Integral Solution Analysis

Integral Solution Analysis: $\int \frac{\tan^3(\ln x)}{x} dx$

This document evaluates the performance of various Large Language Models (LLMs) in solving the indefinite integral:

$$\int \frac{\tan^3(\ln x)}{x}\, dx$$

Correct Solution:

$$\frac{1}{2}\tan^2(\ln x) + \ln|\cos(\ln x)| + C$$

(Equivalent forms using $\sec^2(\ln x)$ *are also accepted).*

1 Summary of Results

Model	Status	Key Issues/Strengths
Gemma3:1b	Incorrect	Failed logic/algebra in integration step.
Gemma3:4b	Partial	Correct derivation, but forgot to substitute $u = \ln x$ back.
Gemma3:12b	Correct	Concise and accurate.
Gemma3:27b	Correct	Used sec substitution effectively.
Cognito (Llama 70B)	Incorrect	Failed substitution strategy; altered the problem logic.
Cognito (Llama 405B)	Correct	Flawless execution.
Mistral 7B	Correct	Correctly handled identities and integration.
Llama 3.3 70B	Incorrect	Integration power rule error ($\int x\, dx \neq x$).
Kimi K2	Correct	Clear step-by-step verification included.
GLM 4.5 Air	Correct	Correct; noted equivalence of log forms.
LongCat Flash	Correct	Correct step-by-step reasoning.
Deepseek R1T/R1T2	Correct	Strong mathematical reasoning; verified results.
Tongyi DeepResearch	Correct	Standard correct derivation.
Grok 4.1 Fast	Correct	Fast, accurate derivation.

2 Detailed Model Analysis

2.1 Small/Edge Models (1B - 7B Parameters)

- **Gemma3:1b: Failed.**
  - *Analysis:* The model correctly identified the u-substitution but hallucinated the algebraic simplification of $\tan^3(u)$.
  - *Context:* 1B models are designed for extreme efficiency (mobile/edge) and often lack the depth for multi-step symbolic math manipulation.
- **Gemma3:4b: Passable (with minor error).**
  - *Analysis:* It struggled mid-generation (detecting an error and retrying), eventually finding the correct integration path. However, it failed the final “clean up” step of substituting u back to $\ln x$ in the boxed answer.

  - *Context:* A good example of "borderline" capability where the reasoning exists but attention span/consistency wavers.

- **Mistral 7B Instruct**: **Success.**
  - *Analysis:* Despite being a smaller model, Mistral 7B accurately handled the trigonometric identities and integration constants.
  - *Context:* Mistral is known for high efficiency and punching above its weight class in reasoning tasks compared to similarly sized models.

2.2 Medium/Large Models (12B - 70B Parameters)

- **Gemma3:12b & 27b**: **Success.**
  - *Analysis:* Both models executed the task flawlessly. The 27B model used a slightly different (but valid) equivalent form involving $\sec^2$.
- **Llama 3.3 70B (and Cognito Llama 70B variant)**: **Failed.**
  - *Analysis:* Surprisingly, both variants of Llama 70B failed. The base Instruct model made a fundamental calculus error (integrating $\sec u \tan u \sec u$ as $\sec u$ instead of $\frac{1}{2}\sec^2 u$). The Cognito variant attempted a bizarre product rule reverse-engineering that didn't work.
  - *Context:* While Llama 3 70B is a general-purpose powerhouse, it can sometimes be prone to "rote" errors in specific step-by-step math procedures if not guided by a Chain-of-Thought (CoT) prompt or specific math fine-tuning.

2.3 Huge/Reasoning-Focused Models

- **Cognito (Llama 405B)**: **Success.**
  - *Analysis:* The massive 405B parameter count ensured high adherence to logical steps, correcting the failures seen in its 70B sibling.
- **Deepseek R1T / Chimera**: **Success.**
  - *Analysis:* These models not only solved the problem but included verification steps (differentiating the answer to check against the prompt).
  - *Context:* Deepseek's "R1" series is heavily optimized for reasoning and mathematics (often using Reinforcement Learning for reasoning paths), making them particularly effective at this type of task.
- **Kimi, GLM, Tongyi, Grok**: **Success.**
  - *Analysis:* These models all handled the standard substitution calculus without issue.

3 Conclusion

For symbolic mathematics and calculus:

1. **Size isn't everything:** The Mistral 7B (small) outperformed the Llama 3.3 70B (large) on this specific task.
2. **Specialization matters:** Deepseek (Math/Code focus) provided very robust, verified answers.
3. **Threshold for Math:** 1B parameters appears insufficient for reliable multi-step algebra, while 4B-7B is the transition zone where capability begins to stabilize.