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

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

39-th AAAI Conference on Artificial Intelligence

Introduction

- **Motivation:** AI models like GPT-4 and Gemini-1.5-Pro excel at solving math problems, but can they **think creatively**?
- **Key Question:** Can LLMs propose **new, innovative mathematical solutions**, or are they just mimicking human approaches?
- **Existing Gap:** Most benchmarks only test correctness, **ignoring creativity in problem-solving**.

👉 We introduce ***CreativeMath***, a dataset and evaluation framework to **assess LLMs' ability to generate novel solutions** after seeing known ones.

Problem Definition

- **Creativity = Novelty + Usefulness** *(Runco & Jaeger, 2012)* [1]
- While **correctness = usefulness, novelty** is harder to measure in mathematics.
- Traditional math AI research **focuses on accuracy**, but we evaluate **solution diversity and originality**.

✳ **Example:** Given a geometry problem with 2 known solutions, can an LLM propose a **different, valid** approach?

CreativeMath: A Benchmark for Mathematical Creativity

Dataset Curation

- **Source:** 6,469 problems & 14,223 solutions from AMC 8, AMC 10, AMC 12, AIME, USAJMO, USAMO, IMO
- **Coverage:**
  - **Difficulty Levels:** Middle school to Olympiad
  - **Topics:** Algebra, Geometry, Combinatorics, Number Theory, etc.
- **Data Source:** Art of Problem Solving (AoPS) – A complete repository of diverse competition problems and human solutions [2].

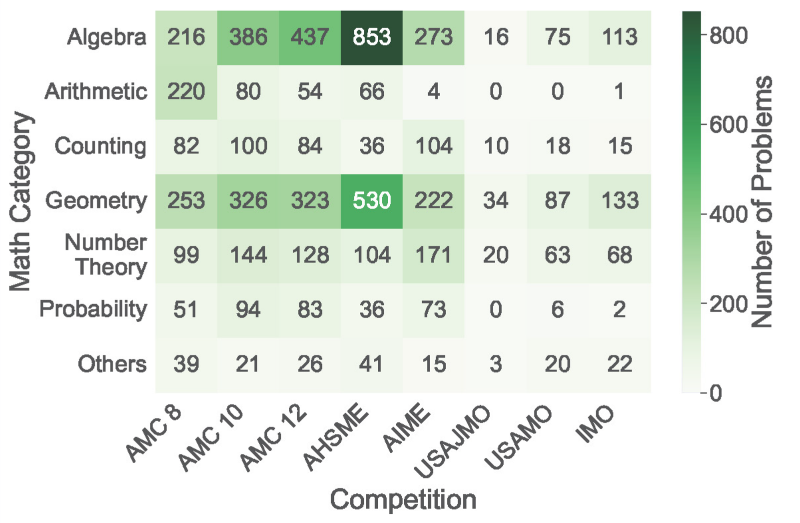

Figure 1: Distribution of problems across different math categories and competitions in the CreativeMath dataset.

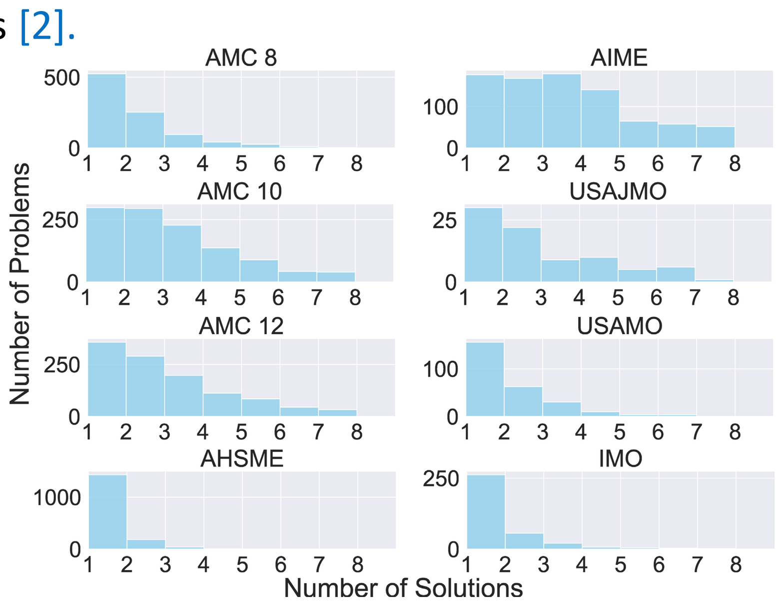

Figure 2: Distribution of the number of solutions per problem across different competitions.

Prompt Templates

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.

Figure 4: The prompt template for generating novel solution.

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.

Figure 5 (top): The prompt templates for evaluating the correctness of the generated solution.

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.

Figure 5 (bottom): The prompt templates for evaluating the novelty of the generated solution.

Reference

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

Method

Goal: Test if LLMs can generate new, correct solutions **distinct** from human-provided ones.

1. **Novel Solution Generation:**
   - Input: A math problem + ***k*** known solutions.
   - LLM generates a **new solution**.
2. **Correctness Check:** Is the new solution valid?
3. **Coarse-Grained Novelty:** Compare against ***k* reference solutions**.
4. **Fine-Grained Novelty:** Compare against **all human solutions (*n* total)**.

