Can Large Language Models Win the International Mathematical Games?

Alessio Cocchieri* **Luca Ragazzi*** **Giuseppe Tagliavini**
Lorenzo Tordi **Antonella Carbonaro** **Gianluca Moro***
Department of Computer Science and Engineering, University of Bologna, Italy
{a.cocchieri, l.ragazzi, giuseppe.tagliavini, antonella.carbonaro, gianluca.moro}@unibo.it lorenzo.tordi2@studio.unibo.it

Abstract

Recent advances in large language models (LLMs) have demonstrated strong mathematical reasoning abilities, even in visual contexts, with some models surpassing human performance on existing benchmarks. However, these benchmarks lack structured age categorization, clearly defined skill requirements, and–crucially–were not designed to assess human performance in international competitions. To address these limitations, we introduce MATHGAMES, a new benchmark of 2,183 high-quality mathematical problems (both text-only and multimodal) in an open-ended format, sourced from an international mathematical games championships. Spanning seven age groups and a skill-based taxonomy, MATHGAMES enables a structured evaluation of LLMs' mathematical and logical reasoning abilities. Our experiments reveal a substantial gap between state-of-the-art LLMs and human participants–even 11-year-olds consistently outperform some of the strongest models–highlighting the need for advancements. Further, our detailed error analysis offers valuable insights to guide future research. The data is publicly available at https://disi-unibo-nlp.github.io/math-games/.

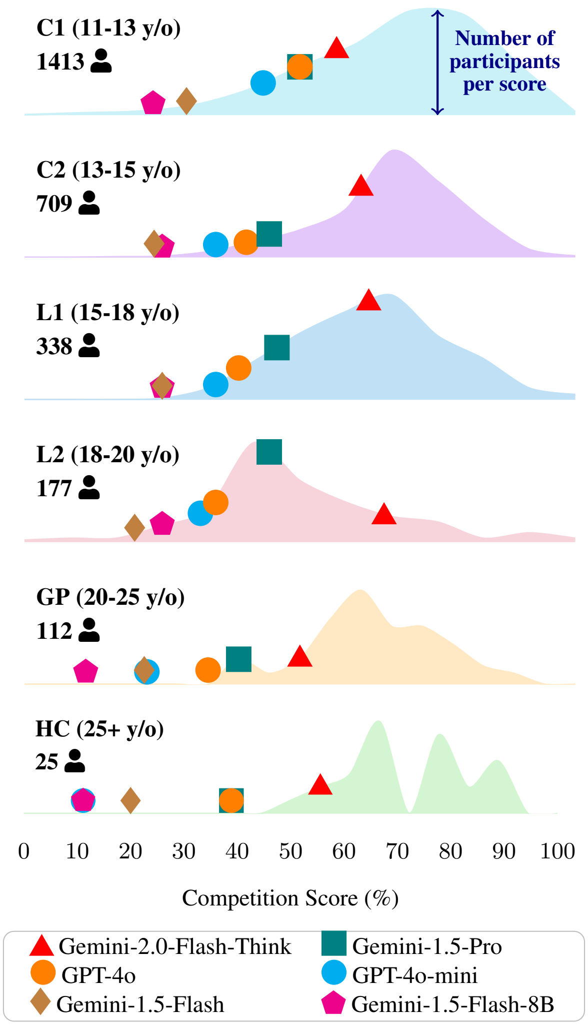


Figure 1: **Human scores by age group in the 2024 Finals of the International Mathematical and Logical Games Championships (MATHGAMES).** Markers indicate model performance for comparison.

1 Introduction

Large language models (LLMs) (Zhao et al., 2023) and large multimodal models (LMMs) (Yin et al., 2023) have made significant progress in mathematical problem-solving (Cobbe et al., 2021; Hendrycks et al., 2021) and complex puzzle interpretation (Ghosal et al., 2024; Giadikiaroglou et al., 2024), raising open questions about their capacity for human-like reasoning and generalization (Cocchieri et al., 2025c), particularly in low-resource scenarios (Domeniconi et al., 2016; Moro and Ragazzi, 2022, 2023). Recent advancements, including models like InternVL (Chen et al., 2023) and Gemini (Reid et al., 2024), have even demonstrated performance surpassing human scores on structured math benchmarks (Lu et al., 2024). Now, with the rise of reasoning-focused models such as OpenAI's o3-mini and DeepSeek-R1 (Guo et al.,

*Equal contribution (co-first authors).

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9657–9683
November 4-9, 2025 ©2025 Association for Computational Linguistics

Benchmark	Size			
MATH (2021)	12,500			
GSM8K (2021)	8,500			
GSM-Plus (2024c)	10,552			
MathBench (2024a)	3,709		✓	
CHAMP (2024)	270			
ConceptMath (2024)	4,011			
OlympiadBench (2024)	8,476	✓		✓
MathVista (2024)	6,141	✓		
MathVision (2024)	3,040	✓		✓
SMART-840 (2024)	840	✓		✓
MathVerse (2024b)	15,000	✓		
MM-Math (2024)	5,929	✓		
MATHGAMES (*ours*)	2,183	✓	✓	✓

Table 1: **Comparison of MATHGAMES with existing math benchmarks.** We highlight size, multimodality, presence of age groups, and competition-specific design.

2025), AI is increasingly designed for complex logical reasoning. However, as these models advance, rigorous evaluation is crucial to determine whether they genuinely exhibit human-like problem-solving skills or merely excel in controlled environments. This raises a key question: are LLMs ready to compete in international competitions designed to assess human logical and mathematical reasoning?

To answer this, we first examined the existing most related math benchmarks (see Table 1) and identified several key limitations: ❶ Most focus solely on text-based problem-solving (Kurtic et al., 2024; Liu et al., 2024a; Mao et al., 2024; Wu et al., 2024), neglecting the visual reasoning required in the real world. ❷ Many recycle existing datasets (Chiang and Lee, 2024; Li et al., 2024a,c), limiting their ability to introduce fresh challenges. ❸ Some lack structured age categories with progressive difficulty scale (Cherian et al., 2024; He et al., 2024; Lu et al., 2024; Zhang et al., 2024b), which are essential for assessing problem-solving development across different expertise levels. ❹ Finally, existing benchmarks favor computation over logical reasoning and were not originally designed to assess human problem-solving abilities.

These limitations may misrepresent model capabilities and skew research priorities. To address this, we introduce MATHGAMES, a new benchmark sourced from a real-world international math competition spanning 20 years (1994-2024). It consists of 2,183 high-quality, playful-style problems in an open-ended format (i.e., without multiple-choice answers), including 1,389 textual and 794 multimodal exercises, ensuring a comprehensive evaluation. These problems span seven age groups with a progressively increasing difficulty, ranging from 8-year-old children to professionals aged 25+. Moreover, we construct a coarse-grained taxonomy capturing the mathematical skills required to solve each problem, enabling a clearer understanding of LLM capabilities. To ensure the highest data quality, all problem-solution pairs were further validated by annotators.

Using MATHGAMES, we benchmarked 28 models, including state-of-the-art LLMs, LMMs, and reasoning-focused models. The results revealed a substantial performance gap between models and human competitors, even in entry-level age group tests (see Figure 1), underscoring critical areas for improvement. We believe that mastering these tasks is crucial for real-world applications in STEM education and scientific research. By mirroring real championship-level settings, MATHGAMES provides a critical foundation for future studies, complementing existing benchmarks with age-related LLM performance and mathematic skill analysis.

In conclusion, our contributions are threefold:

- **MATHGAMES**, a multimodal benchmark grounded in a real-world math competition, featuring problems categorized by age group and mathematical skill.
- **A comprehensive evaluation of 28 models**, revealing that several state-of-the-art models underperform compared to young students.
- **Actionable insights**, through detailed error analysis and a public dataset to support future model development and reasoning research.

2 Related Work

Multimodal Reasoning Benchmarks A variety of benchmarks have been introduced to evaluate the mathematical reasoning capabilities of models, including geometry-focused like GeoEval (Zhang et al., 2024a), as well as multidisciplinary like MMMU (Yue et al., 2024) and OlympicArena (Huang et al., 2024). However, existing math and logic benchmarks lack structured, age-based difficulty progressions, making it hard to assess model performance relative to human cognitive development–as is done in human competitions that identify top performers at each developmental stage. Further, some benchmarks rely on synthetically generated problems (Kurtic et al., 2024; Rahman et al., 2024), limiting realism. The prevalence of multiple-choice formats further reduces their effectiveness, as models can exploit answer elimi-

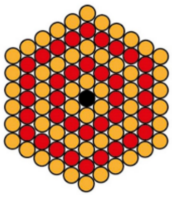
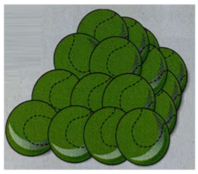
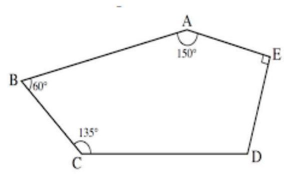
Early Teenager (11-13 y/o)	**Late Teenager (13-18 y/o)**	**Adult (18-25+ y/o)**
		
Question: How many small spheres of different colors are there in the figure?	**Question**: In figure you see tennis balls placed on top of each other, forming at each "plane" of the squares, without holes in the middle. The highest level contains only one ball; the second, coming down, contains 4; the third contains 9 and so on. If you use 7714 balls, how many floors will your pyramid of tennis balls be constituted?	**Question**: In figure you see a pentagonal tile, quite singular, whose sides BC and AE measure 1 dm while AB measures 2 dm. Which is in cm^2, rounded to the nearest cm^2, the area of our tile? (If necessary, use 1,414 for $\sqrt{2}$ and 1,732 for $\sqrt{3}$).
Answer: 91 **Category**: C1	**Answer**: 28 **Category**: C2, L1, L2	**Answer**: 255 cm^2 **Category**: L2, GP, HC

Table 2: **Example of multimodal problems from MATHGAMES, targeting different age groups with increasing levels of difficulty.** Each problem includes the correct answer and corresponding category.

nation strategies rather than demonstrating genuine problem-solving skills. In contrast, MATHGAMES features math exercises with an open-ended resolution format, directly sourced from an international competition and structured by age groups with well-defined difficulty levels. This aligns with prior work on domain transfer in NLP, focused on generalizing beyond training patterns (Cerroni et al., 2013, 2015; Domeniconi et al., 2017; Moro et al., 2018; Frisoni et al., 2024; Cocchieri et al., 2025a,b).

Math-focused LLMs Numerous models have been developed for mathematics. Among closed models, Minerva (Lewkowycz et al., 2022) was an early large-scale math-trained model, followed by leading contenders like GPT-4 (OpenAI, 2023) and Gemini-1.5. Some models have open weights but were partially trained on private data, including Wizard-Math (Luo et al., 2023), DeepSeek-Math (Shao et al., 2024), and Qwen2.5-Math (Yang et al., 2024). Others, such as Llemma (Azerbayev et al., 2024), MetaMath (Yu et al., 2024), and Dart-Math (Tong et al., 2024), rely on open data. Notably, NuminaMath (Li et al., 2024b) introduces a dataset of 860K math competition question pairs, while FineMath (Liu et al., 2024b) curates 34B-54B tokens of mathematical content from CommonCrawl. Recently, focus has shifted toward reasoning-centric LLMs with strong logical capabilities. OpenAI's o1 pioneered this shift, followed by Gemini-Thinking, while DeepSeek-R1 marked a major leap in open-source development.

3 MATHGAMES

We introduce MATHGAMES, a carefully designed benchmark for evaluating the mathematical and logical reasoning abilities of foundation models across both text-only and multimodal problems across different age categories (see examples in Table 2).

3.1 Competition Background

Our benchmark is built on the International Championship of Mathematical and Logical Games, a long-standing annual competition that has engaged participants worldwide since 1994. Each year, thousands of students and adults–from primary school to university level–take part in national qualifying rounds, culminating in an international final. In 2024, over 500 finalists from 15 different countries competed in the final phase. Each country organizes its own preliminary rounds through a dedicated national center, following shared guidelines to ensure consistency in problem difficulty and style across all participants. In this work, we focus on the Italian problem sets curated by the PRISTEM center,[1] the official Italian organizer, from whom we obtained explicit permission to use and translate to English the material in compliance with licensing and copyright regulations. The competition is formally recognized by the Italian Ministry of Education (MIUR) as part of the Program for the Promotion of Excellence.

[1] https://giochimatematici.unibocconi.eu

Age Group y/o		CE 8-10	C1 11-13	C2 13-15	L1 15-18	L2 18-20	GP 20-25	HC 25+
Math Skills								
Arithmetic		25	92	176	168	163	60	25
		2	8	11	8	7	6	4
Logic		30	88	147	145	142	65	36
		28	83	100	91	86	59	40
Pattern Recognition		13	37	65	72	68	35	14
		13	25	35	37	33	16	8
Geometry		5	20	79	101	127	64	33
		13	55	131	150	178	82	58
Combinatorics		22	83	207	228	251	117	74
		38	97	125	122	122	85	54
Algebra		12	85	261	281	294	108	51
		21	46	70	71	69	35	19
Total		222	719	1,407	1,474	1,540	732	416
Avg Len		62.0	69.9	66.8	68.7	70.9	82.1	87.8

Table 3: **Statistics of MathGames, including problem count and word lengths.** The overall count is not the sum of category-specific counts due to overlapping problems (see Figure 2 for a visual reference).

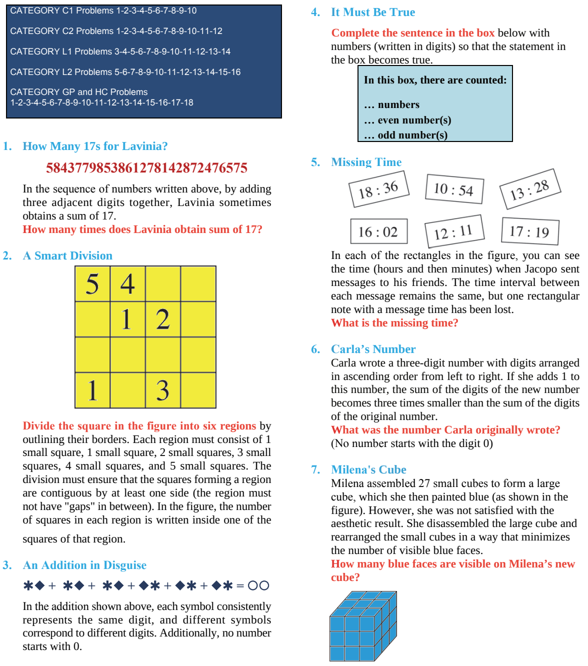


Figure 2: **Example of an English-translated competition exam.** The blue box indicates the exercises to be solved for each age group. Best viewed if zoomed in.

