GeoEval: Benchmark for Evaluating LLMs and Multi-Modal Models on Geometry Problem-Solving

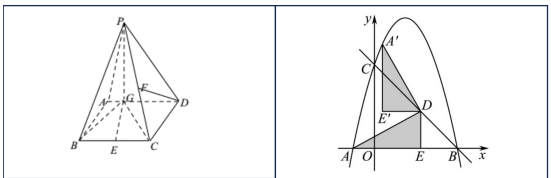
Jiaxin Zhang[§*], **Zhong-Zhi Li**[◇*], **Ming-Liang Zhang**[◇], **Fei Yin**[⊛◇], **Cheng-Lin Liu**[⊛◇†], **Yashar Moshfeghi**[§†]
School of Artificial Intelligence, University of Chinese Academy of Sciences[◇]
MAIS, Institute of Automation of Chinese Academy of Sciences[◇]
Department of Computer & Information Sciences, University of Strathclyde[§]
{jiaxin.zhang, moshfeghi.yashar}@strath.ac.uk[§],
{lizhongzhi2022, zhangmingliang2018}@ia.ac.cn[◇],
{fyin, liucl}@nlpr.ia.ac.cn[⊛]

Abstract

Recent advancements in large language models (LLMs) and multi-modal models (MMs) have demonstrated their remarkable capabilities in problem-solving. Yet, their proficiency in tackling geometry math problems, which necessitates an integrated understanding of both textual and visual information, has not been thoroughly evaluated. To address this gap, we introduce the GeoEval benchmark, a comprehensive collection that includes a main subset of 2,000 problems, a 750 problems subset focusing on backward reasoning, an augmented subset of 2,000 problems, and a hard subset of 300 problems. This benchmark facilitates a deeper investigation into the performance of LLMs and MMs in solving geometry math problems. Our evaluation of ten LLMs and MMs across these varied subsets reveals that the WizardMath model excels, achieving a 55.67% accuracy rate on the main subset but only a 6.00% accuracy on the hard subset. This highlights the critical need for testing models against datasets on which they have not been pre-trained. Additionally, our findings indicate that GPT-series models perform more effectively on problems they have rephrased, suggesting a promising method for enhancing model capabilities.[1]

1 Introduction

Geometry math problems are a key component in assessing the mathematical reasoning skills of K12 students, serving as a critical benchmark for evaluating educational outcomes (Zhang et al., 2023c). The complexity of solving these problems stems from the requirement to interpret both textual and visual information, in addition to applying mathematical reasoning skills. This complexity has made geometry problem-solving a key area of interest for researchers aiming to evaluate the capabilities of AI models in this domain (Chou and Gao, 1996; Ye et al., 2008; Zhang et al., 2023a; Trinh et al., 2024; Zhang et al., 2024; Zhang and Moshfeghi, 2024).



ID: prob_60213
Question: In the pentahedron P-ABCD, the base ABCD is a parallelogram, ⋯ ⋯ E is the midpoint of BC. Find the distance from point D to the plane PBG.
Choice:
(A) $\frac{5}{2}$. (B) $\frac{3}{2}$.
(C) $\frac{1}{2}$. (D) $\frac{1}{5}$.
Answer: (B)

ID: prob_60156
Question: As shown in Figure, given the parabola $y = -x^2 + 3x + 4$ intersecting the x axis at A and B (point A is to the left of point B) ⋯ If the area of ΔADE is S, find the value of m that maximizes S
Choice:
(A) 1.5. (B) 2.5.
(C) 0.5. (D) 3.
Answer: (A)

Figure 1: Examples of the GeoEval benchmark.

In recent years, several datasets, such as Geometry3K (Lu et al., 2021), PGPS9K (Zhang et al., 2023b), and GeomVerse (Kazemi et al., 2023), have been developed to test the proficiency of AI models in solving geometry math problems. Yet, these datasets often lack a standardized format and sufficient diversity, complicating the assessment of models' genuine proficiency in geometry problem-solving. Furthermore, these datasets typically focus on one type of geometry problem, such as flat geometry, overlooking other crucial areas like solid geometry. This oversight limits the ability to conduct a thorough evaluation across the full spectrum of geometry problems.

Simultaneously, advancements in large language models (LLMs) and multi-modal models (MMs) have demonstrated significant potential in handling complex reasoning tasks (Chen et al., 2022b; Wei et al., 2022; Zhang et al., 2023d; Chen et al., 2023). This potential has raised considerable interest in testing these advanced models across a variety of tasks, such as math word problem solving (Lu et al., 2023) and physical problem solving (Sawada et al.,

[*] Equal Contribution
[†] Corresponding Author
[1] The code and data available are at https://github.com/GeoEval/GeoEval.

Findings of the Association for Computational Linguistics ACL 2024, pages 1258–1276

2023). Despite this interest, specific research on evaluating these models' effectiveness in geometry problem-solving remains scarce. Therefore, it is critical to develop a new, comprehensive benchmark that can effectively assess LLMs and MMs in geometry problem-solving, especially considering the potential exposure of existing public datasets during model training (Sainz et al., 2023). Comparing the performance of current LLMs and MMs on such a benchmark is essential, as it could yield valuable insights that further the development of models capable of tackling complex reasoning tasks.

To prompt research towards assessing LLMs' and MMs' proficiency in geometry math problem-solving, we introduce the GeoEval benchmark, a comprehensive collection specifically designed for this task. GeoEval features its *Comprehensive Variety*, sourced from seven public datasets and formatted uniformly to encompass a wide range of geometric shapes. It includes *Varied Problems*, covering flat, solid, and analytic geometry to challenge models comprehensively. GeoEval supports *Dual Inputs*, accommodating both geometric diagrams and textual problem statements, making it suitable for evaluating both LLMs and MMs. To counter the potential overfitting to previously seen datasets, GeoEval introduces *Diverse Challenges* through backward reasoning, augmented, and hard subsets, each designed to test different aspects of models' geometry problem-solving abilities. Additionally, GeoEval is annotated with *Complexity Ratings*, allowing for a fine-grained analysis of model performance across various difficulty levels, thus providing a robust framework for advancing AI capabilities in understanding and solving geometry math problems. Examples of geometry problems from our GeoEval can be found in Figure 1.

In this paper, we conduct extensive experiments using the GeoEval benchmark to evaluate the proficiency of ten LLMs and MMs in solving geometry problems. This includes three LLMs: CodeGen2-16B (Nijkamp et al., 2023), GPT-3.5 (OpenAI, 2022), and GPT-4 (OpenAI, 2023); two LLMs specialized in mathematics: WizardMath-70B and WizardMath-7B-V1.1 (Luo et al., 2023); and five MMs: llava-7B-V1.5 (Liu et al., 2023), Qwen-VL (Bai et al., 2023b), mPLUG-Owl2 (Ye et al., 2023), InstructBLIP (Dai et al., 2023), and GPT-4V (OpenAI, 2023). The findings reveal that GeoEval forms a challenging benchmark, with both LLMs and MMs struggling to resolve its complexities effectively.

Notably, our results indicate that: ① Models pre-trained on mathematical corpora, such as the WizardMath models, deliver superior performance across various GeoEval subsets (Section 4.3.1), establishing new benchmarks in the field. ② One advantage of these models is that they implicitly encompass the required mathematical knowledge demanded to solve geometry math problems (Section 4.6). ③ However, we also find that though pre-training on a mathematical corpus is crucial for solving geometry math problems, it may not be enough (Section 4.3.4). ④ Additionally, we observe that GPT series models exhibit enhanced problem-solving efficiency when tackling geometry questions that they have previously rephrased (Section 4.3.4). ⑤ Further analyses underscore the value of incorporating descriptions of geometric diagrams, which significantly aids LLMs in understanding and solving geometry problems (Section 4.5). ⑥ Finally, our experiments show that the performance of both LLMs and MMs declines as the problem length and complexity of the problem increases (Section 4.7). Through the GeoEval benchmark, we believe this research provides the first comprehensive quantitative assessment of the latest LLMs and MMs in the domain of geometry problem-solving.

2 Related Work

Numerous benchmarks have been developed to assess the capabilities of LLMs in the the geometry problem-solving task. However, these benchmarks face limitations, such as restricted access, like GEOS (Seo et al., 2015) and GeoShader (Alvin et al., 2017) datasets, or insufficient scale, as seen with GEOS++ (Sachan et al., 2017). Although recent efforts have introduced new benchmarks like Geometry3K (Lu et al., 2021), UniGeo (Chen et al., 2022a), and PGPS9K (Zhang et al., 2023b), they still fall short in offering a uniform format and embracing a wide range of problem types. In response, we introduce the comprehensive and challenging GeoEval benchmark, aiming to advance the evaluation of geometry problem-solving abilities.

