NYU

Can LLMs *understand* Math?

Exploring the pitfalls of Mathematical Reasoning

Aditeya Baral, Ayush Rajesh Jhaveri, Tiasa Singha Roy, Yusuf Baig

12/05/2024

Investigating Math Reasoning

Introduction to the problem and Related Work

01

NYU

Can LLMs understand Math?

- LLMs struggle with higher-order mathematical tasks.
- Accuracy does not truly represent mathematical reasoning ability.
- It requires both the answer and the reasoning steps:
  - **Validity**: Determines validity of a step
  - **Redundancy**: Determines redundancy of a step.
  - **Errors**: Identifies different domains of errors
- Need for a **holistic approach** to evaluate mathematical reasoning.

NYU

Question:

Let $a_n = \frac{10^n - 1}{9}$. Define d_n to be the greatest common divisor of a_n and a_{n+1}. What is the maximum possible value that d_n can take on?

Generated Solution:

Final Answer: What is the value of a_n?

Question:

Compute without using a calculator: $9!/8!$

Generated Solution:

To compute $9!/8!$, we need to divide 9 by 8.

$$9! = 9 \times 8!$$

$$\frac{9!}{8!} = \frac{9}{8} = 1$$

Related Work

- Most approaches only use accuracy to evaluate mathematical reasoning
  - Hides unnecessary or incorrect **intermediate** steps.

- ReasonEval[1] presents a methodology to evaluate beyond mere accuracy.
  - Investigates **step-wise redundancy and validity.**
  - Emphasizes the importance of analyzing the reasoning process in mathematical tasks.

NYU

[1] Xia, Shijie, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. "Evaluating Mathematical Reasoning Beyond Accuracy." *arXiv preprint arXiv:2404.05692* (2024)

