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

Anonymous ACL submission

Abstract

Due to the rapid advancements in multimodal large language models, evaluating their multi-modal mathematical capabilities continues to receive wide attention. Despite the datasets like MathVista proposed benchmarks for assessing mathematical capabilities in multimodal scenarios, there is still a lack of corresponding evaluation tools and datasets for fine-grained assessment in the context of K12 education in Chinese language. To systematically evaluate the capability of multimodal large models in solving Chinese multimodal mathematical problems, we propose a Chinese Multi-modal Math Skill Evaluation Benchmark, named CM-MaTH, contraining 23k multimodal K12 math related questions, forming the largest Chinese multimodal mathematical problem benchmark to date. CMMaTH questions from elementary to high school levels, provide increased diversity in problem types, solution objectives, visual elements, detailed knowledge points, and standard solution annotations. We have constructed an open-source tool GradeGPT integrated with the CMMaTH dataset, facilitating stable, rapid, and cost-free model evaluation. Our data and code are available[1]

1 Introduction

Large language models(LLMs) excel in various language tasks, while multimodal models effectively handle visual-language problems. They advance natural language processing and computer vision fields, providing powerful solutions for complex tasks. Multimodal large models demonstrate potential as versatile solvers for multimodal problems.

The systematic evaluation of large models' performance across various mathematical reasoning scenarios has been a subject of extensive research. GSM8K and MATH(Cobbe et al., 2021; Hendrycks et al., 2021b) assessed the ability in 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. (2024b) systematically evaluated the ability of large multimodal models to perceive visual elements and solve corresponding multimodal problems. Shi et al. (2023) 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. (2024a) constructed multidisciplinary Chinese question answering datasets C-Eval and CM-MMU to evaluate the knowledge and reasoning abilities of multimodal large models. However, C-Eval lacks evaluation in multimodal contexts, while CMMMU's dataset has relatively low diversity, consisting of only 540 questions.

Existing Math benchmarks for answer evaluation can be categorized into two types:*Rule-based* (Cobbe et al., 2021; Hendrycks et al., 2021b; He et al., 2024) and *API-based* methods (Lu et al., 2023; Zhang et al., 2024b; Hendrycks et al., 2021a). *API-based* methods are very costly and time-consuming, and they often result in unstable and inconsistent evaluation results. *Rule-based* methods, on the other hand, struggle to handle 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

[1]https://anonymous.4open.science/r/CMMaTH-396B

Figure 1: 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.

answers.

Based on above considerations, we propose a new multimodal mathematical benchmark CM-MaTH. Compared to previous benchmarks, our benchmark demonstrates greater diversity, increased depth of reasoning, and finer-grained knowledge annotation for multimodal models to grasp different levels and types of knowledge. We provided and open-sourced a 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 evaluated mainstream open-source and commercial multimodal large models in Table 4, reporting comprehensive evaluation results along with extensive case analyses. In summary, our paper makes the following contributions:

- We introduce the largest high-quality Chinese multimodal mathematics benchmark with the most detailed annotation granularity to date. We also provide an English version of this dataset. The CMMaTH dataset is a dynamically maintained and will be periodically updated.
- Compared to previous multimodal mathematical benchmarks, our dataset exhibits 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 annotate each question with detailed knowledge points and corresponding skills to evaluate the mastery level of current large models.
- We build an evaluation assistant named GradeGPT on the CMMaTH dataset, which allows for comparing the proximity of model responses to standard answers and assessing the correctness of results and processes. GradeGPT features lightweight open-source characteristics, avoiding the instability and high costs associated with commercial models.
- We conduct a systematic evaluation of existing mainstream multimodal large models, quantitatively and qualitatively comparing with existing models.

2 Related Work

2.1 Assessment of mathematical abilities

To evaluate the performance of large models in mathematical reasoning and examine hallucinations during the reasoning process, numerous benchmarks 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., 2024c) constructed several subsets of datasets to assess whether existing multimodal large models can truly understand mathematical abstract forms.

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). 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, MathVista(Lu et al., 2023) and GeoEval(Zhang et al., 2024b) 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., 2023b; 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 mathematics problems.

Statistic	Number
Total questions	23856
- multiple-choice questions	18191
- Free-form questions	5665
- Questions in the testmini set	1000
Single-choice questions	13706(75.3%)
- 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%)
Multiple-choice & Multi-turn questions	4485(24.7%)
knowledge point number	2299
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 1: Key statistics of CMMaTH. The unit of question length is words.

Unlike PandaLM(Wang et al., 2023b) trying to evalution relative conciseness, clarity and so on, 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 23856 items across 13 visual themes, 5 difficulty levels, and encompassing 150 types of knowledge points. More detailed statistical data can be found in Table 1.

For the convenience of evaluation, we provide a miniaturized test set of CMMaTH, called CMMaTH-testmin, containing 1500 samples. Testmin retains the diversity of the CMMaTH dataset and shows similar overall performance to the entire CMMaTH dataset. Evaluators can conduct quick

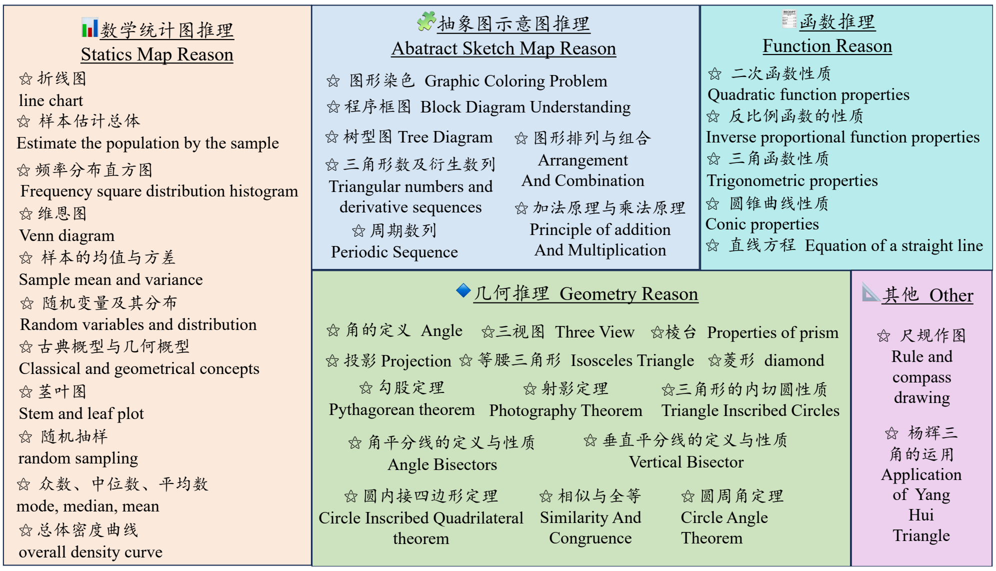


Figure 2: Some of the knowledge points involved in the CMMaTH dataset.

tests and generate preliminary analyses based on CMMaTH-testmin.

