OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems

Anonymous ACL submission

Abstract

Recent advancements have seen Large Language Models (LLMs) and Large Multimodal Models (LMMs) surpassing general human capabilities in various tasks, approaching the proficiency level of human experts across multiple domains. With traditional benchmarks becoming less challenging for these models, new rigorous challenges are essential to gauge their advanced abilities. In this work, we present OlympiadBench, an Olympiad-level bilingual multimodal scientific benchmark, featuring 8,952 problems from Olympiad-level mathematics and physics competitions, including the Chinese college entrance exam. Each problem is detailed with expert-level annotations for step-by-step reasoning. Evaluating top-tier models on OlympiadBench, we implement a comprehensive assessment methodology to accurately evaluate model responses. Notably, the best-performing model, GPT-4V, attains an average score of 17.23% on OlympiadBench, with a mere 11.28% in physics, highlighting the benchmark rigor and the intricacy of physical reasoning. Our analysis orienting GPT-4V points out prevalent issues with hallucinations, knowledge omissions, and logical fallacies. We hope that our challenging benchmark can serve as a valuable resource for helping future AGI research endeavors.

Question: Find all triples (x, y, z) of positive integers such that $x \leq y \leq z$ and $x^3(y^3 + z^3) = 2012(xyz + z)$.

Solution: First note that x divides $2012 \cdot 2 = 2^3 \cdot 503$. If $503 \mid x$ then the right-hand side of the equation is divisible by 503^3, and it follows that $503^2 \mid xyz + 2$. This is false as $503 \mid x$. Hence $x = 2^m$ with $m \in \{0,1,2,3\}$. If $m \geq 2$ then $2^6 \mid 2012(xyz + 2)$. However the highest powers of 2 dividing 2012 and $xyz + 2 = 2^m yz + 2$ are 2^2 and 2^1 respectively. So $x = 1$ or $y = 1$, yielding the two equations

$$y^3 + z^3 = 2012(yz + 2),$$
$$y^3 + z^3 = 503(yz + 1)$$

In both cases ...... It follows that $y \equiv -z (mod\ 503)$ as claimed. Therefore $y + z = 503k$ with $k \geq 1$. In view of $y^3 + z^3 = (y + z)((y - z)^2 + yz)$ the two equations take the form

$$k(y - z)^2 + (k - 4)yz = 8 \quad (1)$$
$$k(y - z)^2 + (k - 1)yz = 1 \quad (2)$$

In (1) we have $(k - 4)yz \leq 8$, which implies k ≤ 4 ......
Therefore (1) has no integer solutions.
Equation (2) implies $0 \leq (k - 1)yz \leq 1$, so that $k = 1$ or $k = 2$.
Also $0 \leq k(y - z)^2 \leq 1$, hence $k = 2$ only if $y = z$. However then $y = z = 1$, which is false in view of $y + z \geq 503$.
Therefore $k = 1$ and (2) takes the form $(y - z)^2 = 1$, yielding $z - y = |y - z| = 1$. Combined with $k = 1$ and $y + z = 503k$, this leads to $y = 251, z = 252$.
In summary the triple (2,251,252) is the only solution.

Final answer: (2,251,252) **Subfield:** Number theory
Answer type: Sequence **Question type:** Open-ended

Figure 1: An example of IMO in OlympiadBench. Solving this example requires AI systems to span different mathematical domains and conduct advanced reasoning.

1 Introduction

Large Language Models(LLMs) have demonstrated remarkable capabilities across various tasks such as text generation (Zhao et al., 2023), code generation (Zan et al., 2023) and mathematical reasoning (Lu et al., 2023; Zhou et al., 2023), garnering significant attention from both academia and industry (Wei et al., 2022; Zhao et al., 2023; Bubeck et al., 2023). The most powerful models such as GPT-4 (OpenAI, 2023a) and Gemini Ultra (Team, 2023) have even surpassed oridinary human level on a wide variety of benchmarks such as MMLU (Hendrycks et al., 2020), MMMU (Yue et al., 2023), and even surpassing human expert in many area. These results show a promising future that LLMs can serve as proficient assistants for human scientists (Nguyen, 2023; Qiu et al., 2023). Among the array of expert-level skills exhibited by LLMs, scientific reasoning consistently emerges as one of the most brilliant, showcasing some of the most distinguished intellectual properties that experts possess. Therefore, this paper primarily focuses on mathematical and physical reasoning.

In recent years, several benchmarks related to mathematics have been proposed, such as the dataset GSM8K (Cobbe et al., 2021) as well as the dataset MATH (Hendrycks et al., 2021). However, these benchmarks, are primarily developed before the advent of highly capable LLMs, and now lack sufficient challenge for the latest models. For instance, GPT-4 with prompting techniques(Zhou et al., 2023) has achieved a 97.0% success rate on GSM8K and 84.3% on MATH. The rapid evolution of LLMs may soon lead to saturated results on these benchmarks. Concurrently, LLMs are not yet fully equipped to assist mathematicians in solving complex problems (Collins et al., 2023; Zhang et al., 2023), nor are they capable of performing expert-level mathematical reasoning independently. This discrepancy underscores the need for more challenging datasets to benchmark future advancements of LLMs in this domain. Similarly, physics presents comparable challenges for AI to those found in mathematics. Nevertheless, existing benchmarks related to physics (Lu et al., 2022; Arora et al., 2023; Wang et al., 2024) are characterized by their relatively low difficulty and limited scope. There is also a significant lack of a rigorous and challenging benchmark in physics.

In addition to the issue regarding the benchmark difficulty, it is important to note that these benchmarks predominantly focus on text. This presents a significant limitation, as a wide range of scientific reasoning contexts require multimodal reasoning abilities. For example, grasping geometry reasoning in mathematics or understanding experiments designs in physics are scenarios where multimodal reasoning capabilities are crucial. Notably, various large multimodal models (LMMs) have been developed (Team, 2023; Liu et al., 2023) and demonstrate proficiency on a variety of tasks (Lu et al., 2022; Yue et al., 2023; Zhang et al., 2024b; Lu et al., 2024), offering the potential for multimodal scientific reasoning. Nevertheless, there is still a lack of sufficient benchmarks to prove whether these LMMs are capable of handling scientific problems. Consequently, a challenging multimodal benchmark is essential for advancing scientific reasoning tasks(Zhang et al., 2024a; Lu et al., 2023).

To address the aforementioned inadequacies, we introduce OlympiadBench, a Olympiad-level bilingual multimodal scientific benchmark. This collection comprises 8,952 math and physics problems sourced from International Olympiads, Chinese Olympiads, and the most challenging segments of the Chinese College Entrance Exam (GaoKao). We download PDF data from official websites and utilize Mathpix[1] for OCR parsing. We meticulously inspect, clean, and revise the data, and further adopt LLMs for deduplication. Finally, we annotate the data with crucial information such as answer types and subfields, yielding a dataset that is clean, accurate, and detailed. As shown in Figure 1, OlympiadBench features numerous distinct characteristics such as difficulty, free-form generation, expert-level solution annotation, detailed labeling of difficulty, wide-coverage of modality and language, etc. These features are summarized more clearly from Table 1.