Recently, LLMs (Peng et al., 2023; Touvron et al., 2023; OpenAI, 2022) and MMs (Liu et al., 2023; Ye et al., 2023; OpenAI, 2023) have achieved impressive results on complex tasks, attracting research into their performance across specialized tasks. Previous work like MathVista (Lu et al.,

2023) have concentrated on scientific domains, likewise SEED (Li et al., 2023) explores models' understanding of temporal and spatial relationships. Despite these advancements, there remains a gap in the examination of models' ability to solve geometry math problems. Through the GeoEval benchmark, we aim to fill this gap by offering a detailed assessment of both LLMs' and MMs' abilities to tackle a variety of geometry math challenges.

3 GeoEval Dataset

The GeoEval benchmark is structured into four subsets: GeoEval-2000, comprising 2,000 problems; GeoEval-backward, with 750 problems; GeoEval-aug, containing 2,000 problems; and GeoEval-hard, including 300 problems. The subsequent sections will detail the collection process for each subset, followed by an explanation of the unique features of the GeoEval benchmark.[2]

3.1 Data Collection

3.1.1 Collection from Diverse Data Sources

We have compiled a comprehensive collection of public geometry math problem datasets, with a total of 24,912 geometry math problems from sources such as Geometry3K (Lu et al., 2021), PGPS9K (Zhang et al., 2023b), UniGeo (Chen et al., 2022a), GeoQA+ (Cao and Xiao, 2022), GeometryQA (Tsai et al., 2021), as well as geometry problems from the MATH (Hendrycks et al., 2021) and MathQA (Amini et al., 2019) datasets. The first four datasets feature geometry questions that include both problem texts and geometric diagrams, whereas the latter three datasets comprise questions that only contain problem texts. Detailed information about all source datasets is available in Appendix B.

Building on the data gathered, we then selected 2,000 geometry math problems to create our GeoEval-2000 subset. This selection process was guided by the aim to inclusively cover a wide range of basic geometric shapes, ensuring a broad representation of geometry concepts. The distribution of geometric shapes within this subset is further detailed in Appendix C.

3.1.2 Backward Data Generation

In contrast to forward problems, backward problems use the answer from forward problems as a starting point, posing a query to determine a specific number that was part of the forward problems but is concealed in the backward problems (Jiang et al., 2023). These types of questions are particularly effective in assessing models' capability for multi-step reasoning. Following the methodology of previous research (Yu et al., 2023), we selected 750 problems from the GeoEval-2000 subset and created corresponding backward questions. This process involved masking a number, the solution of the forward problems, as "X". The prompt "*The correct answer is* ans_{gold}. *Now please answer what is the value of X?*", where ans_{gold} represents the correct answer to the forward problems, is then added. The example of backward problems can be found in Appendix D.

3.1.3 Augmented Data Generation

To evaluate the resilience of current models and mitigate the risk of data leakage that may occur during the pre-training phase, we implement a context learning strategy for rephrasing problems from the GeoEval-2000 subset. Each problem is rephrased into five variant candidates by GPT-3.5 (OpenAI, 2022), ensuring they retain the original problem's semantic essence while varying in lexical structure. Out of these five alternatives, one is selected randomly to substitute the original problems, forming the GeoEval-aug subset.

3.1.4 Hard Data Collection

While the GeoEval-2000 subset comprises geometry problems from a variety of source datasets, it exhibits a lack of diversity in problem categories, notably in solid geometry and analytic geometry. To enhance the diversity of problem categories, we introduce the GeoEval-hard subset, which includes 300 geometry problems specifically focusing on solid geometry and analytic geometry, providing a broader assessment scope. More details regarding the comparison between the GeoEval-hard subset with other datasets are in Appendix E.

The GeoEval-hard subset sources from the copyrighted collection containing 10,000 geometry math problems created based on templates summarized from the online resources. An initial selection is made using a rule-based engine equipped with a keyword list, targeting solid and analytic geometry problems. This step yields around 3,100 potential problems, identified as the GeoEval-hard-raw subset. Next, a manual review further narrows these down to 300 problems related to solid and analytic

[2]Statistics for the GeoEval benchmark are in Appendix A.

Dataset	*Comprehensive Variety*	*Varied Problems*	*Dual Inputs*	*Diverse Challenges*	*Complexity Ratings*
MathQA (Amini et al., 2019)	n/a	flat	text	✗	✗
GeometryQA (Tsai et al., 2021)	n/a	flat	text	✗	✗
Geometry3K (Lu et al., 2021)	n/a	flat	text + diagram	✗	✗
GeoQA+ (Cao and Xiao, 2022)	n/a	flat	text + diagram	✗	✗
MATH (Hendrycks et al., 2021)	n/a	flat	text	✗	✗
UniGeo (Chen et al., 2022a)	n/a	flat	text + diagram	✗	✗
PGPS9K (Zhang et al., 2023b)	n/a	flat	text + diagram	✗	✗
GeomVerse (Kazemi et al., 2023)	n/a	flat	text + diagram	✗	✓
MathVista (Lu et al., 2023)	4	flat	text + diagram	✗‡	✗
GeoEval	7 + 3 (new)	flat, solid, analytic	text + diagram	✓	✓

Table 1: Comparison between GeoEval benchmark and other datasets. Under *Comprehensive Variety*, MathVista and GeoEval stand out as collective datasets, while the rest, are denoted as 'n/a'. GeoEval includes problems from seven public datasets and three newly created ones. *Varied Problems* categorizes problems into "flat geometry", "solid geometry", and "analytic geometry", For *Dual Inputs*, "text" signifies problems presented only in text format, whereas "text + diagram" encompasses problems with both texts and diagrams. In *Diverse Challenges*, the symbol ‡ indicates that MathVista introduces three new datasets, which, however, are unrelated to the geometry problem-solving task.

geometry. The cleaning and manual inspection process is documented in Appendix F.

3.2 Features of GeoEval

The GeoEval benchmark is specifically designed to assess the ability of models to resolve geometric math problems. This benchmark features five characteristics: *Comprehensive Variety*, *Varied Problems*, *Dual Inputs*, *Diverse Challenges*, and *Complexity Ratings*, with each attribute exemplified in the Appendix G. For an insightful contrast, Table 1 offers a comparative analysis of GeoEval against earlier datasets.

Comprehensive Variety GeoEval consists of a diverse collection of geometry problems sourced from the seven most recent datasets. Therefore, the problems in GeoEval cover a wide range of geometric shapes, offering a comprehensive view of varied geometry math challenges.

Varied Problems The GeoEval benchmark encompasses three distinct categories of geometry math problems, namely flat geometry, solid geometry, and analytic geometry.

Dual Inputs GeoEval features problems in two formats: those accompanied by diagrams and those consisting solely of text. This versatility makes it suitable for evaluating models that process diagrams or text-based inputs.

Diverse Challenges In addition to gathering public datasets, GeoEval also generates its out-of-distribution dataset aimed at addressing data leakage problems. This includes a backward reasoning subset, an augmented subset, and a hard subset, all created by us.

Complexity Ratings GeoEval is equipped with annotations indicating the complexity level for each problem, serving as a guideline to evaluate models' proficiency in solving these tasks.[3]

4 Experiments

4.1 Experimental Setup

In this study, we deliberately select state-of-the-art LLMs and MMs that are widely recognized for their advanced capabilities, including:

- **LLMs Specialized in Programming Code**: We include CodeGen2-16B model (Nijkamp et al., 2023), which is renowned for its proficiency in understanding and generating programming code, offering insights into its adaptability to solve geometry math problems.
- **LLMs with a Focus on Mathematics**: This includes WizardMath-7B-V1.1 and WizardMath-70B (Luo et al., 2023), explicitly pre-trained on mathematical corpora. Their inclusion allows for an assessment of models that have been fine-tuned to tackle complex mathematical problems.
- **LLMs Designed for a Broad Range of Topics**: Models such as GPT-3.5 (OpenAI, 2022)

[3]Algorithm for classifying complexity is in Appendix H.

and GPT-4 (OpenAI, 2023) exemplify the advanced commercial LLMs engineered to encompass a broad range of topics.

