

# CMMaTH: A Chinese Multi-modal Math Skill Evaluation Benchmark for Foundation Models

Zhong-Zhi Li<sup>1,2\*</sup>, Ming-Liang Zhang<sup>1,2\*</sup>, Pei-jie Wang<sup>1,2</sup>, Jian Xu<sup>1,2</sup>,  
Rui-Song Zhang<sup>1,2</sup>, Fei Yin<sup>1,2</sup>, Zhi-Long Ji<sup>3</sup>, Jin-Feng Bai<sup>3</sup>, Zhen-Ru Pan<sup>3</sup>,  
Jiixin Zhang<sup>4</sup>, Cheng-Lin Liu<sup>1,2†</sup>

School of Artificial Intelligence, University of Chinese Academy of Sciences<sup>1</sup>  
MAIS, Institute of Automation of Chinese Academy of Sciences<sup>2</sup>  
Tomorrow Advancing Life<sup>3</sup>, University of Strathclyde<sup>4</sup>  
{lizhongzhi2022, zhangmingliang2018}@ia.ac.cn, {fyin, liucl}@nlpr.ia.ac.cn,  
{jizhilong, baijinfeng, panzhenru, jiixinzhang625}@tal.com, {jiixinzhang625}@gmail.com

## Abstract

With the rapid advancements in multimodal large language models, evaluating their multimodal mathematical capabilities continues to receive wide attention. Although datasets such as MathVista have been introduced for evaluating mathematical capabilities in multimodal scenarios, there remains a lack of evaluation tools and datasets tailored for fine-grained assessment in Chinese K12 education. To systematically evaluate the ability of multimodal large models to solve Chinese multimodal mathematical problems, we propose a Chinese Multi-modal Math Skill Evaluation Benchmark (CMMaTH), containing 23,856 multimodal K12 math related questions, making it the largest Chinese multimodal mathematical problem benchmark to date. CMMaTH includes questions ranging from elementary to high school levels, offering greater diversity in problem types, solution goals, visual elements, detailed knowledge points, and standard solution annotations. To facilitate stable, fast, and cost-free model evaluation, we have developed an open-source tool called GradeGPT, which is integrated with the CMMaTH dataset. Our data and code are available at <https://github.com/zzli2022/CMMaTH>.

## 1 Introduction

Large language models (LLMs) excel in various language tasks, while Large Multimodal Models (LMMs) effectively handle visual-language problems. They drive advancements in natural language processing and computer vision, offering powerful solutions for complex tasks, demonstrating potential as versatile solvers for multimodal problems.

The systematic evaluation of LLM/LMMs' performance across various mathematical reasoning scenarios has been an extensive research subject. GSM8K and MATH (Cobbe et al., 2021;

Hendrycks et al., 2021b) assessed the ability for multi-step mathematical reasoning by constructing a high-quality set of elementary school math word problems or various competition mathematics problems. By collecting a diverse set of mathematical problems containing both textual and visual components, Lu et al. (2023), Wang et al. (2024), Zhang et al. (2024c) systematically evaluated the ability of LMMs to perceive visual elements and solve corresponding multimodal problems. Shi et al. (2023a) constructed a multilingual mathematical reasoning dataset, MGSM, for evaluating the LLM reasoning ability in multilingual environments. However, in non-English multimodal contexts, especially in Chinese scenarios, there is still a lack of sufficiently detailed and diverse benchmarks for assessing mathematical abilities. To assess the capability of large language models in non-English contexts, Huang et al. (2023) and Zhang et al. (2024b) constructed multidisciplinary Chinese question answering datasets C-Eval and CMMMU to evaluate the knowledge and reasoning abilities of LMMs. However, C-Eval lacks evaluation in multimodal contexts, while CMMMU's dataset has relatively low diversity, consisting of only 540 questions.

Meanwhile, existing datasets contain numerous problem scenarios that are not typically examined in real K12 education. There are inconsistencies between the assessment results and the actual abilities of students in K12 mathematics. Liu et al. (2024b) has introduced a benchmark for mathematics assessment, which includes questions with detailed annotations of textbook knowledge points. The goal is to address these discrepancies in real K12 educational settings. However, this benchmark is limited to text-only scenarios and does not include multimodal questions and knowledge points.

Existing math benchmarks for answer evaluation can be categorized into two types: *rule-based* (Cobbe et al. (2021), Hendrycks et al. (2021b), He

\* Equal Contribution

† Corresponding Author

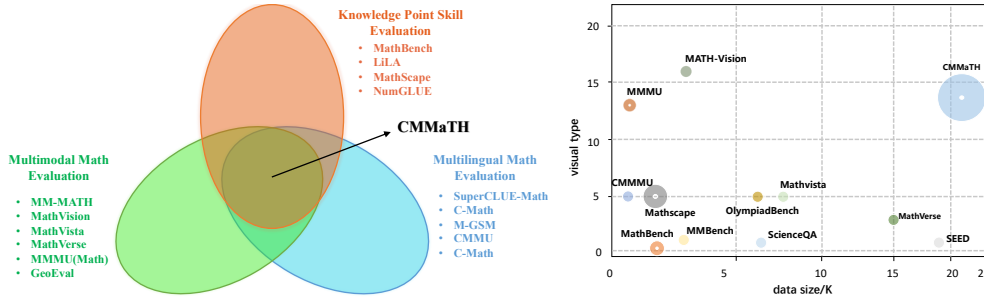


Figure 1: The CMMaTH dataset focuses on the multi-modal mathematical ability assessment of Chinese scenes and has extremely strong diversity and large data scale. CMMaTH combines fine-grained multi-modal mathematical knowledge point annotation. The size of the hollow circle corresponds to the designed number of knowledge points in the dataset.

Dataset	Language Domain	Knowledge Annotation	Knowledge Domain	Knowledge Point Number	Size	Modality	Source	Answer
VQAv2(Goyal et al., 2017)	En	✗	General	–	> 1M	V+T	Annotated	Open/MC/TF
SEED(Li et al., 2023a)	En	✗	General	–	19K	V+T	Annotated	MC
MMBench(Liu et al., 2023b)	En	✗	General	–	3K	V+T	Repurposed	MC
MM-Vet(Yu et al., 2023)	En	✗	General	–	0.2K	V+T	Annotated	Open
ScienceQA(Lu et al., 2022)	En	✗	–	Science	6K	V+T	Textbooks	MC
MMMU(Yue et al., 2023)	En	✗	General	–	11.5K	V(30 Types)+OC	Textbooks	Open/MC
CM3U(Zhang et al., 2024b)	ZH	✗	General	–	< 1K(Math Part)	V(5 Types)+OC	Internet	Open/MC
MathVista(Lu et al., 2023)	ZH/En	✗	Math	–	1K/6K	V(5 Types)+OC	Synthesized	Open/MC/TF
OlympiadBench(He et al., 2024)	ZH/En	✗	Math/Physics	–	6.5K(Math Part)	V(5 Types)	Internet	Open
MathVerse(Zhang et al., 2024c)	ZH/En	✗	Math	–	2.6K/15K	V(3 Types)	Synthesized	MC
MATH-Vision(Wang et al., 2024)	En	✗	Math	–	3K	V(16 Types)+IC	Synthesized	Open/MC
MM-MATH(Sun et al., 2024)	ZH/En	✗	Math	–	5.9K	V(5 Types)	Internet	Open
MathBench(Liu et al., 2024b)	ZH/En	✓	Math	60	3.7K	Text-only	Internet	Open
MathScape(Zhou et al., 2024a)	ZH	✓	K12 Math	107	1.3K	V	Photo-based Paper	Open
CMMaTH	ZH	✓	K12 Math	784	23K	V(14 Types), OC, IC	Internet&Annotated	Open/MC

Table 1: Comparison with other multimodal benchmarks. V: visual input, OC: optical characters caption, T: Question Text, IC: Image Caption, Open: open questions, MC: multiple choice questions, TF: true or false questions.

et al. (2024) and *API-based* methods Lu et al. (2023), Zhang et al. (2024c), Hendrycks et al. (2021a). *API-based* methods leverage the zero-shot in-context learning capabilities of LLM. However, these methods are associated with high costs and significant time consumption, often leading to unstable and inconsistent evaluation outcomes. *Rule-based* methods, aim to derive the correct option from model responses through the design of various extraction systems. Nonetheless, these methods face challenges in handling highly diverse contents of benchmarks. Also, it is difficult to maintain handcrafted rules for dynamically updated benchmarks. Current multimodal math benchmark evaluations often resort to multiple-choice or true/false question formats, using rules or *API-based* LLM to extract options for assessing answers.

Based on the above considerations, we propose a new multimodal mathematical benchmark CMMaTH. Compared to previous benchmarks, our benchmark demonstrates greater diversity and increased depth of reasoning in the Chinese multimodal math context. It also includes finer-grained knowledge annotation to grasp different levels and types of K12 math knowledge. We provide an open-source lightweight answer comparator called GradeGPT, designed to compare the consistency

between outputs from different LLM/LMMs and standard answers, thus avoiding expensive evaluation costs. Leveraging the CMMaTH dataset and GradeGPT tool, we evaluate mainstream open-source and commercial LMMs in Table 3, reporting comprehensive evaluation results along with various and extensive case analyses, and knowledge skill analyses. In summary, our paper makes the following contributions:

- We develop a high-quality multimodal mathematics benchmark specifically tailored for the Chinese language context, featuring detailed knowledge point annotations, extensive quantity, and diversity. It serves as a reference for evaluating the multimodal mathematical reasoning capabilities of foundational models within Chinese language contexts. We also provide an English version of this dataset.
- Compared to previous multimodal mathematical benchmarks, our dataset exhibits a great depth of reasoning and diversity. Our benchmark simulates more realistic educational Q&A scenarios, encompassing a wider variety of question types and answer formats. Additionally, we provide each question with detailed multimodal knowledge points to eval-

uate the mastery level of current large models. The CMMaTH dataset is dynamically maintained and will be periodically updated.

- We build an evaluation assistant named GradeGPT on the CMMaTH dataset, which enables for comparing the proximity of model responses to standard answers and assessing the correctness of results. GradeGPT features lightweight open-source characteristics, avoiding the instability and high costs associated with commercial models.
- We conduct a systematic evaluation of existing mainstream LLM/LMMs, quantitatively and qualitatively comparing with existing models.

## 2 Related Work

### 2.1 Assessment of mathematical abilities

To evaluate the performance of LLM/LMMs in mathematical reasoning and examine hallucinations during the reasoning process, numerous benchmarks (Liu et al., 2023b; Sun et al., 2024; Yue et al., 2023; Yu et al., 2023; Huang et al., 2024b; Bi et al., 2024c) have been proposed for evaluating the mathematical reasoning capabilities of large models. GSM8K (Cobbe et al., 2021) is the first and most widely used mathematical dataset used for large model math evaluation, consisting of 1k math word problem test samples and corresponding answers. The MATH (Hendrycks et al., 2021b) dataset, in comparison to GSM8K, presents a greater challenge in terms of reasoning difficulty. This dataset demands a more profound understanding and intuition in various mathematical domains such as Algebra, Number Theory, and Geometry. MathVista (Lu et al., 2023) is the first dataset used to evaluate the multimodal mathematical capabilities of large models, but it has relatively simple reasoning depth. MATH-VISION (Wang et al., 2024) has richer visual elements and deeper reasoning difficulty. MathVerse (Zhang et al., 2024e) constructs several subsets of datasets to assess whether existing multimodal large models can truly understand mathematical abstract forms. MathBench (Liu et al., 2024b) attempts to assess the level of mastery of specific mathematical skills in existing large models, but this work focuses solely on the pure text domain and annotates a relatively coarse of knowledge points. MM-MATH (Sun et al., 2024) is similar to our work but primarily focuses on English scenarios. The CMMaTH Benchmark, in

comparison to existing works on the evaluation of mathematical proficiency, places a greater emphasis on the analysis of mathematical abilities within the context of the Chinese language. The data distribution of the CMMaTH dataset more closely aligns with the actual distribution found in K12 educational settings, and it provides detailed annotations of mathematical knowledge points to facilitate the assessment of models’ mastery of knowledge and skills.

### 2.2 Large Model Evaluation Tool

Due to their strong generalization capabilities and extensive world knowledge, large language models have achieved outstanding results in tasks such as machine translation (Zhu et al., 2023), question answering (Kamalloo et al., 2023), dialogue (Duan et al., 2023), and so on by generating text. Evaluating the comprehensive abilities of large models, such as clarity, adherence to instructions, comprehensiveness, formality, and mathematical reasoning ability, has received widespread attention (Ke et al., 2023; Mei et al., 2024c,b; Zhou et al., 2024b). Currently, many works opt to use powerful commercial model APIs, such as GPT-4, to assist in evaluating the comprehensive abilities of large models. For instance, in the field of geometric problem solving (Ning et al., 2023; Li et al., 2023c; Zhang et al., 2024d) and multimodal reasoning, MathVista (Lu et al., 2023) and GeoEval (Zhang et al., 2024c) use GPT-4’s API to extract correct answers for evaluation. These methods face several challenges: they are costly and time-consuming, and they struggle to keep up with rapid model iterations. Besides, these methods face challenges in terms of consistency and reproducibility (Wang et al., 2023a; Ke et al., 2023).

Recent methods have proposed using metrics such as BERT score (Zhang et al., 2020) or MAUVE (Pillutla et al., 2021) for evaluation. However, the numerical indicators produced by these methods are difficult to interpret when it comes to the erroneous responses generated by LLM. PandaLM and CritiqueLLM (Wang et al., 2023c; Ke et al., 2023) are similar to our work. They proposed a fine-tuning method based on open-source LLMs, distilling the evaluation capabilities of GPT-3.5 into a series of smaller open-source models. However, they are focused on the automated evaluation of more general text generation tasks, while we are targeting the automated evaluation of responses from large models for multimodal mathe-

Statistic	Number
Total questions	23856
- Choice-mode questions	18191
- Free-form questions	5665
- Questions in the testmini set	1371
Choice-mode questions	18191(76.2%)
- Single-choice questions	13706(57.4%)
- Multiple-choice & Composite questions	4485(18.8%)
Knowledge Point Number	784
Visual Subjects	13
Maximum question length	593
Minimum question length	6
Average question length	75.1
Grade Distribution Elementary(1-6)	800
Junior(7-9)	5082
Senior(10-12)	17972

Table 2: Key statistics of CMMaTH. The unit of question length is words. For more information can refer to Appendix E on the definitions of "Question Difficulty Levels," "Visual Subjects," and "Knowledge Point Number" and "Composite questions"

mathematics problems.

Unlike PandaLM (Wang et al., 2023c) trying to evaluate the relative conciseness, clarity, our evaluation model, GradeGPT, is a dataset-oriented answer comparator that can provide specific reasons based on the standard answer and a model’s response. We distilled the answer comparison capability of GPT-4 using the Cross-Lingual Judge-of-Chain method and enhanced GradeGPT’s answer discrimination ability.

### 3 CMMaTH Dataset

#### 3.1 Overview of CMMaTH

We selected diverse multimodal mathematical problems from a vast pool of K12 educational questions, comprising 23,856 items across 14 visual themes and encompassing 784 types of knowledge points. More detailed statistical data can be found in Table 2 and Appendix E.

#### 3.2 Collection Guidelines

We collected a large number of multimodal mathematics questions from open-source websites, which host a vast collection of K12 math problems. The quality and distribution of the data were guided by the following criteria during collection.

- **Diverse Mathematical Visual Elements.** We have collected solutions to multimodal mathematical problems that rely on understanding image content, especially those containing a

large amount of Chinese visual content such as text and symbols. Table 7 shows some visual elements subject of CMMaTH.

- **High relevance to the K12 math knowledge and skill.** The collector, being knowledgeable in the field, must ensure that each multimodal question targets a specific K12 math concept during the collection process. The dataset primarily includes K12-level math questions, enabling the evaluation of large-scale multimodal models’ potential in mathematics education.
- **High-quality images and answers.** During the collection phase, we instruct collectors to disregard multimodal math questions with erroneous symbols or low-quality images (blurry images). Collectors are required to ensure that the collected questions are generally solvable.

#### 3.3 Data Collections

**Collection from Diverse Multimodal Math Sources** CMMaTH’s is mainly based on “Jiaoyan Yun”<sup>1</sup> and “Zujuan”<sup>2</sup>. These two websites have collected a large number of real K12 education questions with test paper source annotations, grade levels and coarse-grained knowledge point annotation. We purchased the copyright for Jiaoyan Yun’s data and worked with the official website to ensure that it can be open sourced. After preliminary data collection, we compiled about 200,000 preliminary data for each grade. For more data cleaning details, refer to Appendix E.

**Data Filtering** We excluded all questions without images in the question stems, including those non-Chinese language questions, and those solvable without visual content. Due to OCR processing may result in inaccuracies, To ensure the quality of both images and text-based questions, we removed any images with a width or height less than 100 pixels and employed the GPT-4 API to evaluate data quality. For more details on data cleaning, refer to Appendix E.3.

**Data Labeling** We have adopted the current commercial knowledge graph for mathematics education, Jiaoyan Cloud, which has been validated by a large number of users and teachers. We first crawled the knowledge graph involved in "Jiaoyan Cloud", which contains 5531 knowledge points.

<sup>1</sup><https://www.jiaoyanyun.com/>

<sup>2</sup><https://zujuan.xkw.com/>



<p><b>数学统计图推理</b> Statics Map Reason</p> <ul style="list-style-type: none"> <li>☆ 折线图 line chart</li> <li>☆ 样本估计总体 Estimate the population by the sample</li> <li>☆ 频率分布直方图 Frequency square distribution histogram</li> <li>☆ 维恩图 Venn diagram</li> <li>☆ 样本的均值与方差 Sample mean and variance</li> <li>☆ 随机变量及其分布 Random variables and distribution</li> <li>☆ 古典概型与几何概型 Classical and geometrical concepts</li> <li>☆ 茎叶图 Stem and leaf plot</li> <li>☆ 随机抽样 random sampling</li> <li>☆ 众数、中位数、平均数 mode, median, mean</li> <li>☆ 总体密度曲线 overall density curve</li> </ul>	<p><b>抽象图示意图推理</b> Abatract Sketch Map Reason</p> <ul style="list-style-type: none"> <li>☆ 图形染色 Graphic Coloring Problem</li> <li>☆ 程序框图 Block Diagram Understanding</li> <li>☆ 树型图 Tree Diagram</li> <li>☆ 三角形数及衍生数列 Triangular numbers and derivative sequences</li> <li>☆ 周期数列 Periodic Sequence</li> <li>☆ 图形排列与组合 Arrangement And Combination</li> <li>☆ 加法原理与乘法原理 Principle of addition And Multiplication</li> </ul>	<p><b>函数推理</b> Function Reason</p> <ul style="list-style-type: none"> <li>☆ 二次函数性质 Quadratic function properties</li> <li>☆ 反比例函数的性质 Inverse proportional function properties</li> <li>☆ 三角函数性质 Trigonometric properties</li> <li>☆ 圆锥曲线性质 Conic properties</li> <li>☆ 直线方程 Equation of a straight line</li> </ul>
<p><b>几何推理 Geometry Reason</b></p> <ul style="list-style-type: none"> <li>☆ 角的定义 Angle</li> <li>☆ 三视图 Three View</li> <li>☆ 棱台 Properties of prism</li> <li>☆ 投影 Projection</li> <li>☆ 等腰三角形 Isosceles Triangle</li> <li>☆ 菱形 diamond</li> <li>☆ 勾股定理 Pythagorean theorem</li> <li>☆ 射影定理 Photography Theorem</li> <li>☆ 三角形的内切圆性质 Triangle Inscribed Circles</li> <li>☆ 角平分线的定义与性质 Angle Bisectors</li> <li>☆ 垂直平分线的定义与性质 Vertical Bisector</li> <li>☆ 圆内接四边形定理 Circle Inscribed Quadrilateral theorem</li> <li>☆ 相似与全等 Similarity And Congruence</li> <li>☆ 圆周角定理 Circle Angle Theorem</li> </ul>		<p><b>其他 Other</b></p> <ul style="list-style-type: none"> <li>☆ 尺规作图 Rule and compass drawing</li> <li>☆ 杨辉三角的运用 Application of Yang Hui Triangle</li> </ul>

Figure 2: Part of the knowledge points involved in the CMMaTH dataset.

