that might be used inside a larger software system.
2. *Student use case:* Should students use AI when seeking feedback on their proof writing? We chose one of the most popular chatbots, ChatGPT, and used a short and simple prompt like what a student might write quickly.

Motivation

We anticipate that our complete results will be useful for developers, educators, and researchers alike.

If the AI fails to perform in the embedded use case, the investment of time and resources into ongoing research on custom application development is justified. If the AI fails to perform in the student use case, it is not yet ready to give students reliable feedback. Instructors should warn students that misinformation from AI could undermine their mastery of mathematical proof-writing.

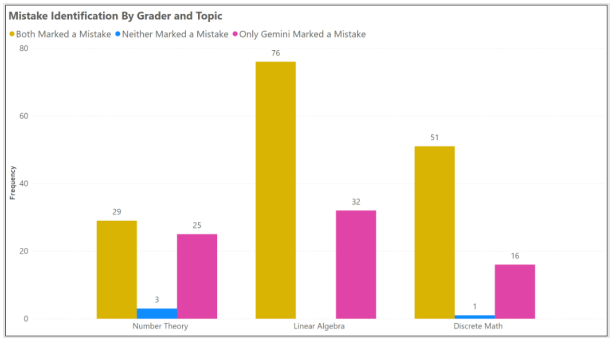
Reference

Samia Kabir et al. "Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, May 2024. doi: 10.1145/3613904.3642596. url: http://dx.doi.org/10.1145/3613904.3642596.

AI and Human Graders Almost Always Disagreed.

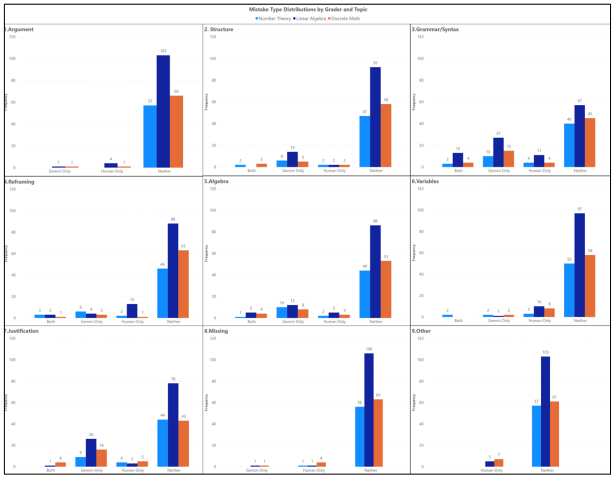
Humans and AI disagreed about 95.7% of the time, whether about the existence of a mistake, or the category to which the mistake belonged.

Summary	Proof Count	% (of 233)
Both find same mistake	6	2.6%
Both find no mistake	4	1.7%
Only human finds a mistake	0	0%
Only AI finds a mistake	73	31.3%
Each finds different mistakes	150	64.4%



AI Critiqued Language Most of All.

Topic most critiqued by AI: Linear Algebra.



Mistake types most commonly marked by AI: grammar/syntax issues, lack of justification, and structure not suitable for the theorem.

Methods

We gathered 784 proofs from real examples of student work donated by 5 instructors at different schools. Several AI tools were tasked with grading all the proofs. We recruited professors with PhDs in statistics or mathematics as expert volunteers who then graded over 250 proofs. We are recruiting expert graders to finalize comparisons--volunteers welcome!

AI did B Work at Best.

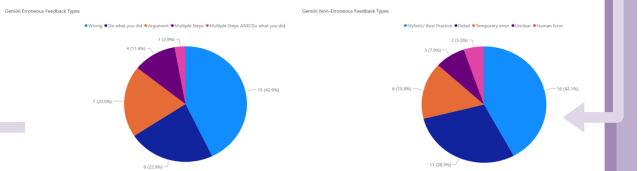
AI was incorrect *at least* 15% of the time, given that at least 35 proofs were erroneously identified by Gemini as containing a mistake.

When Only Gemini Marked a Mistake	Proof Count	% (of 73)
Disagreements caused by AI error	35	47.9%
Disagreements caused by human error	2	2.7%
Disagreements not caused by any error	36	49.3%

Patterns of Disagreement Emerged.

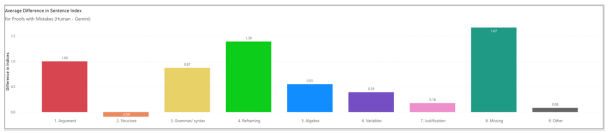
AI feedback on proofs where humans found no error was often incorrect. The feedback was either categorically wrong, directed students to do what they had just done, or rejected a valid argument structure.

At times, AI feedback was not technically incorrect when Gemini disagreed with the human about whether the proof contained an error. In these instances, the AI feedback was typically picky, pedantic, or unforgiving of early errors that students later corrected.



AI Was Quick to Critique.

Humans tended to identify mistakes later in the proofs. AI was quick to mark missing sections, incorrect reframing, invalid arguments, and grammar/syntax issues.



Special Thanks

Data donors, expert graders, researchers, collaborators, and the Bentley University Mathematics Department faculty