39-th AAAI Conference on Artificial Intelligence

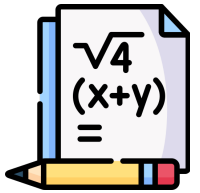
Assessing the Creativity of LLMs in Proposing Novel Solutions to Mathematical Problems

Speaker: Junyi Ye

Junyi Ye[1], Jingyi Gu[1], Xinyun Zhao[1], Wenpeng Yin[2], Guiling Wang[1]

[1]New Jersey Institute of Technology, [2]The Pennsylvania State University

March 2, 2025



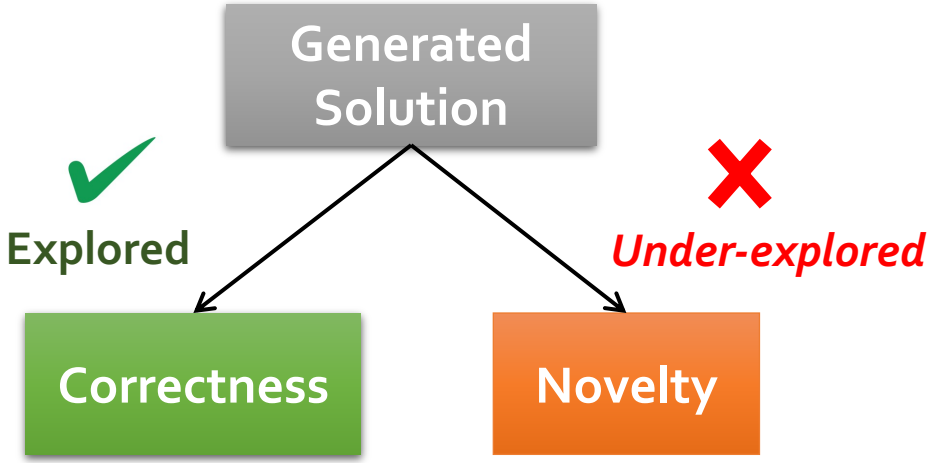
Motivation

- AI advancements, especially in Large Language Models (LLMs), have improved complex problem-solving capabilities.
- LLMs have excelled on challenging mathematical benchmarks like GSM8K, MATH and AIME.

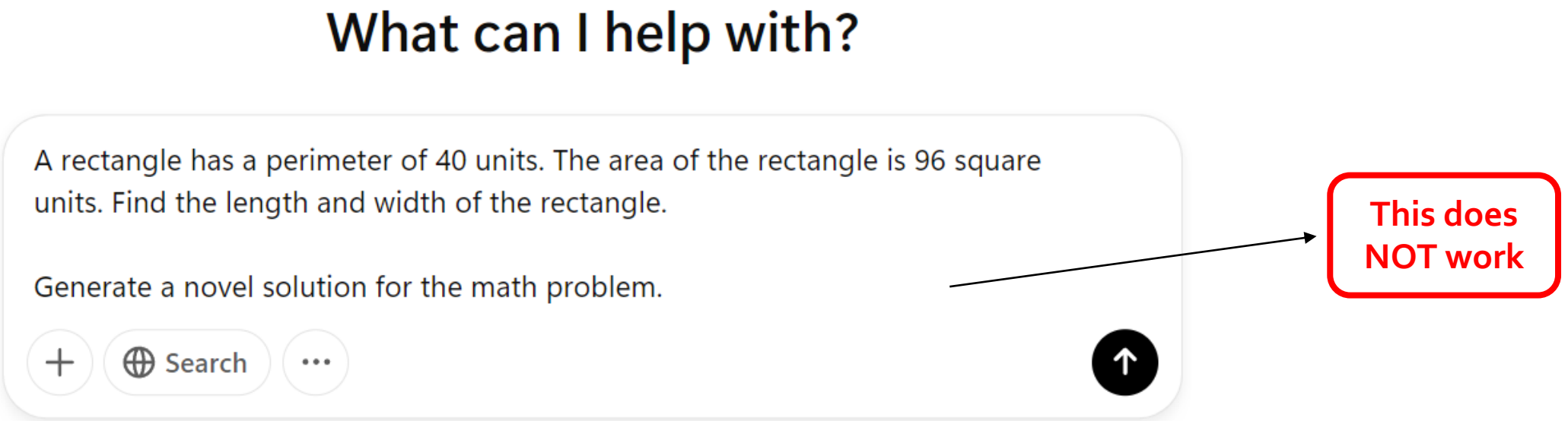
🔷 Potential for creative problem-solving with LLMs remains underexplored.

Standard Definition of Creativity

- **Creativity = Novelty + Usefulness** *(Runco & Jaeger, 2012)*.
- **Usefulness = Correctness.**
- **Novelty** is harder to measure in mathematics.



