

GeoLaux: A Benchmark for Evaluating MLLMs' Geometry Performance on Long-Step Problems Requiring Auxiliary Lines

Yumeng Fu^{1,2}, Jiayin Zhu^{1,2}, Lingling Zhang^{1,2*}, Wenjun Wu^{1,2*}, Bo Zhao^{1,2},
Shaoxuan Ma⁴, Yushun Zhang^{1,2}, Jun Liu^{1,3}

¹School of Computer Science and Technology, Xi'an Jiaotong University

²Ministry of Education Key Laboratory of Intelligent Networks and Network Security, China

³Shaanxi Province Key Laboratory of Big Data Knowledge Engineering, China

⁴School of Software Engineering, Xi'an Jiaotong University

yumfuu@stu.xjtu.edu.cn, nickjunwork@163.com, {zhanglling, liukeen}@xjtu.edu.cn

Abstract

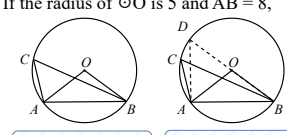
Geometry problem solving (GPS) poses significant challenges for Multimodal Large Language Models (MLLMs) in diagram comprehension, knowledge application, long-step reasoning, and auxiliary line construction. However, current benchmarks lack fine-grained evaluation for long-step problems necessitating auxiliary construction. To address these limitations, we present GeoLaux, a fine-grained annotated dataset comprising 2186 calculation and proof problems. It features long-step reasoning (with an average solution length of 6.51 steps, maximum of 24 steps) and auxiliary line construction (required in 41.8% of problems). Building on the dataset, we conduct a comprehensive five-dimensional evaluation of 23 leading MLLMs. The evaluation yields three pivotal findings: First, models perform significantly worse on long-step problems compared to short-step ones, with 18 models exhibiting a performance drop of over 50%. Second, it is crucial to enhance models' understanding, awareness, and proficiency in auxiliary line construction, which is vital for overall geometric reasoning. Third, limited answer hints effectively improve process correctness, whereas explicit answers lead models to neglect intermediate reasoning steps. These findings position GeoLaux both to benchmark MLLMs geometry reasoning abilities and to guide their improvement. Data and code are available at <https://github.com/Candice-yu/GeoLaux>

1 Introduction

As a cornerstone of mathematics, Geometry Problem Solving (GPS) epitomizes the advanced cognitive patterns of the human mind. It demands the integration of diverse core competencies: extensive geometric knowledge base, rigorous logical reasoning, precise computational skills, spatial visualization ability, and strategic auxiliary line construction

*Corresponding authors.

Question: As shown in the figure, $\odot O$ is the outer circle of $\triangle ABC$, connecting OA , OB . If the radius of $\odot O$ is 5 and $AB = 8$, the value of $\cos \angle ACB$ is?



Answer: 3/5

Type: Calculation Problem [Original Image](#) [Auxiliary Image](#)

Auxiliary Text: Extend BO to meet $\odot O$ at point D , then connect AD .

Reference Solution Process:

- Step1:** \therefore Diameter = $2 \times$ radius, $\therefore BD = 2 \times 5 = 10$.
- Step2:** \therefore The Inscribed Angle Theorem, $\therefore \angle ACB = \angle D$ (both subtended by arc AB).
- Step3:** \therefore Thales' Theorem (diameter subtends right angle), $\therefore \angle DAB = 90^\circ$.
- Step4:** $\therefore \odot O$'s radius = 5 $\therefore BD = 10$.
- Step5:** $\therefore \triangle DAB$ is right-angled, $BD = 10$, and $AB = 8$, $\therefore AD = \sqrt{BD^2 - AB^2} = \sqrt{10^2 - 8^2} = 6$.
- Step6:** $\therefore AD = 6$, $BD = 10$, $\therefore \cos \angle ADB = \text{adjacent/hypotenuse} = AD/BD = 6/10 = 3/5$.
- Step7:** $\therefore \angle ACB = \angle ADB$ (Step2), $\therefore \cos \angle ACB = 3/5$.

Step Length: 7

Figure 1: An illustration of example from GeoLaux.

(Jonsson et al., 2022; Yan et al., 2025). Given these challenging requirements, this task has consistently attracted widespread attention from the community (Lu et al., 2021; Chen et al., 2021; Wu et al., 2024; Trinh et al., 2024; Cheng et al., 2025).