Methodology

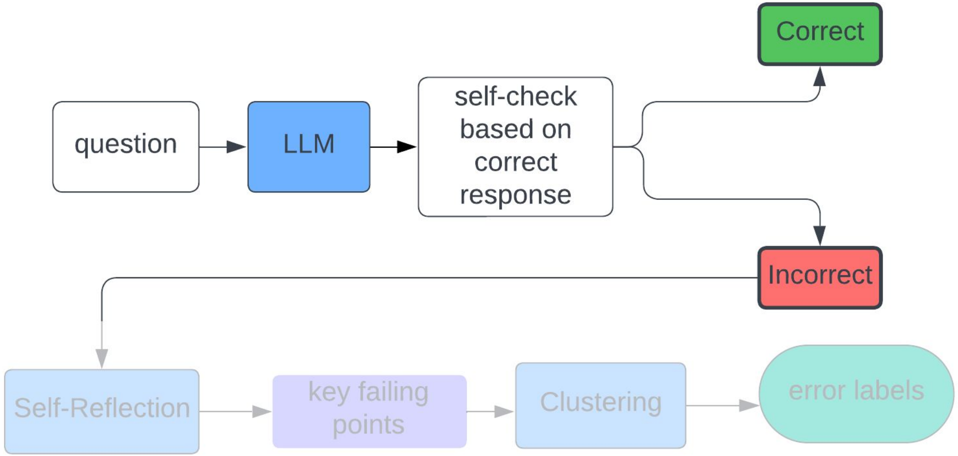
Architecture and Approach

02

NYU

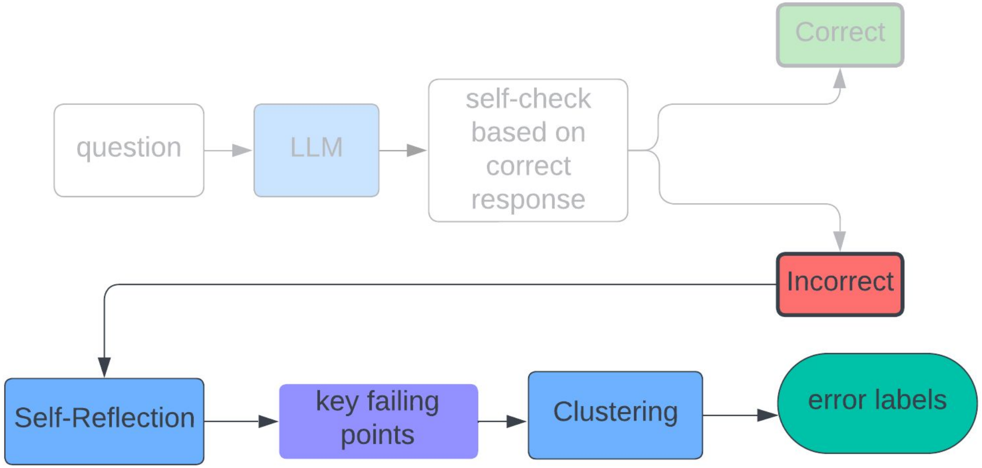
Stage 1 - Evaluating the Final Answer

- Prompt the LLM with the question q_i to generate a solution $\bar{a}_i = \{s_1, s_2, ..., s_n\}$, where s_i is a mathematical reasoning step.
- For each $\bar{a}_i$, induce self-checking in a multi-turn setup with the correct solution.
    - Does not account for reasoning steps.
- Determine the 0-1 accuracy of the final answer and invoke self-reflection in stage 2 for incorrectly generated answers.



Stage 2 - Evaluating the Approach

- Prompt the LLM with the generated solution $\bar{a}_i = \{s_1, s_2, \ldots, s_n\}$ and the actual solution $a_i = \{s_1, s_2, \ldots, s_m\}$
- For each pair (a_i , $\bar{a}_i$) induce self-reflection to highlight the points of **misalignment** of the reasoning steps with the actual solution.
- Analyse the failing points to compile a set of error labels l_i for each incorrectly generated sample.
    - Capture the type of each error - calculation, misinterpretation, etc.



Labels

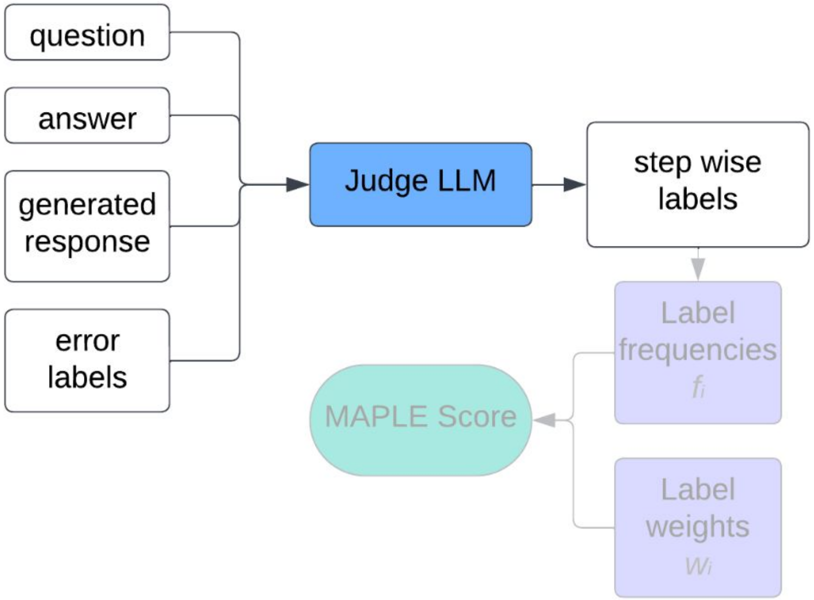
- Complete misunderstanding
- Partial misunderstanding
- Incorrect Method
- Incorrectly Applied Method
- Calculation Error
- Incoherent Output
- No Solution

NYU

Stage 3 - LLM as a Judge

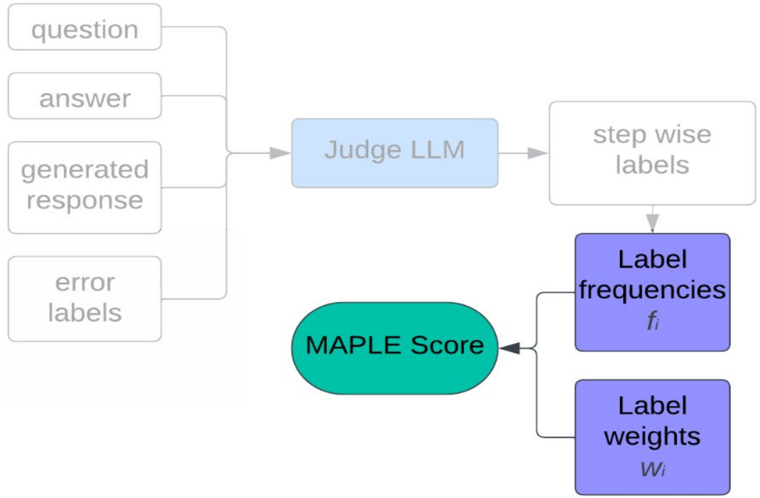
- Prompt the judge LLM with the previously generated error labels, l_i with each sample (q_i , a_i , $\bar{a}_i$) to generate step wise labels matrix.

