Article

Not peer-reviewed version

Benchmarking AI Mathematical Capabilities: A Comprehensive Evaluation of LLMs in Solving Integration Problems with Error Correction Analysis

In Hak Moon *

Posted Date: 5 September 2025

doi: 10.20944/preprints202509.0445.v1

Keywords: large language models; mathematical reasoning; procedural knowledge; conceptual understanding; geometric applications; problem-solving; follow-up prompting; integration techniques; error correction; comprehensive assessment

 doi:10.20944/preprints202509.0445.v1

Article

Benchmarking AI Mathematical Capabilities: A Comprehensive Evaluation of LLMs in Solving Integration Problems with Error Correction Analysis

In Hak Moon

Science Department SUNY Maritime College, 6 Pennyfield Ave., Bronx, NY 10465, USA;
imoon@sunymaritime.edu

Abstract

This study comprehensively evaluated seven leading Large Language Models (Chat GPT 4o, Gemini Advanced with 1.5 Pro, Copilot Pro, Claude 3.5 Sonnet, Meta AI, Mistral AI, and Perplexity) on Calculus II integration problems through a progressive testing framework consisting of three incremental tests and a final examination conducted between September and December 2024. The evaluation covered various integration techniques, including substitution, integration by parts, partial fractions, area and volume calculations, and series convergence tests, totaling 420 possible points. A key methodological innovation was the implementation of structured follow-up prompting to assess models' error correction capabilities. Results showed significant performance variations, with Gemini Advanced with 1.5 Pro achieving the highest score of 91.7% (A-), followed by Chat GPT 4o with 89.3% (B+), and an overall class average of 83.3% (B). While all models demonstrated strong proficiency in basic integration techniques (averaging 95.2%), they struggled significantly with geometric applications requiring spatial reasoning, with failure rates exceeding 70% on volume of revolution problems. The follow-up analysis revealed that computational errors were readily correctable (90%+ success rate), whereas conceptual misunderstandings, particularly in geometric visualization, persisted despite explicit guidance. These findings demonstrate that while current LLMs have achieved undergraduate-level competency in procedural mathematics, they face fundamental limitations in spatial-geometric reasoning and conceptual understanding, suggesting they can serve as valuable computational tools but cannot replace human instruction for developing mathematical intuition.

Keywords: large language models; mathematical reasoning; procedural knowledge; conceptual understanding; geometric applications; problem-solving; follow-up prompting; integration techniques; error correction; comprehensive assessment

1. Introduction

Recent advancements in artificial intelligence, particularly in Large Language Models (LLMs), have demonstrated remarkable capabilities across various domains, including complex problem-solving tasks [1]. While these models have shown proficiency in language understanding, creative writing, and coding, their mathematical reasoning abilities present a unique opportunity for systematic evaluation. Mathematical problems, especially those requiring procedural knowledge and conceptual understanding, serve as an excellent benchmark for assessing the depth of an LLM's reasoning capabilities [2]. The rapid evolution of these models has prompted increasing interest in understanding not just what they can solve, but how they approach mathematical challenges and where their limitations lie.

The field of mathematics education has long recognized calculus as a critical subject that bridges fundamental mathematical concepts with advanced applications in science, engineering, and economics [3–5]. Within this domain, Calculus II—with its focus on various integration techniques

 doi:10.20944/preprints202509.0445.v1

and their applications—presents particularly rich ground for evaluating AI systems. Integration problems demand diverse cognitive skills, from algorithmic procedure following to spatial visualization and conceptual transfer between different representations of mathematical ideas [6]. This complexity makes integration an ideal testing ground for assessing whether AI systems have developed genuine mathematical understanding or merely sophisticated pattern matching capabilities.

The current landscape of AI mathematical assessment has largely focused on isolated problem-solving accuracy, often missing the nuanced aspects of mathematical reasoning that distinguish true understanding from rote application. While previous studies have evaluated LLMs on various mathematical tasks, few have undertaken a systematic, progressive evaluation that mirrors the structure of actual mathematical learning. Furthermore, the ability to correct errors through guided feedback—a crucial aspect of both human learning and potential educational applications—has received limited attention in existing evaluations. This gap in understanding becomes particularly significant as educators and institutions increasingly consider integrating AI tools into mathematical instruction.

This study addresses these limitations by conducting a comprehensive assessment of seven leading LLMs on Calculus II integration problems through a structured, progressive evaluation framework. We evaluate Chat GPT 4o, Gemini Advanced with 1.5 Pro, Copilot Pro, Claude 3.5 Sonnet, Meta AI, Mistral AI, and Perplexity—models selected for their widespread usage, varying architectural approaches, and diverse development teams. Our methodology incorporates not only accuracy assessment but also a novel follow-up prompting approach that examines how models respond to targeted feedback, providing insights into both their initial capabilities and their potential for error correction. This dual focus on performance and adaptability offers a more complete picture of current AI mathematical reasoning capabilities and their implications for educational applications.

2. Methodology

2.1. Selection of Models and Testing Framework

In this study, we selected seven prominent LLMs) that represent the current state-of-the-art in AI capabilities: Chat GPT 4o, Gemini Advanced with 1.5 Pro, Copilot Pro, Claude 3.5 Sonnet, Meta AI, Mistral AI, and Perplexity. These models were chosen based on their widespread usage, varying architectural approaches, and the diversity of their development teams, which provides a broad representation of current AI capabilities in mathematical reasoning.

The assessment was structured as a progressive evaluation consisting of three tests administered throughout the fall of 2024 (September 20, October 18, and November 22), culminating in a comprehensive final examination on December 9. This approach allowed us to evaluate the models' performance across increasingly complex integration problems and techniques. The first test focused on fundamental integration techniques, the second introduced more complex substitutions and applications, and the third emphasized series convergence and divergence analysis. The final examination combined all of these aspects into a comprehensive assessment.

2.2. Problem Selection and Design

Problems were carefully selected to cover the full breadth of a standard Calculus II integration curriculum. Each problem was designed to test specific integration techniques and conceptual understanding, including:

- Basic indefinite and definite integration
- Integration by substitution and parts
- Integration of trigonometric functions
- Partial fraction decomposition
- Improper integrals and convergence analysis
- Applications such as the area between curves and volumes of revolution

 doi:10.20944/preprints202509.0445.v1

- Series convergence using various tests (Integral, Ratio, Limit Comparison)
- Taylor polynomials and power series

The difficulty of the problems increased progressively both within each test and across the sequence of tests. This design allowed us to establish baseline capabilities and then probe the limits of each model's mathematical reasoning abilities. The problems required multi-step solutions, compelling the models to demonstrate not just formula recall but also procedural understanding and application.

2.3. Evaluation Process and Scoring

Each model was evaluated on identical problems under controlled conditions. For each problem, models were specifically instructed to "Show all the steps" to enable an analysis of their solution processes beyond just the final answers.

The scoring system allocated points based on the complexity and importance of each problem: the first test was worth 90 points, the second test 80 points, the third test 100 points, and the final examination 150 points, resulting in a total of 420 possible points. If the correct answer is found on the first attempt, 10 points are added, and if the correct answer is found on the second attempt, 5 points are added. Letter grades were assigned according to standard academic grading scales, with grade boundaries set at traditional percentage thresholds.

2.4. Follow-up Prompting Methodology

A key methodological innovation in this study was the implementation of a structured follow-up prompting approach. When a model initially produced an incorrect answer, we provided targeted follow-up prompts to guide the model toward a more accurate solution. These prompts were designed to be minimally leading while addressing specific errors or misconceptions in the initial response.

For example, if models made computational errors while expanding expressions like "$(-x^2 + 2x + 3)(-x^2 + 2x + 3)$" or incorrectly set up integrations for area problems, we provided specific feedback such as "The calculation was wrong in Step 4" or "Check which curve is above." This approach allowed us to differentiate between fundamental limitations in mathematical understanding and simple computational errors.

In some cases, multiple follow-up prompts were necessary, with each prompt addressing a specific aspect of the solution process. This progressive refinement approach offers insights not only into the models' initial capabilities but also into their ability to correct errors and incorporate feedback, which is crucial for educational applications.

2.5. Documentation and Analysis

All interactions with the models were meticulously documented, including initial prompts, model responses, follow-up prompts, and final answers. For each problem, we recorded:

- The model's initial answer and whether it was correct
- Specific errors or misconceptions in the solution process
- The nature and content of follow-up prompts
- The model's response to these prompts
- Whether the model ultimately arrived at the correct solution

This comprehensive documentation allowed us to analyze not only the models' final performance scores but also patterns in their mathematical reasoning, common error types, and responsiveness to guidance. These qualitative aspects provide deeper insights into the models' underlying mathematical capabilities beyond what raw scores alone can reveal. The results were then compiled into a detailed performance matrix, showing each model's performance on individual tasks.

 doi:10.20944/preprints202509.0445.v1

3. Results

3.1. Overall Performance

	1st Test (90 points)	2nd test (80 points)	3rd Test (100 points)	Final (150 points)	**Total (420 points)**
Chat GPT 4o	**90 (A)**	**70 (B+)**	**85 (B)**	130 (B+)	**375 (89.3: B+)**
Gemini Advanced with 1.5 Pro	**90 (A)**	**70 (B+)**	**85 (B)**	**140 (A)**	**385 (91.7: A-)**
Copilot Pro	85 (A)	**70 (B+)**	80 (B-)	125 (B)	**360 (85.7: B)**
Claude 3.5 Sonnet	75 (B)	60 (C)	80 (B-)	110 (C)	**325 (77.4: C+)**
Meta AI	80 (B+)	60 (C)	70 (C-)	110 (C)	**320 (76.2: C)**
Mistral AI	**90 (A)**	**70 (B+)**	80 (B-)	120 (B-)	**360 (85.7: B)**
Perplexity	**90 (A)**	45 (F)	75 (B)	115 (C+)	**325 (77.4: C+)**
Average	**85.7 (A)**	**63.6 (C+)**	**79.3 (C+)**	**121.4 (B-)**	**2450 (83.3: B)**

A comprehensive assessment of seven leading Large Language Models (LLMs) revealed significant variations in their mathematical reasoning capabilities when tested on Calculus II integration problems. The overall performance table shows that Gemini Advanced with 1.5 Pro achieved the highest cumulative score of 385 points out of a possible 420, translating to 91.7% and earning an A- grade. Chat GPT 4o followed as the second-best performer with 375 points (89.3%, B+), while Copilot Pro and Mistral AI tied for third place with identical scores of 360 points (85.7%, B). The remaining three models scored at the C level: Claude 3.5 Sonnet and Perplexity each scored 325 points (77.4%, C+), and Meta AI received the lowest score of 320 points (76.2%, C). The class average across all seven models was 350 points, representing 83.3% or a B grade. Performance patterns across the four tests varied considerably. The first test yielded exceptional results, with five models achieving perfect or near-perfect scores, averaging 85.7 points out of 90 (95.2%). The second test revealed the most significant performance disparities, with an average score of only 63.6 points out of 80 (79.5%, C+). The third test, which focused on series convergence, showed moderate performance with an average of 79.3 points out of 100 (79.3%, C+). Finally, the comprehensive final examination yielded an average of 121.4 points out of 150 (80.9%, B-). These results demonstrate that while current LLMs have developed substantial capabilities in mathematical reasoning—achieving what would be considered satisfactory performance in undergraduate settings—a notable gap of 15.5 percentage points remains between the highest and lowest performing models. This indicates significant variations in mathematical proficiency among contemporary AI systems.