We evaluate a wide variety of LLMs and LMMs on OlympiadBench. GPT-4V, a fusion of the strongest LLMs and LMMs, achieves a score of 20.35% in mathematics, 11.28% in physics. Importantly, the experiment results show that LMMs still struggle in computational error, incorrect reasoning or induction. For the process involved in the correct responses, the process occasionally includes hallucinated reasoning, or choosing a more complex solution when a simpler solution exists. All these results highlight the substantial challenge our benchmark presents to contemporary models and point the direction of future efforts.

OlympiadBench is inspired by the significant advances made by DeepMind AlphaGeometry (Trinh et al., 2024), which nearly matches the proficiency of International Mathematical Olympiad (IMO) gold medalists in geometry proofs. It is clear that OlympiadBench, along with other challenging datasets like the AI-MO challenge[2], will witness and benchmark the swift progress towards expert-level AI assistants for solving scientific problems.

2 Related Work

This section gives an overview of the existing datasets in solving mathematics and physics problems as well as multimodal datasets.

Mathematics Benchmarks. Solving mathematics problems and proving theorems in natural languages has been a key research focus in machine learning and natural language processing since the 1960s (Bobrow, 1964). Previous benchmarks (Koncel-Kedziorski et al., 2016; Wang et al., 2017; Ling et al., 2017; Amini et al., 2019; Cobbe et al., 2021; Wei et al., 2023) focus predominantly on math word problems (WMPs) which involve four basic arithmetic operations with single or multiple operation steps (Lu et al., 2023). Typically, the GSM8K (Cobbe et al., 2021) dataset targets

[1] https://mathpix.com/

[2] https://aimoprize.com/

elementary-level questions within 8 steps of basic arithmetic operations. However, these problems are typically text-only (Lu et al., 2023) and of lower difficulty, with reasoning limited to a few computations. As the complexity of the problems rises, some works (Hendrycks et al., 2021; Frieder et al., 2023; Arora et al., 2023) introduce competition-level problems integrating mathematical logic and background knowledge. Yet, these challenging datasets are increasingly being surmounted (Zhou et al., 2023). Theorem proving is a problem to demonstrate the truth of a mathematical claim (a theorem) through a sequence of logical arguments (a proof) (Lu et al., 2023). Earlier efforts mainly focused on translating natural language proofs into formal representations, facing significant expertise and labor challenges (Zheng et al., 2022; Welleck et al., 2021). The emergence of LLMs has facilitated notable advancements in the domain of natural language proof (Jiang et al., 2023). OlympiadBench presents mathematical reasoning and theoretical proofs all in natural language with detailed solution annotations.

Physics Benchmarks. Physics questions in SciQ (Welbl et al., 2017) and ScienceQA (Lu et al., 2022) are mainly elementary and high school level multiple-choice questions, lacking complex reasoning and computational tasks. In MMLU-STEM (Hendrycks et al., 2020) and C-Eval-STEM (Huang et al., 2023), physics questions also adopt a multiple-choice format. JEEBench (Arora et al., 2023) extends this format to include multi-step reasoning with physics knowledge, yet it is limited in scope and purely text-only. SciEval (Sun et al., 2023) consists of a total of about 18,000 challenging scientific questions, spanning three important basic science fields: chemistry, physics and biology. SciBench (Wang et al., 2024) and OCW-Courses (Lewkowycz et al., 2022) offer college-level physics questions in free-response formats, where SciBench contains multimodal information. In contrast, OlympiadBench escalates in difficulty, diversifies in question types, and surpasses in volume, setting a new benchmark for complexity and variety in the domain.

Multimodal Benchmarks. For assessing multimodal capability, works such as Geometry3K (Lu et al., 2021), GeoQA (Chen et al., 2021), GeoQA+ (Cao and Xiao, 2022), and UniGeo (Chen et al., 2022) have employed multimodal information for tackling geometric problems, integrating natural language descriptions with diagrams. ScienceQA (Lu et al., 2022), MMMU (Yue et al., 2023), CMMMU (Zhang et al., 2024b) and CMMU (He et al., 2024) are multimodal, multi-discipline evaluation sets, encompassing a broad range of subjects. MathVista (Lu et al., 2024) integrates 28 existing and 3 newly constructed multimodal datasets involving mathematics, aiming to establish a benchmark that encapsulates challenges from a variety of mathematical and visual tasks. However, it does not concentrate on delving into the complexity of mathematics problems.

In summary, we introduce a new benchmark to address these gaps. Table 1 presents a comparison between OlympiadBench and several related benchmarks, highlighting the significant advantages of OlympiadBench across all aspects.

3 The OlympiadBench Dataset

To evaluate the reasoning abilities of LLMs and LMMs in mathematics and physics problems, we have created OlympiadBench, a bilingual and multimodal scientific benchmark at the competition level. This section provides a detailed account of the construction process of OlympiadBench. Summarized statistics of the dataset is shown in Table 2, and more detailed statistics per subject are in Appendix A.2.

3.1 Design Principle

The motivation behind the design of OlympiadBench is to establish a benchmark that represents the pinnacle of human intellectual achievement, thereby encouraging researchers of LLMs to push the boundaries of mathematical and physical reasoning capabilities. To realize this vision, we focus on curating challenges that epitomize the highest level of competition worldwide. Specifically, our benchmark includes:

1. **Inclusion of Olympiad-Level Problems.** The chosen competition problems are of Olympiad caliber, aimed at the most accomplished students in a region's high school education phase. These problems are open-ended, setting them apart from the conventional multiple-choice or fill-in-the-blank formats. This selection is designed to more accurately capture the complexity of advanced mathematical reasoning.

2. **Provision of Detailed Solutions.** Given the advanced difficulty of these problems, which

Benchmark	Subject		Multi-	Detailed	Difficulty	Size		Answer	Language	Question
	Maths	Physics	modal	solution	level	Maths	Physics	type	type	type
SciBench	✓	✓	✓	✓	COL	217	295	Num	EN	OE
MMMU	✓	✓	✓	✓	COL	540	443	Num	EN	MC,OE
MathVista	✓		✓		-	1,000		Num	EN	MC,OE
ScienceQA		✓	✓		H		617		EN	MC
SciEval		✓			-		1,657	Num	EN	MC,FB,J
JEEBench	✓	✓		✓	CEE	236	123	Num	EN	MC,OE
MMLU	✓	✓			COL	948	548		EN	MC
AGIEval	✓	✓			CEE	953	200	Num	EN,ZH	MC,FB,OE
GSM8K	✓			✓	E	1,319		Num	EN	OE
MATH	✓			✓	COMP	5,000		Num,Exp,Tup	EN	OE
OlympiadBench	✓	✓	✓	✓	COMP	6,524	2,428	ALL	EN,ZH	OE

Table 1: For **difficulty level**, COMP: Competition, COL: College, CEE: College Entrance Examination, H: High School, E: Elementary School, and we picked the highest level; For **answer type**, Num: Numeric value, Exp: Expression, Equ: Equation, Int: Interval, Tup: Tuple; For **language type**, EN: English, ZH: Chinese; For **question type**, OE: Open-ended, MC: Multiple-choice, FB: Fill-in-the-blank, J: Judgement. For the statistical analysis of quantity and relevant metrics in AGIEval, we exclude 1,000 questions from the MATH benchmark to facilitate a more accurate comparison. The "-" indicates that it cannot be confirmed. Upon comparison, OlympiadBench leads in all aspects.