Then, we filtered out the multimodal knowledge points that were not involved, and obtained 784 knowledge points. Although the mathematical problems in "Zujuan" do not have knowledge point annotations that can be directly crawled and are not organized in the way of "Jiaoyan Cloud", we have conducted fine-grained annotation of knowledge points on the problem data from the "Zujuan" section. The math problem is first solved through GPT-4. The GPT-4 assisted annotation manual annotation details can be referred to Appendix E.4.

### 3.4 Comparison with Existing Benchmarks

The CMMaTH dataset is primarily used to evaluate multimodal math reasoning capabilities in K12 Chinese educational scenarios. We compared the current mainstream multimodal mathematical datasets and large model benchmarks in Table 1. Compared to existing multimodal benchmarks and multimodal reasoning benchmarks, the CMMaTH dataset has the following characteristics:

**Tailored for real Chinese K12 Multimodal scenarios** MathVista features a substantial number of problems that are associated with natural and synthetic images. However, these images do not accurately represent the genuine data distribution encountered in K12 mathematics educational settings. OlympiadBench is a bilingual multimodal benchmark at the Olympiad level, but it is too challenging and doesn't align well with real K12 multimodal math scenarios. Additionally, the variety of multimodal visual elements is relatively limited. Compared to comprehensive datasets like MMMU and CMMM, CMMaTH is specifically

designed for Chinese subjects and exhibits significant diversity in multimodal Chinese math problems. Instead, We collect multimodal mathematical question data from real Chinese exam aggregation websites, specifically tailored to the Chinese K12 educational context. As illustrated in Figure 7, the questions in the CMMaTH dataset require comprehensive understanding multimodal elements, including Chinese text and math symbols, as well as mastery of K12 knowledge points.

**High-quality Fine-grained Annotation and Evaluation Tool** Every question in our dataset is meticulously annotated with standardized answers, solutions expressed in natural language, associated multimodal knowledge points, visual element categories, and K12 grade levels. This fine-grained annotation enables a more nuanced evaluation of multimodal mathematical proficiency within the K12 educational context. Despite the fact that numerous benchmarks, such as MathVista and GeoEval, rely on GPT-4 for answer extraction and validation, we introduce an open-source model named GradeGPT. GradeGPT stands out by providing a stable, cost-free, and swift accuracy evaluation specifically tailored for the CMMaTH dataset. Also, this approach makes it easier to dynamically maintain our dataset.

**Extreme Diversity** Currently, high-quality Chinese multimodal mathematics datasets are scarce. MATH-VISION lacks Chinese content, MATH-VISTA has only a few Chinese samples, and CMMU includes just 540 math problems, which are not detailed enough. We have included about 23k

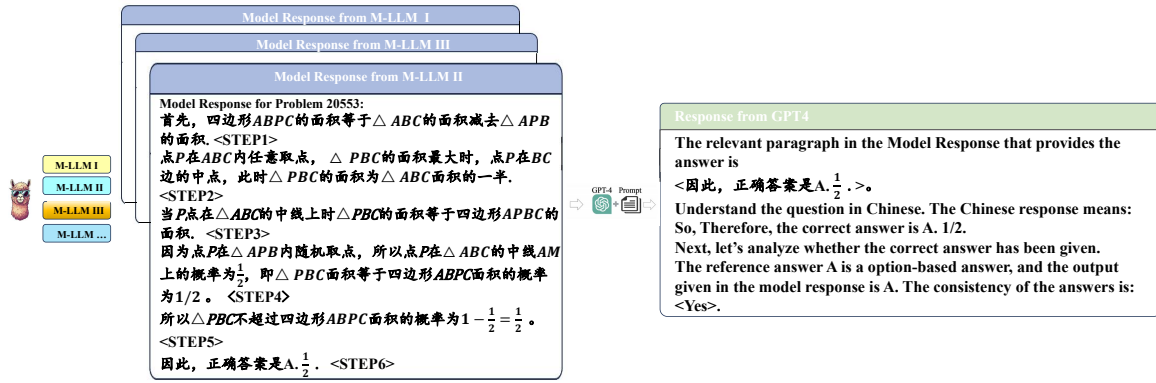


Figure 3: Instruction Construction Pipeline of GradeGPT.

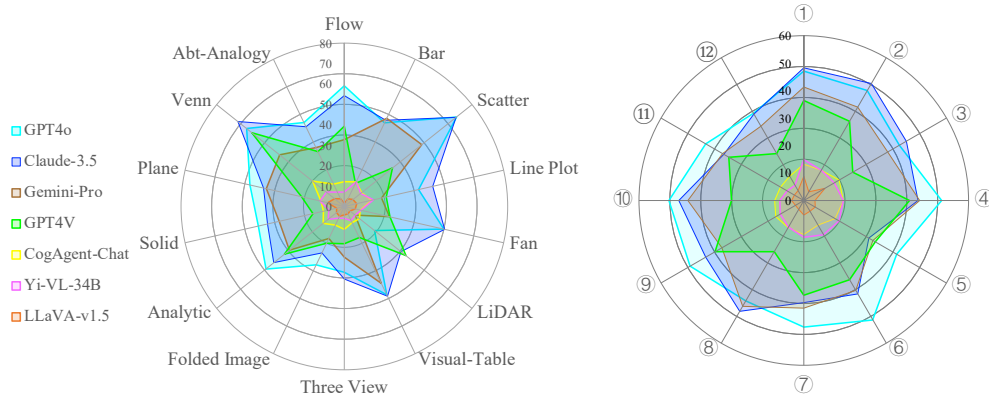


Figure 4: The results of mainstream multimodal large models and pure text large models on the CMMaTH dataset. **Left:** represents the performance evaluation of selected LMMs and LLMs across various Visual Subjects. **Right:** the performance assessment of these models on different educational grade-level questions.

fine-grained multimodal mathematics assessment samples, covering 14 K12 mathematics visual categories, making it the largest known multimodal Chinese dataset to date. The CMMaTH dataset also includes a diverse range of question types, featuring many multi-choice questions, free-form answer questions, and "Composition Questions," as illustrated in Table 2 and Figure 8, which are commonly found in real-world Chinese mathematics education.

## 4 GradeGPT

The CMMaTH dataset encompasses a large variety of problem-solving objectives, such as mathematical expressions, multiple-choice options, numerical outcomes, coordinate points, conclusion figures, and correctness assessments. Traditionally, in reasoning or evaluation contexts, problems have been formulated as multiple-choice or true/false questions to facilitate comparison and to simplify the extraction of results. Also, it is difficult to maintain dynamically updated benchmark. Employing API

models for evaluation is prohibitively expensive, and the resulting evaluations are not consistently stable, which also hampers the iterative development of models on benchmarks, such as hyperparameter selection.

To provide a stable, free, fast, and easy-to-update model response evaluation tool, we introduce GradeGPT, an answer comparison model tailored for the CMMaTH dataset. GradeGPT is designed to receive a question, its standard answers, and a model-generated response. It extracts key steps including results from Chinese output. Determine whether the result is consistent with the standard answer. Our GradeGPT is a streamlined, open-source model. When integrated with frameworks such as vLLM using the 14B model, it can swiftly compare a myriad of model-generated answers, accomplishing a remarkable judgment accuracy of 96.1% for assessing responses comparable with GPT-4 API.

### Prompt Format

In the prompt input of GradeGPT, there are "ques-

Model	Overall	Flow	Bar	Scatter	Line Plot	Fan	LiDAR	Visual-Table	Three View	Folded Image	Analytic	Solid	Plane	Venn	Abt-Analogy
LLMs(Text-only, Zeroshot)															
Baichuan-13B(Yang et al., 2023)	8.4	6.7	4.8	12.2	12.4	13.1	16.2	5.4	4.1	8.5	11.1	6.7	13.7	12.8	9.3
Qwen-14B(Bai et al., 2023)	13.7	15.5	7.3	14.3	16.9	13.6	10.8	11.4	12.8	14.8	15.9	12.7	17.8	20.4	19.3
LLama2-70B(Touvron et al., 2023)	4.5	4.7	2.5	4.4	7.9	7.4	8.1	3.4	5.4	5.1	5.3	4.1	5.3	5.9	4.5
Math LLMs(Text Input, Zeroshot)															
MetaMath-70B(Yu et al., 2024)	5.7	4.6	3.3	6.6	8.7	5.7	0.2	4.2	4.1	8.5	7.2	4.8	8.5	9.8	5.4
DeepSeek-Math-7B(Shao et al., 2024)	14.0	13.4	6.7	14.7	13.1	12.5	12.2	8.1	13.5	12.3	17.2	16.5	21.6	19.5	13.8
InternLM2-Math-20B(Ying et al., 2024)	6.2	4.9	3.4	6.6	9.5	5.7	1.0	4.0	3.9	8.8	8.5	5.3	9.1	9.7	6.0
MAmmoTH2-8x7B(Yue et al., 2024)	16.1	14.1	8.0	15.3	13.4	13.2	12.7	9.4	14.4	12.8	18.0	17.3	21.6	20.3	14.5
Open-source LLMs (Text + Image, Zeroshot)															
LLaVA-v1.5-7B(Liu et al., 2023a)	5.5	1.5	4.2	5.4	6.2	5.4	3.6	4.0	4.2	5.3	4.8	3.9	8.4	6.1	4.2
LLaVA-NEXT-8B(Liu et al., 2024a)	5.4	7.0	2.1	6.8	5.9	1.7	8.1	4.4	6.7	4.2	6.2	4.2	7.7	7.5	6.6
LLaVA-OneVision-7B(Li et al., 2024)	13.7	12.8	5.2	11.4	10.9	9.7	8.1	8.1	13.8	16.5	16.8	13.6	25.2	19.3	13.6
Yi-VL-34B(Young et al., 2024)	8.3	7.1	4.6	10.2	14.6	8.5	6.8	7.7	5.9	6.4	10.1	7.8	12.2	11.3	7.9
CogVLM-18B-Chat(Wang et al., 2023b)	9.4	10.6	4.6	9.5	12.0	7.5	8.4	8.1	10.2	9.7	12.1	9.2	10.2	19.0	10.8
Qwen2-VL-7B(team, 2024)	10.0	10.3	3.5	10.4	6.5	13.6	1.4	7.5	10.2	11.0	10.6	6.8	20.0	13.0	11.8
CogAgent-18B-Chat(Hong et al., 2023)	10.6	12.2	5.2	10.8	13.7	8.0	9.5	8.8	11.2	10.2	13.2	10.5	11.8	19.9	12.2
InternLM-XComposer2-VL(Dong et al., 2024)	3.4	3.3	5.3	3.2	6.2	11.3	6.2	5.4	4.0	0.5	0.4	3.6	1.5	1.8	3.6
InternVL2-8B(Chen et al., 2024b)	23.9	44.6	11.0	22.4	22.6	29.0	14.9	17.4	22.3	14.4	25.0	19.4	30.5	25.1	19.7
Closed-source LLMs (Text + Image, Zeroshot)															
GPT4V(OpenAI, 2023)	27.0	39.3	12.5	30.2	21.0	22.9	38.6	16.9	18.3	20.0	37.5	15.8	21.5	58.0	29.9
Gemini-Pro(Anil et al., 2023)	35.7	58.9	21.1	47.1	31.2	50.6	32.4	27.4	20.3	37.3	37.3	29.5	42.2	53.8	31.5
Claude-3.5(Antropic, 2024)	37.4	63.4	20.9	56.6	44.3	60.2	35.1	31.7	30.8	21.6	37.6	29.1	37.7	59.6	38.8
GPT4o(OpenAI, 2024)	47.8	59.1	45.5	70.3	37.3	50.0	18.9	48.0	32.2	31.8	49.1	43.2	47.4	61.2	45.6
Heuristics baselines															
Random Guess	14.4	13.3	7.9	15.4	13.5	12.4	12.8	7.9	13.2	12.3	17.2	16.4	21.8	19.6	14.3
Frequent Guess	15.1	14.6	7.6	16.0	14.5	13.8	13.5	9.2	15.2	13.8	17.8	17.3	22.6	20.1	14.3
Human Performance															
Human (testmini)	80.1	73.7	78.9	96.2	95.1	57.4	91.7	83.5	69.2	63.2	67.5	51.6	72.1	89.1	83.1

Table 3: Comparison of model performances across various mathematical visual subjects. Visual subjects: Flow: Flow Chart, Bar: Bar Chart, Scatter: Scatter Chart, Line Plot: Line Curve and Plot, Fan: Fan Chart, LiDAR: LiDAR Chart, Visual-Table: Visual-Table Chart, Three View: Three View Graph, Folded Image: Folded Image Graph, Analytic: Analytic Geometry Problem, Solid: Solid Geometry Problem, Plane: Plane Geometry Problem, SolG: Venn: Set Venn Graph, Abt-Analogy: Abstract Analogy Graph. The first and second highest accuracy of LLMs are marked in red and blue, respectively.

tions," "reference answers," and "model output answers." The model is required to provide an answer in the form of "<Yes>" or "<No>" indicating whether the model output answer is equivalent to the standard reference answer. We have designed an instruction format named Cross-Lingual-Judge-of-Chain(CL-JoC) for the purpose of determining answer consistency. CL-JoC first analyzes the model response and finds the key sentences that give the answer in the model response, understand key chinese sentences in English. Then analyze the standard answer, determine the type of the standard answer, and then determine whether the standard answer is included in the model response. More details can be found in Appendix G.

### Instruction Construction

We first generate inference results on CMMaTH using multiple Multimodal LLMs and provide GPT-4 with a detailed few-shot prompt to synthesize answer judgments in the form of a Cross-Lingual Judge-of-Chain response. By employing GPT-4's In-Context Learning, as shown in Figure 3, we have established a procedure for synthesizing instruction data and have produced approximately 56k cross-lingual result judge instruction pairs. Through fine-tuning the model with these instructions, we obtained an expert model, GradeGPT, which possesses the capability to compare answers.

## 5 Experiments

We conducted a comprehensive series of experiments to evaluate various models on the CM-MaTH dataset. Specifically, we assessed multiple LLM/LMMs, including 15 open-source models and 4 API-based closed-source models. We also evaluated the performance variations of different models under conditions augmented with auxiliary information, such as OCR Caption result. Additionally, we investigated the effectiveness of cross-lingual reasoning techniques in enhancing the multimodal mathematical capabilities of LLMs in the Chinese multimodal context. Through systematic experimental design and data analysis, our objective was to elucidate the strengths and weaknesses of these models in handling complex Chinese multimodal contexts. Further details on experiments related to the dependency on visual elements within the CMMaTH dataset, as well as the implementation of contextual learning capabilities in LLMs and evaluation details can be found in Appendix 6.

We also conducted a detailed analysis and evaluation of GPT4o on a random miniset CMMaTH, categorizing errors into four types: Perceptual Errors, Reasoning errors, Calculation errors, and Reject Errors. The error type distribution of GPT-4o on CMMaTH is shown in Figure 6. A more detailed definition of hallucination types can be found in

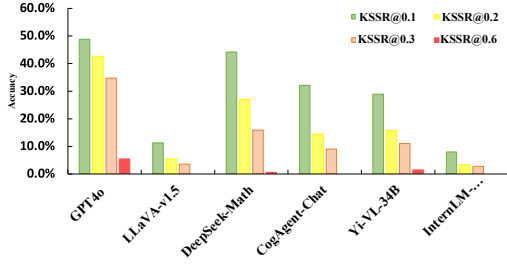


Figure 5: The metrics of different LMM/LLM models about KSSR.

Appendix F.

## 5.1 Main Experiments on LLM/LMMs

We evaluated the results of mainstream multimodal large models and mathematical expert models in Table 3. We analyzed the trend of existing large models in descending with problems and conditions, as well as the effectiveness of techniques such as Cross-Lingual Prompting in solving Chinese multimodal mathematical problems. The experimental results indicate that our data exhibits extremely strong diversity and relatively challenging reasoning depth. Figure 4 and Table 3 show models such as GPT-4o struggle to comprehend our multimodal content and reasoning questions effectively, resulting in significant performance gaps between open-source and proprietary models.

## 5.2 Knowledge Skill Analysis

**Quantitative Analysis** We have formulated a **Knowledge Successful Solve Rate(KSSR)** as a structural metric to gauge the proficiency level of multi-modal extensive models in mastering knowledge points.  $N_{kn}$  is the total number of knowledge points of CMMaTH.  $Acc_{kn_i}$  is the proportion of correct answers to questions labeled as  $i$ 'th knowledge point.  $I$  denotes an indicator function.

$$KSSR@α = \frac{\sum_{i=1}^{N_{kn}} I(Acc_{kn_i} > α)}{N_{kn}} \quad (1)$$

We contend that a knowledge point can be deemed comprehensively understood only when the accuracy rate of solving problems related to that knowledge point surpasses a predefined threshold, denoted as  $α$ . For our investigation, we have established  $α$  at the values of 0.1, 0.2, 0.3, and 0.6 to demarcate the levels of mastery. As shown in Figure 5, our experiment showed that when subjected to a more stringent KSSR metric standard, the most advanced models performed poorly.

**Qualitative Analysis** Based on the fine-grained

LMM	Overall-Acc
LLaVA-v1.5	5.5
InternLM-XComposer2-VL	3.4
Gemini-1.5-Pro	37.2
LLaVA-v1.5 + En-CoT	9.4(+3.9)
InternLM-XComposer2-VL + En-CoT	16.9(+13.5)
Gemini-1.5-Pro + En-CoT	41.1(+3.9)
LLaVA-v1.5 + CLP	12.7(+7.2)
InternLM-XComposer2-VL + CLP	17.1(+13.7)
Gemini-1.5-Pro + CLP	43.8(+6.6)

Table 4: The performance of train-free CoT reasoning techniques on the CMMaTH dataset.