3.2. Performance Analysis by Test

Final Test: 12/9/24

The final examination results from December 9, 2024, provide a comprehensive overview of the models' mathematical capabilities across 15 diverse integration problems. Analyzing the performance on each problem reveals distinct patterns in the strengths and weaknesses of the models. Gemini Advanced with 1.5 Pro and Chat GPT 4o consistently achieved the highest scores, correctly solving most problems either initially or with minimal prompting. Gemini Advanced with 1.5 Pro

 doi:10.20944/preprints202509.0445.v1

earned 140 points (A), while Chat GPT 4o received 130 points (B+). The most challenging problems included F-1 (definite integral with fractional exponents), F-3 (complex substitution integral), F-6 (volume of revolution), and F-9 (convergence of an improper integral). Multiple models required extensive follow-up prompts, and some failed to reach correct solutions even with guidance. In particular, problem F-9, which involved determining the convergence of an improper integral from negative infinity, highlighted critical weaknesses in several models' understanding of limits and infinity. Gemini Advanced, Mistral AI, and Perplexity were unable to correctly evaluate the limit behavior, even after targeted prompts. Problem F-6 focused on finding the volume of a solid of revolution about a horizontal line and exposed widespread difficulties with geometric reasoning. Most models struggled with the proper setup of the washer method and made persistent computational errors when expanding expressions like $(-x^2 + 3x + 4)^2$. In contrast, all seven models demonstrated strong performance on series convergence problems (F-10 through F-13), successfully employing various convergence tests, including the Integral Test, Limit Comparison Test, and Ratio Test. The results for the Taylor polynomial problem (F-14) and the power series radius of convergence (F-15) varied; Perplexity notably struggled with coefficient calculations in the Taylor expansion despite multiple correction attempts. Overall, the final examination reinforced the conclusion that while current language models have achieved competency in procedural calculus techniques and formal convergence tests, they show significant limitations in problems requiring spatial visualization, complex algebraic manipulation, and nuanced understanding of limiting behavior. Success rates varied dramatically based on the type of problem rather than just the level of difficulty.

F-1) Evaluate the definite integral: integration from -1 to 1 [(x-x^2)/(5(x^(1/5)))]. Show all the steps. Calculate each operation separately.

Chat GPT 4o	Found the wrong answer. New prompt: Evaluate the definite integral: integration from -1 to 1 [(x-x^2)/(5(x^(1/5)))]. Show all the steps. Calculate each operation separately. **Don't use even or odd function properties.** Still found the wrong different answers.
Gemini Advanced with 1.5 Pro	Found the correct answer: (2/9).
Copilot Pro	Found the correct answer: (2/9).
Claude 3.5 Sonnet	Found the correct answer: (2/9).
Meta AI	Found the correct answer: (2/9).
Mistral AI	Found the correct answer: (2/9).
Perplexity	Found the correct answer: (2/9).

F-2) Find the indefinite integral: integration [(-3x^3+2x^2+5x-6)/(x^2-3)]. Show all the steps.

Chat GPT 4o	Found the correct answer: (-3x^2)/2 + 2x - 2ln|x^2-3| +C.
Gemini Advanced with 1.5 Pro	Found the correct answer: (-3x^2)/2 + 2x - 2ln|x^2-3| +C.
Copilot Pro	Found the correct answer: (-3x^2)/2 + 2x - 2ln|x^2-3| +C.
Claude 3.5 Sonnet	Found the wrong answer. New prompt: Evaluate the indefinite integral: integration [(-

 doi:10.20944/preprints202509.0445.v1

	3x^3+2x^2+5x-6)/(x^2-3)]. Show all the steps. **Don't use a partial fraction method.** Still found the wrong different answers.
Meta AI	Found the correct answer: (-3x^2)/2 + 2x - 2ln|x^2-3| +C.
Mistral AI	Found the correct answer: (-3x^2)/2 + 2x - 2ln|x^2-3| +C.
Perplexity	Found the wrong answer. New prompt: Evaluate the indefinite integral: integration [(-3x^3+2x^2+5x-6)/(x^2-3)]. Show all the steps. **Don't use a partial fraction method.** Found the correct answer: (-3x^2)/2 + 2x - 2ln|x^2-3| +C.

F-3) Evaluate the definite integral: integration from -1 to 1 [(1+(1/(x^3)))^4(1/(x^4))]. Show all the steps.

Chat GPT 4o	Found the wrong answer. New prompt: Evaluate the definite integral: integration from -1 to 1 [(1+(1/(x^3)))^4(1/(x^4))]. Show all the steps. **Use the u-substitution method.** Found the correct answer: (-32/15).
Gemini Advanced with 1.5 Pro	Found the correct answer: (-32/15).
Copilot Pro	Found the wrong answer. New prompt: Evaluate the definite integral: integration from -1 to 1 [(1+(1/(x^3)))^4(1/(x^4))]. Show all the steps. **Use the u-substitution method.** Found the correct answer: (-32/15).
Claude 3.5 Sonnet	Found the wrong answer. New prompt: Evaluate the definite integral: integration from -1 to 1 [(1+(1/(x^3)))^4(1/(x^4))]. Show all the steps. **Use the u-substitution method.** Still found the wrong answer.
Meta AI	Found the wrong answer. New prompt: Evaluate the definite integral: integration from -1 to 1 [(1+(1/(x^3)))^4(1/(x^4))]. Show all the steps. **Use the u-substitution method.** Still found the wrong answer.
Mistral AI (5/10)	Found the wrong answer. New prompt: Evaluate the definite integral: integration from -1 to 1 [(1+(1/(x^3)))^4(1/(x^4))]. Show all the steps. **Use the u-substitution method.** Found the correct answer: (-32/15).
Perplexity	Found the correct answer: (-32/15).

 doi:10.20944/preprints202509.0445.v1

F-4) Evaluate the definite integral: integration from 16 to 4 [(-1)/(sqrt(x)(1+sqrt(x))^2)]. Show all the steps.

Chat GPT 4o	Found the correct answer: (4/15).
Gemini Advanced with 1.5 Pro	Found the correct answer: (4/15).
Copilot Pro	Found the correct answer: (4/15).
Claude 3.5 Sonnet	Found the wrong answer: (-4/15). However, it evaluated the integration from 4 to 16. New prompt: Definite integral was evaluated from 4 to 16. It was wrong. Evaluate the definite integral from 16 to 4. Found the correct answer: (4/15).
Meta AI	Found the correct answer: (4/15).
Mistral AI	Found the correct answer: (4/15).
Perplexity	Found the correct answer: (4/15).

F-5) Find the area of the region: y=2x^3-3x^2-5x, y=-3x^2+3x. Show all the steps.

Chat GPT 4o	Found the correct answer: 16.
Gemini Advanced with 1.5 Pro	Found the correct answer: 16.
Copilot Pro	Found the wrong answer. It found the wrong intersection points. New prompt: It's wrong. Do it again. It found the correct intersection points, but it still found the wrong answer. New prompt: Find the area separately and add them, but it still found the wrong answer.
Claude 3.5 Sonnet	Found the correct answer: 16.
Meta AI	Found the wrong answer. New prompt: It's wrong. Do it again. Still found the wrong answer.
Mistral AI (5/10)	Found the wrong answer. New prompt: It's wrong. Do it again. Found the correct answer: 16.
Perplexity	Found the wrong answer. It found the wrong intersection points. New prompt: It's wrong. Do it again. It found the correct intersection points, but it still found the wrong answer. New prompt: Find the area separately and add them, but it still found the wrong answer.

F-6) Find the volume of the solid generated by revolving the region bounded by the graphs of y=-x^2+3x+6, y=-x+6 about the line y=2. Show all the steps.

 doi:10.20944/preprints202509.0445.v1

Chat GPT 4o	Found the wrong answer. The calculation was wrong in Step 4, the calculation of (-x^2+3x+4)^2. New prompt: Step 4 was wrong. Do it again. Found the correct answer: (384/5) (3.14).
Gemini Advanced with 1.5 Pro	Found the correct answer: (384/5) (3.14).
Copilot Pro	Found the wrong answer. The setup was wrong in Step 3, (R(x))^2 and (r(x))^2. New prompt: Step 3 was wrong. Do it again. It still found the wrong answer. The calculation was wrong in Step 4, the calculation of (-x^2+3x+4)^2. New prompt: Step 4 was wrong. Do it again. Still found the wrong answer.
Claude 3.5 Sonnet	Found the wrong answer. The calculation was wrong in Step 4, the calculation of (-x^2+3x+4)^2. New prompt: Step 4 was wrong. Do it again. Still found the wrong answer. One more time. New prompt: Step 4 was wrong. Do it again. Still found the wrong answer.
Meta AI	Found the wrong answer. The calculation was wrong in Step 4, the calculation of (-x^2+3x+4)^2. New prompt: Step 4 was wrong. Do it again. Still found the wrong answer. One more time. New prompt: Step 4 was wrong. Do it again. Still found the wrong answer.
Mistral AI	Found the wrong answer. New prompt: It's wrong. Do it again. Still found the wrong answer.
Perplexity	Found the correct answer: (384/5) (3.14).

F-7) Find the indefinite integral: integration ((sec(4x))^4)((tan(4x))^5). Show all the steps.

Chat GPT 4o	Found the correct answer: [((tan(4x))^6)/24] + [((tan(4x))^8)/32] + C.
Gemini Advanced with 1.5 Pro	Found the correct answer: [((tan(4x))^6)/24] + [((tan(4x))^8)/32] + C.
Copilot Pro	Found the correct answer: [((tan(4x))^6)/24] + [((tan(4x))^8)/32] + C.
Claude 3.5 Sonnet	Found the correct answer: [((tan(4x))^6)/24] + [((tan(4x))^8)/32] + C.
Meta AI	Found the correct answer: [((tan(4x))^6)/24] + [((tan(4x))^8)/32] + C.
Mistral AI	Found the correct answer: [((tan(4x))^6)/24] + [((tan(4x))^8)/32] +

 doi:10.20944/preprints202509.0445.v1

	C.
Perplexity	Found the correct answer: [((tan(4x))^6)/24] + [((tan(4x))^8)/32] + C.

F-8) Use the partial fraction to find the indefinite integral: integration (x^2-6x+2)/(x^3+2x^2+x). Show all the steps.

Chat GPT 4o	Found the correct answer: 2ln|x| - ln|x+1| - 9/(x+1) + C.
Gemini Advanced with 1.5 Pro	Found the correct answer: 2ln|x| - ln|x+1| - 9/(x+1) + C.
Copilot Pro	Found the correct answer: 2ln|x| - ln|x+1| - 9/(x+1) + C.
Claude 3.5 Sonnet	Found the correct answer: 2ln|x| - ln|x+1| - 9/(x+1) + C.
Meta AI	Found the wrong answer. The B value, B/(x+1), in Step 2 was wrong. New prompt: The B value in Step 2 was wrong. Do it again. Found the correct answer: 2ln|x| - ln|x+1| - 9/(x+1) + C.
Mistral AI	Found the correct answer: 2ln|x| - ln|x+1| - 9/(x+1) + C.
Perplexity	Found the correct answer: 2ln|x| - ln|x+1| - 9/(x+1) + C.

F-9) Determine whether the improper integral diverges or converges. Evaluate the definite integral if it converges: integration from negative infinity to 1 (1-x)e^(-x). Show all the steps.

Chat GPT 4o	Found the correct answer: It diverges.
Gemini Advanced with 1.5 Pro	Found the wrong answer. As x -> -oo, then e^(x) -> oo. It's correct. So (1-x)e^(-oo) -> 0. It's wrong. **x -> -oo, then (1-x)e^(-x) -> oo. This is correct.** New prompt: Integration was correct, but the evaluation steps were wrong. Do it again. Still found the wrong answer.
Copilot Pro	Found the correct answer: It diverges.
Claude 3.5 Sonnet	Found the correct answer: It diverges. However, some middle steps were wrong. (5/10).
Meta AI	Found the correct answer: It diverges. However, some middle steps were wrong. (5/10).
Mistral AI	Found the wrong answer. As x -> -oo, then e^(-x) -> oo. It's correct. So (1-x)e^(-oo) -> 0. It's wrong. **x -> -oo, then (1-x)e^(-x) -> -oo. This is correct.** New prompt: Integration was correct, but the evaluation steps

 doi:10.20944/preprints202509.0445.v1

	were wrong. Do it again. Still found the wrong answer.
Perplexity	Found the wrong answer. As x -> -oo, then e^(-x) -> oo. It's correct. So (1-x)e^(-oo) -> 0. It's wrong. **x -> -oo, then (1-x)e^(-x) -> -oo. This is correct.** New prompt: Integration was correct, but the evaluation steps were wrong. Do it again. Still found the wrong answer.

