ChatGPT and AI Enhancing Undergrad Math

Matthias Kawski
kawski@asu.edu
School of Mathematical & Statistical Sciences
Arizona State University
Tempe, AZ 85287, USA

Abstract

Generative Artificial Intelligence (AI) systems such as ChatGPT, built on large language models (LLMs), present both challenges and opportunities for teaching and learning mathematics. Although not originally designed for logical reasoning or formal proof-writing, these models are increasingly capable of generating plausible drafts of mathematical arguments. We briefly discuss the surprising successes of generative AI at the International Mathematical Olympiad, as well as more well-established systems that can validate and certify formal proofs.

The primary focus is to promote a powerful and innovative use of publicly available generative AI systems to enhance courses, especially at the mid-undergraduate level, that introduce students to formal proof writing. These may be termed "Introduction to Proof" *or* "Mathematical Structures"*,* "Advanced Calculus" *or* "Introduction to Analysis"*. Combinatorics, Graph Theory, Topology should work equally well. Within this context, we examine the capabilities, and notable limitations, of LLMs.*

Far from being a hindrance, the logical missteps frequently made by these models, ranging from minor inaccuracies to fundamentally flawed strategies, can be powerful tools in an inquiry-based classroom. Students are often more engaged when critiquing imperfect AI-generated arguments than when critiquing drafts written by their peers, making these tools uniquely effective for fostering critical thinking and proof literacy.

1 Introduction

In recent years, the term *"artificial intelligence"* (AI) has dominated media coverage. At the same time, college administrators have increasingly encouraged faculty to explore its integration into the curriculum. Long before the rise of generative AI, topics like *"machine learning"* and *"deep neural networks"* were already popular among students across disciplines. This interest surged with the release of ChatGPT by OpenAI in November 2022. Built on a *"large language model"* (LLM), ChatGPT generates human-like text, and, more recently, images, with applications ranging from drafting business letters and academic essays to analyzing biomedical data and producing functional computer code. Many observers have compared its potential societal impact to that of the steam engine, the microchip, or even the internet.

This article explores and reflects on innovative uses of generative AI systems such as ChatGPT to support and enhance learning in mid-level undergraduate mathematics courses. It is neither a research study in mathematics or mathematics education, nor a technical analysis of AI architectures. Rather, it serves as a progress report from the classroom, highlighting how generative AI can be integrated meaningfully into undergraduate instruction to substantial enhancements for achieving core learning objectives. The examples and reflections presented here aim to demonstrate not only how these tools can support student learning, but also how, by its very nature, mathematics and closely related disciplines are arguably the best places to help students developing critical skills to decide when to trust AI-generated (as well as human-generated) claims. In mathematics we do not argue whether a statement is true or false, either we trust an irrefutable proof, or we acknowledge counterexample. AI-generated proofs are either true or false, no arguing.

In early 2023, concerns about generative AI in academia centered largely on its potential for misuse: cheating on homework, generating essays for college applications, and undermining academic integrity. Yet just two years later, the landscape has shifted dramatically. Using AI systems to draft documents has become routine across disciplines. The focus has now turned toward teaching students how to effectively prompt these systems to generate meaningful and relevant output.

In Section 4.2, we illustrate the early stages of this pedagogical shift in the context of mathematical problem solving and proof writing. We argue that interacting with AI, chatting in the true sense of the word, offers valuable opportunities for students to develop reasoning and communication skills. This form of structured argumentation with a generative system may ultimately help students become more effective thinkers and collaborators, both with machines and with each other. We are only beginning to explore this potential.

Section 4.2 examines a classic math problem, accessible even to middle schoolers, that ChatGPT has consistently failed to solve correctly since its release over two and a half years ago. This example underscores a central theme of the article: large language models (LLMs) like ChatGPT were not designed to construct logically sound arguments, but their persistent reasoning flaws can be turned into powerful teaching tools. See Section 6 "Addendum" for remarks about changes with GPT5.

Since early 2023, the author has actively incorporated AI-generated errors into sophomore- and junior-level courses—such as Mathematical Structures (proof writing), Advanced Calculus, and Topology, as a means to sharpen students' reasoning skills. Many of ChatGPT's missteps closely resemble common student errors and are similar to those found in informal solution sources online.

These flawed outputs are especially effective in inquiry-based, active learning environments where students collaborate, present drafts of solutions, and critique each other's work. One of the biggest challenges in such settings is overcoming students' reluctance to critique peers—either out of insecurity or concern about offending classmates. Using ChatGPT as a "non-human peer" has proven highly successful in lowering these barriers. Since February 2023, this strategy has encouraged more open and critical discussions. Section 3 outlines this approach and provides several examples from the past 30 months.

2 A Brief Intro to AI systems and their capabilities

The main focus of this article is on using publicly available generative AI systems based on large language models (LLMs), such as ChatGPT, offer a unique opportunity to help students write logically correct proofs, to develop greater skepticism, and to sharpen their ability to spot errors in purported proofs or technical arguments.

This section takes a brief detour to reflect on the broader landscape, including AI systems that are significantly more advanced. On one end, we have automated proof verification systems; on another, we have enhanced LLMs that recently made headlines by winning gold medals at the International Mathematical Olympiad, just as we were finalizing this article.

2.1 Generative AI systems and Large Language Models

In the most naive terms, one might think of a pretrained large language model (LLM) as a *transformer* that takes a finite sequence of words (say, about 100) as input and predicts the next word that best continues the sequence. Similar words in similar contexts should yield similar outputs. For a more detailed, but still very accessible, introduction, see, for example, [3].

