Journal of Education and Culture Studies
ISSN 2573-0401 (Print) ISSN 2573-041X (Online)
Vol. 10, No. 4, 2026
www.scholink.org/ojs/index.php/jecs

Original Paper

Judgment Before Answers: Generative AI in First-Year Mathematical Analysis

Saisai Shi[1*]

[1] Institute of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, 233030, P. R. China

[*] Corresponding author, Saisai Shi, E-mail: Saisai_shi@126.com

Received: October 3, 2026 Accepted: October 9, 2026 Online Published: November 13, 2026

doi:10.22158/jecs.v10n4p37 URL: http://dx.doi.org/10.22158/jecs.v10n4p37

Abstract

Generative AI now produces plausible solutions to many standard exercises in first-year Mathematical Analysis, yet the same systems mishandle quantifier scope, omit essential hypotheses, and invite students to substitute fluent text for mathematical judgment. This paper argues that the pedagogical question is not whether GenAI should enter the analysis classroom, but how its use can be designed so that the student, rather than the tool, performs the cognitive labor the course exists to cultivate. Drawing on recent evidence concerning AI in proof-based undergraduate courses, the paper develops a three-principle design framework—tiered access, error-centered inquiry, and judgment-oriented assessment—and instantiates it through worked examples from the first-year syllabus. The examples span limit computations ($\varepsilon - N$ and $\varepsilon - \delta$ arguments, rationalization, boundedness), integral computations (substitution, integration by parts, symmetry, the Riemann integral and its limitations), and the convergence of function sequences. In each case the AI output is treated not as an answer key but as a case to be diagnosed, and the student's task is to supply the judgment that distinguishes a valid argument from a plausible one.

Keywords

Generative AI, Mathematical Mnalysis, Proof Education, Assessment Design

1. Introduction

First-year Mathematical Analysis is not a course in computation. It is a course in the disciplined use of quantification, the construction and critique of proofs, and the formation of the judgment that distinguishes a valid argument from a plausible one. Students learn to control the order of quantifiers, to produce a δ that depends on ε and on nothing else, to recognize which hypothesis of a theorem is

doing essential work, and to see why pointwise convergence is insufficient for the theorems that matter. These capacities are not acquired by watching correct arguments; they are acquired by making judgments—by deciding whether an argument is justified, where it breaks, and what would repair it.

Generative AI enters this pedagogical space with a double face. On one side, publicly available systems can produce competent proofs of sequence convergence, correct $\varepsilon-\delta$ arguments for linear functions, and fluent explanations of the mean value theorem. On the other, the same systems fail in revealing ways: a small change from difference to product in a convergence proof can produce a flawed argument, and topics such as uniform convergence and the completeness of the real numbers are handled poorly, with missing phases and improper application of definitions. This combination of surface competence and structural fragility is not a defect to be waited out; it is a pedagogical resource if the curriculum is designed to exploit it.

The risk is equally clear. When students can query an AI for answers without first committing their own reasoning, the path of least resistance is offloading. Wardat show that unproctored homework in AI-accessible environments increasingly reflects what they call synthetic fluency rather than internalized understanding, and that the alignment between homework and proctored examinations collapses as course content becomes more conceptual. The design implication is not prohibition but a deliberate engineering of the sequence of cognitive labor: the student must reason first, and the AI must respond to that reasoning rather than replace it.

This paper develops a framework for that engineering and illustrates it with worked examples from the first-year Mathematical Analysis syllabus. Section 2 reviews what recent work establishes about GenAI's capabilities and failure modes. Section 3 states the three design principles. Section 4 provides extended examples from limit theory, integral theory, and the theory of convergence. Section 5 discusses conditions and risks, and Section 6 concludes.