Statistics	**Number**
Total Problems	8,952
* Problems with images	5,129 (57%)
* Problems with solutions	8,952 (100%)
Difficulties (CEE: COMP)	66%: 34%
EN: ZH	2,288: 6,664
Open-ended Questions	7,254 (81%)
Theorem Proving	1,698 (19%)
Math: Physics	6,524: 2,428
* Maths with images	3,102
* Physics with images	2,027
Average question tokens	253
Max question tokens	3,701
Average solution tokens	352
Max solution tokens	4,213

Table 2: Statistics of OlympiadBench. When calculating tokens, images are not included.

may exceed the comprehension of individuals without a specialized background in mathematics, each problem is accompanied by expertly crafted solutions that detail the reasoning steps involved. This approach can not only reduces the difficulty of annotation and evaluation but also enhances the accuracy of the solutions provided.

3. **Incorporation of Visuals.** Recognizing the crucial role of visual information in conveying complex ideas, our benchmark incorporates problems that require understanding images, identifying spatial relationships, and other advanced reasoning tasks. This inclusion aims to assess and improve the model's capabilities in interpreting visual data as part of its reasoning process.

4. **Minimization of Data Leakage Risks.** To minimize the risk of data leakage, we have sourced problems from official Olympiad competitions, converting them from their original PDF files provided by official websites to the markdown format required. This strategy is aimed at reducing the likelihood of the data being inadvertently incorporated into the pre-training corpora of models.

Through these meticulously planned criteria, OlympiadBench aspires to not only challenge but also to significantly advance the frontier of LLM capabilities in mathematical and physical reasoning.

3.2 Data Processing

The data processing pipeline is structured into three distinct phases: data collection, format conversion & deduplication, and classification labeling.

Data Collection. OlympiadBench is meticulously compiled from three primary sources: Global Mathematics and Physics Olympiad Problems, Regional and National Chinese Math Competitions, and Gaokao Mock Questions for Mathematics and Physics [3]. Each chosen for its distinct advantages in creating a robust and comprehensive benchmark for evaluating LLMs and LMMs in mathematical and scientific reasoning. Their challenges progressively increase in difficulty, not only distinguishing the reasoning capabilities of models of various sizes but also offering guidance on scaling laws for specialized models in these domains.

Format Conversion and Deduplication. After collecting all PDF files, we utilize the Mathpix [4] tool for OCR recognition and convert them into markdown format. However, no conversion process is flawless, necessitating manual verification by our team members between the original PDF files and the converted Markdown texts. The Markdown texts are further structured into a format akin to "Problem—Solution—Answer", employing its markup language for text organization. Subsequently, we leverage a specialized small-scale language model [5] trained on mathematical symbol corpora for vectorizing the data and performing deduplication based on cosine similarity.

Classification Labeling. We note that both mathematics and physics problems predominantly comprise two types of questions: the open-ended problems and the theorem proving problems. We also note that the dataset, enriched by both Olympiad and national examination questions, covers a broad spectrum of subfields, as illustrated in Figure 2. Therefore, we manually annotate each question with topic and problem type annotations.

Answer type	Example
Numeric	1/4
Expression	$x = (1/2)at^2$
Equation	$x^2 + y^2 = 1$
Tuple	$(x, y, z) = (0, 0, 0)$
Interval	$(-\infty, -1) \cup (1, +\infty)$

Table 3: Examples of the five answer types

3.3 Data Characteristics

In contrast to previous benchmarks, OlympiadBench unveils two unique characteristics within its dataset: the incorporation of Progressive Problems in Physics and the categorization of answers

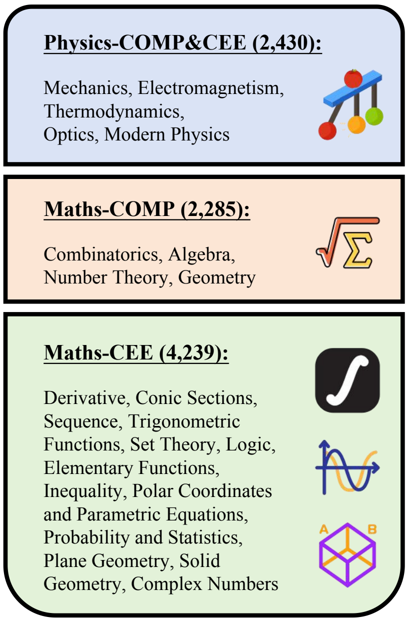

Figure 2: Subfields Distribution of OlympiadBench

to most open-ended questions into a limited number of types.

Progressive Problems in Physics. In physics competitions such as the International Physics Olympiad (IPhO), problems are often structured around a common material or scenario, with subsequent questions potentially relying on the answers or information from previous questions. One example is given in Figure 7 and Figure 8 This design characteristic is commonly referred to as "Progressive Problems." By linking a series of questions together, progressive problems require participants to apply their knowledge and skills comprehensively to gradually solve more complex issues. This type of question design aims to test students' depth of understanding, application capabilities, and innovative thinking, rather than just basic knowledge. To better utilize this feature, we have compiled the material, questions, and their answers that precede each progressive problem into its "Context" field.

Answer Type Classification. Whether in mathematics or physics, the answers to problems requiring definitive responses can largely be categorized into the following types: numeric, expression, equation, interval, and tuple. Simple examples of these can be seen in Table 3.

3.4 Automatic Scoring Pipeline

We design an automated scoring pipeline (see Algorithm 1) to evaluate model-generated answers across complex fields like mathematics and physics,

[3]Due to anonymity of this submission, we can not add concrete source in this version.

[4]https://mathpix.com/

[5]https://huggingface.co/Laurie/Bloom1b7-deepspeed-chat-Chinese-math

where answers vary from numbers to equations. This method simplifies answers into two categories: numeric values, handled through floating-point operations, and symbolic expressions, requiring symbolic computation.

For equations, we ensure all terms are on one side before dividing to check for mathematical equivalence. Intervals and tuples are compared by extracting and evaluating each element. Numeric answers are verified against a small tolerance of error, defaulting to 1e-8 but adjustable for physics problems to allow for a specific error margin. For expressions, we use the SymPy[6] library to confirm if the subtraction of two expressions approaches zero, indicating correctness.

4 Experiments

4.1 Settings

We conduct evaluations of open-source and closed-sourced LMMs that have been selected with consideration of their comprehensive capabilities on OlympiadBench. At the same time, we have selected LLMs with strong mathematical and logical abilities for evaluation on plain text questions.

As no accurate automatic evaluation method for theorem proving exists, we run full experiment on the automatic-scoring-available open-ended problems with answer type included in the Table 3, which is discussed in this section. We do manual sampling check of GPT-4V for theorem proving problems with analysis reported at Section 5.1.

4.1.1 Prompts

We evaluate the models in a zero-shot setting. Due to the high difficulty of the OlympiadBench questions, there should be considerable randomness in the results when using small batch data as the validation set, so we directly use a specific prompt template for all models instead of conducting prompt-engineering for each model respectively. The prompt template for English and Chinese open-ended questions is shown in the figure 3. To ensure the most complete extraction of the model's final results, we explicitly prescribe the types and formats of the answers in the prompt to promote the accuracy of the machine's automatic scoring. In order to test the native mathematical and physical abilities of the models, the prompts used in the test do not introduce knowledge points and other extra information contained in the dataset, but this information can be applied in subsequent research. Note that deepseek-math-7B-RL (Shao et al., 2024) requires the addition of a specific chain-of-thought prompt at the end of the input, which we adhered to during the evaluation.