3.2 Data

Composition Our dataset comprises 2,183 manually curated problems, including 1,389 (63.6%) text-only and 794 (36.4%) visual problems. These span seven age categories. For further details on each category and stage of the competition, please refer to Appendix A. Notably, as shown in Table 3, average problem length increases with age group. We highlight that the dataset provides only final answers, without reasoning steps.

Preparation The competition materials were originally stored as separate PDFs for exercises and solutions. To extract both text and images, we used the pypdf library (see Appendix G for details). Yet, we excluded exams from 1994, 1995, and 1997 from the final dataset due to the absence of solutions. Since the original content was in Italian, we translated it into English using GPT-4o, followed by a careful internal evaluation to ensure accuracy (see Appendix D for details). This step is motivated by the fact that most LLMs are primarily trained to reason in English, struggling in other languages across tasks (Moro et al., 2023a). Notably, Italian candidates receive the problems in Italian even during the international final, as each participant is provided with an identical set of problems translated into their native language.

Contamination The dataset's structure inherently mitigates contamination through several factors: ❶ Sourced exclusively from an official source, the data minimizes the risk of pretraining leakage without explicit consent. ❷ The separation of exercises and solutions reduces the likelihood that models were trained on both together. ❸ The translation process further lowers the chances of models encountering the exact original material. ❹ The lack of explanatory solutions ensures models were never trained on the reasoning path needed to derive the correct answer. ❺ Additionally, following a consolidated approach (Li et al., 2024b; Wang et al., 2024), we analyzed problem similarity in related benchmarks–MathVista and Math-Vision–using n-gram overlap and found no data leakage.

Curation To enhance the quality of our data, we followed a four-stage data curation process: ❶ We manually aligned each problem with its corresponding solution, as they were originally stored in separate files. ❷ We corrected inconsistencies in problem-solution pairs (see Appendix B for a visual representation of detected errors). ❸ We verified potential duplicates and confirm that none exist, as championship rules discourage the recurrence of problems across years. ❹ We categorized exercises by age group, as each document contains multiple problems that must be solved according to specific age categories (see Figure 2 for an example). Problems of different difficulty levels were already stored in separate folders, requiring no further action. Data examples are shown in Appendix I.

Skill-Set Taxonomy To analyze model performance across reasoning types, we defined a coarse-grained taxonomy of six core mathematical skills, consistently observed across years:

- **Arithmetic:** Basic operations, number properties, ratios, and proportions.
- **Logic:** Deductive reasoning, inference, truth/lie problems, and conditionals.
- **Pattern Recognition:** Numerical or visual sequences and structural patterns.
- **Geometry:** Shapes, areas, spatial transformations, and perimeters.
- **Combinatorics:** Counting, arrangements, and strategy-based enumeration.
- **Algebra:** Equations, symbolic manipulation, and relationships between quantities.

We manually annotated a representative sample of 100 problems spanning various years and age groups. To scale this process, we evaluated GPT-4o as an automated annotator on the same sample, achieving high agreement with human labels (macro F1 = 0.94), likely due to the clear-cut nature of the categories. Given this reliability and the cost of manual labeling, we used GPT-4o to classify the full dataset. Each problem is annotated with a single skill tag and age group, enabling dual-perspective evaluation of LLM reasoning abilities (see Table 3 for the number of text-only and multimodal problems across skills).

4 Experiments

We perform a series of experiments to assess model performance on MATHGAMES. Furthermore, we conduct an in-depth error analysis of the best-performing models, examining their error distribution and presenting relevant qualitative examples to illustrate the findings.

Models Our experiments span 28 models, including text-only LLMs and vision-enabled LMMs, varying across key dimensions: (1) backbone architectures, (2) training objectives (general-purpose vs. math-focused instruction tuning), (3) open-source vs. closed-source availability, (4) parameter scale, ranging from 7B to 685B, and (5) optimization for reasoning-intensive tasks vs. general chat-based interactions. A comprehensive description of the models considered is provided in Appendix C.

Human Baseline To establish a reference for human performance, we collected all publicly available data from the official competition website,[2] specifically from the 2024 Italian National Finals. This dataset provides the number of correctly solved problems per participant across all age categories, offering a high-quality and reliable benchmark for comparison. Importantly, this evaluation remains safe from data leakage, as all models considered have a 2023 knowledge cutoff and could not have been exposed to these tests. Since the official scores aggregate performance without distinguishing between text and multimodal problems, we report overall accuracy–consistent with the competition's format, where textual and visual reasoning are treated as a unified whole.

Evaluation Mode We conduct our evaluations in a zero-shot setting, without fine-tuning or few-shot demonstrations. We adopt two standard approaches for mathematical task evaluation: (1) *Chain-of-Thought* (CoT) reasoning (Wei et al., 2022), applied to all models, and (2) *Tool-Integrated Reasoning* (TIR) (Gou et al., 2024), used specifically for math-specialized text-only LLMs. Technically, TIR allows models to leverage a Python interpreter as an auxiliary resource for reasoning tasks (see Appendix E for additional information), enhancing their proficiency in precise calculations. For each model, we use the default system prompt or user instruction guidelines provided by the authors when available. Otherwise, we perform prompt engineering to identify the most effective prompt for the zero-shot setting. Detailed information on the adopted prompts is provided in Appendix H.

Metrics To compare models, we use $pass@1$ and $maj@8$ as reference metrics for both CoT and TIR. The $pass@1$ metric relies on greedy decoding to generate the most probable reasoning path. However, for smaller models, we apply *self-consistency* (Wang et al., 2023), where 8 different reasoning paths are sampled (*maj@8*) with a non-zero temperature, and a majority voting strategy is applied after filtering out ill-formed responses. Since greedy decoding is often suboptimal, self-consistency allows smaller models to generate multiple reasoning paths, leading to more robust predictions and helping them achieve performance closer to larger models while maintaining lower computational costs, in line with efficiency-focused approaches such as knowledge distillation (Italiani et al., 2025) and token pruning (Ragazzi et al.,

[2] Human Results - National Finals 2024

2024). For the self-consistency experiments, we set the temperature t and top_p values according to the model authors' recommendations when available. Otherwise, we adopt nucleus sampling with a t = 0.8 and top_p = 0.95 as the default strategy, following common approaches for reasoning-based sampling in recent literature (Rozière et al., 2023; Gou et al., 2024; Lozhkov et al., 2024). For output parsing, we adopted both automatic evaluation and LLM-as-a-judge approach, using GPT-4o as the evaluator in accordance with standard practices in the literature (Li et al., 2024b) (see Appendix F).

Environmental Setup The experiments were conducted on a workstation equipped with two GPUs: an NVIDIA A100 (80 GB VRAM) for open models with ≥15B parameters and an NVIDIA RTX 3090 (24 GB VRAM) for models with ≤8B parameters. To ensure high-throughput and memory-efficient inference, we used the vLLM library. Models with 70-72B parameters were executed with AWQ quantization to optimize resource usage and reduce generation time. All other open-source models were run with the precision specified in their respective configuration files. Although DeepSeek-R1 and DeepSeek-V3 are open-source, they were executed via the DeepSeek API due to their high computational requirements. OpenAI models were run using the OpenAI Batch API to optimize costs, while Gemini models were accessed via the Gemini API. Additional details, including source references, are shown in Appendix C.

5 Results

In this section, we analyze model performance on MathGames, as shown in Table 4 and Table 5. Notably, we observed no significant performance differences across years (see Figure 3).

In Figure 4, we present the overall performance of the top-performing closed-source LMMs on the full MathGames benchmark (i.e., both on text-only and vision-based exercises), computed as a weighted average across all problems.

5.1 Text-only Problems

Table 4 presents the results for text-based problems. A clear pattern emerges: reasoning-oriented LLMs, such as o3-mini-high, Gemini-2.0-Flash-Think, and DeepSeek-R1–the largest models in terms of parameters–consistently achieve the highest performance. They substantially outperform their chat-based counterparts, such as GPT-4o,

Model	**CE**	**C1**	**C2**	**L1**	**L2**	**GP**	**HC**	**Avg**
Closed-Source								
o3-mini-high 🧠	<u>83.8</u>	**82.0**	**81.7**	**80.6**	**79.2**	**77.1**	**73.3**	79.7
Gemini-2.0-Flash-T 🧠	81.3	71.4	71.0	69.5	66.4	61.9	59.2	68.7
Gemini-2.0-Flash	58.9	56.8	55.4	54.0	51.3	43.7	41.2	51.6
Gemini-1.5-Pro	59.8	54.3	53.2	52.4	50.2	43.9	41.2	50.7
Gemini-1.5-Flash	60.7	49.7	47.4	45.5	42.9	36.9	36.0	45.6
GPT-4o	61.7	50.1	46.2	43.8	42.3	35.0	33.0	44.6
GPT-4o-mini	49.5	42.2	42.8	41.5	39.8	31.4	30.0	40.4
Gemini-1.5-Flash-8B	40.2	35.1	33.9	31.2	29.3	21.8	20.6	31.6
Open-Source > 8B								
DeepSeek-R1 🧠	**85.0**	<u>77.3</u>	<u>75.9</u>	<u>74.7</u>	<u>72.7</u>	<u>69.7</u>	<u>69.0</u>	74.9
DeepSeek-V3	66.4	54.3	52.1	50.7	48.3	40.8	36.5	49.9
Phi-4-14B *	66.4	51.9	48.1	46.0	43.7	37.0	32.2	46.5
Phi-4-14B	59.8	50.4	45.9	43.6	41.1	33.6	30.0	43.5
Qwen2.5-72B	53.3	48.4	45.2	43.3	41.4	34.4	29.6	42.2
QwQ-32B 🧠	56.1	43.5	40.0	37.3	34.4	25.4	23.2	37.1
LLaMA-3.3-70B	44.9	41.5	39.7	37.3	35.7	26.3	26.2	35.9
DeepSeek-R1-Qwen 🧠	44.2	38.7	38.0	36.3	33.4	25.8	19.8	33.7
Open-Source ≤ 8B (Math-Specialized)								
Qwen2.5-Math-7B * 🔧	53.3	47.9	48.3	47.3	46.9	39.8	34.1	45.4
Qwen2.5-Math-7B 🔧	43.9	45.4	44.1	42.2	41.2	33.6	31.3	40.2
NuminaMath-7B * 🔧	43.0	38.8	38.0	36.7	35.3	26.7	24.5	34.7
Qwen2.5-Math-7B *	40.7	37.0	38.3	36.7	35.6	27.5	26.1	34.6
Qwen2.5-Math-7B	40.2	36.5	37.8	36.2	35.1	26.9	24.9	33.9
NuminaMath-7B 🔧	39.2	27.6	31.1	29.4	28.3	19.8	18.0	27.7
NuminaMath-7B	28.8	25.6	24.9	24.1	23.1	25.4	22.4	24.9
Mathstral-7B *	35.5	27.2	26.1	23.6	21.8	16.7	12.4	23.3
NuminaMath-7B *	31.8	25.2	25.4	24.2	22.7	13.1	9.0	21.6
DeepSeek-Math-7B * 🔧	23.4	24.4	23.7	22.9	21.3	15.1	14.2	20.7
Mathstral-7B	27.1	22.0	23.4	21.3	20.1	12.2	11.2	19.6
DeepSeek-Math-7B *	21.3	21.6	21.9	20.7	19.6	13.2	10.2	18.4
DeepSeek-Math-7B 🔧	21.1	21.4	21.7	20.5	19.3	12.8	9.8	18.1
DeepSeek-Math-7B	20.6	21.0	21.4	20.1	18.9	12.5	9.4	17.7
ToRA-7B * 🔧	12.2	11.6	12.1	11.5	11.1	7.6	6.4	10.4
ToRA-7B 🔧	6.5	11.1	12.4	11.8	11.3	9.3	7.7	10.0

🧠 = Reasoning-focused; * = maj@8 instead of pass@1; 🔧 = TIR mode.

Table 4: **Performance of LLMs on text-only problems categorized by age group.** Results highlight differences in model ability across developmental stages. Best and second-best score are bolded and underlined. Models are ordered based on decreasing Avg score.

Gemini-2.0-Flash, and DeepSeek-V3, with an absolute accuracy increase of +35.1%, +17.1%, and +25.0%, respectively. This underscores the complexity of the exercises and the need for improved reasoning capabilities.

Another notable insight comes from smaller math-specialized LLMs, such as Qwen2.5-Math-7B. When combined with TIR and majority voting, this model surpasses larger closed-source alternatives such as GPT-4o, as well as significantly bigger open LLMs like LLaMA-3.3-70B and Qwen2.5-72B (both quantized). This result is particularly important given the cost-effectiveness of a 7B model, demonstrating that smaller, specialized models can achieve competitive performance with low costs.