- Use stepwise labels to compute error metric.



Stage 4 - MAPLE Score

- MAPLE (MAthematical Pitfalls and Logical Evaluation) Score
  - **Novel holistic metric** to quantify errors in mathematical reasoning
- Compute error rate e_i with frequency of each label per sample, $f_i = \{f_1, f_2, \ldots, f_6\}$ from the matrix and their corresponding penalty weight $w_i = \{w_1, w_2, \ldots, w_6\}$.
- Scale value of e_i using redundancy score r_i and validity score v_i.
- Use tanh to normalize final score to a range of [0, 1]



$$e_i = \frac{\sum w_i \cdot log(1 + f_i)}{\sum w_i}$$

$$\text{MAPLE}_{score} = \tanh\left(\frac{e_i \cdot v_i}{r_i}\right)$$

Experiments and Results

Data, Models and Analysis

03

NYU

Experiment Setup - Dataset and Models

Dataset

MATH[1] comprises 12500 math problems distributed across various parameters

- 5 levels of difficulty, L1 to L5
- 7 types of problems: algebra, intermediate algebra, pre-algebra, calculus, pre-calculus, probability and number theory.

Models

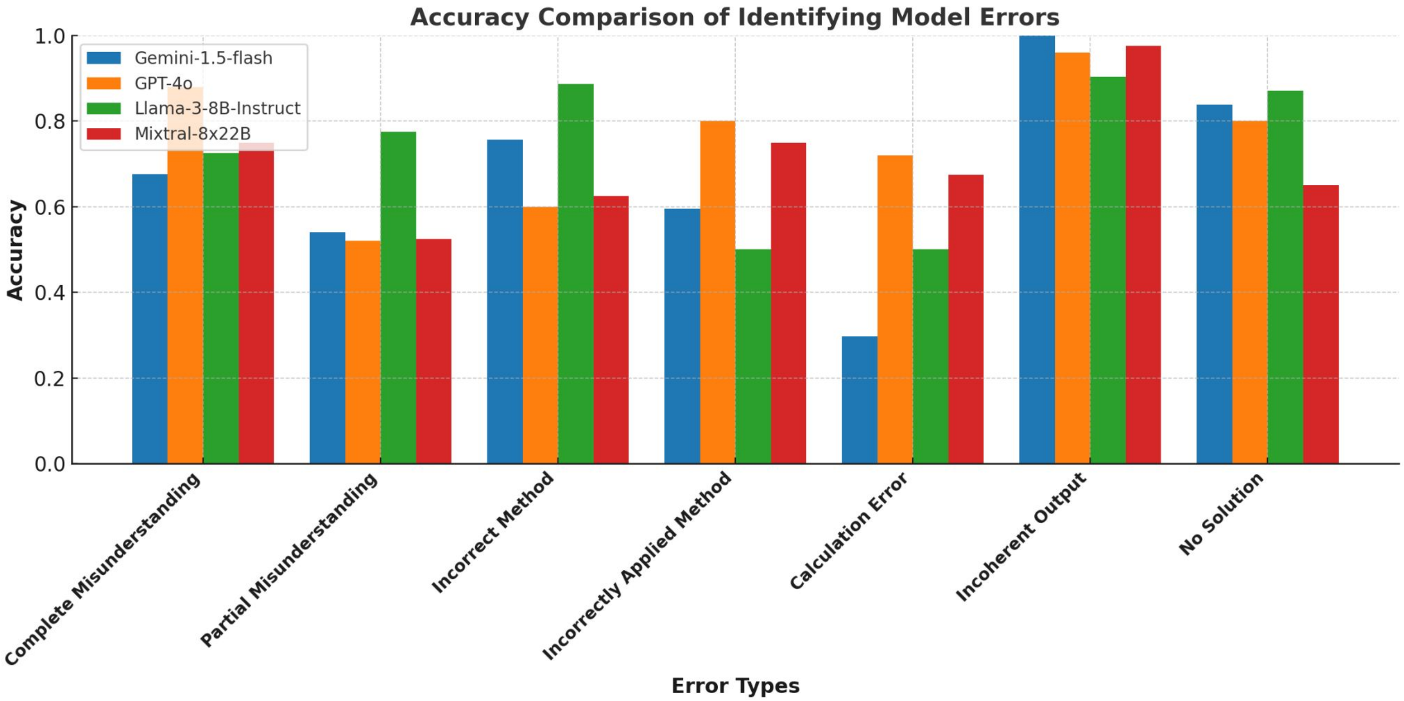
- Llama-3-8B-Instruct
- Gemini-1.5-Flash
- GPT-4o
- Mixtral-8x22B