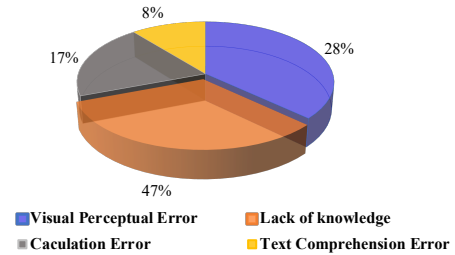


Figure 6: Distribution of Error Types in GPT-4o.

knowledge point annotation, we conducted a detailed knowledge point skill level analysis of current LMMs. We ranked the knowledge points based on their respective problem-solving rates in Appendix B. Our analysis revealed the significant gap between commercial models and open-source models, as well as the distribution of the multimodal knowledge points mastered.

## 5.3 Experiments of Cross-language Reasoning Technology

We also attempted several multilingual Chain-of-Thought approaches such as En-CoT, CLP(Cross-Lingual Prompting) used by Qin et al. (2023) to observe whether multimodal mathematical problems could be enhanced through context learning techniques without training. The results indicate that multilingual CoT methods face challenges in solving, possibly due to the abundance of Chinese contextual text in the image content, which may necessitate the model to demonstrate excellent cross-lingual OCR capabilities. We have included more details on the implementation of Cross-Lingual Prompting and En-CoT on the CMMaTH dataset in Table 4. The prompt implementation detail can refer to Appendix H.3.

## 6 Conclusions

We introduce CMMaTH, a detailed Chinese K12 multimodal reasoning benchmark with diverse question types, visual elements, and complex rea-



son depth. The benchmark includes detailed knowledge points, standard solutions, and grade levels to measure the mastery of knowledge points in the K12 multimodal math skill. To evaluate large multimodal models quickly and affordably, we built GradeGPT, an open-source tool for assessing results. Extensive experimental results on CMMaTH manifest the limitations of current models.

## Limitation & Potential Impact

Our dataset CMMaTH, as a multimodal mathematics dataset aimed at the K12 education sector, can facilitate model evaluation and iteration of multimodal large models in this field, and may promote the research and development of educational artificial intelligence. GradeGPT is a result-oriented, relatively coarse reasoning response evaluator. How to construct a process evaluation model for fine-grained assessment of the reasoning ability of large models remains an area to be explored in the future.

## Acknowledgement

This work has been supported by the National Key Research and Development Program Grant 2020AAA0109700, and the National Natural Science Foundation of China (NSFC) Grant U23B2029.

## References

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. [Gemini: A family of highly capable multimodal models](#). *CoRR*, abs/2312.11805.

Anthropic. 2024. [Claude-3.5](#).

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin,

Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. [Qwen technical report](#). *CoRR*, abs/2309.16609.

Baolong Bi, Shenghua Liu, Lingrui Mei, Yiwei Wang, Pengliang Ji, and Xueqi Cheng. 2024a. [Decoding by contrasting knowledge: Enhancing llms' confidence on edited facts](#). *arXiv preprint arXiv:2405.11613*.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Junfeng Fang, Hongcheng Gao, Shiyu Ni, and Xueqi Cheng. 2024b. [Is factuality enhancement a free lunch for llms? better factuality can lead to worse context-faithfulness](#). *Authorea Preprints*.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Hongcheng Gao, Junfeng Fang, and Xueqi Cheng. 2024c. [Struedit: Structured outputs enable the fast and accurate knowledge editing for large language models](#). *arXiv preprint arXiv:2409.10132*.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Hongcheng Gao, Yilong Xu, and Xueqi Cheng. 2024d. [Adaptive token biaser: Knowledge editing via biasing key entities](#). *arXiv preprint arXiv:2406.12468*.

Yi Chen, Jian Xu, Xu-Yao Zhang, Wen-Zhuo Liu, Yang-Yang Liu, and Cheng-Lin Liu. 2024a. [Recoverable compression: A multimodal vision token recovery mechanism guided by text information](#). *CoRR*, abs/2409.01179.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. [How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites](#). *arXiv preprint arXiv:2404.16821*.

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](#). *CoRR*, abs/2110.14168.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. [Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model](#). *Preprint*, arXiv:2401.16420.

- Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2023. [Botchat: Evaluating llms' capabilities of having multi-turn dialogues](#). *CoRR*, abs/2310.13650.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. [Making the V in VQA matter: Elevating the role of image understanding in visual question answering](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. [Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems](#). *CoRR*, abs/2402.14008.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. [Measuring massive multitask language understanding](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. [Measuring mathematical problem solving with the MATH dataset](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.
- Wenyi Hong, Weihang Wang, Qingsong Lv, Jiazhen Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. [Cogagent: A visual language model for GUI agents](#). *CoRR*, abs/2312.08914.
- Yiming Huang, Xiao Liu, Yeyun Gong, Zhibin Gou, Yelong Shen, Nan Duan, and Weizhu Chen. 2024a. [Key-point-driven data synthesis with its enhancement on mathematical reasoning](#). *CoRR*, abs/2403.02333.
- Yiming Huang, Jianwen Luo, Yan Yu, Yitong Zhang, Fangyu Lei, Yifan Wei, Shizhu He, Lifu Huang, Xiao Liu, Jun Zhao, and Kang Liu. 2024b. [DA-code: Agent data science code generation benchmark for large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13487–13521, Miami, Florida, USA. Association for Computational Linguistics.
- 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](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, et al. 2024a. [Exploring concept depth: How large language models acquire knowledge at different layers?](#) *arXiv preprint arXiv:2404.07066*.
- Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024b. [The impact of reasoning step length on large language models](#). *arXiv preprint arXiv:2401.04925*.
- Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. 2023. [Evaluating open-domain question answering in the era of large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5591–5606. Association for Computational Linguistics.
- Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2023. [Critiquellm: Scaling llm-as-critic for effective and explainable evaluation of large language model generation](#). *CoRR*, abs/2311.18702.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. [Llava-onevision: Easy visual task transfer](#). *Preprint*, arXiv:2408.03326.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. [Seed-bench: Benchmarking multimodal llms with generative comprehension](#). *CoRR*, abs/2307.16125.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. [Evaluating object hallucination in large vision-language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.
- Zhong-Zhi Li, Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. 2023c. [Lans: A layout-aware neural solver for plane geometry problem](#). *arXiv preprint arXiv:2311.16476*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. [Improved baselines with visual instruction tuning](#). *CoRR*, abs/2310.03744.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. [Llava-next: Improved reasoning, ocr, and world knowledge](#).

- Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024b. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*.
- Wentao Liu, Qianjun Pan, Yi Zhang, Zhuo Liu, Ji Wu, Jie Zhou, Aimin Zhou, Qin Chen, Bo Jiang, and Liang He. 2024c. Cmm-math: A chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models. *Preprint*, arXiv:2409.02834.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023b. Mmbench: Is your multi-modal model an all-around player? *CoRR*, abs/2307.06281.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *CoRR*, abs/2310.02255.
- Pan Lu, Swaroop Mishra, Tanglin 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 *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, and Xueqi Chen. 2024a. Slang: New concept comprehension of large language models. *arXiv preprint arXiv:2401.12585*.
- Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Jiayi Mao, and Xueqi Cheng. 2024b. "not aligned" is not "malicious": Being careful about hallucinations of large language models' jailbreak. *arXiv preprint arXiv:2406.11668*.
- Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Ruibin Yuan, and Xueqi Cheng. 2024c. Hiddenguard: Fine-grained safe generation with specialized representation router. *arXiv preprint arXiv:2410.02684*.
- Maizhen Ning, Qiu-Feng Wang, Kaizhu Huang, and Xiaowei Huang. 2023. A symbolic characters aware model for solving geometry problems. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7767–7775.
- OpenAI. 2023. Gpt-4v.
- OpenAI. 2024. Hello gpt-4o.
- Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaïd Harchaoui. 2021. MAUVE: measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 4816–4828.
- Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2695–2709. Association for Computational Linguistics.
- 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. *CoRR*, abs/2402.03300.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023a. Language models are multi-lingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Yongxin Shi, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin. 2023b. Exploring ocr capabilities of gpt-4v(ision) : A quantitative and in-depth evaluation. *Preprint*, arXiv:2310.16809.
- Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. 2024. Mm-math: Advancing multimodal math evaluation with process evaluation and fine-grained classification. *Preprint*, arXiv:2404.05091.
- Qwen team. 2024. Qwen2-vl.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas



- Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *CoRR*, abs/2307.09288.
- Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. [Measuring multimodal mathematical reasoning with math-vision dataset](#). *Preprint*, arXiv:2402.14804.
- Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghui Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. [Large language models are not fair evaluators](#). *CoRR*, abs/2305.17926.
- Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023b. [Cogvlm: Visual expert for pretrained language models](#). *CoRR*, abs/2311.03079.
- Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2023c. [Pandalm: An automatic evaluation benchmark for LLM instruction tuning optimization](#). *CoRR*, abs/2306.05087.
- Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xianrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. [Baichuan 2: Open large-scale language models](#). *CoRR*, abs/2309.10305.
- Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, Yudong Wang, Zijian Wu, Shuaibin Li, Fengzhe Zhou, Hongwei Liu, Songyang Zhang, Wenwei Zhang, Hang Yan, Xipeng Qiu, Jiayu Wang, Kai Chen, and Dahua Lin. 2024. [Internlm-math: Open math large language models toward verifiable reasoning](#). *CoRR*, abs/2402.06332.
- Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. [Yi: Open foundation models by 01.ai](#). *CoRR*, abs/2403.04652.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. [Meta-math: Bootstrap your own mathematical questions for large language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. [Mm-vet: Evaluating large multimodal models for integrated capabilities](#). *CoRR*, abs/2308.02490.
- 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 Wenhua Chen. 2023. [MMM-U: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI](#). *CoRR*, abs/2311.16502.
- Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhua Chen. 2024. [Mammoth2: Scaling instructions from the web](#). *CoRR*, abs/2405.03548.
- 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](#). *arXiv preprint arXiv:2401.13601*.
- 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, Wenhua Chen, and Jie Fu. 2024b. [CMMM-U: A chinese massive multi-discipline multimodal understanding benchmark](#). *CoRR*, abs/2401.11944.
- Jiaxin Zhang, Zhongzhi Li, Mingliang Zhang, Fei Yin, Chenglin Liu, and Yashar Moshfeghi. 2024c. [GeoEval: Benchmark for evaluating llms and multimodal models on geometry problem-solving](#). *CoRR*, abs/2402.10104.
- Ming-Liang Zhang, Zhong-Zhi Li, Fei Yin, Liang Lin, and Cheng-Lin Liu. 2024d. [Fuse, reason and verify: Geometry problem solving with parsed clauses from diagram](#). *arXiv preprint arXiv:2407.07327*.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024e. [Mathverse: Does your multi-modal LLM truly see the diagrams in visual math problems?](#) *CoRR*, abs/2403.14624.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.



Minxuan Zhou, Hao Liang, Tianpeng Li, Zhiyu Wu, Mingan Lin, Linzhuang Sun, Yaqi Zhou, Yan Zhang, Xiaoqin Huang, Yicong Chen, et al. 2024a. [Mathscape: Evaluating mllms in multimodal math scenarios through a hierarchical benchmark](#). *arXiv preprint arXiv:2408.07543*.

Zihao Zhou, Shudong Liu, Maizhen Ning, Wei Liu, Jindong Wang, Derek F Wong, Xiaowei Huang, Qifeng Wang, and Kaizhu Huang. 2024b. [Is your model really a good math reasoner? evaluating mathematical reasoning with checklist](#). *arXiv preprint arXiv:2407.08733*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. [Multilingual machine translation with large language models: Empirical results and analysis](#). *CoRR*, abs/2304.04675.

## A More Related Work

### A.1 Multimodal Large Model Evaluation

The multimodal large models face serious hallucination issues in perceiving objects and executing reason and inference (Zhang et al., 2024a; Mei et al., 2024a; Jin et al., 2024b,a; Shi et al., 2023b). How to evaluate and reduce inference hallucinations of MLLMs (Chen et al., 2024a; Bi et al., 2024a,b) has received widespread attention. To systematically evaluate the various capabilities of multimodal large models, diverse multimodal benchmarks are utilized for assessing the abilities of large models and aiding iterative development. POPE (Li et al., 2023b) is used to evaluate the accuracy of large models in identifying perceptual objects. MMMU and CMMMU (Yue et al., 2023; Zhang et al., 2024b) are comprehensive subject datasets design to assess the proficiency of large models in mastering massive multimodal multi-disciplinary knowledge. SEED-Bench designed 19,000 diverse multimodal questions spanning video and image modalities to evaluate the spatiotemporal capabilities of multimodal large models (Li et al., 2023a). MMVet (Yu et al., 2023) attempts to design datasets to evaluate the integrated capabilities of different multimodal large model systems in combining various Vision-Language skills.

### A.2 Concurrent Work Discussion

MathBench (Liu et al., 2024b) is the first attempt to introduce a fine-grained knowledge point system (Huang et al., 2024a; Bi et al., 2024d) for evaluating large models' mathematical abilities. However, MathBench is a purely text-based mathematical benchmark and does not involve multimodal skills, such as understanding diagrams, and its annotation granularity is relatively coarse.

MathScape (Zhou et al., 2024a) and CMM-Math (Liu et al., 2024c) are concurrent works that were developed after ours, but MathScape only contains 1/10 of the data size and a coarser knowledge point system. MathScape also adopts a method that embeds problem text into images for evaluation, focusing more on end-to-end recognition and solving of problems. CMM-Math provides a dataset for training and evaluation, with around 5k test samples for assessment. In contrast, our CMMaTH dataset is larger in scale, with finer annotation granularity, and is specifically designed for evaluating multimodal mathematical capabilities in Chinese.

## B Analysis of K12 knowledge point mastery level of LMMs

In addition to KSSR, based on the fine-grained knowledge point annotation of CMMaTH, we also clustered several LMMs on the specific level of mastering K12 knowledge points. According to the accuracy of solving the questions corresponding to each knowledge point, for GPT-4o, Gemini-Pro and Claude-3.5, LLaVA-v1.5, we listed the top 20 knowledge points that these models have the best in Figure 16, Figure 17, Figure 18, Figure 19.

At the same time, we identified the K12 knowledge points that these models struggle with and visualized them in Figure 20. Our fine-grained annotation and analysis reveal the shortcomings of current LMMs in solving K12 education problems from the perspective of knowledge points.

Gemini-Pro appears to be optimized for large-scale triangle-related geometry problems but lacks the balanced knowledge and skill level found in models like Claude-3.5 and GPT-4. Additionally, our analysis suggests that open-source models generally struggle with fundamental knowledge skills such as parallelism and positional relationships. We hope that our datasets and tools can promote the data collection and synthesis of existing LMMs from a perspective based on knowledge points.

## C Analysis About Visual/Auxiliary Information for LLMs Inference on CMMaTH

### C.1 The Impact of OCR Information

OCR information includes important information such as Chinese characters on the coordinate axes in mathematical abstract forms, recognized mathematical symbols, etc., and plays a key role in assisting the understanding of visual information. We also evaluated the ability of OCR information for LLM/LMMs to solve CMMaTH Chinese multimodal mathematics questions.

### C.2 The Impact of Visual Content on LMM

In real questions in the real world and academic datasets like MathVista, the text part of many questions contains descriptions of the visual part, which leads to many models that may be able to solve mathematical problems with text reasoning capabilities. In order to evaluate our dataset Regarding the degree of dependence on the visual part, we evaluated the problem-solving capabilities of relatively

Model	Overall	Flow	Bar	Scatter	Line Plot	Fan	LiDAR	Visual-Table	Three View	Folded Image	Analytic	Solid	Plane	Venn	Abt-Analogy
LMMs(Text+Image+OCR Caption, Zeroshot)															
LLaVA-v1.5-7B	5.5	1.5	4.2	5.4	6.2	5.4	3.6	4.0	4.2	5.3	4.8	3.9	8.4	6.1	4.2
Yi-VL-34B	8.3	7.1	4.6	10.2	14.6	8.5	6.8	7.7	5.9	6.4	10.1	7.8	12.2	11.3	7.9
Qwen2-VL-7B	13.7	15.5	7.3	14.3	16.9	13.6	10.8	11.4	12.8	14.8	15.9	12.7	17.8	20.4	19.3
LLaVA-v1.5-7B+OCR Caption	4.9	5.2	1.3	3.4	4.5	5.9	4.4	2.6	3.5	3.9	4.6	3.8	3.7	8.0	5.6
Yi-VL-34B + OCR Caption	10.2	8.3	5.6	11.6	14.8	9.8	7.4	9.2	5.9	6.8	10.8	8.4	13.2	13.0	9.4
Qwen2-VL-7B + OCR Caption	14.1	16.1	8.5	15.1	18.2	14.1	12.3	12.6	13.6	16.3	17.0	14.1	19.3	21.2	20.5
GPT4V	27.0	39.3	12.5	30.2	21.0	22.9	38.6	16.9	18.3	20.0	37.5	15.8	21.5	58.0	29.9
Gemini-Pro	35.7	58.9	21.1	47.1	31.2	50.6	32.4	27.4	24.7	20.3	37.3	29.5	42.2	53.8	31.5
Claude-3.5	37.4	63.4	20.9	56.6	44.3	60.2	35.1	31.7	30.8	21.6	37.6	29.1	37.7	59.6	38.8
GPT4o	47.8	59.1	45.5	70.3	37.3	50.0	18.9	48.0	32.2	31.8	49.1	43.2	47.4	61.2	45.6
Gemini-Pro + OCR Caption	39.7	34.5	50.0	50.9	21.0	25.3	8.5	44.4	26.7	19.5	36.7	35.2	41.0	42.7	33.8
GPT4o + OCR Caption	50.1	61.5	47.7	72.9	39.7	52.6	21.2	50.1	34.8	34.3	51.5	45.0	50.0	63.7	48.0
Claude-3.5 + OCR Caption	46.7	55.6	48.0	70.9	45.9	51.5	36.2	49.8	36.5	25.9	45.1	39.7	41.8	67.8	44.2
LMMs(Text-only, Zeroshot)															
GPT4o-w/o Visual Diagram	17.9	45.0	7.0	39.2	17.4	45.8	19.5	18.6	17.7	10.1	28.4	19.8	19.1	43.2	17.5
Gemini-Pro-w/o Visual Diagram	14.8	25.0	4.3	18.0	7.6	24.4	12.9	14.2	11.3	1.4	18.2	4.9	18.8	31.0	10.0
Claude-3.5-w/o Visual Diagram	19.9	37.0	10.1	32.0	18.1	46.8	17.2	13.4	20.0	7.5	29.8	15.1	20.4	42.2	15.6

Table 5: Model performance of LLMs, Math LLMs, and various commercial API models on CMMaTH with auxiliary OCR caption input. The results of the advanced model removing visual information are also reported in the table. The definitions of different subjects are the same as in Table 3. The first and second highest accuracy of LMMs are marked in red and blue, respectively.

strong interface models, such as GPT-4o, Gemini-Pro, and Claude-3.5, when there is no visual input and only text input. Our results in Table 5 show that GPT-4o and Gemini-Pro and Claude-3.5 suffer huge performance degradation in the absence of visual input. This shows that in addition to understanding the text part, a large number of questions in our benchmark require a full understanding of the corresponding visual elements in order to solve the questions.