3.2 Collectioin Guidelines

We collected a large number of multimodal mathematics questions from a vast K12 educational question bank, including elements such as statistical charts, plane geometry, three-view diagrams, flowcharts, set notation diagrams, etc. 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 2 shows some visual elements subject of CMMaTH.

- High relevance to the K12 math knowledge and skill. The annotator, who is well-versed in knowledge, needs to ensure that the multimodal question assesses a specific K-12 mathematics knowledge point during the question collection process. It primarily includes mathematics questions related to K12 education, facilitating the assessment of the application potential of large-scale multimodal capabilities in the field of 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 data is based on a million-level private database. The private database we used comes from questions collected from the Internet and undergoes rigorous data checking. The project's data has undergone multiple rounds of collection. We first sampled 45,000 multimodal math questions: 14,000 each from elementary, high, and junior high schools. Then, we added 34,000 more questions featuring algorithm block diagrams, statistics, and geometry diagrams to enhance visual diversity.
Data Filtering We filtered out all questions without images in the question stems, including questions with multi-graph reasoning, questions in non-Chinese languages, and questions not relying on visual content to solve. To ensure the quality of the images and text questions, we removed all images whose width and height were less than 100, then used the GPT4 API to score the data quality and filter out questions suspected of being unsolvable and questions with garbled text in the question text.
Data Labeling For K-12 mathematics knowledge points, we have scraped the mathematics section from Jiaoyan Cloud[2] and organized all the knowledge points into a knowledge tree including a total of 5,531 knowledge points. We retained 2,299 knowledge points more relevant to multimodal

[2] https://www.jiaoyanyun.com/

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

Table 2: Primary element types involved in the CMMaTH dataset.

mathematics in K-12. Subsequently, all questions were classified according to knowledge points by GPT-4 and a fine-tuned LLM, followed by manual multi-level verification. Questions that did not match any K-12 multimodal mathematics knowledge points were filtered out.

3.4 Comparison with Existing Benchmarks

The CMMaTH dataset is primarily used to evaluate multimodal reasoning capabilities in K-12 educational scenarios. We compared the current mainstream multimodal mathematical datasets and large model benchmarks in Table 3. Compared to existing multimodal benchmarks and multimodal reasoning benchmarks, the CMMaTH dataset has the following characteristics:

Extreme Diversity Currently, there is a severe lack of high-quality Chinese multimodal mathematics datasets. MATH-VISION lacks a Chinese component, the MATH-VISTA dataset contains only a small number of Chinese samples, and CMMMU contains only 540 math problems, which are not fine-grained and comprehensive enough. We have included about 23k fine-grained multimodal mathematics assessment samples, covering 13 K12 mathematics visual categories, making it the largest known multimodal Chinese dataset to date.

Real and High Quality & Multilingual 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 an Olympiad-level bilingual multimodal benchmark. However, this benchmark is overly challenging and deviates from the application of LMM in real K12 multimodal math scenarios. Additionally, the variety of multimodal visual elements is relatively limited. Instead, we collect multimodal data specifically tailored to the K12 education context. Additionally, MathVista incorporates a significant amount of data from GeoQA and synthetic images, which have relatively poor image quality. Our multimodal visual image elements have all undergone stringent image quality assessments. Unlike CMMMU, CEval, and CMath, our dataset is a bilingual dataset that considers a large number of Chinese scenes. In addition to the text of the questions being in Chinese, the visual elements related to the questions also contain Chinese text/symbols.

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 K-12 grade levels. This fine-grained annotation enables a more nuanced evaluation of multimodal mathematical proficiency within the K-12 educational context. While 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.

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.

Dataset	Size	Image&Supplementary Input	Format	Source	Answer	Knowledge Annotation	Lanugage Domain	Knowledge Domain
VQAv2(Goyal et al., 2017)	> 1M	V	I+T	Annotated	Open/MC/TF	✗	En	General
SEED(Li et al., 2023a)	19K	V	I+T	Annotated	MC	✗	En	General
MMBench(Liu et al., 2023)	3K	V	I+T	Repurposed	MC	✗	En	General
MM-Vet(Yu et al., 2023)	0.2K	V	I+T	Annotated	Open	✗	En	General
ScienceQA(Lu et al., 2022)	6K	V	I+T	Textbooks	MC	✗	En	Science
MathVista(Lu et al., 2023)	1K/6K	V(5 Types)+OC	I+T	Synthesized	Open/MC/TF	✗	En/ZH	Math
MMMU(Yue et al., 2023)	11.5K	V(30 Types)+OC	Interleaved	Textbooks	Open/MC	✗	–	General
CMMMU(Zhang et al., 2024a)	< 1K(Math Part)	V(5 Types)+OC	Interleaved	Internet	Open/MC	✗	ZH	General
OlympiadBench(He et al., 2024)	6.5K(Math Part)	V(5 Types)	Interleaved	Internet	Open	✗	ZH/EH	Math/Physics
MathVerse(Zhang et al., 2024c)	2.6K/15K	V(3 Types)	I+T	Synthesized	MC	✗	ZH/EH	Math
MATH-Vision(Wang et al., 2024)	3K	V(16 Types)+IC	I+T	Synthesized	Open/MC	✗	EH	Math
CMMaTH	23K	V(13 Types)+OC+IC	I+T	Internet/Annotated	Open/MC/TF	✓	ZH	K12 Math

Table 3: Comparison with other multimodal benchmarks. V: visual input, VD: video input, OC: optical characters, IC: Image Caption, I+T: images and text, Open: open questions, MC: multiple choice questions, FIB: fill in the blank questions, TF: true or false questions.

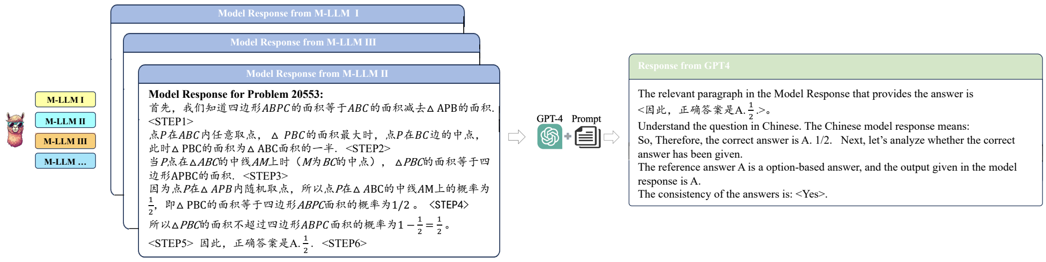


Figure 3: Instruction Construction Pipeline of GradeGPT

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 GPT4 API.