F-10) Use the Integral test to determine the convergence or divergence of the series: summation from n=1 to infinity 3n/(n^2+4). Show all the steps.

Chat GPT 4o	Found the correct answer: The series diverges.
Gemini Advanced with 1.5 Pro	Found the correct answer: The series diverges.
Copilot Pro	Found the correct answer: The series diverges.
Claude 3.5 Sonnet	Found the correct answer: The series diverges.
Meta AI	Found the correct answer: The series diverges.
Mistral AI	Found the correct answer: The series diverges.
Perplexity	Found the correct answer: The series diverges.

F-11) Use the Limit Comparison test to determine the convergence or divergence of the series: summation from n=1 to infinity (-4n^3-2n+3)/(-5n^7+3n^4-6). Show all the steps.

Chat GPT 4o	Found the correct answer: The series converges.
Gemini Advanced with 1.5 Pro	Found the correct answer: The series converges.
Copilot Pro	Found the correct answer: The series converges.
Claude 3.5 Sonnet	Found the correct answer: The series converges.
Meta AI	Found the correct answer: The series converges.
Mistral AI	Found the correct answer: The series converges.
Perplexity	Found the correct answer: The series converges.

F-12) Determine the convergence or divergence of the series: summation from n=1 to infinity

 doi:10.20944/preprints202509.0445.v1

[((-1)^(n+1))n]/(n^2+3). Show all the steps.

Chat GPT 4o	Found the correct answer: The series converges.
Gemini Advanced with 1.5 Pro	Found the correct answer: The series converges.
Copilot Pro	Found the correct answer: The series converges.
Claude 3.5 Sonnet	Found the correct answer: The series converges.
Meta AI	Found the correct answer: The series converges.
Mistral AI	Found the correct answer: The series converges.
Perplexity	Found the correct answer: The series converges.

F-13) Use the Ratio Test to determine the convergence or divergence of the series. If the Ratio Test is inconclusive, determine the convergence or divergence of the series using other methods: summation from n=0 to infinity 2^n/(n+2)!. Show all the steps.

Chat GPT 4o	Found the correct answer: The series converges
Gemini Advanced with 1.5 Pro	Found the correct answer: The series converges
Copilot Pro	Found the correct answer: The series converges
Claude 3.5 Sonnet	Found the correct answer: The series converges
Meta AI	Found the correct answer: The series converges
Mistral AI	Found the correct answer: The series converges
Perplexity	Found the correct answer: The series converges

F-14) Find the n^(th) Taylor polynomial for the function, centered at c=-2: f(x)=1/x^2, n=4. Show all the steps.

Chat GPT 4o	Found the correct answer: p(x)=(1/4) + (x+2)/4 + 3(x+2)^2/16 + (x+2)^3/8 + 5(x+2)^4/64.
Gemini Advanced with 1.5 Pro	Found the correct answer: p(x)=(1/4) + (x+2)/4 + 3(x+2)^2/16 + (x+2)^3/8 + 5(x+2)^4/64.
Copilot Pro	Found the correct answer: p(x)=(1/4) + (x+2)/4 + 3(x+2)^2/16 + (x+2)^3/8 + 5(x+2)^4/64.
Claude 3.5 Sonnet	Found the correct answer: p(x)=(1/4) + (x+2)/4 + 3(x+2)^2/16 +

 doi:10.20944/preprints202509.0445.v1

	(x+2)^3/8 + 5(x+2)^4/64.
Meta AI	Found the correct answer: p(x)=(1/4) + (x+2)/4 + 3(x+2)^2/16 + (x+2)^3/8 + 5(x+2)^4/64.
Mistral AI	Found the correct answer: p(x)=(1/4) + (x+2)/4 + 3(x+2)^2/16 + (x+2)^3/8 + 5(x+2)^4/64.
Perplexity	Found the wrong answer. New prompt: The coefficients of (x+2), (x+2)^3, and (x+2)^4 were wrong. Solve it again. Still found the wrong answer.

F-15) Find the radius of the convergence of the power series: summation from n=0 to infinity ((-1)^n)(x^n)/5^n. Show all the steps.

Chat GPT 4o	Found the correct answer: R=5.
Gemini Advanced with 1.5 Pro	Found the correct answer: R=5.
Copilot Pro	Found the correct answer: R=5.
Claude 3.5 Sonnet	Found the correct answer: R=5.
Meta AI	Found the correct answer: R=5.
Mistral AI	Found the correct answer: R=5.
Perplexity	Found the correct answer: R=5.

1st Test: 9/20/24

The test results from September 20, 2024, show exceptionally strong overall performance across all seven models, with a remarkable average score of 85.7 points out of 90 (95.2%, A grade). Four models—Chat GPT 4o, Gemini Advanced with 1.5 Pro, Mistral AI, and Perplexity—achieved perfect scores of 90 points. Copilot Pro scored 85 points (A), Meta AI earned 80 points (B+), and Claude 3.5 Sonnet received 75 points (B). An analysis of individual problems indicates that models solved problems 1-1 through 1-3 and 1-6 through 1-9 correctly without needing follow-up prompts. This performance demonstrates their mastery of fundamental integration techniques, including basic indefinite integrals, differential equations with initial conditions, summation evaluation, and various substitution methods. However, problems 1-4 and 1-5 identified some weaknesses in the computational abilities of specific models. Problem 1-4, which required evaluating the definite integral of $(x^2-1)/x^3$ from -1 to 1, proved particularly challenging for Claude 3.5 Sonnet, as it failed to arrive at the correct answer of 2/3 even after follow-up prompting. It was the only model unable to solve this problem. Problem 1-5 involved integrating $40(x-x^2)/x^{(1/3)}$ from -1 to 1 and revealed computational errors in the initial attempts of both Copilot Pro and Meta AI. While Copilot Pro successfully corrected its error after prompting, Meta AI continued to make the same mistake despite guidance. The strong performance on this initial test, with six out of seven models scoring in the A or B+ range, suggests that current language models have effectively mastered basic integration

 doi:10.20944/preprints202509.0445.v1

techniques. These include applications of the power rule, u-substitution, integration by substitution with fractional exponents, and handling improper integrals with removable discontinuities. This exceptional baseline performance indicates that fundamental calculus integration skills are well within the capabilities of modern language models, setting a high standard for evaluating their performance on more complex problems in future tests.

1-1) Find the indefinite integral: integration (15x-4)/(x^(1/3)). Show all the steps.

Chat GPT 4o	Found the correct answer: 9x^(5/3)-6x^(2/3)+C.
Gemini Advanced with 1.5 Pro	Found the correct answer: 9x^(5/3)-6x^(2/3)+C.
Copilot Pro	Found the correct answer: 9x^(5/3)-6x^(2/3)+C.
Claude 3.5 Sonnet	Found the correct answer: 9x^(5/3)-6x^(2/3)+C.
Meta AI	Found the correct answer: 9x^(5/3)-6x^(2/3)+C.
Mistral AI	Found the correct answer: 9x^(5/3)-6x^(2/3)+C.
Perplexity	Found the correct answer: 9x^(5/3)-6x^(2/3)+C.

1-2) Find the particular solution of the differential equation that satisfies the initial conditions: f''(x)=5/(x^2), f'(-1)=3, f(1)=-5. Show all the steps.

Chat GPT 4o	Found the correct answer: -5ln|x|-2x-3.
Gemini Advanced with 1.5 Pro	Found the correct answer: -5ln|x|-2x-3.
Copilot Pro	Found the correct answer: -5ln|x|-2x-3.
Claude 3.5 Sonnet	Found the correct answer: -5ln|x|-2x-3.
Meta AI	Found the correct answer: -5ln|x|-2x-3.
Mistral AI	Found the correct answer: -5ln|x|-2x-3.
Perplexity	Found the correct answer: -5ln|x|-2x-3.

1-3) Evaluate: Summation from i=1 to i=8 2i(i-3)^2. Show all the steps.

Chat GPT 4o	Found the correct answer: 792.
Gemini Advanced with 1.5 Pro	Found the correct answer: 792.
Copilot Pro	Found the correct answer: 792.

 doi:10.20944/preprints202509.0445.v1

Claude 3.5 Sonnet	Found the correct answer: 792.
Meta AI	Found the correct answer: 792.
Mistral AI	Found the correct answer: 792.
Perplexity	Found the correct answer: 792.

1-4) Evaluate the definite integral: integration from -1 to 1 (x^2-1/x^3). Show all the steps.

Chat GPT 4o	Found the correct answer: (2/3).
Gemini Advanced with 1.5 Pro	Found the correct answer: (2/3).
Copilot Pro	Found the correct answer: (2/3).
Claude 3.5 Sonnet	Found the wrong answer. New prompt: It's wrong. Solve it again. Still found the wrong answer.
Meta AI	Found the correct answer: (2/3).
Mistral AI	Found the correct answer: (2/3).
Perplexity	Found the correct answer: (2/3).

1-5) Evaluate the definite integral: integration from -1 to 1 40(x-x^2)/x^(1/3). Show all the steps.

Chat GPT 4o	Found the correct answer: 48.
Gemini Advanced with 1.5 Pro	Found the correct answer: 48.
Copilot Pro	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Found the correct answer: 48.
Claude 3.5 Sonnet	Found the correct answer: 48.
Meta AI	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.
Mistral AI	Found the correct answer: 48.
Perplexity	Found the correct answer: 48.

 doi:10.20944/preprints202509.0445.v1

1-6) Find the indefinite integral: integration 6(x^2)((1-x^3)^(1/2)). Show all the steps.

Chat GPT 4o	Found the correct answer: (-4/3)(1-x^3)^(3/2)+C.
Gemini Advanced with 1.5 Pro	Found the correct answer: (-4/3)(1-x^3)^(3/2)+C.
Copilot Pro	Found the correct answer: (-4/3)(1-x^3)^(3/2)+C.
Claude 3.5 Sonnet	Found the correct answer: (-4/3)(1-x^3)^(3/2)+C.
Meta AI	Found the correct answer: (-4/3)(1-x^3)^(3/2)+C.
Mistral AI	Found the correct answer: (-4/3)(1-x^3)^(3/2)+C.
Perplexity	Found the correct answer: (-4/3)(1-x^3)^(3/2)+C.

1-7) Evaluate the definite integral: integration from 2^(1/2) to 6^(1/2) (x)(e^(-(x^(2)/2))). Show all the steps.

Chat GPT 4o	Found the correct answer: (1/e)-(1/e^3).
Gemini Advanced with 1.5 Pro	Found the correct answer: (1/e)-(1/e^3).
Copilot Pro	Found the correct answer: (1/e)-(1/e^3).
Claude 3.5 Sonnet	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Found the correct answer: (1/e)-(1/e^3).
Meta AI	Found the correct answer: (1/e)-(1/e^3).
Mistral AI	Found the correct answer: (1/e)-(1/e^3).
Perplexity	Found the correct answer: (1/e)-(1/e^3).

1-8) Find the indefinite integral: integration (6+(1/(x^3)))^5(1/(x^4)). Show all the steps.