5.2 Multimodal Problems

LMMs perform significantly worse on multimodal tasks than text-only ones. We attribute the higher performance of text-only models to the more consistent patterns found in their pretraining and align-

Model	CE	C1	C2	L1	L2	GP	HC	Avg
Closed-Source								
Gemini-2.0-Flash-T 🧠	**38.3**	**29.3**	**32.2**	**31.5**	**31.3**	**25.4**	**25.1**	30.4
Gemini-1.5-Pro	<u>30.4</u>	<u>25.5</u>	<u>24.2</u>	<u>21.3</u>	<u>20.4</u>	<u>18.4</u>	<u>15.3</u>	22.2
Gemini-1.5-Flash	27.0	19.4	16.1	15.0	15.6	12.7	14.2	17.1
GPT-4o	25.2	20.4	17.8	14.8	12.9	10.2	10.9	16.0
GPT-4o-mini	23.5	18.5	16.1	13.4	12.1	10.2	11.5	15.0
Gemini-1.5-Flash-8B	18.3	14.3	12.3	11.3	11.3	9.2	10.9	12.5
Open-Source > 8B								
InternVL-2.5-38B-MPO	19.1	21.0	19.7	17.1	16.4	12.7	12.0	16.9
InternVL-2.5-38B	14.8	14.7	13.6	11.5	9.9	7.4	6.6	11.2
QVQ-72B 🧠	20.0	11.8	8.9	7.5	7.1	6.7	6.6	9.8
Qwen2-VL-72B	14.8	12.5	11.2	8.8	7.5	6.0	3.8	9.2
Pixtral-12B *	12.2	6.4	5.9	5.2	4.8	5.3	4.4	6.3
Pixtral-12B	11.3	8.9	6.1	4.0	2.6	4.2	3.3	5.8
Open-Source ≤ 8B								
Phi-3.5-4.2B *	24.4	11.2	11.4	11.1	10.3	7.4	7.1	11.5
Qwen2-VL-7B *	13.0	11.5	10.8	10.2	9.3	9.2	7.7	10.2
Qwen2-VL-7B	13.9	9.2	10.0	8.8	7.9	4.6	4.9	8.5
InternVL-2.5-8B *	14.8	9.6	5.7	4.8	6.3	5.7	8.2	7.9
InternVL-2.5-8B *	11.3	9.6	7.8	6.3	5.9	3.5	3.3	6.8
Phi-3.5-4.2B	5.2	7.0	6.8	6.5	6.5	7.4	4.9	6.3

🧠 = Reasoning-focused; * = maj@8 instead of pass@1.

Table 5: **Performance of LLMs on multimodal problems categorized by age group.** Results highlight differences in model ability across developmental stages. Best and second-best score are bolded and underlined. Models are ordered based on decreasing Avg score.

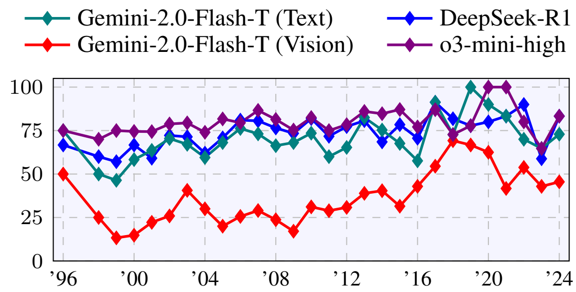


Figure 3: **Overall model accuracy trends over the years.** While text models maintain stable and high performance, vision-based accuracy remains low and more variable, highlighting persistent challenges in visual mathematical reasoning.

ment datasets, facilitating generalization.

The top-performing model, Gemini-2.0-Flash-Think, achieves an overall accuracy of just 30.4%, a striking contrast to its 68.7% accuracy on text-based tasks. This substantial gap highlights the considerable room for improvement in this area. Notably, it is one of the few reasoning-oriented models that supports multimodal inputs, highlighting the advantages of strong reasoning capabilities.

Overall, the performance trend remains similar to that observed in textual problems, with the Gemini family outperforming OpenAI models. Yet, in contrast to previous results, open-source alternatives still lag significantly behind their closed-source counterparts. Even large open-source models, including reasoning-focused ones like QVQ-72B, struggle with these tasks. Smaller models, such as Pixtral-12B and InternVL-2.5-8B, register accuracy scores below 10%, emphasizing the current limitations of open-source LMMs. Moreover, unlike textual tasks, there is a noticeable lack of math-specialized multimodal models. Future research should focus on bridging this gap to enhance performance in math multimodal problems.

5.3 Performance Across Categories

When analyzing performance across different age categories (from CE to HC), we observe a trend that aligns with human behavior. As the difficulty level increases, the performance of both LLMs and LMMs declines. Similarly, models perform worse in advanced categories like HC (professionals, 25+ y/o), highlighting the increasing complexity of these tasks. We conducted a further analysis on the each specific math skill. Figure 5 (top) shows that in text-only tasks, Algebra and Arithmetic lead all models and difficulty levels, with o3-mini and DeepSeek-R1 reaching perfect scores in CE and o3-mini also in HC for Algebra. By contrast, Logic, Geometry, Combinatorics, and Pattern Recognition persistently challenge LLMs: GPT-4o scores below 33% in GP for Arithmetic, Logic and Geometry, Qwen2.5-Math-7B falls below 29% in HC for Combinatorics and Pattern Recognition, and Gemini-2.0-Flash-Think drops from 100% in CE for Algebra and Pattern Recognition to under 51% in HC. Figure 5 (bottom) reveals even steeper declines for multimodal models: Gemini-2.0-Flash-Think attains at best 60% in Algebra and Combinatorics but rarely exceeds 30% in Arithmetic. GPT-4o peaks at 53.85% in CE Geometry before collapsing at higher levels, while Phi-3.5 seldom surpasses 10% in any skill. Across modalities, performance degrades steadily from CE to HC and underscores a persistent gap in abstract, spatial and combinatorial reasoning, especially when visual inputs are involved.

5.4 Human Comparison

We analyze statistics from the 2024 National Finals, considering only LMMs as text-only exercise scores could not be isolated. Results are shown in Figure 1. Despite rapid model advancements, we find that LMMs remain deeply inadequate compared to human participants in our benchmark. Notably, only Gemini-2.0-Flash-Think shows competitive performance, matching top human partici-

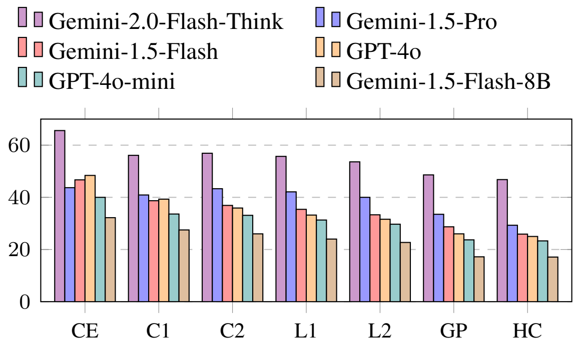


Figure 4: Overall average accuracy of LMMs on both textual and visual problems in MATHGAMES.

pants in the L2 category (18-20 years old). However, even this model falls far behind in all other categories. GPT-4o and most other models fail to even reach the average human score, while Gemini-1.5-Pro is the sole exception, matching the L2 average. Surprisingly, all models perform worse than the average 11-13-year-old, a concerning result that exposes the weaknesses of even the most advanced LMMs in mathematical reasoning and adaptability. While models with stronger CoT abilities, like Gemini-2.0-Flash-Think, show improvements, LMMs still fall short of human-level problem-solving in MATHGAMES and are not yet competitive in international math challenges.

To better understand Gemini-2.0-Flash-Think's relative strength in L2, we examined the 12 exercises in this group, evenly split between textual and multimodal tasks. The *textual problems* cover Combinatorics (3), Algebra (2), and Logic (1), while the *multimodal problems* include Geometry (2) and Algebra (1). These domains align closely with Gemini's strongest reasoning abilities, both textual and visual (see Figure 1, top and bottom), which likely explains the unusually high scores observed in L2. To contextualize this result, we also examined the broader *topic distribution* across multimodal exams in all years. We find that L2 contains a higher proportion of Algebra and Geometry, whereas C1 features more Logic and Combinatorics, areas that are comparatively harder for Gemini. For example, in multimodal tasks L2 includes 178 Geometry and 69 Algebra items versus 122 Combinatorics and 86 Logic, while C1 includes only 55 Geometry and 46 Algebra against 97 Combinatorics and 83 Logic.

Taken together, these findings indicate that Gemini's peak performance in L2 reflects the *exam structure itself*, rather than an artifact of evaluation or systematic model bias.

6 Error Analysis

To conduct error analysis, we randomly selected 25 problems where each top-performing model made incorrect predictions. Since the dataset lacks explanations and reasoning paths for obtaining the correct answers, we manually solved each of these sampled exercises.[3] By comparing our solutions with the models' predictions, we categorized the errors and constructed an error distribution.

For text-only, failures primarily stem from reasoning, comprehension, and calculation errors. For multimodal, the main sources of error are misinterpretation of visual content, flawed image-grounded reasoning, and incorrect text comprehension. The following sections provide an overview of these issues observed in the best-performing models. All examples discussed, along with further details, can be found in Appendix J.

6.1 Errors in Textual Problems

Most errors arise from flawed multi-step reasoning, imperfect comprehension of problem statements, and occasional calculation mistakes. About reasoning, models like GPT-4o, o3-mini and DeepSeek-R1 generally follow plausible inference chains but arrive at incorrect conclusions, whereas Gemini variants often make unpredictable errors: Gemini-2.0-Flash-Think correctly computes an intermediate rounding result before second-guessing itself, and in a date-pattern task it disregards given constraints, selecting an impossible date. Comprehension errors occur when models misinterpret question requirements or ignore explicit constraints. For instance, Gemini-2.0-Flash-Think treats relative changes in a parallelepiped's side lengths as its actual dimensions, and o3-mini identifies the largest rather than the most recent date when asked "What is the last date that owned this property?". Calculation errors, though less common, appear even in larger models: Gemini-1.5-Pro erroneously replaces -19 with $+19$ during equation rewriting, and GPT-4o-mini attempts to divide 2022 by 22, erroneously expecting an integer result.

6.2 Errors in Visual Problems

Multimodal models frequently misinterpret spatial relationships, leading to errors in image content understanding and image-grounded reasoning. For example, Gemini-2.0-Flash-Think reverses layer-

[3]The annotation process was carried out by PhD-level experts with strong mathematical backgrounds.

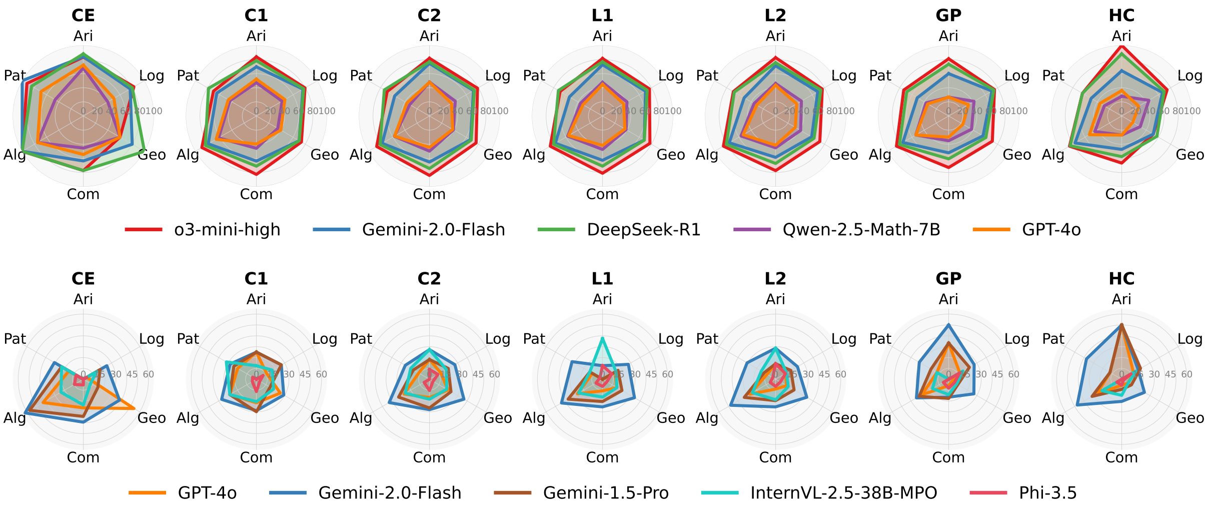


Figure 5: **LLM accuracy across skill categories and age groups.** *Top:* performance on text-only problems. *Bottom:* performance on image-based problems (accuracy scale 0-70 for better visualization). Skills include Arithmetic (Ari), Logic (Log), Geometry (Geo), Combinatorics (Com), Algebra (Alg), and Pattern Recognition (Pat).

ing in overlapping shapes and o3-mini misplaces digits when reconstructing a 3×3 grid from text, while GPT-4o-mini undercounts partially occluded cubes, revealing weaknesses in depth perception and spatial consistency. These failures typically originate in the earliest reasoning steps and propagate through subsequent operations, particularly undermining numerical reasoning when visual and textual information must be integrated. Moreover, models often misread or overconstrain the textual component of a problem accompanying an image–sometimes inventing requirements not present in the prompt–further compounding inaccuracies. Together, these patterns underscore that current LMMs lack reliable strategies for aligning visual cues with textual constraints and maintaining coherent, stepwise reasoning across modalities.

7 Conclusion

We introduce MATHGAMES, a novel benchmark designed to evaluate the readiness of LLMs for international mathematical competitions across various age categories. Our comprehensive evaluation of a diverse set of open-source and closed-source models, coupled with an in-depth error analysis, reveals important insights. Our findings highlight that **current models are not ready to win a mathematical game competition against humans yet**. While advanced reasoning models can handle text-only problems competently, they still face a significant performance gap in multimodal reasoning, highlighting the ongoing challenge of integrating textual and visual information. Through this rigorous evaluation, MATHGAMES aims to foster progress in mathematical reasoning, logic, and multimodal understanding in AI.

An important line of future work concerns extending the dataset to multilingual settings, including the original English versions of the problems. We are currently pursuing authorization to incorporate English and additional languages, which would enable wider applicability and provide a stronger basis for benchmarking multilingual models. Moreover, with the rapid development of advanced reasoning models, it will be equally important to monitor the trade-off between their achieved effectiveness and the associated cost in terms of carbon emissions (Moro and Monti, 2012; Moro et al., 2023b).

Limitations

Despite its contributions, our work has several limitations that warrant further exploration. First, MATHGAMES serves as a benchmark without accompanying training data. This is mainly due to the absence of gold-standard human reasoning annotations, which, while preventing data contamination, also limits the ability to curate high-quality reasoning paths crucial for improving models' mathematical capabilities and foster interpretability (Moro et al., 2024). Future work could focus on constructing such reasoning annotations at scale, possibly through a semi-supervised approach leveraging strong reasoning-centric LLMs. Second, all images in our dataset are abstract, without real-world photographs. While this ensures a controlled evaluation setting, it limits the study of vision-language models in practical mathematical problems.

