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, Newark, USA
[2]The Pennsylvania State University, State College, PA, USA
{jy394, jg95, xz43, gwang}@njit.edu, wenpeng@psu.edu

Abstract

The mathematical capabilities of AI systems are complex and multifaceted. Most existing research has predominantly focused on the *correctness* of AI-generated solutions to mathematical problems. In this work, we argue that beyond producing correct answers, AI systems should also be capable of, or assist humans in, developing *novel solutions* to mathematical challenges. This study explores the creative potential of Large Language Models (LLMs) in mathematical reasoning, an aspect that has received limited attention in prior research. We introduce a novel framework and benchmark, CREATIVEMATH, which encompasses problems ranging from middle school curricula to Olympic-level competitions, designed to assess LLMs' ability to propose innovative solutions after some known solutions have been provided. Our experiments demonstrate that, while LLMs perform well on standard mathematical tasks, their capacity for creative problem-solving varies considerably. Notably, the Gemini-1.5-Pro model outperformed other LLMs in generating novel solutions. This research opens a new frontier in evaluating AI creativity, shedding light on both the strengths and limitations of LLMs in fostering mathematical innovation, and setting the stage for future developments in AI-assisted mathematical discovery.

Code — https://github.com/NJIT-AI-Center/CreativeMath

Introduction

In recent years, artificial intelligence has made significant strides, particularly in the development of Large Language Models (LLMs) capable of tackling complex problem-solving tasks. Models like GPT-4 and Gemini-1.5-Pro have demonstrated impressive proficiency on rigorous mathematical benchmarks (Ahn et al. 2024) such as GSM8K (Cobbe et al. 2021) and MATH (Hendrycks et al. 2021a), underscoring the evolving role of LLMs from simple text generators to sophisticated tools capable of engaging with high-level mathematical challenges. Beyond solving student-oriented math problems, leading mathematicians have begun exploring the use of LLMs to assist in tackling unresolved mathematical challenges (Romera-Paredes et al. 2024; Trinh et al. 2024). Despite these models' success in achieving high accuracy on existing mathematical datasets, their potential for creative problem-solving remains largely underexplored.

The standard definition of creativity, as articulated by (Runco and Jaeger 2012), emphasizes two essential criteria: novelty and usefulness. While correctness aligns with usefulness, evaluating novelty remains a challenge, especially in the domain of mathematics. Mathematical creativity goes beyond solving problems correctly; it involves generating novel solutions, applying unconventional techniques, and offering deep insights—areas traditionally associated with human ingenuity. Yet, most studies have focused primarily on correctness and efficiency, paying little attention to the innovative approaches LLMs might employ. Furthermore, creativity in mathematical problem-solving is rarely integrated into existing benchmarks, limiting our understanding of LLMs' full potential. The current research landscape lacks a comprehensive framework that evaluates both the accuracy and the creative capacity of LLMs. This gap highlights the need for new methodologies and benchmarks specifically designed to assess and cultivate the creative problem-solving abilities of LLMs in mathematics, which is the focus of this paper.

We created the dataset CREATIVEMATH, a comprehensive math benchmark that includes problems from middle school to Olympic-level competitions, each accompanied by multiple high-quality solutions ranging from straightforward to highly innovative approaches. Additionally, we designed a multi-stage framework to rigorously evaluate the creativity of LLMs in generating novel math solutions. This evaluation spans closed-source, open-source, and math-specialized LLMs, assessing both the correctness and novelty of their solutions based on different reference prior solutions.

Our evaluation revealed several interesting key insights: (1) Gemini-1.5-Pro excelled in generating unique solutions, with most correct answers being distinct from the provided references, while smaller and math-specialized models struggled with novelty. (2) Providing more reference solutions generally improved accuracy, with Gemini-1.5-Pro achieving perfect accuracy with four prior solutions. However, increased references made it harder for models to generate unique solutions, indicating a trade-off between leveraging existing knowledge and fostering creativity. (3) As math problem difficulty increased, LLM accuracy declined,

but successful solutions were more likely to be innovative, suggesting that tougher problems encourage creativity. (4) Analysis of solution similarity among different LLMs showed that models like Llama-3-70B and Yi-1.5-34B explored diverse approaches, while others like Mixtral-8x22B produced more similar solutions, highlighting the value of using a diverse set of LLMs to enhance originality.

This study lays the groundwork for future advancements in LLM math creativity. The major contributions include: (1) Introducing a new task—evaluating LLMs' mathematical creativity, (2) Creating the CREATIVEMATH dataset, (3) Developing a framework for assessing mathematical creativity in LLMs, and (4) Evaluating state-of-the-art LLMs, revealing key insights into their strengths and limitations.