2. What GenAI Does and Does Not Do in Analysis

2.1 Competence on Canonical Tasks, Fragility under Variation

The most consistent finding in recent studies is that GenAI performs well on the distribution of problems it has seen and poorly when a problem requires recognizing structural identity across superficial change. ChatGPT can demonstrate that the limit of a sequence exists and that a linear function is continuous using the $\varepsilon-\delta$ definition, and it can produce step-by-step solutions for basic proofs in Real Analysis. Yet the same system struggles with tasks that require sustained structural reasoning: proofs of the completeness of the real numbers and of the uniform convergence of series of functions are frequently incomplete or definitionally loose. The model has learned the surface form of canonical exercises; it has not acquired the capacity to transfer an argument strategy to a context that differs in more than wording.

This pattern matters because Mathematical Analysis is precisely the course in which students are supposed to acquire transferable argument structures. If AI is used only as a solution generator, it

reinforces the misconception that mathematics consists of recognizing problem types and retrieving templates. If it is used as a contrast case—here is an output that looks correct but fails under a small perturbation—it can make visible the difference between template matching and structural reasoning.

2.2 The Pedagogical Value of Flawed AI Arguments

A second line of work turns the model's weaknesses into an asset. Kawski argues that the logical missteps frequently made by generative models, ranging from minor inaccuracies to fundamentally flawed strategies, can be powerful tools in an inquiry-based classroom, and that students are often more engaged when critiquing imperfect AI-generated arguments than when critiquing drafts written by their peers. The engagement is not incidental: the AI's fluency makes its errors less socially fraught to identify, and the exercise positions the student as the bearer of mathematical authority rather than the recipient of a correct answer.

This inversion is pedagogically significant for first-year students, who often treat printed or displayed mathematics as authoritative by default. When the object of critique is an AI output, the question "Is this argument valid?" becomes unavoidable in a way that it is not when the object is a textbook proof. The student must supply the criterion, not merely follow a model.

2.3 The Offloading Risk and the Sequence of Labor

The most serious evidence concerns what happens when AI is available without designed constraints. Wang et al. find that in upper-division courses, the correlation between unproctored homework performance and proctored examination performance collapses, indicating that homework increasingly measures the ability to produce fluent text rather than the ability to reason independently. The finding is not that AI use inevitably harms learning; it is that the default use pattern—query, receive, submit—produces a gap between performance and understanding.

The design response follows directly. If the student must externalize a reasoning attempt before the AI speaks, the AI operates on the student's reasoning rather than replacing it. If the student can query directly for an answer, offloading is the path of least resistance. For Mathematical Analysis, where the goal is proof judgment, this distinction is decisive.

3. Three Design Principles

3.1 Tiered Access

The first principle differentiates tasks by the kind of cognitive work they are meant to elicit and regulates AI access accordingly. The "Traffic Light" model developed in a first-year calculus redesign offers a practical vocabulary: Green tasks permit unrestricted AI use; Yellow tasks permit AI for specified purposes but require the student to document and evaluate what the AI contributed; Red tasks prohibit AI use and require reasoning under conditions where authorship is verifiable.

In Mathematical Analysis the mapping is not arbitrary. Green is appropriate for exploratory work: generating graphs to build intuition about uniform continuity, producing numerical evidence for a limit, or checking routine algebraic manipulation. Yellow is appropriate where AI can extend reach but the

student must retain evaluative authority: asking a model to suggest a proof strategy for a mean value theorem problem and then requiring the student to determine whether the strategy is valid and where it fails. Red is essential for the core proof-writing performances on which the course's credential rests: constructing an $\varepsilon-\delta$ argument unaided, proving that a continuous function on a compact set is uniformly continuous, or showing that pointwise convergence does not imply uniform convergence. The purpose of Red is not to pretend that AI does not exist; it is to create a domain in which the student's own reasoning is the only reasoning available.