Prompt Format

In the prompt input of GradeGPT, there are "questions," "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 for the purpose of determining answer consistency. Cross-Lingual-Judge-of-Chain 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 E

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 GPT4's In-Context Learning, as showned 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 are able to obtain an expert model, GradeGPT, which possesses the capability to compare answers.

5 Experiments

We conducted a series of experiment to evaluate various models on the CMMaTH dataset. We evaluated various LLM/LMM models, including open-source and closed-source models. More model details can be found in Table 13. We employed a method similar to GeoEval and MathVista, generating captions through an GPT4V, and assessed them using MetaMath, and DeepSeekMath equipped with caption information. Our empirical research reveals that even the most advanced models struggle to achieve satisfactory accuracy levels. Furthermore, we conducted an exhaustive error analysis on

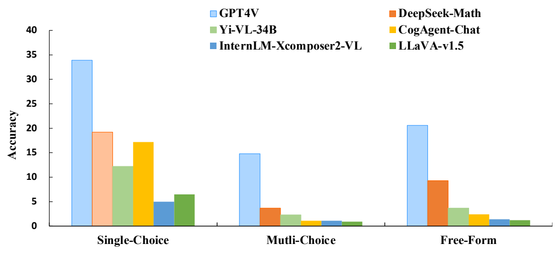


Figure 4: Accuracy of LMMs across different types of problems in CMMaTH Benchmark.

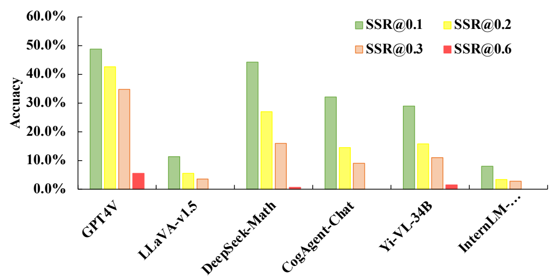


Figure 5: The metrics of different LMMs/LLMs models about SSR.

a sufficiently strong commercial multimodal model, GPT-4V, examining its error distribution and presenting illustrative qualitative examples. Our investigation also revealed that the inclusion of multilingual thought chains does not mitigate the substantial difficulties presented by Chinese multimodal mathematical reasoning scenarios. We postulate that the richness of non-English contextual information contained within the images necessitates models equipped with enhanced multilingual OCR and sophisticated multimodal diagram reasoning capabilities.

5.1 Main Experiments on LLM/LMMs

We evaluated the results of mainstream multimodal large models and mathematical expert models in Table B. 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 in Table 4 results indicates that our data exhibits extremely strong diversity and relatively challenging reasoning depth. Figure 1 and Table 4 shows models such as GPT4V struggle to comprehend our multimodal content and reasoning questions effectively, resulting in significant performance gaps between open-source and proprietary models. In certain rare visual domains, multimodal large models achieve very low reasoning outcomes.

Accuracy on various question types. We evaluated the accuracy of GPT4V on various target-solving tasks in Figure 4. The results indicate that when solving free-form problems, especially those with more diverse targets such as expressions, coordinates, and conclusion judgments, the multimodal large language model shows poorer performance.

Is OCR information sufficient for CMMaTH? We also referred to works like MathVista, attempting to use LLMs combined with OCR information from diagrams to assist in mathematical reasoning in Table 4. We found that, in our benchmark, a small amount of OCR information (such as mathematical symbols in diagrams, axis values, and image titles) made it very difficult to complete our multimodal mathematical reasoning tasks. The results indicate that solving problems in CMMaTH requires stronger multimodal mathematical chart capabilities, beyond just OCR.

K12 Multimodal Knowledge Richness of current LMMs. We systematically evaluated the proficiency of existing multimodal large models in the K12 domain regarding multimodal reasoning skills in Figure 5. The results revealed a significant knowledge gap in existing multimodal K12 educational resources. Compared to other existing LMMs, GPT4V possesses a richer knowledge base, thereby substantially reducing the illusion of reasoning in multimodal mathematical inference.

5.2 Experiments of Cross-language Reason 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 the Table 5.

5.3 Error Analysis

We conducted a detailed analysis and evaluation of GPT4V on CMMaTH-testmin, categorizing errors

Model	Overall	Flow	Bar	Scatter	Line Plot	Fan	LiDAR	Visual-Table	Three View	Folded Image	Analytic	Solid	Plane	Venn	Abt-Analogy
LLMs (Text Only)															
LLama2-70B	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
MetaMath-70B	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	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
Baichuan-13B	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	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
Math LLMs (Text + OCR Caption)															
LLama2-70B	5.6	4.9	2.3	4.8	7.9	7.1	8.0	4.4	6.4	9.1	3.3	4.8	6.3	6.9	5.5
MetaMath-70B	5.1	4.3	3.2	6.9	8.1	5.3	0.0	4.4	4.2	8.8	7.1	4.4	8.3	9.1	5.2
DeepSeek-Math	15.3	13.2	6.9	14.1	12.6	12.3	12.1	8.9	14.4	14.1	17.9	19.3	22.7	21.5	13.9
Baichuan-13B	8.1	6.9	4.3	12.4	11.5	12.3	14.9	3.4	4.4	9.3	11.6	6.8	13.2	12.9	9.9
Qwen-14B	13.3	14.1	7.4	13.3	16.2	13.2	11.8	10.6	11.8	19.8	5.9	11.7	13.8	21.4	16.3
Open-source LMMs (Text + Image)															
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
InternLM-XComposer2-VL	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
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
CogAgent-Chat	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
Closed-source LMMs (Text + Image)															
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
GPT4o	35.2	59.4	18.8	54.5	31.7	58.4	32.4	31.7	28.7	23.8	40.6	31.6	33.6	57.4	29.7
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 4: Comparison of model performances across various mathematical subjects. 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: Abtract Analogy Graph.

LMM	Overall-Acc
LLaVA-v15	4.2
InternLM-XComposer2-VL	3.4
LLaVA-v15 + En-CoT	9.4
InternLM-XComposer2-VL + En-CoT	16.9
LLaVA-v15 + CLP	12.7
InternLM-XComposer2-VL + CLP	17.1

Table 5: The performance of train-free CoT reasoning techniques on the CMMaTH dataset.

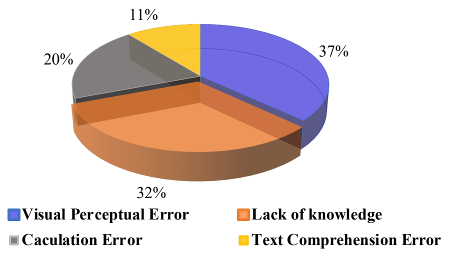


Figure 6: Distribution of Error Types in GPT4V.

into four types: perceptual errors, reasoning errors, calculation errors, and Reject Errors. The error type distribution of GPT4V on CMMaTH is shown in the Figure 6.

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 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.