How to Generate A Novel Solution with LLM?



Challenges:

- Abstract Nature of "Novel".
- Difficult to control the degree of “Novelty”.

How to Generate Novel Solutions with LLM?

- **Our approach:** Generate new, correct solutions **distinct** from human-provided ones.

- **Novel Solution Generation:**
  - Input: A math problem + ***k*** known solutions.
  - LLM generates a **new solution**.

Challenges:

- ~~Abstract Nature of "Novel".~~ **Provide a clearer definition of novel = distinct.**
- ~~Difficult~~ to control the degree of "Novelty". ***k can be used to control the degree.*** **Much easier**

Approach 1 (Algebra)

1. Algebraic Solution:

Let the length be L and the width be W.

- **Perimeter equation:** $2L + 2W = 40 \rightarrow L + W = 20$
- **Area equation:** $L \times W = 96$

From $L + W = 20$, express W in terms of L:

$$W = 20 - L$$

Substitute into the area equation:

$$L(20 - L) = 96$$

$$20L - L^2 = 96$$

$$L^2 - 20L + 96 = 0$$

So:

- $L = \frac{20+4}{2} = 12, W = 8$
- $L = \frac{20-4}{2} = 8, W = 12$

Approach 2 (Geometry)

2. Geometric Solution (Using Graphing or Reasoning):

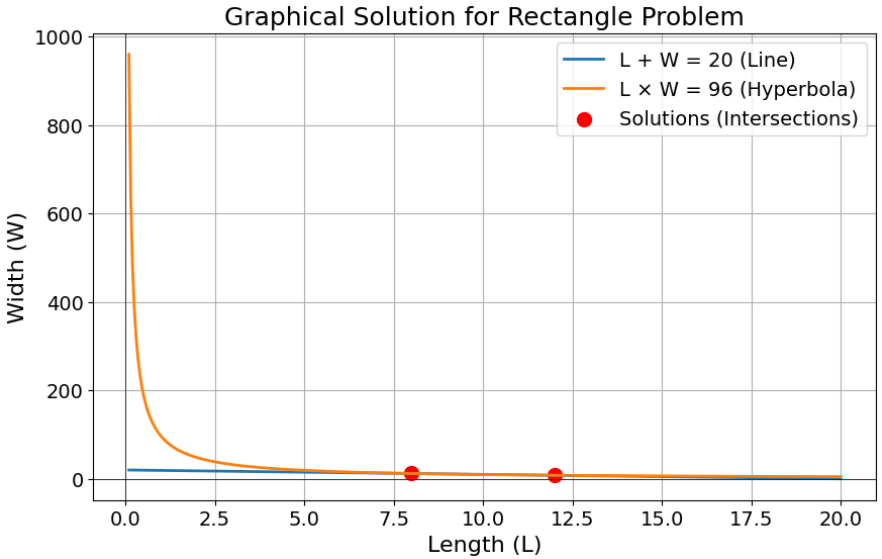
Visualize the problem on a coordinate plane where the sum $L + W = 20$ forms a straight line. The area $L \times W = 96$ forms a hyperbola.

Graphing both equations, the intersection points give the solutions:

- $(12, 8)$
- $(8, 12)$

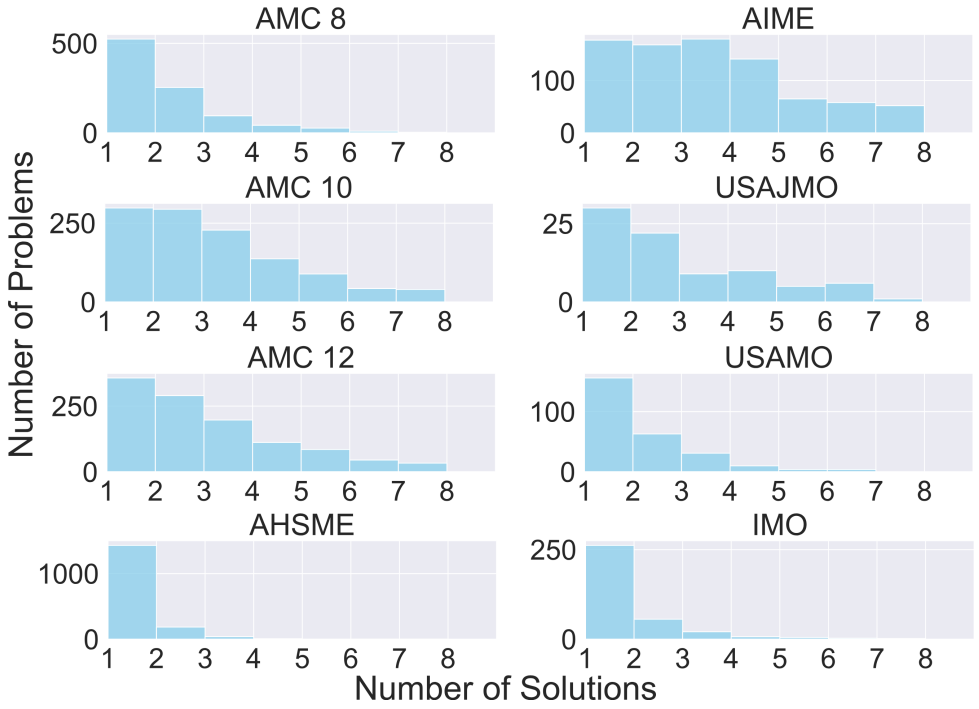
Alternatively, reasoning with rectangle dimensions that multiply to 96 and add up to 20 quickly leads to:

- $12 + 8 = 20$ and $12 \times 8 = 96$



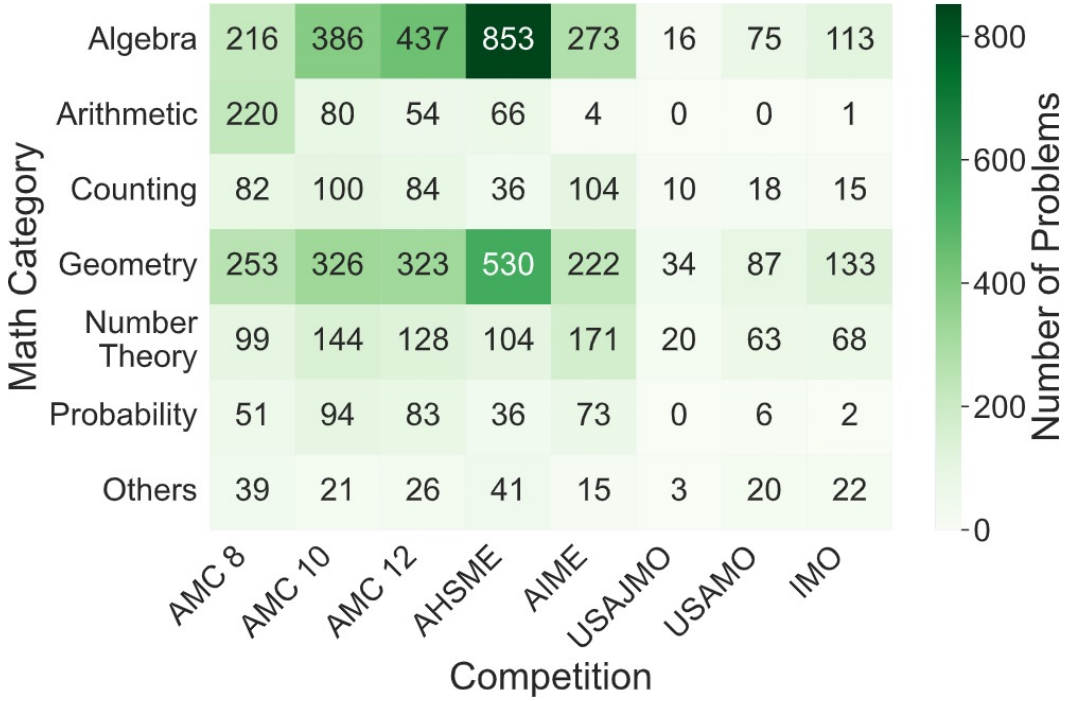
CreativeMath: A Benchmark Dataset

- **CreativeMath** comprises high-quality mathematical problems from various competitions and their numerous solutions.



6,469 problems with 14,223 solutions

- A broad range of ***mathematical topics***, ***problem types***, and covers different ***difficulty levels***.
- **8 major US competitions**: AMC 8, AMC 10, AMC 12, AHSME, AIME, USAJMO, USAMO, and IMO.



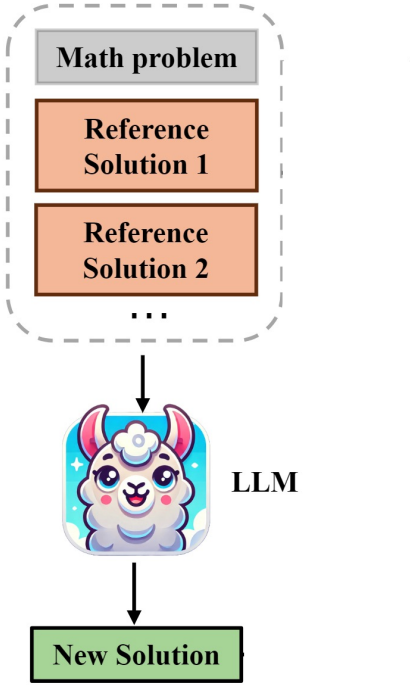
Dataset Creation

Data Collection

- Source: Art of Problem Solving(AoPS).
- Solutions submitted by competition participants.
- Approximate the complete set of viable human solutions for each problem.
- Earlier solutions are often the most common and intuitive, while later ones may build on previous methods, offer improvements, or introduce entirely novel algorithms.