### C.3 Few-Shot Evaluation on CMMaTH Dataset

We evaluated the reasoning capabilities of several advanced API-based models under three few-shot conditions(3-shot) in Table 8. The results indicate that few-shot learning can slightly enhance the performance of current commercial-grade API models. However, there remains a significant gap in effectively addressing multimodal mathematical problems in real-world Chinese multimodal reasoning contexts compared to our human evaluation baseline.

## D Model Generation Details

### D.1 Model Weight Version

We have listed the parameter versions and the Hugging Face repository names of the open-source models and version for API-based model used in Table 16.

### D.2 Model Sampling Parameter

We have listed the corresponding hyperparameters used by the models in Table 15. For API models, we have indicated the corresponding release

Statistic	Number
Single choice's distribution of question options	
- Proportion of answers A	2694(14.8%)
- Proportion of answers B	3903(21.4%)
- Proportion of answers C	3961(21.7%)
- Proportion of answers D	3148(17.5%)
Knowledge point number	784
Levels	5
Visual Subjects	13
Maximum question length	593
Minimum question length	6
Average question length	75.1
Grade Distribution Elementary(1-6)	800
Junior(7-9)	5082
Senior(10-12)	17972

Table 6: Key statistics of CMMaTH. The unit of question length is words.

versions. Models using vLLM for inference are annotated.

## E CMMaTH Dataset Details

### E.1 More Detailed Statistics

For readers' reference and understanding of the characteristics of the dataset, we provide other statistical information on our CMMaTH dataset, statistics on options, number of knowledge points, question difficulty level, number of visual element types in Table 6. We have also provided representative examples in Figure 7 and new question type example "Composition Question" from the dataset in Figure 8.

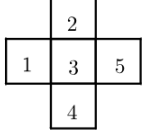
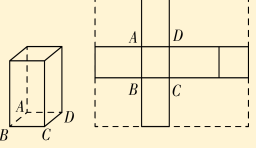
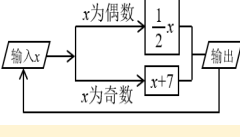
Abstract Analog Graph	Folded Image Graph	Flow chart	Stem-and-Leaf Display								
<p><b>Question:</b> 如图所示, 现有一迷失方向的小青蛙在3处, 它每跳动一次可以等可能地进入相邻的任意一格 (若它在5处, 跳动一次, 只能进入3处, 若在3处, 则跳动一次可以等机会地进入1, 2, 4, 5处), 则它在第三次跳动后首次进入5处的概率是 ( )</p> <p>A. <math>\frac{1}{2}</math>    B. <math>\frac{1}{4}</math>    C. <math>\frac{3}{16}</math>    D. <math>\frac{1}{6}</math></p> 	<p><b>Question:</b> 如图所示, 图2中实线围成的部分是长方体 (图1) 的平面展开图, 其中四边形ABCD是边长为1的正方形, 若向虚线围成的矩形内任意抛掷一质点, 它落在长方体的平面展开图内的概率是 <math>\frac{1}{4}</math>, 则此长方体的体积是_____。</p> 	<p><b>Question:</b> 如下图所示, 这是一个流程图, 在图示的运算程序中, 若开始输入的x的值为100, 我们发现第1次输出的结果为50, 第2次输出的结果为25, 第三次的输出结果为33, 第四次输出结果为40..., 则第2021次输出的结果为_____。</p> 	<p><b>Question:</b> 如图所示的茎叶图记录了甲、乙两组各5名工人某日的产量数据 (单位: 件). 若这两组数据的中位数相等, 且平均值也相等, 那么x和y的值分别为 ( )</p> <p>A. 3,5    B. 5,5    C. 3,7    D. 5,7</p> <table border="1" data-bbox="1077 459 1316 616"> <thead> <tr> <th>甲组</th> <th>乙组</th> </tr> </thead> <tbody> <tr> <td>6</td> <td>5 9</td> </tr> <tr> <td>2 5</td> <td>6 1 7 y</td> </tr> <tr> <td>x 4</td> <td>7 8</td> </tr> </tbody> </table>	甲组	乙组	6	5 9	2 5	6 1 7 y	x 4	7 8
甲组	乙组										
6	5 9										
2 5	6 1 7 y										
x 4	7 8										
<p><b>Answer:</b> C</p> <p><b>Level:</b> High</p>	<p><b>Answer:</b> 3</p> <p><b>Level:</b> High</p>	<p><b>Answer:</b> 8</p> <p><b>Level:</b> Middle</p>	<p><b>Answer:</b> A</p> <p><b>Level:</b> Primary</p>								
<p><b>Knowledge point:</b> 独立事件的概率乘法计算 Probability multiplication of independent events calculate</p>	<p><b>Knowledge point:</b> 计算面积体积相关的几何概率 Calculate geometric probabilities related to area and volume</p>	<p><b>Knowledge point:</b> 规律探索及程序框图 Law exploration and program block diagram</p>	<p><b>Knowledge point:</b> 使用样本估计整体、茎叶图 Use samples to estimate the population, stem and leaf diagrams</p>								

Figure 7: Some samples related to "Folded Image Graph", "Flow chart", "Stem-and-Leaf Display" visual subject on the CMMaTH dataset

## E.2 Data Source Detail

These two electronic data websites, "Jiaoyan Yun"<sup>3</sup> and "Zujuan"<sup>4</sup>, mainly collect electronic exam paper data from primary, middle, and high schools. The electronic test paper data includes detailed information on the grade level of the exam, the exam date, the school administering the exam, and the exam scope. The Jiaoyan Cloud, containing more than 1.2 million Chinese math problems, forms a major part of our private database. Additionally, we have crawled a large number of multimodal math test questions from "Zujuan" totaling around 400k multimodal math questions. We did not crawl all the data, especially since crawling data from Jiaoyan Cloud requires obtaining the corresponding copyrights. Ultimately, the ratio of data in our private database from Jiaoyan Cloud to Zujuan is approximately 7:3.

To format all questions for use, we processed them by OCR engine like Mathpix<sup>5</sup> interface. Due to inherent errors in the OCR engine, we introduced manual checks to ensure the accuracy of parsing results and to verify whether the questions belong to multimodal math problems.

To more clearly elucidate our data collection process, we have depicted the overall pipeline of

<sup>3</sup><https://www.jiaoyanyun.com/>

<sup>4</sup><https://zujuan.xkw.com/>

<sup>5</sup><https://mathpix.com/>

data collection in Figure 15.

## E.3 Quality Check Detail

Since we use GPT-4 as a quality check tool, we provide the prompt for GPT-4 quality check in Figure 9, GPT-4 assigned a confidence level to the data on a scale of 1 to 5. Data with a confidence level lower than 3 were filtered out. To ensure the high quality of the final data, we conducted sampling and manual verification. We performed three random samples, each consisting of 500 multimodal samples, to check the data quality and ensure the consistency of the knowledge points and data. When verifying whether problem is solvable, we use multiple closed-source interface models (GPT-4o, Gemini-Pro and Claude-3.5-Sonnet) to solve each problem. For problems that cannot be solved, we perform manual checks to compare whether the reference analysis given is correct and whether the problem can be solved.

## E.4 Knowledge Point Assistant Labeling Detail

The data from Jiaoyan Yun already includes detailed knowledge point classifications. We have also provided fine-grained knowledge point annotations for the questions sourced from Zujuan. The GPT-4 prompt for knowledge point labeling includes a detailed documentation of knowledge



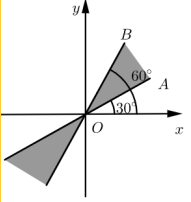
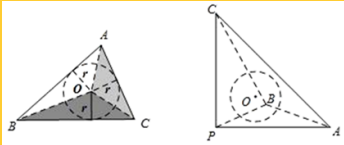
Plane Geometry(Middle)	Solid Geometry(High)
<p><b>Question:</b> 如图, 分别写出适合下列条件的角的集合。</p> <p>(1) 终边落在射线OB上: _____。</p> <p>(2) 终边落在直线OA上: _____。</p> <p>(3) 终边落在阴影区域内(含边界): _____。</p> 	<p><b>Question:</b> 设O是<math>\triangle ABC</math>的内切圆圆心, <math>r_{内}</math>是<math>\triangle ABC</math>的内切圆半径, 设S是<math>\triangle ABC</math>的面积, <math>l_{ABC}</math>是<math>\triangle ABC</math>的周长。</p> <p>(1) 与三棱锥的四个面都相切的球叫做三棱锥的内切球。设三棱锥的体积是V, 表面积是S, 请用类比推理思想, 写出三棱锥的内切球的半径公式<math>R_{内}</math></p> <p>(2) 如图2, 在三棱锥<math>P-ABC</math>中, <math>PA, PB, PC</math>两两垂直, 且<math>PA = PB = PC = 1</math>, 求三棱锥<math>P-ABC</math>的内切球半径和外界球的半径之比。</p> 
<p><b>Answer:</b></p> <p>(1) <math>\{\alpha   \alpha = 60^\circ + k \cdot 180^\circ, k \in \mathbb{Z}\}</math></p> <p>(2) <math>\{\alpha   \alpha = 30^\circ + k \cdot 180^\circ, k \in \mathbb{Z}\}</math></p> <p>(3) <math>\{\alpha   30^\circ + k \cdot 180^\circ \leq \alpha \leq 60^\circ + k \cdot 180^\circ, k \in \mathbb{Z}\}</math></p>	<p><b>Answer:</b></p> <p>(1) <math>R_{内} = \frac{3V}{S}</math>      (2) <math>\frac{\sqrt{3}-1}{3}</math></p>
<p><b>Knowledge point:</b> 根据终边位置描述角的范围 Describes the extent of an angle in terms of its terminal edge position</p>	<p><b>Knowledge point:</b> 合情推理与演绎推理 Plausible reasoning and deductive reasoning</p>

Figure 8: Examples of Composition questions question types

<p><b>GPT-4 Data Quality Check Prompt Example:</b></p> <p><b>Role: System</b></p> <p>You are a data quality scoring assistant. Please refer to the scoring criteria and provide a reference score for a Chinese math Question. The scoring criteria for reference are as follows:</p> <p>1: This question has serious grammar issues due to the lack of coherence between the preceding and following sentences.</p> <p>2: The sentences in this question are generally smooth, but there are formatting issues with the LaTeX formula and table.</p> <p>3: The sentences in this question are generally smooth, and the format of the LaTeX formula and table is basically correct. However, there may be some missing conditions, redundant noise text, or contradictory conditions.</p> <p>4: The sentence of this question is smooth, the problem description is basically clear, the LaTeX formula is correct, and the conditions given in the question are not contradictory.</p> <p>5: The sentence of this question is fluent, the problem description is basically clear, the LaTeX formula is expressed correctly, the conditions given in the question are not contradictory, and it is preliminarily estimated that they can be solved.</p> <p><b>Role: User.</b></p> <p>Math Question: {question}</p> <p>Estimated Score:</p>
--

Figure 9: Prompt used for quality inspection with GPT-4.

point types. Considering GPT-4's context length limitation and the large number of knowledge points, we assess whether a problem belongs to a specific category in batches. A GPT-4 Knowledge point classification prompt in Figure 10.

We used GPT-4 to formalize the standard form of knowledge points as a classification problem. After meticulous prompt processing, GPT-4 achieved an accuracy rate of 92% (We validated the preliminary accuracy of knowledge point annotation using GPT on a relatively 500 test set with only a few hundred questions). But this still did not meet our requirements. Therefore, GPT-4 is used here solely to assist with annotation and speed up the process. Afterward, manual verification is performed to en-

sure the accuracy of the annotations.

### E.5 Visual Subjects Detail

We provide the Chinese and English explanations of the Visual Subjects involved in the topics in the dataset in Table 7.

The definition of "Visual Subject" was extracted and screened from the knowledge point names of "Jiaoyan Yun". The classification of "Visual Subject" is coarser compared to the Jiaoyan Yun system and is based solely on the types of images used in multimodal math problems. Jiaoyan Yun's question knowledge point annotations involve the examination of specific "Visual Subjects." For example, for the "Visual-Table" subject, it involves Jiaoyan

**Prompt Example:**  
**Role: System**  
 You are a math problem analysis assistant. Please determine whether the knowledge points tested by a math problem belong to a series of knowledge points based on the definition of the corresponding knowledge points, If there are no matching knowledge points here, provide output <No\_Match>:  
 List of reference knowledge points are:  
 Chinese Knowledge Point 1: “根据折线图计算中位数” which means Calculate the median based on the line chart  
 Chinese Knowledge Point 2: “等腰三角形外角的性质：三角形的一个外角等于与其相邻的两个内角之和” which means: The properties of the outer angle of an isosceles triangle: One outer angle of a triangle is equal to the sum of two inner angles that are not adjacent to it.  
 Chinese Knowledge Point 3: .....  
 .....  
 Chinese Knowledge Point 10:  
 For Example:  
 Example 1:  
 Math Question: 如图,  $AD$ 是 $\triangle ABC$ 的 $BC$ 边上的高,  $AE$ 平分 $\angle BAC$ , 若 $\angle B=42^\circ$ ,  $\angle C=70^\circ$ , 则 $\angle AEC$ 的度数是         .  
 Question Analysis:  
 $\angle B=42^\circ$ ,  $\angle C=70^\circ$ ,  $\angle BAC=180^\circ - \angle B - \angle C=68^\circ$ ,  $AE$ 是角平分线,  $\angle EAC=\frac{1}{2}\angle BAC=34^\circ$ .  
 $AD$ 是高,  $\angle C=70^\circ$ ,  $\angle DAC=90^\circ - \angle C=20^\circ$ ,  $\angle EAD=\angle EAC - \angle DAC=34^\circ - 20^\circ=14^\circ$ ,  $\angle AEC=90^\circ - 14^\circ=76^\circ$ .  
 Corresponding Knowledge Point: 等腰三角形外角的性质：三角形的一个外角等于与其相邻的两个内角之和  
 ....  
**Role: User**  
 Math Question: {question}  
 Corresponding Knowledge Point:

Figure 10: Prompt form used to annotate knowledge points.

Yun’s knowledge point "calculation of median and mode based on statistical tables." For "Solid Geometry", it involves "volume calculation of solid geometric shapes".

**E.6 Knowledge Point Detail**

The CMMaTH dataset mainly adopts the knowledge point system of "Jiaoyan Yun". "Jiaoyan Yun" is a relatively mature commercial knowledge graph widely used in the field of Chinese mathematics education. It has undergone long-term user validation in the mathematics education sector. Compared to the coarse-grained knowledge point classification structures used by MathBench and MM-Math, "Jiaoyan Yun" employs a very rich knowledge structure with a massive scale of knowledge points, which better meets the needs of real-world educational scenarios in the industry.

MathBench(Liu et al., 2024b) is another related work trying to provide knowledge point label in math evaluation benchmark. However, their knowledge system is build by "Subject Area" and "Topic", which has coarse-grained classification. We provide some comparisons of some knowledge points here and the one-to-many correspondence between them in Figure 12 and Figure 13.

We provided detailed annotations of knowledge points for our dataset and conducted preliminary clustering of these knowledge points in "Jiaoyan Yun". The distribution of knowledge points in different clusters is showed in Figure 11.

**E.7 Question Level Detail**

The "Level" is a reference question difficulty provided by the Teaching and Research Cloud, which can be obtained through OCR tags during crawling. It is manually marked by teachers and corresponds to difficulty levels 1-5: "very simple", "simple", "moderate difficulty", "relatively difficult", "difficult".

**E.8 Characteristics Of Annotators**

We utilized a standard team of eight people, who spent two weeks annotating the data. All annotators have a university undergraduate education and are well-versed in basic knowledge of the K12 education field. To ensure quality, each question was verified by at least two people.

**E.9 Heuristics baselines Detail**

Similar to MathVista, we added two heuristic Baselines. These two heuristic strategies can only han-

Image Type	#Num	Image Type	#Num	Image Type	#Num	Image Type	#Num
视觉表格 <i>Visual-Table</i>	1513	折叠展开图 <i>Folded Image Graph</i>	235	立体几何图 <i>Solid Geometry</i>	2054	解析几何图 <i>Analatic Geometry</i>	3060
流程图 <i>Flow Chart</i>	3120	条形图 <i>Bar Chart</i>	4924	散点图 <i>Scatter Chart</i>	517	平面几何图 <i>Plane Chart</i>	3834
折线图 <i>Line Chart</i>	846	饼状图 <i>Fan Chart</i>	175	雷达图 <i>LiDAR Chart</i>	73	抽象类比图 <i>Abstract Analog Graph</i>	440
三视图 <i>Three View Graph</i>	22	枝页图 <i>Stem-and-Leaf display</i>	23	其他 <i>Other Image type</i>			240

Table 7: Primary visual element types involved in the CMMaTH dataset.