6 Conclusions

We introduce CMMaTH, a detailed Chinese math reasoning benchmark with diverse question types, vivid visuals, and complex reasoning. The benchmark includes detailed knowledge points, standard thought processes, and grade levels to measure the mastery of knowledge points in the K-12 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 in multilingual, multimodal math reasoning.

Limitation & Potential Impact

Our dataset CMMaTH, as a multimodal mathematics dataset aimed at the K-12 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. CMMaTH primarily consists of single-image problems, without considering multi-image contextual reasoning or scenarios requiring auxiliary line drawing and similar tasks. 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 can continue to be explored in the future.

References

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.

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*.

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*.

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.

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.

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. 2023. 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*.

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.

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. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. *CoRR*, abs/2305.17926.

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. 2023b. Pandalm: An automatic evaluation benchmark for LLM instruction tuning optimization. *CoRR*, abs/2306.05087.

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 Wenhu Chen. 2023. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. *CoRR*, abs/2311.16502.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, Haoran Zhang, Xingwei Qu, Junjie Wang, Ruibin Yuan, Yizhi Li, Zekun Wang, Yudong Liu, Yu-Hsuan Tsai, Fengji Zhang, Chenghua Lin, Wenhao Huang, Wenhu Chen, and Jie Fu. 2024a. CMMMU: 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. 2024b. Geoeval: Benchmark for evaluating llms and multimodal models on geometry problem-solving. *CoRR*, abs/2402.10104.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024c. 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.

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 About Multimodal Large Model Evaluation

The multimodal large models face serious hallucination issues in perceiving objects and executing inference. 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., 2024a) are comprehensive subject datasets designed 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.

B Model Generation Details

B.1 Model Weight Version

We evaluated models on CMMaTH, including open-source models such as LLaVA-v1.5, Deepseek-Math, InternLM-XComposer2-VL, Yi-VL-34B, CogAgent-Chat, MetaMath-70B, LLama-70B, Baichuan-13B and Qwen-14B as well as state-of-the-art commercial models GPT4V. We have listed the parameter versions and the Hugging Face repository names of the open-source models used in Table 12.

B.2 Model Sampling Parameter

We have listed the corresponding hyperparameters used by the models in Table 13. For API models, we have indicated the corresponding release versions. Models using vLLM for inference are annotated.

B.3 Data quality control

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.

C Prompt Details

C.1 Prompt For Step Response Generation

When evaluating hallucinations during the assessment process, we use a few-shot prompt format to elicit step-by-step outputs from the model as showed in Table 6.

C.2 Prompt For GradeGPT

We also listed the prompts used by GradeGPT in Tables 7.

C.3 Prompt For Cross-Lingual Prompting and En-CoT

We have listed the specific prompts used for En-CoT and Cross-Lingual Prompt during actual execution in Table 11. 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.

D CMMaTH Dataset Details

D.1 Data Collection Details

To more clearly elucidate our data collection process, we have depicted the overall pipeline of data collection in Figure 7.

D.2 Knowledge Point Details

We provided detailed annotations of knowledge points for our dataset and conducted preliminary clustering of these knowledge points. The distribution of knowledge points in different clusters is as follows: We have formulated a Knowledge **S**uccessful **S**olve **R**ate*(SSR)* 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 point of CMMaTH. Acc_{kn_i} is the $Acc_{outcome}$ of questions about i'th knowledge point. I denotes an indicator function.

$$SSR@\alpha = \frac{\sum_{i=1}^{N_{kn}} I(Acc_{kn_i} > \alpha)}{N_{kn}} \tag{1}$$

It is our contention 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 the purpose of 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.

D.3 Characteristics Of Annotators

We utilized a standard team of four 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 GradeGPT details

E.1 GradeGPT Prompt Detail

We have listed detailed Fewshot Examples using the GPT4-generated GradeGPT model responses in Table 11. Through this table, you can observe the specific form of the Cross-Lingual-Judge-of-Chain that we have used.

E.2 GradeGPT Performance Metric

GradeGPT 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

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	**Question:** 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。 **Model response:** 首先计算小明买苹果的开销为2乘10等于20元. <STEP1> 用2000元减去所有的开销, 2000-20=1980. <STEP2> 所以最终的答案为1980元. <STEP3>
Example2	**Question:** 如图，在△ABC中，∠C=90°，∠B=30°，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$\frac{1}{2}$MN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∠BAC的平分线；②∠ADC=60°；③点D在AB的中垂线上；④BD=2CD．A.4 B.3 C.2 D.1 **Model response:** 首先确定△ABC的几何性质。<STEP1>∠C=90°，∠B=30°，∠A=120°。<STEP2>△ABC是一个等腰三角形。<STEP3>根据等腰三角形的性质，AD是∠BAC的平分线。<STEP4>①AD是∠BAC的平分线，正确。<STEP5>② ∠ADC=60°，错误。<STEP6>③ 点D在AB的中垂线上，错误。<STEP7>④BD=2CD，正确。<STEP8>因此，正确的个数为4。<STEP9>答案为A.4。<STEP10>请你根据这个例子，解决下面的数学题。问题：在△ABC中，∠C=90°，∠B=30°，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$\frac{1}{2}$MN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∠BAC的平分线；②∠ADC=60°；③点D在AB的中垂线上；④BD=2CD．A.4 B.3 C.2 D.1 求解步骤：首先确定△ABC的几何性质。∠C=90°，∠B=30°，∠A=120°。△ABC是一个等腰三角形。根据等腰三角形的性质，AD是∠BAC的平分线。①AD是∠BAC的平分线，正确。②∠ADC=60°，错误。③点D在AB的中垂线上，错误。④BD=2CD，正确。因此，正确的个数为4。答案为A.4。

Table 6: Prompt for all model to generate step-by-step answer.

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	**Question:** 如图所示，AD是△ABC的中线，点E在AC上，BE交AD于点F．当$\frac{AF}{FD}=\frac{1}{2}$时，$\frac{AE}{AC}=$ **Model Response:** 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，$AD^2 = AD \times AC^2$。设$AD = x, AC = y$，则有$x^2 = xy^2$。<STEP1> 解这个方程，我们得到 $y = \sqrt{x}$。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。 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$F = (x+y)/2$。<STEP2> 现在我们可以计算AF和FD的长度。$AF = x - F, FD = FD - F$ 将已知的x和y代入公式，得到：$AF = x - (x+y)/2FD = (x+y)/2 - (x+y)/2$ 简化得到：$AF = x - xFD = x - x$ 这意味着$AF = FD$。<STEP3> 现在我们可以求解题目中的条件：$AE/AC = 1/2$ 将已知的AF和FD代入公式，得到：$AE/AC = (x-x)/(\sqrt{x}y)$化简得到：$AE/AC = 1/2$所以，答案为$AE/AC = 1/2$ <STEP4> **Reference Response:** $\frac{4\pi}{3} - 2\sqrt{3}$ **Judge Result:** The relevant paragraph in the Model Response that provides the answer is 所以，答案为AE / AC = 1 / 2. Understand the question in Chinese. The Chinese model response means: Therefore, the answer is AE / AC = 1 / 2. Next, let's analyze whether the correct answer has been given. The reference answer $\frac{1}{5}$ is a numerical answer, and the output given in the model response is 1/2. The consistency of the answers is: <No>.