4.1.2 Evaluation Workflow

We first apply each model to generate answers for questions in OlympiadBench using prompts formed by prompt template, with open-ended models running on NVIDIA A800 GPUs. Then, we run the automatic scoring pipeline to judge the correctness of the answers as described in subsection 3.4. Finally, we calculate the micro-average accuracy as the comparing metric. The code of the whole workflow is provided in the supplementary material.

4.2 Baselines

In our study, we evaluate the performance of current leading bilingual large multimodal models (LMMs), as well as bilingual large language models (LLMs) that has strong mathematical and reasoning abilities. We take both open- and closed-source models into consideration, and use the largest and latest released checkpoint or the best-performing official API that can be achieved.

For LMMs, we selected GPT-4V(GPT-4-Vision) (OpenAI, 2023b), Gemini-Pro-Vision (Team, 2023), Qwen-VL-Max (Bai et al., 2023) for closed-source models, while Yi-VL-34B (01-ai, 2024) and LLaVA-NeXT-34B (Liu et al., 2024) for open-source models. For models that demand compulsory image input, we take their LMM counterpart (corresponding text-model api or base LLM) for evaluation.To examine the impact of replacing LMM with base LLM for processing text-only data, we subsequently compare the performance differences between GPT-4V and GPT-4 [7] on text-only questions in OlympiadBench.

For LLMs, we select DeepSeekMath-7B-RL (Shao et al., 2024) as the primary baseline for text-only questions, and report the results of the selected LMMs (or their LLM counterparts) on the text-only questions for comparison, and additionally evaluate GPT-4 as described above.

4.3 Main Results

The overall experiment result is shown in table 4. Based on the results, our key findings can be sum-

[6] https://www.sympy.org

[7] The version of GPT-4 and GPT-4V are both "0125-preview".

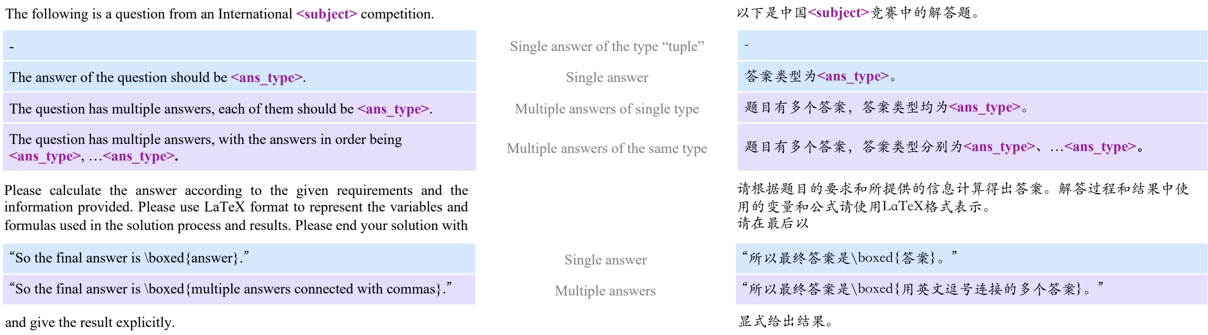

Figure 3: The template of the construction of the prompt for English(left) and Chinese(right) open-ended questions, among which <subject>, <ans_type>, and whether there are multiple answers can all be obtained from the data items in OlympiadBench dataset.

Models	Maths					Physics			Avg.
	En_COMP	Zh_COMP	Zh_CEE	Avg.		En_COMP	Zh_CEE	Avg.	
LLaVA-NEXT-34B† [8]	3.88	2.43	4.85	4.36	-	2.12	1.69	1.80	3.60
Yi-VL-34B† [9]	4.38	5.11	4.68	4.67	-	0.95	1.75	1.55	3.37
Gemini-Pro-Vision [10]	6.98	2.38	5.36*	5.35	-	3.01*	2.39	2.89	4.38
Qwen-VL-Max	10.80*	13.20	13.27*	12.76	-	3.97*	6.40*	4.44	10.31
GPT-4V	27.56*	14.94	19.15	20.35	-	11.41	10.74	11.28	17.23
Experiment with text-only									
LLaVA-NEXT-34B	3.64	2.71	9.88	6.39	-	2.16	5.45	1.98	5.80
Yi-VL-34B[9]	4.18	5.30	9.08	6.69	-	0.87	9.08	1.21	5.95
DeepSeekMath-7B-RL	19.73	2.71	27.54	20.25	-	6.49	15.45	9.38	18.73
Gemini-Pro-Vision[10]	7.62	2.71	10.26*	7.96	-	4.76	7.27	5.56	7.63
Qwen-VL-Max	11.68*	14.29	27.34	19.83	-	4.37*	21.82	10.03	18.47
GPT-4V	29.51*	16.01	38.31*	31.20	-	12.99	25.45	16.98	29.24
GPT-4	30.69*	16.50	38.48	31.75	-	11.44	28.18	16.84	29.50

Table 4: Experimental results. En_COMP: COMP problems in English, Zh_COMP: COMP problems in Chinese, Zh_CEE: CEE problems in Chinese. For closed-source models, the responses for some problems are not available, we mark the results with * (all of the proportion of missing answers are less than 5%). The causes are further described in Appendix B.3. Moreover, LLaVA-NEXT-34B and Yi-VL-34B only accepts input with single image, we mark results from only one image input with †.

marized as the following:

OlympiadBench is more challenging than existing benchmarks, which provides new perspective to compare LMMs. As shown in table 7, the most advanced model only achieves an average accuracy of 17.23% on OlympiadBench, which is much lower than that of existing benchmarks. Moreover, the gap between the models has been widened, thereby becoming more significant, which helps people to compare the differences in capabilities between different models more accurately.

There still exists a huge difference between the most powerful closed-source models and open-source models, but a large model size is needed. The average accuracy of GPT-4V is more than 5 times larger than the best-performing open-source model (Yi-VL-34B). But Gemini-Pro-Vision, being closed-source models of the second-tier size, is much less compatible on complicated tasks such as OlympiadBench, for it achieves an average accuracy that is only slightly higher than open-source model.

The challenge lies more on question-with-images, Physics and none-English text. The model performance on text-only questions is significantly above average, showing the challenging spirit of multi-modal questions. Meanwhile, Physics questions, especially Physics questions with images, are more challenging than math questions, as they require knowledge of the laws of Physics as well as other world knowledge besides mathematical abilities such as calculation and reasoning. Moreover, LMMs with a focus on bilingual image-text training data, such as Qwen-VL-Max and Yi-VL-34B, perform better on Chinese questions then English questions.

Open source LLMs is catching at fast speed in the area of maths and physics. Although with a relatively small size, DeepSeekMath-7B-RL outperforms or is on par with Gemini-Pro-Vision and Qwen-VL-Max on the text-only part of OlympiadBench, especially in Math problems, showing promising future of open-source model of pre-training and fine-tuning on fine-grained mathematical and reasoning data.

Multi-modal training slightly hurts performance on text-only math and physics tasks, but may also bring some improvement. The text-only version GPT-4 performs slightly better on all datasets of OlympiadBench, except for the **En_COMP** dataset. We hypothesis that the improvement in the **En_COMP** dataset shows an enhancement of long-context text reasoning capabilities, which is discussed in Appendix B.2.