Related Work

LLMs have demonstrated significant advancements in both mathematical reasoning and creative capabilities, making them increasingly powerful tools in a variety of domains. In the realm of mathematical reasoning, techniques such as prompt engineering, Chain-of-Thought (CoT) prompting, and program-aided language modeling have notably enhanced LLMs' abilities to solve complex problems (Brown 2020; Wei et al. 2022; Zhou et al. 2023). These approaches enable models to break down problems into more manageable steps, thereby improving their accuracy and reasoning depth. Moreover, specialized models like MathVerse (Zhang et al. 2024) and Internlm-Math (Ying et al. 2024), which are trained on extensive mathematical corpora, have achieved significant improvements in mathematical problem-solving performance (Lewkowycz et al. 2022; Ying et al. 2024). Benchmarks such as GSM8K and MATH further provide a structured means to evaluate and compare these advancements, highlighting the continuous progress in this area (Cobbe et al. 2021; Hendrycks et al. 2021b).

In terms of creativity, LLMs have shown remarkable prowess across diverse fields. They have excelled in generating high-quality, human-like content, ranging from code generation (Ni et al. 2023; Liu et al. 2024a) and music composition (Yuan et al. 2024) to literature (Gómez-Rodríguez and Williams 2023; Liu et al. 2024b) and educational tools (Lan and Chen 2024; Orenstrakh et al. 2023). Creativity in LLMs is often evaluated using frameworks like Margaret Boden's taxonomy (Boden 2004), which categorizes creativity into combinational, exploratory, and transformational types. While LLMs perform well in combinational creativity, achieving true transformational creativity remains a significant challenge (Franceschelli and Musolesi 2023). Psychological metrics such as the Torrance Tests of Creative Thinking (TTCT) (Torrance 1966), where LLMs have demonstrated high fluency, originality, and flexibility. However, the applicability of these traditional creativity metrics to AI systems is still a topic of debate, as they were originally designed to assess human creativity (Zhao et al. 2024).

Techniques such as associative thinking have been employed to enhance the creative output of LLMs further, although challenges remain in ensuring that these models can meaningfully integrate unrelated concepts (Mehrotra, Parab, and Gulwani 2024). The ethical and legal implications of AI-generated creativity continue to be a significant area of concern, underscoring the need for ongoing research to refine evaluation methods and address societal impacts (Lofstead 2023).

CREATIVEMATH Curation

This section details the creation, collection, and processing of our dataset **CreativeMath**, which comprises high-quality mathematical problems from various competitions and their numerous solutions. The dataset is diverse, encompassing a broad range of mathematical topics and problem types, and covers difficulty levels from middle school to Olympiad level. It includes problems from eight major US competitions: AMC 8, AMC 10, AMC 12, AHSME, AIME, USAJMO, USAMO, and IMO[1].

Data Collection. The dataset was sourced from the Art of Problem Solving (AoPS)[2], a platform offering the most comprehensive collection of problems from various math competitions, along with multiple solutions contributed by participants over the years. As the most popular and sought-after resource for math competitors, AoPS effectively functions as a natural crowdsourcing platform. It uniquely approximates the complete set of viable human solutions for each problem, with later contributors often building on earlier ones.

We meticulously scraped data from eight competitions, ranging from middle school level to Olympic-level, to capture the breadth of mathematical challenges and the depth of solution strategies available.

Data Cleaning. To ensure the integrity and reliability of the dataset, we conducted a rigorous data cleaning procedure. We accurately extracted LaTeX-formatted problems and solutions from HTML, ensuring their correct representation. Irrelevant comments were removed to make each problem and solution clear and self-sufficient. Samples with images, problems without solutions, or incomplete entries were manually removed from the dataset. After this process, the dataset comprises 6,469 mathematical problems and 14,223 solutions. Each problem in the dataset is tagged with detailed metadata, including difficulty level, math category, and problem type. Difficulty levels and problem types were assigned based on official competition data, while the math category were determined using the Llama-3-70B model.

Dataset Analysis. As shown in Figure 1, the problem distribution inside CreativeMath reveals that Algebra and Geometry are the most represented categories across all com-

[1]AMC 8: American Mathematics Competition for grade 8 and below, AMC 10: American Mathematics Competition for grade 10 and below, AMC 12: American Mathematics Competition for grade 12 and below, AHSME: American High School Mathematics Examination, AIME: American Invitational Mathematics Examination, USAJMO: USA Junior Mathematical Olympiad, USAMO: USA Mathematical Olympiad, IMO: International Mathematical Olympiad.

[2]Art of Problem Solving. "AoPS Wiki", https://artofproblemsolving.com/wiki/.

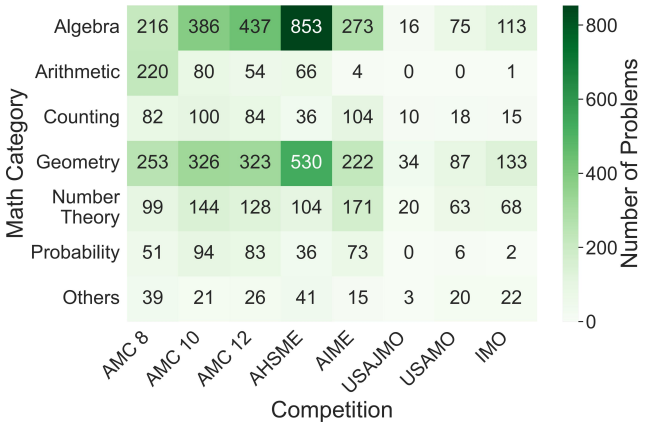

Figure 1: Distribution of problems across different math categories and competitions in the CreativeMath dataset.