Table 7: Detail prompt for GradeGPT and it's instruction construction prompt by GPT4.

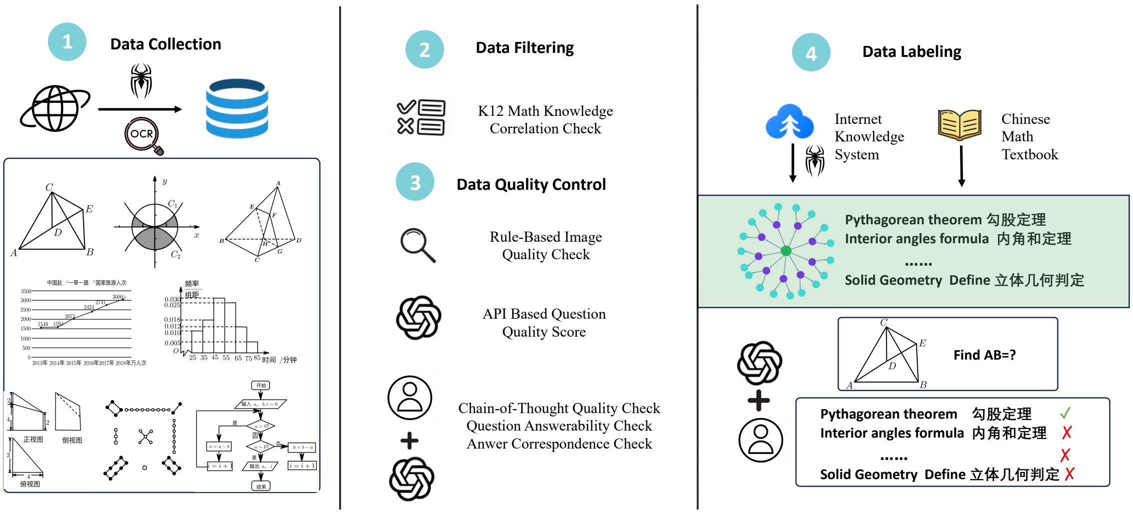


Figure 7: Overall Data Collection Pipeline of CMMaTH.

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.

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

E.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 9.

E.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.

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
GPT4(4-Shot)	97.2

Table 8: Ablation study on the instruction fine-tuning of GradeGPT commands

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 GPT4 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 GPT4. The proposed strategy can significantly enhance GradeGPT's ability to judge results. It is only slightly weaker than the performance of GPT4(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 GPT4.

The Baseline we compared, Qwen-7B/14B(4-Shot), GPT4(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

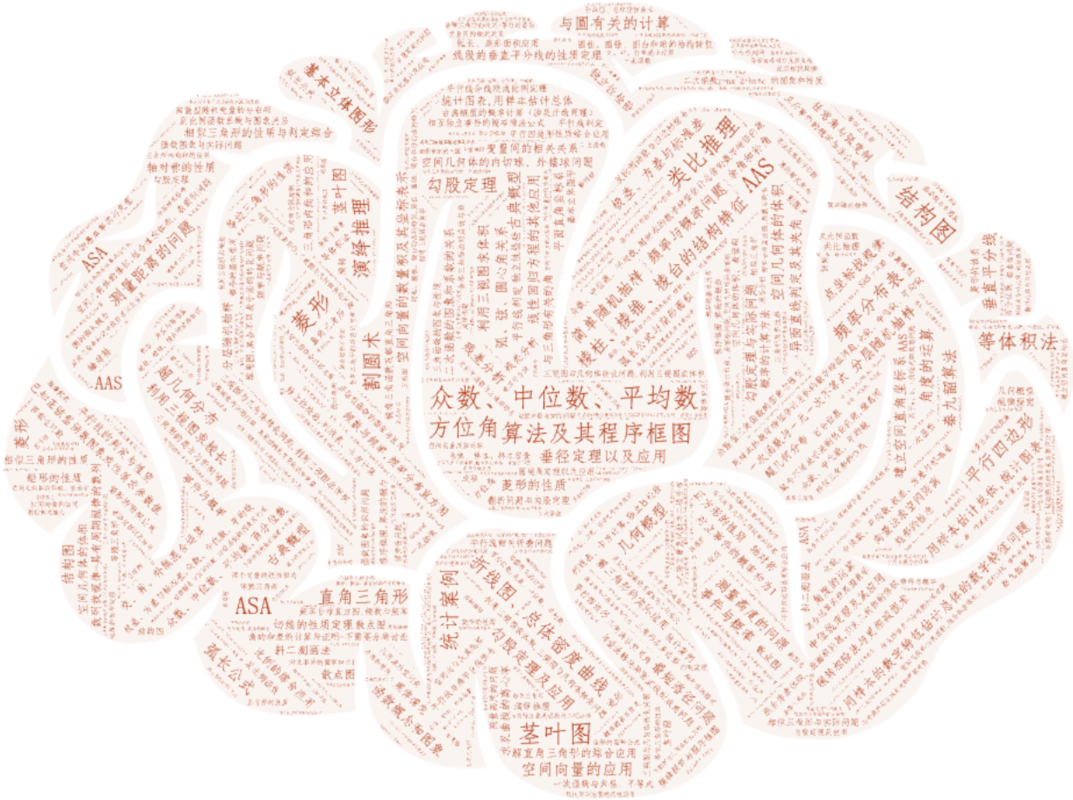


Figure 8: Cloud diagram of the knowledge points contained in the CMMaTH dataset.

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 outputs 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 GPT4.

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

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, which showed the impact of using different LLM models on constructing diverse and effective instruction data.

F Case Study

We conducted a fine-grained manual evaluation of GPT4V's output on CMMaTH, with the results shown in Figure 10 to Figure 22.

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ΔABE≅ΔACD，∠A= 50°， ∠C=20°，则∠ADC=?.
Reference Answer: 110°
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： ∠A + ∠B + ∠C = 180° 。 50° + 20° + ∠ADC = 180° 。接下来，我们需要找到另一个角度，即∠ADC。我们可以通过将内角和减去已知角度来得到它: ∠ADC = 180° - ∠A+ ∠B+ ∠C. ∠ADC = 180° - (50° + 20°)。 ∠ADC = 180° - 70° ,∠ADC = 110° 。所以，三角形ΔABE与ΔACD的另一个角度是∠ADC=110° 。
Judgment Result:
Role: Assistant
The relevant paragraph in the Model Response that provides the answer is
所以，三角形ΔABE与ΔACD的另一个角度是∠ADC=110°
Understand the question in Chinese. The Chinese model response means: Therefore, another angle of triangle Δ ABE and Δ ACD is ∠ADC= 110°.
Next, let's analyze whether the correct answer has been given. The reference answer 110° is a numerical answer, and the output given in the model response is 110° . The consistency of the answers is: <Yes>.