- **Multi-Modal Models (MMs) with Diverse Decoders**: Given the ubiquity of ViT architecture (Dosovitskiy et al., 2021) as the vision encoder in MMs, we select models that integrate ViT with various LLMs as decoders. This includes llava-7B-V1.5 (Liu et al., 2023) with Vicuna (Peng et al., 2023), Qwen-VL (Bai et al., 2023b) using Qwen (Bai et al., 2023a), mPLUG-Owl2 (Ye et al., 2023) with LLaMA (Touvron et al., 2023), InstructBLIP (Dai et al., 2023) with Vicuna (Peng et al., 2023), and GPT-4V (OpenAI, 2023).

These models are evaluated through a zero-shot approach, utilizing straightforward instruction prompts to directly assess their geometry problem-solving capabilities without further fine-tuning specifically for our benchmark.[4]

4.2 Evaluation Metric

Building upon the approach by MathVista (Lu et al., 2023), we first input the generated sequence from the model into GPT-4 to extract the target value or option letter. To enhance the precision of our answer extraction, we formulate intricate rules for post-processing the outcomes in cases where GPT-4 falls short. Specifically, our extraction pipeline involves two steps: firstly, using a prompt to extract the answer. Secondly, employing regular expressions to extract any remaining answers that couldn't be obtained from the first step. Please refer to Table 4 and Table 5 in Appendix for the extraction instruction and the constructed samples. This approach has enabled us to attain an extraction accuracy surpassing 97%[5], similar to the success rate reported in MathVista (Lu et al., 2023).

The extracted results are compared against the golden answers to determine the final performance metric. Given the model's intention to produce responses in varying formats, either as the precise answer (for instance, "3.15") or as the corresponding option letter (such as "A"), we regard a prediction as accurate if it either matches the golden answer or the golden option letter.

[4]Details on the prompt design and the hyper-parameters used for these models are available in Appendix I.

[5]We assess the accuracy of the extraction by manually reviewing 200 uniformly sampled examples.

4.3 Experimental Results

In this section, we present the accuracy achieved by models on our GeoEval benchmark. Table 2 highlights that models pre-trained on a math-specific corpus tend to outperform others. Furthermore, except for llava-7B-V1.5 and Qwen-VL, multi-modal models (MMs) generally exceed the performance of large language models (LLMs). Notably, InstructBLIP exhibits exceptionally high accuracy scores across all subsets, yet its results raise some concerns, and we have chosen to exclude the InstructBLIP model. The rationale behind this decision is detailed in Appendix J.

4.3.1 Comparison among LLMs

When reviewing the performances of LLMs as detailed in Table 2, it becomes evident that models pre-trained on mathematical corpora demonstrate superior efficacy in solving geometry math problems compared to those trained on general corpora. Specifically, evaluating on all problems of the GeoEval-2000 subset (marked as "A" in the table), WizardMath-70B leads with an accuracy of 55.67%, while WizardMath-7B-V1.1 closely follows with a 54.78% accuracy, outperforming other LLMs. Conversely, GPT-4, GPT-3.5, and CodeGen2-16B report notably lower accuracies, all under 30.00%. Focusing on questions solely based on problem text within the GeoEval-2000 subset (indicated as "T" in the table), GPT-4 emerges as the frontrunner, securing the highest accuracy of 43.86%, with WizardMath models also surpassing the 32.00% accuracy. These findings underscore the enhanced proficiency of models pre-trained on math-specific corpora in tackling geometry math problems, particularly when problems are well-described textually, as evidenced by GPT-4's leading performance.

In the GeoEval-backward subset, WizardMath-7B-V1.1 excels with the highest accuracy of 32.66%, closely followed by WizardMath-70B at 28.66%. This significant drop in performance across all LLMs, compared to the GeoEval-2000 results, highlights a collective weakness in backward reasoning capabilities. For the GeoEval-aug subset, WizardMath-7B-V1.1 again tops the leaderboard with an accuracy of 47.75%, with GPT-4 not far behind at 45.75% accuracy. Lastly, within the GeoEval-hard subset, all models, excluding GPT-3.5, exhibit relatively low accuracies, indicating a broad difficulty in addressing the most

Model	GeoEval-2000 A (%)	GeoEval-2000 T (%)	GeoEval-backward A (%)	GeoEval-aug A (%)	GeoEval-hard A (%)
CodeGen2-16B ♢	28.76	22.06	5.10	8.50	5.66
GPT-3.5 ♢	24.71	21.27	22.66	41.25	22.33
GPT-4 ♢	27.95	43.86	26.00	45.75	10.10
WizardMath-70B ♢	**55.67**	34.20	28.66	37.75	6.00
WizardMath-7B-V1.1 ♢	54.78	32.76	32.66	**47.75**	6.00
llava-7B-V1.5	12.80	21.01	11.33	20.25	20.30
Qwen-VL	25.60	25.97	5.66	22.25	21.66
mPLUG-Owl2	37.76	n/a	**35.33**	38.00	**22.66**
InstructBLIP †	52.18	n/a	15.66	35.00	70.30
GPT-4V	37.22	**43.86** ‡	26.00	45.75	10.10

Table 2: Accuracy scores of models on our GeoEval benchmark. The "♢" refers to all LLMs. The "A" signifies the overall accuracy across all problems, while "T" denotes the accuracy for problems containing only texts without diagrams. The "n/a" indicates that scores are unavailable due to models cannot process text-only inputs. The "†" shows our doubt on the high accuracy rates reported by the IntructBLIP model, our point is elaborated in Section 4.3. The "‡" notes that the accuracy figures for GPT-4V are derived from GPT-4, as GPT-4V does not support image-free inputs. Detailed reporting on model performance, segmented by dataset origins, is available in Appendix K.

challenging solid geometry and analytic geometry problems. To investigate the reason for GPT-3.5 achieves better performance than GPT-4 on the GeoEval-hard subset, we find that GPT-4 tends to generate verbose solutions, often accompanied by code, which causes it to either terminate before solving the problem or enter a self-cycling loop of generating redundant information, failing to provide a final answer. In contrast, GPT-3.5 adopts a more concise approach, consistently producing option letters (e.g., "A") following the reasoning steps as solutions. We believe this concise solution generation strategy contributes to GPT-3.5's relatively better performance on the GeoEval-hard subset.

4.3.2 Comparison among Multi-Modal Models

Table 2 shows that among the MMs, GPT-4V and mPLUG-Owl2 consistently outperform their counterparts across all subsets. Specifically, within the GeoEval-2000 subset, mPLUG-Owl2 leads with an accuracy of 37.76%, closely followed by GPT-4V at 37.22%, with the remaining MMs falling behind at lower accuracies. Specifically, Qwen-VL and llava-7B-V1.5 achieve accuracies of 25.60% and 12.80%, respectively. When examining problems that only involve texts, GPT-4V achieves a 43.86% accuracy, significantly surpassing llava-7B-V1.5 (21.01%) and Qwen-VL (25.97%).

In the GeoEval-backward subset, mPLUG-Owl2 tops with the accuracy of 35.33%, with GPT-4V following at 26.00% accuracy. This performance shows a notable lack of backward reasoning skills, as illustrated by the diminished results of llava-7B-V1.5 and Qwen-VL in this category. Moving to the GeoEval-aug subset, GPT-4V leads with an impressive 45.75% accuracy, with mPLUG-Owl2 in second place with 38.00% accuracy. Both Qwen-VL and llava-7B-V1.5 show comparable performances in this subset. Lastly, within the GeoEval-hard subset, mPLUG-Owl2 demonstrates the highest efficacy with a 22.66% accuracy, closely followed by Qwen-VL and llava-7B-V1.5. Surprisingly, GPT-4V records a lower accuracy of just 10.10%, highlighting the challenging nature of the GeoEval-hard subset and the varied capabilities of MMs in addressing the most difficult problems.

4.3.3 Comparison between LLMs and Multi-Modal Models

In the GeoEval-2000 subset, specifically for problems that only include texts, GPT-4's performance exceeds the top MMs, Qwen-VL, by 17.89%. This is attributed to the MMs' inability to access geometric diagrams, which likely hinders their comprehension of the problems. Moreover, when evaluating all problems of the GeoEval-2000 subset, WizardMath-70B surpasses the best MMs, Qwen-VL, by 17.91% in accuracy. However, MMs like

GPT-4V and mPLUG-Owl2 achieve significantly higher accuracy than LLMs not pre-trained on mathematical content. This underscores the value of mathematical pre-training for excelling in geometry problem-solving. Notably, GPT-4V's accuracy on all GeoEval-2000 problems is 9.27% higher than GPT-4's, suggesting GPT-4V's superior capability in solving geometry problems with diagrams.