Data Cleaning

- HTML to latex
- Remove incomplete problem and solutions
- Remove problems with images



Novel Solution Generation

STAGE 1:

Novel Solution Generation

- Generate a new solution that is distinct from *k* reference solutions.
- *k solutions are sequentially selected based on the order in which competitors uploaded their solutions on the website.*
- *When k increases, the difficulty in generating novel solutions also increases.*

Criteria for evaluating the difference between two mathematical solutions include:
1. If the methods used to arrive at the solutions are fundamentally different, such as algebraic manipulation versus geometric reasoning, they can be considered distinct;
2. Even if the final results are the same, if the intermediate steps or processes involved in reaching those solutions vary significantly, the solutions can be considered different;
3. If two solutions rely on different assumptions or conditions, they are likely to be distinct;
4. A solution might generalize to a broader class of problems, while another solution might be specific to certain conditions. In such cases, they are considered distinct;
5. If one solution is significantly simpler or more complex than the other, they can be regarded as essentially different, even if they lead to the same result.

Given the following mathematical problem:
{problem}

And some typical solutions:
{solutions}

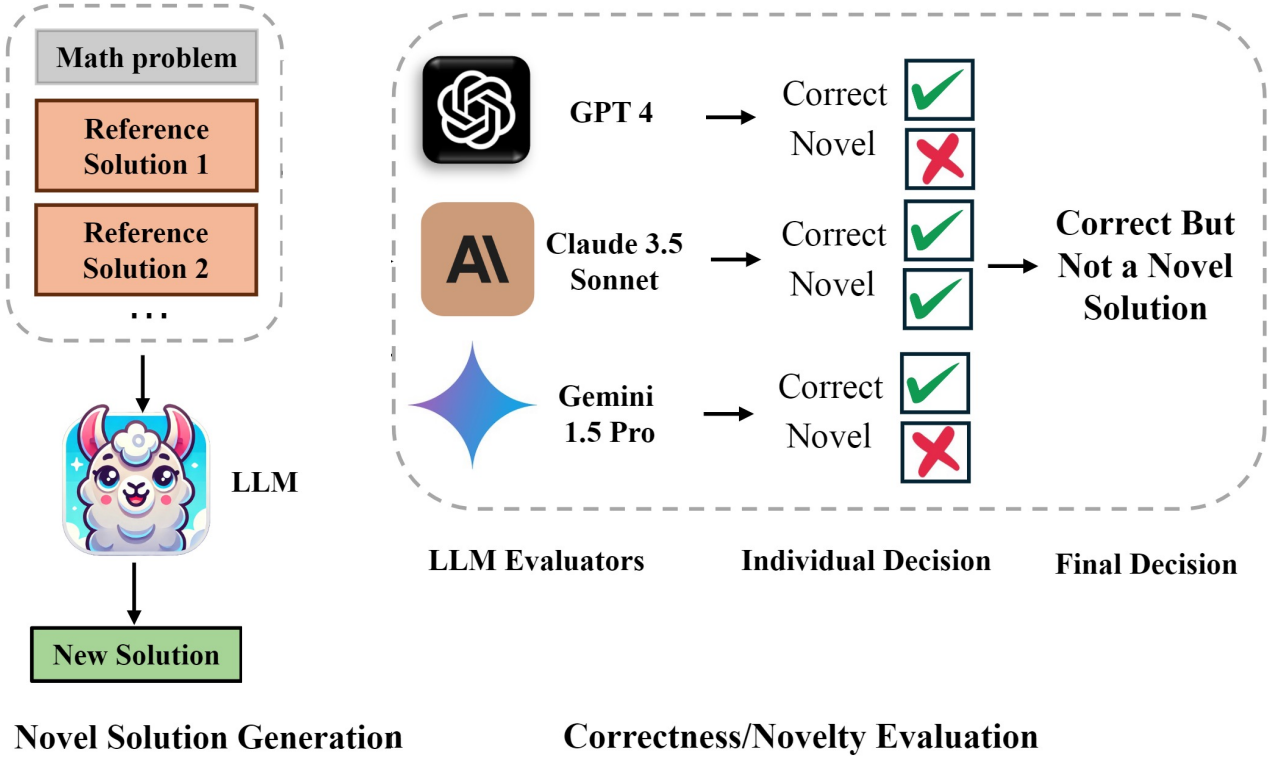
Please output a novel solution distinct from the given ones for this math problem.