Figure 9: A instruction example to finetune GradeGPT.

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	**Question:** Xiaoming has 2000 yuan, an apple costs 2 yuan, Xiaoming bought 10 apples. How much money does Xiaoming have left? **Model response:** First, calculate the cost of the apples Xiaoming bought, which is 2 times 10 equals 20 yuan. <STEP1> Subtract the total cost from 2000 yuan, 2000-20=1980. <STEP2> So the final answer is 1980 yuan. <STEP3>

Table 10: 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	**Request:** 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? **Reason Result:** 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.

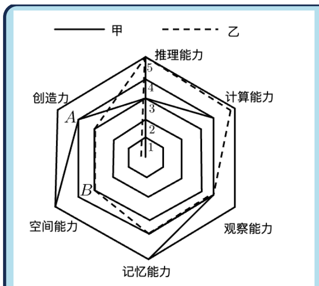
Table 11: Detail prompt for Cross-Lingual Prompt(CLP) Reason technology for our Experiment.

Model	Name
Llama-2-70B	meta-llama/Llama-2-70b-hf
MetaMath-70B	meta-math/MetaMath-70B-V1.0
DeepSeek-Math-7B	deepseek-ai/deepseek-math-7b-instruct
Baichuan-13B	baichuan-inc/Baichuan2-13B-Chat
Qwen-14B	Qwen/Qwen-14B-Chat
LLaVA-v1.5	liuhaotian/llava-v1.5-13b
InterLM-XComposer2-VL	internlm/internlm-7b
Yi-VL-34B	01-ai/Yi-VL-34B
CogAgent-Chat	THUDM/cogagent-chat-hf

Table 12: LLMs used in our experiments and their corresponding names in Huggingface Hub.

Model Name	Generation Parameters	Comments
Llama-2-70B	do_sample=True, top_k=0.5, top_p=0.5, max_tokens=512	model=""Salesforce/codegen2-16B"
GPT-4	temperature=0.2, max_tokens=2048	version="gpt-4-1106-preview"
llava-7B-V1.5	temperature=0.2, max_new_tokens=2048	llava package
DeepSeek-Math-7B	temperature=0.2, max_new_tokens=2048	vllm package
Baichuan-13B	temperature=0.2, max_new_tokens=2048	vllm package
Qwen-14B	temperature=0.2, max_new_tokens=2048	vllm package
InterLM-XComposer2-VL	temperature=0.2, max_new_tokens=2048	Huggingface
Yi-VL-34B	temperature=0.2, max_new_tokens=2048	Huggingface
CogAgent-Chat	temperature=0.2, max_new_tokens=2048	Huggingface
GPT4V	temperature=0.2, max_tokens=2048	version="gpt-4-vision-2023-05-15"
GPT4o	temperature=0.2, max_tokens=2048	version="gpt-4o-2024-02-01"

Table 13: 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.



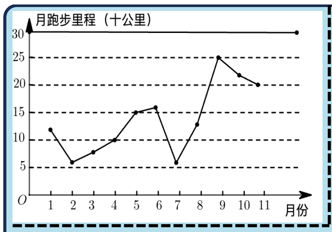
Question:
某综艺节目为比较甲、乙两名选手的各项能力（指标值满分为5分，分值高者为优）绘制了如图所示的六维能力雷达图，图中点A表示甲的创造力指标值为4，点B表示乙的空间能力指标值为3，则下面叙述正确的是（ ）．
A.乙的记忆能力优于甲
B.乙的创造力优于观察能力
C.甲的六大能力整体水平优于乙
D.甲的六大能力中记忆能力最差

Knowledge: 归纳推理

Analysis:
由六维能力雷达图，得：在A中，乙的记忆能力是4，甲的记忆能力是5，故乙的记忆能力劣于甲，故A错误；
在B中，乙的创造能力是3，乙的观察能力是4，故乙的创造力劣于观察能力，故B错误；在C中，甲的六大能力之和为：25，乙的六大能力之和为：24，∴甲的六大能力整体水平优于乙，故C正确；在D中，甲的六大能力中推导能力最差，故D错误．故选C

GPT4V:
首先，我们需要根据雷达图来分析甲、乙两名选手的各项能力值。<STEP1>
从图中可以看出，甲的各项能力值都在点A的外侧，而乙的各项能力值都在点B的内侧。这意味着甲在所有能力上的指标值都大于等于4，而乙在所有能力上的指标值都小于等于3。<STEP2>
A选项，乙的记忆能力优于甲，这是错误的，因为甲的记忆能力指标值至少为4，而乙的所有能力指标值都不超过3。<STEP3>
B选项，乙的创造力优于观察能力，无法从图中直接得出结论，因为图中没有直接比较乙的创造力和观察能力的数据。<STEP4>
C选项，甲的六大能力整体水平优于乙，这是正确的，因为甲的所有能力指标值都在乙的能力指标值之上。<STEP5>
D选项，甲的六大能力中记忆能力最差，无法从图中直接得出结论，因为图中没有显示甲各项能力的具体数值，只能知道它们都不低于4分。<STEP6>
综上所述，正确答案是C选项，甲的六大能力整体水平优于乙。<STEP7>
Comment: Visual Perception Error
Misinterpreted the ability levels represented by A and B in the radar chart in <STEP5>.

Figure 10: Case of GPT4V. The red ones are marked as generated inference hallucinations.



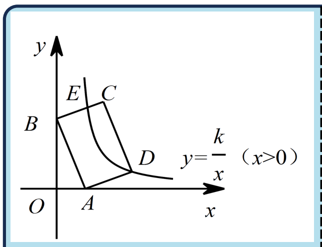
Question:
"悦跑圈"是一款基于社交型的跑步应用，用户通过该平台可查看自己某时间段的运动情况，某人根据2019年1月至2019年11月期间每月跑步的里程（单位：十公里）的数据绘制了下面的折线图，根据该折线图，下列结论正确的是（ ）．
A.月跑步里程逐月增加
B.月跑步里程最大值出现在9月
C.月跑步里程的中位数为8月份对应的里程数
D.1月至5月的月跑步里程相对于6月至11月波动性更小，变化比较平稳

Knowledge: 折线图、总体密度曲线

Analysis:
A:根据题意，依次分析选项：2月跑步里程比1月的小，7月跑步里程比6月的小，10月跑步里程比9月的小，故A错误；
B:月跑步里程9月最大，故B正确；
C:月跑步平均里程的月份从高到底依次为：9月，10月，11月，6月，5月，8月，1月，8月恰好在中间位置，故其中位数为8月份对应的里程数，故C正确；
D:1月至5月的月跑步平均里程相对于6月至11月，波动性更小，变化比较平稳，故D正确．