This strategy has proven extremely powerful and effective in domains such as writing essays or business letters. However, one would not expect it to perform equally well in mathematics, whether in arithmetic, algebra, or formal proofs. Indeed, even two years ago, we saw ChatGPT produce surprisingly coherent attempts at writing simple proofs (e.g., the irrationality of certain square roots). Yet, the final steps often broke down, especially in basic algebra. The article [9] highlights just how much difficulty ChatGPT most recently still had with routine algebraic and calculus exercises.

More importantly, LLMs naturally lack any sense of situational awareness. Their internal "state" is simply the string of successive inputs. To illustrate this limitation in an extreme way, consider training an LLM on a large number of chess games represented as strings of words, such as:

1.e4 c5 2.Nf3 e6 3.d4 cxd4 4.Nxd4 Nc6 5.Nb5 d6 6.c4 Nf6 7.N1 c3a6 8.Na3 d5 9.cxd5 exd5

(This example is from the beginning of a famous game between Karpov and Kasparov in 1985.) Clearly, if the model is working only from the most recent N moves, without recalling the full game state, or if there is even a small typo or substitution of a single word, the result can be catastrophic. Indeed, it is almost comical to watch LLMs like ChatGPT and Gemini play chess against each other, as in various tournaments, easily found by www-searches. Their moves often reveal a complete lack of awareness of the actual state of the board.

Given this extreme example of where pure LLMs fail, it is all the more astonishing how far these systems have advanced in mathematics—particularly in the short time since their first public release less than three years ago. This limitation also hints at where LLMs are likely to succeed in mathematics and where they may struggle. Over the years, we have seen some bizarre solutions to classic calculus problems, for example, the well-known optimization problem where a swimmer (or dog) wants to choose the fastest route involving swimming toward a straight shoreline and then running along it to reach a destination. If the same problem is reworded in terms of, say, swamps, dollars, and pipelines, the model tends to perform even worse. The core issue is that LLMs do not form a mental picture of the geometry involved in such problems.

Indeed, the examples in Sections 3 and 4 illustrate this point clearly: LLMs perform quite well on proofs that follow a natural, linear sequence of arguments, where each step builds directly and necessarily on the previous ones. By contrast, current systems perform very poorly when a proof involves multiple intertwined ideas that must come together for a logical conclusion. To the author's best understanding, this is precisely the area where the most advanced (yet unreleased) LLMs are making significant progress—namely, in incorporating deeper mathematical and logical reasoning capabilities, and in handling multiple parallel lines of reasoning simultaneously.

LLMs have been advancing at breakneck speed, and integrating genuine reasoning abilities is now one of the field's central goals. To assess this progress, a wide variety of benchmarks have been proposed. Among the more popular benchmark suites [1], over a dozen datasets have a distinctly mathematical focus—including one derived from the annual American Invitational Mathematics Examination (AIME) [18], as described in [17]. For a more formal and comprehensive bench-marking study of LLMs' proof-writing abilities, particularly in the context of textbook problems from undergraduate and graduate courses—see [6].

2.2 Logical reasoning and Proof Verification Systems

AI-based proof verification systems are of great interest not only at the highest levels of mathematical research but also for the mid-level undergraduate courses we focus on here. One obvious appeal is the potential to automate the grading of student work written in essay form, as is typical for most proofs. Beyond offering significant time and cost savings, such systems could also improve grading accuracy and fairness. Current automated or online grading systems for algebra and (pre-)calculus assignments and exams have very limited capabilities. Unfortunately, this has led to a noticeable decline in the sophistication of questions typically posed to students at these levels. By contrast, we are especially excited about the potential for students to use AI systems to check their own proof drafts, written in natural language, and receive immediate, constructive feedback.

At the highest levels of mathematical research, arguably the most prominent AI-based proof verification system is LEAN. For a recent overview of LEAN's structure and applications, see [2]. Progress in this area is advancing at breakneck speed—consider, for example, the recent formalization of the proof of the Polynomial Freiman-Ruzsa (PFR) conjecture in Lean 4, as announced in [22], along with Tao's 2023 commentary and the striking dependency graph featured on the first page. Another major tool in this space is the Isabelle proof assistant, widely used for theoretical research and development of formal verification. More directly relevant to undergraduate education is Lurch+ [15], which brings this kind of AI-based formal reasoning much closer to accessibility at the mid-undergraduate level. The author almost incorporated Lurch+ into his courses in Spring 2025. Its modular design is easy to use—for instance, instructors and students can load packages containing basic axioms, definitions, and theorems (on logic, sets, functions, etc.) and create new custom environments as needed.

However, two key issues prevented us from using it in our setting. First, Lurch+ employs nonstandard conventions for the scoping of quantifiers. For example, does the expression $(\forall x)P(x) \longrightarrow Q$ mean $((\forall x)P(x)) \longrightarrow Q$ or $(\forall x)\,(P(x) \longrightarrow Q))$. Second, there is a workflow overhead when building a sequence of logically related theorems. In particular, to use a newly proven "if and only if" statement in a subsequent proof, the student must first manually define two separate inference rules for each direction.