3.2 Error-centered Inquiry

The second principle makes the model's fragility the substance of the task. Rather than asking students to use AI to find correct answers, the task asks them to adjudicate AI outputs that are plausible but flawed. The errors that GenAI makes in analysis are frequently the same errors that students make or are about to make. A scope error in quantifier handling—treating N_1 and N_2 as fixed after they have been chosen—is a canonical student error, and it is also an error that AI outputs can exhibit depending on problem framing. A false claim that every continuous function extends continuously to the closure of its domain targets exactly the conceptual gap that the course exists to repair.

Error-centered inquiry reverses the authority relation. The AI is not the answer key; it is a case to be diagnosed. The instructor's role shifts from transmitting correct proofs to designing productive wrong proofs and equipping students with the criteria to distinguish them from correct ones.

3.3 Judgment-oriented Assessment

The third principle follows from the offloading evidence. If unproctored homework increasingly measures synthetic fluency, the assessment system must include performances that cannot be outsourced. A short oral component—asking a student to explain why a particular δ works, or what happens to a theorem if a hypothesis is dropped—provides evidence of internalized understanding that a written problem set cannot. Process artifacts offer a second route: when a Yellow task permits AI use for strategy suggestion, the student submits the interaction trace, a brief evaluation of which suggestions were adopted and why, and a revision of the proof after an error is discovered. The assessment then evaluates the student's judgment, not the AI's output.

4. Worked Examples from the First-Year Syllabus

The framework becomes concrete only when instantiated in specific content. The following examples are designed for a first-year course and assume that students have access to a general-purpose language model. They move from limit computations through integral computations to the convergence of function sequences.

4.1 A Sequence Limit with a Nontrivial N : $\lim_{n\to\infty}\frac{2n^2+1}{n^2+3}=2$

A Green-tier exploratory task asks students to have the AI compute the limit numerically for several values of n and to produce a table. The AI will report values approaching 2 and may state the limit

correctly. The learning begins with the Yellow-tier question: given that the limit is 2, produce an $\varepsilon - N$ proof, and then compare your proof with the AI's.

The student's derivation proceeds as follows. For any $\varepsilon > 0$ we require

$$\left|\frac{2n^2+1}{n^2+3}-2\right| = \left|\frac{2n^2+1-2n^2-6}{n^2+3}\right| = \frac{5}{n^2+3} < \varepsilon.$$

Since $n^2+3 > n^2$, it suffices to require $5/n^2 < \varepsilon$, that is $n > \sqrt{5/\varepsilon}$. We therefore choose $N = \lfloor \sqrt{5/\varepsilon} \rfloor + 1$. Then for all $n > N$,

$$\frac{5}{n^2+3} < \frac{5}{n^2} < \varepsilon,$$

which establishes the limit.

The AI will typically produce this argument correctly. The pedagogical value lies in the follow-up. The instructor presents a second AI-generated proof: "We need $5/(n^2+3) < \varepsilon,$ so $n^2+3 > 5/\varepsilon,$ hence $n > \sqrt{5/\varepsilon - 3}$. Take $N = \lfloor \sqrt{5/\varepsilon - 3} \rfloor$." Students are asked: is this proof valid for every $\varepsilon > 0$? The answer is no—when $\varepsilon > 5/3,$, the quantity $5/\varepsilon - 3$ is negative, and the square root is undefined, so the choice of N fails for large ε. The first proof avoided this by using the coarser but always-available bound $n^2+3 > n^2$. The student must decide which proof is correct, and why the coarser bound is not a defect but a necessity. This is a genuine judgment task: both proofs look like $\varepsilon - N$ arguments, and only one is valid for all ε.

4.2 A Definite Integral by Substitution $\int_0^1 \frac{x}{1+x^2}dx$

A Green-tier task asks the AI to compute this integral. The AI will typically set $u = 1+x^2$, $du = 2xdx$, and obtain

$$\int_0^1 \frac{x}{1+x^2}dx = \frac{1}{2}\int_1^2 \frac{du}{u} = \frac{1}{2}\ln 2.$$

The computation is correct. The Yellow-tier follow-up asks students to check the limits of integration: when $x = 0, u = 1$; when $x = 1, u = 2$. The AI may or may not have made this explicit. Students are asked to rewrite the computation with full attention to the change of limits, and then to explain why failing to change the limits and instead substituting back in x would also give the correct answer here but would fail in other cases. This is a small task, but it targets a procedural habit that AI outputs can obscure: the substitution rule for definite integrals requires either changing the limits or returning to the original variable before evaluation.