GPT4V:
求解步骤：首先观察折线图，可以看到每个月的跑步里程数。<STEP1>
A选项说月跑步里程逐月增加，但从图中可以看出，有些月份的里程数比前一个月少，所以A选项错误。<STEP2>
B选项说月跑步里程最大值出现在9月，从图中可以看到8月的里程数是最高的，所以B选项正确。<STEP3>
C选项说月跑步里程的中位数为8月份对应的里程数，由于共有11个月的数据，中位数应该是第6个月的数据，即6月份的数据，而不是8月份的数据，所以C选项错误。<STEP4>
D选项说1月至5月的月跑步里程相对于6月至11月波动性更小，变化比较平稳，从图中可以看出1月至5月的数据变化确实比6月至11月的数据变化要小，所以D选项正确。<STEP5>
综上所述，正确的答案是是D。<STEP6>
Comment：Visual Perception Error
Misinterpreted the peak quantities represented in the line chart in <STEP3>.

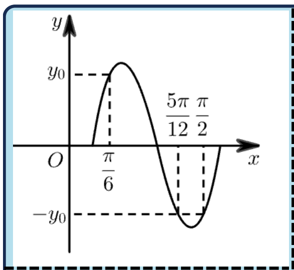
Figure 11: Case of GPT4V. The red ones are marked as generated inference hallucinations.



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$\measuredangle OBA = \measuredangle FAD$，结合$\measuredangle AOB = \measuredangle DFA = 90°$可得出$\Delta AOB \backsim \Delta DFA$，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可以看作是把边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 12: Case of GPT4V. The red ones are marked as generated inference hallucinations.



Question: 若函数$y = \sin(wx + \varphi)$,$\left(w > 0, |\varphi| < \frac{\pi}{2}\right)$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×$(\frac{11\pi}{24} - \frac{5\pi}{24}) = \frac{\pi}{2}$，故$w = \frac{2\pi}{T} = 4$，故$y = \sin(4x + \varphi)$，由上求得的对称轴知$\sin\left(4\times\frac{5\pi}{24} + \varphi\right) = 1$，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2π。 <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φ 。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 13: Case of GPT4V. 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

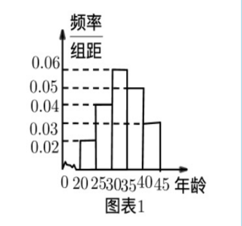
Analysis: 根据中位数、平均数、方程、百分位数等知识对选项进行分析，从而确定正确答案.

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 14: Case of GPT4V. The red ones are marked as generated inference hallucinations.



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

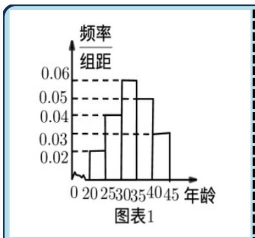
GPT4V:

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。这个频率是指在这个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>
现在我们有两个方程：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>现在我们有两个包含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 15: Case of GPT4V. The red ones are marked as generated inference hallucinations.



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

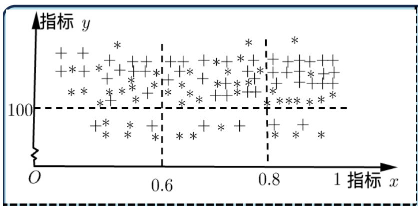
GPT4V:

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>
现在我们有两个方程：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>现在我们有两个包含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 16: Case of GPT4V. The red ones are marked as generated inference hallucinations.



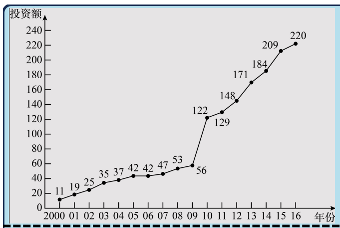
Question:
在贯彻中共中央国务院关于精准扶贫政策的过程中，某单位定点帮扶甲、乙两个村各50户贫困户．为了做到精准帮扶，工作组对这100户村民的年收入情况、劳动能力情况、子女受教育情况、危旧房情况、患病情况等进行调查，并把调查结果转化为各户的贫困指标x和y，制成下图，其中“"*”表示甲村贫困户，“+”表示乙村贫困户。
若 $0 < x < 0.6$，则认定该户为“绝对贫困户”，若 $0.6 \leq x \leq 0.8$，则认定该户为“相对贫困户”，若 $0.8 \leq x \leq 1$，则认定该户为“低收入户”；若 $y \leq 100$，则认定该户为“今年能脱贫户”，否则为“今年不能脱贫户”．
（1）从甲村 50 户中随机选出一户，求该户为“今年不能脱贫的绝对贫困户”的概率．
（2）若从所有“今年不能脱贫的非绝对贫困户”中选 3 户，用 X_i表示所选 3 户中乙村的户数，求 X_i的分布列和数学期望 $E(X_i)$.
（3）试比较这 100 户中，甲、乙两村指标 y 的方差的大小(只需写出结论)．

Knowledge: 离散型随机变量的数学期望,离散型随机变量的分布列,古典概型的概率计算（不涉及计数原理）

GPT4V: （1）**首先，我们需要从图中确定甲村中“今年不能脱贫的绝对贫困户”的数量。根据题目描述，绝对贫困户是指贫困指标x在0<x<0.6的范围内，而“今年不能脱贫户”是指贫困指标y小于100的户。 <STEP1>**
从图中可以看出，甲村中标记为“*”的点，位于x轴的$0 < x < 0.6$区间内且y轴小于100的点的数量即为“今年不能脱贫的绝对贫困户”的数量。数一下这样的点的个数。
<STEP2>设甲村中“今年不能脱贫的绝对贫困户”的数量为A户。 <STEP3>
那么，从甲村50户中随机选出一户，该户为“今年不能脱贫的绝对贫困户”的概率为$\frac{A}{50}$。 <STEP4>
（2）对于第二问，我们需要确定“今年不能脱贫的非绝对贫困户”的数量，这些户是指贫困指标x在$0.6 \leq x \leq 1$的范围内且y轴小于100的户。 <STEP5>
从图中可以看出，标记为“*”和“+”的点，位于x轴的$0.6 \leq x \leq 1$区间内且y轴小于100的点的数量即为“今年不能脱贫的非绝对贫困户”的数量。数一下这样的点的个数，并分别计算出甲村和乙村的数量。 <STEP6>
设甲村中“今年不能脱贫的非绝对贫困户”的数量为B户，乙村中“今年不能脱贫的非绝对贫困户”的数量为C户。 <STEP7>
从所有“今年不能脱贫的非绝对贫困户”中选3户，可以有以下几种情况：选出的3户都来自乙村，选出的2户来自乙村和1户来自甲村，选出的1户来自乙村和2户来自甲村，或者选出的3户都来自甲村。 <STEP8>
根据组合数学原理，我们可以计算出每种情况的概率，并据此得到X的分布列。 <STEP9>
数学期望$E(x_i)$可以通过分布列计算得出，即$E(x_i) = \sum x_i P(x_i)$，其中$P(x_i)$是x_i取某个值的概率。 <STEP10>
（3）对于第三问，我们需要比较甲、乙两村指标$$y$$的方差的大小。方差是衡量一组数据离散程度的统计量，计算公式为$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$
，其中$\bar{y}$是y值的平均数，n是数据的数量。 <STEP11>
从图中可以看出，甲村和乙村的y值分布情况。我们可以观察到，甲村的y值分布是否比乙村的更加集中或者更加分散。 <STEP12>
根据观察结果，我们可以得出结论，哪个村的y值的方差更大。 <STEP13>
由于题目中没有给出具体的数值，我们无法计算具体的概率和方差，但以上步骤展示了如何根据图表信息解决问题。实际操作中，需要根据图中的具体数据进行计算。