2.3 Generative AI Systems' Uses in Learning Mathematics

In a completely different direction, while writing this manuscript, the author was taken by surprise just last week by the announcements in [4] and [23]: two LLM-based generative AI systems had performed at gold medal level at the 2025 International Mathematical Olympiad (IMO). The IMO is a global competition for high school students, but only for the very best on the planet. To appreciate the difficulty of the problems and the depth of reasoning required to solve them, see for example [5, 19]. Unlike in 2024, the best AI systems in 2025 no longer required any human intervention to translate natural language into a formal language that the systems could process. Instead, these AI models now operate autonomously, using natural language as both input and output.

It is truly remarkable how quickly these systems have evolved, adding genuine reasoning ability to large language models. OpenAI's AlphaGeometry 2 and AlphaProof are two prominent examples of such advances. The ideal, perhaps even the "holy grail," is a future in which creative generative models are integrated with rigorous proof verification systems, capable of discovering new theorems, generating original proofs, and validating them automatically.

While many of the current tools remain proprietary or under active development, the pace of progress suggests that we may soon see new publicly available systems, like ChatGPT, incorporating these capabilities. Notably, OpenAI employee Alexander Wei recently revealed that GPT-5 will be released soon. However, the IMO gold-medal model remains part of an experimental research study and is not scheduled for public release in the coming months.

3 Learning from false proofs generated by ChatGPT

The main focus of this article is to promote a powerful and innovative use of publicly available generative AI systems to enhance courses especially at the mid-undergraduate level, often termed *"Introduction to Proof"*, *"Mathematical Structures"*, *"Advanced Calculus"*, *"Introduction to Analysis"*. Combinatorics, Graph Theory, Topology should work equally well.

In our inquiry-based classes, we consider it essential that students share their (drafts of) proofs with one another, critique each others' work, and then collaboratively improve it. This process begins in small homework study groups or in-class teams of two to four students. But the big challenge is to get this done with the whole class together. As hard as it is, employing many tricks and strategies, to get students to volunteer to present, it is even even much harder to get their peers to politely, but uncompromisingly point out factual mistakes. Students often explain their silence by saying they *"did not want to hurt their peers' feelings."* There are many ways to continue addressing these issues, but the rise of publicly available generative AI systems offers a simple, and perhaps fleeting, opportunity: Instead of critiquing each other's work, students collaboratively scrutinize purported proofs produced by generative AI.

It turns out that ChatGPT and similar systems often make the same kinds of mistakes that our students make. As a result, students still learn what not to do: they become better at identifying common errors and, in turn, improve their own writing. Some of the most frequent mistakes addressed at this level include:

- using symbols without first declaring their meaning (e.g., failing to assign values or properly quantify them, or confusing free and bound variables),
- confusing necessary and sufficient conditions,

- incorrectly negating implications (especially when hidden universal quantifiers are involved),
- failing to consider all possible cases exhaustively,
- misusing quantifiers in induction proofs, and many more of a similar nature.

prove: the difference of two converging sequences conerges

We are to **prove** that the difference of two converging sequences converges.

Theorem:

Let (a_n) and (b_n) be sequences of real numbers such that

$$\lim_{n\to\infty} a_n = A \quad \text{and} \quad \lim_{n\to\infty} b_n = B.$$

Then the sequence $(a_n - b_n)$ converges, and

$$\lim_{n\to\infty}(a_n - b_n) = A - B.$$

Proof:

Let $\epsilon > 0$. Since $(a_n) \to A$, there exists $N_1 \in \mathbb{N}$ such that for all $n \geq N_1$,

$$|a_n - A| < \frac{\epsilon}{2}.$$

Since $(b_n) \to B$, there exists $N_2 \in \mathbb{N}$ such that for all $n \geq N_2$,

$$|b_n - B| < \frac{\epsilon}{2}.$$

Let $N = \max(N_1, N_2)$. Then for all $n \geq N$,

$$|(a_n - b_n) - (A - B)| = |(a_n - A) - (b_n - B)| \leq |a_n - A| + |b_n - B| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Hence, $(a_n - b_n) \to A - B$, which proves the claim.

✅ **Conclusion:** The difference of two converging sequences converges to the difference of their limits.

Figure 1: ChatGPT is nearly perfect at many textbook exercises and standard proofs

Generative AI systems get an impressive number of proofs right. In one of the first tasks in the lower level class, students are asked to prove that *"there is no rational number whose square is 2"*. In the higher level class, they are instead asked to *"use the real number axioms to prove that there exists a real number whose square equals 2."* For the latter, it took only minimal coaxing to prompt ChatGPT to improve a vague, continuity-based argument into one that gets the algebra right, presenting an explicit formula for δ in terms of ε. In figure 1 we share a near perfect response for a question at the transition to *Advanced Calculus: "Prove that the difference of two converging sequences converges."*

Except for the tacit assumption, suggestion by the prompt, that the sequences are real-valued, this is practically perfect. Inventing names for variables, and introducing them in the correct order: (a_n), (b_n), A, B, ε, N_1, N_2, N will make any teacher happy. Some most common mistakes by students stem from restating the definitions for convergence of the given sequences, often using *"for all* ε_1, ε_2 *there exist* N_1, N_2, *such that ...* and believing that N_1, N_2 are bound variables now. But they are not, their scope has expired, and declaring $N = \max\{N_1, N_2\}$ is meaningless. Of course, the key is to first fix an arbitrary $\varepsilon > 0$, then define ε_1, ε_2 of ε and only after these introduce N_1, N_2, and N. So ChatGPT is already too smart to provide a *desired* false argument here. But the perfection does no go far beyond – already the proof of a similar prompt with difference replaced by product gets ChatGPT into trouble.