This pattern persists in the GeoEval-aug subset, where WizardMath-7B-V1.1, a model trained on a mathematical corpus, achieves the highest accuracy at 47.75%. Conversely, mPLUG-Owl2 leads in the GeoEval-backward and GeoEval-hard subsets, with accuracies of 35.33% and 22.66%, respectively. Given that GeoEval-aug rephrases questions from GeoEval-2000, it implies both subsets might have been exposed to the models during their pre-training phase. In contrast, GeoEval-backward and GeoEval-hard subsets are less likely to have been previously exposed. This suggests that WizardMath-7B-V1.1 excels with familiar geometry math problems, while mPLUG-Owl2 demonstrates a robust capability in tackling unseen geometry problems. This is further evidenced by the low performance of WizardMath models on the GeoEval-hard subset, where both models only achieve an accuracy of 6.00%.

4.3.4 Analysis on the Best Model

Table 2 shows that GPT-4, the leading LLMs, records the highest accuracy on the GeoEval-aug subsets, though it only secures a 27.95% accuracy on the GeoEval-2000 subset. A similar pattern of improvement is noted for the GPT-3.5 model, which sees its accuracy jump from 24.71% on the GeoEval-2000 subset to 41.25% on the GeoEval-aug subset. This improvement aligns with the involvement of GPT-3.5 in generating the GeoEval-aug subset, suggesting that the capabilities of GPT-3.5 and GPT-4 in addressing geometry math problems significantly benefit from their use in rephrasing geometry question texts.

While WizardMath-70B and WizardMath-7B-V1.1, both pre-trained on a mathematical corpus, demonstrate superior performance on the GeoEval-2000 subset, they show a marked decline in accuracy across the other subsets, with the most significant decreases observed on the GeoEval-hard subset. This indicates that although pre-training on a mathematical corpus is crucial for solving geometry math problems, it may not be enough.

In contrast to the significant variances in accuracy observed among LLMs across different subsets, the top-performing multi-modal model, mPLUG-Owl2, maintains relatively stable accuracies with scores of 37.76% on the GeoEval-2000, 35.33% on the GeoEval-backward, and 38.00% on the GeoEval-aug subsets. Additionally, the performance of GPT-4V on the GeoEval-aug subset surpasses its accuracy on the GeoEval-2000 subset, mirroring the trends observed with GPT-4 and GPT-3.5, further illustrating the enhanced effectiveness of GPT-series models when engaged in rephrasing the content of geometry questions.

4.4 Results Across Different Subjects

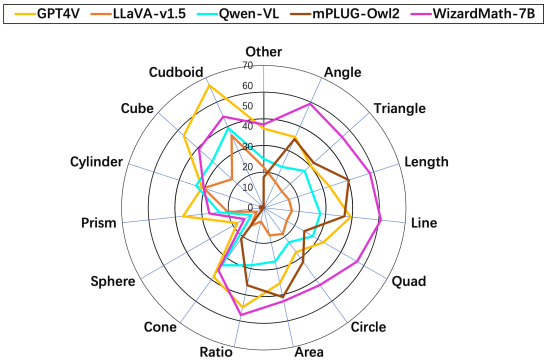


Figure 2: Detailed accuracy scores for models across various academic subjects.

Figure 2 displays the performance of models across various subjects, revealing distinct strengths. The WizardMath-7B model significantly outperforms others in flat geometry problems, such as length and lines. Conversely, in solid geometry problems like cuboids and spheres, GPT-4V surpasses WizardMath-7B, indicating its superior capability in addressing solid geometry questions.

4.5 Benefit from the Geometric Diagram Descriptions

Models	✗	✓
GPT-4V	40.28	45.61 (+*5.33*)
WizardMath-7B	38.10	56.83 (+*18.73*)

Table 3: Comparison of models with (✓) and without (✗) geometric diagram descriptions.

To assess the impact of including geometric diagram descriptions on models' ability to comprehend geometric diagrams and solve related problems, we selected a sample of 300 questions with geometric diagram descriptions from the GeoEval-

2000 subset. We then evaluated the performance of two models, GPT-4V and WizardMath-7B-V1.1, on these questions, both with and without the use of geometric diagram descriptions, which describe the geometric shapes and relations encapsulated in the diagram. The results in Table 3 indicate that GPT-4V's accuracy decreases by 5.33% without the diagram descriptions. More significantly, WizardMath-7B's accuracy falls by 18.73% in the absence of these descriptions. This evidence suggests that supplemental geometric diagram descriptions significantly enhance models' efficiency in solving geometry math problems, particularly benefiting LLMs.

4.6 External Constants Required for Solving the Problems

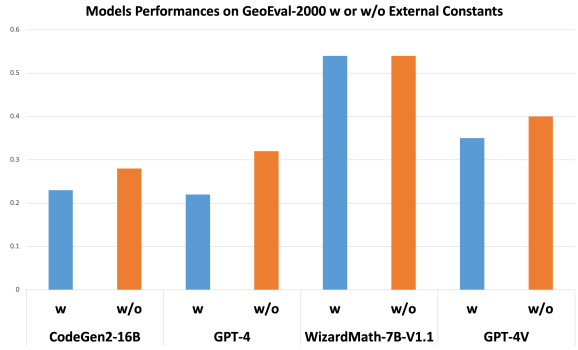


Figure 3: Comparison of models requiring external constants ("w" in blue color) and those do not ("w/o" in orange color).

In the GeoEval benchmark, certain questions require external constants, such as the value of π, which is not typically included in the problem text. This necessitates models to have pre-existing knowledge to accurately solve these problems. Figure 3 assesses the performance of four models on problems differentiated by the need for external constants, identified through a heuristic approach that classifies problems according to whether their solutions require constants.

Figure 3 shows that the WizardMath-7B-V1.1 model maintains consistent accuracy on the GeoEval-2000 subset, regardless of the requirement for external constants, unlike other models, which perform better on problems without such requirements. This consistency in WizardMath-7B-V1.1's performance is likely due to its pre-training on a math-specific corpus, providing it with the necessary knowledge to resolve geometry math problems effectively. In contrast, models trained on general corpora may not possess this specialized mathematical knowledge, hindering them from using external constants to solve the problems correctly.

4.7 Performances According to Different Problem Lengths and Varied Complexities

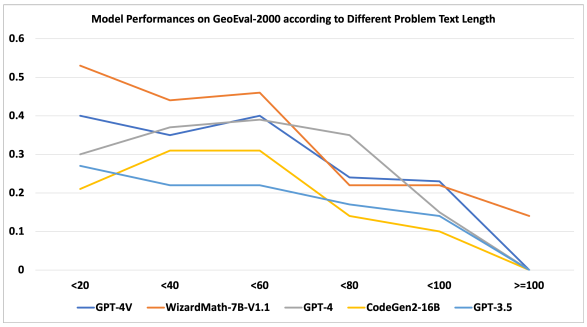


Figure 4: Models performances on GeoEval-2000 subset according to different question lengths.

Figure 4 shows how models perform with inputs of different lengths. Performance slightly varies for problems ranging from 80 to 100 characters, but there's a clear trend of decreasing accuracy as problem length increases. This is expected, as longer questions typically involve more complex geometry math problems, challenging the models more as the length grows. The figure also points out that the WizardMath-7B-V1.1 model is notably more adept at handling longer questions, with GPT-4V and GPT-4 showing relatively stable accuracy for increased question lengths. On the other hand, GPT-3.5 and CodeGen2-16B perform less effectively on lengthy questions.

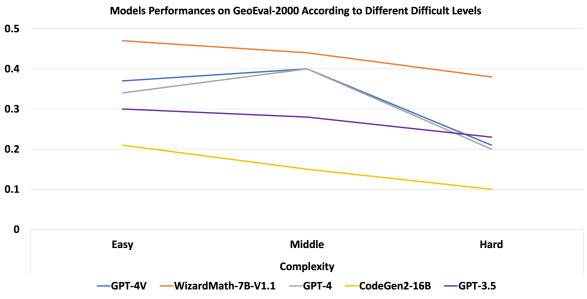


Figure 5: Model performances on GeoEval-2000 subset according to different complexity levels.