Chat GPT 4o	Found the correct answer: (-1/18)(6+(1/x^3))^(6)+C.
Gemini Advanced with 1.5 Pro	Found the correct answer: (-1/18)(6+(1/x^3))^(6)+C.
Copilot Pro	Found the correct answer: (-1/18)(6+(1/x^3))^(6)+C.
Claude 3.5 Sonnet	Found the correct answer: (-1/18)(6+(1/x^3))^(6)+C.
Meta AI	Found the correct answer: (-1/18)(6+(1/x^3))^(6)+C.

 doi:10.20944/preprints202509.0445.v1

Mistral AI	Found the correct answer: (-1/18)(6+(1/x^3))^(6)+C.
Perplexity	Found the correct answer: (-1/18)(6+(1/x^3))^(6)+C.

1-9) Evaluate the definite integral: integration from 4 to 1 (-1)/(x^(1/2)(1+x^(1/2))^2). Show all the steps.

Chat GPT 4o	Found the correct answer: (1/3).
Gemini Advanced with 1.5 Pro	Found the correct answer: (1/3).
Copilot Pro	Found the correct answer: (1/3).
Claude 3.5 Sonnet	Found the correct answer: (1/3).
Meta AI	Found the correct answer: (1/3).
Mistral AI	Found the correct answer: (1/3).
Perplexity	Found the correct answer: (1/3).

2nd Test: 10/18/24

The results from the second test, conducted on October 18, 2024, reveal significant performance disparities among the models. The class average dropped to 63.6 points out of 80 (79.5%, C+), marking the lowest average performance across all four assessments. Four models—Chat GPT 4o, Gemini Advanced with 1.5 Pro, Copilot Pro, and Mistral AI—achieved relatively strong scores of 70 points each (B+). In contrast, Claude 3.5 Sonnet and Meta AI scored 60 points (C), while Perplexity dramatically underperformed with only 45 points (F). The test consisted of eight problems that exposed critical weaknesses in various areas of mathematical reasoning. Problem 2-1, which involved L'Hôpital's rule for evaluating a limit as x approaches negative infinity, proved challenging for most models. Only Gemini Advanced and Copilot Pro provided correct initial solutions, while Chat GPT 4o, Claude 3.5 Sonnet, and Meta AI needed a follow-up prompt to arrive at the correct answer. Mistral AI and Perplexity failed to reach the correct solution, even after additional guidance. Problem 2-6, which required the integration of $\sec^3(4x)\tan^3(4x)$, emerged as particularly difficult. Only Chat GPT 4o and Mistral AI solved it correctly on their first attempts. Copilot Pro and Perplexity needed assistance to find the solution, while Gemini Advanced, Claude 3.5 Sonnet, and Meta AI could not arrive at the correct answer despite receiving specific hints about substitution methods. The improper integral in problem 2-8 further highlighted the differences among the models, with Perplexity unable to correctly evaluate convergence even after prompting, contributing to its failing grade. Problems involving integration by parts (2-4), trigonometric integration (2-5), and partial fractions (2-7) showed more consistent performance; however, Perplexity struggled across multiple problem types. The sharp decline in average performance from the first test's 95.2% to this test's 79.5% indicates that while the models have mastered basic integration techniques, they face significant challenges with more complex applications requiring sophisticated substitution strategies, limit evaluation at infinity, and advanced trigonometric manipulations. The 25-point spread between the highest and lowest scores reveals substantial variations in how well different models handle increased mathematical complexity.

2-1) Evaluate the limit, using L'Hôpital's Rule if necessary: limit x goes negative infinity (-e^(x^2))/(1-x^3). Show all the steps.

 doi:10.20944/preprints202509.0445.v1

Chat GPT 4o	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Found the correct answer: - oo.
Gemini Advanced with 1.5 Pro	Found the correct answer: - oo.
Copilot Pro	Found the correct answer: - oo.
Claude 3.5 Sonnet	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Found the correct answer: - oo.
Meta AI	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Found the correct answer: - oo.
Mistral AI	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.
Perplexity	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.

2-2) Find the indefinite integral: integration (x^3-4x^2-4x+20)/(x^2-5). Show all the steps.

Chat GPT 4o	Found the correct answer: (x^2/2)-4x+(ln|x^2-5|/2)+C.
Gemini Advanced with 1.5 Pro	Found the correct answer: (x^2/2)-4x+(ln|x^2-5|/2)+C.
Copilot Pro	Found the correct answer: (x^2/2)-4x+(ln|x^2-5|/2)+C.
Claude 3.5 Sonnet	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Found the correct answer: (x^2/2)-4x+(ln|x^2-5|/2)+C.
Meta AI	Found the correct answer: (x^2/2)-4x+(ln|x^2-5|/2)+C.
Mistral AI	Found the correct answer: (x^2/2)-4x+(ln|x^2-5|/2)+C.
Perplexity	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer. (It used partial fractions.)

2-3) Find the indefinite integral: integration ln(x^2)/x^3. Show all the steps.

 doi:10.20944/preprints202509.0445.v1

Chat GPT 4o	Found the correct answer: -(ln(x)/x^2)-1/(2x^2)+C.
Gemini Advanced with 1.5 Pro	Found the correct answer: -(ln(x)/x^2)-1/(2x^2)+C.
Copilot Pro	Found the correct answer: -(ln(x)/x^2)-1/(2x^2)+C.
Claude 3.5 Sonnet	Found the correct answer: -(ln(x)/x^2)-1/(2x^2)+C.
Meta AI	Found the correct answer: -(ln(x)/x^2)-1/(2x^2)+C.
Mistral AI	Found the correct answer: -(ln(x)/x^2)-1/(2x^2)+C.
Perplexity	Found the correct answer: -(ln(x)/x^2)-1/(2x^2)+C.

2-4) Find the indefinite integral: integration (-3x)/(e^(2x)). Show all the steps.

Chat GPT 4o	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Found the correct answer: (3/(4e^(2x)))(2x+1)+C.
Gemini Advanced with 1.5 Pro	Found the correct answer: (3/(4e^(2x)))(2x+1)+C.
Copilot Pro	Found the correct answer: (3/(4e^(2x)))(2x+1)+C.
Claude 3.5 Sonnet	Found the correct answer: (3/(4e^(2x)))(2x+1)+C.
Meta AI	Found the correct answer: (3/(4e^(2x)))(2x+1)+C.
Mistral AI	Found the correct answer: (3/(4e^(2x)))(2x+1)+C.
Perplexity	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.

2-5) Find the indefinite integral: integration [((sin(2x))^4)((cos(2x))^3)]. Show all the steps.

Chat GPT 4o	Found the correct answer: ((1/10)(sin2x)^5)-((1/14)(sin2x)^7)+C.
Gemini Advanced with 1.5 Pro	Found the correct answer: ((1/10)(sin2x)^5)-((1/14)(sin2x)^7)+C.
Copilot Pro	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.
Claude 3.5 Sonnet	Found the correct answer: ((1/10)(sin2x)^5)-((1/14)(sin2x)^7)+C.

 doi:10.20944/preprints202509.0445.v1

Meta AI	Found the correct answer: ((1/10)(sin2x)^5)-((1/14)(sin2x)^7)+C.
Mistral AI	Found the correct answer: ((1/10)(sin2x)^5)-((1/14)(sin2x)^7)+C.
Perplexity	Found the correct answer: ((1/10)(sin2x)^5)-((1/14)(sin2x)^7)+C.

2-6) Find the indefinite integral: integration [((sec(4x))^3)((tan(4x))^3)]. Show all the steps. They found different types of answers. We need to verify each problem separately.

Chat GPT 4o	Found the correct answer: ((sec(4x))^5)/20)-((sec(4x))^5)/12)+C.
Gemini Advanced with 1.5 Pro	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.
Copilot Pro	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Found the correct answer: ((sec(4x))^5)/20)-((sec(4x))^5)/12)+C.
Claude 3.5 Sonnet	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.
Meta AI	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.
Mistral AI	Found the correct answer: (1/(20(cos(4x))^5))-(1/(12(cos(4x))^5))+C.
Perplexity	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Found the correct answer: ((sec(4x))^5)/20)-((sec(4x))^5)/12)+C.

2-7) Use the partial fraction to find the indefinite integral: integration (x^2-4x+2)/(x^3-2x^2+x). Show all the steps.

Chat GPT 4o	Found the correct answer: 2ln|x|-ln|x-1|+(1/(x-1))+C.
Gemini Advanced with 1.5 Pro	Found the correct answer: 2ln|x|-ln|x-1|+(1/(x-1))+C.
Copilot Pro	Found the correct answer: 2ln|x|-ln|x-1|+(1/(x-1))+C.
Claude 3.5 Sonnet	Found the correct answer: 2ln|x|-ln|x-1|+(1/(x-1))+C.
Meta AI	Found the wrong answer.

 doi:10.20944/preprints202509.0445.v1

	New prompt: The answer is wrong. Solve the problem again. Found the correct answer: 2ln|x|-ln|x-1|+(1/(x-1))+C.
Mistral AI	Found the correct answer: 2ln|x|-ln|x-1|+(1/(x-1))+C.
Perplexity	Found the correct answer: 2ln|x|-ln|x-1|+(1/(x-1))+C.

2-8) Determine whether the improper integral diverges or converges. Evaluate the definite integral if it converges: integration from 1 to infinity (1-x)e^(-x). Show all the steps.

Chat GPT 4o	Found the correct answer: (-1/e).
Gemini Advanced with 1.5 Pro	Found the correct answer: (-1/e).
Copilot Pro	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Found the correct answer: (-1/e).
Claude 3.5 Sonnet	Found the correct answer: (-1/e).
Meta AI	Found the correct answer: (-1/e).
Mistral AI	Found the correct answer: (-1/e).
Perplexity	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.

3rd Test: 11/22/24

The results of the third test conducted on November 22, 2024, focused on applications of integration and series convergence, yielding a class average of 79.3 points out of 100 (79.3%, C+). Chat GPT 4o and Gemini Advanced with 1.5 Pro led with scores of 85 points (B), followed by Copilot Pro, Claude 3.5 Sonnet, and Mistral AI, each scoring 80 points (B-). Perplexity achieved 75 points (B), while Meta AI trailed with 70 points (C-). The test structure revealed a significant contrast in performance between geometric applications and series convergence problems. For example, in problems 3-1 through 3-3, which involved determining areas between curves and volumes of revolution, most models exhibited severe limitations in geometric reasoning. In problem 3-1, which required finding the area between the curves defined by $y = -x^2 - 2x + 2$ and $y = x + 2$, most models ultimately reached the correct answer of 9/2. However, many had initially set up the integration incorrectly, confusing which function was above the other, and needed multiple prompts to correct their approach. Problem 3-2 was even more challenging. Only Chat GPT 4o found the correct area of 253/12 initially, while Gemini Advanced, Claude 3.5 Sonnet, Meta AI, and Perplexity failed to arrive at the correct answer, even with guidance on using two separate integrations and checking the positions of the curves. In problem 3-3, which involved calculating the volume of revolution about the line $y = 2$, the geometric reasoning deficits were even more pronounced. Claude 3.5 Sonnet was the only model to solve it correctly without prompting, while Chat GPT 4o, Copilot Pro, Meta AI, Mistral AI, and Perplexity all failed to find the solution, despite receiving multiple hints about the proper setup using the washer method. In contrast, problems 3-4 through 3-10, which assessed series convergence through various

 doi:10.20944/preprints202509.0445.v1

tests (including geometric series, the Integral Test, the Limit Comparison Test, the Alternating Series Test, and the Ratio Test), demonstrated perfect performance from all seven models, with no follow-up prompting needed. This stark contrast—with failure rates exceeding 70% on geometric problems and 100% success on series convergence—suggests that while language models (LLMs) have thoroughly grasped algorithmic convergence tests and formal mathematical procedures, they struggle with spatial visualization and geometric intuition necessary for setting up area and volume integrals. This highlights a fundamental limitation in their current training, which emphasizes symbolic manipulation over geometric understanding.

3-1) Find the area of the region bounded by the graphs of the equations: y=-x^2-2x+2, y=x+2. Show all the steps.