Ethical Considerations

The MATHGAMES dataset consists of problems from publicly available sources, with prior consent obtained for research use. Its primary goal is to evaluate LLMs' mathematical reasoning and support advancements in AI-driven problem-solving. By providing a structured benchmark, we aim to facilitate future studies and contribute to the broader scientific community. While the dataset is not intended for training models to assist in academic dishonesty, we acknowledge the potential risks associated with its misuse. Legally, all problems originate from PRISTEM, which holds the rights to these materials. For over two decades, they have served as a reference for students preparing for mathematical competitions. Their inclusion in MATHGAMES aligns with this purpose, offering a standardized benchmark for AI research while ensuring compliance with ethical and legal guidelines.

Acknowledgements

Research partially supported by AI-PACT project (CUP B47H22004450008, B47H22004460001); National Plan PNC-I.1 DARE initiative (PNC0000002, CUP B53C22006450001); PNRR Extended Partnership FAIR (PE00000013, Spoke 8); 2024 Scientific Research and High Technology Program, project "AI analysis for risk assessment of empty lymph nodes in endometrial cancer surgery", the Fondazione Cassa di Risparmio in Bologna; Chips JU TRISTAN project (G.A. 101095947).

References

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, et al. 2024. Llemma: An open language model for mathematics. In ICLR 2024. OpenReview.net.

Walter Cerroni, Gianluca Moro, Roberto Pasolini, and Marco Ramilli. 2015. Decentralized detection of network attacks through P2P data clustering of SNMP data. Comput. Secur., 52:1–16.

Walter Cerroni, Gianluca Moro, Tommaso Pirini, and Marco Ramilli. 2013. Peer-to-peer data mining classifiers for decentralized detection of network attacks. In Proceedings of the Twenty-Fourth Australasian Database Conference - Volume 137, ADC '13, page 101–107, AUS. Australian Computer Society, Inc.

Zhe Chen, Jiannan Wu, Wenhai Wang, et al. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. CoRR, abs/2312.14238.

Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, et al. 2024. Evaluating large vision-and-language models on children's mathematical olympiads. In NeurIPS 2024.

Cheng-Han Chiang and Hung-yi Lee. 2024. Over-reasoning and redundant calculation of large language models. In EACL 2024, pages 161–169. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. 2021. Training verifiers to solve math word problems. CoRR, abs/2110.14168.

Alessio Cocchieri, Giacomo Frisoni, Marcos Martínez Galindo, Gianluca Moro, Giuseppe Tagliavini, and Francesco Candoli. 2025a. Openbioner: Lightweight open-domain biomedical named entity recognition through entity type description. In NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025, pages 818–837. Association for Computational Linguistics.

Alessio Cocchieri, Marcos Martínez Galindo, Giacomo Frisoni, Gianluca Moro, Claudio Sartori, and Giuseppe Tagliavini. 2025b. Zeroner: Fueling zero-shot named entity recognition via entity type descriptions. In ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 15594–15616. Association for Computational Linguistics.

Alessio Cocchieri, Luca Ragazzi, Paolo Italiani, Giuseppe Tagliavini, and Gianluca Moro. 2025c. "What do you call a dog that is incontrovertibly true? Dogma": Testing LLM generalization through humor. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 22922–22937. Association for Computational Linguistics.

Giacomo Domeniconi, Gianluca Moro, Andrea Pagliarani, and Roberto Pasolini. 2017. On deep learning in cross-domain sentiment classification. In IC3K, Funchal, Madeira, Portugal, November 1-3, 2017, pages 50–60. SciTePress.

Giacomo Domeniconi, Konstantinos Semertzidis, Vanessa López, Elizabeth M. Daly, Spyros Kotoulas, and Gianluca Moro. 2016. A novel method for unsupervised and supervised conversational message thread detection. In DATA 2016 - Proceedings of 5th International Conference on Data Management Technologies and Applications, Lisbon, Portugal, 24-26 July, 2016, pages 43–54. SciTePress.

Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gianluca Moro, and Zaiqiao Meng. 2024. To generate or to retrieve? on the effectiveness of artificial contexts for medical open-domain question answering. In ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 9878–9919. Association for Computational Linguistics.

Deepanway Ghosal, Vernon Toh Yan Han, Yew Ken Chia, et al. 2024. Are language models puzzle prodigies? algorithmic puzzles unveil serious challenges in multimodal reasoning. CoRR, abs/2403.03864.

Panagiotis Giadikiaroglou, Maria Lymperaiou, Giorgos Filandrianos, et al. 2024. Puzzle solving using reasoning of large language models: A survey. In EMNLP 2024, pages 11574–11591. Association for Computational Linguistics.

Zhibin Gou, Zhihong Shao, Yeyun Gong, et al. 2024. Tora: A tool-integrated reasoning agent for mathematical problem solving. In ICLR 2024. OpenReview.net.

Daya Guo, Dejian Yang, Haowei Zhang, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Chaoqun He, Renjie Luo, Yuzhuo Bai, et al. 2024. Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In ACL 2024, pages 3828–3850. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, et al. 2021. Measuring mathematical problem solving with the MATH dataset. In NeurIPS Datasets and Benchmarks 2021.

Zhen Huang, Zengzhi Wang, Shijie Xia, et al. 2024. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent AI. In NeurIPS 2024.

Paolo Italiani, Gianluca Moro, and Luca Ragazzi. 2025. Enhancing legal question answering with data generation and knowledge distillation from large language models. Artificial Intelligence and Law.

Eldar Kurtic, Amir Moeini, and Dan Alistarh. 2024. Mathador-lm: A dynamic benchmark for mathematical reasoning on large language models. In EMNLP 2024, pages 17020–17027. Association for Computational Linguistics.

Aitor Lewkowycz, Anders Andreassen, David Dohan, et al. 2022. Solving quantitative reasoning problems with language models. In NeurIPS 2022.

Chengpeng Li, Zheng Yuan, Hongyi Yuan, et al. 2024a. Mugglemath: Assessing the impact of query and response augmentation on math reasoning. In ACL 2024, pages 10230–10258. Association for Computational Linguistics.

Jia Li, Edward Beeching, Lewis Tunstall, et al. 2024b. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. Hugging Face repository, 13:9.

Qintong Li, Leyang Cui, Xueliang Zhao, et al. 2024c. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers. In ACL 2024, pages 2961–2984. Association for Computational Linguistics.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, et al. 2024a. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. In ACL 2024, pages 6884–6915. Association for Computational Linguistics.

Yan Liu, Renren Jin, Lin Shi, et al. 2024b. Finemath: A fine-grained mathematical evaluation benchmark for chinese large language models. CoRR, abs/2403.07747.

Anton Lozhkov, Raymond Li, and Loubna Ben Allal. 2024. Starcoder 2 and the stack v2: The next generation. CoRR, abs/2402.19173.

Pan Lu, Hritik Bansal, Tony Xia, et al. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In ICLR 2024. OpenReview.net.

Haipeng Luo, Qingfeng Sun, Can Xu, et al. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. CoRR, abs/2308.09583.

Yujun Mao, Yoon Kim, and Yilun Zhou. 2024. CHAMP: A competition-level dataset for fine-grained analyses of llms' mathematical reasoning capabilities. In ACL 2024, pages 13256–13274. Association for Computational Linguistics.

Gianluca Moro and Gabriele Monti. 2012. W-grid: A scalable and efficient self-organizing infrastructure for multi-dimensional data management, querying and routing in wireless data-centric sensor networks. J. Netw. Comput. Appl., 35(4):1218–1234.

Gianluca Moro, Andrea Pagliarani, Roberto Pasolini, and Claudio Sartori. 2018. Cross-domain & in-domain sentiment analysis with memory-based deep neural networks. In IC3K, Seville, Spain, September 18-20, 2018, pages 125–136. SciTePress.

Gianluca Moro, Nicola Piscaglia, Luca Ragazzi, et al. 2023a. Multi-language transfer learning for low-resource legal case summarization. Artificial Intelligence and Law, pages 1–29.

Gianluca Moro and Luca Ragazzi. 2022. Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, pages 11085–11093. AAAI Press.

Gianluca Moro and Luca Ragazzi. 2023. Align-then-abstract representation learning for low-resource summarization. Neurocomputing, 548:126356.

Gianluca Moro, Luca Ragazzi, and Lorenzo Valgimigli. 2023b. Carburacy: Summarization models tuning and comparison in eco-sustainable regimes with a novel carbon-aware accuracy. In Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Washington, DC, USA, February 7-14, 2023, pages 14417–14425. AAAI Press.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, Fabian Vincenzi, and Davide Freddi. 2024. Revelio: Interpretable long-form question answering. In The Second Tiny Papers Track at ICLR 2024, Tiny Papers @ ICLR 2024, Vienna, Austria, May 11, 2024. OpenReview.net.

OpenAI. 2023. GPT-4 technical report. CoRR, abs/2303.08774.

Luca Ragazzi, Paolo Italiani, Gianluca Moro, and Mattia Panni. 2024. What are you token about? differentiable perturbed top-k token selection for scientific document summarization. In Findings of the Association for Computational Linguistics: ACL 2024, pages 9427–9440, Bangkok, Thailand. Association for Computational Linguistics.

A M. Muntasir Rahman, Junyi Ye, Wei Yao, et al. 2024. From blind solvers to logical thinkers: Benchmarking llms' logical integrity on faulty mathematical problems. CoRR, abs/2410.18921.

Machel Reid, Nikolay Savinov, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. CoRR, abs/2403.05530.

Baptiste Rozière, Jonas Gehring, and Fabian Gloeckle. 2023. Code llama: Open foundation models for code. CoRR, abs/2308.12950.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. CoRR, abs/2402.03300.

Kai Sun, Yushi Bai, Ji Qi, et al. 2024. MM-MATH: advancing multimodal math evaluation with process evaluation and fine-grained classification. In EMNLP 2024, pages 1358–1375. Association for Computational Linguistics.

Yuxuan Tong, Xiwen Zhang, Rui Wang, et al. 2024. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. In NeurIPS 2024.

Ke Wang, Junting Pan, Weikang Shi, et al. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. In NeurIPS 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, et al. 2023. Self-consistency improves chain of thought reasoning in language models. In ICLR. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In NeurIPS 2022.

Yanan Wu, Jie Liu, Xingyuan Bu, et al. 2024. Conceptmath: A bilingual concept-wise benchmark for measuring mathematical reasoning of large language models. In ACL 2024, pages 6815–6839. Association for Computational Linguistics.

An Yang, Beichen Zhang, Binyuan Hui, et al. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. CoRR, abs/2409.12122.

Shukang Yin, Chaoyou Fu, Sirui Zhao, et al. 2023. A survey on multimodal large language models. CoRR, abs/2306.13549.

Longhui Yu, Weisen Jiang, Han Shi, et al. 2024. Metamath: Bootstrap your own mathematical questions for large language models. In ICLR 2024. OpenReview.net.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, et al. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In IEEE/CVF 2024, pages 9556–9567. IEEE.

Jiaxin Zhang, Zhongzhi Li, Ming-Liang Zhang, et al. 2024a. Geoeval: Benchmark for evaluating llms and multi-modal models on geometry problem-solving. In ACL 2024, pages 1258–1276. Association for Computational Linguistics.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, et al. 2024b. MATHVERSE: does your multi-modal LLM truly see the diagrams in visual math problems? In ECCV 2024, volume 15066 of Lecture Notes in Computer Science, pages 169–186. Springer.

Wayne Xin Zhao, Kun Zhou, Junyi Li, et al. 2023. A survey of large language models. CoRR, abs/2303.18223.

A Competition Structure and Categories

Categories The *International Mathematical and Logical Games Championships* are structured to accommodate a wide range of participants through a categorization system based on educational level. Each participant competes within one of the following seven categories, each tailored to match the expected mathematical background and cognitive development of the group:

- **CE:** for pupils in the 4th and 5th grades of primary school (typically ages 9-10).
- **C1:** for students in the 6th and 7th grades (corresponding to the first and second years of lower secondary school).
- **C2:** for students in the 8th grade and the 9th grade (third year of lower secondary and first year of upper secondary school).

Event	Count	Categories
Autumn Games	478	CE, C1, C2, L1, L2, HC
Team Games	357	C2, L1, L2
Rosi's Games	275	C2, L1, L2
Quarterfinals	16	C1, C2, L1, L2
Semifinals	314	C1, C2, L1, L2, GP
Final	1073	CE, C1, C2, L1, L2, GP, HC
International Final	421	CE, C1, C2, L1, L2, GP, HC

Table 6: Distribution of competition phases in MATHGAMES and associated category levels.

- **L1:** for students in the 10th, 11th, and 12th grades (second through fourth year of upper secondary school).
- **L2:** for students in the final year of upper secondary school (13th grade) and those enrolled in the first two years of university.
- **GP (Grand Public):** for adult participants, including those in their third year of university and beyond, up to 99 years of age.
- **HC (High Competition):** for adults–post-university and beyond–who placed first, second, or third in the national finals of the L2 or GP categories in any of the past ten editions.

Phases The championship is composed of four competitive phases, each progressively narrowing the pool of participants based on performance: Quarterfinals, Semifinals, National Final, International Final. In addition to the official competition stages, a number of preliminary activities are organized to foster engagement and preparation. These include training events such as the *Autumn Games, the Rosi's Games* and *Team Games*, which offer an opportunity for students and educators to familiarize themselves with the types of problems and reasoning skills featured in the main competition. These preliminary rounds serve not only as practice, but also as an inclusive entry point for participants of all skill levels. For an overview of the number of occurrences of each competition phase in MATHGAMES, along with the associated categories identified, see Table 6.

B Data Curation

To ensure the quality of our MATHGAMES benchmark, we manually review and revise each problem-solution pair. Figure 6 and Figure 7 show different examples of inconsistencies.

Original problem:
Now the triangle is arbitrary, and the measures of its three sides are expressed (in meters) by consecutive integers. How much is the area of the triangle in m^2, knowing that it is equal to 2/5 of the product of the lengths of its two longest sides?

Answer provided:
The area is $\frac{2}{5} \times 14 \times 15 = 84\,\text{cm}^2$.

Rewritten problem:
Now, consider a triangle with side lengths given by three consecutive integers (in centimeters). How much is the area of the triangle in cm^2, knowing that it is equal to 2/5 of the product between the measurements of its two longest sides?