Multimodal large language models (MLLMs), represented by GPT-4o (Hurst et al., 2024), have recently emerged as a significant focus of research attention. By combining the inherent strengths of Large Language Models (LLMs) in knowledge, reasoning, and calculation with visual modules, MLLMs demonstrate remarkable performance in multimodal reasoning (Driess et al., 2023; Yin et al., 2024). These characteristics reflect their potential for GPS, leading to numerous studies (Anand et al., 2024a; Zhang et al., 2024a) evaluating MLLMs' geometric reasoning abilities.

Table 1 summarizes existing benchmarks for evaluating MLLMs' geometric reasoning, which generally exhibit three main limitations: (1) *Ab-*

Benchmark	Size	Type	Steps Length Avg. / Max	Long-step Prob. Num.	Ultra-long Prob. Num.	Auxiliary Lines	Source	Evaluation		
								Answer	Error	Step
Geometry3K (2021)	601	C	-	-	-	×	S	✓	×	×
GeoQA (2021)	755	C	1.96 / 4	0	0	×	S	✓	×	×
UniGeo (2022)	1447	C P	-	-	-	×	S	✓	×	×
PGPS9K (2023)	1000	C	2.43 / -	0	0	×	P S	✓	×	×
IMO-AG-30 (2024)	30	P	-	-	-	✓	S	✓	×	×
GPSM4K (2024b)	200	C P	-	-	-	×	S A	✓	×	×
GeoEval (2024a)	2000	C	-	-	-	×	P A S	✓	✓	×
GeoSense (2025)	1789	C	5.70 / 16*	169*	14*	×	P S	✓	✓	×
SolidGeo (2025b)	3113	C	-	-	-	×	S P	✓	✓	×
GeoLaux (ours)	2186	C P	6.51 / 24	292	208	✓	S	✓	✓	✓

Table 1: Comparison with other geometry benchmarks. **Type:** C=Calculation, P=Proving. **Source:** S=Self-Sourced, P=Collected from Public Datasets, A=Augmented from Existing Data. *: Derived from the official dataset as originally textually unreported. Long-step problems: 9–12 steps. Ultra-long problems: over 13 steps.

reasoning is essential for solving complex geometric problems, making accurate long-step evaluation crucial for advancing model performance. However, constrained by limited step length, current benchmarks cannot fully assess MLLMs’ long-step reasoning, illustrated by SolidGeo (Wang et al., 2025b) having only 6.7% multi-step questions and GeoSense (Xu et al., 2025) with just 14 problems exceeding 13 steps. (2) *Absence of auxiliary line evaluation.* Constructing correct auxiliary lines is a vital evaluation dimension that critically tests MLLMs’ deep understanding of both geometric diagrams and textual problems. Given an image with n geometric primitives (i.e. points, lines and circles), there exist n^3 possible auxiliary line constructions (Marinković, 2017), posing a significant challenge to spatial reasoning capabilities of models. However, benchmarks evaluating this aspect are currently lacking. (3) *Coarse-grained process evaluation.* While recent LLM research highlights the necessity of fine-grained trajectory analysis (e.g., error localization and quality assessment) for evaluating long-step reasoning (Zhang et al., 2025; Yan et al., 2026; Xu et al., 2026), existing GPS benchmarks solely use answer correctness as the success criterion, with process analysis limited to error classification (Zhang et al., 2024a; Xu et al., 2025). Such coarse-grained evaluation fails to identify model weaknesses and guide improvement.

To address these, we present a plane geometry problem dataset GeoLaux, which comprises 2186 problems collected from Zhongkao mathematics papers across 34 provincial-level regions in China over past two years. This dataset exhibits three key characteristics: (1) **long-step reasoning** with problems averaging 6.51 solution steps (up to 24 steps), (2) **annotated auxiliary lines** including both de-

tailed construction methods and resulting geometric diagrams, and (3) **dual problem types** comprising 1,418 calculation and 768 proving problems. As shown in Figure 1, we annotate step-by-step solution process for each problem, establishing foundation for fine-grained process evaluation.