NYU

[1] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. NeurIPS, 2021

Results: LLM as a Judge

We use LLM as a Judge to generate step-wise errors and compare the performance with human grounding.

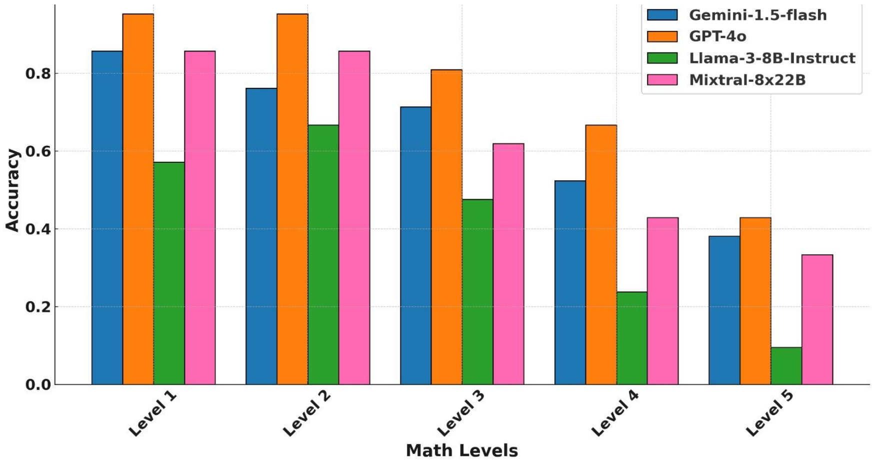


NYU

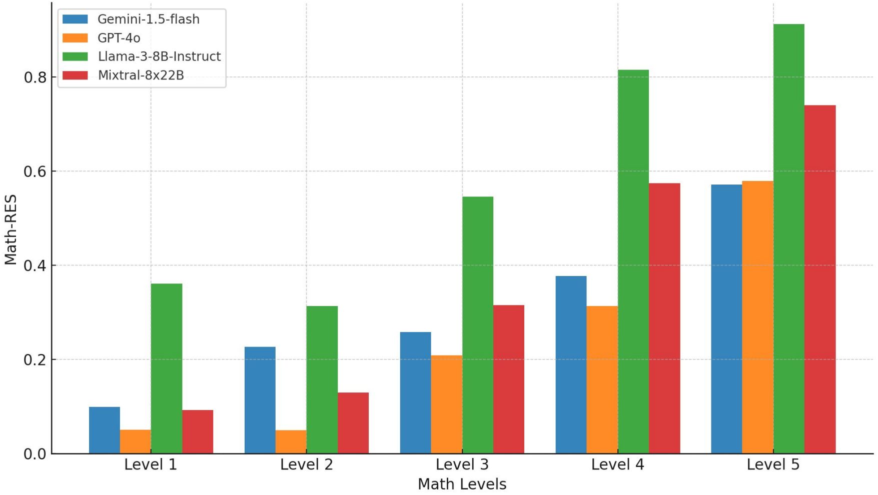
Results: Level-wise Analysis

- Accuracy helps judge whether the answer to the math question is right or wrong.
- With MAPLE Score, we can quantitatively determine the *incorrectness* of the answer.