Chat GPT 4o	Found the correct answer: 9/2, but some steps are inaccurate because they switched the upper and lower functions. New prompt: The answer is correct, but some steps are not correct. Solve the problem again. Found the correct answer, but some steps are still not correct. New prompt: The answer is wrong. Solve the problem again. Found the correct answer: 9/2.
Gemini Advanced with 1.5 Pro	Found the correct answer: 9/2.
Copilot Pro	Found the correct answer: 9/2, but some steps are inaccurate because they switched the upper and lower functions. New prompt: The answer is correct, but some steps are not correct. Solve the problem again. Found the correct answer, but some steps are still not correct. New prompt: The answer is wrong. Solve the problem again. Found the correct answer, but some steps are still not correct.
Claude 3.5 Sonnet	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.
Meta AI	Found the correct answer: 9/2, but some steps are inaccurate because they switched the upper and lower functions. New prompt: The answer is correct, but some steps are not correct. Solve the problem again. Found the correct answer, but some steps are still not correct. New prompt: The answer is wrong. Solve the problem again. Found the correct answer, but some steps are still not correct.
Mistral AI (5/10)	Found the correct answer: 9/2, but some steps are inaccurate because they switched the upper and lower functions. New prompt: The answer is correct, but some steps are not correct. Solve the problem again. Found the correct answer, but some steps are still not correct.

 doi:10.20944/preprints202509.0445.v1

	New prompt: The answer is wrong. Solve the problem again. Found the correct answer, but some steps are still not correct.
Perplexity	Found the correct answer: 9/2, but some steps are inaccurate because they switched the upper and lower functions. New prompt: The answer is correct, but some steps are not correct. Solve the problem again. Found the correct answer, but some steps are still not correct. New prompt: The answer is wrong. Solve the problem again. Found the correct answer: 9/2.

3-2) Find the area of the region bounded by the graphs of the equations: y=x^3-3x^2-4x, y=-2x^2+2x. Show all the steps.

Chat GPT 4o	Found the correct answer: 253/12.
Gemini Advanced with 1.5 Pro	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.
Copilot Pro	Found the correct answer: 253/12, but some steps are inaccurate because they switched the upper and lower functions. New prompt: The answer is correct, but some steps are not correct. Solve the problem again. Found the correct answer, but some steps are still not correct. New prompt: The answer is wrong. Solve the problem again. Found the correct answer, but some steps are still not correct.
Claude 3.5 Sonnet	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.
Meta AI	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.
Mistral AI (5/10)	Found the correct answer: 253/12, but some steps are inaccurate because they switched the upper and lower functions. New prompt: The answer is correct, but some steps are not correct. Solve the problem again. Found the correct answer, but some steps are still not correct. New prompt: The answer is wrong. Solve the problem again. Found the correct answer, but some steps are still not correct.
Perplexity	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again.

 doi:10.20944/preprints202509.0445.v1

	Still found the wrong answer.

3-3) Find the volume of the solid generated by revolving the region bounded by the graphs of y = - x^2 + 2 x + 5, y = - x + 5 about the line y=2. Show all the steps.

Chat GPT 4o	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.
Gemini Advanced with 1.5 Pro	Found the correct answer: (108/5)(3.14), but some steps are inaccurate because they switched the upper and lower functions. New prompt: The answer is correct, but some steps are not correct. Solve the problem again. Found the correct answer, but some steps are still not correct. New prompt: The answer is wrong. Solve the problem again. Found the correct answer, but some steps are still not correct.
Copilot Pro	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.
Claude 3.5 Sonnet	Found the correct answer: (108/5)(3.14).
Meta AI	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.
Mistral AI	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.
Perplexity	Found the wrong answer. New prompt: The answer is wrong. Solve the problem again. Still found the wrong answer.

3-4) Determine the convergence or divergence of the sequence with the given n^th term. If the sequence converges, find its limit: a_n=n^3/((3^n)-1). Show all the steps.

Chat GPT 4o	Found the correct answer: Converges to 0.
Gemini Advanced with 1.5 Pro	Found the correct answer: Converges to 0.
Copilot Pro	Found the correct answer: Converges to 0.
Claude 3.5 Sonnet	Found the correct answer: Converges to 0.

 doi:10.20944/preprints202509.0445.v1

Meta AI	Found the correct answer: Converges to 0.
Mistral AI	Found the correct answer: Converges to 0.
Perplexity	Found the correct answer: Converges to 0.

3-5) Determine the convergence or divergence of the series: summation from n=1 to n=positive infinity ((4^n)+3)/(4^(n+1)). Show all the steps.

Chat GPT 4o	Found the correct answer: Diverges.
Gemini Advanced with 1.5 Pro	Found the correct answer: Diverges.
Copilot Pro	Found the correct answer: Diverges.
Claude 3.5 Sonnet	Found the correct answer: Diverges.
Meta AI	Found the correct answer: Diverges.
Mistral AI	Found the correct answer: Diverges.
Perplexity	Found the correct answer: Diverges.

3-6) Use the Integral Test to determine the convergence or divergence of the series: summation from n=1 to n=positive infinity n^2/(n^3+1). Show all the steps.

Chat GPT 4o	Found the correct answer: Diverges.
Gemini Advanced with 1.5 Pro	Found the correct answer: Diverges.
Copilot Pro	Found the correct answer: Diverges.
Claude 3.5 Sonnet	Found the correct answer: Diverges.
Meta AI	Found the correct answer: Diverges.
Mistral AI	Found the correct answer: Diverges.
Perplexity	Found the correct answer: Diverges.

3-7) Use the Limit Comparison Test to determine the convergence or divergence of the series: summation from n=1 to n=positive infinity (2n^2-1)/(3n^5+4n^3+1). Show all the steps.

Chat GPT 4o	Found the correct answer: Converges.
Gemini Advanced with 1.5	Found the correct answer: Converges.

 doi:10.20944/preprints202509.0445.v1

Pro	
Copilot Pro	Found the correct answer: Converges.
Claude 3.5 Sonnet	Found the correct answer: Converges.
Meta AI	Found the correct answer: Converges.
Mistral AI	Found the correct answer: Converges.
Perplexity	Found the correct answer: Converges.

3-8) Determine the convergence or divergence of the series: summation from n=1 to n=positive infinity ((-1)^n)n)/ln(n+1). Show all the steps.

Chat GPT 4o	Found the correct answer: Diverges.
Gemini Advanced with 1.5 Pro	Found the correct answer: Diverges.
Copilot Pro	Found the correct answer: Diverges.
Claude 3.5 Sonnet	Found the correct answer: Diverges.
Meta AI	Found the correct answer: Diverges.
Mistral AI	Found the correct answer: Diverges.
Perplexity	Found the correct answer: Diverges.

3-9) Determine the convergence or divergence of the series: summation from n=1 to n=positive infinity ((-1)^n)(3^(n+1))/(n+1)!. Show all the steps.

Chat GPT 4o	Found the correct answer: Converges.
Gemini Advanced with 1.5 Pro	Found the correct answer: Converges.
Copilot Pro	Found the correct answer: Converges.
Claude 3.5 Sonnet	Found the correct answer: Converges.
Meta AI	Found the correct answer: Converges.
Mistral AI	Found the correct answer: Converges.
Perplexity	Found the correct answer: Converges.

3-10) Determine the convergence or divergence of the series: summation from n=1 to n=positive infinity (n+1)^4/3^(n+2). Show all the steps.

 doi:10.20944/preprints202509.0445.v1

Chat GPT 4o	Found the correct answer: Converges.
Gemini Advanced with 1.5 Pro	Found the correct answer: Converges.
Copilot Pro	Found the correct answer: Converges.
Claude 3.5 Sonnet	Found the correct answer: Converges.
Meta AI	Found the correct answer: Converges.
Mistral AI	Found the correct answer: Converges.
Perplexity	Found the correct answer: Converges.

3.3. Response to Follow-up Prompt

A notable aspect of the results is the varying ability of different models to correct errors when given follow-up prompts. Gemini Advanced with 1.5 pro exhibited exceptional error correction capabilities, successfully incorporating feedback to arrive at correct answers in most instances. Both Chat GPT 4o and Copilot Pro also demonstrated strong adaptability to feedback.

In contrast, certain models faced persistent difficulties in specific problem areas despite targeted prompts. For example, Perplexity consistently struggled with Taylor polynomial expansions (problem F-14) and improper integrals (problem F-9), failing to correct errors even after multiple follow-up prompts.

The follow-up prompting methodology revealed that errors typically fell into three categories: computational mistakes (e.g., incorrectly expanding expressions), conceptual misunderstandings (e.g., improper setup of area calculations between curves), and errors in technique application (e.g., incorrect substitution in trigonometric integrals). Most models could easily correct computational errors with simple prompts, but conceptual misunderstandings often persist despite explicit guidance.

3.4. Problem Type Analysis

Analyzing performance by problem type reveals distinct patterns in the mathematical capabilities of large language models (LLMs):

1. Basic Indefinite Integration: All models demonstrated strong proficiency, achieving near-perfect results on problems involving polynomial, rational, and simple transcendental functions.

2. Trigonometric Integration: Performance varied significantly among models. Complex trigonometric substitutions (problems 2-5, 2-6, F-7) exposed limitations in several of them.

3. Area and Volume Applications: Problems in this category (3-1, 3-2, 3-3, F-5, F-6) consistently challenged all models. Many required multiple follow-ups prompts to accurately configure integration regions and apply the appropriate methods.

4. Improper Integrals: Problems involving the convergence analysis of improper integrals (2-8, F-9) showed significant variation across models, highlighting differences in their understanding of limits and infinity.

5. Series Convergence: All models performed exceptionally well on problems F-10 through F-13, demonstrating strong capabilities in applying various convergence tests.

6. Taylor Polynomials: Problem F-14, which involved deriving a Taylor polynomial, revealed computational limitations in most models, with several making errors in calculating derivatives and coefficients.

 doi:10.20944/preprints202509.0445.v1

These results provide valuable insights into the current state of mathematical reasoning in LLMs, highlighting both impressive strengths in formal calculus techniques and ongoing limitations in geometric intuition and complex problem setup.