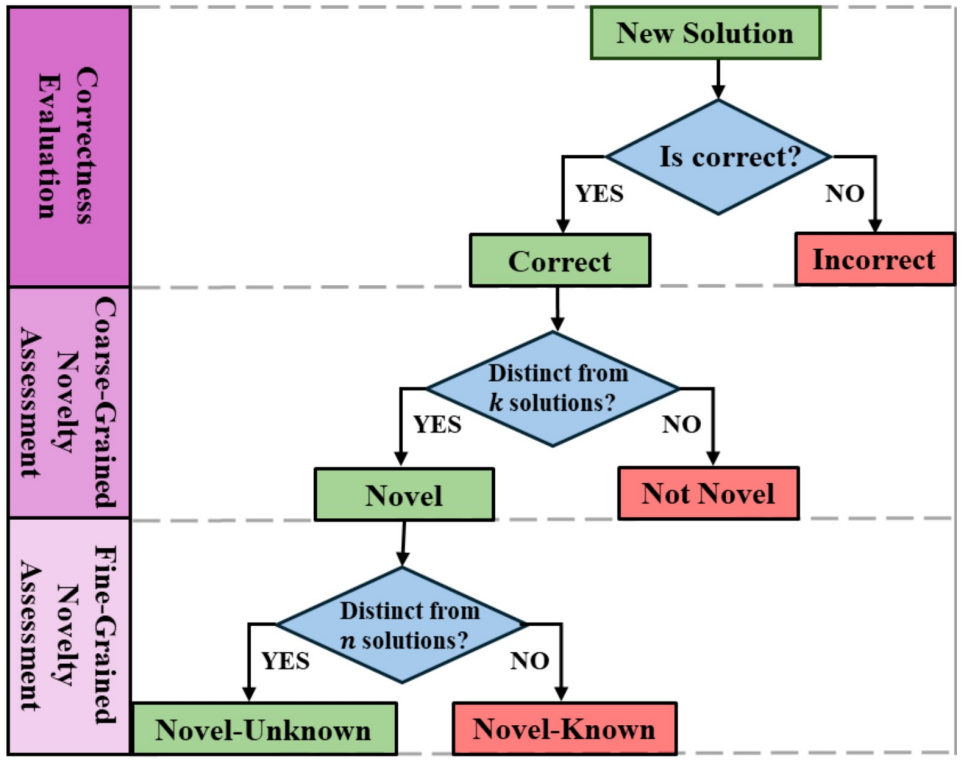
- *k ranges from 1* to *n*.
- *n is the total number of available reference solutions.*

STAGE 2:

Correctness and Novelty Evaluation

- 2.1 Correctness Evaluation
- 2.2 Coarse-Grained Novelty Assessment
- 2.3 Fine-Grained Novelty Assessment