Figure 6: **Example illustrating inconsistencies between the original problem and the provided solution.** Modifications, highlighted in red, were made to align the problem with the green-highlighted corrected answer.

Original problem (Italian):
Completate la frase tra virgolette con un numero scritto in lettere, in modo che la frase risulti vera : «In questa frase, potete contare _____ lettere "e" ».
Solution: dieci *(ten)*

Rewritten problem (English):
Fill in the blank with a number written in words so that the statement remains true: "In this sentence, you can count _____ letters 'e'."
Solution: six

Figure 7: This example demonstrates how a word puzzle involving counting the letter "e" requires manual adjustment when translated from Italian to English due to differences in letter frequency.

C Models

We evaluated 28 distinct LLMs, encompassing both vision-enabled and text-only variants. Given the focus of our benchmark, we prioritized models with advanced reasoning capabilities and those specifically fine-tuned for mathematical problem-solving. Table 7 summarizes all information related to the source of the models tested in our experiments.

OpenAI We evaluated the latest available versions of OpenAI models at the time of writing. Specifically, we included **OpenAI o3-mini**, released on January 31, 2025,[4] the most recent and cost-efficient model in OpenAI's reasoning series. Additionally, we evaluated **GPT-4o**, an advanced

[4] https://openai.com/index/openai-o3-mini/

Model	Source	URL
GPT-4o	gpt-4o-2024-08-06	https://platform.openai.com/
GPT-4o-mini	gpt-4o-mini-2024-07-18	https://platform.openai.com/
o3-mini-high	o3-mini-2025-01-31	https://platform.openai.com/
Gemini-1.5-Flash-8B	gemini-1.5-flash-8b-001	https://ai.google.dev/
Gemini-1.5-Flash	gemini-1.5-flash-002	https://ai.google.dev/
Gemini-1.5 Pro	gemini-1.5-pro-002	https://ai.google.dev/
Gemini-2.0-Flash	gemini-2.0-flash-001	https://ai.google.dev/
Gemini-2.0-Flash-Thinking	gemini-2.0-flash-thinking-exp-01-21	https://ai.google.dev/
DeepSeek-V3	deepseek-chat	https://api.deepseek.com
DeepSeek-R1	deepseek-reasoner	https://api.deepseek.com
DeepSeek-Math-7B	local checkpoint	https://huggingface.co/deepseek-ai/deepseek-math-7b-instruct
DeepSeek-R1-Distill-Qwen-32B	local checkpoint	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B
Phi-4	local checkpoint	https://huggingface.co/microsoft/phi-4
Phi-3.5-Vision	local checkpoint	https://huggingface.co/microsoft/Phi-3.5-vision-instruct
Qwen2.5-Math-7B	local checkpoint	https://huggingface.co/Qwen/Qwen2.5-Math-7B-Instruct
Qwen2.5-72B	local checkpoint	https://huggingface.co/Qwen/Qwen2-72B-Instruct
Qwen2-VL-7B	local checkpoint	https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct
Qwen2-VL-72B	local checkpoint	https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct-AWQ
QVQ-72B	local checkpoint	https://huggingface.co/kosbu/QVQ-72B-Preview-AWQ
QwQ-32B	local checkpoint	https://huggingface.co/Qwen/QwQ-32B-Preview
Llama-3.3-70B	local checkpoint	https://huggingface.co/casperhansen/llama-3.3-70b-instruct-awq
Mathstral-7B	local checkpoint	https://huggingface.co/mistralai/Mathstral-7B-v0.1
NuminaMath-7B-TIR	local checkpoint	https://huggingface.co/AI-MO/NuminaMath-7B-TIR
NuminaMath-7B-CoT	local checkpoint	https://huggingface.co/AI-MO/NuminaMath-7B-CoT
ToRA-7B	local checkpoint	https://huggingface.co/llm-agents/tora-7b-v1.0
Pixtral-12B	local checkpoint	https://huggingface.co/mistral-community/pixtral-12b
IntenVL-2.5-8B	local checkpoint	https://huggingface.co/OpenGVLab/InternVL2_5-8B
IntenVL-2.5-38B	local checkpoint	https://huggingface.co/OpenGVLab/InternVL2_5-38B
IntenVL-2.5-8B-MPO	local checkpoint	https://huggingface.co/OpenGVLab/InternVL2_5-8B-MPO
IntenVL-2.5-38B-MPO	local checkpoint	https://huggingface.co/OpenGVLab/InternVL2_5-38B-MPO

Table 7: The source of the models used in our evaluation.

multimodal LLM that delivers faster performance and lower costs compared to GPT-4-Turbo. We also included **GPT-4o-mini**, a highly cost-efficient model that surpasses GPT-3.5-Turbo.

DeepSeek The DeepSeek family consists of open-source LLMs excelling in both Chinese and English tasks. For our work, we used multiple models from this series. We first considered **DeepSeek-V3**, a 685B parameter Mixture-of-Experts (MoE) model with 37B active parameters per token. It employs an auxiliary-loss-free load-balancing strategy and a multi-token prediction objective, enhancing overall performance. Next, we evaluated **DeepSeek-R1**, which builds on V3 but specializes in logical reasoning and problem-solving. It follows a two-phase training strategy, integrating cold-start reinforcement learning (RL) with supervised fine-tuning, and applies Group Relative Policy Optimization (GRPO) for improved reasoning. Additionally, we used **DeepSeek-Math-7B-Instruct**, a math-specialized model leveraged in both CoT and Tool-Integrated Reasoning (TIR) modes. Finally, we included **DeepSeek-R1-Distill-Qwen-32B**, a dense models distilled from DeepSeek-R1, designed for cost-effective reasoning tasks.

Google Gemini Gemini is Google's flagship language model family, first released in December 2023. From the Gemini-1.5 series, we evaluated multimodal models capable of reasoning over fine-grained information from millions of tokens, including long documents and hours of video and audio. Specifically, we used **Gemini-1.5-Pro**, the strongest model, and **Gemini-1.5-Flash**, a lightweight variant optimized for efficiency with minimal quality trade-offs, both launched in December 2024. Additionally, we explored models from the latest Gemini-2.0 family, including **Gemini-2.0-Flash** and **Gemini-2.0-Flash-Thinking**, the latter being the most advanced in this lineup, offering exceptional capabilities for complex reasoning and multimodal tasks.

Qwen Qwen refers to the LLM family built by Alibaba Cloud, first introduced in September 2023. For text-only problems, we evaluated **Qwen2.5-72B** and **Qwen2.5-Math-7B-Instruct**, both supporting Chinese and English while exhibiting strong mathematical reasoning capabilities. The latter incorporates TIR, beyond CoT, for enhanced problem-solving. For multimodal tasks, we explored the **Qwen2-VL** series, which introduces the Naive Dynamic Resolution mechanism, enabling flexible visual processing by dynamically converting images into varying numbers of visual tokens. Specifically, we evaluated **Qwen2-VL-72B** and **Qwen2-VL-7B**. Further, we tested **QwQ-32B** for textual tasks and **QVQ-72B** for vision-related tasks–early Qwen experiments in reasoning-focused models with strong CoT capabilities.

InternVL The InternVL-2.5 family is an advanced series of LMMs that builds on the InternVL-2.0 architecture, introducing significant improvements in training, evaluation strategies, and data quality. InternVL-2.5 performs competitively alongside leading models like GPT-4o. In our study, we evaluated **InternVL-2.5-8B** and **InternVL-2.5-38B**, along with the Mixed Preference Optimization (MPO) variants **InternVL-2.5-MPO-8B** and **InternVL-2.5-38B-MPO**, which enhance performance in multimodal CoT reasoning tasks. We omitted results from **InternVL-2.5-8B-MPO** as no improvements were observed over the non-MPO version, thus avoiding redundancy.

LLaMA-3 The LLaMA-3 family debuted on April 18, 2024, followed by the release of LLaMA-3.1 on July 23, 2024. LLaMA-3 models represent a huge advancement over LLaMA-2, setting a new state-of-the-art for LLMs at these scales due to improvements in both pre- and post-training. In our study, we evaluated **LLaMA-3.3-70B-instruct**, the leading model in the series, which delivers performance comparable to the larger LLaMA-3.1-405B Instruct, while enhancing efficiency.

Mathstral & Pixtral The **Mathstral-7B** model, developed by Mistral,[5] is designed to address advanced mathematical problems requiring complex, multi-step reasoning. It achieves state-of-the-art performance in its size category, excelling on industry-standard benchmarks, and is used in our text-based problems. For image-grounded problems, we evaluated **Pixtral-12B**, Mistral's first multimodal model. Designed to excel in both textual and visual tasks, it features a custom vision encoder that processes images at their natural resolution and aspect ratio, supporting up to 128K tokens and multiple images. Pixtral-12B outperforms similarly sized models like LLaMA-3.2-11B and Qwen-2-VL-7B, and even surpasses much larger models like LLaMA-3.2-90B, despite being 7x smaller.

Phi-3.5 & 4 The Phi-3 family models are among the most capable and cost-effective small language models available today. The **Phi-3.5-Vision** model, with 4.2B parameters and derived from Phi-3.5-mini, excels in reasoning tasks, handling both single- and multi-image, as well as text prompts. We also tested **Phi-4**, released on December 14, 2024, a 14B-parameter model developed with a data-quality-focused training approach. Phi-4 underwent rigorous enhancement and alignment, combining supervised fine-tuning with preference optimization to ensure precise instruction adherence and safety. It significantly outperforms its predecessor in STEM-focused question-answering tasks.

NuminaMath NuminaMath is a series of language models specifically trained to solve math problems using CoT and TIR. The **NuminaMath-7B TIR** model won the first progress prize at the AI Math Olympiad (AIMO), scoring 29/50 on both the public and private test sets. It is trained on the largest math dataset ever released in the field. In this study, we evaluate both versions of the model for our text-only problems.

ToRA ToRA is a series of models specifically designed for solving challenging mathematical problems. It was the first to introduce the TIR paradigm, through interaction with tools such as computation libraries and symbolic solvers. We tested both **ToRA-7B** and **ToRA-Code-7B**, but ultimately report only the performance of ToRA-7B, as both models performed similarly. ToRA-7B is fine-tuned from LLaMA-2 using the ToRA-Corpus 16K, which includes TIR trajectories from MATH and GSM8K, generated by GPT-4. After supervised fine-tuning, the model undergoes *output space shaping* to enhance its TIR capabilities.

D Translation Quality

To ensure the reliability and usability of the LLM-generated translations, each problem was independently reviewed by a fluent Italian-English bilingual speaker with a PhD-level background. Reviewers evaluated translation quality across four key dimensions: *semantic fidelity* (faithfulness to the original meaning and logical structure), *mathematical clarity* (accuracy and readability of mathematical content), *terminological consistency* (uniform and appropriate use of domain-specific terms), and *linguistic fluency* (naturalness and grammatical correctness in English). Each translation was then assigned a holistic quality score ranging from 1 (Poor) to 3 (Excellent), following the rubric in

[5] https://mistral.ai/en/news/mathstral

Translation Quality Assessment Scale

Evaluate translations based on the following dimensions:

- **Semantic Fidelity:** Does the translation preserve the intended meaning and logic of the original problem?
- **Mathematical Clarity:** Are formulas, operations, and problem constraints clearly and accurately conveyed?
- **Consistency of Terminology:** Are domain-specific terms (e.g., geometric figures, logical relations) translated uniformly and appropriately?
- **Linguistic Fluency:** Does the English text read naturally, with correct grammar, syntax, and idiomatic phrasing?

Assign a single overall score from 1 to 3 based on the criteria above:

1 (Poor): Translation contains significant errors that compromise meaning, mathematical clarity, or specialized terminology. Mathematical concepts may be incorrectly rendered, and/or the text reads as awkwardly translated with unnatural phrasing and structural problems.

2 (Acceptable): Translation adequately preserves the original meaning with only minor semantic shifts. Mathematical notation and concepts are generally accurate. Terminology is mostly consistent with occasional minor lapses. The text is comprehensible but may contain phrasing that reveals its translated nature.

3 (Excellent): Translation demonstrates exceptional fidelity to source content while achieving natural expression in the target language. Mathematical concepts are rendered with perfect accuracy and clarity. Terminology is consistently appropriate throughout, and the text reads as if originally written in the target language.

Evaluation Process:

- Reference the original text when encountering ambiguities or domain-specific phrasing.
- Consider all aspects defined above holistically
- Apply a conservative judgment approach that prioritizes preservation of technical meaning.
- Provide corrections to the translation **only** when assigning a score of 1 (Poor).

Figure 8: Human evaluation rubric used to assess the quality of LLM-generated English translations from Italian problem statements.

Figure 8. In cases of ambiguity or domain-specific phrasing, reviewers consulted the original Italian version to preserve the intended structure and meaning. Edits were applied conservatively–only when necessary to correct substantial errors in translations rated as Poor.

E Tool-Integrated Reasoning (TIR)

We follow a structured approach to generate and assess code solutions for TIR with self-consistency:

1. For each problem, the input is duplicated N times to define the initial batch of prompts for `vLLM`, effectively determining the number of candidates used for majority voting.
2. The model samples N diverse completions until it produces a complete Python block.
3. Each generated Python block is executed, and its output–along with tracebacks–is captured.
4. This process is repeated M times, producing a batch of generations of size N and depth M. This iterative approach enables the model to self-correct code errors using traceback information. If a sample fails to yield a valid output (e.g., incomplete code blocks), it is pruned.
5. The remaining solution candidates undergo postprocessing, followed by majority voting to determine the final answer.

For our experiments, we set $N = 8$ candidates with a depth of $M = 4$ to compute @8, similarly to recent works such as ToRA (Gou et al., 2024) and NuminaMath.[6] Instead, for $pass@1$ calculation, each batch contains different input problems, generating a single solution path per problem using greedy decoding.