The general experience from the past two and a half years of using this approach is that public versions of generative AI systems have progressively improved. They can now solve many—if not most—of the standard exercises found in mainstream textbooks at this level. One gets the impression that solution manuals for such texts may have been part of the training data.

However, it often takes only slight rewording of a question to throw off the system. A popular and amusing example from induction proofs is the assertion: *"All rabbits (in every set of n rabbits) have the same color.* Change it to *"All kangaroos have the same father,"* and experiment how soon the *Large Language Model* will fail to see that mathematically this is the same question.

We conclude this section with a false proof, confidently provided by ChatGPT, of a major (but, as stated, incorrect) theorem: *Every continuous function can be continuously extended to the closure of its domain.* This issue is often first encountered in precalculus courses. For example, given the function $f\colon \mathbb{R} \setminus \{2\} \mapsto \mathbb{R}$, $f(x) = (x^2 - 4)/(x - 2)$, students are asked to extend the function so that its graph does not have a hole. On the graduate level and beyond, this issue becomes central in the theory of PDEs and in functional analysis, e.g. when relaxing regularity hypotheses on boundary conditions, or when *"inventing"* notions of generalized, or *"weak"* solutions.

This incorrect theorem is a central topic of discussion in the first *Advanced Calculus* course, serving as a key motivation for introducing the concepts of uniform continuity, compactness, and sequential compactness. To the author's delight, and to the benefit of his students, as recently as a few weeks ago, ChatGPT still confidently produced a flawed proof of this incorrect assertion (see figure 2). For the complete ChatGPT-generated response, see [12].

Statement: Suppose f is a continuous function defined on a domain D, and let p be a limit point of D. Then f has a continuous extension at p, meaning there exists a function g defined on $D \cup \{p\}$ such that g is continuous at p and $g(x) = f(x)$ for all $x \in D$.

Proof:

Since p is a limit point of D, there exists a sequence (x_n) in D such that $x_n \to p$ as $n \to \infty$.

Consider the limit:

$$L = \lim_{x \to p} f(x)$$

if it exists.

Step 1: Show that L exists

Since f is continuous on D, for any sequence (x_n) in D with $x_n \to p$, we have:

$$\lim_{n \to \infty} f(x_n) = L$$

if the limit exists. To show the existence of L, we must show that for every sequence (x_n) in D converging to p, the corresponding sequence $(f(x_n))$ has the same limit.

Since f is continuous on D, its values do not oscillate wildly. That is, if two sequences (x_n) and (y_n) in D converge to p, then $f(x_n)$ and $f(y_n)$ must approach the same value (otherwise, f would fail to be well-behaved near p, contradicting continuity). This implies that $f(x)$ has a unique limit as $x \to p$.

Thus, L is well-defined:

$$L = \lim_{x \to p} f(x)$$

exists.

Step 2: Define the Extension

Define a function $g : D \cup \{p\} \to \mathbb{R}$ by:

$$g(x) = \begin{cases} f(x), & x \in D \\ L, & x = p \end{cases}$$

where $L = \lim_{x \to p} f(x)$.

Step 3: Show that g is Continuous at p

Figure 2: Wrong proof of continuous extension to the closure of its domain

4 A long chat with AI about Proving Tiling Theorems

4.1 The problems: Tiling with bent and straight triominos

For well over two decades the author has used the following problem for extended formal proof-writing projects (with some problem solving) in the second or third year transition class *Mathematical Structures*. The particular worksheet [10] was first used for informal arguments at a Julia Robinson Math Festival for middle and high school age students, but has proven very stable and useful for many years, and serves as a scaffold for the writing project.

Problem 1 *For which positive integers n can every $n \times n$ board from which any one square has been removed (no matter its location) be tiled with (L-shaped) bent triominos?*

We usually request that students make an effort to provide reasonable precise definitions for *"board"*, and *"tiling"* – this is part of the exercise. Rather than considering graphs, most people and the AI systems, prefer to work with rectangular regions in the plane with positive integer side lengths. Then a bent triomino is a 2×2 board from which any one 1×1 square has been removed. A (possibly *"mutilated"*) board is tiled when it is completely *"covered"* by the specified tiles *"without overlaps"*, except possibly along the *"edges"* and *"vertices"*. These few comments already suggest that this invites, but also leaves much room, for students as well as for AI, to provide less or more precise definitions. A fun playground for both, and for the instructor.

One key objective in this class is careful attention to alternating quantifiers, here the key is *"no matter which square has been removed"*. AI systems, just like students often present *proofs by example* (like in the game Solitaire), which are inadmissible as proofs of universally quantified theorems. Conversely, both AI and students, often jump to the inadmissible conclusion that a board cannot be tiled just because after a *reasonable effort* they could not find a tiling

While the problem most certainly is much older, the articles [7, 8] are widely regarded as having much popularized this question, with focus on mutilated $2^k \times 2^k$ boards. The arguments invite for many students one of their first and cleanest induction proofs (no distracting algebra) There are at least two very different ways to handle the induction.

For general n, one usually also proceeds by induction, but usually in steps $n \longrightarrow n + 3$ (or $n \longrightarrow n + 6$), as an obvious necessary condition for tileability is that 3 does not divide n. Proceeding in two (or four) parallel chains of induction arguments, with at least as many distinct starting cases, invites for some fun problem solving, and very careful writing.