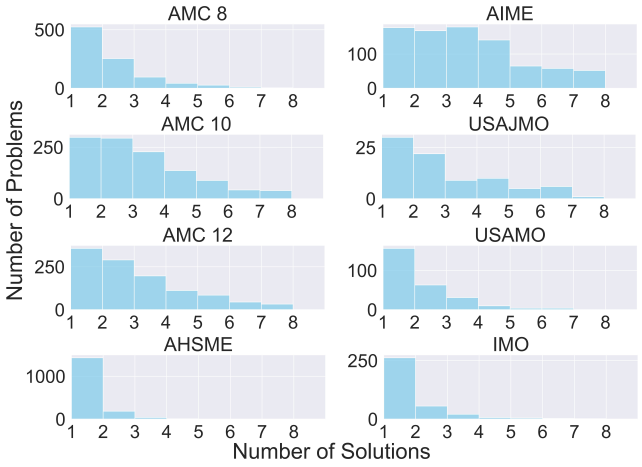

Figure 2: Distribution of the number of solutions per problem across different competitions.

petitions. The number of solutions across different competitions, as depicted in Figure 2, reflects the varying complexity of the problems. Medium-difficulty competitions like AMC 10, AMC 12, and AIME typically have a larger number of solutions, as these problems allow for a variety of approaches. In contrast, simpler competitions like AMC 8 tend to have fewer solutions due to the straightforward nature of the problems, which often have limited methods of solving. Olympic-level competitions such as USAJMO, USAMO, and IMO also see fewer solutions, likely due to the high complexity of the problems, which limits the number of viable solving strategies.

Methods

Our approach consists of a multi-stage pipeline designed to evaluate the novelty of mathematical solutions generated by an LLM. The methodology is structured into four key stages: *Novel Solution Generation*, *Correctness Evaluation*, *Coarse-Grained Novelty Assessment*, and *Fine-Grained Novelty Assessment*. This comprehensive pipeline illustrated in Figure 3 ensures that the generated solutions are not only correct but also exhibit a meaningful degree of novelty relative to the reference solutions. The sample prompts and LLMs' responses are provided in the Appendix.

Novel Solution Generation

The first stage of the methodology aims to generate novel solutions for the given mathematical problem using LLM. For each problem, a subset of k reference solutions (where k ranges from 1 to n, with n representing the total number of available reference solutions) is sequentially selected based on the order in which competitors uploaded their solutions on the website. Earlier solutions are often the most common and intuitive, while later ones may build on previous methods, offer improvements, or introduce entirely novel algorithms. Consequently, as k increases, the difficulty in generating new and innovative solutions also increases.

To ensure clarity and consistency in both prompting and evaluating the novelty of generated solutions, we define a set of criteria agreed upon in consultation with several mathematicians. These criteria guide both the generation and the evaluation process and are used to assess the distinctiveness of the solutions. The criteria are as follows:

- **Methodological Differences**: If the methods used to arrive at the solutions are fundamentally different (e.g., algebraic manipulation versus geometric reasoning), the solutions are considered distinct.
- **Intermediate Step Variation**: Even if the final results are identical, if the intermediate steps or processes involved in reaching those solutions differ significantly, the solutions are considered novel.
- **Assumptions and Conditions**: Solutions that rely on different assumptions, initial conditions, or constraints are treated as distinct.
- **Generalization**: A solution that generalizes to a broader class of problems is considered novel compared to one that is specific to certain conditions.
- **Complexity**: If one solution is notably simpler or more complex than another, they are regarded as different, even if they lead to the same final result.

These criteria, also illustrated in Figure 4, are embedded into the prompt used to guide the LLM in generating novel solutions. The reference solutions provided to the model aim to capture a variety of approaches, and the LLM is instructed to output a new solution that is distinct according to the defined criteria. The prompt emphasizes generating solutions that use different problem-solving methods, distinct intermediate steps, and variations in assumptions or generalizability.

As part of this process, to avoid influencing the judgment of evaluators during the subsequent evaluation stage, transition sentences and justifications explaining why the new solution is distinct from the reference solutions are manually removed. Only the newly generated solution is presented for evaluation.

