MathNet: A GLOBAL MULTIMODAL BENCHMARK FOR MATHEMATICAL REASONING AND RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Mathematical problem solving remains a challenging test of reasoning for large language and multimodal models, yet existing benchmarks are limited in size, language coverage, and task diversity. We introduce *MathNet*, a large-scale, high-quality, multilingual, and multimodal dataset of Olympiad-level problems. MathNet spans 40 countries, 10 languages, and two decades of competitions, comprising 17,512 **expert-authored problems with solutions** across diverse domains.

MathNet supports three tasks: (i) *mathematical comprehension*, (ii) *mathematical retrieval*, an underexplored but essential capability , and (iii) Math RAG to test how retrieval augmented generation can improve problem solving. For retrieval, we construct 39K pairs of mathematically equivalent problems to enable equivalence-based evaluation, in addition 70 pairs of expert curated. Experimental results show that even state-of-the-art reasoning models (76.8% for GPT-5 and 46.8% for Claude 4.5 Opus) are challenged, while embedding models struggle to retrieve equivalent problems. Finally, we show that LLM performance in RAG-based math problem solving varies noticeably with the quality of retrieved context, which shows that more community effort is needed in this domain.

MathNet provides the largest high-quality Olympiad dataset and the first retrieval benchmark for problem equivalence. We publicly release both the dataset and benchmark at http://mathnet.netlify.app.

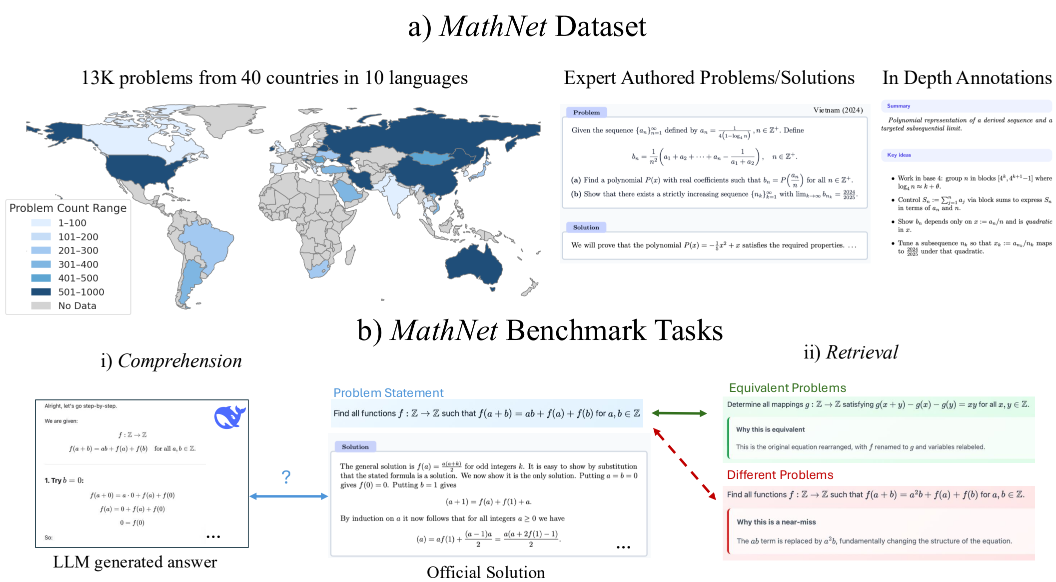


Figure 1: **Overview of MathNet.** (a) Dataset of 17K Olympiad-level problems across 40 countries, 10 languages, and 700 competitions with expert-authored solutions. (b) Benchmark tasks: comprehension (solution generation) and retrieval (equivalence-based problem matching).

1 INTRODUCTION

Recent LLMs and LMMs have made rapid strides on mathematical reasoning benchmarks, from grade-school arithmetic to competition mathematics (Cobbe et al., 2021; Hendrycks et al.; Achiam et al., 2023). This year, public reports claimed unprecedented gold-medal–level performance at the International Mathematical Olympiad (IMO) by advanced AI systems (Luong et al., 2025; Zhihong Shao, 2025). Moreover, there have been multiple incidents of AI systems reportedly solving open mathematical problems (Nie et al., 2025; Feldman & Karbasi, 2025).

Despite these advances, progress in the research community remains constrained by the lack of open, high quality, and diverse benchmarks. Existing Olympiad-level datasets are typically drawn from community platforms such as AoPS and are predominantly sourced from few competitions in the U.S and China, and they are small in scale (see Table 1). To address this gap, we present *MathNet*: the first large-scale, multilingual, and multimodal dataset of Olympiad-level problems. Curated over two decades from 40 countries and spanning 10 languages, *MathNet* comprises 17,512 problems with official solutions written by experts across a wide range of mathematical domains. Its scale, diversity, and expert quality provide an unprecedented foundation for exploring mathematical generalization and analogical reasoning.

Building on this foundation, we focus on two main tasks *Math Comprehension*: the ability of solving mathematical problems similar to all previous benchmark, and *Math Retrieval*, a fundamental yet underexplored capability. *Math Retrieval* is the abality of retrieving "mathematically" equivalent or related problems. Unlike existing semantic retrieval (Izacard et al., 2021; Khattab & Zaharia, 2020; Formal et al., 2021), mathematical retrieval must be sensitive to symbolic structure, invariances, and transformations. For example, the problem of solving $x^2 + y^2 = 1$ is equivalent to one that poses $\sqrt{a^2 + b^2} = 1$, or to a geometric formulation constraining a 2D vector to unit norm $|u|^2 = 1$. Crucially, however, these are not equivalent to solving $x + y = 1$. Current retrieval models fail to make this distinction: due to superficial lexical overlap (Das et al., 2025), they often rank a problem containing $x + y = 1$ as closer to $x^2 + y^2 = 1$ than to the truly equivalent formulations.

This challenge is evident in mathematical practice. In the IMO pipeline, more than a hundred countries propose original problems annually; a shortlist is debated and six are ultimately selected. Despite rigorous vetting, near-duplicates and thematic overlaps occasionally emerge, since existing tools cannot reliably surface mathematical equivalences across languages, formats, and notations. More broadly, contest success is often viewed as only weakly correlated with the deeper, sustained reasoning required for research mathematics (Gemstones, 2020).

A similar difficulty arises for research search. For example, a mathematician interested in finding a bound for consecutive primes might want to check if someone showed a result like $p_{n+1} - p_n \leq C(\log p_n)^2$, where p_n denotes the n-th prime number and C is a constant. In their search, they must typically look using paraphrases like "upper bounds on prime gaps" rather than by the symbolic form itself. Existing MathIR and formula-search systems attempt to bridge this gap: for example, Vemuganti et al. (2025) explore structural enrichments in formula only retrieval, however, it's not built to support more complex langague interleaved with math.

This paper introduces *MathNet*, a benchmark designed to evaluate *math-aware retrieval* and its role in reasoning. Our contributions are:

1. **Dataset.** A 17K-problem corpus of Olympiad-style math with aligned LaTeX and natural-language statements, expert solutions, and metadata spanning 40+ countries, 10 languages.
2. **New Annotations and Similarity Axes.** 39,078 synthetic problem pairs that are mathematically equivalent, in addition to 70 curated problem pairs by Olympiad experts that appeared in real competitions and are conceptually similar.
3. **Large-Scale Evaluation.** Benchmarking across 27 models on three primary tasks that measure mathematical comprehension, retrieval quality, and analogical reasoning (MathRAG) using both automatic grading and human expert grading.
4. **Analysis: Solving vs. Retrieving.** We demonstrate a sharp divergence between problem solving and retrieval: even state-of-the-art models struggle with mathematical retrieval. Moreover, retrieval-augmented generation (RAG) improves reasoning only when retrievers surface *structure-aligned*, mathematically relevant neighbors.

2 Related Work

Mathematical problem solving has long been a core benchmark for evaluating AI intelligence. Early efforts focused on text-based arithmetic problems, while recent research has expanded to competition-level reasoning, theorem proving, and multimodal problem-solving. Existing datasets can be broadly categorized into text-only benchmarks, multimodal benchmarks, and large-scale aggregates.

Text-Only Mathematical Benchmarks. Several datasets evaluate LLMs' mathematical reasoning using text-only problems. Cobbe et al. (2021) introduced **GSM8K**, grade-school level problems for elementary arithmetic reasoning. Hendrycks et al. proposed **MATH**, which consists of problems spanning high school to competitive mathematics. Gao et al. (2024b) presented **Omni-MATH**, with 4,428 Olympiad-level problems. He et al. (2024) and Wang et al. (2024) further extend coverage with bilingual and competition-level datasets, though most are limited in scale, language diversity, or structured similarity annotations.

Multimodal Mathematical Benchmarks. Multimodal benchmarks integrate visual information with textual descriptions, primarily for geometry or diagram-based reasoning. Datasets such as **MATH-Vision** (Wang et al., 2024) and **MathVista** (Lu et al., 2024) incorporate broad visual contexts, including charts and diagrams. Despite this added modality, these datasets remain comparatively easy and do not capture the full difficulty of Olympiad-level problem solving.

Large-Scale Aggregates. Large datasets aggregate problems from multiple sources such as Numina-Math (Li et al., 2024b) and (Li et al., 2025). Although valuable for large-scale training and evaluation, these datasets typically lack curated multimodal content, multi-lingual coverage, and fine-grained annotations.

Math Retrieval There has been work on formula-aware indexing (Zanibbi et al., 2025), but such systems predate LLMs and typically operate at the formula level, missing broader conceptual and structural similarities expressed in natural language. Meanwhile, modern IR excels at semantic paraphrase but is often *blind* to symbolic equivalence and cross-modal cues.

Limitations and Motivation for *MathNet*. Despite these advances, current benchmarks exhibit thre main limitations: (i) limited detailed solutions written by experts, (ii) restricted visual multilingual content, especially for high-difficulty problems, and (iii) no focus on retrieving mathematically equivalent or related problems. *MathNet* addresses these gaps by offering a large-scale, multilingual, multimodal dataset of 17,512 Olympiad-level problems. It includes expert-validated problem pairs and a fine-grained taxonomy of mathematical similarity, enabling rigorous study of retrieval-augmented reasoning, analogical problem solving, and cross-lingual generalization in LLMs and LMMs.

Benchmark	Size	Languages	Evaluation Type	M	Source	Difficulty
GSM8k Cobbe et al. (2021)	8,500	EN	Numeric Answer	×	Crowdsourced problems	Grade School
MATH Hendrycks et al.	12,500	EN	Numeric Answer	×	Competitions / textbooks	High School
MATH-Vision Wang et al. (2024)	3,040	EN	Expression / Proof	✓	Math Competitions	High School
CMMLU Li et al. (2024a)	11,528	ZH	MCQ	×	Chinese exam materials	High School / College
MMLU Hendrycks et al. (2021)	15,908	EN	MCQ	×	College / professional exams	College-Level
C-Eval Huang et al. (2023)	13,948	ZH	MCQ	×	Chinese college exams	College Entrance
MMMU Yue et al. (2024)	11,500	EN	MCQ / Expression	✓	Multimodal academic exams	College-Level
AGIEval Zhong et al. (2024)	3,300	EN & ZH	MCQ / Expression	×	College entrance exams	College Entrance
JEEBench Arora et al. (2023)	515	EN	MCQ / Numeric Answer	×	Indian JEE Advanced	JEE Advanced Exam
OlympiadBench He et al. (2024)	6,142	EN & ZH	Proof / Expression	✓	Official Websites	Olympiad Level
OlympicArena Huang et al. (2024)	3,233	EN & ZH	Proof / Process	✓	Official Websites	Olympiad Level
Omni-Math Gao et al. (2024b)	4,428	EN	Proof / Process	×	AoPS Forum / Contest Pages	Olympiad Level
IneqMath Sheng et al. (2025)	1,552	EN	Proof / Analytical Tools	×	Curated Inequalities Problems	Olympiad Level
OlymMATH Sun et al. (2025)	200	EN & ZH	Numeric Answer	×	AoPS Forum/Official Websites	Olympiad Level
LiveAoPS Mahdavi et al. (2025)	-	EN	Numeric / Expression	×	AoPS Forum (rolling snapshot)	Olympiad Level
MathArena Balunović et al. (2025)	162	EN	Final Answer / Proof	✓	Newly released competitions	Olympiad Level
IMOBench Luong et al. (2025)	460	EN	Numeric / Proof	×	IMO & national archives	Olympiad Level
MathNet (ours)	**17,152**	EN, ZH, ES RU, AR, RO DE, FA, …	Expression / Proof	✓	Printed Official Country Booklets/ International and National Contests	Olympiad Level

Table 1: Comparison of mathematical reasoning benchmarks across different sizes, languages, evaluation types, and difficulty levels. We include both unimodal and multimodal datasets, spanning grade-school to Olympiad-level mathematics. Our proposed **MathNet** expands coverage to 10 languages and focuses on proof- and process-based evaluation with authentic national contest problems.