Paper & Code

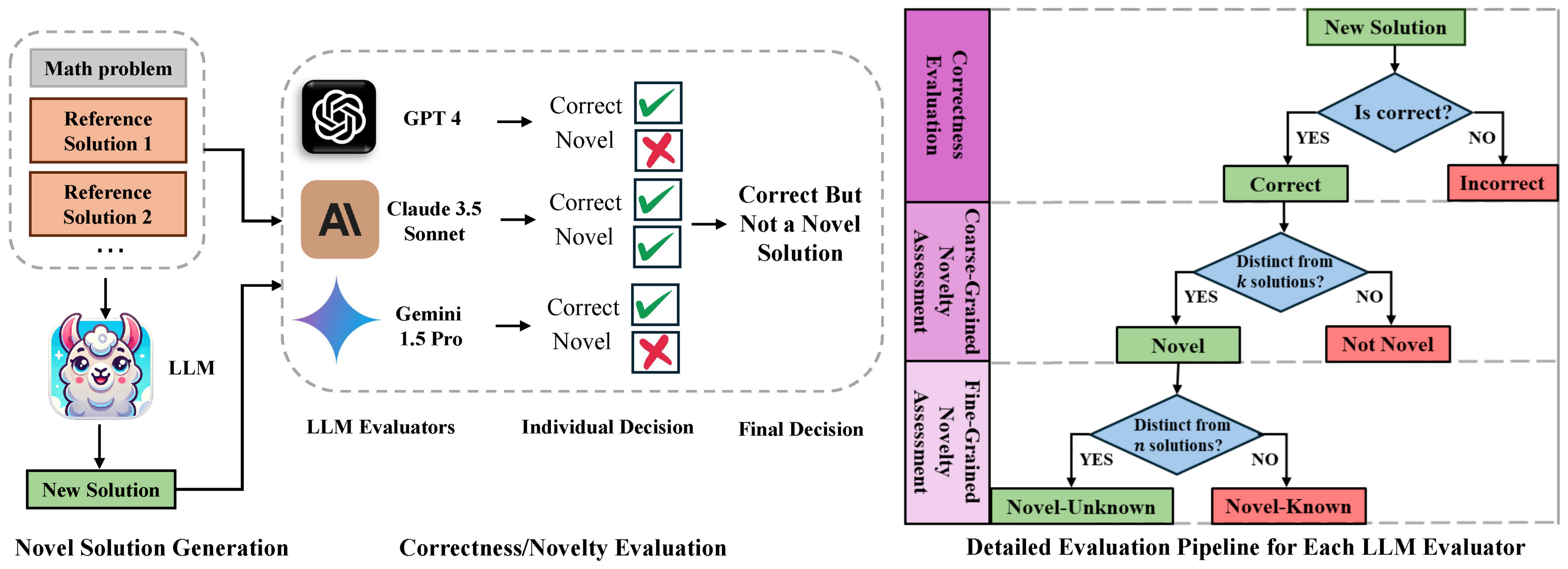

Figure 3: The framework includes solution generation (left) and the evaluation pipeline (middle). The flowchart of the detailed evaluation pipeline is illustrated on the right.

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.

Results & Key Findings

How effectively can the 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)

Table 2: Experimental results for various closed-source and open-source LLMs on the CreativeMath subset (↑ indicates that higher is better).

Key Insights:

- 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 the performance?

Correctness Ratio 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

Table 3: Correctness Ratio (C) across different models with varying numbers of reference solutions (*k*).

Novelty Ratio 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

Table 4: Novelty-to-Correctness Ratio (N/C) for different models based on the degree of solution availability (*n*−*k*).

- When *k* increases, the correctness ratio increases. (Align with few-shot learning).
- When *n-k* decreases, novelty-to-correctness ratio drops. This indicates tightening the constraints, making it harder for the model to generate new solutions.

How does difficult affect the 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

decreases

increases

Table 5: Average Correctness (C) and Novelty-to Correctness Ratio (N/C) for all LLMs when solving math problems of varying difficulty levels, with k = 1 across all competitions.

- 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 between novel solutions generated by LLMs

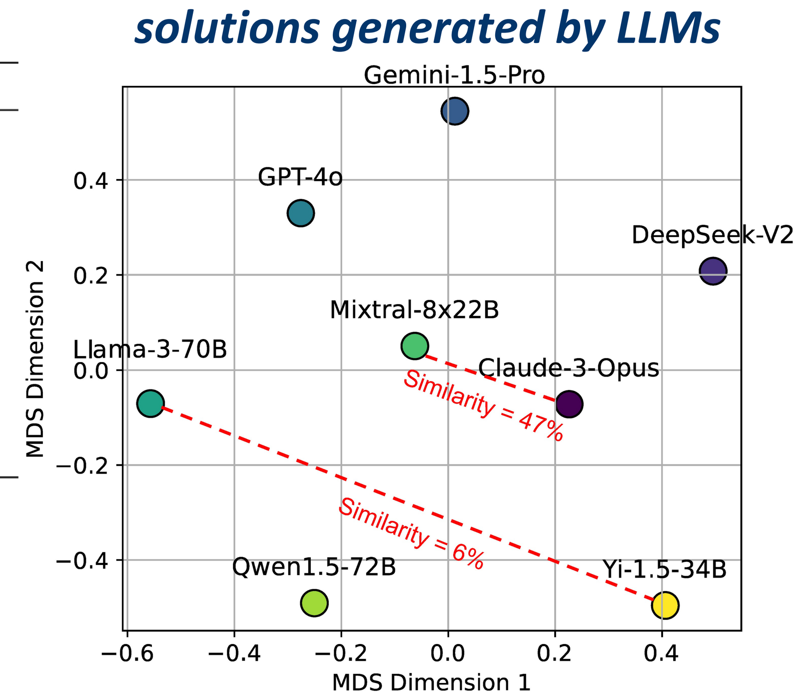

Figure 6: Similarity map between the novel solutions generated by different LLMs.

- 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.