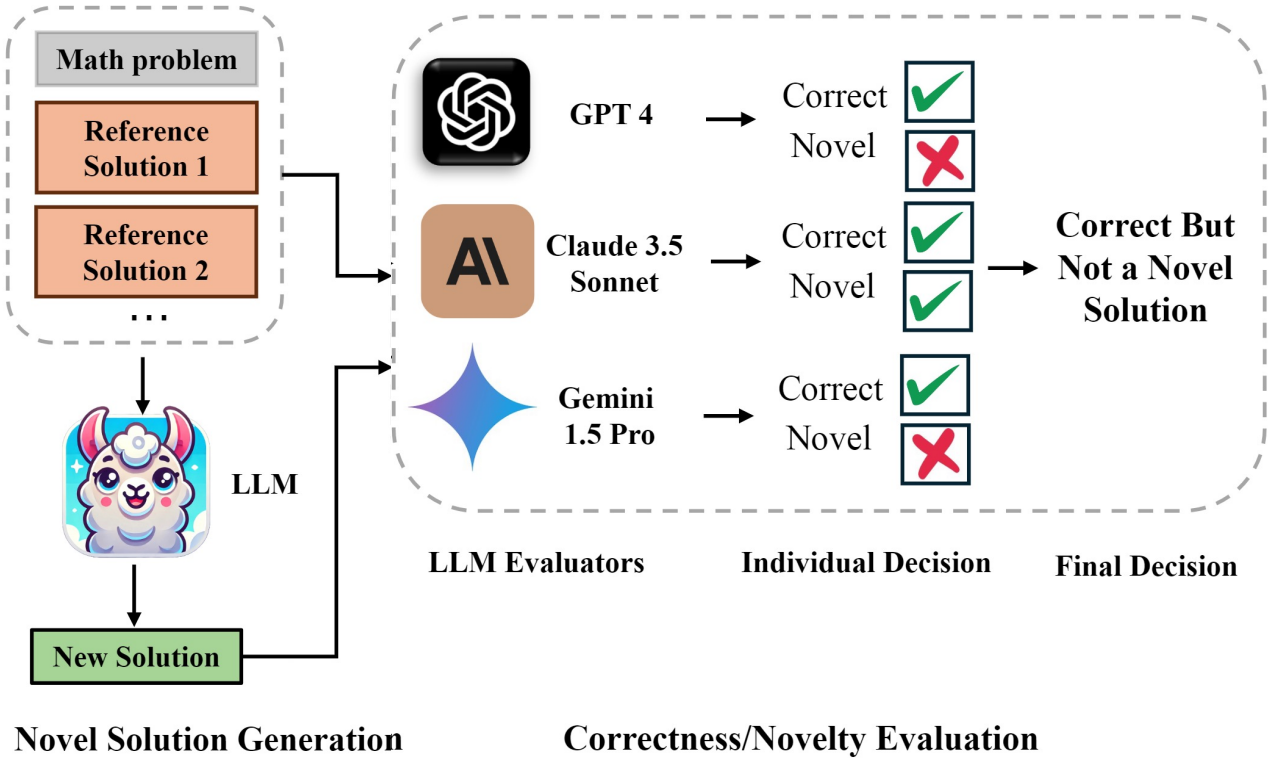




Detailed Evaluation Pipeline for Each LLM Evaluator

STAGE 2:

Correctness and Novelty Evaluation



Given the following mathematical problem:
{problem}

Reference solutions:
{solutions}

New solution:
{new solution}

Please output YES if the new solution leads to the same result as the reference solutions; otherwise, output NO.

Criteria for evaluating the novelty of a new mathematical solution include:
1. If the new solution used to arrive at the solutions is fundamentally different…
...

Given the following mathematical problem:
{problem}

Reference solutions:
{solutions}

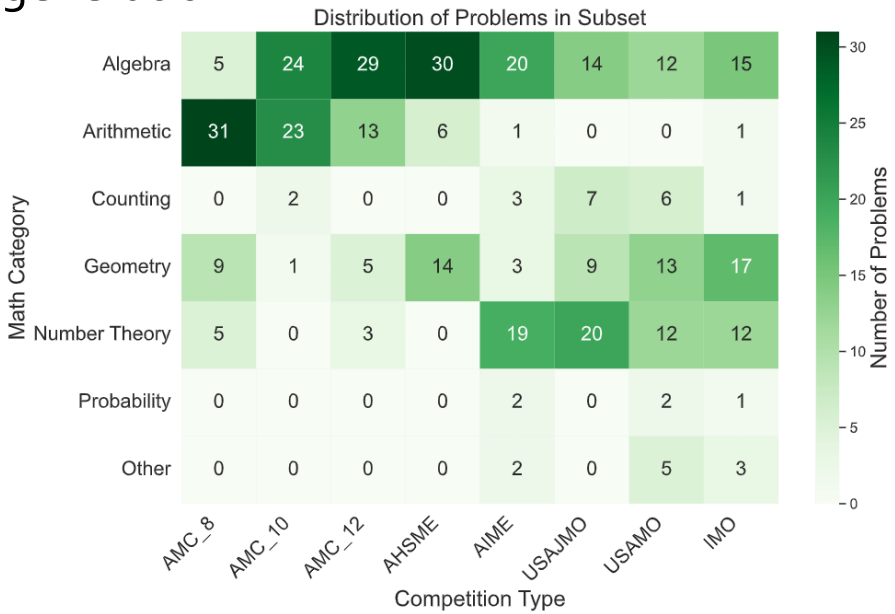
New solution:
{new solution}

Please output YES if the new solution is a novel solution; otherwise, output NO.

Experiment Setting

Dataset: CreativeMath Subset

- Randomly selected 50 problems/competition (400 math problems and 605 solutions with k varying from 1 to at most 5)
- Limit prompt length to 3K tokens
- 1K tokens are reserved for new solution generation.



Evaluation Metrics

Symbol	Metric Definition
C	**Correctness Ratio**: The proportion of solutions that are valid and can solve the problem correctly.
N	**Novelty Ratio**: The proportion of solutions that are both correct and distinct from the provided k reference solutions.
N_{u}	**Novel-Unknown Ratio**: The proportion of solutions that are both correct and unique compared to all known human-produced solutions n.
N/C	**Novelty-to-Correctness Ratio**: The ratio of novel solutions to all correct solutions.
N_{u}/N	**Novel-Unknown-to-Novelty Ratio**: The ratio of Novel-Unknown solutions to all available novel solutions.