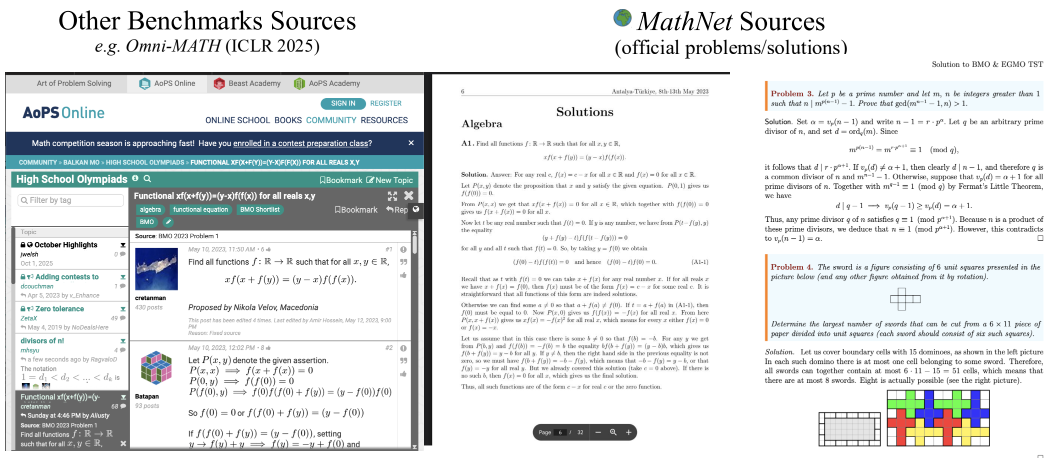


Figure 2: **MathNet is a collection of official Olympiad documents sourced directly from national problem booklets**. This example shows a BMO 2023 problem that appears in both MathNet and Omni-MATH Gao et al. (2024a) While Omni-MATH relies on the AoPS discussion shown on the left, MathNet provides the official problem and solution on the right.

3 Dataset

We introduce *MathNet*, a large-scale benchmark designed to evaluate the cognitive reasoning and retrieval abilities of large language models (LLMs) and large multimodal models (LMMs). The benchmark contains both text-only and interleaved text–image problems, supporting multi-lingual presentation to broaden accessibility and inclusivity. In total, *MathNet* comprises 17,512 problems with expert-written solutions, spanning 40 countries and 300 distinct competitions.

A key feature of *MathNet* is its fine-grained taxonomy of mathematical similarity, which enables systematic analysis of model performance across varying levels of structural and semantic overlap. To complement the dataset, we define a novel retrieval task that measures a model's ability to identify related problems based on deeper structural relationships rather than surface-level features. We further provide baseline models and evaluations, demonstrating the benchmark's utility in assessing both problem-solving accuracy and mathematical understanding.

3.1 Data Collection, Extraction and Annotation

Data sources. Each year, participating countries in the International Mathematical Olympiad (IMO) contribute original problems for use in their national contests and team selection examinations. To construct our benchmark, we curated a collection of official problem booklets from 40 countries spanning 2006–2025, comprising 739 PDF volumes and more than 25,000 pages in total. Unlike prior math benchmarks that often rely on community-sourced platforms such as AoPS, *MathNet* is built exclusively from officially published national materials. All included problems and solutions are authored and disseminated by national teams themselves, ensuring expert-level quality, consistency in style, and immunity from noisy or informal annotations. For more details, refer to section A.1.

Problems Extraction. We first convert all contest booklets into a Markdown format using `dots-ocr` dot (2025), which is a multilingual document parsing framework (see Appendix 9). This step establishes a uniform input format for downstream processing. The underlying source material spans a wide range of formats: recent volumes are digitally typeset, while older archives are only available as scanned copies, and many booklets are bilingual. By leveraging the multilingual recognition and layout analysis capabilities of `dots-ocr`, our pipeline robustly handles this variation, ensuring consistent and faithful text extraction across diverse document types.

Problems Solution Matching and Annotation Extracting aligned problem–solution pairs from parsed contest booklets poses a significant challenge due to the heterogeneity of source documents. Some booklets present problems and solutions in separate sections (see Appendix 9), while others interleave them. Numbering schemes and naming conventions vary not only across countries but

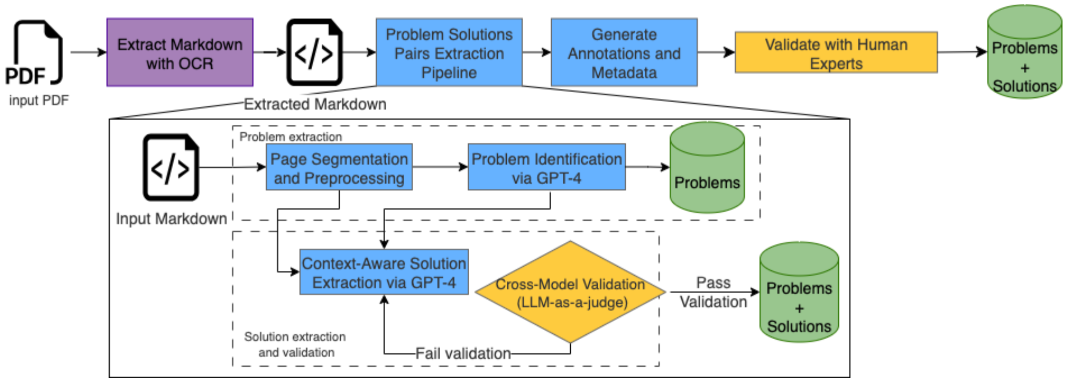


Figure 3: Problem–solution Extraction and Validation Pipeline.

often within a single document. These inconsistencies render traditional parsing techniques (e.g., regex-based heuristics) brittle and non-scalable.

To address this, we designed a tailored LLM-based pipeline for problem–solution alignment (illustrated in Figure 3). Our approach operates in three stages:

Document Ingestion and Problem Extraction. We preprocess each contest booklet by segmenting it into page-level units, which are then provided as input to **GPT-4.1** for problem identification and extraction in LaTeX format. For each extracted statement, we additionally record the source file and page number to maintain provenance metadata.

Solution Retrieval. Since solutions never precede their corresponding problems, the system begins searching only after the identified problem page. We slide an overlapping window of four consecutive pages, pairing the problem text with these candidate pages, and prompt **GPT-4.1** to extract the corresponding solution as shown in Appendix 1. This strategy balances robustness to noisy formatting with efficiency in long documents.

Semantic Verification. We subsequently evaluate each extracted problem–solution pair with two independent LLMs: **GPT-4.1** and **Claude 4 Opus**. We prompt both models to act as judges (see Appendix 2), assessing (i) the correctness of alignment and whether the solution corresponds to the intended problem (ii) the completeness of coverage—whether the entire solution is captured. The system accepts a pair into the dataset only when both LLMs independently agree on its validity, thereby providing cross-model consensus that mitigates single-model bias or hallucination (Gu et al., 2024).

Through this multi-stage design, our pipeline achieves high recall and precision across diverse document structures, enabling the construction of a clean, large-scale dataset of expert-authored problems and solutions.

3.2 Data Quality Verification

Human Validation of Problem and Solution Extraction. To obtain a reliable estimate of extraction quality, we randomly sampled 100 problem–solution pairs from the dataset and conducted a controlled human evaluation. We recruited 20 annotators with academic backgrounds in mathematics, computer science, and engineering, and instructed them to independently assess each problem-solution pair along two dimensions: (i) the correctness of alignment, i.e., whether the solution corresponds to the intended problem, and (ii) the completeness of coverage, i.e., whether the solution is fully captured. To facilitate human validation, we developed and publicly released a lightweight web-based interface that supports multimodal display of problems, solutions, and provenance metadata (e.g., source document and page number).

LLM-Based Stress Testing with Distractors. To assess dataset robustness and potential leakage, we employed a large language model (GPT-4.1) to generate a set of "distractor" problems. For each problem, we prompted the model to produce five plausible but incorrect statements and then instructed it to identify the correct problem from a mixture of its own distractors and the true related problems

Mode	Problem A	Problem B
Invariance		
Syntactic Equivalence	Find $f : \mathbb{R} \to \mathbb{R}$ such that $f(x^2 - y^2) = (x - y)(f(x) + f(y))$.	Find $g : \mathbb{R} \to \mathbb{R}$ such that $(g(a) + g(b))(a - b) = g(a^2 - b^2)$.
Reformulation	Let $a_i > 0$. Prove $\sum_{i=1}^{n} \frac{a_i}{a_i^2 + a_{i+1}a_{i+2}} \leq \sum_{i=1}^{n} \frac{1}{a_i + a_{i+1}}$	Let $a_i > 0$. Prove $\sum_{i=1}^{n} \frac{a_i^2}{a_i^2 + a_{i+1}a_{i+2}} \geq \frac{1}{2}$
Transformational	Find all $x \in \mathbb{R}$ such that $4^x + 6^x = 9^x$.	Find all $x \in \mathbb{R}$ such that $(2/3)^x + (3/2)^x = 5/2$.
Structural Resonance		
Generalization/Specialization	For $k \geq 1$, prove that k divides $\binom{n}{k}$ for all $n \geq k$.	Show that $\binom{n}{m} \equiv \prod \binom{n_i}{m_i} \pmod{p}$, where $n = \sum n_i p^i$, $m = \sum m_i p^i$.
Common Lemma	Prove that $4^n + 2^n + 1$ is never a prime number.	Prove that $2^{2n} + 2^n + 1$ is divisible by 3 for all n.
Structural Reduction	If $ab + 1 \| a^2 + b^2$, show that $\frac{a^2+b^2}{ab+1}$ is a perfect square.	If $a^2 + b^2 + c^2 = k(ab + bc + ca)$, show that $k \in \{1, 2, 3\}$.
Affinity		
Thematic	Show that the largest prime factor of $\binom{2n}{n}$ is greater than $n^{2/3}$.	For every $n > 1$, there is a prime p such that $n < p < 2n$.

Table 2: Taxonomy of mathematical similarity with Olympiad-style examples. Invariance captures strict equivalence under reformulation, Structural Resonance reflects shared lemmas or reductions, and Affinity denotes looser thematic clustering.

from our dataset. The model's low success rate indicates that the annotated problem connections in *MathNet* are non-trivial and cannot be inferred through simple surface-level patterns, thereby reinforcing the quality of our annotations (see Section 4.3).

Expert Review of Similarity Annotations. As an additional validation step, we asked experts to review a subset of 500 sampled problems with their associated distractors. At least two annotators independently assessed each problem–distractor set, and a senior expert resolved any disagreements through consensus. This procedure confirmed that the similarity annotations capture genuine mathematical structure rather than superficial lexical overlap, providing a complementary layer of assurance beyond the LLM-based evaluation.