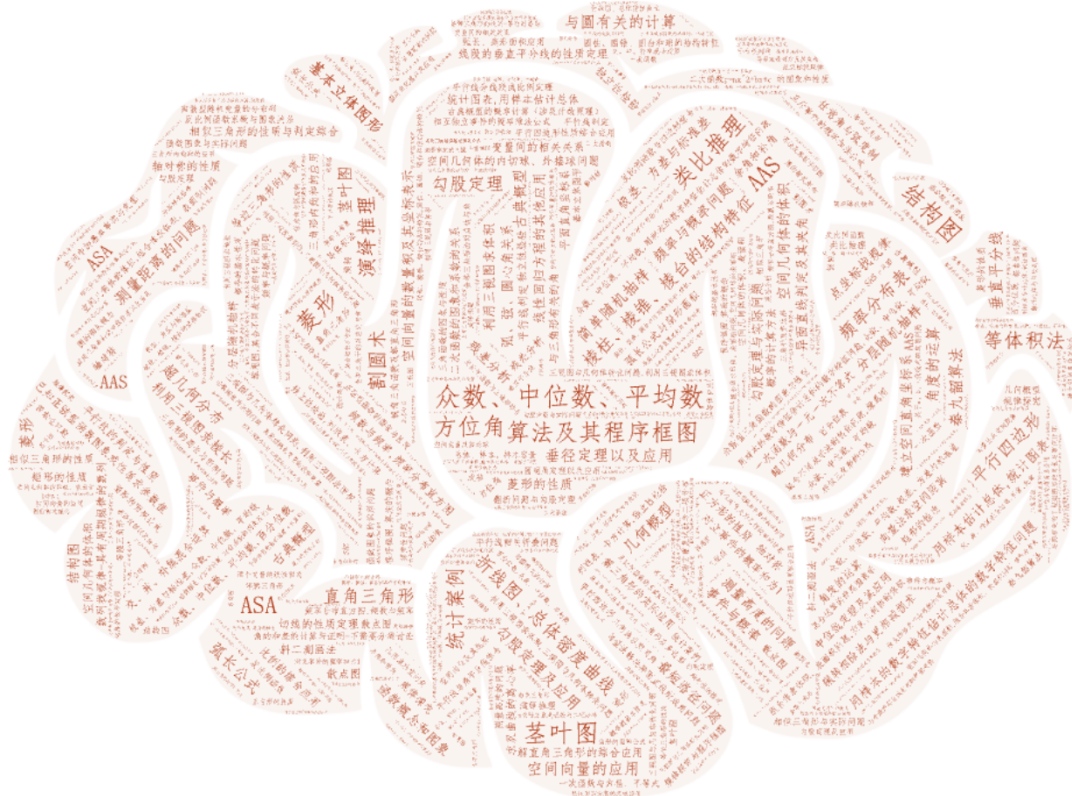


Figure 11: Cloud diagram of the knowledge points contained in the CMMaTH dataset.

dle multiple-choice questions in the dataset. "Random Guess" selects one from the options with equal probability each time, and "Frequent guess" follows the options. The proportion in the dataset serves as a prior probability to sample an option. For the evaluation of human performance, we used a subset of the CMMaTH dataset, consisting of approximately 1,500 samples. The participants were high school students from three groups, each from a top high school. Every student in each group was required to answer the questions in the subset, and we reported the average accuracy of the answers for the three groups.

## F Hallucinations Types Definition of Human Evaluation

In our study, we employed a detailed typology of hallucinations for human evaluation on the CMMaTH subset (about 500 examples).

### Perception Errors

Perception Error refers to the model's erroneous interpretation and utilization of diagram content during reasoning. For example, incorrect OCR, misidentification of numerical relationships, geometric relationships, logical relationships, etc.

### Reasoning Errors

Reasoning Error are quite common during the solving process. For instance, the model may misinterpret symbols or use incorrect logic or

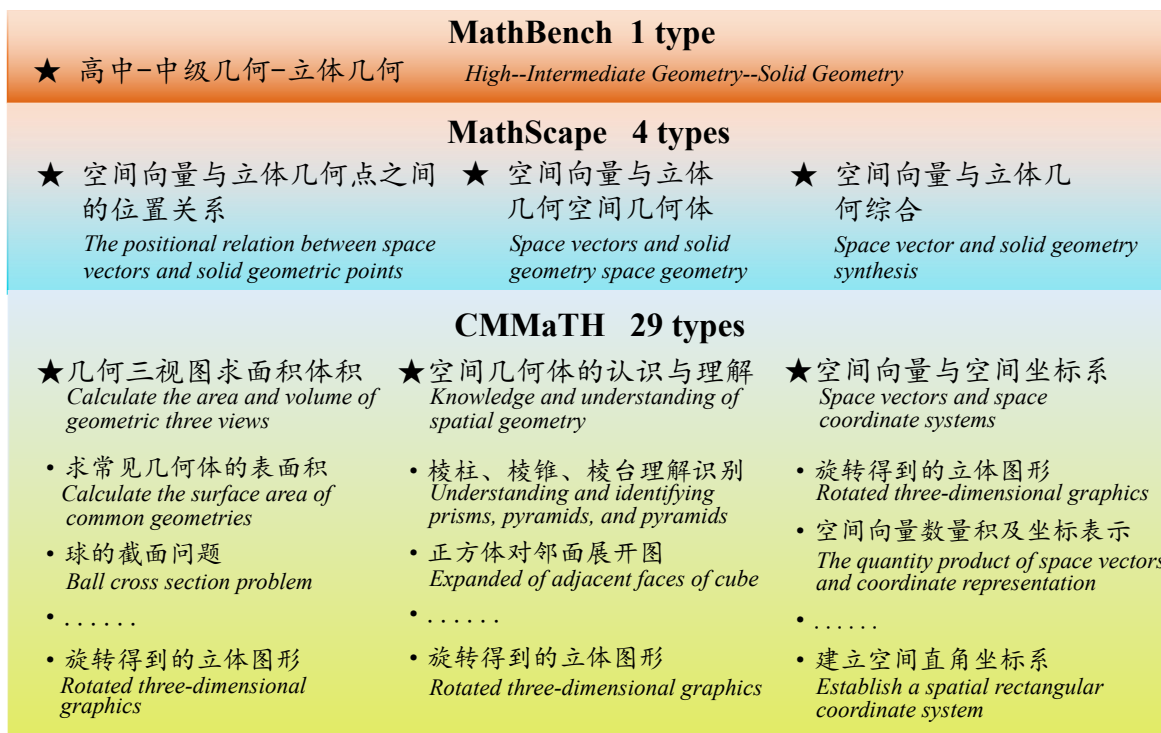


Figure 12: Knowledge Point Annotation Comparisons in Topic Plane Circle.

Model	Overall	Flow	Bar	Scatter	Line Plot	Fan	LiDAR	Visual-Table	Three View	Folded Image	Analytic	Solid	Plane	Venn	Abt-Analogy
LMMs(Text+Image, Zeroshot)															
GPT4V	27.0	39.3	12.5	30.2	21.0	22.9	38.6	16.9	18.3	20.0	37.5	15.8	21.5	58.0	29.9
Gemini-Pro	35.7	58.9	21.1	47.1	31.2	50.6	32.4	27.4	24.7	20.3	37.3	29.5	42.2	53.8	31.5
Claude-3.5	37.4	63.4	20.9	56.6	44.3	60.2	35.1	31.7	30.8	21.6	37.6	29.1	37.7	59.6	38.8
GPT4o	47.8	59.1	45.5	70.3	37.3	50.0	18.9	48.0	32.2	31.8	49.1	43.2	47.4	61.2	45.6
LMMs(Text+Image, Few shot)															
Gemini-Pro(3-Shot)	39.3	34.9	49.7	50.5	20.8	24.9	9.0	44.1	26.6	19.7	36.5	34.6	41.0	42.5	34.0
Claude-3.5(3-Shot)	48.8	57.9	50.4	73.8	47.9	53.5	38.6	52.0	38.6	28.8	47.8	42.6	43.8	70.2	47.2
GPT4o(3-Shot)	52.2	63.9	50.4	75.3	42.9	55.0	24.0	53.6	37.0	36.8	54.3	48.6	52.6	66.1	50.5

Table 8: Model performance of few-shot(3-shot) experiment on CMMaTH.

knowledge for inference. The frequency of Reasoning Errors reflects the model’s logical and mathematical reasoning capabilities.

### Calculation Errors

Calculation Error refers to the model performing incorrect mathematical operations, such as writing equations or solving equations incorrectly.

### Reject Errors

Reject Error refers to the model’s inability to solve a problem that is actually solvable. The frequency of such errors reflects the model’s ability to follow instructions.

## F.1 Case Study

We conducted a fine-grained manual evaluation of GPT-4V’s output on CMMaTH, with the results shown in Figure 21 to Figure 31.

## G GradeGPT details

### G.1 GradeGPT Prompt Detail

We have listed detailed Fewshot Examples using the GPT-4-generated GradeGPT model responses in Table 14. Through this table, you can observe the specific form of the Cross-Lingual-Judge-of-Chain that we have used.

### G.2 GradeGPT Performance Metric

GradeGPT’s performance evaluation metric is precision in comparison. We constructed a model that responds to a test set containing outputs from various large models (including both correct and incorrect model outputs). Each output is labeled as correct or incorrect based on its result. GradeGPT is tasked with assessing whether the model responses are correct or incorrect, and this performance evaluation metric is a binary classification metric.



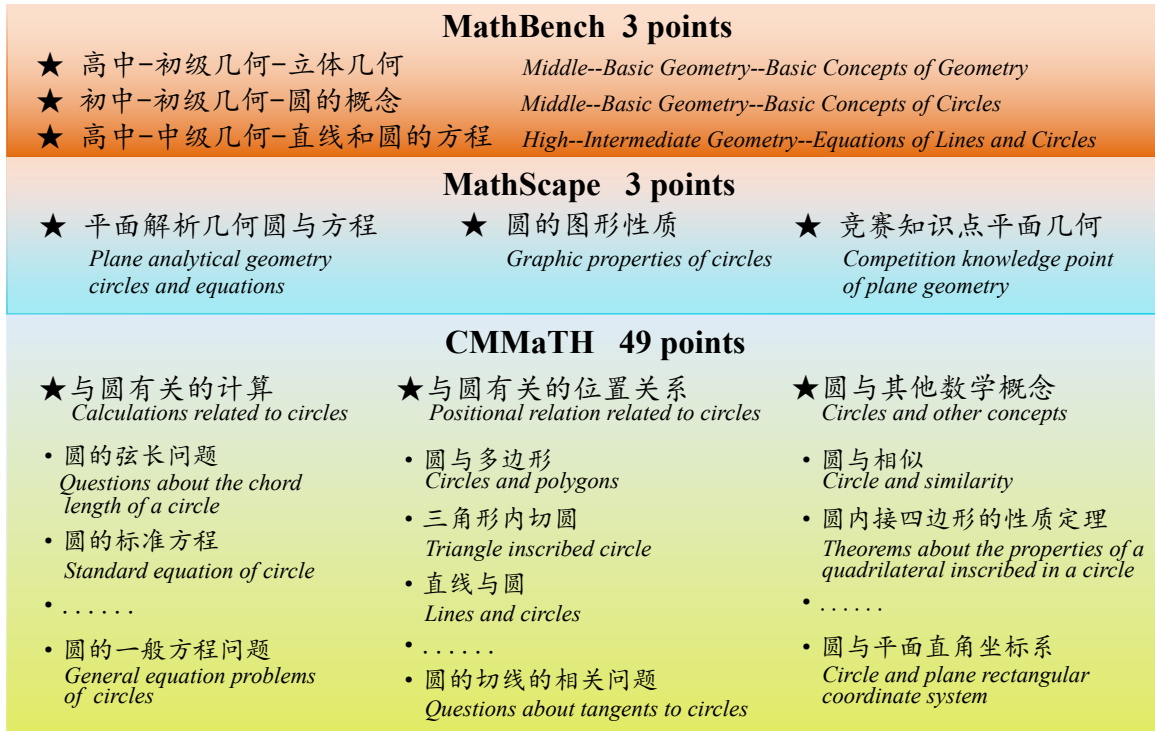


Figure 13: Knowledge Point Annotation Comparisons in Topic Solid Geometry.

### G.3 GradeGPT Training Details

We generated cross-lingual evaluation instruction pairs using the outputs from InternLM-XComposer, LLaVA-v1.5, CogAgent-18B and Yi-VL-34B. These outputs were produced using GPT-4 Fewshot. The generated evaluation instructions were filtered based on specific rules, retaining only those responses from GPT-4 that contained the fields: <Yes>/<No>. Ultimately, we constructed a cross-lingual format instruction set comprising 56k instruction pairs.

GradeGPT was trained on 8 H800, with the Qwen-14B-Chat version used as the base model. The model’s batch size was set to 16. The learning rate was set to 1e-4, and the gradient accumulation step was set to 16. It was trained for 10 epochs on a 40k bilingual Judge-of-Chain dataset. A detail example of instruction can refer to Figure 14.

### G.4 Futher More Ablation Study

We conducted experiments on a development set comprising outputs from a 0.5k model. The development set was sampled from a subset of 0.5k questions on CMMaTH. Each question was accompanied by answers provided by GPT-4V, GPT-4o, and middle school students. Each answer was manually annotated to indicate whether it was correct. We use 2 to measure the

answer judgment capability of different LMMs, including Zershot LMMs and LLMs after Finetune.

**Ablation On Instruction Format** We conducted experiments on various instruction enhancement techniques used by GradeGPT and compared the results with GPT-4 in Table 9. The results suggest that after various instruction enhancements, the accuracy of GradeGPT in model response judgment on CMMaTH can be improved to 96.1%, significantly surpassing the accuracy of GPT-4. The proposed strategy can significantly enhance GradeGPT’s ability to judge results. It is only slightly weaker than the performance of GPT-4(Fewshot) executed with a large number of examples. Our GradeGPT, as an open-source parameter model of approximately 14B, can serve as a stable, low-cost, and efficient alternative to GPT-4.

The Baseline we compared, Qwen-7B/14B(4-Shot), GPT-4(4-Shot), *Naive Outcome Finetune*, *Judge-of-Chain*. In the *Naive Outcome Finetune* format of instructions, the model is required to output its results indicating whether they are correct in the form of "<Yes>"/"<No>". *Judge-of-Chain* also includes the understanding of results and natural language descriptions of model outputs, but does not include the part of extracting key Chinese out-

**Instruction Example:****Role: System**

You are an answer evaluation assistant. Based on the following questions and reference answers, determine if the model's output is the same answer. When making the determination, please provide the analysis process. Note: for multiple-choice questions, it is only correct if the output includes all the corresponding options. Note: if the model response does not specify a particular option, and its correctness cannot be determined, the answer consistency should be: <No>.

**Role: User**

以下是某道题目，答案和一条模型响应。请确定模型响应中是否给出了正确的参考答案。

Question: 如图：若 $\triangle ABE \cong \triangle ACD$ ， $\angle A = 50^\circ$ ， $\angle C = 20^\circ$ ，则 $\angle ADC = ?$ 。

Reference Answer:  $110^\circ$

Model Response: 首先，我们需要了解这个问题的背景和目标。题目给出了三角形的三个角度，并要求我们找到另一个角度。我们可以通过三角形的内角和来解决这个问题。首先，我们计算三角形的内角和： $\angle A + \angle B + \angle C = 180^\circ$ 。 $50^\circ + 20^\circ + \angle ADC = 180^\circ$ 。接下来，我们需要找到另一个角度，即 $\angle ADC$ 。我们可以通过将内角和减去已知角度来得到它： $\angle ADC = 180^\circ - \angle A + \angle B + \angle C$ 。 $\angle ADC = 180^\circ - (50^\circ + 20^\circ)$ 。 $\angle ADC = 180^\circ - 70^\circ$ ， $\angle ADC = 110^\circ$ 。所以，三角形 $\triangle ABE$ 与 $\triangle ACD$ 的另一个角度是 $\angle ADC = 110^\circ$ 。

Judgment Result:

**Role: Assistant**

The relevant paragraph in the Model Response that provides the answer is

所以，三角形 $\triangle ABE$ 与 $\triangle ACD$ 的另一个角度是 $\angle ADC = 110^\circ$

Understand the question in Chinese. The Chinese model response means: Therefore, another angle of triangle  $\triangle ABE$  and  $\triangle ACD$  is  $\angle ADC = 110^\circ$ .

Next, let's analyze whether the correct answer has been given. The reference answer  $110^\circ$  is a numerical answer, and the output given in the model response is  $110^\circ$ . The consistency of the answers is: <Yes>.

Figure 14: A instruction example to finetune GradeGPT.

puts and translating them into English. Compared to having the language model directly predict the <Yes>/<No> judgment labels, directly using Chinese Judge-of-Chain to construct Judge-of-Chain improves the model's performance in answer evaluation. However, they didn't yield good results. The performance disparity may stem from an imbalance in the quality of Chinese and English components within some bilingual base models due to the training corpus.

Our findings show that fine-tuning with Cross-Lingual-Judge-of-Chain for detailed thought chain refinement significantly improves the performance of open-source models in outcome analysis tasks. Additionally, we discovered that using bilingual thought chains instead of Chinese-only thought chains for base model fine-tuning effectively enhances performance in outcome determination tasks. By using and synthesizing the instructions in the form of Cross-Lingual-Judge-of-Chain that we designed, we are able to efficiently distill the answer reviewing capabilities of GPT-4.

**Ablation On Instruction Data Source** The instruction data for Cross-Lingual Judge-of-Chain Prompts comes from outputs of various LLMs on CMMaTH. We conducted ablation experiments on the sources of instruction data in Table 10, which showed the impact of using different LLM models in constructing diverse and effective instruction

LLM	$Acc_{outcome}$
Qwen-7B-Chat(4-Shot)	35.1
+Naive Outcome Finetune	51.5
+Judge-of-Chain	65.3
+Cross-Lingual-Judge-of-Chain	85.1
Qwen-14B-Chat(4-Shot)	43.7
GradeGPT(14B)	96.1
GPT-4(4-Shot)	97.2

Table 9: Ablation study on the instruction fine-tuning of GradeGPT commands

Model Response Source	$Acc_{outcome}$
LLaVA-v1.5 response	77.2
+InterLM-XComposer2-VL response	83.1
+CogAgent-Chat response	84.2
+Yi-VL-34B response	85.1

Table 10: Ablation study on the instruction fine-tuning of GradeGPT commands. All experiments were conducted using the GradeGPT (7B) version.

data.

$$Acc_{outcome} = \frac{I(GradeGPT(R_i), Overcome_{GT})}{N_{response}} \times 100 \quad (2)$$

## H Prompt Details

### H.1 Prompt For Step Response Generation

When evaluating hallucinations during the assessment process, we use a few-shot prompt format

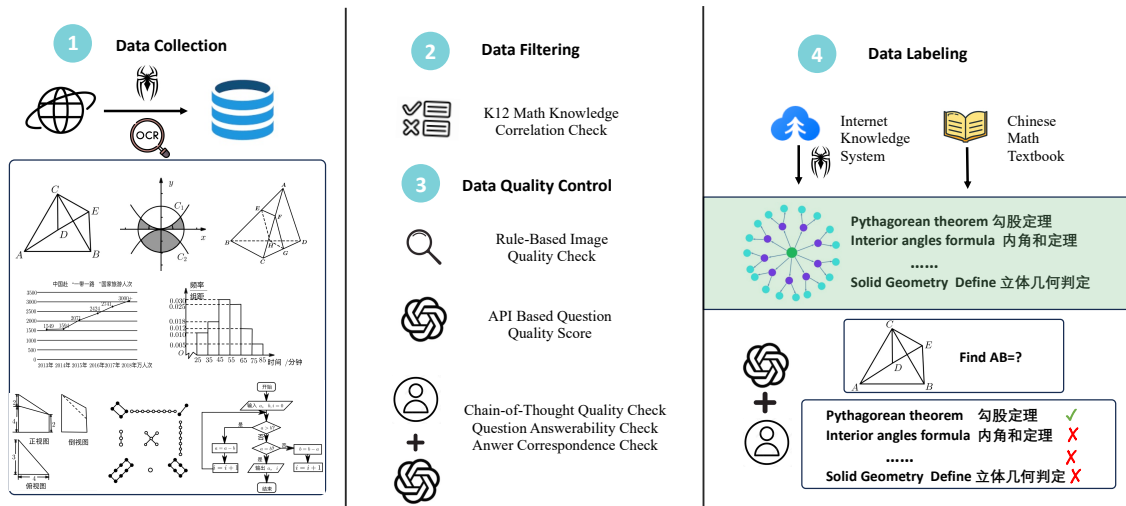


Figure 15: Overall Data Collection Pipeline of CMMaTH.