The choices of multiple induction starts are made more complicated by the fact that the usual general arguments for augmenting the size by 3 do not apply for the special case when going from 4 to 7 (one special case of the location of the *removed square*), and the observation that mutilated 5×5 boards can be tiled or cannot, depending on the specific location of the *removed square*.

To understand the latter, first consider a complete (not mutilated) 3×3 board. To tile it completely with (bent) triominos, one needs $(3 \times 3)/3 = 3$ tiles. But the board has four corner squares, and each tile can cover at most one corner. Hence a complete tiling is impossible.

To extend this kind of argument to the mutilated 5×5 board consider a so-called *gingham pattern* resulting from intersecting horizontal and vertical strips, both directions colored in an alternating fashion white and some semi-transparent color. More precisely, if we label the rectangular grid of squares by pairs (i, j) of integers, $1 \le i \le 5$ and $1 \le j \le 5$, then we color the squares with two odd coordinates red, and leave all other squares white. (Note that we are labelling the squares, not the corners as we would do if the board was a rectangular subset of the Cartesian coordinate plane – another place that invites for very clear language from the students as well as from the AI).

The mutilated 5×5 board has 24 squares, demanding 8 bent triominos. But each of these can cover at most on of the squares that are marked red. Hence the mutilated board can only be tiled if the removed square is one of those that have been marked red. This is a necessary condition – but it is easy to show that in all other cases, one can indeed tile the mutilated 5×5 board. Utilizing rotational and reflective symmetries one only needs to consider three cases.

We note on the side, that a closely related problem uses straight triominos, a problem which ChatGPT apparently was trained on. The tiles are 3×1 rectangular boards (which may be rotated as usual). A popular task is to cover, or *tile* a mutilated 8×8 board with these. Rather than using a checker-board two-coloring (as when tiling with dominos), here a similar popular necessary condition utilizes two [!] tri-colorings of the square board. Again with coordinates as above use the colorings defined by coloring the tile (i, j) by the *color* $c \in \{0,\ 1,\ 2\}$ if and only if $i \pm j \equiv c$ which each of the cases + and − giving rise to a diagonal rainbow pattern. It turns out that the board of 64 squares will consist of 22, 21, 21 squares of the respective colorings. Since each straight triomino covers exactly one square of each color (in either coloring), the mutilated board can be tiled only if (again, just a necessary condition) in both [!] colorings the removed square has the color that 22 squares have. This restricts the possible locations of the removed square to the locations $(3, 3)$, $(6, 3)$, $(3, 6)$, $(6, 6)$. Sufficiency is not hard to prove – just present one tiling (utilizing rotational symmetry).

This problem is fun, somewhat challenging, but nowhere of the difficulty of Olympiad problems. Thus it is quite amazing that even 2 1/2 years after their first public releases none of the standard, publicly available generative AI systems, that we tried, such as ChatGPT, Gemini DeepMind, DeepSeek, Claude, etc. do not produce reasonably correct answers, without laborious interventions, the subject of the next subsection.

4.2 The next step: chatting and cooperative proving

Over the last few years we have asked several times to have ChatGPT to respond to Problem 1. There has been surprisingly little change using the basic public version. In July 2025, we also tried several other generative AI systems with similarly abysmal results. See Section 6 "Addendum" for remarks about changes with GPT5, which now quickly comes up with the correct solution, yet still repeats most of the mistakes describe in the sequel when asked to justify its answer. This subsection presents some persistent false arguments that these systems generate when prompted with this problem.

The main difference as opposed to the previous section is that we tried to tease out the *"chat"* component of ChatGPT. Indeed, this appears to be the current recommended primary use of such LLMs, e.g., when generating business letters. While generally the current versions of the most prominent LLMs are pretrained (the *"P"* in ChatGPT) and do not learn new responses based on our dialogues with them, they substantially improve their responses upon given feedback in a *chat* or dialogue.

Following are commented excerpts of a 2-hour dialogue with ChatGPT in July 2025 (using

GPT4.o) which is publicly available and archived online in its full length [11] (as a .pdf document it is 63 pages long).

This dialogue or chat is quite typical of several ones we have had with various LLMs, and it exemplifies Terence Tao's description [21] of such a discussion (albeit by a much more naive interviewer, this author) as in this excerpt

```
The experience seemed roughly on par with trying to advise a mediocre,
but not completely incompetent, (static simulation of a) graduate student.
However, this was an improvement over previous models, whose capability was
closer to an actually incompetent (static simulation of a) graduate student.
```

First, a few noteworthy observations: We deliberately provided a very imprecise question with any undefined terms. Indeed, this is the way most people start on truly new problems, from research mathematicians and Math Circles [24], to the general public including the way many students talk and think. Formalizing the problem with precise definitions, and possibly axioms and major theorems, to be used is a major step that is not to be underestimated.

Key facts and constraints:

1. **Size and divisibility:**

- Each bent triomino covers exactly 3 squares.
- The mutilated board has $n^2 - 1$ squares.
- For it to be tileable, $n^2 - 1$ must be divisible by 3.

This gives the divisibility condition:

$$3 \mid (n^2 - 1)$$

Rewrite $n^2 - 1 = (n-1)(n+1)$.

Since 3 is prime, for 3 to divide $(n-1)(n+1)$, at least one of $n-1$ or $n+1$ must be divisible by 3.

- So, either $n \equiv 1 \pmod 3$ or $n \equiv 2 \pmod 3$.
- In other words, $n \not\equiv 0 \pmod 3$.