3.3 What makes problems similar?

Mathematical progress often depends on recognizing when different problems share common structure. Similarity is not a single notion but can take several forms, from strict equivalence to looser thematic connections. We distinguish three modes of similarity: *Invariance*, *Resonance*, and *Affinity* (see Table 2).

Invariance refers to strict equivalence under transformation. Two problems are invariant when they differ only in representation but share the same underlying structure. Examples include syntactic renaming, algebraic reformulation, geometric re-characterization, or cross-domain isomorphism.

Resonance refers to partial similarity. Problems are not identical, but they can be addressed using the same idea, proof strategy, or structural analogy. Resonance highlights opportunities to transfer tools or insights across contexts.

Affinity refers to a broad sense of relatedness without structural equivalence. Problems may belong to the same conceptual or disciplinary area (e.g., number theory, geometry) even if they do not share a method or solution strategy. Affinity provides a way to group problems by theme, context, or historical development.

3.4 How are similar problem pairs constructed?

To asses the limitations of current embedding models, we designed a *Problem Retrieval* task that aims to distinguish between surface-level lexical overlap and deep mathematical equivalence. We construct three types of paired data points:

a) Synthetic Equivalent Pairs. We generated equivalent versions of anchor problems via variable renaming (e.g., $x \to a$), algebraic manipulation, and paraphrasing using GPT-4.1 (prompt details more details Appendix). For example, the functional equation $f(x) + f(y) = f(x + y)$ is paired with an algebraically equivalent variant such as $g(a) - g(a + b) = -g(b)$.

b) Hard Negatives (Near-Misses). To asses how much models can rely solely on token overlap, we generated adversarial "hard negatives" that mimic the syntax of the anchor but differ mathematically (e.g., $f(x^2) + f(y) = f(x - y)$). These serve as near-miss distractors that require genuine mathematical understanding to avoid.

To ensure correctness of the synthetic problems, we performed a human verification pass on all generated samples, where 500 pairs of synthetic samples were verified.

c) Expert-Curated Conceptual Pairs. We curated 70 pairs from real Olympiad problems over the past 20 years. These belong to the *Structural Resonance* category of our taxonomy of pairs exhibiting conceptual similarity such as generalization/specialization relations, shared key lemmas, or one problem being a reduction of the other. These expert-curated pairs capture similarity that goes beyond algebraic transformations.

3.5 Data Preparation and Release

Our benchmark contains 17,512 problems, with 5,500 designated for model-based evaluation as *MathNet-test-large*, and 140 curated hard problems as *MathNet-test-small*. *MathNet-test-small* is organized into pairs consisting of a problem and a conceptually related problem (Invariance and Structural Resonance), both of which appeared in real competitions. The dataset will be publicly released.

3.6 Data Analysis

Figure 6b illustrates the diversity of our dataset across mathematical domains. Notably, Number Theory and Combinatorics account for a large share of the most difficult problems, reflecting their inherent complexity. In addition, the dataset is multilingual, with problems provided in ten different languages (see Appendix Table 11), which makes it particularly well-suited for evaluating cross-lingual reasoning.

4 Experiments

4.1 Experimental Setup

We evaluate 25 models on *MathNet* under two benchmarks: (a) **Math Comprehension** and (b) **Math Retrieval**.

For **Math Comprehension**, we evaluate two types of models: (i) **LLMs and LMMs**, including `gpt-4o`, `Llama-4-Maverick-17B-128E-Instruct-FP8`, `Grok-4.1`, `Grok-3`. For models that accept images, we provide both the text and image as input; otherwise, we supply a text-only description of the image. (ii) **LLMs and LMMs with CoT Reasoning**, including `gpt-5`, `gpt-5-mini`, `gpt-5-nano`, `gemini-3-pro`, `gemini-2.5-pro`, `gemini-2.5-flash`, `DeepSeek-V3.2-Speciale`, `DeepSeek-V3`, `DeepSeek-R1`, `claude-opus-4.5`.

For **Math Retrieval**, we evaluate retrieval performance using embeddings derived from a diverse set of state-of-the-art models, including `all-mpnet-base-v2`, `multi-qa-mpnet-base-dot-v1`, `cohere-embed-v4.0`, `qwen3-embedding-4B`, `gemini-embedding-001`, `text-embedding-ada-002`, `text-embedding-3-small`, and `text-embedding-3-large`. We compute similarities between problem statements using cosine similarity over the embedding representations.

4.2 Evaluation Protocol

Math Comprehension. For evaluation, following the protocol proposed by IMO-Bench Luong et al. (2025), we adopt a model-based evaluation using `Gemini-2.5-Pro`, which was shown in IMO-Bench to achieve a Pearson correlation of 0.87 with human graders. For each problem, `Gemini-2.5-Pro` is provided with the problem statement, the reference solution, and the model-generated solution, and is asked to judge whether the output is consistent with the correct answer using a numeric system 0-7 binarized to score=7 (full correctness) and score != 7. This allows us to distinguish between models that arrive at the correct final answer by coincidence versus those that demonstrate consistent reasoning ability. We also report performance by subject domain (algebra, geometry, combinatorics, number theory), enabling a fine-grained analysis of model strengths and weaknesses.

Math Retrieval. The primary evaluation metric for our retrieval task is **Recall@k**, which measures whether any of the top-k retrieved problems correspond to a "correct" match from our equivalent versions of each problem. We report Recall@1, Recall@5, and Recall@10. To better understand embedding behavior, we further analyze cosine similarity distributions between equivalent problem pairs, unrelated pairs, and near misses (hard negatives), highlighting cases where models struggle to separate fine-grained distinctions.

Retrieval Augmented Generation. To assess the impact of retrieval RAG, we evaluate how retrieval quality affects downstream mathematical problem solving. We adopt a controlled setup: we report solver performance under three conditions: (a) zero-shot, (b) RAG with a standard off-the-shelf retriever, and (c) RAG with an oracle retriever that supplies a ground-truth similar problem from MathNet. This allows us to directly quantify how much math-aware retrieval contributes to solution accuracy.

We report the relative gains between (b) and (c), highlighting how structural alignment in retrieval enables models to make effective use of retrieved context, and we identify cases where irrelevant retrieval harms performance. This protocol provides a clear empirical demonstration of the value of math-aware retrieval.

4.3 Main Results

Math Comprehension Table 4 summarizes accuracy across four mathematical domains. Baseline LLMs such as *Llama-4-Maverick-17B* and *DeepSeek-V3* achieve modest macro-averages in the mid-40s, indicating that direct pattern matching and shallow heuristics are insufficient for Olympiad-level problem solving.

Math Comprehension Results on Test-Set-Small (70 samples)

Model	RD	Human Grading			LLM Grading		
	(2025)	zero shot	embed-RAG	expert-RAG	zero shot	embed-RAG	expert-RAG
DeepSeek-V3.2-Speciale	01 Dec	84.8%	89.5%	97.3%	82.23%	87.87%	89.03%
Claude-4.5-Opus	24 Nov	46.8%	55.5%	52.4%	45.97%	50.34%	56.43%
oLMO-3-Think	20 Nov	45.2%	54.6%	47.6%	49.49%	45.56%	51.07%
Grok-4.1-Fast	19 Nov	75.4%	83.8%	83.2%	73.06%	67.66%	69.11%
Gemini-3-Pro	18 Nov	89.1%	92.9%	87.5%	73.16%	70.54%	76.43%
GPT-5	07 Aug	76.8%	75.2%	86.6%	87.09%	81.81%	85.76%
Phi-4-Reasoning Plus	30 Apr	15.1%	14.3%	16.7%	24.06%	19.64%	30.04%

Table 3: Performance of evaluated language models on the Math Comprehension Test-Set-Small (70 samples). The table reports human and average LLM grading accuracy under three prompting and retrieval configurations: zero-shot, embed-RAG, and expert-RAG. RD=release date.

Reasoning-augmented models (e.g., *GPT 5* and *Gemini 2.5 Flash*) substantially improve performance, with macro-averages around above 60%. However, their accuracy remains uneven across domains: while Algebra shows steady gains, **Geometry and Discrete Math remain the hardest categories**, reflecting difficulty with abstract reasoning, non-obvious solution paths, and combinatorial structures. For more breakdown analysis of Language and Multimodality sensitivity refer to Table 9 and Table 10 in Appendix A.4.

Math Retrieval As shown in Table 5, retrieval on *MathNet* remains highly challenging at the top-1 level, with even the strongest models (`Qwen3-embedding-4B` and `Gemini-embedding-001`) achieving only ∼5% Recall@1. Performance improves markedly at higher cutoffs, with Recall@10 exceeding 80% in several domains. Among all models, Gemini-embedding-001 provides the most consistent gains, delivering the highest Recall@5 and Recall@10 across domains and the strongest aggregate performance (68.88% and 83.79%, respectively). In contrast, legacy embedding models such as `text-embedding-ada-002` and `text-embedding-3-small` perform substantially worse across all settings.

These results suggest that current general-purpose embedding models fail to capture the deep structural and symbolic relationships that define mathematical equivalence. A critical failure mode is that both LLMs and LMMs often rely on superficial textual overlap (e.g., matching on keywords such as "triangle" or "polynomial") rather than reasoning over the underlying mathematical concepts. The weak top-1 retrieval performance highlights that these models lack a robust internal representation of mathematical knowledge that would support analogical reasoning across problem variants. This gap

Zero-Shot Math Comprehension Results on Test-Set-Large (5500 samples)						
	Algebra	Number Theory	Geometry	Discrete Math	**Macro Avg**	**Micro Avg**
LLMs (Text-only)						
Mistral-3B	8.87% ± 0.54	5.99% ± 0.53	1.28% ± 0.17	4.51% ± 0.46	5.16% ± 1.37	4.60% ± 0.20
DeepSeek-V3.2	11.03% ± 0.60	12.11% ± 0.73	1.74% ± 0.20	5.83% ± 0.52	7.68% ± 2.08	6.66% ± 0.24
Grok-3	22.13% ± 0.79	18.23% ± 0.87	3.14% ± 0.27	12.97% ± 0.74	14.12% ± 3.56	12.41% ± 0.31
LVLMs (Vision-enabled)						
Llama-4-Maverick-17B	25.72% ± 0.83	22.08% ± 0.93	5.84% ± 0.36	10.67% ± 0.68	16.08% ± 4.05	14.60% ± 0.34
GPT-4.1	41.71% ± 0.94	41.34% ± 1.11	12.75% ± 0.51	33.91% ± 1.05	32.43% ± 5.89	29.00% ± 0.43
GPT-4o	19.22% ± 0.75	15.77% ± 0.82	3.13% ± 0.27	12.05% ± 0.72	12.54% ± 3.00	11.04% ± 0.30
LLMs + Reasoning (Text-only)						
DeepSeek-R1	32.43% ± 0.90	27.21% ± 1.00	5.18% ± 0.34	16.69% ± 0.83	20.38% ± 5.22	18.07% ± 0.37
LVLMs + Reasoning (Vision + deliberate reasoning)						
Gemini-2.5-Flash	51.42% ± 0.95	52.18% ± 1.02	56.73% ± 0.88	42.57% ± 0.91	50.03% ± 3.20	49.61% ± 0.45
Claude-4-Opus	40.27% ± 0.90	20.44% ± 0.75	79.12% ± 1.10	26.39% ± 0.82	41.08% ± 4.10	40.22% ± 0.50
GPT-5	92.41% ± 0.50	89.06% ± 0.70	64.23% ± 0.73	85.28% ± 0.78	**82.74% ± 5.49**	**79.63% ± 0.38**
GPT-5-mini	87.14% ± 0.64	88.81% ± 0.71	65.43% ± 0.73	79.13% ± 0.90	80.13% ± 4.62	77.58% ± 0.40
GPT-5-nano	77.59% ± 0.80	78.05% ± 0.93	51.92% ± 0.76	66.81% ± 1.04	68.59% ± 5.31	65.75% ± 0.45

Table 4: Experimental results on *MathNet-Test-Large*, which consists of 5500 problems. Results are expressed as percentages, with the highest score in each setting underlined and the highest scores across all settings bolded.