Level-wise Accuracy Comparison



Level-wise MAPLE score Comparison

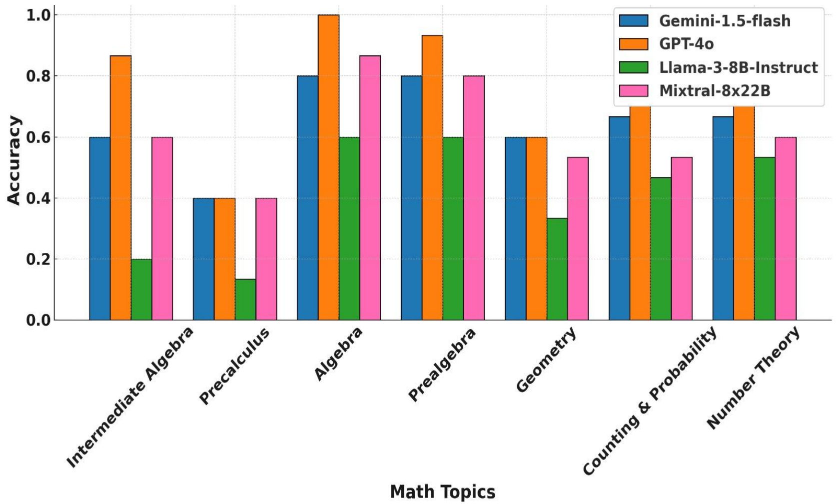


NYU

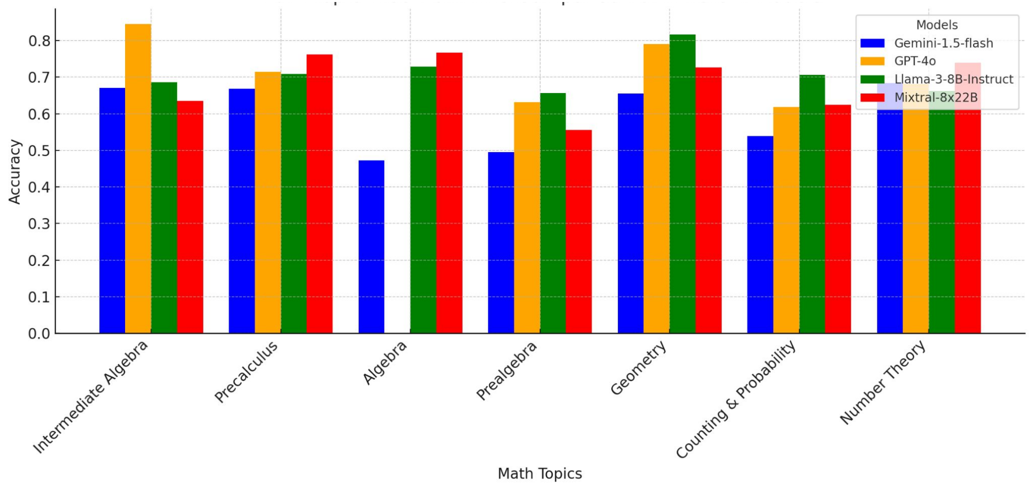
Results: Topic-wise Analysis

- Accuracy helps judge whether the answer to the math question is right or wrong.
- With MAPLE Score, we can quantitatively determine the *incorrectness* of the answer.

Topic-wise Accuracy Comparison



Topic-wise MAPLE score Comparison



NYU

Conclusion

Future Work

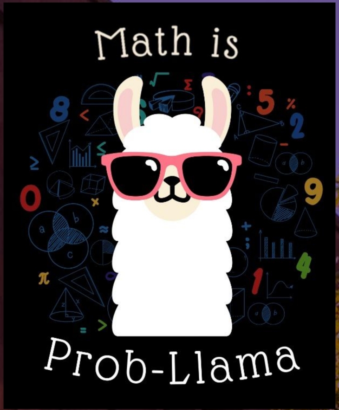
04

NYU

Future Work

- Expand our framework to consider an exhaustive range of errors
  - Consider **topic-specific** reasoning errors.
- Handle potential **hallucination** in LLMs to create stronger human aligned judgement.
  - Fine-tune LLMs for evaluation-specific tasks and explore alternatives to LLM as a Judge.
- Incorporate ranking of labels in final scoring to address their **relevance** to a sample.
- Test evaluation framework on a broader range of models and datasets.

Thank You :)



NYU