Correctness and Novelty Evaluation

To rigorously evaluate the correctness and novelty of the generated solutions, we employ three leading LLMs—GPT-

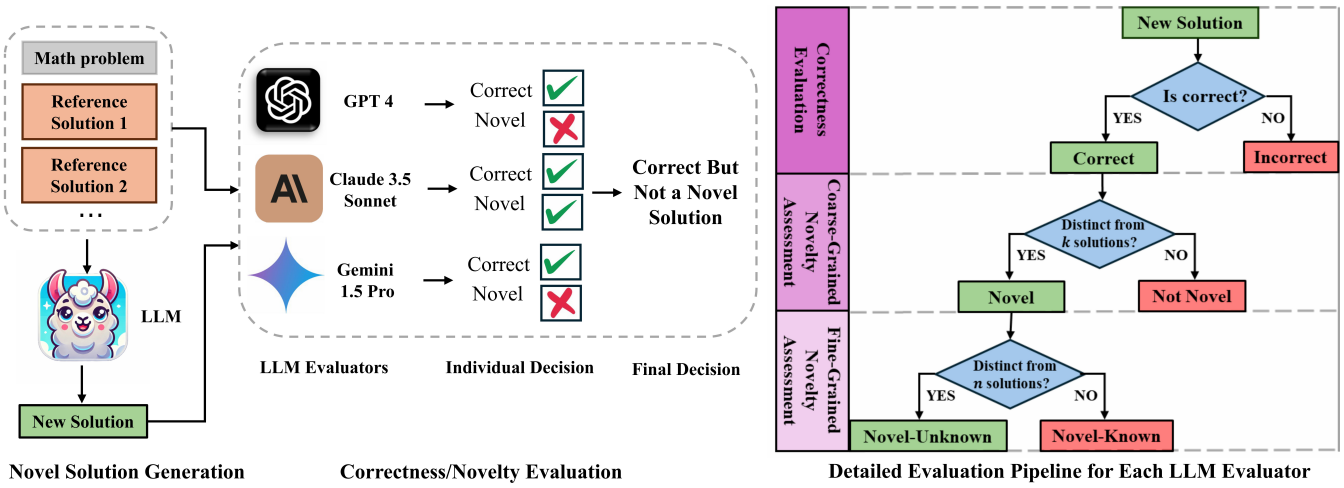

Figure 3: The framework includes solution generation (left) and the evaluation pipeline (middle). The flowchart of the detailed evaluation pipeline is illustrated on the right.

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.

4, Claude 3.5 Sonnet, and Gemini 1.5 Pro—as **LLM Evaluators**, recognized among the strongest models available. These LLM Evaluators collaboratively assess the solutions following the framework illustrated in Figure 3 (middle). Each LLM Evaluator adheres to the flowchart depicted in Figure 3 (right) to systematically evaluate the generated solutions across three dimensions:

- **Correctness**: The solution must first be validated for correctness, ensuring it produces the correct result for the problem. Only correct solutions proceed to the novelty assessment stages.
- **Coarse-Grained Novelty**: If the solution is correct, it is then evaluated for novelty against a subset of k reference solutions. A solution is deemed **novel** if it is distinct from these k solutions.
- **Fine-Grained Novelty**: A solution deemed novel in the coarse-grained assessment undergoes further evaluation against the entire set of n human-provided solutions. This stage distinguishes between:
  - **Novel-Unknown**: A solution that is distinct from all n human-generated solutions, representing a truly original contribution.
  - **Novel-Known**: A solution that is distinct from the k reference solutions but similar to others in the remaining $n - k$ solutions.

Evaluation Strategy We apply different strategies for correctness and novelty evaluation to ensure both rigor and practicality. For correctness, only solutions unanimously deemed correct by all LLM Evaluators proceed to the novelty assessment, ensuring that only fully reliable solutions are considered. Given the subjective nature of assessing novelty, we use a majority voting strategy, which balances diverse perspectives and effectively identifies genuinely innovative solutions without being overly restrictive.

Correctness Evaluation Once a solution is generated, the first essential step is to verify its correctness. The newly generated solution, along with the original problem and a set of reference solutions, is evaluated by the LLM Evaluators using the prompt shown in Figure 5, top. The LLM Evaluators determine if the solution leads to the correct outcome, with responses of "YES" indicating correctness and "NO" indicating otherwise. Only solutions unanimously validated as correct by all LLM Evaluators advance to the novelty assessment stages.

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, such as algebraic manipulation versus geometric reasoning, they can be considered distinct;
...

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: The prompt templates for evaluating the correctness (top) and novelty (bottom) of the generated solution. The criteria for evaluating the novelty are rephrased from the same criteria applied during the novel solution generation process to ensure alignment.

Coarse-Grained Novelty Assessment After correctness is established, the next step is to evaluate the solution's novelty at a coarse level. This involves comparing the generated solution against the k reference solutions. The LLM Evaluators assess whether the solution employs distinct approaches or methods that differentiate it from the provided references, using the prompt (Figure 5, bottom). If the solution is considered novel relative to the k reference solutions, it is marked as "YES" and proceeds to the fine-grained novelty assessment.

Fine-Grained Novelty Assessment In the final stage, the solution undergoes a fine-grained novelty evaluation to determine its originality in comparison to all n human-generated solutions. This assessment uses the same prompt as the coarse-grained novelty assessment but changes the reference solutions from the subset 1 to k to the complementary set $k+1$ to n. The evaluation focuses on whether the solution introduces new insights, methods, or approaches that surpass existing human solutions in terms of innovation, complexity, or generalizability. The outcome categorizes the solution as either a unique contribution or as similar to existing human-generated solutions.