Table 5: Experimental results on *MathNet*, expressed as percentages for Recall@1 and Recall@5. The highest score in each setting is underlined, and the highest overall scores are bolded.

	Algebra		Number Theory		Geometry		Discrete Mathematics		All	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
all-mpnet-base-v2	4.54%	73.06%	4.67%	82.54%	4.37%	74.76%	4.25%	75.38%	3.78%	57.7%
multi-qa-mpnet-base-dot-v1	4.0%	69.4%	3.73%	80.76%	3.88%	71.73%	3.98%	73.4%	3.27%	55.08%
cohere-embed-v4.0	2.73%	59.85%	2.67%	68.85%	2.35%	59.87%	2.78%	63.4%	2.24%	44.81%
qwen3-embedding-4B	5.24%	78.74%	4.62%	86.43%	5.6%	79.05%	5.96%	81.5%	**4.96%**	64.95%
gemini-embedding-001	5.5%	81.62%	4.95%	87.43%	5.49%	81.86%	5.35%	82.8%	4.83%	**68.88%**
text-embedding-ada-002	2.05%	54.94%	2.22%	63.35%	2.16%	55.07%	2.71%	57.51%	1.94%	42.02%
text-embedding-3-small	2.1%	47.47%	1.89%	54.62%	2.1%	47.61%	2.84%	50.12%	1.98%	35.49%
text-embedding-3-large	3.19%	68.18%	2.73%	75.25%	3.2%	68.18%	3.35%	69.52%	2.74%	54.23%

underscores the need for embeddings explicitly trained to encode mathematical structure, rather than depending on incidental surface-level cues.

To further illustrate this issue, Figure 4 shows the distribution of cosine similarities between equivalent and non-equivalent problems. Surprisingly, non-equivalent pairs often exhibit higher similarity scores than equivalent ones. This counterintuitive trend highlights that embeddings frequently capture superficial lexical or symbolic overlap rather than true structural relationships, leading models to mis-rank distinct problems as closer than genuinely equivalent ones. This explains the weak Recall@1 performance observed in Table 5.

MathRAG. As shown in Table 3, providing these ground-truth pairs as retrieval context (*expert-RAG*) yields consistent gains over zero-shot sttings under both human and LLM grading. Improvements are largest for lower- and mid-tier solvers, indicating that math-aware retrieval supplies structure-aligned hints that current models do not reliably surface on their own. For the strongest systems, we observe occasional small dips (e.g., *Gemini-3-Pro* with human grading; *GPT-5* with LLM grading), which we attribute to over-conditioning on partially relevant context. In the embed-RAG setting, we see high variance across results: when it retrieves structure-aligned neighbors it helps, but near-miss distractors often degrade performance.

Together, these results show that retrieval can meaningfully boost Olympiad problem solving but only when the retrieved context is truly *structurally* similar. Progress in retrieval-enhanced systems will hinge on retrievers attuned to mathematical structure rather than surface lexical overlap. *MathNet* with expert-aligned pairs and hard negatives offers a controlled setting to develop and rigorously evaluate such math-aware retrieval for RAG.

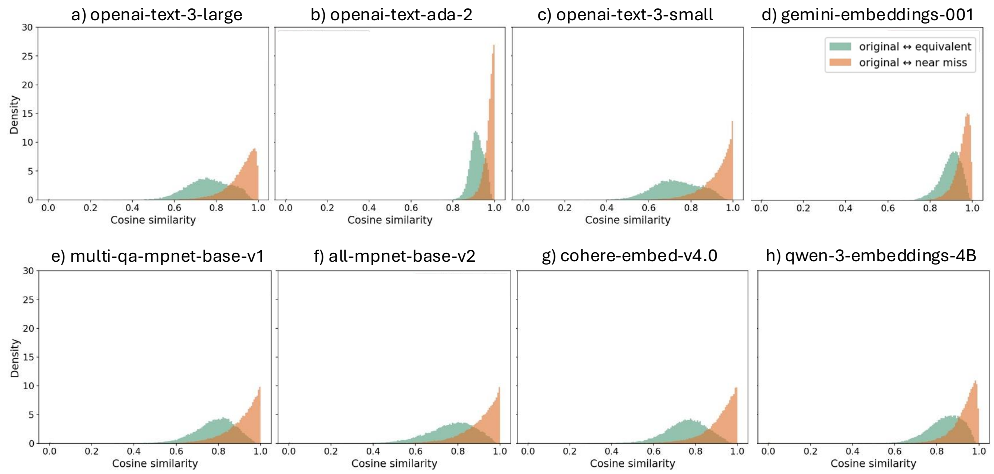


Figure 4: Cosine similarity distributions for equivalent (green) and near-miss/hard negatives (orange) problem pairs across different embedding models. Higher separation between the two distributions indicates a model's ability to distinguish structurally identical problems from those with small but critical alterations.

5 Discussion

Results on *MathNet* reveal a clear gap between the problem-solving ability of modern LLMs/LMMs and their understanding of mathematical structure. While models achieve impressive scores on answer-generation benchmarks, our retrieval task shows they lack a generalizable grasp of equivalence and analogy. The limited gains from visual augmentation further suggest that multimodal integration for symbolic tasks remains underdeveloped.

The strong performance of the formula-aware baseline indicates that structured, non-textual representations are crucial for retrieval. Progress in true mathematical reasoning may require moving beyond next-token prediction toward architectures that explicitly integrate symbolic reasoning.

6 Conclusion

In this work, we introduced *MathNet*, the first large-scale, multilingual, multimodal benchmark for mathematical reasoning and retrieval. By providing a rich dataset of 17,512 problems with a fine-grained taxonomy of equivalence, we enabled a rigorous study of mathematical generalization and analogical reasoning. To ensure reliability, we complemented automated extraction with systematic human validation: expert annotators reviewed problem similarity labels, and student evaluators assessed the alignment and completeness of extracted problem–solution pairs. These human contributions establish a strong ground-truth foundation, ensuring that *MathNet* captures deep mathematical structure rather than superficial overlap.

Our comprehensive evaluations show that while frontier models can solve complex problems, they struggle with a fundamental yet overlooked task: retrieving mathematically equivalent or related problems from large corpora. This deficiency in retrieval highlights a key limitation in their ability to form a robust, internally consistent representation of mathematical knowledge. We hope *MathNet* will serve as a valuable resource for the community, paving the way for research into improved retrieval-augmented reasoning, symbolic AI, and ultimately, more capable and reliable problem-solving models.

BIBLIOGRAPHY

dots.ocr: Multilingual document layout parsing in a single vision-language model. https://github.com/rednote-hilab/dots.ocr, 2025.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Daman Arora, Gaurav Goyal, Harshit Arora, et al. Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions. *arXiv preprint arXiv:2505.23281*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Debrup Das, Sam O' Nuallain, and Razieh Rahimi. Rader: Reasoning-aware dense retrieval models. *arXiv preprint arXiv:2505.18405*, 2025.

Moran Feldman and Amin Karbasi. Godel test: Can large language models solve easy conjectures? *arXiv preprint arXiv:2509.18383*, 2025.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*, 2021.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*, 2024a.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*, 2024b.

Mathematical Gemstones. Olympiad vs. higher math. https://www.mathematicalgemstones.com/misc/olympiad-vs-higher-math/, 2020. Accessed: 2025-09-25.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

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. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Yuzhuo Huang, Xunzhi Bai, Yifan Li, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2023.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, et al. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent AI. *arXiv preprint arXiv:2406.12753*, 2024.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *SIGIR*, 2020.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. CMMLU: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024a.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Kashif Rasul, Longhui Yu, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *arXiv preprint arXiv:2409.02834*, 2024b.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Kashif Rasul, Longhui Yu, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Big-math: A large-scale, high-quality math dataset for reinforcement learning. *arXiv preprint arXiv:2502.17387*, 2025.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2024.

Minh-Thang Luong, Dawsen Hwang, Hoang H Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, et al. Towards robust mathematical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35406–35430, 2025.

Sadegh Mahdavi, Muchen Li, Kaiwen Liu, Christos Thrampoulidis, Leonid Sigal, and Renjie Liao. Leveraging online olympiad-level math problems for llms training and contamination-resistant evaluation. *arXiv preprint arXiv:2501.14275*, 2025.

Fan Nie, Ken Ziyu Liu, Zihao Wang, Rui Sun, Wei Liu, Weijia Shi, Huaxiu Yao, Linjun Zhang, Andrew Y Ng, James Zou, et al. Uq: Assessing language models on unsolved questions. *arXiv preprint arXiv:2508.17580*, 2025.

Jiayi Sheng, Luna Lyu, Jikai Jin, Tony Xia, Alex Gu, James Zou, and Pan Lu. Solving inequality proofs with large language models. *arXiv preprint arXiv:2506.07927*, 2025.

Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models. *arXiv preprint arXiv:2503.21380*, 2025.

Praneeth Vemuganti, Samir Chhatkuli, and Yashwanth Reddy. Advancing math formula search using diverse structural features. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, pp. 101–115. Springer, 2025. doi: 10.1007/978-3-031-88708-6_8.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Richard Zanibbi, Behrooz Mansouri, Anurag Agarwal, et al. Mathematical information retrieval: Search and question answering. *Foundations and Trends® in Information Retrieval*, 19(1-2):1–190, 2025.

Chengda Lu Z.Z. Ren Jiewen Hu Tian Ye Zhibin Gou Shirong Ma Xiaokang Zhang Zhihong Shao, Yuxiang Luo. Deepseekmath-v2: Towards self-verifiable mathematical reasoning, 2025.

Wanjun Zhong, Ruixiang Cui, Sai Liang, et al. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024.

A Appendix

The appendix provides additional tables, figures, prompts, and implementation details to support reproducibility and further analysis.

A.1 Overview of competitions covered by MathNet

This section lists the national and regional competitions represented in *MathNet*, along with years covered and document sources, to clarify the dataset's institutional breadth.