4. Discussion

4.1. Correct Solutions

Chat GPT 4o

Original command:	**Follow-up commands:**
3-3) Find the volume of the solid generated by revolving the region bounded by the graphs of y = - x^2 + 2x + 5, y = -x + 5 about the line y = 2. Show all the steps. Still found the wrong answer.	3-3) **(-x^2+2x+3)(-x^2+2x+3).** The answer is x^4-4x^3-2x^2+12x+9. **(-x^2+2x+3)(-x^2+2x+3)-(-x+3)(-x+3).** The answer is x^4-4x^3-3x^2+18x. **integration from 0 to 3 x^4-4x^3-3x^2+18x.** The answer is (108/5). **Found the correct answer: (108/5)(3.14).**
F-1) Evaluate the definite integral: integration from -1 to 1 [(x-x^2)/(5(x^(1/5)))]. Show all the steps. Calculate each operation separately. Still found the wrong answer.	F-1) **Integrate x^(4/5)-x^(9/5).** Show all the steps. (1/5)[(5/9)x^(9/5)-(5/14)(x^(14/5)]. **f(x)=x^9, g(x)=x^14, f(-1)=? and g(-1)=?** f(-1)=-1, g(-1)=1. **(-1)^(1/5)=? 1^(1/5)=?** (-1)^(1/5)=(-1), 1^(1/5)=(1). **(1/5)(5/9)[1-(-1)]=?** **Found the correct answer: (2/9).**

Gemini Advanced with 1.5 Pro

Original command:	**Follow-up commands:**
2-6) Find the indefinite integral: integration [((sec(4x))^3)((tan(4x))^3)]. Show all the steps. Still found the wrong answer.	2-6) **Let u=sec(4x). Put (sec(4x))^3=(sec(4x)^2(sec(4x)).** Solve the problem again. Show all the steps. **Found the correct answer: [(sec(4x)^5)/20]-[(sec(4x)^3)/12]+C**
3-2) Find the area of the region bounded by the graphs of the equations: y=x^3-3x^2-4x, y=-2x^2+2x. Show all the steps. Still found the wrong answer.	3-2) **Must use two integrations instead of one integration. In each integration, check which curve is above.** Solve the problem again. Show all the steps. The answer is wrong. **(16/3)+(63/4): The answer is wrong. What is the answer for (16/3)+(63/4)?** **Found the correct answer: (253/12).**
3-3) Find the volume of the solid generated by	3-3) **R(x) and r(x) answers are wrong.** Do it

 doi:10.20944/preprints202509.0445.v1

revolving the region bounded by the graphs of y = - x^2 + 2x + 5, y = -x + 5 about the line y = 2. Show all the steps. Still found the wrong answer.	again. Show all the steps. **Found the correct answer: (108/5)(3.14).** * R(x): Outer radius. r(x): Inner radius.
F-9) Determine whether the improper integral diverges or converges. Evaluate the definite integral if it converges: integration from negative infinity to 1 (1-x)e^(-x). Show all the steps. Still found the wrong answer.	F-9) Step 4. Evaluate the limit: limit t -> -infinity, (1/e^(t)) -> 0: The answer is wrong. **limit t -> -infinity, (te^(-t)): Answer is wrong. Do it again. Use the L'Hôpital's rule.** Show all the steps. **Found the correct answer: It diverges to -oo.**

Copilot Pro

Original command:	**Follow-up commands:**
3-1) Find the area of the region bounded by the graphs of the equations: y=-x^2-2x+2, y=x+2. Show all the steps. Found the correct answer, but some steps are still not correct.	3-1) **Area setup was wrong. Check which curve is above.** Solve the problem again. Show all the steps. **Found the correct answer with the correct steps: (9/2).**
3-2) Find the area of the region bounded by the graphs of the equations: y=x^3-3x^2-4x, y=-2x^2+2x. Show all the steps. Still found the wrong answer.	3-2) **Must use two integrations instead of one integration. In each integration, check which curve is above.** Solve the problem again. Show all the steps. The answer is wrong. **In Area 1, y=x^3-3x^2-4x is above y=-2x^2+2x.** **In Area 2, y=-2x^2+2x is above y=x^3-3x^2-4x.** Solve the problem again. Show all the steps. The answer is wrong. **Area 1 answer is wrong. Find the Area 1 answer, again.** **Add Area 1 and Area 2.** **Found the correct answer: (253/12).**
3-3) Find the volume of the solid generated by revolving the region bounded by the graphs of y = - x^2 + 2x + 5, y = -x + 5 about the line y = 2. Show all the steps. Still found the wrong answer.	3-3) **R(x) and r(x) are wrong. They must be switched.** The answer is wrong. **(-x^2+2x+3)(-x^2+2x+3).** The answer is x^4-4x^3-2x^2+12x+9. **(-x^2+2x+3)(-x^2+2x+3)-(-x+3)(-x+3).** The answer is x^4-4x^3-3x^2+18x. **integration from 0 to 3 x^4-4x^3-3x^2+18x.** The answer is (108/5). **Found the correct answer: (108/5)(3.14).**

 doi:10.20944/preprints202509.0445.v1

F-5) Find the area of the region: y=2x^3-3x^2-5x, y=-3x^2+3x. Show all the steps. Still found the wrong answer.	F-5) **Must use two integrations instead of one integration. In each integration, check which curve is above.** Solve the problem again. Show all the steps. **Found the correct answer: 16.**
F-6) Find the volume of the solid generated by revolving the region bounded by the graphs of y = -x^2 + 3x + 6, y = -x + 6 about the line y = 2. Show all the steps. Still found the wrong answer.	F-6) **Use a Washer method.** The answer is wrong. **R(x) and r(x) are wrong. They must be switched.** The answer is wrong. **(-x^2+3x+4)(-x^2+3x+4) is wrong. Do it again.** **Found the correct answer: (384/5)(3.14).**

Claude 3.5 Sonnet

Original command:	**Follow-up commands:**
1-4) Evaluate the definite integral: integration from -1 to 1 (x^2-1/x^3). Show all the steps.	1-4) Use an antiderivative method. **Find the correct answer: (2/3).**
2-6) Find the indefinite integral: integration [((sec(4x))^3)((tan(4x))^3)]. Show all the steps. Still found the wrong answer.	2-6) **Let u=sec(4x). Put (sec(4x))^3=(sec(4x)^2(sec(4x)).** Solve the problem again. Show all the steps. **Found the correct answer: [(sec(4x)^5)/20]-[(sec(4x)^3)/12]+C.**
3-1) Find the area of the region bounded by the graphs of the equations: y=-x^2-2x+2, y=x+2. Show all the steps. Found the correct answer, but some steps are still not correct.	3-1) **Area setup was wrong. Check which curve is above.** Solve the problem again. Show all the steps. The answer is wrong. **Area setup was right, but the answer is wrong.** Solve the problem, agin. Show all the steps. **Found the correct answer with the correct steps: (9/2).**
3-2) Find the area of the region bounded by the graphs of the equations: y=x^3-3x^2-4x, y=-2x^2+2x. Show all the steps. Still found the wrong answer.	3-2) **Must use two integrations instead of one integration. In each integration, check which curve is above.** Solve the problem again. Show all the steps. The answer is wrong. **In step 6, the second integral answer is wrong. Calculate the second integral, again.** The answer is wrong. **(-2x^2+2x)-(x^3-3x^2-4x): the answer is wrong. Calculate it again.** **(-2x^2+2x)-(x^3-3x^2-4x): the answer is right. The second integral answer is still wrong.**

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2025 doi:10.20944/preprints202509.0445.v1

	Integrate it again. The answer is wrong. **(3^3/3) is not 27. Integrate it again.** **Found the correct answer: (253/12).**
F-2) Find the indefinite integral: integration [(-3x^3+2x^2+5x-6)/(x^2-3)]. Show all the steps. Still found the wrong answer.	F-2) **The long division result is wrong.** Show all the steps. The answer is wrong. **Don't use the partial fraction method.** **Found the correct answer: (-3x^2)/2 + 2x - 2ln|x^2-3| +C.**
F-3) Evaluate the definite integral: integration from -1 to 1 [(1+(1/(x^3)))^4(1/(x^4))]. Show all the steps. Still found the wrong answer.	F-3) **Use the substitution method. Let u=(1+(1/x^3)).** Show all the steps. **Found the correct answer: (-32/15).**
F-6) Find the volume of the solid generated by revolving the region bounded by the graphs of y = -x^2 + 3x + 6, y = -x + 6 about the line y = 2. Show all the steps. Still found the wrong answer.	F-6) **(-x^2+3x+4)(-x^2+3x+4)** x^4-6x^3+x^2+24x+16. **Solve the problem again. Show all the steps** The answer is wrong. **Integrate from 0 to 4 x^4-6x^3+32x.** **Found the correct answer: (384/5)(3.14).**
F-9) Determine whether the improper integral diverges or converges. Evaluate the definite integral if it converges: integration from negative infinity to 1 (1-x)e^(-x). Show all the steps. Still found the wrong answer.	F-9) **Integrate (1-x)e^(-x). Must use integration by parts.** Show all the steps. xe^(-x) + C. **f(x)=xe^(x). What is f(1)-f(b)?** (1/e)-(be^(-b)). **As b goes -infinity where does (1/e)-[1/e^(b)] go?** **Found the correct answer: It diverges to -oo.** *** We must prompt each step correctly to get the correct answer.**

Meta AI

Original command:	**Follow-up commands:**
1-5) Evaluate the definite integral: integration from -1 to 1 40(x-x^2)/x^(1/3). Show all the steps. Still found the wrong answer.	1-5) (-1)^(5/3) value is wrong. Solve the problem again. Show all the steps. **Found the correct answer: (48).**
2-6) Find the indefinite integral: integration [((sec(4x))^3)((tan(4x))^3)]. Show all the steps. Still found the wrong answer.	2-6) **Let u=sec(4x). Put (sec(4x))^3=(sec(4x)^2(sec(4x)). Replace (tan(4x))^2=((sec(4x)^2)-1).** Solve the problem again. Show all the steps. **Found the correct answer: [(sec(4x)^5)/20]-[(sec(4x)^3)/12]+C.**

 doi:10.20944/preprints202509.0445.v1

3-1) Find the area of the region bounded by the graphs of the equations: y=-x^2-2x+2, y=x+2. Show all the steps. Found the correct answer, but some steps are still not correct.	3-1) **Area setup was wrong. Check which curve is above.** Solve the problem again. Show all the steps. The answer is wrong. **The area setup was right, but the answer was wrong.** Solve the problem, again. Show all the steps. The answer is wrong. **Calculate: (((-3)^3)/3)+((3(-3)^2)/2)** **Found the correct answer with the correct steps: (9/2).**
3-2) Find the area of the region bounded by the graphs of the equations: y=x^3-3x^2-4x, y=-2x^2+2x. Show all the steps. Still found the wrong answer.	3-2) **In the first integration, the upper function and lower function setup were wrong. Do the first integration again.** Show all the steps. The answer is wrong. **(x^3-3x^2-4x)-(-2x^2+2x): the answer is wrong. Calculate it again.** **Found the correct answer: (253/12).**
3-3) Find the volume of the solid generated by revolving the region bounded by the graphs of y = - x^2 + 2x + 5, y = -x + 5 about the line y = 2. Show all the steps. Still found the wrong answer.	3-3) **(-x^2+2x+3)(-x^2+2x+3)** x^4-4x^3-2x^2+12x+9. **Solve the problem again. Show all the steps.** **Found the correct answer: (108/5)(3.14).**
F-3) Evaluate the definite integral: integration from -1 to 1 [(1+(1/(x^3)))^4(1/(x^4))]. Show all the steps. Still found the wrong answer.	F-3) **Use the substitution method.** Show all the steps. **Found the correct answer: (-32/15).**
F-5) Find the area of the region: y=2x^3-3x^2-5x, y=-3x^2+3x. Show all the steps. Still found the wrong answer.	F-5) **Must use two integrations instead of one integration. In each integration, check which curve is above.** Solve the problem again. Show all the steps. The final answer is wrong. * In Area A1, (2x^3-3x^2-5x)-(-3x^2+3x) answer is wrong. In Area A2, (-3x^2+3x)-(2x^3-3x^2-5x) answer is wrong. **In Area A1, (2x^3-3x^2-5x)-(-3x^2+3x).** Area A1 is 8. **In Area A2, (-3x^2+3x)-(2x^3-3x^2-5x).** Area A2 is 8. **A1+A2.** **Found the correct answer: 16.**
F-6) Find the volume of the solid generated by revolving the region bounded by the graphs of	F-6) **(-x^2+3x+4)(-x^2+3x+4)** x^4-6x^3+x^2+24x+16.

 doi:10.20944/preprints202509.0445.v1

y = -x^2 + 3x + 6, y = -x + 6 about the line y = 2. Show all the steps. Still found the wrong answer.	**Solve the problem again. Show all the steps.** The answer is wrong. **Integrate from 0 to 4 x^4-6x^3+32x.** **Found the correct answer: (384/5)(3.14).**
F-9) Determine whether the improper integral diverges or converges. Evaluate the definite integral if it converges: integration from negative infinity to 1 (1-x)e^(-x). Show all the steps.	F-9) **Integrate (1-x)e^(-x). Must use integration by parts.** Show all the steps. xe^(-x) + C. **f(x)=xe^(x). What is f(1)-f(b)?** (1/e)-(be^(-b)). **As b goes -infinity where does (1/e)-[1/(e^(b))] go?** **Found the correct answer: It diverges to -oo.** *** We must prompt each step correctly to get the correct answer.**