Experiment

In this section, we conduct extensive experiments and analyses to show the performance of ten the-state-of-the-art LLMs in math problem solving. We also address several research questions.

Dataset

We selected a subset from our **CreativeMath** dataset for this study. For each competition, 50 samples were randomly chosen to ensure a representative evaluation of the LLMs' performance. The datasets were meticulously curated to ensure that when the problem and all reference solutions were included in the novel solution generation prompt, the total token count did not exceed 3K tokens. This approach allowed for 1K tokens to be reserved for generation, accommodating the token limits of models like DeepSeek-Math-7B-RL, which has a 4K-token capacity. In total, the dataset comprises 400 math problems and 605 solutions, forming 605 distinct samples with k varying from 1 to 5.

Large Language Models

In this study, we explore the ability of various LLMs to generate novel and creative solutions in mathematical problem-solving. The LLMs selected for this research have demonstrated superior performance on key mathematical benchmarks, such as GSM8K and the MATH dataset, outperforming other models of similar parameter scale. We include three leading close-sourced models—GPT-4o (Version 2024-05-13) (OpenAI 2024), Claude-3-Opus (Version 2024-02-29) (Anthropic 2024), and Gemini-1.5-Pro (Reid et al. 2024)—which are renowned for their excellence in complex mathematical reasoning. To ensure a comprehensive evaluation, we also incorporate five top-ranking open-source instruction-tuned LLMs in math reasoning: Llama-3-70B (Meta AI 2024), Qwen1.5-72B (Bai et al. 2023), Yi-1.5-34B (Young et al. 2024), Mixtral-8x22B-v0.1 (Mistral AI 2024), and DeepSeek-V2 (DeepSeek-AI 2024). Furthermore, two specialized mathematical instruction LLMs, DeepSeek-Math-7B-RL (Shao et al. 2024) and Internlm2-Math-20B (Ying et al. 2024), are included for their advanced capabilities in mathematical reasoning. By selecting these models, we aim to gain a comprehensive understanding of whether their demonstrated excellence in math benchmarks also reflects an enhanced capacity for generating novel solutions.

Implementation Details

For the closed-source LLMs and DeepSeek-V2, we utilized API calls provided by their respective platforms. Open-source LLMs were run using the Hugging Face library on one to four NVIDIA A100 (80G) GPUs, depending on the model's memory requirements. To ensure reproducibility, all experiments were conducted using the greedy decoding strategy, adhering to the recommended settings provided on the official Hugging Face pages or the models' respective papers. The system prompt followed the guidelines outlined in the models' documentation, with the maximum number of new tokens set to 1024. This standardized approach ensures

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.

consistent and reliable evaluation across all models used in our study.

Evaluation Metrics

To assess the effectiveness of LLMs in generating novel solutions, we define several evaluation metrics, as outlined in Table 1. These metrics capture key aspects of the solutions, including correctness, different levels of novelty, and the relationship between novelty and correctness. Importantly, novelty is only considered if the solution is correct, and the Correctness Ratio, Novelty Ratio, and Novel-Unknown Ratio are calculated based on all generated solutions to ensure a consistent evaluation.

Results & Discussions

We introduce our results in the context of each of our four research questions and discuss our main findings.

$\mathbb{Q}_1$: Given a math problem with n known solutions, and an LLM provided with the problem along with k of those solutions, how effectively can the LLM generate a novel solution?

Analysis of Coarse-Grained Novelty Table 2 demonstrates the superior performance of Gemini-1.5-Pro across all evaluated metrics, particularly in its ability to generate novel solutions. With a Novelty Ratio (N) of 66.94% and a Correctness Ratio (C) of 69.92%, Gemini-1.5-Pro not only generates a high number of correct solutions but also ensures that most of these are novel. The model's Novelty-to-Correctness Ratio (N/C) of 95.75% indicates that nearly all correct solutions it produces are distinct from the provided reference solutions.

Llama-3-70B and Claude-3-Opus also perform well in terms of N, with Llama-3-70B achieving a noteworthy N/C of 82.87%. This contrasts sharply with models like GPT-4o, DeepSeek-V2, and Mixtral-8x22B, which, despite similar C values, have N/C ratios below 50%. This discrepancy highlights significant differences in the ability of LLMs to generate novel solutions, even when their correctness levels are comparable. Notably, Llama-3-70B outperforms closed-source models Claude-3-Opus and GPT-4o, suggesting that open-source LLMs can achieve competitive novelty generation capabilities.

In contrast, smaller models like Yi-1.5-34B and specialized math-tuned models such as Deepseek-Math-7B-RL and Internlm2-Math-20B exhibit lower C and N/C ratios. This outcome is consistent with scaling laws (Kaplan et al. 2020), where large models generally outperform compared to small ones. The low N/C in these math-specialized models suggests that their fine-tuning for mathematical tasks may limit their adaptability in generating novel solutions outside of their specialized domain.