F Output Parsing

To ensure consistent parsing of model-generated answers, we instructed all tested models to provide their final response within \boxed{ } (see Appendix H), following standard practice for math-specialized LLMs. We then processed outputs as follows: for expected numeric solutions, we first verify whether the model's output can be converted into a numerical value. If conversion is successful, we apply an exact match criterion. Otherwise, the answer is evaluated by GPT-4o, which determines its correctness. On average, only 20% of the total answers require GPT-4o's judgment. The prompt used for this evaluation is shown in Figure 9. To assess the reliability of this automatic evaluation and enable reproducible future assessments without human supervision, we manually reviewed the GPT-4o judgments of the best five models, finding a misjudgment rate of only 1 or 2 cases per model among the processed answers. We corrected these cases to ensure accuracy in the reported results of the main paper. However, even without human intervention in edge cases, GPT-4o remains a highly reliable evaluator. The minor errors would have a negligible impact on final accuracy and do not compromise result validity, establishing a reliable standard for future evaluations.

G Image Extraction Pipeline

Our code relies on the modern `.images` interface available in recent versions of `PyPDF`, which abstracts away the manual handling of content streams and provides a high-level, reliable way to extract images.

Listing 1: Example of the `.images` interface in PyPDF.

```python
from pypdf import PdfReader

reader = PdfReader("example.pdf")
page = reader.pages[0]
for image_file_object in page.images:
    with open(image_file_object.name, "wb") as f:
        f.write(image_file_object.data)
```

This method:

- Automatically detects embedded images without manual stream parsing;
- Extracts the original binary data without re-encoding, preserving native formats (JPEG, PNG, etc.);
- Avoids quality loss during extraction;
- Handles compression and decoding internally, minimizing the risk of artifacts.

To validate this approach, we manually inspected a subset of extracted images and confirmed they matched the quality of the embedded originals. We also explored OCR-based alternatives such as the `MathPix` API, following practices adopted in works like *NuminaMath*. Applied on a subset of problems, this approach did not yield improvements over our PyPDF-based pipeline. Finally, in a few cases, older PDFs contain lower-resolution images. Here, the reduced quality is inherent to the source files, not introduced by the extraction method. We argue that retaining these examples remains valuable, as it reflects real-world conditions in which LLMs must reason over imperfect or noisy data.

System:
Given the primary question, compare the gold answer with the student's final answer to determine if they are equivalent. First, provide a concise rational, without trying to redo the problem, then respond with 'yes' or 'no' in the exact format below:

Rationale: [your rationale]
Answer: [yes/no]

User:
Question: `{problem_question}`
Gold answer: `{gold}`
Final answer: `{final_answer}`

Figure 9: Prompt used to guide GPT-4o in evaluating the equivalence between the generated answer and the reference solution.

H Model Prompts

All the instructions used to guide the models in our experiments are provided in Table 11 for text-only problems and in Table 12 for multimodal inputs.

I Example Data

Examples of text-only problems within MATH-GAMES can be found in Table 8 and Table 10. The first table presents sample questions for different

[6] https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf

Child (9-11 y/o)	**Teenager** (11-18 y/o)	**Adult** (18-25+ y/o)
Question: Desiderio and Liliana are very good at peeling potatoes. Today, they have to peel 2,400 kg. Desiderio, if he worked alone, would take 30 minutes. Liliana is faster and, alone, it would take 20 minutes. How many minutes do they take working together?	**Question:** A clock emits a beep every 10 hours. Now it's exactly 10:00 a.m. and he emits a beep. How many hours must go to a minimum for the clock to issue a beep again at 10 a.m. on a next day (in the morning or in the evening)?	**Question:** Now, consider a triangle with side lengths given by three consecutive integers (in centimeters). How much is the area of the triangle in cm^2, knowing that it is equal to 2/5 of the product between the measurements of its two longest sides?
Answer: 12 minutes	**Answer:** 60 hours	**Answer:** 84 cm^2
Year: 2018	**Year:** 2016	**Year:** 2017
Category: CE	**Category:** C2, L1	**Category:** L2, GP, HC
Question: To each letter of the alphabet Nathan wants to associate a number according to the rule that begins like this: A=1 B=A+2 C=B+3 D=C+4 What number will it associate with the letter G?	**Question:** A snail fell at the bottom of a well of 24 meters deep and now wants to climb up. However, it takes an hour to climb up 3 meters; then, tired, rests and falls asleep for an hour, but in this way it descends by 2 meters. How many hours will it take to return to the surface?	**Question:** Nando have fun adding the whole numbers: 0+1+2+3+4+... The calculator however, starting from the second number beaten by Nando, and before he beats the next, displays the provisional sum: 0, 1, 3, 6, 10, 15, ... At this point Nando observes that some provisional results can split into two numbers with the first double of the second, as happens in the case of 21, 105, 2211, 9045 etc. What is the greater of the numbers of six digits that enjoy this property? (The split into two numbers cannot generate a second number that starts with 0).
Answer: G=28	**Answer:** 43	**Answer:** 890445
Year: 2023	**Year:** 2023	**Year:** 2024
Category: CE	**Category:** C1, C2, L1	**Category:** GP, HC

Table 8: **Example of text-only problems within MathGames divided by age groups.** Each problem includes its correct answer, competition year, and corresponding category.

age groups, while the second illustrates how a question from a specific category (e.g., C1) evolves based on the competition's difficulty level. Similarly, Table 9 demonstrates this progression for an image-grounded question.

J Example Errors

We provided several examples of common errors made by the tested models during our experiments. Table 13, Table 14, and Table 15 present typical mistakes in text-only problems made by Gemini models, OpenAI models, and DeepSeek models, respectively. On the other hand, Table 16, Table 17, and Table 18 present common errors in multimodal problems made by Gemini models, OpenAI models, and open-source models, respectively.

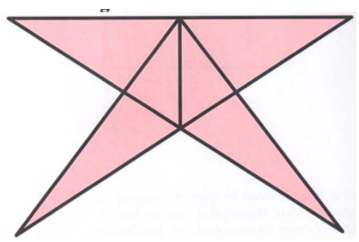
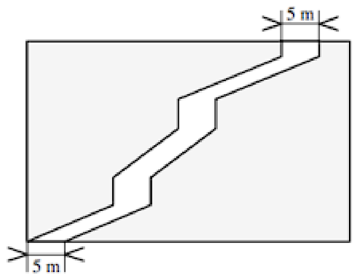
Preliminary	**Semifinals**	**Finals**
		
Question: How many triangles can you see in the picture?	**Question**: "I have to cut the grass of the garden and are 1100 m2! There is no way I'll do it!", yes Renato complains with his boss. "Don't be cunning: on the path (marked in white in the figure) there is no grass – the boss replies – and the square meters of the garden, in to cut the grass, are therefore only 975". What are the dimensions of the garden?	**Question**: During the train stop I observe, one meter from my eyes, a digital clock, which indicates the hours and the minutes. There are four digits, two for the hours (from 00 to 23), two for the minutes (from 00 to 59). They are formed by illuminating appropriate segments between the seven that form the two squares superimposed, according to the diagrams alongside, which indicate the ten digits from 0 to 9. I enjoy counting the segments that change state (turn on or off) at each change of minute. For the set of four digits, counted successively, during the first four minute changes:4 segments changing state; 1 segment that changes state; 11 segments that change by state; 4 segments that change state. What time will the clock show after the fifth change of minute?
Answer: The triangles are 15	**Answer**: The size of the garden is 25m and 44m	**Answer**: 2 solutions : 00 h 02 min and 10 h 02 min
Year: 2016	**Year**: 2013	**Year**: 2005
Category: C2 L1 L2	**Category**: C2 L1 L2	**Category**: C2 L1 L2

Table 9: **Illustration of stage progression for math problems within the same category for multimodal problems.** Each problem includes its correct answer, competition year, and corresponding category.

Preliminary	**Semifinals**	**Finals**
Question: Find a positive integer number n so that 2n+3 is a divider of 6n+43.	**Question:** In the lottery organized for the beginning of the school year, the major prize was won by the ticket holder with the number 205 . Even the holders of numbers 025, 052, 250, 502, 520 won a nice prize, but a little smaller. Finally, a consolation prize was won by the holders of a ticket on which there was a number that contained two of the three digits 2-0-5 of the number 205 (for example, 002 and 570). How many numbers gave entitlement to this consolation prize?	**Question:** In the convent of the City of Mathematics, if you meet two nuns, taken at random in the whole of the nuns who stay there, you have exactly one chance in two that they are both brown. How many nuns stay in the convent?
Answer: n = 7	**Answer:** 144 numbers	**Answer:** 4 and 21
Year: 2004	**Year:** 2021	**Year:** 2005
Category: C2 L1 L2	**Category**: C2 L1 L2	**Category**: C2 L1 L2

Table 10: **Illustration of stage progression for math problems within the same category for text-only problems.** Each problem includes its correct answer, competition year, and corresponding category.

Model	System Prompt / Instruction
Qwen2.5-Math-7B-Instruct (CoT)	`Please reason step by step, and put your final answer within \boxed{}.`
Qwen2.5-Math-7B-Instruct (TIR)	`Please integrate natural language reasoning with programs to solve the problem above, and put your final answer within \boxed{}.`
Qwen2.5-72B-Instruct (CoT)	`You are a helpful and harmless assistant. You should think step-by-step. Put your final answer within \boxed{}.`
Mathstral-7B (CoT)	`Please reason step by step, and put your final answer within \boxed{}.`
DeepSeek-Math-7B-Instruct (CoT)	`Please reason step by step, and put your final answer within \boxed{}.`
DeepSeek-Math-7B-Instruct (TIR)	`You are an expert programmer. Solve the above mathematical problem by writing a Python code. Express your answer as a numeric type or a SymPy object.`
DeepSeek-R1 (CoT)	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
DeepSeek-V3 (CoT)	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
DeepSeek-R1-Distill-Qwen-32B (CoT)	`Please reason step by step, and put your final answer within \boxed{}.`
QwQ-32B (CoT)	`You are a helpful and harmless assistant. You should think step-by-step. Put your final answer within \boxed{}.`
LLaMA-3.3-70B (CoT)	`You are a mathematical expert. Solve the given problem by reasoning step by step. Please, for the validity of the answer, enclose your final answer within \boxed{}.`
Phi-4 (CoT)	`You are a helpful and harmless assistant. You should think step-by-step. Put your final answer within \boxed{}.`
GPT-4o (CoT)	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
GPT-4o-mini (CoT)	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
OpenAI-o3-mini (CoT)	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
Gemini-1.5-Flash (CoT)	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
Gemini-1.5-Pro (CoT)	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
Gemini-1.5-Flash-8B (CoT)	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
Gemini-2.0-Flash (CoT)	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
Gemini-2.0-Flash-Thinking (CoT)	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
NuminaMath-7B-CoT	`NaN`
NuminaMath-7B-TIR	`NaN`
ToRA-7B (TIR)	`NaN`

Table 11: **Prompts used for text-based mathematical reasoning.** Instructions were incorporated as system prompts when supported by the tokenizer's chat template. When system prompts were not available, instructions were provided as user instructions. "NaN" indicates models that were specifically fine-tuned to receive the mathematical problem directly as user input, without additional instructions.

Model	System Prompt/ Instruction
Gemini-2.0-Flash-Thinking	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
Gemini-1.5-Pro	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
Gemini-1.5-Flash	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
Gemini-1.5-Flash-8B	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
GPT-4o	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
GPT-4o-mini	`You are a mathematical expert. Solve the user's problem by reasoning step by step, and enclose the final answer in \boxed{}.`
InternVL-2.5-38B-MPO	`You are a mathematical expert. Solve the given problem by reasoning step by step. Please, for the validity of the answer, enclose your final answer within \boxed{}.`
InternVL-2.5-38B	`You are a mathematical expert. Solve the given problem by reasoning step by step. Please, for the validity of the answer, enclose your final answer within \boxed{}.`
InternVL-2.5-8B	`You are a mathematical expert. Solve the given problem by reasoning step by step. Please, for the validity of the answer, enclose your final answer within \boxed{}.`
Phi-3.5	`You are a mathematical expert. Solve the given problem by reasoning step by step. Please, for the validity of the answer, enclose your final answer within \boxed{}.`
Qwen2-VL-7B	`You are a mathematical expert. Solve the given problem by reasoning step by step. Please, for the validity of the answer, enclose your final answer within \boxed{}.`
Qwen2-VL-72B	`You are a mathematical expert. Solve the given problem by reasoning step by step. Please, for the validity of the answer, enclose your final answer within \boxed{}.`
QVQ-72B	`You are a helpful and harmless assistant. Solve the given problem by thinking step-by-step. Please, ensure to enclose your final answer within \boxed{}.`
Pixtral-12B	`You are a mathematical expert. Solve the given problem by reasoning step by step. Put your final answer within \boxed{}.`

Table 12: **Prompts used for multimodal mathematical reasoning.** Instructions were incorporated as system prompts when supported by the tokenizer's chat template. When system prompts were not available, instructions were provided as user instructions. The image placeholder is then prepended or appended to the input according to the specific chat template considered.