5 Analysis

In this section, we conduct analysis on the GPT-4V's answers of specific open-ended questions that have been sampled, as well as giving preliminary examination of theorem proving questions.

5.1 Examination of Theorem Proving Questions

For GPT-4V, we do manual sampling check to evaluate the mathematical theorem proving questions. In the questions drawn according to the knowledge point distribution, GPT-4V only answers 6 out of 81 questions correctly in Math-Zh_COMP, all of which are relatively simple and classic conclusions (e.g. AM-GM inequality), or involved only simple computational derivations, and was basically unable to complete the proof within the token limitation in Math-En_COMP, indicating that existing models still cannot effectively solve lengthy reasoning and proofs, which is consistent with the conclusions in existing papers (Trinh et al., 2024).

In solving proof problems, GPT-4V exposes several important issues, including: inability to fully utilize image information (figure 6 as an example); tending to make mistakes in simplifying and transforming algebraic expressions; proposing simple, basic incorrect conclusions;struggling with classification discussions, etc. Detailed examples can be found in the Appendix C.

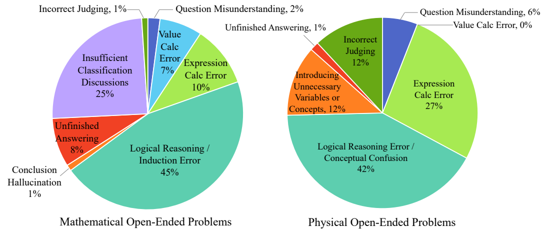

Figure 4: Distribution of the error occurring in GPT-4V's solving process of 164 sampled Olympic-level open-ended problems.

5.2 Mistake Analysis of GPT-4V

We manually sample and check 97 maths (55 for English and 42 for Chinese) and 67 physics Olympics-level open-ended problems that GPT-4V fails, and analyze the type of mistakes, the overall results are shown in figure 4. In maths problems, the typical errors of GPT-4V include: insufficient classification discussion, especially in combinatorial problems; poor performance in problems requiring large calculations (e.g. conic curve problems), manifests as a lack of logic in the calculation process, resulting in the model being unable to provide a reasonable answer. However, we also found that GPT-4V has strong abilities in solving quadratic equations and derivative problems. In physics problems, GPT-4V tends to fall in conceptual confusion, or introduce unnecessary variables or concepts, but its capability to simplify and transform algebraic expressions is stronger than in purely mathematical situations, with nearly no numerical calculation errors.

6 Conclusion

We create OlympiadBench, an Olympiad-level bilingual multimodal scientific benchmark to assess the capabilities of large models in mathematics and physics reasoning. Each problem is detailed with expert-level annotations for step-by-step reasoning. In our benchmarking, we provide a detailed analysis of model performance, pinpointing prevalent error types and potential areas for enhancement. This significant and challenging effort fills a notable void, and we intend to open-source the benchmark to advance AGI and scientific reasoning research. Future efforts will focus on gathering more challenging questions and broadening the range of subjects to further develop rigorous scientific benchmarks.

Ethical Considerations

In this paper, we introduce OlympiadBench, a highly challenging bilingual, multimodal scientific benchmark aimed at evaluating the mathematical and physical reasoning of large models now and AGI in the future. The paper outlines the dataset construction, including data gathering, OCR processing, cleansing, deduplication, and detailed annotation. OlympiadBench's data, derived exclusively from official sources, substantially reduces the likelihood of pre-training data leakage. We offer precise annotations for each problem and have implemented an exhaustive evaluation script for more accurate model performance assessment. Additionally, being bilingual and providing expert-level reasoning annotations for every question, OlympiadBench serves as a crucial resource for propelling AGI's prowess in scientific reasoning. Committed to environmental sustainability, we intend to release the dataset and accompanying scripts publicly to cut down on unnecessary carbon footprint. In experiments, we comply with all licenses for models and data.

Limitations

In pursuit of understanding the logical reasoning abilities of LLMs and LMMs within the multimodal domains of mathematics and physics, we develop OlympiadBench, a challenging bilingual multimodal scientific benchmark. Despite filling a notable void, this work acknowledges inherent limitations. First, in the OlympiadBench, some questions feature answers that require categorical discussions or textual descriptions, such as proofs, which currently cannot be assessed using regular expressions or tools like SymPy at the code level and necessitate manual review. However, this data holds significant research value. Secondly, the automated scoring system we propose cannot perform specific analysis based on the particulars of each question. It makes logical judgments solely based on the two symbols or numerical expressions inputted, without integrating any special constraints that may exist within the actual problem context. What's more, the development of datasets for multimodal scientific reasoning requires extensive manual effort in gathering and annotating data, which constrains the diversity and difficulty of multimodal scientific challenges. As a result, this hampers AI's capacity to learn from and address more intricate scenarios.

References

01-ai. 2023. Yi-34b-chat model card.

01-ai. 2024. Yi-vl-34b model card.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Daman Arora, Himanshu Gaurav Singh, and Mausam. 2023. Have llms advanced enough? a challenging problem solving benchmark for large language models.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Daniel G. Bobrow. 1964. Natural language input for a computer problem solving system. Technical report, USA.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Jie Cao and Jing Xiao. 2022. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3313–3323, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, Online. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias

Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Katherine M Collins, Albert Q Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B Tenenbaum, William Hart, et al. 2023. Evaluating language models for mathematics through interactions. *arXiv preprint arXiv:2306.01694*.

Simon Frieder, Luca Pinchetti, Alexis Chevalier, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, and Julius Berner. 2023. Mathematical capabilities of chatgpt.

Zheqi He, Xinya Wu, Pengfei Zhou, Richeng Xuan, Guang Liu, Xi Yang, Qiannan Zhu, and Hua Huang. 2024. Cmmu: A benchmark for chinese multi-modal multi-type question understanding and reasoning.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models.

Albert Q. Jiang, Sean Welleck, Jin Peng Zhou, Wenda Li, Jiacheng Liu, Mateja Jamnik, Timothée Lacroix, Yuhuai Wu, and Guillaume Lample. 2023. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. MAWPS: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157, San Diego, California. Association for Computational Linguistics.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786, Online. Association for Computational Linguistics.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023. A survey of deep learning for mathematical reasoning.

Ha-Thanh Nguyen. 2023. A brief report on lawgpt 1.0: A virtual legal assistant based on gpt-3. *arXiv preprint arXiv:2302.05729*.

NousResearch. 2023. Nous-hermes-2-yi-34b model card.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. Gpt-4v(ision) system card.

Jielin Qiu, William Han, Jiacheng Zhu, Mengdi Xu, Michael Rosenberg, Emerson Liu, Douglas Weber, and Ding Zhao. 2023. Transfer knowledge from natural language to electrocardiography: Can we detect cardiovascular disease through language models? *arXiv preprint arXiv:2301.09017*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2023. Scieval: A multi-level large language model evaluation benchmark for scientific research.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024. Scibench: Evaluating college-level scientific problem-solving abilities of large language models.

Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 845–854, Copenhagen, Denmark. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. 2023. Cmath: Can your language model pass chinese elementary school math test?

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hannaneh Hajishirzi, Yejin Choi, and Kyunghyun Cho. 2021. Naturalproofs: Mathematical theorem proving in natural language.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.