Analysis of Fine-Grained Novelty The high average Novel-Unknown to Novelty Ratio (N_u/N) of 95% across models indicates that the vast majority of novel solutions generated are distinct from any available human solutions. This suggests a substantial potential for these models to contribute genuinely original and innovative solutions that extend beyond the existing human knowledge base. The ability to produce solutions that are not only correct but also novel, surpassing human ingenuity, underscores the LLMs' capacity to explore new solution spaces. This makes them powerful tools for advancing fields that demand creative problem-solving.

Distinctions Between Novel Solution Generation and Math Problem Solving Novel solution generation and traditional math problem-solving differ fundamentally in their structure and evaluation criteria. In traditional math problem solving, typically using few-shot settings with $k = 4$ fixed reference examples, the task is to solve a new, unseen problem with correctness as the sole criterion. The provided examples consist of different problems and their solutions, with the solution to the target problem being unknown.

In contrast, novel solution generation involves a fixed problem where the model is given varying numbers of reference solutions with $1 \leq k \leq n$. Here, the solution is known, and the model must not only solve the problem correctly but also generate solutions that are distinct from the provided references. This requirement for distinctiveness adds a layer of complexity, challenging the model's ability to innovate beyond mere correctness.

This distinction is evident in the evaluation metrics. For example, while GPT-4o achieves a high accuracy of 76.6% on the MATH benchmark in Table 2, it performs poorly on novelty metrics, indicating a limited ability to generate distinct solutions despite its problem-solving accuracy. This contrast underscores the more stringent demands of novel solution generation, where models must demonstrate creativity and innovation in addition to correctness.

$\mathbb{Q}_2$: How does the number of provided solutions, k, affect the LLM's performance in generating new solutions?

Impact of k on Correctness This section examines how increasing the number of provided reference solutions (k) affects the correctness of generated solutions, as shown in Table 3. Across most models, there is a clear trend of improved correctness with larger k values. For example, Gemini-1.5-Pro reaches 100% correctness at $k = 4$, demonstrating its ability to effectively utilize additional examples. This trend is consistent with findings in few-shot learning, where more examples typically lead to better model performance (Brown 2020). Models like Llama-3-70B and DeepSeek-V2 show moderate improvements with increased k, though the gains

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 MultiMath subset (↑ indicates that higher is better). The best-performing models in the open-source and closed-source categories for each evaluation metric are respectively highlighted. MATH column represents the accuracy on MATH datasets with 4-shot (CoT) setting as reported by the corresponding papers or websites of the LLMs. Refer to Table 1 for detailed definitions of the evaluation metrics used.

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). Sample sizes for $k=1$ to $k=4$ are 400, 154, 42, and 8, respectively.

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 ($n-k$). Higher values of $n-k$ indicate scenarios with fewer provided solutions, which are easier for the LLM.

are less pronounced compared to Gemini-1.5-Pro. In contrast, models like Qwen1.5-72B and Yi-1.5-34B show minimal increases in correctness, potentially due to variability introduced by smaller sample sizes at higher k values.

Impact of the Degree of Solution Availability ($n-k$) on Novelty The degree of solution availability, denoted by $n-k$, represents the gap between the total available solutions and those provided to the model. A higher $n-k$ means fewer distinct solutions are given, leaving more room for the model to explore and innovate. This typically results in fewer constraints, facilitating the generation of novel outputs. As k increases and $n-k$ decreases, the model is exposed to more reference solutions, tightening the constraints and making it harder to generate novel solutions. This pattern is evident in Table 4, where models generally show higher Novelty-to-Correctness Ratios (N/C) at higher $n-k$ values, with Gemini-1.5-Pro achieving a perfect N/C at $n-k=2$. However, as $n-k$ decreases, the ability to produce novel solutions diminishes. This mirrors human problem-solving, where creativity often diminishes when more examples are provided, as the model (or individual) must work within tighter constraints.

$\mathbb{Q}_3$***: How does the creativity of LLMs vary when solving math problems of varying difficulty levels?***

We analyzed the correctness (C) and Novelty-to-Correctness Ratio (N/C) of all LLMs across competitions of different difficulty levels, focusing on problems where $k=1$ to ensure consistency. As shown in Table 5, as problem difficulty increases, the correctness of LLMs consistently decreases, dropping from 71.80% on AMC 8 problems to around 35% on more challenging competitions like USAMO and IMO. Conversely, the N/C ratio increases with difficulty, from 55.39% on easier problems to 83.01% on the most difficult ones. This suggests that while LLMs struggle with accuracy on harder problems, they are more likely to generate novel solutions when they do succeed. The observed trend indicates a shift in the balance between familiarity and innovation: as problem difficulty rises, LLMs are pushed to rely less on familiar strategies and more on creative problem-solving. This complex interplay between familiarity and innovation becomes more pronounced with increasing problem difficulty, leading to a higher likelihood of novel solutions.