Figure 3: ChatGPT starts the necessary conditions involving divisibility

Like many students, ChatGPT spend much time and effort at laying out lots of details of the obvious necessary condition of divisibility: The total number of squares covered by m tiles each containing k squares is mk.

Step 2: Color argument

Let's use a **coloring argument** to check if tiling is possible.

Color the $n \times n$ board in a **3-color pattern** with colors 0,1,2 cyclically:

$$\text{color}(i, j) = (i + j) \mod 3$$

Properties of bent triominoes on this coloring:

- Each bent triomino covers **exactly one square of each color**.

Why? Because the shape is an "L" with squares arranged so that the three cells have all distinct $(i + j)$ $\text{mod } 3$ values.

Implication of the coloring for tiling:

- The total count of each color on the mutilated board must be the same, since each triomino covers one of each color.
- The full board has:

$$\#\text{color } c = \begin{cases} \frac{n^2}{3}, & \text{if } 3 \mid n \\ \text{two colors have } \lceil \frac{n^2}{3} \rceil, \text{ one has } \lfloor \frac{n^2}{3} \rfloor, & \text{otherwise} \end{cases}$$

- But one square is removed.

For the counts of colors after removing one square to remain balanced (equal counts for all three colors), **the removed square must be of the color which appears one more time than the others.**

Figure 4: ChatGPT insists on forcing arguments with 3-colorings, which can't work and the counting is still wrong of the frequency of $\lceil x \rceil$ and $\lfloor x \rfloor$

Step 4: Does this hold for all bent triomino orientations?

Bent triominoes come in four orientations (rotations by 0°, 90°, 180°, 270°). The above argument works similarly for each orientation, because the squares always occupy coordinates that differ by 1 in each direction, so their $(i + j) \mod 3$ values are three consecutive residues mod 3.

For example, if the bent triomino is oriented as:

```nginx
X X
  X
```

Positions could be $(i, j), (i, j + 1), (i + 1, j + 1)$:

- Colors: $(i + j) \mod 3, (i + j + 1) \mod 3, (i + j + 2) \mod 3$ — again distinct.

Summary:

Because the bent triomino covers three cells whose $(i + j)$ sums differ by 0, 1, and 2 modulo 3, the three covered squares always have three distinct colors in the coloring $(i + j) \mod 3$.

Figure 5: ChatGPT insists on forcing arguments with 3-colorings

In a very strange way, not only ChatGPT, but also several other generative AI systems insist to again and again utilize coloring arguments. Indeed, tri-colorings are very useful when tiling with straight triominos but close to useless for bent triominos. They all make lots of false assertions and claims, and they keep repeating them even when given specific prompts.

Step 6: Result

It turns out **the 3×3 square cannot be tiled completely by bent triominoes.**

Step 7: Why is this important?

- If you cannot tile the 3×3 square, **you cannot tile larger boards whose areas are multiples of 9 by bent triominoes either** without mutilation.

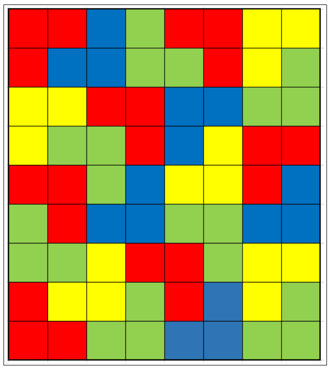

Figure 6: Even children can tile a complete 9×9 board.

In order to help ChatGPT to more carefully investigate tiling mutilated 5×5 board, we prompted it to first address the problem of tiling a complete 3×3 board with bent triominos. It was a very long process to coax ChatGPT towards eventually writing a meaningful correct statement. However, this was immediately followed be the above abstruse comment about the impossibility of tiling the 3×3 board having consequences for the tileability of a 9×9 board. Such *hallucinations* of generative AI systems can hardly be even clearer than in mathematics – LLM just making things up without any reasoning or proof.

Applying the pigeonhole principle:

- Total number of bent triominoes to cover the board: 8 (since $24/3 = 8$).
- The maximum number of black squares that can be covered by 8 bent triominoes is **8** (since each covers at most one black square).
- The 5×5 board has **9 black squares**.

So:

- If the **removed square is black**, then 8 black squares remain.
- Then 8 triominoes can cover these 8 black squares (at most one black square per triomino), so coverage is possible.
- If the **removed square is white**, then all 9 black squares remain.
- But only 8 triominoes are available, so at most 8 black squares can be covered.
- One black square would remain uncovered, making tiling impossible.

Final conclusion:

> **The mutilated 5×5 board can be tiled by bent triominoes if and only if the removed square is black (both coordinates odd).**

Figure 7: Finally – after over 20 pages of refutations and tutoring

Being reminded of the quote from Terence Tao at the beginning of the section, this is like pulling teeth: It took about 20 pages of refutations until ChatGPT eventually came up with the correct conclusion about tiling mutilated 5×5 boards with bent triominos.

Statement to prove:

> For every $n \times n$ board with $n \geq 7$ odd and $n \not\equiv 0 \pmod 3$, **any mutilated board** (one square removed) can be tiled by bent triominoes.

Figure 8: Finally ChatGPT arrives at stating the correct theorem

After two hours, and 63 pages, ChatGPT finally stated the correct answer. It really should not yet call it a theorem, but rather a conjecture at this point, as so far it really only has a very rough idea for a proof. Given the preceding experiences, it likely will take several more hours of coaxing, likely a few hundred pages of attempts, to eventually make this into a proof that would be acceptable in the author's junior level college class.