Daoguang Zan, Bei Chen, Fengji Zhang, Dianjie Lu, Bingchao Wu, Bei Guan, Wang Yongji, and Jian-Guang Lou. 2023. Large language models meet NL2Code: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7443–7464, Toronto, Canada. Association for Computational Linguistics.

Cedegao E Zhang, Katherine M Collins, Adrian Weller, and Joshua B Tenenbaum. 2023. Ai for mathematics: A cognitive science perspective. *arXiv preprint arXiv:2310.13021*.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, Haoran Zhang, Xingwei Qu, Junjie Wang, Ruibin Yuan, Yizhi Li, Zekun Wang, Yudong Liu, Yu-Hsuan Tsai, Fengji Zhang, Chenghua Lin, Wenhao Huang, Wenhu Chen, and Jie Fu. 2024b. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2022. Minif2f: a cross-system benchmark for formal olympiad-level mathematics.

Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2023. Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification.

A Dataset Details

A.1 Data Sources

Our data sources can be split into the following three parts:

1. **Global Mathematics and Physics Olympiad Problems.** The Mathematics and Physics Olympiad problems are globally recognized for their complexity and quality. These problems often require multiple methods of solution and the ability to integrate sub-disciplines from within the broader fields of mathematics and physics. The participants in these competitions represent some of the most proficient individuals worldwide in logical reasoning within mathematics and physics. This not only sets a high standard for problem-solving but also fosters a diverse set of analytical skills that are crucial for the advancement of large models.

2. **Regional and National Chinese Mathematics Competitions.** In addition to maintaining

Subject	Source	Coverage Years	Number
Maths	IMO	2006-2022	509
	RMM	2011, 2013, 2015-2019, 2021, 2023	53
	ARML	2009-2014, 2019, 2023	505
	EMC	1998-2023	364
	EGMO	2013, 2015-2023	64
Physics	IPhO	1984, 1986-1990, 2008-2012, 2014-2016, 2018-2019, 2021	381
	APhO	2013-2015	200
	EPhO	2019-2022	17
	USAPhO	2017-2021	113
	PUPC	2020-2022	65
	OPhO	2020-2023	132

Table 5: Summary of Problems in Maths and Physics Competitions, the full name of each acronym is given in Table 6

a high level of difficulty, regional competitions and the CMO introduce elements specific to the Chinese context. This inclusion is instrumental in furthering the development and research of Chinese-oriented and multilingual large models. By encompassing a wide array of mathematics and physics problems, these competitions provide a unique opportunity to develop models that are adaptable and proficient across different mathematical queries, enhancing their versatility and effectiveness.

3. **Gaokao Mock Questions for Mathematics and Physics [11].** Given that the resolution of Olympiad-level problems typically necessitates models with substantial parameter sizes, we also incorporate Gaokao simulation problems to evaluate smaller models' capabilities in answering free-form mathematics and physics questions.

The integration of data from Gaokao simulation problems, regional and national competitions, to the global Olympiads constructs a smooth difficulty transition curve. This methodology not only distinguishes the mathematical and physical problem-solving capabilities of different models but also provides guidance on the scaling laws for models specialized in these domains.

A.2 Data Curation Process

Our initial step involves a comprehensive survey of well-known Olympiad competitions, and the list of which is accessible through the AoPS community platform [12]. We cataloged these competitions based on several criteria: difficulty level, volume of questions, availability of materials in public domains, language, discipline, and coverage years. Following the design principles outlined in Section 3.1, we meticulously select specific contests and years that not only adhere to our dataset design criteria but also try to span the widest possible range of years (Table 5).

In the format conversion phase, we also manually annotated the subfield of each question in maths or physics, with their distribution detailed in Table 9.

A.3 Example of Progressive Problem in Physics

Figures 7 and 8 present a sequential challenge from the International Physics Olympiad (IPhO) 2021, illustrating the intricacies of progressive problem-solving in a competitive context. This particular problem set exemplifies a common trait in advanced physics competitions: the dependency of many questions on the solutions and materials of preceding ones. These dependencies are sometimes explicit, but most are implicit, weaving a complex web of interconnected knowledge and reasoning.

An explicit instance of this dependency can be observed in problem C.2, where the prompt directly requires the use of the symbol β defined in B.1 for the calculation of an unknown quantity. This requirement not only tests the participants' ability to understand and apply physical concepts but also assesses their skill in navigating through and linking

[11]Due to anonymity of this submission, we can not add concrete source in this version.

[12]https://artofproblemsolving.com/community/c13

various parts of a problem set. Such explicit instructions are crucial for guiding participants through the logical progression of the problems, yet the majority of dependencies remain implicit, demanding a deeper level of comprehension and integration of the material.

This structure of problem-solving reflects a realistic scientific inquiry, where discoveries and solutions often rely on previously established knowledge. The explicit mention of β in C.2 as derived from B.1 is emblematic of this educational approach, aiming to foster a holistic understanding and the ability to build upon existing information to solve complex problems. It underscores the importance of thorough comprehension of earlier sections for successful problem-solving in later sections, simulating real-world scientific challenges where new solutions are often predicated on a foundation of established knowledge.

B Evaluation Details

B.1 Details of the Evaluated Models

B.1.1 LMMs

We have selected current mainstream LMMs that have performed the best on past scientific multi-modal datasets for evaluation.

The closed-source models include: GPT-4V (OpenAI, 2023b), developed by OpenAI, which is currently the most powerful multimodal model. Gemini (Team, 2023) is the LMM series developed by Google Deepmind, with Gemini-Ultra-Vision being purported to have surpassed GPT-4V on datasets like MMMU. However the unavailability of Google's API for Gemini Ultra, we test the accessible Gemini-Pro-Vision as an alternative. Qwen-VL-Max (Bai et al., 2023), developed by Alibaba, is the largest LMM, and stands on par with GPT-4V and Gemini-Ultra in multi-modal tasks. Due to the large proportion of Chinese data used in its training, Qwen-VL-Max has a certain advantage in Chinese language ability.

The open-source models include: Yi-VL-34B (01-ai, 2024) is the first open-source 34B multi-modal model that has demonstrated satisfying performance on several latest datasets. With Chinese text-image pairs included in the training process, Yi-VL-34B offers adequate multilingual support. LLaVA-NeXT-34B (Liu et al., 2024) claims to be the strongest open-source LMM, with enhancements in reasoning, OCR, and world knowledge. Despite being trained exclusively with English multi-modal data, it demonstrates an emergent zero-shot Chinese multi-modal capability on Chinese benchmarks.

It should be noted that an image must be passed for Gemini-Pro-Vision, LLaVA-NeXT, and Yi-VL during inference. Therefore, for the text-only questions in OlympiadBench dataset, we use the corresponding text-model api (for closed-source models), or their base LLM (for open-source models). To examine the impact of replacing LMM with base LLM for processing text-only data, we subsequently compare the performance differences between GPT-4V and GPT-4 on text-only questions in OlympiadBench.

B.1.2 LLMs

The field of LLM starts early in scientific areas such as mathematics and physics, with models specifically trained occurring. We select DeepSeekMath-7B-RL (Shao et al., 2024) as the primary baseline for text-only questions. DeepSeekMath-7B-RL is pre-trained on 120B math-related data and enhanced chain-of-thought (CoT) reasoning capabilities using reinforcement learning, in the result scoring close to GPT-4 and Gemini-Ultra on the MATH (Hendrycks et al., 2021) dataset. We report the results of the selected LMMs (or their LLM counterparts) on the text-only questions for comparison, and additionally evaluate GPT-4 in order to compare with GPT-4V [13].