$\mathbb{Q}_4$***: When different LLMs are given the same math problem and k solutions, how likely are the new solutions generated by these LLMs to be identical or distinct? Additionally, how does the pairwise similarity between the so-***

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

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.

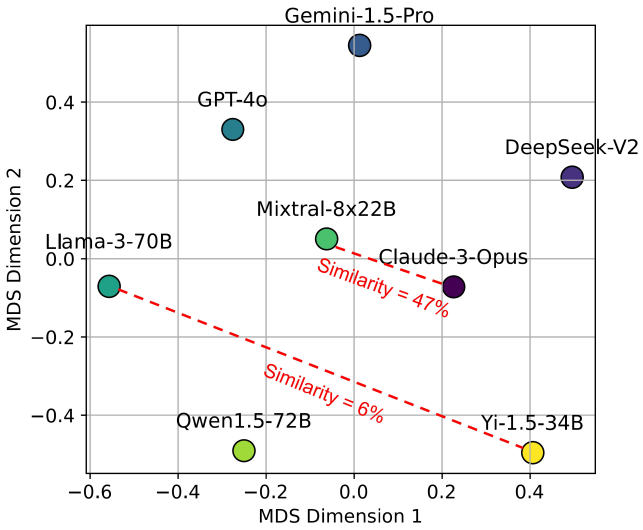

Figure 6: Similarity map between the novel solutions generated by different LLMs.

lutions generated by different LLMs inform us about their tendencies to produce similar outputs?

To explore the tendency of different LLMs to generate novel solutions, we first measured pairwise similarity between the outputs of various models. We conducted an experiment using 17 samples where all included LLMs were capable of generating novel solutions. Math-specialized LLMs were excluded due to their low novelty ratios. For each pair of LLMs, we used the same prompt as in the novelty assessment, but replaced the reference solution with the solution generated by one LLM and the new solution with that generated by another LLM. The pairwise similarity was determined based on whether the solutions were distinct ("YES") or similar ("NO"). The similarity score for each LLM pair was computed as the ratio of similar solutions to the total number of samples (17).

We applied Multidimensional Scaling (MDS) to the pairwise similarity matrix, mapping the LLMs into a two-dimensional space. As illustrated in Figure 6, the similarity map reveals a general trend of low similarity between the novel solutions generated by different LLMs. The most distinct pair, Llama-3-70B and Yi-1.5-34B, shows only a 6% similarity, indicating that these models explore vastly different solution spaces. On the other hand, the most similar pairs—Mixtral-8x22B with GPT-4o and Mixtral-8x22B with Claude-3-Opus—each show a 47% similarity. Mixtral-8x22B, positioned centrally in the similarity map, tends to produce solutions that are slightly more similar to those of other models. This analysis suggests that leveraging multiple LLMs positioned on the periphery of the similarity map could be a promising approach to generate diverse novel solutions. These models, exploring vastly different solution spaces, are likely to enhance the efficiency and breadth of problem-solving strategies.

Conclusion

In this study, we introduced the **CreativeMath** dataset and developed a comprehensive framework that encompasses both the generation of novel solutions by LLMs and their rigorous evaluation. This framework is designed to assess the creative potential of LLMs in mathematical problem-solving, systematically distinguishing between solutions that are merely correct and those that offer genuinely innovative approaches. Our findings reveal significant variability in the creative abilities of state-of-the-art LLMs, emphasizing the importance of advancing AI systems that not only solve problems accurately but also contribute original insights. We encourage future research to delve deeper into methodologies for uncovering and assessing the creative capabilities of LLMs, particularly in complex and abstract domains like mathematics.

Acknowledgments

We would like to thank Suraj Patel, Venkata Sai Lakshman Palli, and Aadish Jain from New Jersey Institute of Technology for their assistance with data cleaning and dataset statistical analysis.

References

01-ai. 2024. Hugging Face: Yi-1.5-34B-Chat. https://huggingface.co/01-ai/Yi-1.5-34B-Chat/. Accessed: 2024-08-13.

Ahn, J.; Verma, R.; Lou, R.; Liu, D.; Zhang, R.; and Yin, W. 2024. Large Language Models for Mathematical Reasoning: Progresses and Challenges. In Falk, N.; Papi, S.; and Zhang, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024: Student Research Workshop, St. Julian's, Malta, March 21-22, 2024*, 225–237. Association for Computational Linguistics.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Accessed: 2024-08-13.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Boden, M. A. 2004. The creative mind: Myths and mechanisms.

Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

DeepSeek-AI. 2024. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. arXiv:2405.04434.

Franceschelli, G.; and Musolesi, M. 2023. On the creativity of large language models. *arXiv preprint arXiv:2304.00008*.