Comment: Perception Error
GPT4V cannot fully analyze the content in the scatter plot, and cannot calculate the corresponding probability, mathematical expectation, and variance according to the requirements of the question.
nd c does not equal 83.

Figure 17: Case of GPT4V. The red ones are marked as generated inference hallucinations.



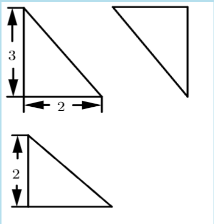
Question:
如图是某地区2000年至2016年环境基础设施投资额y(单位：亿元)的折线图．则下列结论中表述不正确的是( )
A.从2000年至2016年，该地区环境基础设施投资额逐年增加
B.2011年该地区环境基础设施的投资额比2000年至2004年的投资总额还多
C.2012年该地区基础设施的投资额比2004年的投资额翻了两番t
D.为了预测该地区2019年的环境基础设施投资额，根据2010年至2016年的数据(时间变量的值依次为1，2，⋯，7)建立了投资额y与时间变量t的线性回归模型$\hat{y} = 99 + 17.5t$，根据该模型预测该地区2019的环境基础设施投资额为256.5亿元．

Knowledge: 两个变量的线性相关,选择回归模型问题

GPT4V: 首先，我们需要根据折线图来分析每个选项的正确性。
A. 选项A说从2000年至到2016年的每一年，投资额都在增加，所以A选项是正确的。B. 选项B说2011年该地区环境基础设2016年，该地区环境基础设施投资额逐年增加。根据图表，我们可以看到从2000年施的投资额比2000年至2004年的投资总额还多。根据图表，2011年的投资额是122亿元，**而2000年至2004年的投资额分别是11，19，20，35，37亿元，总和为122亿元。**因此，B选项是错误的，因为2011年的投资额与2000年至2004年的总和相等，而不是更多。
C. 选项C说2012年该地区基础设施的投资额比2004年的投资额翻了两番。根据图表，2012年的投资额是129亿元，而2004年的投资额是37亿元。129亿元不是37亿元的两倍，所以C选项是错误的。D. **选项D说根据2010年至2016年的数据建立了投资额y与时间变量t的线性回归模型，预测2019年的环境基础设施投资额为256.5亿元。根据给出的模型，当t=9（2019年是2010年后的第9年）时，预测的投资额为99+17.5*9=256.5亿元。所以D选项是正确的**。综上所述，不正确的结论是B选项。

Comment: Reason Error
For option B, the content in the table is incorrectly parsed. For option D, GPT4V cannot correctly verify whether the parsed expression is correct, causing it to think that option D is correct.

Figure 18: Case of GPT4V. The red ones are marked as generated inference hallucinations.



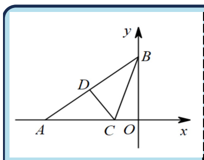
Question:
右图为某几何体的三视图，则该几何体外
接球的表面积为（ ）.
A.17 π
B.68 π
C.13 π
D.23 π

Knowledge: 两个变量的线性相关,选择回归模型问题

GPT4V: 求解步骤：
首先观察三视图，可以看出这是一个直角三角形的棱柱，其底面直角三角形的两条直角边长分别为2和3. <STEP1>
由于直角三角形的斜边是几何体的直径，我们可以使用勾股定理计算斜边的长度，即直径d。
d^2 = 2^2 + 3^2 = 4 + 9 = 13. 因此，$d = \sqrt{13}$. <STEP2>
球的半径r是直径的一半，所以$r = d/2 = \sqrt{13}/2$. <STEP3>
球的表面积公式是4πr^2，将r代入公式得到球的表面积$S = 4\pi(\sqrt{13}/2)$^2 $= 4\pi(13/4) = 13\pi$. <STEP4>
所以最终的答案为C.13 π. <STEP5>

Comment: Reason Error
GPT4V incorrectly infers the solid geometry as a triangular prism based on three views. In fact, the solid geometry is a triangular pyramid.

Figure 19: Case of GPT4V. The red ones are marked as generated inference hallucinations.



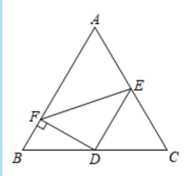
Question:
如图，将一个直角三角形纸片ABO，放置在平面直角坐标中，点$A(-8,0)$，点$B(0,6)$，点$O(0,0)$，C是OA上的点，CD⊥AB，沿着CD折叠该纸片，点A正好落在B处，则线段CD的长为？.

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 20: Case of GPT4V. The red ones are marked as generated inference hallucinations.



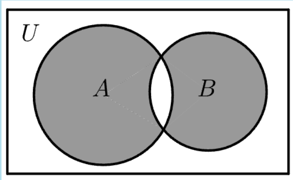
Question:
如图，在边长为4的等边三角形ABC中，D，E分别是边BC，AC的中点，DF ⊥ AB于点F，连结EF，则EF的长为？

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 = √(AD^2 − AF^2) = √(12 − 2^2) = √(12 − 4) = √8 = 2√2. <STEP6>
现在我们有了直角三角形DEF，其中DE=2，DF=2√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√2)^2 = 4 + 8 = 12. <STEP8>
所以，EF = √12 = 2√3. <STEP9>因此，EF的长度是2√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 21: Case of GPT4V. The red ones are marked as generated inference hallucinations.



Question:已知全集 $U=R$，集合 $A=\{x|0\le x\le 2\}$，$B=\{x|x^2-x>0\}$，则图中的阴影部分表示的集合为？
$A.\{x|x\le 1$或$x>2\}$
$B.\{x|x<0$或$1<x<2\}$
$C.\{x|1\le x<2\}$
$D.\{x|1<x\le 2\}$

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$0\le x\le 2$。 <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$1<x\le 2$。 <STEP4>
所以最终的答案是D.$\{x|1<x\le 2\}$。 <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 22: Case of GPT4V. The red ones are marked as generated inference hallucinations.