Inductive proof idea:

1. **Partition the $n' \times n'$ board into:**
   - An $n \times n$ inner square.
   - A surrounding border of width 3 on all sides (so total $n' = n + 6$).
2. **Place the missing square:**
   - If the removed square lies in the inner $n \times n$ board, then by inductive hypothesis, the inner $n \times n$ board (with that square removed) can be tiled.
   - The surrounding border (width 3) is a frame of known dimensions.
3. **Tile the border frame:**
   - The border frame is a $n' \times n'$ board with an $n \times n$ hole.
   - Because its width is fixed (3), the border can be tiled by bent triominoes (or decomposed into 2×3 or 3×2 rectangles, each tileable by two bent triominoes).
   - This tiling of the border does not interfere with the inner board.
4. **If the removed square lies in the border:**
   - Then tile the inner $n \times n$ board fully (no missing square).
   - Tile the border with the missing square removed, using similar decomposition.

Figure 9: Making this rough idea into an actual proof may take hours more of chatting

The way we like to run our classes, do mathematics: Try to find proofs which may be refuted and do it again, as most beautifully laid out in the famous book [14]. In this case, the author, with perfect knowledge of the problem and many solutions, managed to painfully guide ChatGPT to eventually *make* the right assertion. But it is doubtful whether a struggling student would have been helped by such a mediocre *AI study partner*. In summary, ChatGPT *confidently* makes lots of unfounded assertions.

5 Summary, conclusions, and outlook

Generative AI systems are advancing at a remarkable pace. New developments, such as Multi-Threading Inference Systems and Adaptive Parallel Reasoning [20], are reshaping what these tools can do. As of this writing (August 1), Google DeepMind has announced Gemini 2.5 Deep Think, its first publicly available multi-agent model and a variant of the system that earned a gold medal at the 2025 International Mathematical Olympiad.

For now, however, we can reasonably assume that our students will not yet have access to these cutting-edge systems. Most will continue using freely available or low-cost generative LLMs, which still frequently produce flawed reasoning and incorrect proofs. These shortcomings, far from being a drawback, offer valuable pedagogical opportunities.

By engaging students in the critique and correction of AI-generated arguments, they may become not only more skilled proof writers, but also more critically minded citizens. The goal is to cultivate confidence grounded in understanding: self-assured students who trust the correctness of their work not because it matches an answer key, but because they can identify and flawed reasoning. This shift places a new responsibility on faculty: to rethink the kinds of problems we assign. While routine exercises still have their place, we should increasingly emphasize problems that expose the limitations of AI-generated reasoning and which spark meaningful mathematical discussion.

6 Addendum

Right after the submission of the original version of this article, GPT-5 was released with much fanfare. The new model did produce a reasonably correct *answer* to Problem 1, far superior to the 68-page dialogue in [11]. However, upon closer scrutiny, it failed just as abysmally when it came to providing proofs for its claims. Figure 10 illustrates its complete lack of understanding of what it means to tile a mutilated 5×5 board with bent triominos. GPT-5 even went so far as to claim: *"Each letter is one L-triomino; you can check each letter appears exactly three times and forms an L."*

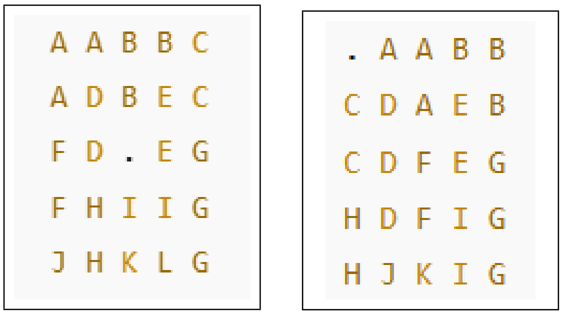

Figure 10: Mutilated 5×5 board.

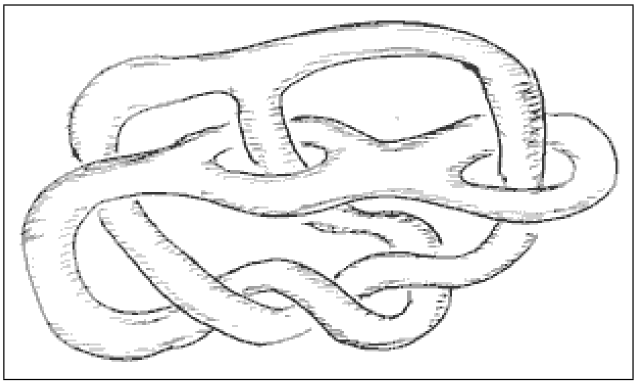

Figure 11: Find the genus of a pictured surface.

Other notable recent experiments included feeding GPT-5 a .jpg file of the logo for our undergraduate topology class (see Figure 11) and asking it for the genus of the surface. The system responded with two beautifully written pages surveying the relevant theory and explaining how to apply it to the given surface, and then it confidently concluded [13]: *"In your case: From the loops depicted, it's clear the surface has two distinct handles—each corresponding to a "hole" like on a double donut—hence genus 2."* This is false. When it comes to solutions of routine textbook exercises, one striking example is GPT-5's entirely incorrect answer to a slight variation of Exercise 16 in Section 17 of Munkres' *Topology* [16]. The task is to find the closure of a particular subset of the *ordered square*, that is, the set $[0,1] \times [0,1] \subseteq \mathbb{R}^2$ equipped with the order topology inherited from the lexicographic order. GPT-5's response displayed faulty reasoning and an apparently illicit interchange of existential and universal quantifiers.