Upon the analysis in Figure 5, similar to the observations made in Figure 4 regarding input lengths, we delve into the models' performances as they relate to the complexity of geometry math problems. Figure 5 presents the performance of models across varying levels of problem complexity. It is evident

that as the complexity of geometry problems escalates, the accuracy of the models correspondingly diminishes.

5 Conclusion

In this study, we present GeoEval, a benchmark developed to assess the geometry problem-solving capabilities of large language models (LLMs) and multi-modal models (MMs). GeoEval comprises four distinct subsets, each designed to facilitate a thorough evaluation. Through our assessment of ten cutting-edge LLMs and MMs using the GeoEval benchmark, we underscore the critical role of mathematical corpus pre-training for effective geometry problem resolution. This is exemplified by the WizardMath model's leading performance on the GeoEval-2000 subset, achieving an accuracy of 55.67%. However, the WizardMath model's challenges with the GeoEval-hard subset suggest a need for enhanced reasoning skills. Additionally, our analysis reveals that GPT-series models exhibit improved performance on geometry problems they have rephrased, pointing to the potential benefits of self-rephrasing in problem-solving.

6 Limitations

This study, while providing significant insights into the capabilities of large language models (LLMs) and multi-modal models (MMs) in solving geometry problems, has several limitations.

One primary constraint is that our evaluation predominantly focuses on quantitative metrics of accuracy, potentially overlooking qualitative aspects of model reasoning and explanation that are crucial for educational applications. The performance of models on the hard subset also highlights a gap in advanced reasoning abilities, suggesting that current LLMs and MMs, including those pre-trained on mathematical corpora, may still struggle with highly complex or novel problem types. In addition, we conduct experiments focusing on testing the models' ability to recall and effectively utilize knowledge about mathematical constants. However, a comprehensive evaluation of external knowledge utilization is required, such as theorems and principles, which are beyond just mathematical constants. We plan to explore models' abilities to leverage diverse forms of external knowledge in future work.

Moreover, this work reveals the effectiveness of rephrased problems by GPT-series models and suggests a specific interaction effect that may not generalize across all types of geometry problems or other LLMs and MMs, indicating a need for broader research to fully understand the implications of rephrasing on model performance.

Acknowledgments

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

Chris Alvin, Sumit Gulwani, Rupak Majumdar, and Supratik Mukhopadhyay. 2017. Synthesis of solutions for shaded area geometry problems. In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017*, pages 14–19. AAAI Press.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2357–2367. Association for Computational Linguistics.

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. 2023a. Qwen technical report. *CoRR*, abs/2309.16609.

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

Jie Cao and Jing Xiao. 2022. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Ko-*

rea, October 12-17, 2022, pages 1511–1520. International Committee on Computational Linguistics.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022a. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3313–3323. Association for Computational Linguistics.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2022b. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *CoRR*, abs/2211.12588.

Shang-Ching Chou and Xiao-Shan Gao. 1996. Automated generation of readable proofs with geometric invariants i. multiple and shortest proof generation. *J. Autom. Reason.*, 17(3):325–347.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. 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. 2021. 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*.

Weisen Jiang, Han Shi, Longhui Yu, Zhengying Liu, Yu Zhang, Zhenguo Li, and James T. Kwok. 2023. Forward-backward reasoning in large language models for mathematical verification.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. 2023. Geomverse: A systematic evaluation of large models for geometric reasoning. *CoRR*, abs/2312.12241.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *CoRR*, abs/2307.16125.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744.

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, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6774–6786. Association for Computational Linguistics.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *CoRR*, abs/2308.09583.

Erik Nijkamp, Hiroaki Hayashi, Caiming Xiong, Silvio Savarese, and Yingbo Zhou. 2023. Codegen2: Lessons for training llms on programming and natural languages. *CoRR*, abs/2305.02309.

OpenAI. 2022. gpt-3.5-turbo-0125.

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

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. *CoRR*, abs/2304.03277.

Mrinmaya Sachan, Avinava Dubey, and Eric P. Xing. 2017. From textbooks to knowledge: A case study in harvesting axiomatic knowledge from textbooks to solve geometry problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 773–784. Association for Computational Linguistics.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10776–10787. Association for Computational Linguistics.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J. Nay, Kshitij Gupta, and Aran Komatsuzaki. 2023. ARB: advanced reasoning benchmark for large language models. *CoRR*, abs/2307.13692.

Min Joon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1466–1476. The Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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

Shih-hung Tsai, Chao-Chun Liang, Hsin-Min Wang, and Keh-Yih Su. 2021. Sequence to general tree: Knowledge-guided geometry word problem solving. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 964–972. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. 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*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *CoRR*, abs/2311.04257.

Zheng Ye, Shang-Ching Chou, and Xiao-Shan Gao. 2008. An introduction to java geometry expert - (extended abstract). In *Automated Deduction in Geometry - 7th International Workshop, ADG 2008, Shanghai, China, September 22-24, 2008. Revised Papers*, volume 6301 of *Lecture Notes in Computer Science*, pages 189–195. Springer.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *CoRR*, abs/2309.12284.

Jiaxin Zhang, Yinghui Jiang, and Yashar Moshfeghi. 2024. GAPS: geometry-aware problem solver. *CoRR*, abs/2401.16287.

Jiaxin Zhang and Yashar Moshfeghi. 2024. Gold: Geometry problem solver with natural language description.

Ming-Liang Zhang, Zhong-Zhi Li, Fei Yin, and Cheng-Lin Liu. 2023a. LANS: A layout-aware neural solver for plane geometry problem. *CoRR*, abs/2311.16476.

Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. 2023b. A multi-modal neural geometric solver with textual clauses parsed from diagram. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 3374–3382. ijcai.org.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023c. Evaluating the performance of large language models on GAOKAO benchmark. *CoRR*, abs/2305.12474.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023d. Multimodal chain-of-thought reasoning in language models. *CoRR*, abs/2302.00923.

A Statistic Analysis

Table 6 presents a statistical breakdown of the GeoEval benchmark. This benchmark encompasses a total of 5,050 geometry math problems, categorized into four subsets: GeoEval-2000 (2,000 problems), GeoEval-backward (750 problems), GeoEval-aug (2,000 problems), and GeoEval-hard (300 problems). Besides the problem text, each problem in the dataset includes at least one of the following: a geometric diagram, a description of the diagram, or both. The majority of the correct answers are numerical values, with a minority comprising expressions, coordinates, or option letters, primarily in the GeoEval-hard subset.

B Source Datasets

Table 7 provides details on the source datasets that contribute to the GeoEval-2000 subset, including information on their content and characteristics. Meanwhile, Figure 6 visualizes the proportional contributions of these source datasets to the GeoEval-2000 subset, showcasing the variety and scope of the geometry problems collected from each source.

Element	Prompt
Task description	You are a result extraction bot. I will provide you with geometry questions and a model output, and you will help me extract the reference answers from the model's output.
Example 1	**Question:** As shown in the figure, in triangle ABC, $AB = AC$, $\angle A = 40°$, DE is the perpendicular bisector of AB. What is the degree measure of $\angle DBC$? () **Choices:** (A) 30.0 (B) 20.0 (C) 15.0 (D) 10.0 **Model response:** Given that $\angle A = 40°$, $AB = AC$, we can conclude that $\angle ABC = \angle ACB = 70°$. Also, since DE is perpendicular and bisects AB, we have $DB = AD$ and $\angle ABD = \angle A = 40°$. Therefore, $\angle DBC = \angle ABC - \angle ABD = 70° - 40° = 30°$. The answer is option A. **Extracted answer: A**
Example 2	**Question:** As shown in the diagram, the lateral surface area of a cone with a radius of the base r equals 6 cm and a height h equals 8 cm is (). **Choices:** (A) 94.2 (B) 150.8 (C) 188.5 (D) 251.3 **Model response:** Given $h = 8$ cm and $r = 6$ cm, the length of the generatrix of the cone can be set as l. Using the Pythagorean theorem, we have $l = \sqrt{6^2 + 8^2} = 10$ cm. The lateral area of the unfolded cone is $S_{\text{side}} = 0.5 \times 2 \times 6\pi \times 10 = 60\pi$ cm^2. Therefore, the lateral area of the cone is 60π cm^2. Therefore, the answer is C. **Extracted answer: C**
Example 3	**Question:** In triangle ABC, F is the midpoint of BC and point E is on the AC side. AC = 10. What is the length of AE? **Choices:** (A) 3.0 (B) 4.0 (C) 5.0 (D) 4.5 **Model response:** Since F is the midpoint of BC, EF is parallel to AB, so EF is the median of triangle ABC. Therefore, point E is the midpoint of AC. Therefore, AE = 0.5 × AC. Since AC = 10, AE = 5. Therefore, the answer is C. **Extracted answer: C**

Table 4: Task-specific instructions for extracting the answer. The table shows three examples with answers that can be extracted using the prompt.