Mistral AI

Original command:	**Follow-up commands:**
2-1) Evaluate the limit, using L'Hôpital's Rule if necessary: limit x goes negative infinity (-e^(x^2))/(1-x^3). Show all the steps. Still found the wrong answer.	2-1) Evaluate the limit, using **L'Hôpital's Rule twice**: limit x goes negative infinity (-e^(x^2))/(1-x^3). Show all the steps. **Found the correct answer: -oo.**
2-6) Find the indefinite integral: integration [((sec(4x))^3)((tan(4x))^3)]. Show all the steps. Still found the wrong answer.	2-6) **Let v=sec(u). Put (sec(u))^3=(sec(u))^2(sec(u)).** Solve the problem again. Show all the steps. **Found the correct answer: [(sec(4x)^5)/20]-[(sec(4x)^3)/12]+C.**
3-1) Find the area of the region bounded by the graphs of the equations: y=-x^2-2x+2, y=x+2. Show all the steps. Found the correct answer, but some steps are still not correct.	3-1) **Area setup was wrong. Check which curve is above.** Solve the problem again. Show all the steps. **Found the correct answer with the correct steps: (9/2).**
3-2) Find the area of the region bounded by the graphs of the equations: y=x^3-3x^2-4x, y=-2x^2+2x. Show all the steps. Still found the wrong answer.	3-2) **Must use two integrations instead of one integration. In each integration, check which curve is above.** Solve the problem again. Show all the steps. **Found the correct answer: (253/12).**
3-3) Find the volume of the solid generated by revolving the region bounded by the graphs of y = - x^2 + 2x + 5, y = -x + 5 about the line y = 2. Show all the steps.	3-3) **(-x^2+2x+3)(-x^2+2x+3)-(-x+3)(-x+3).** The answer is x^4-4x^3-3x^2+18x. **integration from 0 to 3 x^4-4x^3-3x^2+18x.** The answer is (108/5).

 doi:10.20944/preprints202509.0445.v1

Still found the wrong answer.	**Found the correct answer: (108/5)(3.14).**
F-6) Find the volume of the solid generated by revolving the region bounded by the graphs of y = -x^2 + 3x + 6, y = -x + 6 about the line y = 2. Show all the steps. Still found the wrong answer.	3-3) **(-x^2+3x+4)(-x^2+3x+4)-(-x+4)(-x+4).** The answer is x^4-6x^3+32x. **integration from 0 to 3 x^4-6x^3+32x.** The answer is (384/5). **Found the correct answer: (384/5)(3.14).**
F-9) Determine whether the improper integral diverges or converges. Evaluate the definite integral if it converges: integration from negative infinity to 1 (1-x)e^(-x). Show all the steps. Still found the wrong answer.	F-9) **Integration (1-x)e^(-x) by parts.** The answer is wrong. **The answer is wrong.** The answer is correct: xe^(-x). limit b goes to (negative infinity) [(1)e^(1)-(b)e^(-b)]. **Found the correct answer: It diverges to -oo.**

Perplexity

Original command:	**Follow-up commands:**
2-1) Evaluate the limit, using L'Hôpital's Rule if necessary: limit x goes negative infinity (-e^(x^2))/(1-x^3). Show all the steps. Still found the wrong answer.	2-1) Evaluate the limit, using **L'Hôpital's Rule twice**: limit x goes negative infinity (-e^(x^2))/(1-x^3). Show all the steps. **Found the correct answer: -oo.**
2-2) Find the indefinite integral: integration (x^3-4x^2-4x+20)/(x^2-5). Show all the steps. Still found the wrong answer.	2-2) After division, use **a u-substitution instead of the partial fraction.** Solve the problem again. Show all the steps. **Found the correct answer: (x^2/2)-4x+(ln|x^2-5|/2)+C.**
2-8) Determine whether the improper integral diverges or converges. Evaluate the definite integral if it converges: integration from 1 to infinity (1-x)e^(-x). Show all the steps. Still found the wrong answer.	2-8) **Integration (1-x)e^(-x) by parts.** **Found the correct answer: (-1/e).**
3-2) Find the area of the region bounded by the graphs of the equations: y=x^3-3x^2-4x, y=-2x^2+2x. Show all the steps. Still found the wrong answer.	3-2) **Must use two integrations instead of one integration. In each integration, check which curve is above.** Solve the problem again. Show all the steps. **Found the correct answer: (253/12).**
3-3) Find the volume of the solid generated by revolving the region bounded by the graphs of	3-3) **R(x) and r(x) are wrong.** R(x) and r(x) are correct, but the final answer is wrong.

 doi:10.20944/preprints202509.0445.v1

y = - x^2 + 2x + 5, y = -x + 5 about the line y = 2. Show all the steps. Still found the wrong answer.	**(-x^2+2x+3)(-x^2+2x+3)** The answer is wrong. **(-x^2+2x+3)(-x^2+2x+3) is wrong** x^4-4x^3-2x^2+12x+9. **Solve the problem again. Show all the steps.** **Found the correct answer: (108/5)(3.14).**
F-5) Find the area of the region: y=2x^3-3x^2-5x, y=-3x^2+3x. Show all the steps. Still found the wrong answer.	F-5) **Must use two integrations instead of one integration. In each integration, check which curve is above.** Solve the problem again. The answer is wrong. **Areas set up, A1 and A2, were wrong. Check which curve is above.** **Found the correct answer: 16.**
F-9) Determine whether the improper integral diverges or converges. Evaluate the definite integral if it converges: integration from negative infinity to 1 (1-x)e^(-x). Show all the steps. Still found the wrong answer.	**F-9) Step 1 answer is wrong. Integration (1-x)e^(-x) by parts.** The answer is wrong. **Step 5 answer is wrong. As t -> -infinity, te^(-t) goes to where? Use L'Hôpital's rule.** **Found the correct answer: It diverges to -oo.**
F-14) Find the n^(th) Taylor polynomial for the function, centered at c=-2: f(x)=1/x^2, n=4. Show all the steps. Still found the wrong answer.	F-14) **The coefficients of (x+2), (x+2)^3, and (x+2)^4 were wrong.** Solve the problem again. Show all the steps. The answer is wrong. **The coefficients of (x+2)^3, and (x+2)^4 were wrong.** Solve the problem again. Show all the steps. The answer is wrong. **The coefficient of (x+2)^4 was wrong.** Solve the problem again. Show all the steps. **Found the correct answer: p(x)=(1/4) + (x+2)/4 + 3(x+2)^2/16 + (x+2)^3/8 + 5(x+2)^4/64.**

4.2. Interpretation of Key Findings

The follow-up prompt data indicates that mathematical reasoning in large language models (LLMs) operates on a spectrum of correctability, with distinct patterns emerging based on the types of errors and the architectures of the models. A particularly striking finding is that Gemini Advanced with 1.5 Pro demonstrated superior error recovery capabilities, successfully incorporating feedback to arrive at correct solutions for complex problems like 2-6, 3-2, and F-9, where other models failed despite receiving similar guidance. This suggests that mathematical competence in LLMs consists of two key components: initial problem-solving ability and adaptive error correction capacity.

Chat GPT 4o exhibited the highest initial accuracy but occasionally required extensive step-by-step prompting, as illustrated in problems 3-3 and F-1. This indicates that even top-performing models can have blind spots in specific mathematical contexts. The data also reveals a clear hierarchy of problem difficulty that extends beyond simple computational complexity. Geometric problems (such as 3-3 and F-6) required the most extensive prompting across all models. Many of these problems necessitated explicit algebraic expansion prompts, like "$(-x^2 + 2x + 3)(-x^2 + 2x + 3)$," to

 doi:10.20944/preprints202509.0445.v1

overcome computational errors. This pattern suggests that current LLMs face significant challenges when problems require simultaneous management of spatial reasoning, algebraic manipulation, and integral setup. Interestingly, Claude 3.5 Sonnet uniquely solved problem 3-3 without any prompting, despite having a lower overall score. This indicates that mathematical capabilities in LLMs may be more modular than previously thought, with different models excelling in various mathematical domains.

A critical insight from the follow-up data is that prompt engineering significantly impacts model performance, but its effectiveness varies dramatically by problem type. For computational errors, simple directives like "Step 4 was wrong. Do it again" often sufficed. However, for conceptual errors, even detailed guidance sometimes failed. For example, Perplexity persistently struggled to solve problems 2-1, 2-2, and F-14 despite multiple attempts. This suggests fundamental differences in how models encode mathematical knowledge, with some models possessing more robust internal representations that can be activated through appropriate prompting, while others may lack the necessary underlying framework entirely.

4.3. Comparative Analysis of Model Strengths and Weaknesses

The follow-up analysis of the models' responses indicates that each one has unique behavior patterns when given mathematical guidance, which reflects their underlying architectures and training methods. Gemini Advanced with 1.5 Pro demonstrated the most advanced error recovery mechanisms, particularly evident in problem 3-3, where it succeeded with minimal prompting ("R(x) and r(x) answers are wrong. Do it again"). In contrast, other models required extensive algebraic breakdowns. Additionally, Gemini Advanced showed superior capabilities in evaluating limits in problem F-9, ultimately recognizing divergence after prompts regarding the application of L'Hôpital's rule.

Chat GPT 4o, while initially performing well, often needed exhaustive step-by-step guidance. This was particularly clear in problem F-1, where four separate prompts were necessary to guide it through basic algebraic evaluations. Mistral AI and Copilot Pro exhibited remarkably similar performance patterns. Both succeeded with standard techniques but required similar prompting strategies for more complex problems. Their parallel struggles with geometric applications—both failing problem 3-3 even with extensive guidance—suggest that they may have similar training methodologies or architectural constraints. Meta AI displayed an interesting pattern of computational fragility, frequently making basic arithmetic errors, such as incorrectly calculating (-1)^(5/3) in problem 1-5. However, it showed strong conceptual understanding once these errors were addressed. This suggests that while Meta AI may have a solid mathematical framework, its attention mechanisms for ensuring computational accuracy are weaker.

The most revealing comparisons emerged in problems where models exhibited differing success rates despite receiving the same prompts. For instance, in problem F-6 (volume of revolution), the prompt "$(-x^2 + 3x + 4)(-x^2 + 3x + 4)$" led to success for Chat GPT 4o, Meta AI, Claude, and Mistral AI, while Copilot Pro failed even with this explicit guidance. Similarly, in problem 2-6 (trigonometric integration), the prompt "Let u = sec(4x)" enabled most models to find the solution, but Gemini Advanced—typically the top performer—could not solve it after multiple attempts. These divergent responses to identical prompts suggest that the models have fundamentally different internal representations of mathematical concepts, making certain guidance strategies effective for some architectures but not for others.

4.4. Error Patterns and Mathematical Reasoning

The follow-up prompt data offers valuable insights into the types of mathematical errors found in large language models (LLMs). It reveals three distinct categories of errors, each with different profiles for correction. The most common errors were computational errors, such as incorrectly expanding $(-x^2 + 2x + 3)^2$. These were also the easiest to correct, with success rates exceeding 90% after a single prompt. These computational errors appear to result from lapses in attention rather than

 doi:10.20944/preprints202509.0445.v1

gaps in knowledge, as models consistently demonstrated the ability to perform the computation correctly when explicitly instructed to retry. This pattern was observed in problems like 3-2, F-6, and F-1, where models made arithmetic mistakes in intermediate steps but quickly recovered with minimal guidance.

Conceptual errors were much more challenging to address, particularly in areas like geometric visualization and limit evaluation. For instance, the washer method setup in problems 3-3 and F-6 highlighted fundamental limitations, with models often confusing the inner and outer radii even after explicit correction. The prompt "R(x) and r(x) are wrong. They must be switched" worked for some models but not for others, suggesting that geometric intuition cannot easily be prompted if the underlying representation is lacking. A particularly telling example was in problem F-9 (improper integral), where models like Gemini Advanced and Perplexity mistakenly asserted convergence despite it being mathematically impossible. They were unable to correctly evaluate limits at negative infinity, even with explicit instructions regarding L'Hôpital's rule.