B.2 Detailed Experiment Result

The comparison of the performance of mainstream closed-ended models on different datasets are shown in Table 7.

To further discuss the performance difference between GPT-4 and GPT-4V on the Physics-En_COMP, we split the **En_COMP** dataset into two sub-datasets, with **normal-PhO** being normal PhO questions, and **long-PhO** being PhO questions that show in a relational series, therefore having long context. As shown in table 8, GPT-4 keeps performing slightly better on **normal-PhO**, but lags much behind on **long-PhO**, which may indicate improvement of long-context text reasoning capabilities after multimodal training.

[13] The version of GPT-4 and GPT-4V are both "0125-preview".

Subject	Acronym	Full name
Maths	IMO	International Mathematical Olympiad
	RMM	Romanian Master of Mathematics
	ARML	American Regions Mathematics League
	EMC	Euclid Mathematics Competition
	EGMO	European Girls' Mathematical Olympiad
Physics	IPhO	International Physics Olympiad
	APhO	Asian Physics Olympiad
	EPhO	European Physics Olympiad
	USAPhO	USA Physics Olympiad
	PUPC	Princeton University Physics Competition
	OPhO	Online Physics Olympiad

Table 6: Full names of all competitions'acronyms used in this paper

Benchmark	**GPT-4(V)**	**Qwen VL-Max**	**Gemini Pro**
MATH	52.9	-	32.6
MathVista(testmini)	49.9	50.0	45.2
OlympiadBench	17.23	10.31	4.38

Table 7: Comparison of Performance on Different Benchmarks. The values for MATH and MathVista are obtained from Gemini and Qwen's report.

	long-PhO (157)	**normal-PhO (74)**
GPT-4V	18.47	1.35
GPT-4	14.92	4.05

Table 8: Average accuracy of GPT-4V and GPT-4 for the En_COMP dataset

B.3 Unavailable Responses for Closed-Source Models

As described in table 4, the response for some problems are not available, the main causes are as follows:

1. Exceeding input limit: Some of the context of the problems are too long, which exceed the input token limitation for the API. This case mainly occurs in Physics-En_COMP that contains long-context problems of over 6,000 tokens.

2. Inappropriate response: Some problems trigger inappropriate response, which are banned by the API to return.

3. No response: Some problems continuously get no or empty response from the API.

4. Request timed out: Some problems continuously fail to get a response.

We removed the problems with unavailable response when calculating the accuracy.

C Additional Analysis and Examples

C.1 Detailed Description of the Error Types in GPT-4V's solving process

The error types are as follows:

1. Question Misunderstanding

2. Value Calculation Error

3. Expression Calculation Error

4. Logical Reasoning / Induction Error / Conceptual Confusion

5. Introducing Unnecessary variables or concepts

6. Conclusion Hallucination

7. Unfinished Answering: sometimes GPT-4V says the question have confliction in settings, or degenerates after some tokens.

8. Insufficient Classification Discussions

9. Incorrect Judging: Calls for future work of automatically deciding required precision of the answer, or automatically judging expressions such as $a\sqrt{b}$ and $\sqrt{a^2b}$ with $a \geq 0$.

C.2 Other Observation for Open-Ended Problems

1. Given a simple solution, GPT-4V may choose a more complex method to solve the problem (see Figure 5).

2. Models may give correct answers with a false process. Mainly observed for problems with a simple answer, such as the variables takes 0 as the answer.

3. GPT-4V may not fully utilize the information from the image (see Figure 6).

D Automatic Scoring Pipeline

Algorithm 1: Auto Scoring Judge

```
Input: GroundTruth, ModelOutput;
Output: Boolean value indicating match;
Preprocess GroundTruth and ModelOutput;
if GroundTruth equals ModelOutput then
    return True;
else
    if GroundTruth and ModelOutput are intervals or tuples then
        if GroundTruth equals ModelOutput as intervals or tuples then
            return True;
        else
            return False;
        end
    else
        if GroundTruth equals ModelOutput numerically then
            return True;
        else
            if Both GroundTruth and ModelOutput contain "=" then
                return EquationEqual(GroundTruth, ModelOutput);
            else
                return ExpressionEqual(GroundTruth, ModelOutput);
            end
        end
    end
end
```

<table>
<tr><th>Question</th><th>GPT-4V’s Solution</th></tr>
<tr><td>已知正数 x,y 满足: $x+y=1$. 求 $\frac{1}{x}+\frac{4}{y}$的最小值.</td><td rowspan="3">由已知条件 $x+y=1$，我们可以将 $\frac{1}{x}+\frac{4}{y}$表达为 x 和 y 的函数。首先，我们将 y 用 x 表示出来，即 $y=1-x$。然后将 y 代入 $\frac{1}{x}+\frac{4}{y}$中，得到：$\frac{1}{x}+\frac{4}{y}=\frac{1}{x}+\frac{4}{1-x}$
为了找到这个表达式的最小值，我们可以使用微分法。首先，我们对上述表达式求导：$\frac{d}{dx}\left(\frac{1}{x}+\frac{4}{1-x}\right)=\frac{1}{x^2}+\frac{4}{(1-x)^2}$.
为了找到极值点，我们令导数等于零：⋯. 解这个方程，⋯我们得到：$3x^2+2x-1=0$. 这是一个二次方程，我们可以使用求根公式来解它：$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$, 其中$a=3,b=2,c=-1$，代入求根公式得到：$x=\cdots=\frac{-2\pm 4}{6}$. 我们得到两个解，分别是：$x=\frac{1}{3}$或$x=-1$.
由于 x 是正数，我们排除 $x=-1$ 的解，所以 $x=\frac{1}{3}$. 将 x 的值代入 $y=1-x$，得到⋯。
现在我们可以计算 $\frac{1}{x}+\frac{4}{y}$的值：$\frac{1}{x}+\frac{4}{y}=\frac{1}{\frac{1}{3}}+\frac{4}{\frac{2}{3}}=\cdots=9$.
所以最终答案是 9。</td></tr>
<tr><th>Standard Solution</th></tr>
<tr><td>由 $a+b-2\sqrt{ab}=\left(\sqrt{a}-\sqrt{b}\right)^2\geq 0$, 故 $a+b\geq 2\sqrt{ab}$. 所求 $\frac{1}{x}+\frac{4}{y}=\left(\frac{1}{x}+\frac{4}{y}\right)(x+y)=1+4+\frac{y}{x}+\frac{4x}{y}\geq 5+2\sqrt{\frac{y}{x}\cdot\frac{4x}{y}}=9$, 等号在 $x=\frac{1}{3},y=\frac{2}{3}$处取到, 故最小值为 9.</td></tr>
</table>

Figure 5: An example of GPT-4V giving a solution that is much complex than that of the standard solution, although reaching the correct answer eventually.