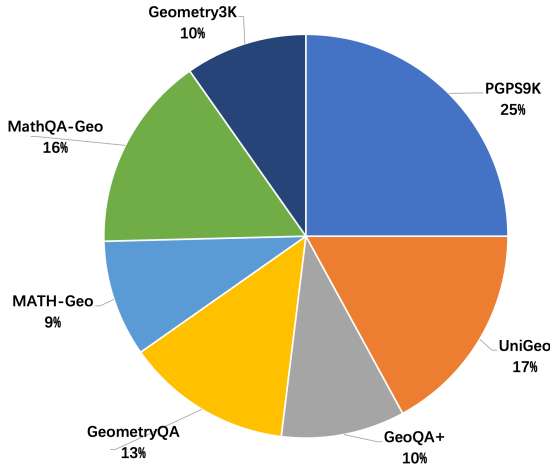


Figure 6: Distributions of source datasets.

C Distributions of Different Geometric Shapes

Figure 7 illustrates the varied distribution of geometric shapes within the GeoEval-2000 subset, highlighting the diversity of geometry concepts represented in this collection.

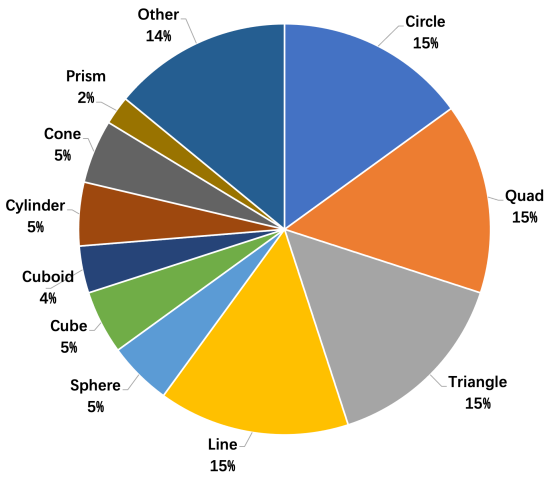


Figure 7: Distributions of different geometric shapes.

Regular expressions	Demonstration Examples
`value of (\w+) is\s*([\d.]+)`	The value of x is 3.5.
`correct answer is\s*(.+).`	correct answer is C."
`answer is\s*([\d.]+)`	answer is 17.1."
`answer should be\s*(.+) degrees`	Therefore, the answer should be choice D."
`answer to (.+) is (.+) degrees`	The answer to the angle ABC is 60°
`answer to the problem is\s*(.+)`	The correct answer to problem is $y = x^2 + 2x + 3$."
`The closest (.+) is (.+).`	So we got the area is 13.1. The closest answer is D."
`the (.+) is equal to (.+).`	The degree measure of angle ABC is 35 degrees.
`(.+) is approximately (.+) units`	So, the length of the line segment is approximately 10 units."

Table 5: Regular expressions used for extracting the answers that first step extraction cannot handle. The "Demonstration Examples" columns display corresponding examples that the regular expressions can match.

Total Numbers	
- GeoEval-2000	2,000
- GeoEval-backward	750
- GeoEval-aug	2,000
- GeoEval-hard	300
Input Types	
- text + description	1,120
- text + diagram	1,120
- text + description + diagram	1,166
Answer Types	
- number	5,050
- expression	232
- coordinate	68
Problem Types	
- flat geometry	5,050
- solid geometry	272
- analytic geometry	28
Others	
- average problem length	28
- average description length	34
- geometry shapes	12

Table 6: Statistics of GeoEval benchmark.

Source Dataset	Diagram	Diagram Descriptions	Quantity
Geometry3K	✓	✓	3001
PGPS9K	✓	✓	9022
UniGeo	✓	✗	4998 †
GeoQA+	✓	✗	2518
GeometryQA	✗	✗	1398
MATH	✗	✗	1349 ‡
MathQA	✗	✗	2625 ‡

Table 7: The information of source datasets for GeoEval-2000 dataset. The "†" symbol indicates that proving problems from the UniGeo dataset have been excluded. The "‡" sign specifies that the count only pertains to geometry problems within the dataset, focusing on problems directly relevant to the GeoEval-2000's scope.

D Backward Question Example

Figure 8 is an example from the GeoEval-backward subset.

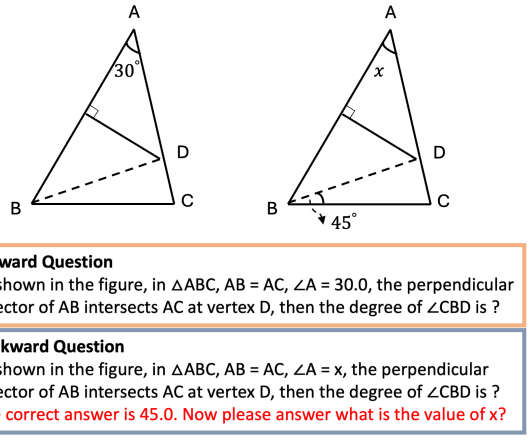


Figure 8: Example for the backward question. The left and right figures are diagrams for the forward question and backward question, respectively.

E Comparison between GeoEval-hard subset and other public datasets

To thoroughly assess the models' abilities in grasping concepts of solid and analytic geometry, the GeoEval-hard subset was created to include a diverse range of visual elements, such as three-dimensional views, across a spectrum of topics in solid geometry. The distinctions between the GeoEval-hard subset and other publicly available datasets are detailed in Table 8, demonstrating the unique coverage and complexity of the GeoEval-hard subset in comparison.

F Inspection of GeoEval-hard subset

To ensure the GeoEval-hard dataset's high quality and accuracy, and prevent the LLMs and MMs from recalling or inferring solutions to problems

Dataset	Solid Geometry		Analytic Geometry		
	#solid geometry shapes	#question type	#geometry curve knowledge	#question types	#grade
UniGeo (Chen et al., 2022a)	✗	calculate/prove	✗	–	6-12
GeoQA (Cao and Xiao, 2022)	✗	calculate	✗	–	6-12
Geometry3K (Lu et al., 2021)	✗	calculate	✗	–	6-12
PGPS9K (Zhang et al., 2023b)	✗	calculate/judge	✗	–	6-12
MathVista(Geometry Part) (Lu et al., 2023)	✗	calculate/judge	✗	–	–
MathVista(FunctionQA Part) (Lu et al., 2023)	✗	calculate/judge	✓	judge	–
GeoEval-hard	✓	judge/calculate/reason	✓	judge/calculate/reason	9-12

Table 8: Comparison between GeoEval-hard with other public datasets.

with similar structures or templates as those seen during training, we made several modifications to the original problems sourced from the GeoEval-hard-raw subset, where these modifications were made to prevent direct recall or plagiarism from the training data, while still preserving the underlying mathematical reasoning required. Specifically, we have adopted these steps:

1. We changed the variable names and numerical values used in the problem statements.

2. We swapped the original question with the provided conditions.

3. We used the original answer as a new condition in the reformulated problem statement.

Finally, we thoroughly analyze the revised reformulated problem to ensure it can be solved. Specifically, we form a team of six reviewers, each holding at least a Master's degree, to scrutinize every question. This evaluation process is structured in three phases: individual review, swap review, and candidate review. The primary focus lies on two key standards: the completeness and relevance of the geometric diagrams, and the reasonableness of the answers provided.

In the first phase, "individual review", each reviewer is randomly assigned 50 geometry math problems from the GeoEval-hard dataset. Their task is to assess the geometry math problems based on the standards, marking any that fail to meet these standards. During the "swap review" phase, these sets of 50 geometry math problems are exchanged among reviewers for a second evaluation. To ensure unbiased assessment, we hide the results of the initial review. Here, reviewers again highlight geometry math problems not conforming to the standards. The final phase, "candidate review", involves selecting geometry math problems for the dataset based on the outcomes of the first two phases. Geometry math problems unmarked in both phases are retained, those marked in both are discarded, and those highlighted in only one phase undergo further examination by the entire review team, with the majority decision determining their inclusion.