Model	Error type	Frequency	Question example	Gold	Answer	Reasoning	Comment
GEMINI-2.0-FLASH-THINKER	Reasoning	Almost always	Carla's calculator rounds up the results that it obtains, writing only the first digit after the comma. If, for example, the result of a calculation is 34,143, the calculator writes the number 34,1 Carla sets the number 73,5 and then asks the calculator to perform three successive divisions by 2, What is the last result that the calculator will give?	9,1	9,2	Let's consider the possibility that "rounds up" is a typo and it should be "rounds down" or "truncates". If we use truncation, we get 9.1. [...] Rounded to one decimal place is 9.2. New number is 9.2. The last result is 9.2. Final answer should be 9.2	The model initially provides the correct answer, but then retracts and gives another one
	Reading comprehension	Rare	C was once a parallelepiped. By increasing its smaller size of 3 cm and decreasing the larger one of 5 cm, however, it has been transformed into a cube that has the same volume as the initial parallelepiped. What is the volume of c ui we speak?	The volume is 421875 cm3	$\frac{3375}{8}$	Let the dimensions of the parallelepiped be a, b, and c. We are given that the smaller size is 3 cm and the larger size is 5 cm. [...]	The model doesn't understand the text and considers the increase in sides as the direct measurement
GEMINI-2.0-FLASH	Reasoning	Almost always	On 29/09/2011 (September 29, 2011) is written using four digits, each of them employed twice. What will be the next date (written as the previous one, as day. month. year) which has the same property to be written with four digits, each of them employed twice?	The next date will be the 13th of 03-2012	"2/2/2022"	Let's analyze the given date, 29/09/2011. The digits used are 1, 2, 9, and 0, each appearing twice. We are looking for the next date that can be written using four digits, each appearing twice. [...] Let's start by considering the year. The year 2011 uses 1, 0, and 2. We need to find a year that uses four digits, each twice. The next possible year would be 2022. The digits are 0 and 2, each appearing twice. [...] Final Answer: The next date is $\boxed{02/02/2022}$	The model directly considers the year 2022, making an error
	Reading comprehension	Rare	The solid of the year is a convex polyhedron. The sum of the measurements of all its angles, except one, is of 2014 degrees. It has three triangular faces. How many vertices from which they depart three edges has? Note: the added angles are those on the polygonal faces. A grade is worth n/180 radiants.	2 answers : 6 or 7	V=8, E=12	Let P be the convex polyhedron. Let V be the number of vertices, E the number of edges, and F the number of faces. [...] The number of vertices is 8. The number of edges is 12. Final Answer: The number of vertices is 8 and the number of edges is 12. $\boxed{V = 8, E = 12}$	The question only asked for the number of vertices in a specific situation, but the model provides the general number of vertices and sides instead
GEMINI-1.5-PRO	Calculation	Rare	Find all pairs of positive integers x and y that match the same 9x+9y - 2xy+19=0	Couples (x;y) are: (5;64) + (8;13) + (13;8) + (64;5)	(5, 26), (26, 5)	[...] We can rewrite the equation as $2xy - 9x - 9y + 19 = 0$. [...] Final Answer: The final answer is $\boxed{(5, 26), (26, 5)}$	The model makes a mistake and puts +19 instead of -19
	Reasoning	Almost always	Consider the date of 31/12/2014: the sum of its digits is equal to 14 (3+1+1+1+2 +2+0+1+4 = 14). What is the first subsequent date for which the sum of the digits is equal to ncora 14?	The date is: 05/1/2015	29/12/2023	[...] We can start by incrementing the day. 01/01/2015: $0+1+0+1+2+0+1+5 = 10$ 02/01/2015: $0+2+0+1+2+0+1+5 = 11$ ... 10/01/2015: $1+0+0+1+2+0+1+5 = 10$ 11/01/2015: $1+1+0+1+2+0+1+5 = 11$ ... 31/01/2015: $3+1+0+1+2+0+1+5 = 13$ 01/02/2015: $0+1+0+2+2+0+1+5 = 11$ [...] Final Answer: The final answer is $\boxed{29/12/2023}$	The model completely skips the analysis of some dates, missing the correct answer as well
GEMINI-1.5-FLASH	Reading comprehension	Rare	Liliana celebrated her birthday on March 28, with two days late. Desire celebrated it the same day, even if her birthday falls 10 days after that of Liliana. What is the day of the birthday of Desire?	Desiderio's birthday is April 5th	april 7	Liliana's birthday is on March 28. She celebrated it two days late, which means she celebrated it on March 28 + 2 days = March 30. [...] Therefore, Desire's birthday is April 7. $\boxed{April 7}$	The model doesn't understand that the birthday was celebrated on the 28th, with a 2-day delay
	Reasoning	Almost always	Insert three addition signs (+) between the digits that go from 1 to 7, without changing their order, so that the result of the calculation is equal to 190: 1 2 3 4 5 6 7	1 2 3 + 4 + 5 6 + 7=190	No Solution	We are given the digits 1, 2, 3, 4, 5, 6, 7 and we need to insert three addition signs (+) between them such that the sum is 190. Let's explore possible combinations. [...] If we try 123 + 45 + 6 + 7, we get 171. Too low. If we try 1234 + 5 + 6 + 7, we get 1252. Too high. [...] Let's try 123 + 45 + 22 = 190. This doesn't work because we only have one 2. There seems to be no solution using only three plus signs. There is no solution to this problem.	The model tries to make some attempts, but then gets stuck and says that it cannot be solved

Table 13: Examples of errors made by Gemini models in text-only problems within MATHGAMES.

Model	Error type	Frequency	Question example	Gold	Answer	Reasoning	Comment
GPT-4O	Reasoning	Always	Renato is six years old, Amerigo is two years younger. What will be the age of Amerigo when Renato is ten times as old as he is now?	Amerigo will be 58 years old	56	To solve this problem, we need to determine Amerigo's age when Renato is ten times his current age. 1. **Determine Renato's current age:** Renato is currently 6 years old. 2. **Calculate Renato's age when he is ten times his current age:** Renato's current age is 6, so ten times his current age is: $10 \times 6 = 60$ Therefore, Renato will be 60 years old. 3. **Determine the age difference between Renato and Amerigo:** Amerigo is two years younger than Renato. Therefore, the age difference is: $6 - 2 = 4$ So, Amerigo is currently 4 years old. 4. **Calculate Amerigo's age when Renato is 60:** Since the age difference between Renato and Amerigo is 4 years, when Renato is 60, Amerigo will be: $60 - 4 = 56$ Thus, Amerigo will be $\boxed{56}$ years old when Renato is ten times as old as he is now.	In this case, the model always makes reasoning errors, as being riddles, a resolution might seem sensible but isn't correct in the end
GPT-4O MINI	Calculation	Rare	The 2010 divisible years are divisible by 10 (the number formed by its last two digits); in the same way 2016 is divisible by 16; Find the first two years, after 2016, whose number is divisible by that formed by the last two digits.	2020 2025 (both solutions, the order is not relevant)	2020 and 2022	6. **Year 2022**: - Last two digits: 22 - Check: $2022 \mod 22$ - Calculation: $2022 \div 22 \approx 91$ (an integer) - Result: Divisible. [...] $\boxed{\text{2020 and 2022}}$	2022 is not divisible by 22
	Reasoning	Almost always	That's how Carla and Milena have the passion of numbers. Carla writes the following sequence 4 12 6 18 9 9 3 15 and observes that each of her numbers is a divider or multiple of the previous one. Milena does not want to be less and writes then another sequence of seven numbers (other than 0 and all different between them) that ends with 8, the largest of the numbers written by Milena, in which once again each element is a divider or multiple of the previous one. Write the Milena sequence.	Two sequences: 5136248 or 7136248	1, 2, 4, 8, 6, 3, 12	To create a sequence of seven numbers that ends with 8, where each number is either a divisor or a multiple of the previous one, we can start with the number 8 and work backwards to find the previous numbers in the sequence. [...] Now we have the sequence: 1, 2, 3, 4, 6, 8, 12. This sequence has 7 different numbers, ends with 8, and each number is either a divisor or a multiple of the previous one.	The model claims that the sequence it provided ends with 8, but that's not true
O3 MINI	Reasoning	Almost always	Desire and Nando have booked two seats on the same plane. The automatic procedure that has served them the place has put them at random, but in a same row made up of six seats. Keeping in mind that the six seats are all occupied, what is the probability (expressed by an irreducible fraction) that Desire and Nando are the one beside the other or without one or more travellers being seated among them? N.B. It is assumed that Desiderio and Nando are sitting next to each other, not that they were divided by the corridor center ale.	The probability is 1/3	$\frac{4}{15}$	We are told that Desire and Nando are seated at random in a row of six seats that is fully occupied. However, because of the layout of the plane, the six seats are arranged as two groups of 3 (separated by the central aisle). In this configuration, even though seats 3 and 4 are consecutive in the row, a traveler seated in seat 3 is not next to the traveler in seat 4 (they are separated by the aisle). In other words, sentarse uno junto al otro (sitting next to each other) is only possible if they are consecutive seats on the same side. (The phrase sin uno o más viajeros entre ellos means that there is no other passenger sitting between them, i.e. they appear consecutively.) Step 1. Total possible pairs of seats Since Desire and Nando get two different seats out of 6, the number of equallylikely pairs is C(6,2) = 15. Step 2. Count favorable pairs (neighbors on the same side) Divide the row into two blocks: Left block: 3 seats arranged in order; the adjacent pairs here are (seat1, seat2) and (seat2, seat3). That gives 2 pairs. Right block: 3 seats arranged in order; the adjacent pairs here are (seat4, seat5) and (seat5, seat6). That gives 2 pairs. Thus the total favorable pairs = 2 + 2 = 4. Step 3. Form the probability The probability that Desire and Nando are seated with no other passengers between them (i.e. next to each other in the proper sense) is Probability = (Number of favorable pairs)/(Total number of pairs) = 4/15. Thus, the final answer is $\boxed{\frac{4}{15}}$.	Incorrectly assuming that any consecutive seats in the row of six constitute "sitting next to each other," without considering the separation by the aisle and the definition that adjacency must be on the same side of the aisle.
	Reading comprehension	Rare	On 17 June 2345 it will be a particular date because it will be possible to write with eight digits all different: 1 7 0 6 2 3 4 5 What is the last date that owned this property (to be able to write in the form of a number of eight digits all different from each other)?	The date is: 25th 06th 1987	30/12/9876	We wish to find, among all valid dates written in the two?digit day, two digit month, four digit year format (dd/mm/yyyy), the very last (i.e. chronologically greatest) date for which the eight digits are all different. [...]	The model understands that it should take the largest date, not the most recent one

Table 14: Examples of errors made by OpenAI models in text-only problems within MATHGAMES.

Model	Error type	Frequency	Question example	Gold	Answer	Reasoning	Comment
DEEPSEEK-R1	Reasoning	Always	On February 20, 2002 was a date palindrome because it is written as 20/02/2002 ch and reads the same way from left to right and from right to left. What was the next date palindrome?	1/2/2010	11/2/2011	Okay, so I need to find the next date palindrome after February 20, 2002, which is written as 20/02/2002. Hmm, let me start by understanding what a date palindrome is.[...] So that is a palindrome. So, the key is that when the date is written as an 8-digit number (DDMMYYYY), it reads the same forwards and backwards. [...]So, the date is 02/01/2010. Wait, but written as DDMMYYYY: 02012010. Reverse of this is 01020102, which is not the same as 02012010,[...] Wait, this is getting me more confused. [...] Therefore, to find the next palindrome date after 20/02/2002, we need to find the next occurrence where reverse(DDMM) = YYYY, and the date is valid. [...] Let's think. First, note that the day and month (DDMM) must form a four-digit number which, when reversed, gives the year. So, reverse(DDMM) = YYYY. Therefore, YYYY = MM DD reversed. Wait, reversing the four-digit number DDMM gives us MM DD. So, for example, if DDMM is 1102, reversing gives 2011, so the year would be 2011. Therefore, the date 11/02/2011 would be a palindrome: 11022011, which reversed is 11022011. Wait, let me check: 11022011 reversed is 11022011. Yes, that's a palindrome. So, that date is 11/02/2011. [...]	the model begins to think, also claiming to be confused, and in the end finds a valid year but skipping 2010
DEEPSEEK-V3	Reasoning	Almost always	With two equal circles, you can completely cover a square of 10 cm. side. How much is the radius of the two circles worth, at least? (Give the answer in mm., rounded to the nearest mm. If necessary, in the result replaced 1,414 to ?2; 1,732 to ?3; 2,236 to ?5).	The radius is worth 56mm.	112 mm	To determine the minimum radius of two equal circles that can completely cover a square with a side length of 10 cm, we need to analyze the geometric arrangement of the circles and the square. [...] Rounding to the nearest millimeter: ≈ 112mm	The model gives the measurement of a radius that is larger than the side of the square, which is not the correct reasoning to follow
	Calculation	Rare	Find all pairs of positive integers x and y that match the same 9x+9y - 2xy+19=0	Couples (x;y) are: (5;64) + (8;13) + (13;8) + (64;5)	(5, 83) and (83, 5)	To solve the equation $9x+9y-2xy+19=0$ for positive integers x and y, we will rearrange and factor the equation step by step. [...] Add $81/4$ to both sides to complete the rectangle (this is a standard technique for factoring equations of this form): $2xy-9x-9y+\frac{81}{4}=19+\frac{81}{4}$ The left-hand side can now be factored as: $(2x-9)(2y-9)=\frac{157}{4}$ [...]	The factorization is wrong because when solved, it leads to 4xy - 18x - 18y + 81
DEEPSEEK-MATH-7B	Calculation	Rare	Consider the date of 31/12/2014: the sum of the- its digits is equal to 14 (3+1+1+1+2+2+0+1+4 = 14). What is the first subsequent date for which the sum of the digits is equal to 20?	The date is 29/1/2015	"9/1/2015"	[...] 8. 08/01/2015: The sum of the digits is 8+1+1+2+0+1+5 = 19, which is not equal to 20. 9. 09/01/2015: The sum of the digits is 9+1+1+2+0+1+5 = 20, which is equal to 20. So, the first subsequent date for which the sum of the digits is equal to 20 is $\boxed{09/01/2015}$.	The model says that 8+1+1+2+0+1+5 = 19 or that 9+1+1+2+0+1+5 = 20
	Reasoning	Almost always	Find all pairs of positive integers x and y that match the same 9x+9y - 2xy+19=0	Couples (x;y) are: (5;64) + (8;13) + (13;8) + (64;5)	there are no solutions	To solve this problem, we need to find all pairs of positive integers x and y that satisfy the equation 9x + 9y - 2xy + 19 = 0. [...] x(9 - 2y) + y(9 + 19/y) = 0 Since x and y are both positive integers, the expression inside the parentheses must be equal to zero: 9 - 2y = 0 9 + 19/y = 0 Solving these equations, we get: y = 9/2 y = -19 Since y is a positive integer, the second equation has no solution. Therefore, the only solution for y is y = 9/2. However, we are looking for pairs of positive integers x and y, so y cannot be a fraction. Therefore, there are no pairs of positive integers x and y that satisfy the given equation. So, the answer is there are no solutions.	The model reasons incorrectly and arrives at a wrong conclusion, without making calculation or similar errors

Table 15: Examples of errors made by DeepSeek models in text-only problems within MATHGAMES.