Gómez-Rodríguez, C.; and Williams, P. 2023. A confederacy of models: A comprehensive evaluation of LLMs on creative writing. *arXiv preprint arXiv:2310.08433*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021a. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021b. Measuring Mathematical Problem Solving With the MATH Dataset. *NeurIPS*.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Lan, Y.-J.; and Chen, N.-S. 2024. Teachers' agency in the era of LLM and generative AI. *Educational Technology & Society*, 27(1): I–XVIII.

Lewkowycz, A.; Andreassen, A.; Dohan, D.; Dyer, E.; Michalewski, H.; Ramasesh, V.; Slone, A.; Anil, C.; Schlag, I.; Gutman-Solo, T.; et al. 2022. Solving quantitative reasoning problems with language models, 2022. *URL https://arxiv. org/abs/2206.14858*.

Liu, J.; Xia, C. S.; Wang, Y.; and Zhang, L. 2024a. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*, 36.

Liu, Y.; Chen, S.; Cheng, H.; Yu, M.; Ran, X.; Mo, A.; Tang, Y.; and Huang, Y. 2024b. How ai processing delays foster creativity: Exploring research question co-creation with an llm-based agent. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–25.

Lofstead, J. 2023. Economic, Societal, Legal, and Ethical Considerations for Large Language Models. In *2023 Fifth International Conference on Transdisciplinary AI (TransAI)*, 155–162. IEEE.

Mehrotra, P.; Parab, A.; and Gulwani, S. 2024. Enhancing Creativity in Large Language Models through Associative Thinking Strategies. *arXiv preprint arXiv:2405.06715*.

Meta AI. 2024. Meta LLaMA 3. https://ai.meta.com/blog/meta-llama-3/. Accessed: 22-October-2024.

Mistral AI. 2024. Mixtral 8x22B: The New Frontier in AI Models. https://mistral.ai/news/mixtral-8x22b/. Accessed: 2024-08-13.

Ni, A.; Iyer, S.; Radev, D.; Stoyanov, V.; Yih, W.-t.; Wang, S.; and Lin, X. V. 2023. Lever: Learning to verify language-to-code generation with execution. In *International Conference on Machine Learning*, 26106–26128. PMLR.

OpenAI. 2024. Hello GPT-4o! https://openai.com/index/hello-gpt-4o/. Accessed: 2024-08-13.

Orenstrakh, M. S.; Karnalim, O.; Suarez, C. A.; and Liut, M. 2023. Detecting llm-generated text in computing education: A comparative study for chatgpt cases. *arXiv preprint arXiv:2307.07411*.

Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillicrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Romera-Paredes, B.; Barekatain, M.; Novikov, A.; Balog, M.; Kumar, M. P.; Dupont, E.; Ruiz, F. J. R.; Ellenberg, J. S.; Wang, P.; Fawzi, O.; Kohli, P.; and Fawzi, A. 2024. Mathematical discoveries from program search with large language models. *Nat.*, 625(7995): 468–475.

Runco, M. A.; and Jaeger, G. J. 2012. The standard definition of creativity. *Creativity research journal*, 24(1): 92–96.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Zhang, M.; Li, Y.; Wu, Y.; and Guo, D. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Torrance, E. P. 1966. Torrance tests of creative thinking. *Educational and psychological measurement*.

Trinh, T. H.; Wu, Y.; Le, Q. V.; He, H.; and Luong, T. 2024. Solving olympiad geometry without human demonstrations. *Nat.*, 625(7995): 476–482.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Ying, H.; Zhang, S.; Li, L.; Zhou, Z.; Shao, Y.; Fei, Z.; Ma, Y.; Hong, J.; Liu, K.; Wang, Z.; Wang, Y.; Wu, Z.; Li, S.; Zhou, F.; Liu, H.; Zhang, S.; Zhang, W.; Yan, H.; Qiu, X.; Wang, J.; Chen, K.; and Lin, D. 2024. InternLM-Math: Open Math Large Language Models Toward Verifiable Reasoning. arXiv:2402.06332.

Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Li, H.; Zhu, J.; Chen, J.; Chang, J.; et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Yuan, R.; Lin, H.; Wang, Y.; Tian, Z.; Wu, S.; Shen, T.; Zhang, G.; Wu, Y.; Liu, C.; Zhou, Z.; et al. 2024. ChatMusician: Understanding and Generating Music Intrinsically with LLM. *arXiv preprint arXiv:2402.16153*.

Zhang, R.; Jiang, D.; Zhang, Y.; Lin, H.; Guo, Z.; Qiu, P.; Zhou, A.; Lu, P.; Chang, K.-W.; Gao, P.; et al. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*.

Zhao, Y.; Zhang, R.; Li, W.; Huang, D.; Guo, J.; Peng, S.; Hao, Y.; Wen, Y.; Hu, X.; Du, Z.; et al. 2024. Assessing and understanding creativity in large language models. *arXiv preprint arXiv:2401.12491*.

Zhou, J.; Rao, S.; Gao, L.; Zhang, C.; Tang, H.; Li, Y.; and Chan, F. T. 2023. Solving many-task optimization problems via online intertask learning. *Expert Systems with Applications*, 225: 120110.