G Examples from GeoEval Representing Five Features

G.1 Comprehensive Variety

Figure 9 presents sample data from the GeoEval-2000 subset, illustrating its diversity in terms of data sources.

G.2 Varied Problems

Figure 10 displays examples of three distinct problem types in the GeoEval benchmark: flat geometry, analytic geometry, and solid geometry.

G.3 Dual Inputs

Figure 9 shows that the GeoEval benchmark comprises geometry math problems that contain both diagrams and textual descriptions, as well as problems that include textual descriptions alone.

G.4 Diverse Challenges

Figure 11 showcases examples from the GeoEval-2000, GeoEval-backward, GeoEval-aug, and GeoEval-hard subsets, illustrating the diverse challenges within the GeoEval benchmark.

G.5 Complexity Ratings

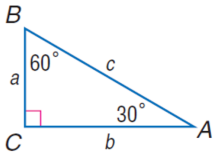
Every problem in the GeoEval benchmark is annotated with a complexity rating, indicating the level of skill necessary to solve it, as shown in Figure 12.

H Algorithm for Classifying Geometry Math Problems Complexity

Algorithm 1 details our methodology for classifying each geometry math problem into distinct levels of complexity.

GeoEval – 2000 from PGPS9K

- Problems: If c = 5, find b.
- Diagram Descriptions:
  - structure: ["line B A", "line C A", "line B C"],
  - semantic: ["CA \\perp BC on C", "BA = c", "BC = a", "AC = b", "m \\angle ABC = 60", "m \\angle BAC = 30"]
- Choice List: A. 1.7, B. 2.6, C. 3.5, D. 4.3
- Solution Program: Equal c N0 Gsin N3 N1 N4 Get b



GeoEval – 2000 from MathQA

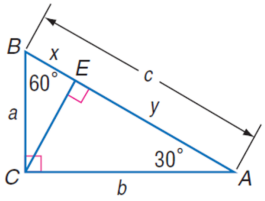
- Problems: eight cubes , each with a volume of 512 cm ^ 3 , are joined to form one large cube . what is the surface area of the large cube ?
- Diagram Descriptions: n/a
- Choice List: A. 4096 sq cm, B. 1536 sq cm, C. 1024 sq cm, D. 2048 sq cm, E. 512 sq cm
- Program: multiply(const_2,const_4)|multiply(n0,#0)|cube_edge_by_volume(#1)|surface_cube(#2)

GeoEval – 2000 from GeometryQA

- Problems: The length, width, and height of a rectangular box are (4/5) meter, (3/4) meter, and (1/6) meter, respectively. What is the maximum amount of water that can be held in this box?
- Diagram Descriptions: n/a
- Choice List: A. 1/10, B. 1/12, C. 1/15, D. 1/20
- Program: x=cuboid_volume((4/5), (3/4), (1/6))\nx=(4/5)*(3/4)*(1/6)

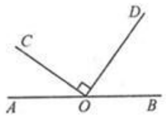
GeoEval – 2000 from Geometry3K

- Problems: If x = 7\\sqrt{3}, find b.
- Diagram Descriptions:
  - structure: ["line B C", "line C A", "line B E A", "line E C"],
  - "semantic": ["BA \\perp EC on E", "BC \\perp CA on C", "BC = a", "CA = b", "BE = x", "AE = y", "BA = c", "m \\angle EAC = 30", "m \\angle EBC = 60"]
- Choice List: A. 7.0, B. 12.1, C. 24.2, D. 42.0
- Program: Equal x N0 Gtan V0 N3 N7 Gsin V0 N2 N6 Get b



GeoEval – 2000 from UniGeo

- Problems: As shown in the figure, point O is on the straight line AB and OC ⊥ OD, if ∠COA = 36.0, then the size of ∠DOB is ()
- Diagram Descriptions: n/a
- Choice List: A. 36.0, B. 54.0, C. 64.0, D. 72.0
- Solution Program: g_minus C_3 N_0 g_minus V_0 C_2



GeoEval – 2000 from Math-Geometry

- Problems: Compute $\\sin 60^\\circ$.
- Diagram Descriptions: n/a
- Choice List: A. \frac{\sqrt{1}}{2}, B. \frac{\sqrt{3}}{2}, C. \frac{\sqrt{5}}{8}, D. \frac{\sqrt{1}}{7}
- Solution Program: n/a

GeoEval – 2000 from GeoQA+

- Problems: As shown in the diagram, the diagonals AC and BD of square ABCD intersect at point O. Point M is on side AD, and OM is connected. A perpendicular line is drawn from point O to OM, intersecting CD at point N. If the area of quadrilateral MOND is 2, then what is the length of BD?
- Diagram Descriptions: n/a
- Choice List: A. 1.4, B. 2.0, C. 2.8, D. 4.0
- Solution Program: g_double N_0

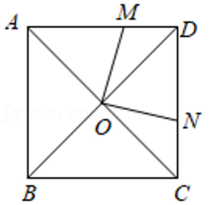
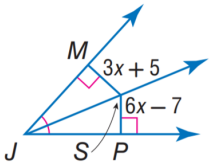


Figure 9: Examples from GeoEval-2000 dataset. The golden answer choice is highlighted in red color.

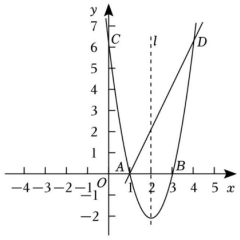
GeoEval – 2000 flat geometry

- Problems: If ST = 8, TR = 4, and PT = 6, find QR.
- Diagram Descriptions:
  - "structure": ["line Q R", "line S T R", "line Q P S", "line P T"],
  - "semantic": ["PT \\parallel QR"]
- Choice List: A. 6.0, B. 8.0, C. 9.0, D. 10.0
- Solution Program: g_double N_0



GeoEval – 2000 analytic geometry

- Problems: The graph shows a parabola $y=a x^{2}+b x+c \\quad(a, b, c$ are constants, and $a \\neq 0)$ passing through three points $A(1,0), B(3,0), C(0,6)$. Find the expression of the parabola.
- Diagram Descriptions: n/a
- Choice List: A. $y=2 x^{2}-8 x+6$, B. $y=-2 x^{2}+8 x+6$, C. $y=2 x^{2}+8 x+6$, D. $y=-2 x^{2}-8 x+6$
- Solution Program: n/a



GeoEval – 2000 solid geometry

- Problems: What is the volume of the triangular pyramid P-B1AM in the figure below? ABC-A1B1C1 is a right prism with angle ABC equal to 90 degrees, AB=BC=2, AA1=2, M is the midpoint of BC, N is the midpoint of A1C1, point P is on the line segment B1N, point Q is also on the line segment B1N, but first on the line segment AM, and AQ=2/3AM. S is the intersection of AC1 and A1C. If PS is parallel to the plane B1AM..
- Diagram Descriptions: n/a
- Choice List: A. 2/3, B. 1/3, C. 2/5, D. 3/7
- Solution Program: n/a

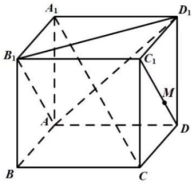
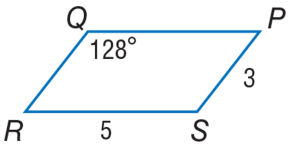


Figure 10: Examples of the flat geometry problem, the analytic geometry problem, and the solid geometry problem in GeoEval benchmark.