Country	Years	Competitions
Argentina	2003–2023	Cono Sur MO; Argentine National Olympiad; Rioplatense Olympiad; Iberoamerican MO; Olimpiada de Mayo;
Australia	2010–2024	AMOC Senior Contest; APMO; AIMO; Australian MO; EGMO (TST); IMO (TST); MCYA
Austria	2010–2024	Austrian MO – Regional; Austrian MO – Junior Regional; Austrian MO – National; National Olympiad – Preliminary; National Olympiad – Final; Beginners' Competition; EGMO (TST); IMO (TST)
Balkans	2010–2025	Balkans Mathematical Olympiad (BMO)
Baltics	2009–2023	Baltic Way; Baltic Way Shortlist
Belarus	2010–2024	Belarusian MO; IMO (TST)
Brazil	2006–2012	OBM
Bulgaria	2007–2024	Bulgarian MO – Regional; Bulgarian MO – Final; Bulgarian Autumn Competition; Bulgarian Spring Competition; Bulgarian Winter Competition (Rousse, Varna, National); IMO (TST); BMO (TST); Other Bulgarian Competitions; JBMO (TST)
Canada	2010–2017	CMO
China	2007–2025	AMC 10/12; AIME; CMO (China); Chinese MO; China Southeastern MO; CWMO; CGMO; Hua Luogeng Cup; IMO (TST); Soviet Mathematical Competition; Russian Mathematical Competition; Putnam (China ed.)
Croatia	2010–2019	Croatian MO; National Olympiad – City; National Olympiad – County; National Olympiad – Final; MEMO; IMO/MEMO (TST)
Czech Republic	2000–2025	Czech MO – School; Czech MO – District; Czech MO – Regional; Czech MO – Final; Czech–Polish–Slovak Match; Czech–Slovak–Polish Match; Czech–Austrian–Polish–Slovak Match; CAPS Match; Olympiad Corner; IMO/EGMO/MEMO (TST)
Slovakia	2000–2025	Slovak MO – School; Slovak MO – District; Slovak MO – Regional; Slovak MO – Final; Czech–Slovak Match; Czech–Polish–Slovak Match; Czech–Slovak–Polish Match; Czech–Austrian–Polish–Slovak Match; CAPS Match; Olympiad Corner; IMO/EGMO/MEMO (TST)
Poland	2004–2025	Polish MO; Czech–Polish–Slovak Match; Czech–Slovak–Polish Match; Czech–Austrian–Polish–Slovak Match; CAPS Match; Olympiad Corner; IMO/EGMO/MEMO (TST)
Estonia	2010–2025	Estonian MO; Kangaroo; IMO (TST); Other Estonian Open Contests; EGMO (TST)
Greece	2007–2024	Hellenic MO – Archimedes; National Competition – Thales; National Competition – Euclides; BMO; JBMO; Mediterranean Competition; EGMO (TST); IMO (TST); JBMO (TST)
Hong Kong	2014–2017	Hong Kong MO; Hong Kong Team Selection Test; Preliminary Selection – IMO; IMO (TST); APMO; CHKMO
India	2006–2023	INMO; RMO; TSTs (IMO/EGMO/RMM); EGMO (TST); RMM (TST); IMO (TST); USA TST Exchange; ISL/ELMO (training/mock)
Iran	2010–2024	Iranian MO; IMO (TST)
Ireland	2007–2025	Irish MO; IMO (TST)
Japan	2006–2025	JMO; JJMO; IMO/EGMO (TST)
Mongolia	2009–2025	Mongolian MO; Mongolian National MO; IMO (TST); EGMO (TST)
Netherlands	2019–2025	Dutch MO; Junior MO; Kangaroo; Pythagoras Olympiad; BxMO; BxMO/EGMO (TST); IMO (TST)

Continued on next page

Country	Years	Competitions
North Macedonia	2008–2023	Macedonian MO; Macedonian Junior MO; National Olympiad – Regional; National Olympiad – Final; BMO; JBMO; Mediterranean Competition; EGMO (TST); IMO (TST); BMO (TST)
Romania	2010–2025	Romanian MO – District; Romanian MO – Final; RMM; BMO; JBMO; EGMO; IMAR Competition; Stars of Mathematics; Danube Competition; Clock-Tower School Competitions; IMO/BMO/JBMO/EGMO/RMM (TST)
Russia	2009–2025	Russian MO – Regional; Russian MO – Final; Euler Olympiad; All-Russian Olympiad (district, regional, national); IMO/EGMO (TST)
Saudi Arabia	2010–2025	Saudi MO; APMO (TST); EGMO (TST); IMO (TST); BMO (TST); JBMO (TST)
Singapore	2010–2025	SMO (Junior, Senior, Open); SIMOC Camp Quizzes; National Olympiad – Round 2 (all); IMO/EGMO (TST)
Slovenia	2008–2016	Slovenian National MO; International Kangaroo; IMO (TST)
South Africa	2010–2024	SAMO; National Olympiad – Senior; University Training Camps; Talent Search; Monthly Problem Sets; IMO (TST)
South Korea	2004–2024	KMO; National Olympiad; IMO (TST)
Spain	2012–2023	Spanish MO; National Olympiad – First Phase; National Olympiad – Final Phase; Iberoamerican MO; Mediterranean MO; Barcelona Contest; BarcelonaTech Math Contest; Arhimede Contest; IMO (TST)
Taiwan	2012–2024	Taiwan MO; National Olympiad Training Camps (Independent Study, Mock Exams, International Practice); IMO (TST)
Thailand	2007–2017	Thailand MO; TMO; IMO (TST)
Turkey	2008–2024	Turkish MO; Junior Turkish MO; National Olympiad; IMO (TST); JBMO (TST); EGMO (TST); Silk Road Mathematical Competition
UK	2006–2022	BMO (Rounds 1 & 2); BMO; EGMO (TST); IMO (TST); RMM (TST); CGMO (TST); Mathematics Ashes
USA	2001–2025	AMC 10/12; AIME; USAMO; USAJMO; IMO (TST); EGMO (TST); RMM (TST)
Ukraine	2005–2023	Ukrainian National MO; Regional Olympiads; Kyiv City Olympiad; Ukrainian Tournament of Mathematical Battles; Ukrainian Mathematical Competitions; Online Olympiads (Algebra, Combinatorics, Number Theory); Ukrainian Summer School Competitions; EGMO (TST); IMO (TST); RMM (TST); EMC
Vietnam	2001–2024	VMO; Vietnamese National Olympiad; IMO (TST)

A.2 Taxonomy of Topics Commonly used in Math Olympiad

We provide the curated taxonomy used for labeling domains, subjects, topics, and subtopics. These labels ground our analyses and enable consistent cross-competition comparisons.

Sub-subtopic	Key Concepts
Geometry	
Plane Geometry	
Triangles	Centroid, incenter, circumcenter, orthocenter, ex-centers, Euler line, nine-point circle; geometric inequalities; trigonometry (metric relations)
Quadrilaterals	Cyclic, inscribed/circumscribed, Complete quadrangle, perpendicular diagonals
Circles	Angles, coaxal, tangents, radical axis, metric relations, Apollonius circle
Concurrency / Collinearity	Theorems of Ceva, Menelaus, Pappus, Desargues
Transformations	Translation, rotation, homothety, spiral similarity, inversion, the method of moving points
Advanced Configurations	Simson line, Miquel, Napoleon / Fermat / Brocard points, symmedians, polar triangles, harmonic/isogonal/isotomic conjugates, barycentric coordinates
Geometric Inequalities	Classical and advanced

Continued on next page

Sub-subtopic	Key Concepts
Combinatorial Geometry	Helly, Sylvester, convex hulls, Pick theorem, Minkowski theorem, convex figures
Analytic / Coordinate Methods	Complex numbers, Cartesian coordinates, vectors, trigonometric relations
Miscellaneous	Angle/distance chasing, constructions, loci
Solid Geometry	
3D Shapes	Polyhedra, prisms, pyramids, spheres, cylinders, cones
Volume	Cavalieri's principle, Formulae and problem-solving
Surface Area	Formulae and applications
Other 3D problems	Mixed problems, reducing the problem into a plane geometry problem
Differential Geometry	
Curvature	Gaussian, mean
Manifolds	Surfaces, parametric
Geodesics	Shortest paths, great circles
Non-Euclidean Geometry	
Spherical Geometry	Spherical triangles, angles, area
Hyperbolic Geometry	Lines, models, inequalities

Algebra	
Prealgebra / Basic Algebra	
Integers	Sets of integers, Divisibility, primes, the Greatest Common Divisor (GCD), the Least Common Multiplier (LCM)
Fractions	Operations, simplification, comparison
Decimals	Conversion, operations, rounding
Simple Equations	Linear equations, word problems
Other	Number properties, prime factorization, divisors
Algebraic Expressions	
Polynomials	Operations, factorization, Algebraic identities, symmetric functions, Vieta's formula, interpolation formulae, complex numbers, roots of unity, Chebyshev polynomials and other trigonometric polynomials, irreducibility of polynomials, Descartes rule of signs, rootso of polynomials, Intermediate Value Theorem (IVT)
Sequences / Series	Recurrences, Charachteristic equations, monotonocity, boundedness, periodicity, convergence and divergence, floors/ceilings, sums/products, telescoping sums, Abel summation
Functional Equations	Substitution, defining a new function, Cauchy's equations, Injectivity/surjectivity, Periodicity, application of Calculus and Mathematical Analysis, iterations
Inequalities	
Functional considerations	Linear/Quadratic solving techniques
Classical inequalities	Cauchy-Schwarz, QM-AM-GM-HM, Power Mean, Jensen's Inequality, smoothing, Muirhead, Chebyshev's inequality, majorization, combinatorial optimization
Discrete Mathematics	
Graph Theory	
Basic concepts	Vertices, edges, path, connected graphs, cycles, Hamiltonian cycle and path, trees
Matchings	Marriage Lemma, Tutte's theorem
Connectivity	Menger, max-flow min-cut
Extremal	Turán
Euler characteristic	$V - E + F$
Combinatorics	
Enumeration	Symmetry, basic counting techniques, recursion, bijection, inclusion-exclusion, double counting
Probability	Expected values, probabilistice methods, partitions, generating functions
Binomial coefficients	Algebraic properties

Continued on next page

Sub-subtopic	Key Concepts
Pigeonhole principle	Applications
Invariants / Monovariants	Problem-solving
Coloring / Extremal	Graph problems
Induction	Standard and smoothing
Games / Greedy	Strategies, combinatorial games
Logic / Algorithms / Other	
Logic	Propositional/predicate logic, truth tables
Algorithms	Sorting, searching, Dynamic Programming (DP), greedy
Other	Miscellaneous problems, strategy development problems, inter-deciplinary problems

Number Theory	
Divisibility / Factorization	
Primes	Properties, sieves, prime numbers tests
GCD	Euclidean algorithm; linear combinations; Bezout's identity
LCM	Computation; relation with GCD
Factorization	Trial, Fermat, Pollard
Modular Arithmetic	
Basic operations $\pmod{n}$, inverses $\pmod{n}$	Existence (when $\gcd(a, n) = 1$); computation (extended Euclidean algorithm)
Chinese Remainder Theorem (CRT)	Solving systems of congruences; applications in number theory and cryptography
Fermat / Euler / Wilson	Theorems; proofs; problem-solving applications
Polynomials mod p	Roots, factorization; applications to number theory problems
Residues / Primitive Roots	
Primitive roots	Existence modulo primes; modulo p^n; computation
Quadratic residues	Properties; Legendre symbol; Euler's criterion
Quadratic reciprocity	Law of quadratic reciprocity; applications
Multiplicative order $\pmod{n}$	Definition; computation; relation with primitive roots and cyclic groups
Diophantine Equations	
Factorization Methods	Difference of squares, Sophie Germain identity, special factorizations; Unique Factorization Domains (Gaussian, Eisenstein integers); Norms in algebraic number fields; Vieta jumping
Modular Arithmetic & Congruences	Reductions modulo primes or powers; Quadratic residues, Legendre symbol; Multiplicative order & primitive roots; Hensel lifting; Local–global principles (solvability mod p)
Parametrization of Solutions	Pythagorean triples; Rational parametrization of conics (general quadratics); Higher-degree parametrizations (elliptic curves, quartics)
Inequalities & Size Arguments	Bounding arguments; Infinite descent; Minimal solutions (no smaller solution possible)
Special Equations	Pell's equation: continued fractions, fundamental solution, recurrence; Fermat-type: $x^4 + y^4 = z^2$,
Descent & Structural Methods	Infinite descent; Descent on elliptic curves; Geometry of numbers
Arithmetic Functions	
Euler's totient's function	Properties, applications
Number / Sum of divisors	Computation, properties
Sum of digits	Basic properties
Möbius inversion	Definition, applications
Algebraic Number Theory	
Algebraic numbers	Minimal polynomials, field extensions, solving Diophantine equations

A.3 BENCHMARKING LLM GRADERS VS HUMAN EXPERT GRADERS

We benchmark the accuracy of a wide range of LLM graders and compare their judgments to human expert grading on `Testset-Small`. This evaluation quantifies how reliably current models can act as automatic graders for Olympiad-level mathematical reasoning. For each model, we report performance under three settings: zero-shot, embed-RAG, and expert-RAG. This measure both cross-model grading consistency and alignment with human scoring.