A third category of errors, technique selection errors, emerged from the data. In these cases, models chose inappropriate methods despite knowing the correct procedures. Problem 2-2 illustrated this perfectly; several models attempted to use partial fractions when polynomial division was needed. The prompt "Don't use a partial fraction method" successfully redirected most models, indicating they had knowledge of alternative techniques but struggled with method selection. This pattern repeated in problem F-2, where models defaulted to partial fractions for rational functions that actually required simple polynomial division. The data suggest that while LLMs have internalized a broad repertoire of mathematical techniques, they lack the meta-cognitive skills needed to select the most appropriate method for a given problem, relying instead on pattern matching that can be adjusted with explicit instruction.

4.5. Implications for Educational Applications

The follow-up prompt data reveals both significant opportunities and critical limitations for the deployment of large language models (LLMs) in mathematics education. The high success rate of targeted prompts for identifying computational errors suggests that LLMs can act as effective "mathematical debugging partners" for students, aiding them in identifying and correcting arithmetic mistakes through interactive dialogue. The models' capability to provide step-by-step solutions with increasing accuracy after prompting reflects effective tutoring strategies, where instructors guide students to recognize their own errors instead of simply providing answers. However, the data also indicates that this scaffolding approach primarily works for procedural errors, while conceptual misunderstandings often persist despite extensive guidance.

The varying responsiveness to prompts across different models has significant implications for selecting educational tools. Gemini Advanced's superior error correction capabilities make it more suited for independent student use, as it can often self-correct with minimal guidance. In contrast, models like Perplexity or Meta AI, which demonstrated persistent errors despite prompting, may frustrate students or reinforce misconceptions. The data suggests a tiered approach to educational deployment: high-performing models like Gemini Advanced with 1.5 Pro and Chat GPT 4o for independent practice, mid-tier models like Copilot Pro for supervised homework help, and a cautious avoidance of models that exhibit persistent conceptual errors in specific domains. Notably, Claude 3.5 Sonnet uniquely solved the volume problem without prompting, despite overall lower scores, indicating that educators might benefit from utilizing multiple models to provide diverse problem-solving perspectives.

Most critically, the prompt analysis reveals that LLMs cannot replace human instruction when it comes to developing geometric intuition and spatial reasoning. The universal difficulty with problems 3-3 and F-6, where even detailed prompting often failed, shows that these models lack the visual-spatial frameworks that human instructors naturally employ. The persistence of errors in setting up area integrals—specifically in determining which function is "above"—even after explicit guidance suggests that students using these tools without proper conceptual grounding may develop

 doi:10.20944/preprints202509.0445.v1

procedural skills while missing fundamental understanding. This implies that LLMs should supplement human instruction in areas where they excel, such as procedural verification, computational checking, and providing multiple solution methods, while human teachers focus on developing conceptual understanding, geometric intuition, and problem-solving heuristics that current AI systems cannot adequately convey.

4.6. Methodological Considerations and Limitations

The follow-up prompting methodology used in this study highlights both the advantages and limitations of interactive assessments for evaluating AI mathematical capabilities. This approach revealed details about model performance that simple right/wrong scoring might overlook. For instance, Gemini Advanced with 1.5 Pro demonstrated superior adaptability, even though it had similar initial error rates compared to other models. However, the subjective nature of prompt construction introduces potential bias; for example, the phrasing "Check which curve is above" might appeal to certain model architectures more than "Determine the upper and lower functions," despite both expressing the same mathematical concept. Future studies would benefit from the development of standardized prompt libraries that test various phrasings to ensure a robust assessment of error correction capabilities rather than relying on prompt-specific pattern recognition.

A significant limitation arises from the text-only interaction used in this study, particularly evident in geometric problems where visual representations would greatly enhance communication. The extensive prompting required for problems 3-3 and F-6 could have been replaced by simple diagram annotations in a multimodal environment. Describing geometric relationships verbally, such as stating "R(x) is the outer radius," introduces ambiguity that visual communication would eliminate. Furthermore, the sequential nature of prompting does not accurately reflect real mathematical problem-solving, as students often revisit earlier steps based on insights gained later. This linear prompt-response structure may disadvantage models that could excel with a more holistic and iterative problem-solving approach.

Additionally, the study's emphasis on prompting effectiveness raises questions about ecological validity. In actual educational settings, students rarely receive such targeted and mathematically precise feedback. Prompts like "$(-x^2+3x+4)(-x^2+3x+4)$" offer a level of specificity that tends to provide partial solutions rather than genuine guidance. Consequently, this methodology might overestimate the practical utility of these models in education, where feedback is usually less precise and more conceptual. Moreover, the variation in prompting aggressiveness—some problems receiving up to five follow-up attempts while others were abandoned after two—introduces inconsistency that could affect model rankings. Despite these limitations, the methodology offers valuable insights into the adaptability of model responses and the depth of their mathematical understanding.

4.7. Implications for LLM Development and Future Directions

The follow-up prompt analysis provides a roadmap for improving mathematical capabilities in future large language model (LLM) architectures. The significant difference in high success rates for computational corrections compared to persistent failures in geometric reasoning indicates that current training paradigms effectively capture procedural mathematics but struggle to develop spatial-mathematical representations. The data shows that models like Gemini Advanced with 1.5 Pro have established more robust error-correction mechanisms, likely due to reinforcement learning techniques that explicitly reward self-correction. Future models should include training objectives that emphasize not only initial accuracy but also the ability to recognize and correct errors through dialogue, mirroring the iterative nature of human mathematical learning.

The problem-specific failure patterns identified by the prompts suggest targeted improvements for different mathematical domains. The universal difficulty with volume integration setup, even with explicit guidance, indicates that current architectures lack the implicit coordinate system representations that humans naturally use. Incorporating multimodal training with visual representations of mathematical concepts could help address this gap. Additionally, the persistent

 doi:10.20944/preprints202509.0445.v1

errors in evaluating limits at infinity suggest a need for better representations of asymptotic behavior and infinite processes. The success of technique-specific prompts (e.g., "Use u-substitution method") indicates that models possess diverse procedural knowledge but lack the meta-cognitive frameworks necessary for method selection. This suggests that future training should focus not only on solving problems but also on explaining why specific approaches are chosen.

Perhaps most importantly, the varying responsiveness to identical prompts across models suggests that mathematical capability in LLMs is not a singular skill but rather emerges from complex interactions between architecture, training data, and optimization objectives. The fact that Claude 3.5 Sonnet uniquely solved certain geometric problems while struggling in other areas implies that different architectural choices may provide advantages in specific mathematical domains. This leads to the possibility of specialized mathematical AI systems, where different models excel in various mathematical tasks, potentially combined using ensemble methods.

The prompt analysis also reveals that current evaluation benchmarks, which typically focus on final answers, overlook the valuable insights contained in error patterns and correction capabilities. Future development should incorporate interactive evaluation metrics that assess not only accuracy but also mathematical dialogue capability, error recognition, and conceptual flexibility—skills that the follow-up prompt data indicates are crucial for practical mathematical problem-solving applications.

5. Conclusions

The study's primary findings establish a clear performance hierarchy among the tested models. Gemini Advanced with 1.5 Pro achieved the highest score of 91.7% (A-), followed by Chat GPT 4o at 89.3% (B+). The overall class average was 83.3% (B), indicating that contemporary large language models (LLMs) have reached an undergraduate-level proficiency in formal calculus techniques. This performance level marks a significant milestone in AI development, as these models can successfully tackle structured mathematical problems that require both procedural knowledge and conceptual understanding. The consistency of performance patterns across various problem types suggests that these models have developed coherent mathematical frameworks rather than simply memorizing solution patterns. However, the 15.5 percentage point gap between the highest and lowest performing models highlights substantial variations in mathematical proficiency among different AI architectures.

A critical insight from the study is the stark contrast in models' performance on procedural versus conceptual mathematical tasks. While all models demonstrated near-perfect accuracy on basic integration techniques (averaging 95.2% on the first test), their performance sharply declined on problems requiring geometric intuition and spatial reasoning, particularly in areas such as volumes of revolution and area calculations between curves. The follow-up prompting methodology revealed that computational errors were easily correctable with simple guidance (over a 90% success rate). However, conceptual misunderstandings, especially regarding geometric visualization, often persisted despite detailed instructions. This pattern mirrors human learning trajectories in mathematics and highlights fundamental limitations in current AI architectures, likely stemming from their predominantly textual training, which provides limited exposure to the visual and spatial reasoning that underpin human mathematical intuition.

The educational implications of these findings present a nuanced picture of how LLMs could be integrated into mathematics instruction. The models' strong performance on procedural tasks and their ability to provide step-by-step solutions suggest they could serve as valuable supplementary tools for students practicing standard techniques or seeking alternative explanations. Their varying responsiveness to corrective prompts indicates that models like Gemini Advanced with 1.5 Pro, with superior error correction capabilities, may be more suitable for independent student use, while others might require more supervised implementation. However, the persistent challenges in geometric reasoning and spatial visualization, even with explicit guidance, highlight that these tools cannot replace human instruction in developing deep mathematical understanding and intuition. Educators

 doi:10.20944/preprints202509.0445.v1

should view LLMs as tools for computational verification and sources of procedural practice, rather than comprehensive learning solutions, with human teachers remaining essential for fostering conceptual understanding, geometric intuition, and the meta-cognitive skills needed for effective problem selection and approach.

Looking ahead, the study's findings provide clear directions for improving the mathematical capabilities of AI systems. The significant performance gap between procedural and conceptual tasks suggests that future models should incorporate multimodal training approaches that blend visual understanding with symbolic reasoning, particularly for geometric applications. The varying successes of different models on specific problem types indicate that mathematical capability in LLMs stems from complex interactions among architecture, training data, and optimization objectives, suggesting the potential for specialized mathematical AI systems optimized for different domains. Furthermore, the importance of error correction capabilities, as revealed through follow-up prompting, suggests that future training should emphasize not only initial accuracy but also the ability to recognize and correct errors through dialogue. As AI systems continue to advance, this study establishes a baseline for mathematical reasoning capabilities while highlighting specific areas—such as geometric intuition, complex problem setup, and conceptual understanding—where significant improvements are needed before these tools can fully support advanced mathematical learning and practice.

References

1. Hagos D. H., Battle R., Rawat D. B. (2024). Recent Advances in Generative AI and Large Language Models: Current Status, Challenges, and Perspectives. arXiv preprint arXiv:2407.14962. https://doi.org/10.48550/arXiv.2407.14962
2. Dagley M. A., Gill M., Saita E., Moore B., Chini J., Li X. (2018). Using Active Learning Strategies in Calculus to Improve Student Learning and Influence Mathematics Department Cultural Change. Proceedings of the Interdisciplinary STEM Teaching and Learning Conference, 2018, Volume 2. https://doi.org/10.20429/stem.2018.020108
3. Mishra A. K. (2023). An Introduction to Calculus: Fundamental Concepts and Applications. International Journal of Creative Research Thoughts, February 2023, Volume 11, Issue 2, f536-f541.
4. Bailey J. D., Claridge J., Partner A. (2024). Investigating students' perception of the importance of calculus: a cross-discipline comparison to inform module development. MSOR Connections, Volume 22, Number 1, 5–27. https://doi.org/10.21100/msor.v22i1.1457
5. Spresser D. M. (1981). High school calculus and achievement in university engineering and applied science courses. International Journal of Mathematical Education in Science and Technology, Volume 12, Issue 4, 453-459. https://doi.org/10.1080/0020739810120415
6. Kiat S. S. (2005). Analysis of Students' Difficulties in Solving Integration Problems. The Mathematics Educator, Volume 9, Number 1, 39-59.