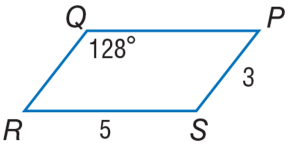
GeoEval – 2000

- Problems: Use parallelogram PQRS to find m \\angle R.
- Diagram Descriptions:
  - "structure": ["line R S", "line P S", "line Q P", "line Q R"],
  - "semantic": ["RS = 5", "SP = 3", "m \\angle RQP = 128"]
- Choice List: A. 3.0, B. 5.0, C. 52.0, D. 128.0



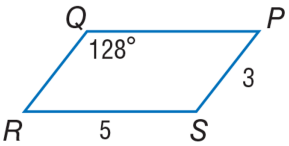
GeoEval – backward

- Problems: Use parallelogram PQRS to find m \\angle R. The correct answer is 52.0. Now please answer what is the value of x?
- Diagram Descriptions:
  - "structure": ["line R S", "line P S", "line Q P", "line Q R"],
  - "semantic": ["RS = 5", "SP = 3", "m \\angle RQP = x"]
- Choice List: A. 3.0, B. 5.0, C. 52.0, D. 128.0



GeoEval – aug

- Problems: What is the value of angle R in the provided parallelogram PQRS?
- Diagram Descriptions:
  - "structure": ["line R S", "line P S", "line Q P", "line Q R"],
  - "semantic": ["RS = 5", "SP = 3", "m \\angle RQP = 128"]
- Choice List: A. 3.0, B. 5.0, C. 52.0, D. 128.0



GeoEval – hard

- Problems: In $\\triangle ABO$ shown below, with $\\angle B=90^{\\circ}, AO=5, AB=4$, what is the sine of $\\angle A$?
- Diagram Descriptions: n/a
- Choice List: A. \frac{3}{4}, B. \frac{4}{3}, C. \frac{3}{5}, D. \frac{4}{5}

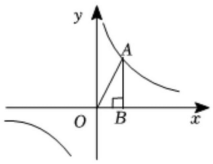


Figure 11: Examples of GeoEval-2000, GeoEval-backward, GeoEval-aug, GeoEval-hard subsets.

GeoEval – 2000

- Problems: As shown in the figure, a large parasol can be approximately regarded as a conical shape when the umbrella surface is expanded. The length of its generatrix is 5.0 and the bottom radius is 3.0. The area of fabric required to make this parasol is () square (Seams are not counted)
- Diagram Descriptions: n/a
- Choice List: A. 25.1, B. 37.7, C. 47.1, D. 62.8
- Complexity: 0.15

Figure 12: Example for a problem annotated with complexity in the GeoEval benchmark.

I Evaluation Details

I.1 Model Hyper-parameters

Table 9 presents the complete list of hyper-parameters applied to the models throughout the evaluation phase.

I.2 Instruction Prompt Used for Evaluating Models

Before employing instruction prompts to steer model responses, we combine the problem texts, diagram descriptions, and choice lists from an example, as depicted in the "Merge" row of Table 10. Following this combination, as illustrated in the "Instruction" row of Table 10, we incorporate instruction prompts into the merged texts and then forward these to the models to generate responses.

J Reason for Removing InstructBLIP from the Comparison

As shown in Figure 13, InstructBLIP's responses on the GeoEval-2000 subset are typically scalar, lacking any intermediate reasoning steps. This suggests that InstructBLIP may have been exposed to GeoEval-2000 questions during its pre-train phase, leading to the memorization of answers. This is supported by the observed performance decline from GeoEval-2000 to GeoEval-aug, which falls from 52.18% to 35.00%. Additionally, InstructBLIP tends to directly generate option letters (e.g., "A") for the GeoEval-hard subset without any reasoning process, resulting in an improbably high accuracy rate of 70.30% for this subset. Consequently, in our subsequent analysis and discussions, we have chosen to exclude the InstructBLIP model.

K Models Performances across Different Data Sources

Table 11 shows model performances on the GeoEval-2000 subset according to the different original datasets. We can observe that WizardMath models still achieve the best accuracy scores on almost all datasets.

Model Name	Generation Parameters	Comments
CodeGen2-16B	do_sample=True, top_k=0.5, top_p=0.5, max_tokens=512	model=""Salesforce/codegen2-16B"
WizardMath-7B-V1.1	temperature=0.0, top_p=1, max_tokens=1024	vLLM package
WizardMath-70B	temperature=0.0, top_p=1, max_tokens=1024	vLLM package
GPT-3.5	temperature=0.7, max_tokens=512	version="gpt-3.5-turbo-0125"
GPT-4	temperature=0.7, max_tokens=512	version="gpt-4-1106-preview"
llava-7B-V1.5	temperature=0.0, max_new_tokens=512	llava package
Qwen-VL	temperature=0.0, max_new_tokens=512	model="Qwen/Qwen-VL"
mPLUG-Owl2	do_sample=True, top_p=0.7, max_tokens=512	model="mPLUG-Owl2"
InstructBLIP	do_sample=False, num_beams=5, max_tokens=512, top_p=0.9, temperature=1.0	model="Salesforce/instructblip-vicuna-7b"
GPT-4V	temperature=0.0, max_tokens=512	version="gpt-4-vision-preview"

Table 9: The hyper-parameters 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.

	Template	Example
Merge	Here are the basic description of the diagram: ${diagram descriptions}, ${problems texts}, The Choices are: ${choice list}	Please solve this math problem: Here are the basic description of the diagram: line B A, line C A, line B C\nCA \\perp BC on C, BA = c, BC = a, AC = b, m \\angle ABC = 60, m \\angle BAC = 30\nIf c = 5, find b. The Choices are: [1.7, 2.6, 3.5, 4.3]
Instruction	Please solve this math problem: ${Merge} ### Problem-solving Bot:	Please solve this math problem: Here are the basic description of the diagram: line B A, line C A, line B C\nCA \\perp BC on C, BA = c, BC = a, AC = b, m \\angle ABC = 60, m \\angle BAC = 30\nIf c = 5, find b. The Choices are: [1.7, 2.6, 3.5, 4.3] ### Problem-solving Bot:

Table 10: Templates and examples provided illustrate the process of merging and instruction creation. The placeholder "${Merge}" represents the combined texts of "diagram descriptions," "problems texts", and "choice list". In cases where "diagram descriptions" are absent, the phrase "Here are the basic description of the diagram:" is omitted.

InstructBLIP Prediction Example on GeoEval – 2000 subset

- Problems: As shown in the figure, a large parasol can be approximately regarded as a conical shape when the umbrella surface is expanded. The length of its generatrix is 5.0 and the bottom radius is 3.0. The area of fabric required to make this parasol is () square (Seams are not counted)
- Choice List: A. 25.1, B. 37.7, C. 47.1, D. 62.8
- InstructBLIP predictions: The answer is 47.1.

Figure 13: One example of InstructBLIP prediction on GeoEval-2000 subset.

Models	GeoEval-2000 Data Sources					
	MATH (Geometry) (%)	GeometryQA (%)	GeoQA+ (%)	PGPS9K (%)	UniGeo (%)	MathQA (Geometry) (%)
CodeGen2-16B	0.36	0.35	0.44	0.18	0.41	0.25
GPT-3.5	0.35	0.31	0.19	0.27	0.23	0.26
GPT-4	0.58	0.74	0.27	0.28	0.27	0.44
WizardMath-7B-V1.1	0.58	0.53	0.59	0.55	0.54	0.35
WizardMath-70B	0.54	0.58	0.62	0.54	0.57	0.35
llava-7B-V1.5	0.26	0.4	0.12	0.15	0.12	0.19
Qwen-VL	0.29	0.46	0.27	0.22	0.32	0.24
mPLUG-Owl2	0.27	n/a	0.29	0.46	0.27	0.0
InstructBLIP	0.0	n/a	0.59	0.48	0.57	0.0
GPT-4V	0.45	0.61	0.34	0.38	0.45	0.38

Table 11: The accuracy scores achieved by models on different sources datasets constituting the GeoEval-2000 subset.

Algorithm 1: Algorithm for classifying geometry math problems complexity

Input: All Problem Texts T, All Diagram Descriptions D, All Golden Solution Programs S
Output: Complexity for each problem
$len_{T,D} = 0$;
$len_S = 0$;
for t ***in*** T, d ***in*** D, s ***in*** S **do**
  $len_{T,D} += len_t + len_d$; `/* sum up the length of problem texts and the length of diagram descriptions. */`
  $len_S += len_s$; `/* sum up the length of golden solution programs. */`
end
for t ***in*** T, d ***in*** D, s ***in*** S **do**
  $C_{t,d,s} \leftarrow \alpha \times \frac{len_t + len_d - \min(len_{T,D})}{\max(len_{T,D}) - \min(len_{T,D})} + (1-\alpha) \times \frac{len_s - \min(len_S)}{\max(len_S) - \min(len_S)}$;
  if $0.0 \leq C_{t,d,s} \leq 0.2$ **then**
    Complexity $\leftarrow$ Easy; `/* classify the problem as Easy problem. */`
  else if $0.2 < C_{t,d,s} \leq 0.6$ **then**
    Complexity $\leftarrow$ Middle; `/* classify the problem as Middle problem. */`
  else if $0.6 < C_{t,d,s} \leq 1.0$ **then**
    Complexity $\leftarrow$ Hard; `/* classify the problem as Hard problem. */`
  end
end