Cross-Model Grading + Human Scores						
Model	**LLaMA-4**	**DeepSeek-V3**	**GPT-4.1**	**GPT-4o**	**Average**	**Human Expert**
Zero Shot						
`claude-opus-4.5`	72.243	41.557	31.371	38.700	45.971	46.8%
`deepseek-v3.2-speciale`	96.186	74.286	85.457	73.014	82.229	84.8%
`gemini-3-pro-preview`	94.700	72.357	71.686	53.500	73.057	89.1%
`gpt-5`	98.057	85.014	83.157	82.143	87.086	76.8%
`grok-4.1-fast`	92.657	63.357	76.529	59.700	73.060	75.4%
`olmo-3-32b-think`	70.243	44.286	35.200	48.214	49.486	45.2%
`phi-4-reasoning-plus`	46.186	23.529	6.629	19.900	24.057	15.1%
embed-RAG						
`claude-opus-4.5`	64.93%	59.69%	40.30%	36.49%	50.34%	55.5%
`deepseek-v3.2-speciale`	94.21%	78.37%	92.21%	86.73%	87.89%	89.5%
`gemini-3-pro-preview`	95.91%	71.43%	68.63%	46.17%	70.54%	92.9%
`gpt-5`	93.23%	73.94%	80.96%	79.11%	81.81%	75.2%
`grok-4.1-fast`	88.57%	61.13%	72.23%	48.67%	67.66%	83.8%
`olmo-3-32b-think`	66.89%	38.57%	31.49%	45.29%	45.56%	54.6%
`phi-4-reasoning-plus`	36.64%	12.70%	08.16%	21.04%	19.64%	14.3%
expert-RAG						
`claude-opus-4.5`	77.06%	55.51%	53.57%	39.54%	56.43%	52.4%
`deepseek-v3.2-speciale`	97.14%	83.41%	89.29%	86.24%	89.03%	97.3%
`gemini-3-pro-preview`	99.63%	70.03%	72.70%	63.37%	76.43%	87.5%
`gpt-5`	97.29%	77.93%	82.60%	85.20%	85.76%	86.6%
`grok-4.1-fast`	92.50%	55.66%	74.23%	54.09%	69.11%	83.20%
`olmo-3-32b-think`	74.67%	48.99%	33.16%%	47.44%	51.07%	47.6%
`phi-4-reasoning-plus`	48.57%	31.26%	09.70%	30.64%	30.04%	16.7%

Table 8: Breakdown of cross-model grading performance under Zero-Shot, embed-RAG, and expert-RAG configurations, augmented with human evaluation scores for each model.

A.4 PERFORMANCE SENSITIVITY TO LANGUAGE, IMAGE PRESENCE AND TOPICS

We analyze how model accuracy varies with two factors: (1) the presence of figures in the test sample and (2) the language of the sample.

Table 9: Average accuracy by model (best in **bold**, second best <u>underlined</u>)

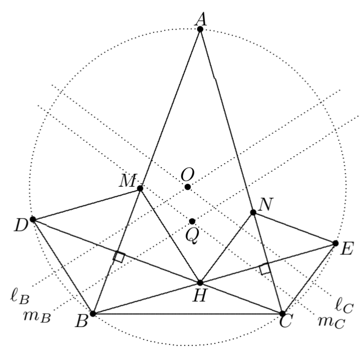
Model	**Has figures**	**No figures**
`Ministral-3B`	4.72%	6.95%
`grok-3`	12.25%	19.39%
`gpt-5-mini`	<u>75.51%</u>	<u>85.20%</u>
`gpt-5-nano`	63.24%	74.14%
`gpt-4.1`	26.56%	39.29%
`gpt-4o`	9.38%	17.03%
`gpt-5`	**76.56%**	**88.50%**
`DeepSeek-R1`	16.48%	26.53%
`DeepSeek-V3-0324`	5.53%	10.14%
`Llama-4-Maverick-17B`	13.94%	21.43%

We find that all models experience some drop in accuracy when problems include figures, regardless of their overall capability. This suggests that visual–text integration remains a challenging dimension across the board.

However, the magnitude of the drop differs substantially by model strength. Less capable models (e.g., Ministral-3B, DeepSeek-V3) not only perform poorly overall but also show consistently low accuracy on figure-based questions, indicating limited visual understanding. In contrast, stronger models such as GPT-5 and GPT-5 mini still exhibit a small decline when figures are present, but they maintain high absolute performance in both settings, demonstrating comparatively robust multimodal reasoning.

Problem with Figure

Problem. Let $\triangle ABC$ be an acute triangle satisfying $AB > BC$ and $AC > BC$. Let O and H be the circumcentre and orthocentre of $\triangle ABC$. The circumcircle of $\triangle AHC$ meets AB again at $M \neq A$, and the circumcircle of $\triangle AHB$ meets AC again at $N \neq A$. Prove that the circumcentre of $\triangle MNH$ lies on the Euler line OH.



GPT-5 Solution.

Let $\triangle ABC$ be acute with $AB > BC$ and $AC > BC$. Define M and N as in the problem.
1) Angles from the defining circles. Because M lies on the circumcircle of AHC and N on that of AHB,

$$\angle AMH = \angle ANH = 90^\circ - A.$$

This yields

$$\angle MHN = 180^\circ - A = \angle BHC.$$

2) Similarity. Angles at M and B are both 90°; similarly for N and C. Thus

$$\triangle HMN \sim \triangle HBC,$$

implying a spiral similarity at H.
3) Circumcentres. Let X be the circumcentre of $\triangle HMN$ and U that of $\triangle HBC$. Spiral similarity about H sends U to X, hence H, U, X are collinear. Since U lies on OH, we conclude $X \in OH$.

$$\boxed{X \in OH.}$$

Mistral-3B Solution.

1. Identify points: O = circumcentre of ABC, H = orthocentre, M, N from intersecting the circles of AHC and AHB with AB, AC.
2. Geometry facts: Since M, N lie on those circles, they relate to the orthocentric configuration. OH is the Euler line of ABC.
3. Euler line property (stated): The model asserts that circumcentres of triangles formed from points on such circles and H lie on OH, though this is *not* a valid theorem.
4. Conclusion (unjustified): Therefore, the circumcentre of $\triangle MNH$ lies on OH.

$$\boxed{\text{Claim: } X \in OH.}$$

(Note: This argument is not mathematically valid.)

Official Solution.

Extend CH and BH to meet the circumcircle of ABC at D and E. From cyclic quadrilateral $AMHC$ and the perpendicularity relations,

$$\angle BMH = \angle MBH,$$

so CD is a symmetry axis of $\triangle BMH$. Similarly AB is a symmetry axis of $\triangle DBH$. Thus $HBDM$ is a rhombus; likewise $HCEN$ is a similar rhombus.
Let ℓ_B, ℓ_C be the perpendicular bisectors of BD and CE; they meet at O. Let m_B, m_C be the perpendicular bisectors of MH and NH; they meet at the circumcentre Q of $\triangle MNH$.
A dilation about H maps $BD \to MH$ and $CE \to NH$, thus mapping $\ell_B \to m_B$ and $\ell_C \to m_C$. Hence the intersection O maps to Q.
Since H is the centre of dilation, points H, O, Q are collinear.

$$\boxed{Q \in OH.}$$

Why This Problem Is Difficult for LLMs. This geometry problem requires a long, multi-step chain of reasoning. The figure encodes critical structural cues, as a result, frontier models like `gpt-5` can reconstruct the full Olympiad-style argument, while weaker models like Mistral-3B fail to produce a valid proof.

Table 10: Average accuracy by model and language (best in **bold**, second best underlined).

Model	en	zh	es	mn
Ministral-3B	6.22%	1.55%	4.60%	0%
grok-3	17.00%	7.40%	10.34%	15.38%
gpt-5-mini	82.06%	65.62%	77.65%	50.00%
gpt-5-nano	70.56%	51.36%	65.12%	49.65%
gpt-4.1	34.79%	18.35%	50.00%	0%
gpt-4o	14.35%	2.72%	21.61%	15.38%
gpt-5	**84.27%**	**74.38%**	**78.16%**	**63.64%**
DeepSeek-R1	23.23%	5.43%	15.91%	14.29%
DeepSeek-V3-0324	8.49%	3.53%	12.79%	0%
Llama-4-Maverick	18.97%	5.47%	13.10%	0%

While several models perform reasonably in English, many degrade sharply in non-English settings—especially Mongolian, where multiple models score 0%. Even high-performing models show reduced accuracy in Mongolian, but the relative drop is far smaller: GPT-5 and GPT-5 mini remain the strongest models across all tested languages and are the least affected by cross-lingual shifts. This indicates that although Mongolian remains an especially difficult language for current LLMs, frontier-tier models exhibit significantly improved multilingual robustness.

A.5 Error Analysis

We present both quantitative and qualitative analyses of model performance. First, we engaged human graders to record observations of failure cases across 1,470 generated solutions. Second, we measured model performance across 82 distinct skills and topics within the Math Olympiad curriculum (see Appendix A.2).

Based on grader feedback, the models demonstrate high proficiency in predicting the final answer (87.3% average accuracy); however, they struggle to generate coherent, rigorous proofs. The most common failure mode is attempting to generalize from specific examples, assuming this constitutes a sufficient proof. For instance, in Number Theory, models tend to verify cases modulo different primes and conclude the proof is complete without rigorous generalization. In Functional Equations, models often identify simple candidate solutions (e.g., linear, constant, or quadratic forms) and assume these are the unique solutions, failing to prove that no other solutions exist.

Regarding the skillset breakdown, we found that LLMs struggle most significantly with Combinatorics problems that require clever construction and cannot be solved via brute force. Number Theory also presents significant challenges. Conversely, models perform best in Algebraic problems that can be expressed purely through equation manipulation.

Finally, we observed a specific failure case in our MathRAG experiments: performance degrades when the retrieved problem (via embed-RAG) is irrelevant. This distraction causes a performance drop in 22% of such cases.

A.6 Dataset Statistics and Examples

We report summary statistics including per-language and per-domain distributions, subtopic frequencies, and problem/solution length profiles, with additional visualizations. For access to full dataset, refer to http://mathnet.netlify.app/.

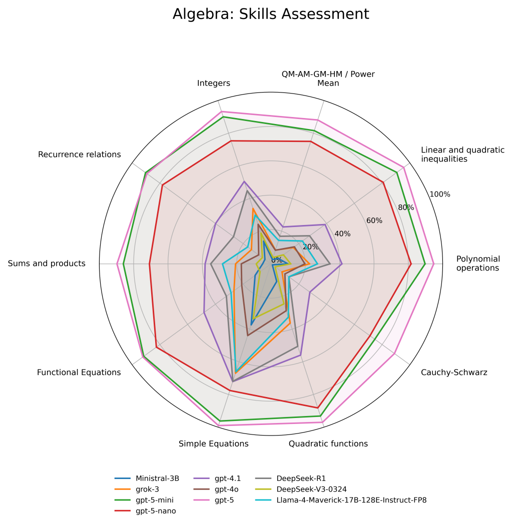
Target Problem and Expert RAG Problem

Target Problem. Show that there are no 2-tuples (x, y) of positive integers satisfying

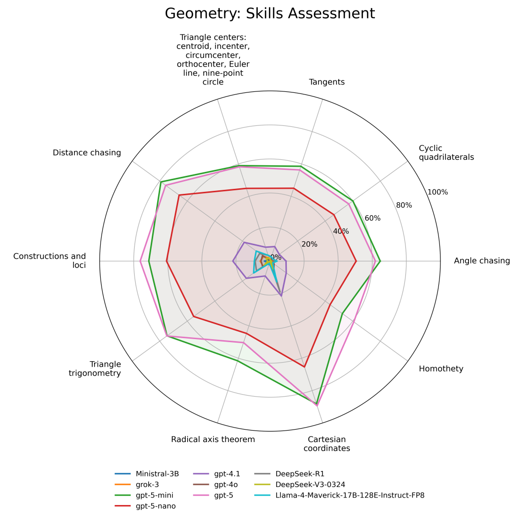
$$(x+1)(x+2)\cdots(x+2014) \;=\; (y+1)(y+2)\cdots(y+4028).$$

Source: 2014 Chinese TST

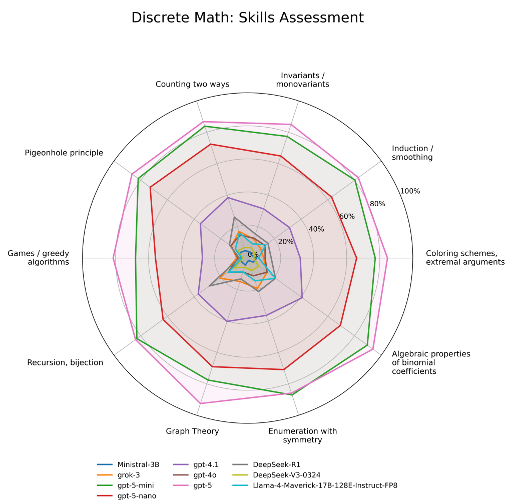
RAG-Expert Problem. Alireza multiplied one billion consecutive natural numbers, while Matin multiplied two million consecutive natural numbers. Prove that their two products cannot be equal; therefore, if they claim to have obtained the same number, at least one of them must have made a mistake.