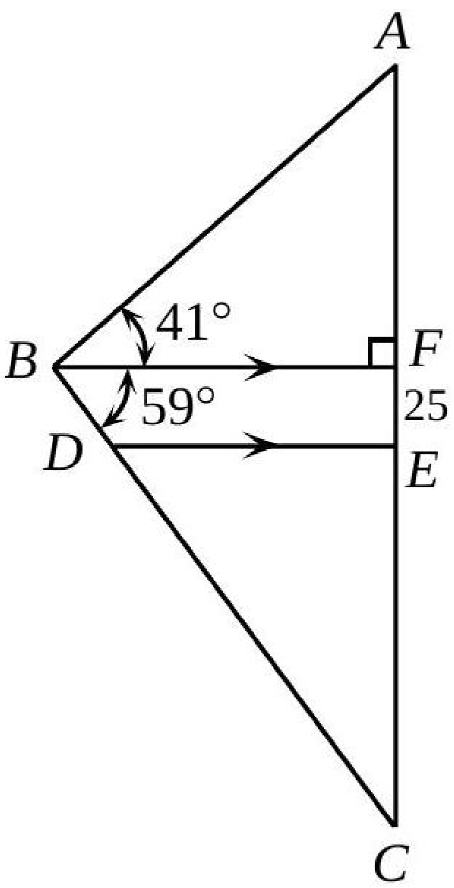

Figure 6: An example of GPT-4V’s failure in utilizing image information from Math-Zh_COMP. GPT-4V starts proving with "we have $\angle DEF=\angle FBC=59°$", which is an error that can evidently be identified from the image, showing insufficient comprehension of the given plane geometry figure.

Subset	Subfield	Number
CEE Math	Derivative	334
	Conic Sections	351
	Sequence	273
	Trigonometric Functions	244
	Set Theory	25
	Logic	3
	Elementary Functions	167
	Inequality	139
	PC&PE	95
	Probability and Statistics	865
	Plane Geometry	831
	Solid Geometry	1375
	Complex Numbers	8
COMP Math	Combinatorics	406
	Algebra	567
	Number Theory	295
	Geometry	544
CEE&COMP Physics	Mechanics	1040
	Electromagnetism	756
	Thermodynamics	257
	Optics	157
	Modern Physics	220

Table 9: Statistics of subfield in Mathematics and Physics. PC&PE stands for Polar Coordinates and Parametric Equations.

Electrostatic lens (10 points)

Consider a uniformly charged metallic ring of radius R and total charge q. The ring is a hollow toroid of thickness $2a \ll R$. This thickness can be neglected in parts A, B, C, and E. The xy plane coincides with the plane of the ring, while the z-axis is perpendicular to it, as shown in Figure 1. In parts A and B you might need to use the formula (Taylor expansion)

$$(1+x)^{\varepsilon} \approx 1 + \varepsilon x + \frac{1}{2}\varepsilon(\varepsilon - 1)x^2, \text{ when } |x| \ll 1.$$

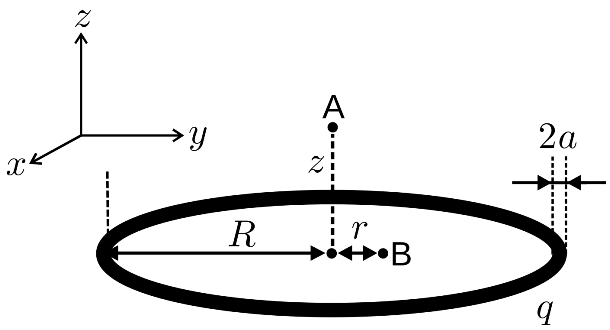

Figure 1. A charged ring of radius *R*.

Part A. Electrostatic potential on the axis of the ring (1 point)

A.1	Calculate the electrostatic potential $\Phi(z)$ along the axis of the ring at a z distance from its center (point A in Figure 1).	0.3pt

A.2	Calculate the electrostatic potential $\Phi(z)$ to the lowest non-zero power of z, assuming $z \ll R$.	0.4pt

A.3	An electron (mass m and charge $-e$) is placed at point A (Figure 1, $z \ll R$). What is the force acting on the electron? Looking at the expression of the force, determine the sign of q so that the resulting motion would correspond to oscillations. The moving electron does not influence the charge distribution on the ring.	0.2pt

A.4	What is the angular frequency ω of such harmonic oscillations?	0.1pt

Part B. Electrostatic potential in the plane of the ring (1.7 points)

In this part of the problem you will have to analyze the potential $\Phi(r)$ in the plane of the ring ($z = 0$) for $r \ll R$ (point B in Figure 1). To the lowest non-zero power of r the electrostatic potential is given by $\Phi(r) \approx q(\alpha + \beta r^2)$.

B.1	Find the expression for β. You might need to use the Taylor expansion formula given above.	1.5pt

Figure 7: An example illustrating the first section of Problem 2 in IPhO 2021.

B.2	An electron is placed at point B (Figure 1, $r \ll R$). What is the force acting on the electron? Looking at the expression of the force, determine the sign of q so that the resulting motion would correspond to harmonic oscillations. The moving electron does not influence the charge distribution on the ring.	0.2pt

Part C. The focal length of the idealized electrostatic lens: instantaneous charging (2.3 points)

One wants to build a device to focus electrons—an electrostatic lens. Let us consider the following construction. The ring is situated perpendicularly to the z-axis, as shown in Figure 2. We have a source that produces on-demand packets of non-relativistic electrons. Kinetic energy of these electrons is $E = mv^2/2$ (v is velocity) and they leave the source at precisely controlled moments. The system is programmed so that the ring is charge-neutral most of the time, but its charge becomes q when electrons are closer than a distance $d/2$ $(d \ll R)$ from the plane of the ring (shaded region in Figure 2, called "active region"). In part C assume that charging and de-charging processes are instantaneous and the electric field "fills the space" instantaneously as well. One can neglect magnetic fields and assume that the velocity of electrons in the z-direction is constant. Moving electrons do not perturb the charge distribution on the ring.

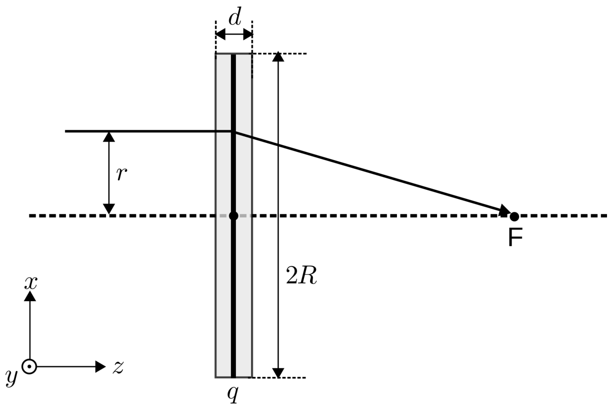

Figure 2. A model of an electrostatic lens.

C.1	Determine the focal length f of this lens. Assume that $f \gg d$. Express your answer in terms of the constant β from question B.1 and other known quantities. Assume that before reaching the "active region" the electron packet is parallel to the z-axis and $r \ll R$. The sign of q is such so that the lens is focusing.	1.3pt

In reality the electron source is placed on the z-axis at a distance $b > f$ from the center of the ring. Consider that electrons are no longer parallel to the z-axis before reaching the "active region", but are emitted from a point source at a range of different angles $\gamma \ll 1$ rad to the z-axis. Electrons are focused in a point situated at a distance c from the center of the ring.

C.2	Find c. Express your answer in terms of the constant β from question B.1 and other known quantities.	0.8pt

Figure 8: An example illustrating the second section of Problem 2 in IPhO 2021.