Element	Prompt
system-prompt	你是一个数学题解题助手，你的输入是一道数学题，以及这道题目的图像，你的任务是输出这道题目的求解思路以及答案。输出格式是一步一步的方式。
Example1	<p><b>Question:</b> 小明有2000元，一个苹果2元，小明买了10个苹果。小明还有多少钱。</p> <p><b>Model response:</b> 首先计算小明买苹果的开销为2乘10等于20元。&lt;STEP1&gt; 用2000元减去所有的开销，2000-20=1980。&lt;STEP2&gt; 所以最终的答案为1980元。&lt;STEP3&gt;</p>
Example2	<p><b>Question:</b> 如图，在<math>\triangle ABC</math>中，<math>\angle C=90^\circ</math>，<math>\angle B=30^\circ</math>，以A为圆心，任意长为半径画弧分别交AB、AC于点M和N，再分别以M、N为圆心，大于<math>MN/2</math>的长为半径画弧，两弧交于点P，连结AP并延长交BC于点D，则下列说法中正确的个数是（ ）①AD是<math>\angle BAC</math>的平分线；②<math>\angle ADC=60^\circ</math>；③点D在AB的中垂线上；④<math>BD=2CD</math>。A.4 B.3 C.2 D.1</p> <p><b>Model response:</b> 首先确定<math>\triangle ABC</math>的几何性质。&lt;STEP1&gt;<math>\angle C=90^\circ</math>，<math>\angle B=30^\circ</math>，<math>\angle A=120^\circ</math>。&lt;STEP2&gt;<math>\triangle ABC</math>是一个等腰三角形。&lt;STEP3&gt;根据等腰三角形的性质，AD是<math>\angle BAC</math>的平分线。&lt;STEP4&gt;①AD是<math>\angle BAC</math>的平分线，正确。&lt;STEP5&gt;②<math>\angle ADC=60^\circ</math>，错误。&lt;STEP6&gt;③点D在AB的中垂线上，错误。&lt;STEP7&gt;④<math>BD=2CD</math>，正确。&lt;STEP8&gt;因此，正确的个数为4。&lt;STEP9&gt;答案为A.4。&lt;STEP10&gt;请你根据这个例子，解决下面的数学题。问题：在<math>\triangle ABC</math>中，<math>\angle C=90^\circ</math>，<math>\angle B=30^\circ</math>，以A为圆心，任意长为半径画弧分别交AB、AC于点M和N，再分别以M、N为圆心，大于<math>MN/2</math>的长为半径画弧，两弧交于点P，连结AP并延长交BC于点D，则下列说法中正确的个数是（ ）①AD是<math>\angle BAC</math>的平分线；②<math>\angle ADC=60^\circ</math>；③点D在AB的中垂线上；④<math>BD=2CD</math>。A.4 B.3 C.2 D.1 求解步骤：首先确定<math>\triangle ABC</math>的几何性质。<math>\angle C=90^\circ</math>，<math>\angle B=30^\circ</math>，<math>\angle A=120^\circ</math>。<math>\triangle ABC</math>是一个等腰三角形。根据等腰三角形的性质，AD是<math>\angle BAC</math>的平分线。①AD是<math>\angle BAC</math>的平分线，正确。②<math>\angle ADC=60^\circ</math>，错误。③点D在AB的中垂线上，错误。④<math>BD=2CD</math>，正确。因此，正确的个数为4。答案为A.4。</p>

Table 11: Prompt for all model to generate step-by-step answer.

to elicit step-by-step outputs from the model as showed in Table 11.

## H.2 Prompt For GradeGPT

We also listed the prompts used by GradeGPT in Table 12.

## H.3 Prompt For Cross-Lingual Prompting and En-CoT

We have listed the specific prompts used for En-CoT in Table 13 and Cross-Lingual Prompt during actual execution in Table 14. Unlike the original

Cross-Lingual Prompt paper, for experimental simplicity, we only adopted a single-turn format. However, this suffices to illustrate the varying inferential capabilities across different languages in current LMMs.

Element	Prompt
system-prompt	You are an answer evaluation assistant. Based on the following questions and reference answers, determine if the model's output is the same answer. When making the determination, please provide the analysis process in the following format. Note, for multiple-choice questions, it is only correct if the output includes all the corresponding options. (For example: If the reference answer is: A, C, and the model response is: C, the answer consistency is: No) Note, if the model response does not specify a particular option, and its correctness cannot be determined, the answer consistency should be: No.
Example	<p><b>Question:</b> 如图所示, AD是<math>\triangle ABC</math>的中线, 点E在AC上, BE交AD于点F. 当<math>\frac{AF}{FD} = \frac{1}{2}</math>时, <math>\frac{AE}{AC} =</math></p> <p><b>Model Response:</b> 首先, 我们需要找到中线AD, 然后找到点E在AC上的位置。找到中线AD, 可以通过勾股定理求解, <math>AD^2 = AD \times AC^2</math>。设<math>AD = x, AC = y</math>, 则有<math>x^2 = xy^2</math>。&lt;STEP1&gt;解这个方程, 我们得到<math>y = \sqrt{x}</math>。现在我们知道点E在AC上, 所以E的横坐标为y, 纵坐标为x。</p> <p>接下来, 我们需要找到点F在BD上的位置。由于F点在BD上, 我们可以设F点在BD的一半, 即<math>F = (x + y)/2</math>。&lt;STEP2&gt;现在我们可以计算AF和FD的长度。<math>AF = x - F, FD = FD - F</math> 将已知的x和y代入公式, 得到: <math>AF = x - (x + y)/2, FD = (x + y)/2 - (x + y)/2</math> 简化得到: <math>AF = x - x, FD = x - x</math> 这意味着<math>AF = FD</math>。&lt;STEP3&gt;现在我们可以求解题目中的条件: <math>AE/AC = 1/2</math> 将已知的AF和FD代入公式, 得到: <math>AE/AC = (x - x)/(\sqrt{xy})</math>化简得到: <math>AE/AC = 1/2</math>所以, 答案为<math>AE/AC = 1/2</math> &lt;STEP4&gt;</p> <p><b>Reference Response:</b> <math>\frac{4\pi}{3} - 2\sqrt{3}</math></p> <p><b>Judge Result:</b> The relevant paragraph in the Model Response that provides the answer is 所以, 答案为<math>AE/AC = 1/2</math>。 Understand the question in Chinese. The Chinese model response means: Therefore, the answer is <math>AE/AC = 1/2</math>. Next, let's analyze whether the correct answer has been given. The reference answer <math>\frac{1}{5}</math> is a numerical answer, and the output given in the model response is <math>1/2</math>. The consistency of the answers is: &lt;No&gt;.</p>

Table 12: Detail prompt for GradeGPT and its instruction construction prompt by GPT4.

Element	Prompt
system-prompt	You are a math problem-solving assistant. Your input is a math problem and an image of the problem. Your task is to provide the solution steps and the final answer. The output format should be step-by-step.
Example	<p><b>Question:</b> Xiaoming has 2000 yuan, an apple costs 2 yuan, Xiaoming bought 10 apples. How much money does Xiaoming have left?</p> <p><b>Model response:</b> First, calculate the cost of the apples Xiaoming bought, which is 2 times 10 equals 20 yuan. &lt;STEP1&gt; Subtract the total cost from 2000 yuan, <math>2000 - 20 = 1980</math>. &lt;STEP2&gt; So the final answer is 1980 yuan. &lt;STEP3&gt;</p>

Table 13: Prompt for En-CoT in Our Cross-Lingual Experiment.

Element	Prompt
system-prompt	Please act as an expert in multi-lingual understanding in Chinese. After understanding, you should act as an expert in arithmetic reasoning in English. Let's resolve the task you understand above step-by-step! Finally, you should format your answer as 'Answer: [num]'.
Example	<p><b>Request:</b> 玛丽比琼小两岁, 琼比杰莎大五岁。如果杰莎20岁, 这三个女孩的年龄总和是多少?</p> <p><b>Reason Result:</b> Let's understand the task in English step-by-step! 1. "玛丽比琼小两岁" means "Mary is two years younger than Joan." 2. "琼比杰莎大五岁" means "Joan is five years older than Jessie." 3. "如果杰莎 20 岁" means "If Jessie is 20 years old." Then, Let's solve the task step-by-step: 1. According to the first sentence, Mary is two years younger than Joan. Let's suppose Joan's age is J, then Mary's age is J - 2. 2. According to the second sentence, Joan is five years older than Jessie. If Jessie's age is x, then Joan's age is x + 5. Answer: 68.</p>

Table 14: Detail prompt for Cross-Lingual Prompt(CLP) Reason technology for our Experiment.

Model Name	Generation Parameters	Comments
MAmmoTH-70B	temperature=0.2, max_new_tokens=2048	vllm package
InternLM-Math	temperature=0.2, max_new_tokens=2048	vllm package
MetaMath-70B	temperature=0.2, max_new_tokens=2048	vllm package
DeepSeek-Math-7B	temperature=0.2, max_new_tokens=2048	vllm package
Llama-2-70B	do_sample=True, top_k=0.5, top_p=0.5, max_tokens=512	vllm package
Baichuan-13B	temperature=0.2, max_new_tokens=2048	vllm package
Qwen-14B	temperature=0.2, max_new_tokens=2048	vllm package
llava-7B-V1.5	temperature=0.2, max_new_tokens=2048	llava package
Yi-VL-34B	temperature=0.2, max_new_tokens=2048	Huggingface
LLaVA-NEXT-34B	temperature=0.2, max_new_tokens=2048	Huggingface
LLaVA-OneVision	temperature=0.2, max_new_tokens=2048	Huggingface
CogAgent-Chat	temperature=0.2, max_new_tokens=2048	Huggingface
Qwen2-VL	temperature=0.2, max_new_tokens=2048	Huggingface
InternVL2	temperature=0.2, max_new_tokens=2048	Huggingface
InterLM-XComposer2-VL	temperature=0.2, max_new_tokens=2048	Huggingface
CogVLM	temperature=0.2, max_new_tokens=2048	Huggingface
GPT-4	temperature=0.2, max_tokens=2048	version="gpt-4-1106-preview"
Gemini	temperature=0.2, max_tokens=2048	version="gemini-1.5-Pro-2023-05-15"
Claude	temperature=0.2, max_tokens=2048	version="claude-3.5-sonnet-2024-05-24"
GPT-4V	temperature=0.2, max_tokens=2048	version="gpt-4-vision-2023-05-15"
GPT-4o	temperature=0.2, max_tokens=2048	version="gpt-4o-2024-05-14"

Table 15: The hyperparameters for the models used in the evaluation are detailed. When the "comments" section includes the format `model = ""`, it signifies that the model was loaded from the transformer package. The vLLM package indicates that models are implemented by the vLLM package, where more details can be found in <https://github.com/vllm-project/vllm>. For models other than OpenAI's GPT, custom codes were utilized for evaluation unless specified otherwise in the comments.



Model	Name
Llama-2-70B	meta-llama/Llama-2-70b-hf
Baichuan-13B	baichuan-inc/Baichuan2-13B-Chat
Qwen-14B	Qwen/Qwen-14B-Chat
InternLM-Math	internlm/internlm2-math-20b
DeepSeek-Math-7B	deepseek-ai/deepseek-math-7b-instruct
MetaMath-70B	meta-math/MetaMath-70B-V1.0
MAMmoTH2-8x7B	TIGER-Lab/MAMmoTH2-8x7B
Yi-VL-34B	01-ai/Yi-VL-34B
Qwen2-VL	Qwen/Qwen2-VL-7B-Instruct
LLaVA-v1.5	liuhaotian/llava-v1.5-13b
LLaVA-NEXT-8B	llava-hf/llama3-llava-next-8b-hf
LLaVA-OneVision	lmms-lab/llava-onevision-qwen2-7b-ov
InternVL2-8B	OpenGVLab/InternVL2-8B
InterLM-XComposer2-VL	internlm/internlm-7b
CogVLM	THUDM/cogvlm-chat-hf
CogAgent	THUDM/cogagent-vqa-hf
Gemini	<Gemini 1.5 Pro>
Claude	<Claude 3.5 Sonnet>
GPT4o	<GPT4o-20240201>

Table 16: LLMs used in our experiments and their corresponding names in Huggingface Hub.

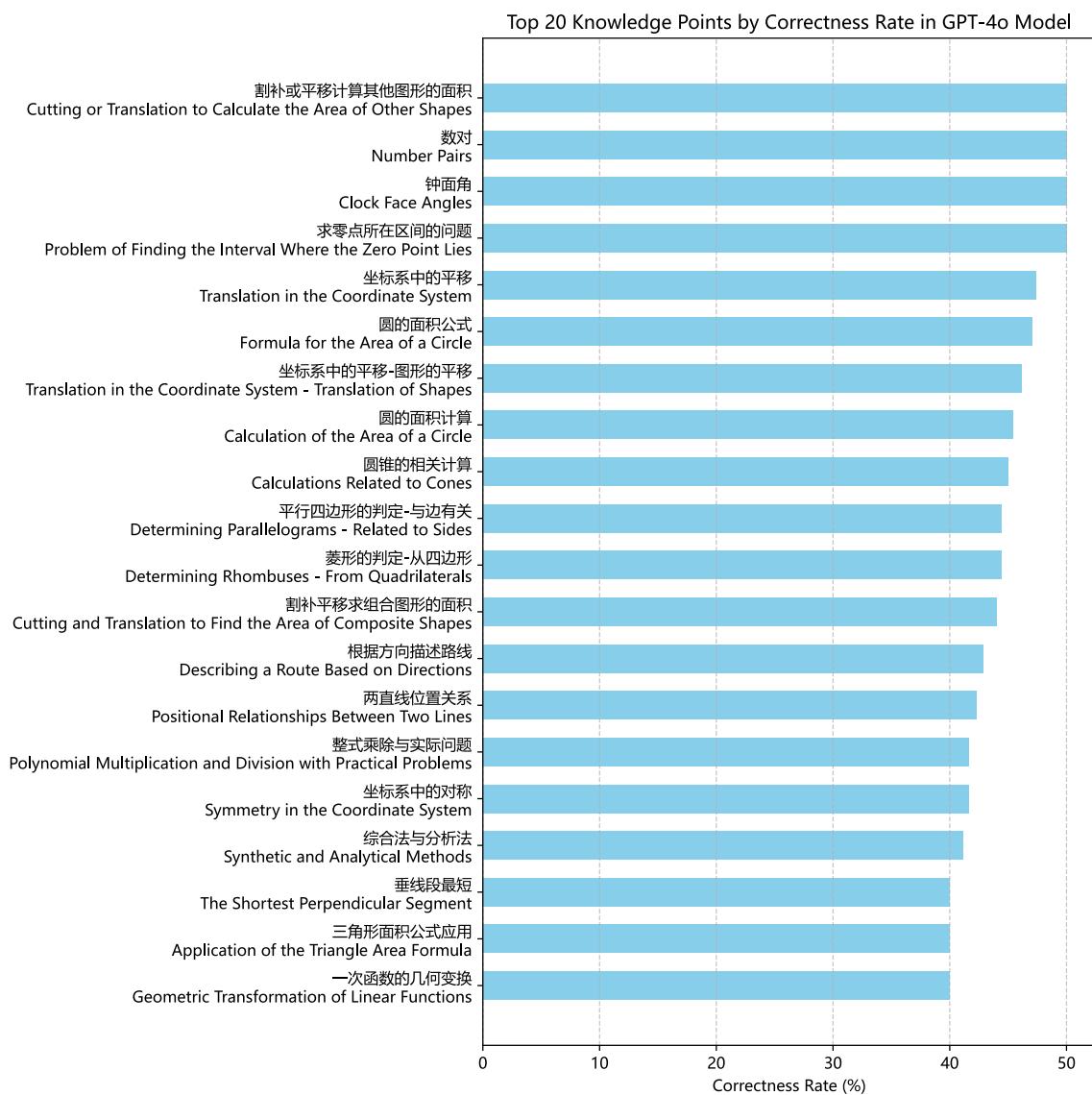


Figure 16: Top 20 Knowledge Point Correctness of GPT-4o.

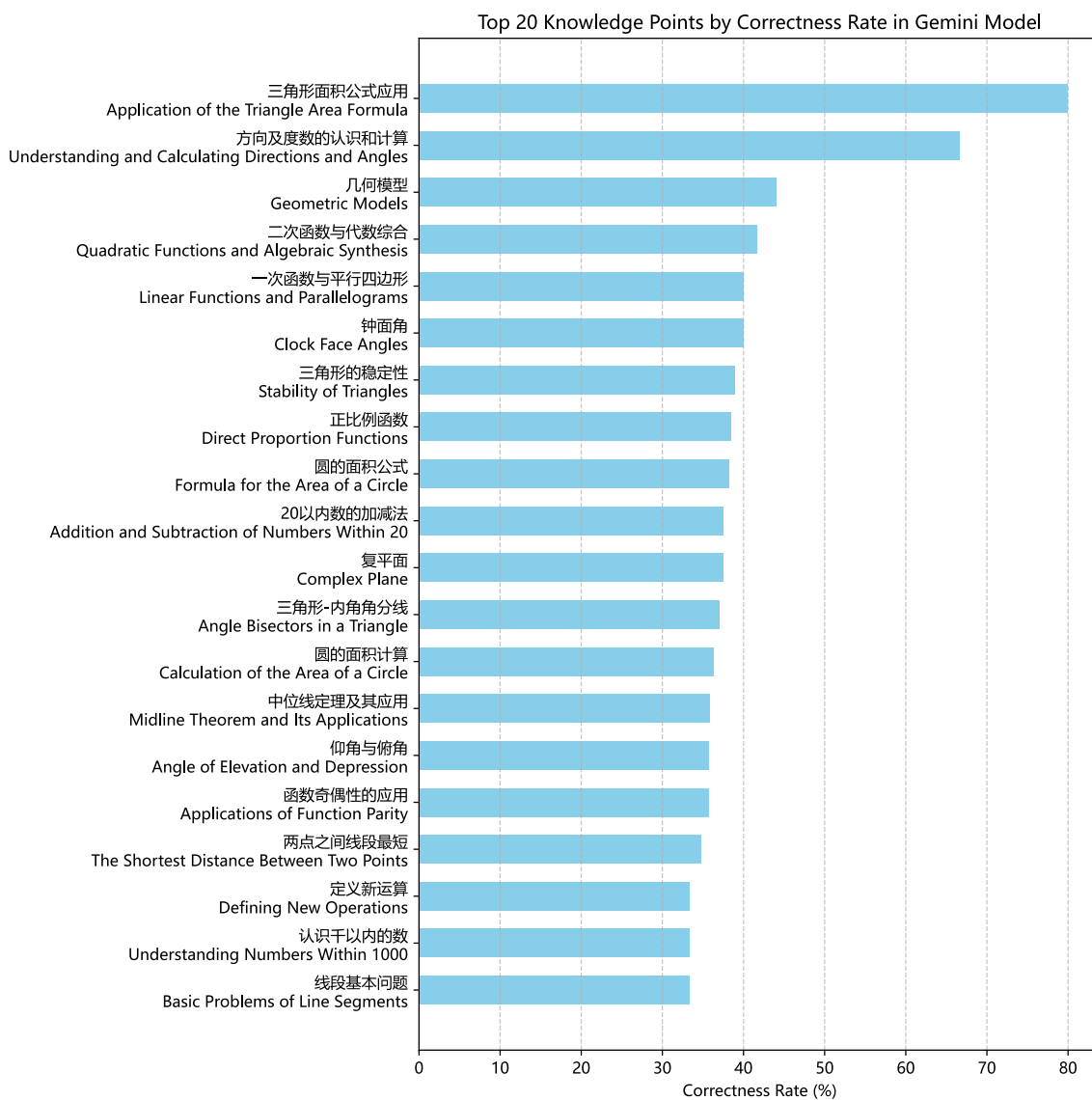


Figure 17: Top 20 Knowledge Point Correctness of Gemini-Pro.