In summary, although generative AI systems are improving at a breakneck pace, they remain a long way from exhibiting sound logical reasoning.

7 Acknowledgments.

The authors gratefully thank the two anonymous referees for their careful reading of the manuscript, their encouragement, and for providing many valuable suggestions for its improvement.

In addition to standard spelling and grammar checkers, the author also *thanks* ChatGPT which he asked to also improve the wording and flow of many paragraphs.

References

[1] Evidently AI. 200 LLM benchmarks and evaluation datasets, 2025. URL: https://www.evidentlyai.com/llm-evaluation-benchmarks-datasets.

[2] Jeremy Avigad, Leonardo de Moura, and Soonho Kong. Theorem Proving in LEAN, Oct 21, 2024. URL: https://leanprover.github.io/theorem_proving_in_lean/theorem_proving_in_lean.pdf.

[3] Nathan Carter. ChatGPT gets all the attention. *Math Horiz.*, 32(2):24–27, 2024. doi:10.1080/10724117.2024.2396266.

[4] Davide Castelvecchi. DeepMind and OpenAI Models Solve Maths Problems at Level of Top Students, July 24, 2025. doi:10.1038/d41586-025-02343-x\bigwedge.

[5] International Mathematical Olymiad Foundation. IMO 2025 Problems, 2025. URL: https://www.imo-official.org/problems.aspx.

[6] Simon Frieder, Luca Pinchetti, Alexis Chevalier, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, and Julius Berner. Mathematical Capabilities of ChatGPT. *NeurIPS 2023 Datasets and Benchmarks*, 2023. doi:10.48550/arXiv.2301.13867.

[7] Martin Gardner. L-tromino Tiling of Mutilated Chessboards. *College Math. J.*, 40(3):162–168, 2009. doi:10.4169/193113409X469352.

[8] Solomon Wolf Golomb. Checker Boards and Polyominoes. *Amer. Math. Monthly*, 61:675–682, 1954. doi:10.2307/2307321.

[9] Kris H. Green. Does ChatGPT know Calculus? *J. Humanist. Math.*, 14(1):248–257, 2024. URL: https://scholarship.claremont.edu/jhm/vol14/iss1/14, doi:10.5642/jhummath.ILCN1025.

[10] Matthias Kawski. Tiling Problems, 2010. URL: https://math.la.asu.edu/~kawski/classes/mat300/18smr/various/JRMF-tiling.pdf.

[11] Matthias Kawski and openAI ChatGPT. ChatGPT Tiling Mutilated $n \times n$ Boards with Bent Triominos, July 2025. URL: https://tinyurl.com/y6smbprt.

[12] Matthias Kawski and openAI ChatGPT. Continuous Functions have Continuous Extensions, July 2025. URL: https://tinyurl.com/4826vymc.

[13] Matthias Kawski and openAI ChatGPT. Determine the Genus of a Depicted Surface, September 2025. URL: https://tinyurl.com/32zyst3n.

[14] Imre Lakatos. *Proofs and Refutations: The Logic of Mathematical Discovery.* Cambridge Philosophy Classics. Cambridge University Press, 1976. URL: https://books.google.com/books?id=1n6SFdXCOBQC.

[15] Ken Monks. Proof Verification with Lurch Plus. URL: https://lurch.plus/.

[16] James R. Munkres. *Topology.* Featured Titles for Topology. Prentice Hall, Incorporated, 2000. URL: https://books.google.com/books?id=XjoZAQAAIAAJ.

[17] Art of Problem Solving and Maxwell-Jia. AIME 2024 Dataset. 2024. URL: https://huggingface.co/datasets/Maxwell-Jia/AIME_2024.

[18] Art of Problem Solving and Maxwell-Jia. American Invitational Mathematics Examination (AIME). 2024. URL: https://artofproblemsolving.com/wiki/index.php/American_Invitational_Mathematics_Examination.

[19] Art of Problem Solving Online. IMO Problems and Solutions, 2025. URL: https://artofproblemsolving.com/wiki/index.php/IMO_Problems_and_Solutions.

[20] Jiayi Pan, Xiuyu Li, Long Lian, Charlie Snell, Yifei Zhou, Adam Yala, Trevor Darrell, Kurt Keutzer, and Alane Suhr. Learning Adaptive Parallel Reasoning with Language Models, 2025. URL: https://arxiv.org/abs/2504.15466.

[21] Terrence Tao. OpenAI's new Iteration GPT-o1, 13 September, 2024. URL: https://mathstodon.xyz/@tao/113132502735585408.

[22] Terrence Tao. Formalizing the Proof of PFR in LEAN4 using Blueprint: a Short Tour, 18 November, 2023. URL: https://terrytao.wordpress.com/2023/11/18/formalizing-the-proof-of-pfr-in-lean4-using-blueprint-a-short-tour/.

[23] Alexander Wei. Gold Medal-Level Performance ... International Math Olympiad, July 2025. URL: https://x.com/alexwei_/status/1946477742855532918.

[24] wikipedia. Math Circle: Content Choices. URL: https://en.wikipedia.org/wiki/Math_circle#Content_choices.