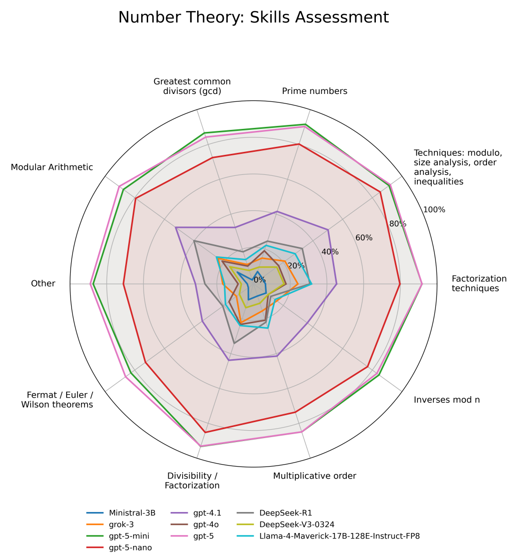
(a) Algebra



(b) Geometry



(c) Discrete Mathematics



(d) Number Theory

Figure 5: Breakdown of performance across four domains: (a) Algebra, (b) Geometry, (c) Discrete Mathematics, (d) Number Theory.

Language	English	Spanish	Arabic	Russian	Roman	Bulgarian	Persian	German	Chinese	Ukrainian	Mongolian
Count	16154	242	200	180	60	52	70	23	418	83	30

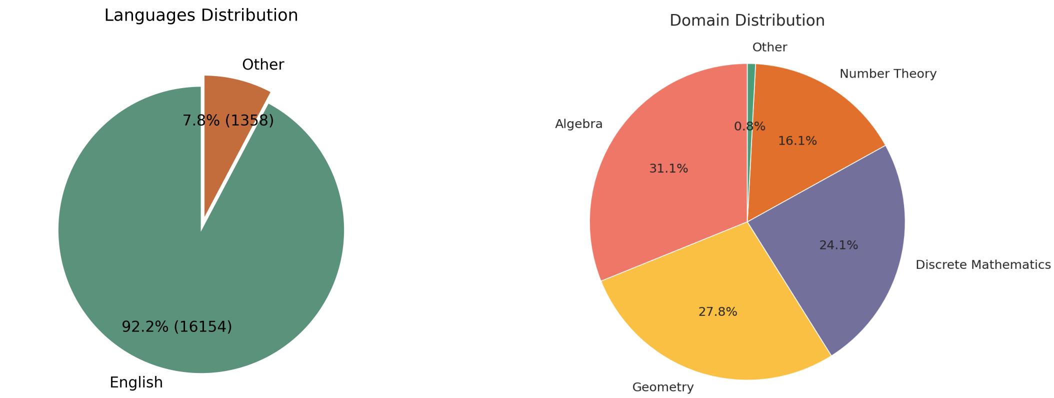
Table 11: Problems Distribution per Language

A.7 PROMPTS

We include the core prompts used for extraction, evaluation, and metadata classification. These are the exact versions used in our experiments.

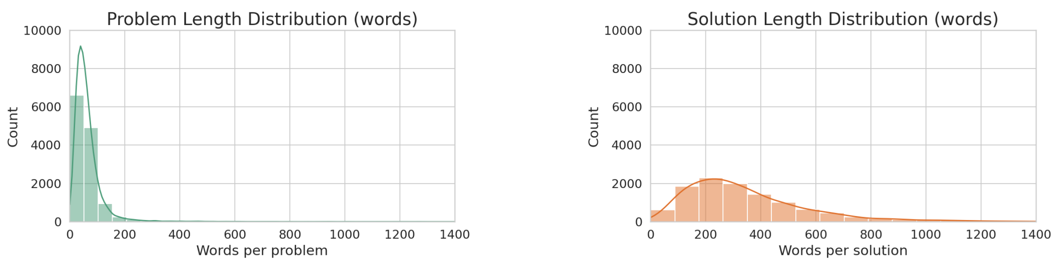
Listing 1: System prompt for solution extraction

```
sys_prompt = """
    You are an expert in extracting mathematical problems and solutions.
    I will provide you with:
        - One math problem
```



(a) Problems Distribution per Language (b) Problems Distribution per Domain

Figure 6: Distribution of problems across languages and domains.



(a) Problem Length Distribution (Words) (b) Solution Length Distribution (Words)

Figure 7: Problems vs Solutions (Length Distribution) (words)

```
    - Multiple pages
Extract the solution that matches the problem
Important instructions:
- If the problem statement is split into multiple numbered points,
extract the solution in multiple points
- Never leave 'solution_text' empty. If no solution can be found,
write '"Not found"' as the value.
- If solution contains imgs make sure to extractt image path such as:
 ![](images/Argentina2022_p5_data_8b4126bbff.png)
- If solution coontains tables make sure to extract the tables such
as: <table><thead><tr><th>Team</th><th>T1</th><th>T2</th><th>T3</th><
th>T4</th><th>T5</th><th>T6</th><th>T7</th><th>T8</th><th>Total</th></
tr></thead><tbody><tr><td>T1</td><td>-</td><td>2</td><td>2</td><td>2</
td><td>2</td><td>2</td><td>2</td><td>2</td><td>14</td></tr><tr><td>T2
</td><td>0</td><td>-</td><td>2</td><td>2</td><td>2</td><td>2</td><td
>2</td><td>2</td><td>12</td></tr><tr><td>T3</td><td>0</td><td>0</td><
td>-</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td><td>10</td
></tr><tr><td>T4</td><td>0</td><td>0</td><td>0</td><td>-</td><td>2</
td><td>2</td><td>2</td><td>2</td><td>8</td></tr><tr><td>T5</td><td>0</
td><td>0</td><td>0</td><td>0</td><td>-</td><td>2</td><td>2</td><td>2</
td><td>6</td></tr><tr><td>T6</td><td>0</td><td>0</td><td>0</td><td>0</
td><td>0</td><td>-</td><td>2</td><td>2</td><td>4</td></tr><tr><td>T7</
td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>-</
td><td>2</td><td>2</td></tr><tr><td>T8</td><td>0</td><td>0</td><td>0</
td><td>0</td><td>0</td><td>0</td><td>0</td><td>-</td><td>0</td></tr></
tbody></table>
- Follow the JSON schema below precisely:
```json
{
```
```

(a) Algebra

(b) Geometry

(c) Discrete Mathematics

(d) Number Theory

Figure 8: Domain subtopic distribution

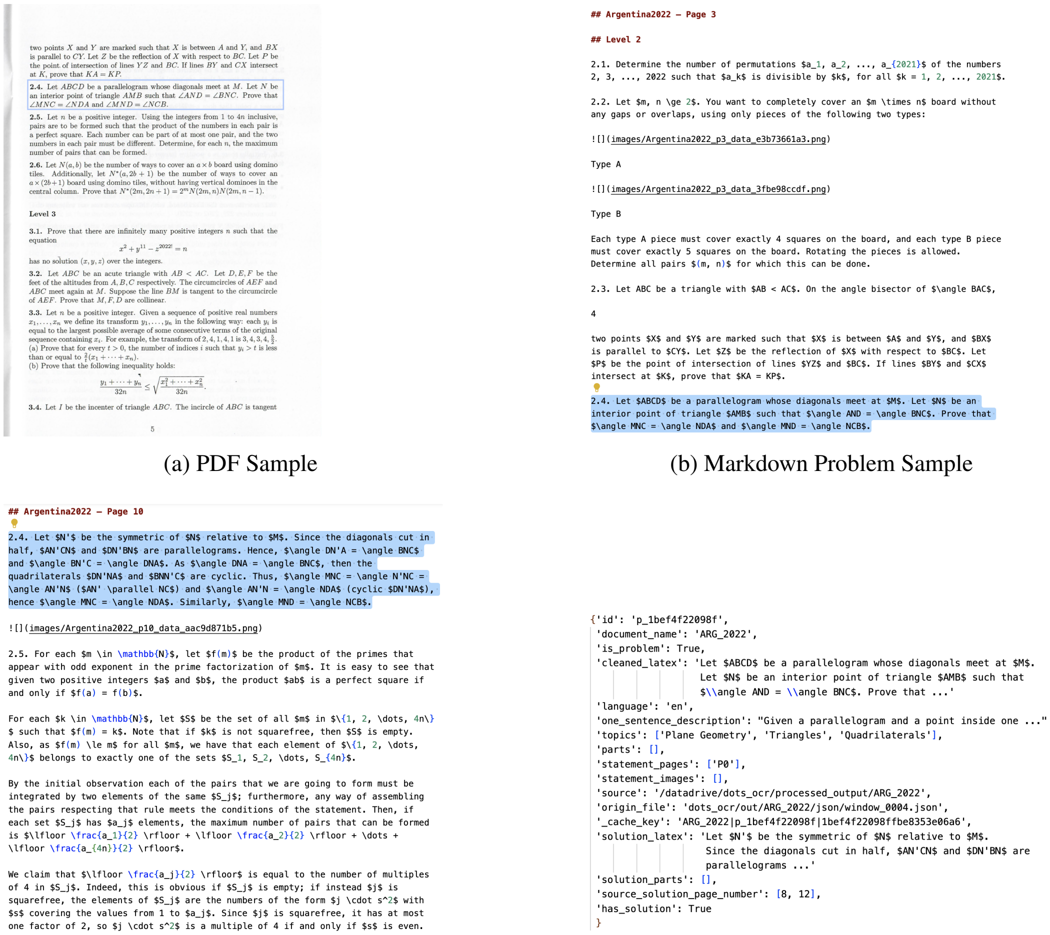
```
        "has_solution": "bool, if solution was found and extracted set to
 true, else false"
        "solution_page_number": "the page number where the solution is
found"
        "solution_latex": "extracted solution in latex format"
        "solution_parts": [
            "part_label": "label of the part"
            "part_latex": "extracted part solution in latex format"
        ]
    }
    ```
"""
```