4.3 Symmetry and the Definite Integral $\int_{-1}^{1} \frac{x^2}{1+x^4} dx$

The AI may compute this integral by substitution and obtain zero, or it may state that the integrand is odd and the interval symmetric, so the integral is zero. Both are correct. The Yellow-tier task asks students to determine which argument is more robust: the substitution argument requires the antiderivative $(1/4)\ln(1+x^4)$ and evaluation at the endpoints; the symmetry argument requires only the observation that $f(-x) = f(x)$. Students are asked to explain why the symmetry argument is valid and to identify the general principle: if f is odd and integrable on $[-a, a]$, then $\int_{-a}^{a} f = 0$. The AI can state this principle; the student must see why it is true—because the integral over $[-a,0]$ equals the negative of the integral over $[0, a]$ under the substitution $x \to -x$.

A second task asks students to evaluate $\int_{-1}^{1} \frac{x^2}{1+x^4} dx$ and to determine whether symmetry helps. The integrand is even, so the integral equals $2\int_{0}^{1} \frac{x^2}{1+x^4} dx$. The AI may produce this reduction correctly. Students are asked to explain why the even case reduces to twice the half-integral while the odd case gives zero, and to state the general principle for even and odd functions. The task is small but targets a distinction that students often apply mechanically without understanding.

4.4 The Riemann Integral and Its Limits: The Dirichlet Function

A Yellow-tier task presents an AI-generated argument that every bounded function on a closed interval is Riemann integrable. The argument invokes upper and lower Darboux sums and asserts that refinement eventually closes the gap. Students are asked to locate the flaw. The Dirichlet function—1 on the rationals, 0 on the irrationals—is bounded on $[0,1]$, but every upper sum equals 1 and every lower sum equals 0, because every subinterval contains both rationals and irrationals. The gap never closes, and the function is not Riemann integrable.

The AI may or may not produce this counterexample when prompted. The student's task is to determine whether the AI's argument actually establishes the claim, and if not, to construct the object that defeats it. A follow-up asks students to explain why the Dirichlet function is nevertheless Lebesgue integrable, and why the Riemann integral's reliance on partitions into intervals is the source of the limitation. This connects the first-year computation of integrals to the conceptual question of what integration is for.

4.5 A Limit of Integrals $\lim_{n\to\infty} \int_{0}^{1} \frac{x^n}{1+x} dx$

The sequence $f_n(x) = x^n/(1+x)$ converges pointwise to 0 on $[0,1)$. The question is whether the integrals converge to 0. The AI may assert that the limit of the integrals equals the integral of the limit,

invoking some form of convergence theorem. Students are asked to justify this without invoking a theorem they have not proved: since $0 \le x^n/(1+x) \le x^n$ on $[0,1]$, we have

$$0 \le \int_0^1 \frac{x^n}{1+x}dx \le \int_0^1 x^n dx = \frac{1}{n+1} \to 0.$$

The squeeze theorem gives the result. The AI may produce this argument, or it may invoke uniform convergence, which is not available here (the convergence is not uniform on $[0,1)$). Students must determine which argument is valid and why the direct estimate is preferable. The task makes visible the difference between a theorem-based justification and a direct estimate, and it shows that the failure of uniform convergence does not preclude the convergence of the integrals.

5. Conditions, Risks, and What Would Count as Success

The framework proposed here is not a solution to the problem of GenAI in mathematics education; it is a set of design commitments that make the problem tractable. Several conditions must be met for it to function.