Table 1: Evaluation Metrics and Their Definitions

How Effectively Can LLM Generate A Novel Solution?

Source	Model	C (%) ↑	N (%) ↑	N/C (%) ↑	N_u(%) ↑	N_u/N (%) ↑	MATH (%) ↑
Closed-source	Gemini-1.5-Pro	**69.92**	**66.94**	**95.75**	**65.45**	**97.78**	67.7 (Reid et al. 2024)
	Claude-3-Opus	59.84	44.63	74.59	42.98	96.30	61.0 (Anthropic 2024)
	GPT-4o	60.83	30.08	49.46	27.60	91.76	**76.6** (OpenAI 2024)
Open-source	Llama-3-70B	58.84	**48.76**	**82.87**	**46.94**	96.27	50.4 (Meta AI 2024)
	Qwen1.5-72B	47.44	33.06	69.69	32.40	**98.00**	41.4 (DeepSeek-AI 2024)
	DeepSeek-V2	**63.47**	30.91	48.70	29.09	94.12	43.6 (DeepSeek-AI 2024)
	Yi-1.5-34B	42.98	29.09	67.69	28.43	97.73	50.1 (01-ai 2024)
	Mixtral-8x22B	56.03	27.27	48.67	25.62	93.94	41.8 (Mistral AI 2024)
	Deepseek-Math-7B-RL	38.35	12.56	32.76	11.57	92.11	**51.7** (Shao et al. 2024)
	Internlm2-Math-20B	40.17	11.90	29.63	11.07	93.06	37.7 (Ying et al. 2024)

Key Findings:

- Gemini-1.5-Pro excels in generating novel solutions.
- Smaller and math-specialized models show lower performance in novelty generation.
- A clear distinction between traditional math problem-solving and novel solution generation.

How Does *k* Affect LLM's Performance?

Impact of *k* on Correctness

Correctness increases →

Model	$k=1$	$k=2$	$k=3$	$k=4$
Gemini-1.5-Pro	**68.00**	**70.78**	**78.57**	**100**
Llama-3-70B	55.00	66.23	64.29	75.00
Claude-3-Opus	55.00	66.88	76.19	75.00
Qwen1.5-72B	43.75	55.19	57.14	37.50
DeepSeek-V2	61.00	66.88	71.32	75.00
GPT-4o	58.25	64.94	66.67	75.00
Yi-1.5-34B	42.75	42.21	47.62	50.00
Mixtral-8x22B	53.50	60.39	64.28	62.50
Deepseek-Math-7B-RL	35.50	40.91	52.38	50.00
Internlm2-Math-20B	38.00	42.21	47.62	62.50

◆ When *k* increases, the correctness ratio increases. (Align with few-shot learning).

Impact of the Degree of Solution Availability (*n* – *k*) on Novelty

Novelty decreases →

Model	$n-k=2$	$n-k=1$	$n-k=0$
Gemini-1.5-Pro	**100**	**95.92**	**95.10**
Llama-3-70B	87.50	85.26	81.03
Claude-3-Opus	91.67	72.94	73.68
Qwen1.5-72B	85.00	70.15	68.37
DeepSeek-V2	36.00	54.17	47.84
GPT-4o	57.69	53.33	47.35
Yi-1.5-34B	52.38	52.87	46.43
Mixtral-8x22B	33.33	35.48	56.07
Deepseek-Math-7B-RL	27.78	25.86	35.10
Internlm2-Math-20B	15.00	27.69	32.89

◆ When *n*-*k* decreases, novelty-to-correctness ratio drops.

How Does Difficulty Level Affect LLM's Performance?

Competition	Difficulty	k	Average C	Average N/C
AMC 8	1-1.5	1	71.80	55.39
AMC 10	1-3	1	67.20	59.96
AHSME	1-4	1	65.08	63.11
AMC 12	2-4	1	60.40	54.05
AIME	3-6	1	35.80	55.55
USAJMO	6-7	1	37.00	77.23
USAMO	7-9	1	35.00	83.01
IMO	5.5-10	1	35.60	78.86

Difficulty decreases ↓

Correctness decreases ↓

Novelty increases ↓

Findings:

- LLMs struggle with accuracy on harder problems, they are more likely to generate novel solutions when they do succeed.
- A shift in the balance between familiarity and innovation