Listing 2: System prompt for evaluation

```
sys_prompt_eval = """
    You are an expert in evaluating mathematical problems and solutions.

    I will supply you with a problem and its solution(s), including
alternative solutions if available.
    Your task is to evaluate based on the following criteria:

    1. **Extraction completeness:** All main parts of the solution must
have been correctly extracted. Missing or truncated content should be
 noted.
    2. **Problem-solution match:** Ensure that the solution corresponds
correctly to the provided problem. If they are mismatched or
unrelated, it should be noted.
```
```



two points X and Y are marked such that X is between A and Y, and BX is parallel to CY. Let Z be the reflection of X with respect to BC. Let P be the point of intersection of lines YZ and BC. If lines BY and CX intersect at K, prove that $KA = KP$.

2.4. Let $ABCD$ be a parallelogram whose diagonals meet at M. Let N be an interior point of triangle AMB such that $\angle AND = \angle BNC$. Prove that $\angle MNC = \angle NDA$ and $\angle MND = \angle NCB$.

2.5. Let n be a positive integer. Using the integers from 1 to $4n$ inclusive, pairs are to be formed such that the product of the numbers in each pair is a perfect square. Each number can be part of at most one pair, and the two numbers in each pair must be different. Determine, for each n, the maximum number of pairs that can be formed.

2.6. Let $N(a, b)$ be the number of ways to cover an $a \times b$ board using domino tiles. Additionally, let $N^*(a, 2b+1)$ be the number of ways to cover an $a \times (2b+1)$ board using domino tiles, without having vertical dominoes in the central column. Prove that $N^*(2m, 2n+1) = 2^m N(2m, n) N(2m, n-1)$.

Level 3

3.1. Prove that there are infinitely many positive integers n such that the equation

$$x^2 + y^{11} - z^{2022!} = n$$

has no solution (x, y, z) over the integers.

3.2. Let ABC be an acute triangle with $AB < AC$. Let D, E, F be the feet of the altitudes from A, B, C respectively. The circumcircles of AEF and ABC meet again at M. Suppose the line BM is tangent to the circumcircle of AEF. Prove that M, F, D are collinear.

3.3. Let n be a positive integer. Given a sequence of positive real numbers $x_1, \dots, x_n$ we define its transform $y_1, \dots, y_n$ in the following way: each y_i is equal to the largest possible average of some consecutive terms of the original sequence containing x_i. For example, the transform of $2, 4, 1, 4, 1$ is $3, 4, 3, 4, \frac{5}{2}$.
(a) Prove that for every $t > 0$, the number of indices i such that $y_i > t$ is less than or equal to $\frac{2}{t}(x_1 + \cdots + x_n)$.
(b) Prove that the following inequality holds:

$$\frac{y_1 + \cdots + y_n}{32n} \le \sqrt{\frac{x_1^2 + \cdots + x_n^2}{32n}}.$$

3.4. Let I be the incenter of triangle ABC. The incircle of ABC is tangent

5

(a) PDF Sample

```
## Argentina2022 – Page 3

## Level 2

2.1. Determine the number of permutations $a_1, a_2, ..., a_{2021}$ of the numbers
2, 3, ..., 2022 such that $a_k$ is divisible by $k$, for all $k = 1, 2, ..., 2021$.

2.2. Let $m, n \ge 2$. You want to completely cover an $m \times n$ board without
any gaps or overlaps, using only pieces of the following two types:

![](images/Argentina2022_p3_data_e3b73661a3.png)

Type A

![](images/Argentina2022_p3_data_3fbe98ccdf.png)

Type B

Each type A piece must cover exactly 4 squares on the board, and each type B piece
must cover exactly 5 squares on the board. Rotating the pieces is allowed.
Determine all pairs $(m, n)$ for which this can be done.

2.3. Let ABC be a triangle with $AB < AC$. On the angle bisector of $\angle BAC$,

4

two points $X$ and $Y$ are marked such that $X$ is between $A$ and $Y$, and $BX$
is parallel to $CY$. Let $Z$ be the reflection of $X$ with respect to $BC$. Let
$P$ be the point of intersection of lines $YZ$ and $BC$. If lines $BY$ and $CX$
intersect at $K$, prove that $KA = KP$.

2.4. Let $ABCD$ be a parallelogram whose diagonals meet at $M$. Let $N$ be an
interior point of triangle $AMB$ such that $\angle AND = \angle BNC$. Prove that
$\angle MNC = \angle NDA$ and $\angle MND = \angle NCB$.
```

(b) Markdown Problem Sample

```
## Argentina2022 – Page 10

2.4. Let $N'$ be the symmetric of $N$ relative to $M$. Since the diagonals cut in
half, $AN'CN$ and $DN'BN$ are parallelograms. Hence, $\angle DN'A = \angle BNC$
and $\angle BN'C = \angle DNA$. As $\angle DNA = \angle BNC$, then the
quadrilaterals $DN'NA$ and $BNN'C$ are cyclic. Thus, $\angle MNC = \angle N'NC =
\angle AN'N$ ($AN' \parallel NC$) and $\angle AN'N = \angle NDA$ (cyclic $DN'NA$),
hence $\angle MNC = \angle NDA$. Similarly, $\angle MND = \angle NCB$.

![](images/Argentina2022_p10_data_aac9d871b5.png)

2.5. For each $m \in \mathbb{N}$, let $f(m)$ be the product of the primes that
appear with odd exponent in the prime factorization of $m$. It is easy to see that
given two positive integers $a$ and $b$, the product $ab$ is a perfect square if
and only if $f(a) = f(b)$.

For each $k \in \mathbb{N}$, let $S$ be the set of all $m$ in $\{1, 2, \dots, 4n\}
$ such that $f(m) = k$. Note that if $k$ is not squarefree, then $S$ is empty.
Also, as $f(m) \le m$ for all $m$, we have that each element of $\{1, 2, \dots,
4n\}$ belongs to exactly one of the sets $S_1, S_2, \dots, S_{4n}$.

By the initial observation each of the pairs that we are going to form must be
integrated by two elements of the same $S_j$; furthermore, any way of assembling
the pairs respecting that rule meets the conditions of the statement. Then, if
each set $S_j$ has $a_j$ elements, the maximum number of pairs that can be formed
is $\lfloor \frac{a_1}{2} \rfloor + \lfloor \frac{a_2}{2} \rfloor + \dots +
\lfloor \frac{a_{4n}}{2} \rfloor$.

We claim that $\lfloor \frac{a_j}{2} \rfloor$ is equal to the number of multiples
of 4 in $S_j$. Indeed, this is obvious if $S_j$ is empty; if instead $j$ is
squarefree, the elements of $S_j$ are the numbers of the form $j \cdot s^2$ with
$s$ covering the values from 1 to $a_j$. Since $j$ is squarefree, it has at most
one factor of 2, so $j \cdot s^2$ is a multiple of 4 if and only if $s$ is even.
```

(c) Markdown Solution Sample


```
{'id': 'p_1bef4f22098f',
 'document_name': 'ARG_2022',
 'is_problem': True,
 'cleaned_latex': 'Let $ABCD$ be a parallelogram whose diagonals meet at $M$.
                   Let $N$ be an interior point of triangle $AMB$ such that
                   $\\angle AND = \\angle BNC$. Prove that ...'
 'language': 'en',
 'one_sentence_description': "Given a parallelogram and a point inside one ..."
 'topics': ['Plane Geometry', 'Triangles', 'Quadrilaterals'],
 'parts': [],
 'statement_pages': ['P0'],
 'statement_images': [],
 'source': '/datadrive/dots_ocr/processed_output/ARG_2022',
 'origin_file': 'dots_ocr/out/ARG_2022/json/window_0004.json',
 '_cache_key': 'ARG_2022|p_1bef4f22098f|1bef4f22098ffbe8353e06a6',
 'solution_latex': 'Let $N'$ be the symmetric of $N$ relative to $M$.
                    Since the diagonals cut in half, $AN'CN$ and $DN'BN$ are
                    parallelograms ...'
 'solution_parts': [],
 'source_solution_page_number': [8, 12],
 'has_solution': True
 }
```


(d) LLM Output Sample

Figure 9: Sample Input Data

```
3. **Solution completeness:** Check if the reasoning is fully present
 in the extracted solution. Minor implicit steps are acceptable, but
missing entire parts should be flagged.

**Important:**
- Since these problems and solutions are authored by experts, do NOT
reject for correctness. Focus only on extraction issues or mismatches.

- Do NOT reject for minor omissions, terseness, formatting, or style.
- Only reject if there is a **clear, significant extraction issue**
or the solution does not match the problem.
- Always provide a clear reason if rejecting, mentioning which
criteria are affected

Provide your evaluation strictly in JSON format:

```json
{
  "final_verdict": "accept" or "reject",
  "reason": "A concise explanation for your decision, mentioning
which criteria failed if rejected"
}
    ```
Do not add any extra commentary outside the JSON.
```

```
    """
```

Listing 3: System prompt for topics, final answer, and metadata extraction

```
SYSTEM_PROMPT = r"""
You are a rigorous matholympiad content analyzer.

You will be given one problem package containing:
- A problem statement
- One or more official solutions (labeled Solution 1, Solution 2, )
- Optional final answers

Your tasks are:

======================================================================
1. TOPIC EXTRACTION
======================================================================
- Assign the problem its most specific topics from the taxonomy.
- Each topic path must be an array of strings from general  specific.
- Include ALL paths relevant to the problem or solutions.
- Every topic path must be a verbatim copy of a path from the taxonomy.
- No paraphrasing, renaming, reordering, or combining nodes.
- Every topic must begin with "Topics".


======================================================================
2. MAIN IDEAS / TRICKS / TOOLS
======================================================================
- Produce a bullet list of the key structural insights or tools used.
- Examples:
  - Techniques used
  - Classical lemmas or theorems applied
  - Core inequality strategies
  - Key constructions or combinatorial ideas
- Do NOT retell the whole solution; extract the essential tools.

======================================================================
3. NATURAL-LANGUAGE PROBLEM DESCRIPTION
======================================================================
- Summarize the core task of the problem in normal English.
- NO mathematical symbols at all (no variables, no equations, no angle
   notation, etc.)
- A high-level, intuitive, short description.

======================================================================
4. PROBLEM TYPE CLASSIFICATION
======================================================================
Classify the problem into exactly one of the following:

- "proof only": no explicit final numeric/closed-form answer is required.
- "final answer only": problem only asks for a value/choice with no proof
    required.
- "proof and answer": requires both reasoning and a final value/statement.

- "MCQ": problem requires choosing from given options.

======================================================================
5. FINAL ANSWER EXTRACTION
======================================================================
- If the problem requires a final numeric/closed-form expression, value,
   or choice, extract it.
- If the problem's nature does NOT require a final answer (e.g., proof-
   only), output `null`.

Specific rules:
```

```
- If multiple solutions exist, the final answer must match the official
    answer section if present.
- Accept integers, expressions, ranges, choices, constructed forms, etc.
- For MCQ, return the *selected option* if identifiable; otherwise null.

======================================================================
TAXONOMY BLOCK
======================================================================
Use this taxonomy for the topics field.
Each topic path must follow the hierarchy strictly.

 Topics
  Geometry
   Plane Geometry
    Triangles
     Triangle centers: centroid, incenter,
       circumcenter, orthocenter, Euler line,
       nine-point circle
     Triangle inequalities
     Triangle trigonometry
    Quadrilaterals
     Cyclic quadrilaterals
     Inscribed/circumscribed quadrilaterals
     Quadrilaterals with perpendicular diagonals
    Circles
     Coaxal circles
     Tangents
     Radical axis theorem
     Circle of Apollonius
    Concurrency and Collinearity
     Cevas theorem
     Menelaus theorem
        ... (more topics here)
======================================================================
OUTPUT FORMAT (STRICT JSON)
======================================================================

Return ONLY a JSON object:

{
  "topics": [
      ["Topics", "...", "..."],
      ["Topics", "...", "..."]
  ],
  "main_ideas": [
      "key idea 1",
      "key idea 2",
      "key idea 3"
  ],
  "natural_language_description": "...",
  "final_answer": "... or null",
  "problem_type": "proof only | final answer only | proof and answer |
    MCQ",
  "confidence": 0.01.0
}

Rules:
- NO text outside the JSON.
- NO markdown in the output.
- natural_language_description must contain zero mathematical symbols.
- Confidence reflects how certain you are about the classification.
"""
```

A.8 LLMs Usage in the Paper

The authors made use of large language models (LLMs) primarily to support the writing process, including polishing the text for clarity and readability. In addition, LLMs were employed to assist in refining the design of the project website as well as the interface used by annotators.