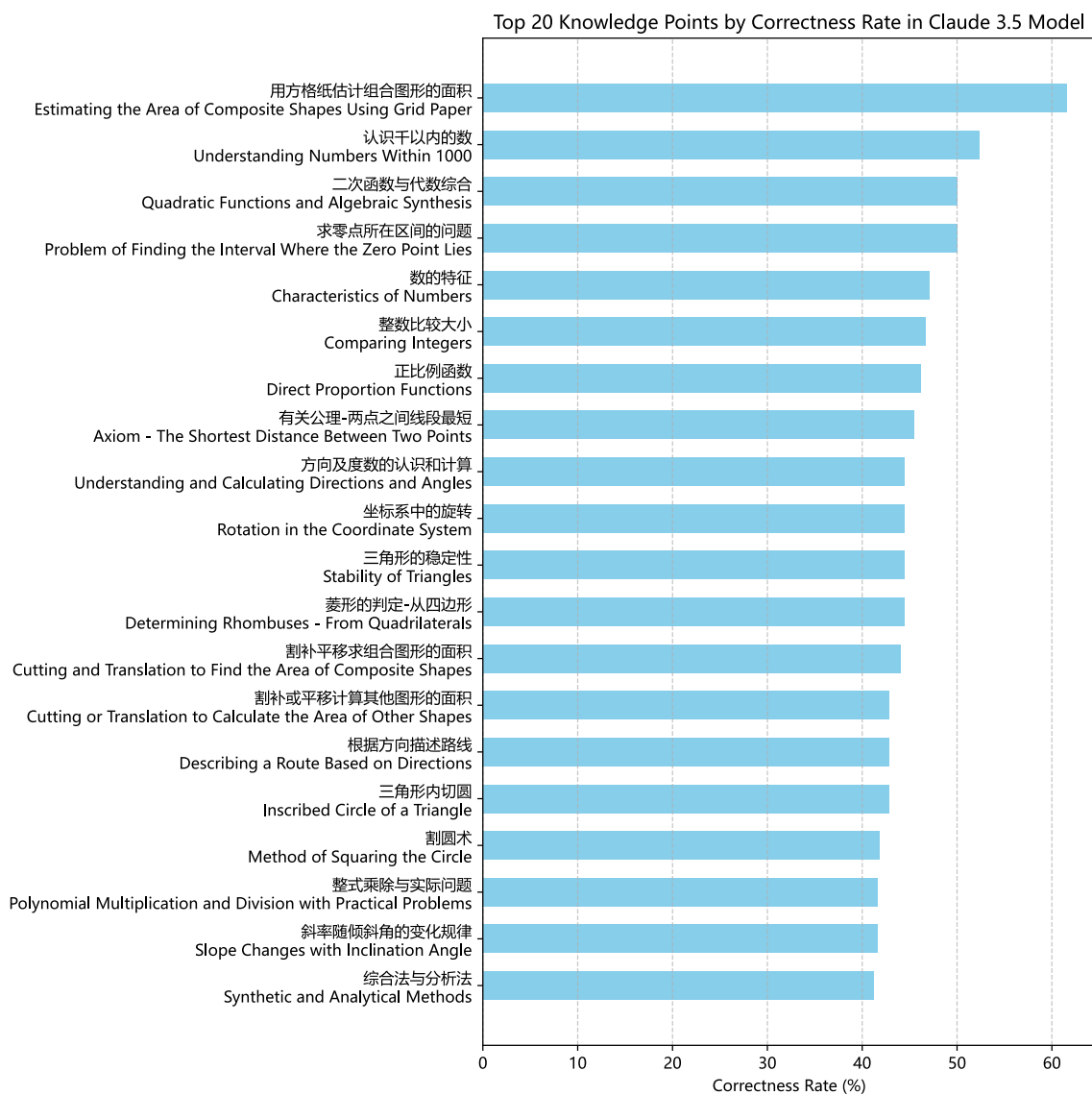


Figure 18: Top 20 Knowledge Point Correctness of Claude-3.5.



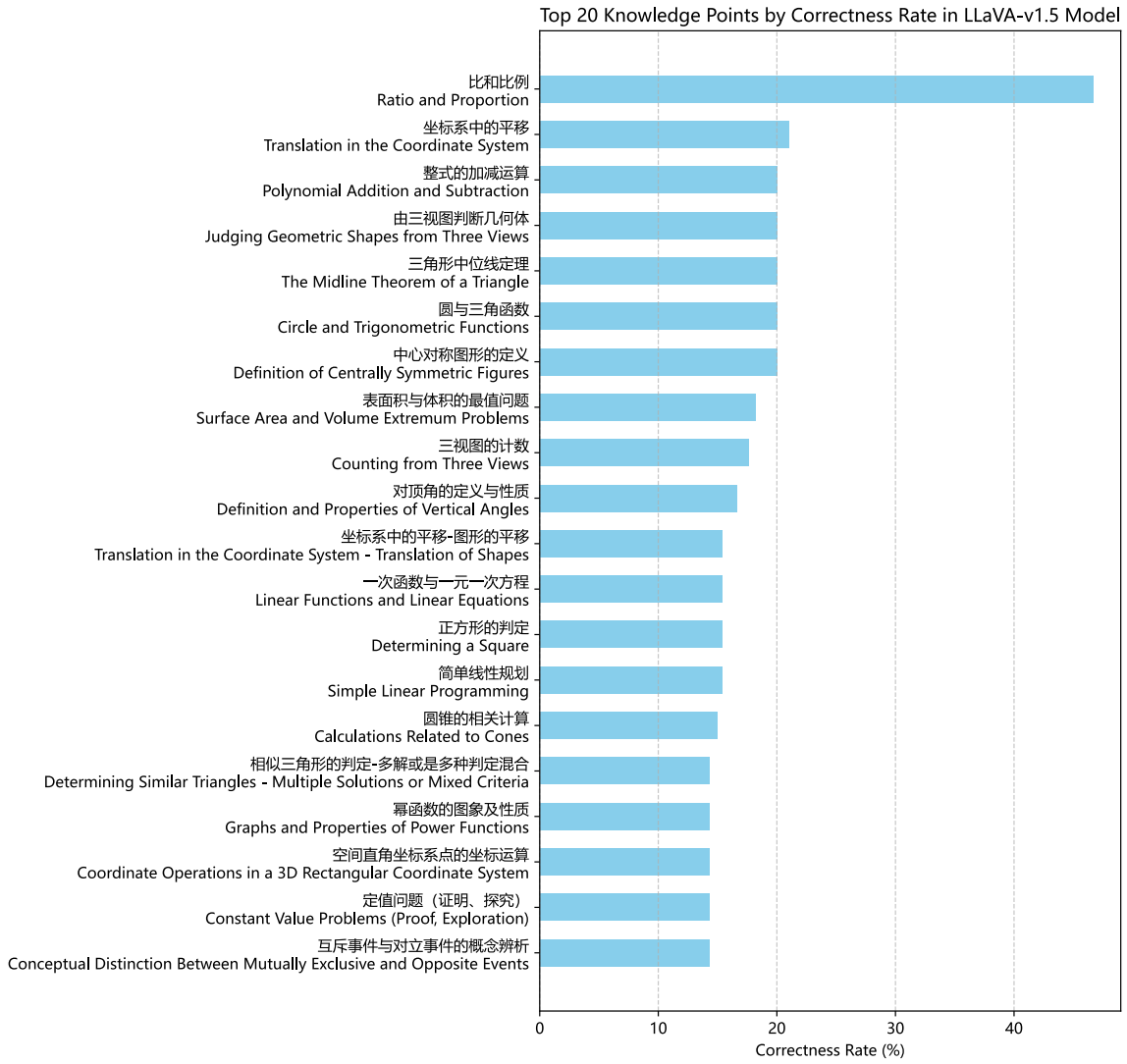


Figure 19: Top 20 Knowledge Point Correctness of LLaVA-v1.5.



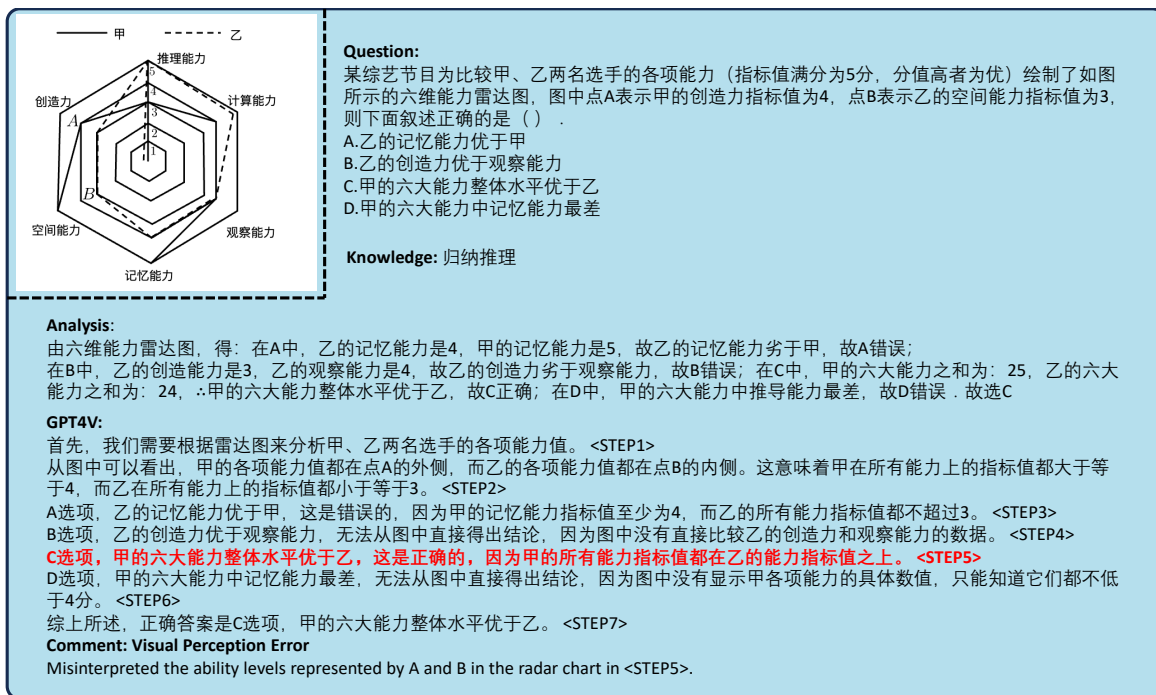


Figure 21: Case from GPT-4V/o. The red ones are marked as generated inference hallucinations.

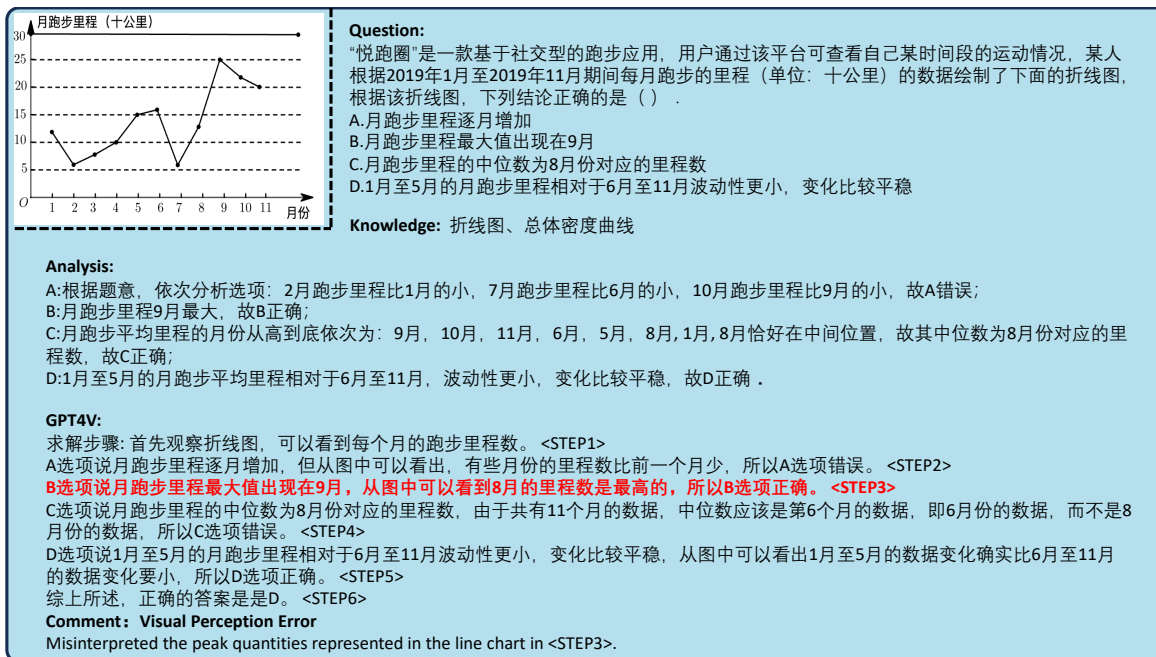
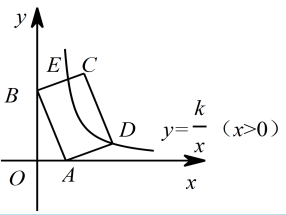


Figure 22: Case from GPT-4V/o. The red ones are marked as generated inference hallucinations.



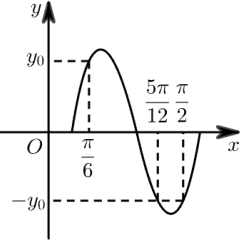
**Question:**  
如图，在平面直角坐标系中，矩形 ABCD 的边  $AB:BC = 3:2$ ，点  $A(3, 0)$ ， $B(0, 6)$  分别  $x$  轴， $y$  轴上，反比例函数  $y = \frac{k}{x} (x > 0)$  的图象经过点  $D$ ，且与边  $BC$  交于点  $E$ ，则点  $E$  的坐标为？

**Knowledge:** 反比例函数

**Analysis:** 过点  $D$  作  $DF \perp x$  轴于点  $F$ ，由同角的余角相等可得出  $\angle OBA = \angle FAD$ ，结合  $\angle AOB = \angle DFA = 90^\circ$  可得出  $\triangle AOB \sim \triangle DFA$ ，根据相似三角形的性质结合点  $A$ 、 $B$  的坐标，即可得出  $AF$ 、 $DF$  的长度，进而可得出点  $D$  的坐标，再利用反比例函数图象上点的坐标特征即可求出  $k$  值，可得反比例函数解析式，由四边形  $ABCD$  是矩形，边  $BC$  可以看作是边  $AD$  平移得到的，即可求得点  $C$  的坐标，即可求得  $BC$  所在直线的解析式，联立两函数解析式，解方程组，即可求解。

**GPT4V:** 首先，我们需要确定矩形 ABCD 的各个顶点的坐标。由于点 A 和点 B 的坐标已知，我们可以利用矩形的性质来确定点 C 和点 D 的坐标。  
 <STEP1> 由于  $AB:BC = 3:2$ ，我们可以设 AB 的长度为  $3x$ ，BC 的长度为  $2x$ 。因为点 A 的坐标为  $(3, 0)$ ，点 B 的坐标为  $(0, 6)$ ，所以 AB 的长度为  $6$ ，BC 的长度为  $4$ 。所以  $x = 2$ 。<STEP2>  
 因此，点 C 的坐标为  $(0, 0)$ ，点 D 的坐标为  $(3, -4)$ 。<STEP3>  
 接下来，我们需要找到反比例函数  $y = k/x$  的  $k$  值。由于图象经过点 D，我们可以将点 D 的坐标代入函数中求解  $k$ 。<STEP4>  
 代入点 D 的坐标  $(3, -4)$  得到  $-4 = k/3$ ，解得  $k = -12$ 。所以反比例函数为  $y = -12/x$ 。<STEP5>  
 最后，我们需要找到点 E 的坐标。由于点 E 在  $y$  轴上，所以它的  $x$  坐标为  $0$ 。将  $x = 0$  代入反比例函数中，我们发现这是不可能的，因为反比例函数的定义域为  $x > 0$ 。因此，我们需要找到反比例函数与 BC 边的交点。<STEP6>  
 由于 BC 边在  $y$  轴上，其方程为  $x = 0$ 。但是反比例函数的定义域为  $x > 0$ ，所以它们不可能有交点。这里似乎有一个矛盾，因为题目中提到反比例函数与边 BC 交于点 E。<STEP7>  
 我们重新审视题目，发现点 E 实际上是反比例函数与边 CD 的交点，而不是与边 BC 的交点。因此，我们需要找到反比例函数与边 CD 的交点。<STEP8>  
 边 CD 的方程为  $y = x - 4$ 。将反比例函数  $y = -12/x$  代入得到  $-12/x = x - 4$ 。解这个方程得到  $x^2 - 4x + 12 = 0$ 。这是一个二次方程，我们可以用求根公式来解它。<STEP9>  
 解得  $x = 2 \pm \sqrt{4 - 12}$ ，这里判别式小于  $0$ ，所以方程没有实数解。这意味着反比例函数与边 CD 实际上没有交点。<STEP10>  
 因此，根据题目描述，似乎存在一些错误或遗漏，因为根据给定的条件和函数，点 E 的坐标无法确定。我们需要更多信息或者更正题目描述才能找到点 E 的坐标。<STEP11>  
**Comment: Reason Error**  
 In <STEP2>, the length of AB is not 6; it should be calculated using the Pythagorean theorem. The model didn't correctly invoke theorem knowledge.

Figure 23: Case from GPT-4V/o. The red ones are marked as generated inference hallucinations.