First, the instructor must be able to curate AI outputs with pedagogically useful errors. General-purpose models are unstable: the same prompt may yield a correct proof on one occasion and a flawed one on another, and model updates change behavior. This instability is manageable if instructors archive outputs that illustrate specific error types—the unconstrained δ, the false claim about integrability and primitives, the circular integration by parts—and treat them as reusable materials rather than as fixed textbook content.

Second, the tiered model requires student buy-in. Students who have internalized the idea that AI is either forbidden or freely available may resist a regime in which the same tool is sometimes required, sometimes restricted, and sometimes prohibited. Clear communication of the rationale is essential: the tiers protect the development of judgment rather than punish tool use. The first-year calculus redesign that developed the Traffic Light model found that students' strongest demand was not unrestricted access but guidance on effective use, which suggests that students recognize the distinction between using AI to learn and using AI to avoid learning even when their behavior does not always reflect it.

Third, the assessment redesign is resource-intensive. Oral spot-checks and process-artifact grading require instructor time that large first-year courses may not have. Hybrid solutions—short oral defenses for a sample of students, peer evaluation of AI critiques, or structured interaction logs—can reduce the burden, but none eliminates it.

Fourth, the offloading risk is redistributed rather than eliminated. A student can still use AI to generate a critique of an AI output, producing a meta-level synthetic fluency. The defense against this is not technological but pedagogical: the Red-tier performances and oral defenses remain the ultimate check on whether reasoning has been internalized.

What would count as evidence of success? The outcomes most directly tied to the course's purposes are diagnosable and assessable. Students should improve in their capacity to distinguish an asserted δ from a justified one; to identify which hypothesis of a theorem is doing essential work; to recognize a scope error in a quantifier argument; to construct a counterexample to a plausible but false claim; and to choose an argument that establishes a result rather than one that merely surrounds it with notation. These capacities can be assessed through targeted written tasks, through oral defenses, and through the quality of student critiques of AI-generated arguments.

6. Conclusion

The question for first-year Mathematical Analysis is not whether generative AI can produce correct answers—it increasingly can, within limits—but whether its presence can be made to serve the formation of judgment rather than substitute for it. The evidence reviewed here suggests that the answer depends less on the technology than on the design of the learning environment. When AI is positioned as an answer source, it invites offloading and produces a gap between performance and understanding. When it is positioned as a case to be judged—an output whose plausibility must be tested, whose errors must be diagnosed, and whose limitations must be mapped—it can become a cognitive partner in the sense that matters for this course: it expands the space of arguments a student can encounter while requiring the student to supply the judgment that distinguishes valid from invalid reasoning.

The worked examples from limit theory, integral theory, and convergence show that the framework is not abstract. It lives in the specific tasks of deriving a δ that works for every ε of rationalizing a difference of square roots rather than asserting that it tends to zero, of locating the sign error in an integration by parts, of constructing the Dirichlet function as a counterexample to a false claim about integrability, and of explaining why the convergence of x^n on $[0,1]$ is not uniform. If GenAI is to earn a place in the first-year analysis classroom, it will be because it has been made to serve these tasks, not because it has made them easier to avoid.

Fund Project

This research was supported by the Anhui Provincial Quality Engineering Teaching Research General Project (No. 2025jyxm0162) and by the Key Teaching Research Project of the University-Level Undergraduate Teaching Quality and Teaching Reform Program at Anhui University of Finance and Economics (No. acjyzd2025015).

References

Kawski M. (2025). ChatGPT and AI enhancing undergrad math. *Proceedings of the Asian Technology Conference in Mathematics*.

Sánchez-Ruiz L. M., Moll-López S., Nuñez-Pérez A., Moraño-Fernández J. A., & Vega-Fleitas E. (2023). ChatGPT challenges blended learning methodologies in engineering education: A case study on mathematics exams. *Applied Sciences*. https://doi.org/10.3390/app13106039

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*. https://doi.org/10.29333/ejmste/13272