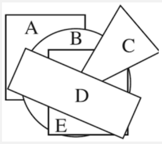
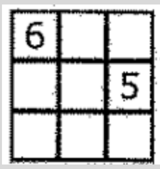
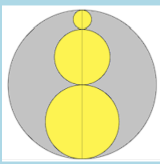
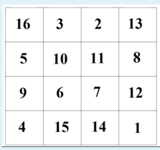
Model	Error type	Freq.	Question example	Image	Reasoning	Comment
GEMINI-2.0-FLASH-THINKER	Image comprehension	Often	Lavinia puts on the table (in a certain order) a rectangle of paper, a circle, two squares and a triangle always of paper. In figure you see the composition that it gets. What is the order in which Lavinia put the various forms of paper on the table? (Start writing the letter of form that Lavinia put first on the table and then away the others, until what Lavinia put last on the table). **Gold:** The order is ABECD **Answer:** EDABC		Let's analyze the image to determine the order in which the shapes were placed. [...] - Shape C is on top of Shape D (Rectangle) in the top-right part of C. [...] Shape C is on top of ALL other shapes it intersects. So, Shape C is the last one to be placed. Shape D (Rectangle): - Shape D is under Shape A (Square) in the top-left part of A. [...]So the order of letters is E, D, A, B, C.	The model claims that C is above all when it's not, or that D is below A
	Reasoning	Often	The grid must contain all numbers from 1 to 9 (one for box). 5 and 6 have already been written in figure. The eight sums of the three numbers placed on each of the three lines, of the three columns and of the two diagonals must all be different and must give all values from 10 to 18, except 13. Complete the grid. **Gold:** 2 solutions: 6 2 8 9 1 5 3 7 4 - 6 4 8 9 3 5 1 7 2 **Answer:** No solutions		The problem is to complete a 3x3 grid with numbers from 1 to 9, using each number exactly once. Numbers 5 and 6 are already placed. The sums of the three rows, three columns, and two diagonals must be distinct and take values from the set 10, 11, 12, 14, 15, 16, 17, 18. We are given that the sums are all different and take values from 10 to 18, except 13. Let the grid be a b c d e f g h i We are given a = 6 and f = 5. So the grid is 6 b c d e 5 g h i [...]	The model understood the text of the question and the image correctly, but was unable to find a final solution when there were two possible ones
GEMINI-1.5-PRO	Image comprehension	Often	Each of the four lamps in the figure is connected to a switch, and only to one. Each of the four switches lights a lamp, and only one, in one of its A or B positions, and turns it off in the other. A switch is not necessarily placed below the lamp it controls. Lucia wants to turn on the four lamps n ello same time. She made three attempts and the figure represents the result: a drawing on a white background represents a lamp on, one on a grey background one off. From left to right, how will Lucia have to place (A or B) the switches? **Gold:** BABA **Answer:** AAAB		[...] In the first attempt, the switches are in positions B, B, A, A. The first two lamps are off, and the last two lamps are on. In the second attempt, the switches are in positions B, A, A, B. The first and last lamps are off, and the second and third lamps are on. In the third attempt, the switches are in positions A, A, A, A. The first, second, and third lamps are on, and the fourth lamp is off. [...]	The model doesn't correctly understand the lights being off and on
	Reasoning	Often	Angelo is a great pastry chef who prepares large circular cakes (with a diameter of 74 cm). Lately, however, he has decided to diversify his production and, in the circular box of 74 cm of diameter of the large cake, he puts three cakes of different diameters but always expressed by a whole number of cm. The three cakes fall exactly in the package of the large cake, as he sees it in figure (where the proportions do not make text though). Angelo sells the package of the three cakes at the same price as the big cake but the surface of the three cakes, put together, is half that of the big cake. How many cm do the diameters of the three cakes measure? **Gold:** Diameters measure: 9 cm, 16 cm; 49 cm **Answer:** 4, 25, 25		Let D be the diameter of the large cake, so $D = 74$ cm. Let d_1, d_2, d_3 be the diameters of the three smaller cakes. The area of the large cake is $A = \pi(\frac{D}{2})^2 = \pi(\frac{74}{2})^2 = \pi(37)^2 = 1369\pi$. The sum of the areas of the three smaller cakes is half the area of the large cake, so $\pi(\frac{d_1}{2})^2+\pi(\frac{d_2}{2})^2+\pi(\frac{d_3}{2})^2 = \frac{1}{2}(1369\pi)$ $d_1^2+d_2^2+d_3^2 = \frac{1}{2}(4)(1369) = 2(1369) = 2738$. Also, the sum of the diameters of the three smaller cakes is equal to the diameter of the large cake, so $d_1+d_2+d_3 = 74$. We are given that d_1, d_2, d_3 are integers. We can try to find integer solutions for d_1, d_2, d_3. Let's assume $d_1 \leq d_2 \leq d_3$. If $d_1 = d_2 = d_3$, then $3d_1 = 74$, so $d_1 = \frac{74}{3}$, which is not an integer. [...] If $d_1 = 18$, $d_2 = 26$, $d_3 = 30$, then $18 + 26 + 30 = 74$. $18^2 + 26^2 + 30^2 = 324 + 676 + 900 = 1900 \neq 2738$. If $d_1 = 22$, $d_2 = 24$, $d_3 = 28$, then $22 + 24 + 28 = 74$. $22^2 + 24^2 + 28^2 = 484 + 576 + 784 = 1844 \neq 2738$. If $d_1 = 24$, $d_2 = 24$, $d_3 = 26$, then $24+24+26 = 74$. $24^2 + 24^2 + 26^2 = 576+576+676 = 1828 \neq 2738$. If $d_1 = 24, d_2 = 25, d_3 = 25$, then $24 + 25 + 25 = 74$. $24^2+25^2+25^2 = 576+625+625 = 1826 \neq 2738$. [...] Final Answer: The final answer is $\boxed{24, 25, 25}$	The error is in concluding that 24, 25, 25 is the final answer after demonstrating that it does not satisfy the derived equations.
	Reading comprehension	Rare	Dürer's magic square, contained in its famous "Melancholia", has a peculiarity: the sum of two symmetrical numbers compared to the center of the square is always equal to half of the magic sum (i.e. to 17). Find yourself another 4x4 magic square, with the numbers 15 and 14 written in this order in the center of the fourth line (figure), which enjoys the same property. **Gold:** 3 solutions: 13 3 2 16 8 10 11 5 12 6 7 9 1 15 14 4 - 16 3 2 13 9 6 7 12 5 10 11 8 4 15 14 1 - 13 3 2 16 12 6 7 9 8 10 11 5 1 15 14 4 **Answer:** The problem is ill-posed		Let the magic square be represented by the matrix A: $A = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \\ m & 15 & 14 & n \end{pmatrix}$ The magic sum is $1 + 2 + ... + 16 = \frac{16 \cdot 17}{2} = 136$. [...]	The model doesn't understand that the magic sum is 17

Table 16: Examples of errors made by Gemini models in multimodal problems within MATHGAMES.

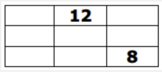
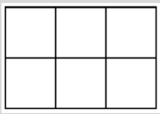
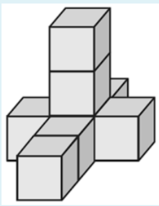
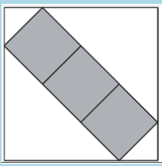
Model	Error type	Freq.	Question example	Image	Reasoning	Comment
GPT-4o	Image comprehension issues	Often	Fill the free boxes of the grid in figure with positive integer numbers so that in each row, in each column and in each of the two diagonals the middle number is the arithmetic mean of those at its extremes. In how many different ways can you do it? **Gold:** In 7 ways **Answer:** 15		[...]Let's denote the grid as follows: a | 12 | b c | d | e f | 8 | g [..]	The model places the 8 in the wrong position
	Reasoning issues	Often	How many squares can you see in the picture? **Gold:** The number of squares is 8 **Answer:** 9		To find the total number of squares in the grid, we need to consider squares of different sizes. [...] 3. **3x2 Square:** - The entire grid itself is a 3x2 square. Adding these up, we have: - 6 (1x1 squares) - 2 (2x2 squares) - 1 (3x2 square) Total number of squares = 6 + 2 + 1 = 9 Thus, the total number of squares is $\boxed{9}$.	In the reasoning, the model incorrectly considers the 2x3 rectangle as a square
GPT-4o MINI	Image comprehension	Often	Desire built the solid you see in the figure, pasting some white cubes between them. Then, he painted blue all the faces of the solid, including those of its lower base. In the end, taken from a raptus, he again separated the various initial cubes. How many of them have exactly one and only one white face? **Gold:** The cubes are 5 **Answer:** 0		[..] 1. **Identify the structure**: The solid consists of several cubes arranged in a specific configuration. From the image, we can see that there are: - 1 cube on top (the highest cube). - 3 cubes in a horizontal row at the bottom. - 1 cube in the middle of the vertical stack. 2. **Count the cubes**: - There are 5 cubes in total. [...] Thus, the number of cubes that have exactly one and only one white face is: $\boxed{0}$	The model only sees 5 cubes
	Reasoning	Often	The centers of the smaller square (dark in figure), located within the larger square, are al-lined along one of the two diagonals of the larger square. What is the area of this square, knowing that each of the square-dark you has an area of 17 cm2? **Gold:** Large square area measure 136 cm2 **Answer:** 76.5		[...] 3. **Calculate the total length along the diagonal:** Since there are three smaller squares, the total length along the diagonal is: Total diagonal length $= 3 \times \sqrt{17}\,\mathrm{cm}$ [...]	The model incorrectly reasons about the calculation of the square's diagonal

Table 17: Examples of errors made by OpenAI models in multimodal problems within MATHGAMES.

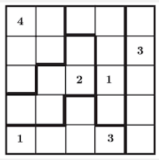
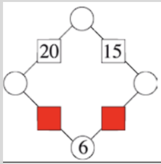
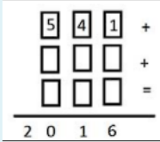
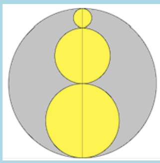
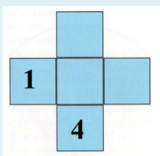
Model	Error type	Freq.	Question example	Image	Reasoning	Comment
QWEN2-VL-72B	Image compre-hension	Almos Al-ways	Fill the boxes of the square with the digits 1, 2, 3, 4, 5 and in particular write on the sheet-answer the digits of the first line (top), from left to right. However: each of these digits must appear one and only once in each row, in each column and in each of the five pieces in which the square was divided. **Gold:** 4 1 3 5 2 **Answer:** 4 1 2 3 5		[...] Let's start by analyzing the given quadrato: $A = \begin{pmatrix} 4 & \_ & \_ & \_ & \_ \\ \_ & \_ & \_ & \_ & \_ \\ \_ & \_ & \_ & \_ & \_ \\ \_ & \_ & \_ & \_ & \_ \\ \_ & \_ & 3 & \_ & \_ \end{pmatrix}$ [...]	The model misreads the table in the image
	Reasoning	Somet	In the square you see in figura, all the numbers are larger than 1. The numbers placed at the ends of each side of the square are multiplied between them and the result is written in the square placed in the middle of the side. Taking into account the three numbers that already appear in figura, you will find those to svere in the two red squares. **Gold:** In the square on the left 24 in the square on the right 18 **Answer:** 2, 3		[...] From the figure, we have the following information: - The product of the numbers at the ends of the top side is 15. - The product of the numbers at the ends of the left side is 20. - The product of the numbers at the ends of the bottom side is 6. [...]	The model correctly associates the two numbers above in the figure with their meaning, but reasons incorrectly with the 6, associating it with the multiplication result instead of a factor
INTERNVL 2.5-8B	Image compre-hension	Often	Fill the three free boxes of the cross with the numbers 2, 3, 5, so that, adding the three numbers of the horizontal axis, you get the same result obtained by adding the three numbers of the vertical axis. **Gold:** 2 1 3 5 4 **Answer:** -3, 1, 0		[...] We are given the following equations based on the sums of the rows and columns: 1. $1+x+4=1+y+z$ 2. $1+x+5=1+4+z$ 3. $4+x+5=1+y+5$ [...]	The model misinterprets the rows and columns of the image
	Reasoning	Often	Angelo is a great pastry chef who prepares large circular cakes (with a diameter of 74 cm). Lately, however, he has decided to diversify his production and, in the circular box of 74 cm of diameter of the large cake, he puts three cakes of different diameters but always expressed by a whole number of cm. The three cakes fall exactly in the package of the large cake, as he sees it in figure (where the proportions do not make text though). Angelo sells the package of the three cakes at the same price as the big cake but the surface of the three cakes, put together, is half that of the big cake. How many cm do the diameters of the three cakes measure? **Gold:** Diameters measure: 9 cm, 16 cm; 49 cm **Answer:** 1, 2, 3		Let's denote the diameters of the three smaller cakes as a, b, and c. The area of the large cake is $\pi \times (37)^2 = 1369\pi$ square cm. The combined area of the three smaller cakes is half of this, so it is $1369\pi/2 = 684.5\pi$ square cm. The areas of the cakes are proportional to the squares of their diameters, so we have: $\frac{a^2}{1369} + \frac{b^2}{1369} + \frac{c^2}{1369} = \frac{1}{2}$ Simplifying, we get: $a^2 + b^2 + c^2 = 684.5$ Given that a, b, and c are whole numbers, we need to find three whole numbers that satisfy this equation. The only possible solution is $a = 1$, $b = 2$, and $c = 3$, as these are the only whole numbers that satisfy the equation. Thus, the diameters of the three cakes are 1 cm, 2 cm, and 3 cm. The answer is: $\boxed{1, 2, 3}$.	(1) Incorrectly simplifying the relationship between diameters and the area condition, leading to a drastically wrong target sum of squares. (2) Making an unsubstantiated claim that 1, 2, 3 is the "only possible solution" and failing to verify if it even satisfies their own wrongly simplified equation (it doesn't) and ignoring other possible constraints.
	Reading compre-hension	Rare	The six white boxes must be completed with different digits, so that the addition shown is correct and that in each column, from top to bottom, the grid numbers are written from the smallest to the largest. Write in particular the digits of the second line. **Gold:** 6 8 2 **Answer:** 5, 4, 1, 6, 9, 0		To solve the problem, we need to fill in the six white boxes with different digits so that the sum of the numbers in each column from top to bottom is increasing. [...]	The model understands that the sums of the numbers in the columns must be equal

Table 18: Examples of errors made by open-source models in multimodal problems within MATHGAMES.