**Question:** 若函数  $y = \sin(wx + \varphi)$ ， $(w > 0, |\varphi| < \frac{\pi}{2})$  的部分图像如图所示，那么  $w = ?$ ， $\varphi = ?$

**Knowledge:** 已知正弦型函数图象或性质求参数值

**Answer:** 由图像可知函数的一条对称轴为  $x = \frac{5\pi}{12} + \frac{\pi}{24} = \frac{11\pi}{24}$ ，另一条对称轴为  $x = (\frac{\pi}{6} + \frac{3\pi}{24})/2 = \frac{5\pi}{24}$ ，则  $T = 2 \times (\frac{11\pi}{24} - \frac{5\pi}{24}) = \frac{\pi}{2}$ ，故  $w = \frac{2\pi}{T} = 4$ ，故  $y = \sin(4x + \varphi)$ ，由上求得的对称轴知  $\sin(4 \times \frac{5\pi}{24} + \varphi) = 1$ ，又  $|\varphi| < \frac{\pi}{2}$ ， $\therefore \varphi = -\frac{\pi}{3}$

**GPT4V:**  
 首先，我们观察给定的函数图像。根据正弦函数的性质，我们知道它的周期是  $2\pi$ 。<STEP1>  
 从图中可以看出，正弦波从  $x = \frac{\pi}{6}$  开始，到  $x = \frac{5\pi}{12}$  结束，完成了一个完整的周期。因此，周期  $T$  是  $\frac{5\pi}{12} - \frac{\pi}{6} = \frac{\pi}{4}$ 。<STEP2>  
 我们知道周期  $T$  和角频率  $w$  之间的关系是  $T = \frac{2\pi}{w}$ 。由此我们可以解出  $w$ ： $w = \frac{2\pi}{T} = \frac{2\pi}{\pi/4} = 8$ 。<STEP3>  
 接下来，我们需要确定相位  $\varphi$ 。由于正弦波在  $x = \frac{\pi}{6}$  时开始，而标准的  $\sin(x)$  波在  $x = 0$  时开始，所以相位偏移是  $\varphi = \frac{\pi}{6}$ 。<STEP4>  
 综上所述，我们得到  $w = 8$  和  $\varphi = \frac{\pi}{6}$ 。<STEP5>  
 所以答案是： $w = 8$ ， $\varphi = \frac{\pi}{6}$ 。<STEP6>  
**Comment: Reason Error**  
 The model made an incorrect observation about the period of the function graph in <STEP1>, exhibiting perceptual errors and drawing overconfident conclusions.

Figure 24: Case from GPT-4V/o. The red ones are marked as generated inference hallucinations.



甲队		乙队
7	0	8 9
2	6	1 9 7
0	2	2 7 8
1	3	

**Question:** 如图所示的茎叶图记录了甲、乙两支篮球队各6名队员某场比赛的得分数据（单位：分），则下列说法正确的是（ ）  
 A. 甲队数据的中位数大于乙队数据的中位数；  
 B. 甲队数据的平均值小于乙队数据的平均值；  
 C. 甲队数据的标准差大于乙队数据的标准差；  
 D. 乙队数据的第75百分位数为27.

**Knowledge:** 众数、中位数、平均数、极差、方差与标准差、百分位数

**Analysis:** 根据中位数、平均数、方差、百分位数等知识对选项进行分析，从而确定正确答案。

**GPT4V:** 首先，我们需要从茎叶图中提取出甲、乙两队的得分数据。 <STEP1>  
**甲队的得分数据为：7, 16, 20, 22, 23, 31.** <STEP2>  
 乙队的得分数据为：8, 19, 27, 27, 28, 39. <STEP3>  
 接下来，我们分别计算甲、乙两队的中位数、平均值、标准差和第75百分位数。 <STEP4>  
 甲队的中位数是(22+23)/2=22.5. <STEP5>\n乙队的中位数是(27+27)/2=27. <STEP6>  
 甲队的平均值是(7+16+20+22+23+31)/6=19.83（保留两位小数）。 <STEP7>  
 乙队的平均值是(8+19+27+27+28+39)/6=24.67（保留两位小数）。 <STEP8>  
 甲队的标准差需要计算每个数值与平均值的差的平方，然后求和，再除以数据个数，最后取平方根。 <STEP9>  
 乙队的标准差同理。 <STEP10>  
 甲队的标准差是 $\sqrt{((7-19.83)^2+(16-19.83)^2+(20-19.83)^2+(22-19.83)^2+(23-19.83)^2+(31-19.83)^2)/6}$ ≈8.07. <STEP11>  
 乙队的标准差是 $\sqrt{((8-24.67)^2+(19-24.67)^2+(27-24.67)^2+(27-24.67)^2+(28-24.67)^2+(39-24.67)^2)/6}$ ≈10.13. <STEP12>  
 甲队的第75百分位数是在23和31之间，可以通过插值法计算，但题目没有要求。 <STEP13>  
 乙队的第75百分位数是在27和28之间，可以通过插值法计算，但题目没有要求。 <STEP14>  
 根据以上计算，我们可以得出：A选项错误，因为甲队的中位数是22.5，而乙队的中位数是27。 <STEP15>  
 B选项错误，因为甲队的平均值是19.83，而乙队的平均值是24.67。 <STEP16>  
 C选项正确，因为甲队的标准差是8.07，而乙队的标准差是10.13。 <STEP17>  
 D选项错误，因为乙队的第75百分位数是在27和28之间，不是准确的27。 <STEP18>  
 所以最终的答案是C选项。 <STEP19>  
**Comment: Visual Perception Error**  
 Incorrect scoring data was obtained from the stem-and-leaf plot, making it difficult to interpret the content of the plot.

Figure 25: Case from GPT-4V/o. The red ones are marked as generated inference hallucinations.

图表1

**Question:** 张家口市为迎接冬奥会，计划招募各类志愿者2000人。为做好宣传工作，招募小组对20-45岁的人群随机抽取了100人，回答“冬奥会”的有关知识，根据统计结果制作了如下的统计图表1、表2，分别求出表2中的a, x的值。

组号	按年龄分组	回答完全正确人数	回答完全正确人数占本组频率
1	[20,25)	5	0.5
2	[25,30)	18	x
3	[30,35)	a	0.9
4	[35,40)	9	0.36
5	[40,45)	3	0.2

**Knowledge:** 众数、中位数、平均数、极差、方差与标准差、百分位数

**GPT4V:**  
 求解a和x的值：首先，我们需要理解表中的“回答完全正确人数占本组频率”这一列的含义。这个频率是指在该年龄组中回答完全正确的人数占该年龄组总人数的比例。对于a，我们知道在30-35岁组中的频率是0.9，这意味着在这个年龄组中，有90%的人回答完全正确。由于这个年龄组的总人数是100人的一部分，我们可以通过以下步骤求解a：设30-35岁组的总人数为b，则有a/b=0.9。 <STEP1>  
 由于总人数是100人，我们可以通过其他组的人数来计算b。 <STEP2>  
 我们知道其他组的人数分别是：20-25岁组5人，25-30岁组未知，35-40岁组9人，40-45岁组3人。 <STEP3>  
 所以b=100-(5+未知+9+3)。 <STEP4>  
 由于我们不知道25-30岁组的人数，我们需要先求出x。 <STEP5>  
 对于x，我们知道在25-30岁组中的频率是x，这意味着在这个年龄组中，有x\*100%的人回答完全正确。由于回答完全正确的人数是18，我们可以通过以下步骤求解x：设25-30岁组的总人数为c，则有18/c=x。 <STEP6>  
 由于我们不知道c的具体数值，我们需要通过其他组的人数来计算c。 <STEP7>  
 我们知道其他组的人数分别是：20-25岁组5人，30-35岁组未知，35-40岁组9人，40-45岁组3人。 <STEP8>  
 所以c=100-(5+未知+9+3)。 <STEP9>由于未知的部分是30-35岁组的人数b，我们可以将c表示为c=100-(5+b+9+3)。 <STEP10>  
 现在我们有方程：a/b=0.9和18/c=x。 <STEP11>  
 我们可以通过解这两个方程来求解a和x。 <STEP12>  
**由于我们知道b+c=100-(5+9+3)=83，我们可以将b表示为b=83-c。 <STEP13>**  
 将b代入a/b=0.9得到a=(0.9)(83-c)。 <STEP14>将c代入18/c=x得到x=18/(83-a)。 <STEP15>现在我们有包含a和c的方程，我们可以通过解这两个方程来求解a和x。 <STEP16>由于a=(0.9)(83-c)，我们可以将a表示为a=74.7-0.9c。 <STEP17>将a代入x=18/(83-a)得到x=18/(83-(74.7-0.9c))。 <STEP18>解这个方程得到c=20，所以x=18/20=0.9。 <STEP19>将c代入a=74.7-0.9c得到a=74.7-0.9\*20=56。 <STEP20> 所以a的值是56，x的值是0.9。  
**Comment: Reason Error**  
 Misunderstanding of the concept of frequency and sampling. The sum of b and c does not equal 83.

Figure 26: Case from GPT-4V/o. The red ones are marked as generated inference hallucinations.

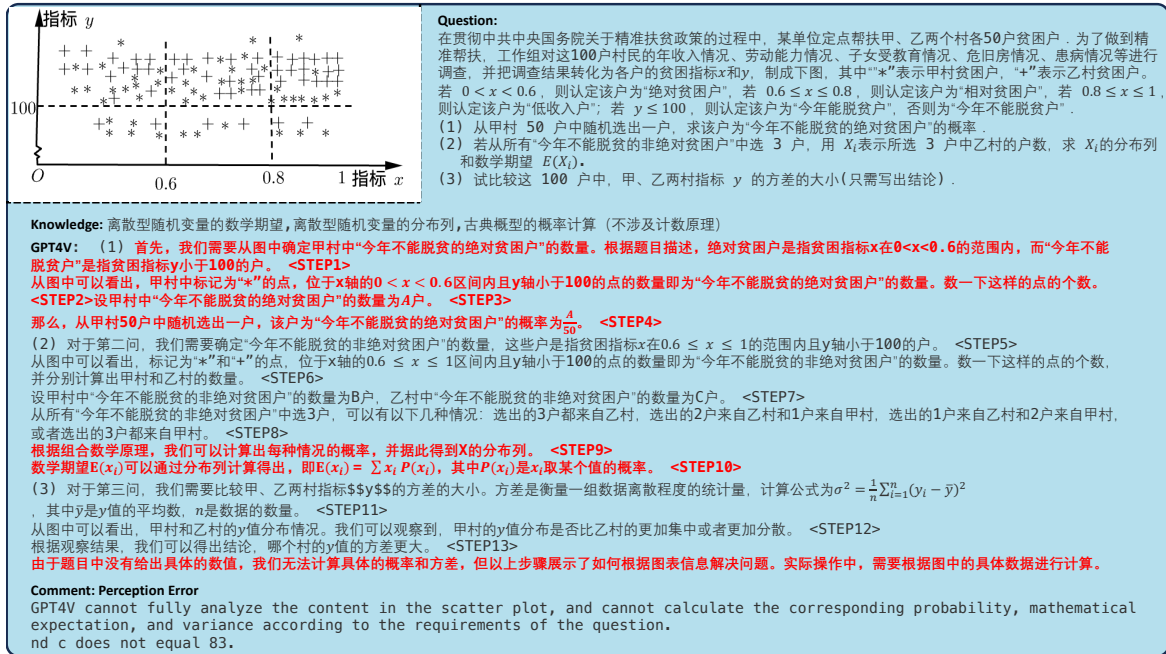


Figure 27: Case from GPT-4V/o. The red ones are marked as generated inference hallucinations.

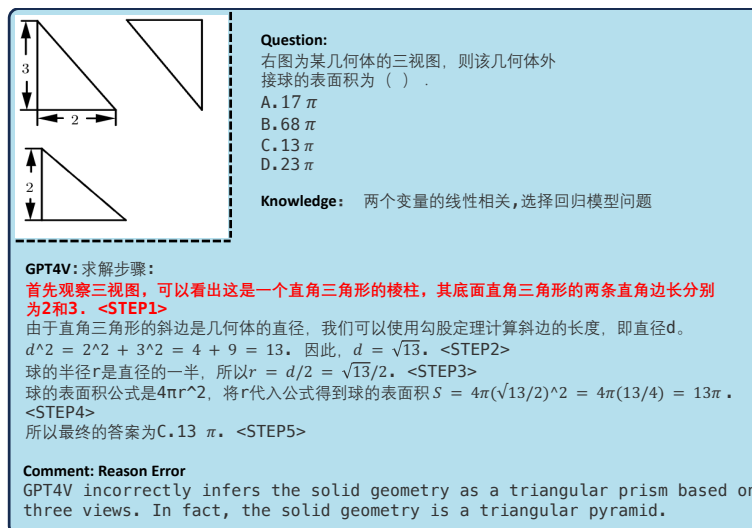


Figure 28: Case from GPT-4V/o. The red ones are marked as generated inference hallucinations.

**Question:**  
如图, 将一个直角三角形纸片ABO, 放置在平面直角坐标中, 点A(-8,0), 点B(0,6), 点O(0,0), C是OA上的点, CD⊥AB, 沿着CD折叠该纸片, 点A正好落在B处, 则线段CD的长为? .

**Knowledge:** 轴对称, 几何变换

**GPT4V: 求解步骤:**  
首先, 我们需要找到点C的坐标。由于C在OA上, OA是x轴的一部分, 所以C的y坐标是0。我们设C的x坐标为x。 <STEP1>  
接下来, 我们需要找到点D的坐标。由于CD垂直于AB, 我们可以使用斜率的概念。直线AB的斜率是 $(6-0)/(0-(-8))=6/8=3/4$ 。因此, CD的斜率是 $-4/3$  (垂直线段斜率的乘积为-1)。 <STEP2>  
由于点D在AB上, 我们可以用点斜式方程来表示直线AB:  $y - 6 = (3/4)(x - 0)$ 。  
<STEP3>  
点D也在直线CD上, 其方程可以表示为 $y - 0 = (-4/3)(x - x)$ , 其中x是点C的x坐标。  
<STEP4>  
由于点A在折叠后落在点B处, 点D实际上是点A和点B的中点。因此, D的坐标是 $((-8+0)/2, (0+6)/2)=(-4, 3)$ 。 <STEP5>  
现在我们有了解点D的坐标, 我们可以解出点C的x坐标。将D的坐标代入直线CD的方程中, 我们得到 $3 = (-4/3)(-4 - x)$ 。解这个方程, 我们得到 $x = -4 - (3/(-4/3)) = -4 + 3/(4/3) = -4 + 9/4 = -7/4$ 。 <STEP6>  
因此, 点C的坐标是 $(-7/4, 0)$ 。 <STEP7>  
最后, 我们可以计算CD的长度。由于C和D的y坐标相同, CD的长度就是D和C的x坐标之差的绝对值。所以CD的长度是 $|-4 - (-7/4)| = |(-16/4) + (7/4)| = |(-9/4)| = 9/4$ 。 <STEP8>  
所以线段CD的长度是9/4。 <STEP9>

**Comment: Perception Error**  
Wrong perception of the slope of CD. The length of CD should be calculated using the distance formula between two points.

Figure 29: Case from GPT-4V/o. The red ones are marked as generated inference hallucinations.

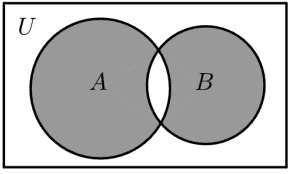
**Question:**  
如图, 在边长为4的等边三角形ABC中, D, E分别是边BC, AC的中点, DF ⊥ AB于点F, 连接EF, 则EF的长为?

**Knowledge:** 等腰三角形的性质

**GPT4V: 求解步骤:**  
首先, 由于D和E分别是边BC和AC的中点, 根据中线定理, DE的长度是BC的一半, 即 $DE=BC/2=4/2=2$ 。 <STEP1>  
接着, 因为三角形ABC是等边三角形, 所以 $AB=AC=BC=4$ 。 <STEP2>  
由于DF垂直于AB, 并且D是BC的中点, 所以F是AB的中点,  $AF=FB=AB/2=4/2=2$ 。 <STEP3>  
在直角三角形ADF中, 根据勾股定理, 我们有 $DF^2 = AD^2 - AF^2$ 。  
<STEP4>  
由于AD是等边三角形ABC的高, 可以通过勾股定理计算出AD的长度。在等边三角形中, 高也是中线, 所以 $AD=BD=CD$ , 且 $AD^2 = AB^2 - (AB/2)^2 = 4^2 - (4/2)^2 = 16 - 4 = 12$ 。 <STEP5>  
因此,  $DF = \sqrt{AD^2 - AF^2} = \sqrt{12 - 2^2} = \sqrt{12 - 4} = \sqrt{8} = 2\sqrt{2}$ 。 <STEP6>  
现在我们有了解直角三角形DEF, 其中 $DE=2$ ,  $DF=2\sqrt{2}$ , 我们可以使用勾股定理来找到EF的长度。 <STEP7>  
 $EF^2 = DE^2 + DF^2 = 2^2 + (2\sqrt{2})^2 = 4 + 8 = 12$ 。  
<STEP8>  
所以,  $EF = \sqrt{12} = 2\sqrt{3}$ 。 <STEP9>因此, EF的长度是 $2\sqrt{3}$ 。  
<STEP10>

**Comment: Reason Error**  
GPT4V misunderstood the role corresponding to point F. Point F is a vertical foot on AB rather than the midpoint.

Figure 30: Case from GPT-4V/o. The red ones are marked as generated inference hallucinations.



**Question:**已知全集  $U = R$ , 集合  $A = \{x|0 \leq x \leq 2\}$ ,  $B = \{x|x^2 - x > 0\}$ , 则图中的阴影部分表示的集合为?

A.  $\{x|x \leq 1 \text{ 或 } x > 2\}$   
 B.  $\{x|x < 0 \text{ 或 } 1 < x < 2\}$   
 C.  $\{x|1 \leq x < 2\}$   
 D.  $\{x|1 < x \leq 2\}$

**Knowledge:** 集合的概念与表示方法

**GPT4V: 求解步骤:**  
 首先解集合B的不等式  $x^2 - x > 0$ , 可以分解为  $x(x - 1) > 0$ , 得到  $x > 1$  或  $x < 0$ 。 <STEP1>  
 集合A是从0到2的闭区间, 即  $0 \leq x \leq 2$ 。 <STEP2>  
**图中阴影部分表示的是集合A和B的交集, 即  $A \cap B$ 。** <STEP3>  
 根据集合A和B的定义, 交集部分是  $1 < x \leq 2$ 。 <STEP4>  
 所以最终的答案是D.  $\{x|1 < x \leq 2\}$ 。 <STEP5>

**Comment: Reason Error**  
 Wrong understanding of the contents of the shaded part. The intersection of sets A and B does not represent the area of the shaded part.

Figure 31: Case from GPT-4V/o. The red ones are marked as generated inference hallucinations.