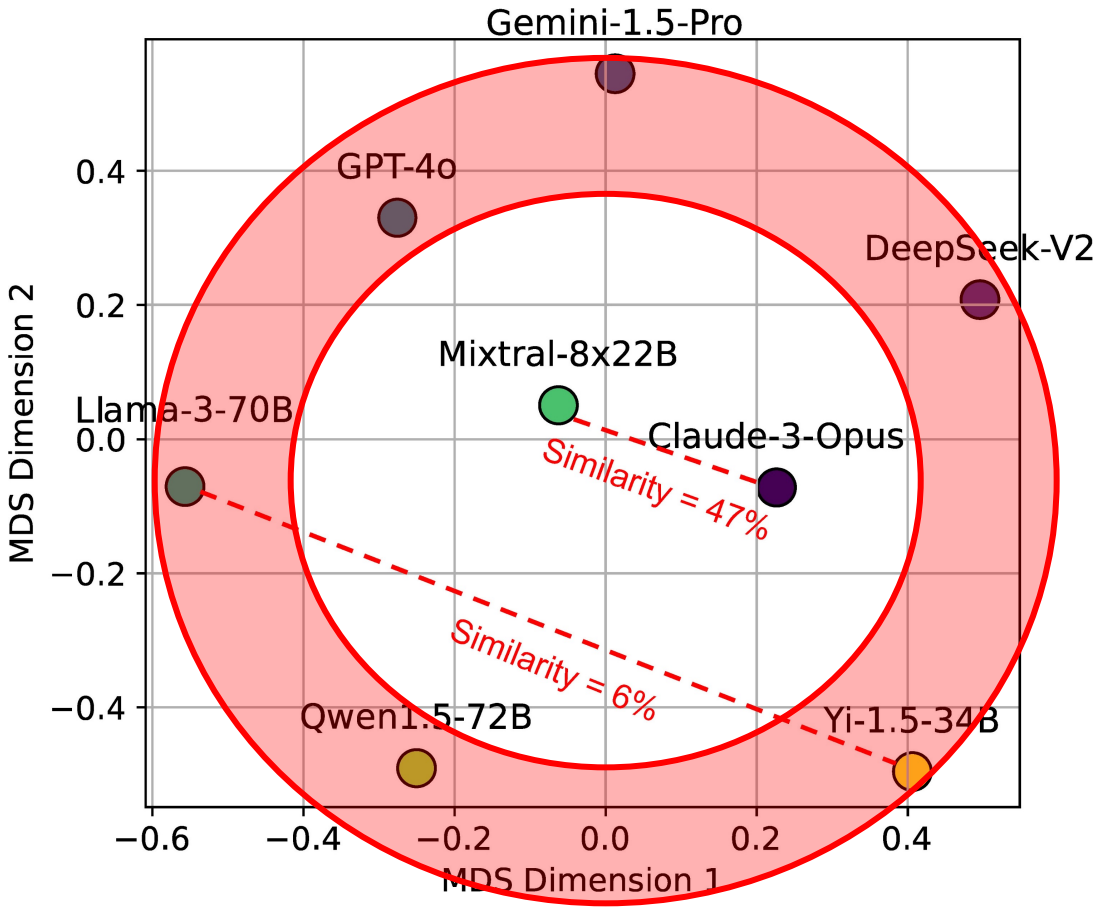
Similarity Map Between Novel Solutions Generated By Different LLMs

Step 1: Measure pairwise similarity between the outputs of various LLMs.

Step 2: Map similarity matrix into 2D plane with Multidimensional Scaling (MDS).

Findings:

- Low similarity between the novel solutions generated by different LLMs.
- Leverage LLMs on the periphery to generate diverse solutions.



Conclusion

- **CreativeMath Dataset:** Introduced a dataset to assess LLMs' creative problem-solving.
- **Framework:** Developed a system to generate novel solutions and measure both accuracy and innovation.
- **Key Findings:** Found significant variability in LLMs' creative abilities.
- **AI Advancement:** Stressed the need for AI to offer original insights, not just correct answers.
- **Future Research:** Encouraged deeper exploration of LLM creativity in complex domains like math.

Thank You

Guiling Wang

Wenpeng Yin

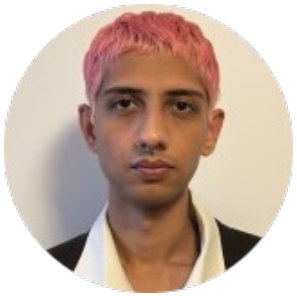
Junyi Ye

Jingyi Gu

Xinyun Zhao

Suraj Patel

Venkata Sai Lakshman Palli



Aadish Jain

Paper & Code

NJIT New Jersey Institute of Technology

PennState

Reference

1. Runco, M. A.; and Jaeger, G. J. 2012. The standard definition of creativity. Creativity research journal, 24(1): 92–96.
2. Art of Problem Solving. “AoPS Wiki”, https://artofproblemsolving.com/wiki/.