Besides the dataset, we design a fine-grained framework to evaluate MLLM problem-solving. Specifically, beyond general **answer correctness** (measured by ACS) and **error type**, we introduce three additional dimensions tailored for complex reasoning: **solution process correctness** (PCS), **solution process quality** (PQS), and **auxiliary line construction**, totaling 5 dimensions and 3 metrics. Leveraging this evaluation framework, we assess 23 state-of-the-art MLLMs, including 10 thinking models and 13 non-thinking models. The results demonstrate that Gemini-2.5 Pro (Deepmind, 2025) achieves the highest overall performance, followed by o3 (OpenAI, 2025a) and Qwen3-VL-32B-thinking (Yang et al., 2025). Our analysis reveals three critical findings:

- **MLLM performance degrades significantly on long-step problems compared to shorter ones:** Nine models show a performance drop of over 50% from short-step to ultra-long-step problems, with some exceeding 90% or even reaching 100%.
- **MLLMs lack the awareness and capability to proactively construct auxiliary lines:** They fail to recognize when such constructions are necessary and struggle to generate correct lines when attempted. Yet, their strong performance gains under standard construction prompting underscore the vital role of accurate auxiliary lines.
- **MLLMs perform better with limited answer hints, but decline when answers are clearly disclosed:** Compared to free-response problems,

models often achieve higher process correctness scores on multiple-choice questions, yet lower on proving problems. This implies options act as incentives for reasoning, while explicit answers lead models to disregard process correctness.

In this way, we establish a fair evaluation benchmark that not only assesses MLLMs' reasoning capabilities on long-step auxiliary line problems, but also provides clear guidance for enhancing their geometry reasoning performance.

2 Related Work

Prior to the rapid development of MLLMs, several established benchmarks existed for evaluating traditional geometric problem solving methods, including Geometry3K (Lu et al., 2021), GeoQA (Chen et al., 2021), and UniGeo (Chen et al., 2022). Typically featuring low difficulty and a limited variety of problem types, they are inadequate for meeting current evaluation demands. Consequently, several specialized benchmarks for evaluating MLLMs' math reasoning capabilities have emerged in recent years, such as MathVista (Lu et al., 2023), MathVerse (Zhang et al., 2024b), We-Math (Qiao et al., 2025), but these works have not focused on analyzing the aspect of GPS. Among works focused on geometry, IMO-AG-30 (Trinh et al., 2024) provide 30 Olympiad-level problems, but exclusively address theorem proving without algebra; GeoEval (Zhang et al., 2024a) restructures existing problems into a unified format to assess answer accuracy; GeoSense (Xu et al., 2025) evaluates recognition of geometric principles but overlooks key dimensions such as diagram comprehension and auxiliary line construction; SolidGeo (Wang et al., 2025b) focuses on solid geometry problems, yet lacks attention to critical auxiliary line construction and includes only 6.7% multi-step questions. Similarly, while the recent Geoint benchmark (Wei et al., 2025) introduces Lean 4 to formalize auxiliary lines, its evaluation relies on rigid code-level matching, falling short of deeply assessing the models' dynamic geometric reasoning capabilities.

In conclusion, current MLLM geometry benchmarks lack fine-grained process evaluation, auxiliary line assessment, and multi-step reasoning evaluation, necessitating new evaluation standards.

3 GeoLaux Data Set

GeoLaux is a challenging plane geometry dataset comprising 2186 fully verified problems, divided

into 1418 calculation and 768 proving problems. Of the calculation problems, 522 are multiple-choice (single-answer) and 896 are free-response questions. As shown in Figure 1, each problem in our dataset contains 8 annotated elements: problem text, geometric diagram, type (proving or calculation), answer, step-by-step solution, step length, auxiliary line construction text, and auxiliary line construction image. This section describes its semi-automated construction pipeline, which consists of three main stages: data acquisition, step segmentation, and auxiliary line extraction.

Data Acquisition. To ensure data authenticity and comprehensiveness, we systematically select plane geometry problems from the High School Entrance Examination (HSEE/Zhongkao) mathematics papers across China's 34 provincial-level regions as our raw dataset. These questions comprehensively cover the core plane geometry knowledge required in secondary education. Beyond the original problems' texts and diagrams, we further incorporate expert-curated standard answers and solutions from official exam materials. Every problem is carefully verified for: (1) diagram clarity, (2) text-diagram correspondence, (3) answer accuracy, and (4) detailed annotation of solution processes. These validated real-world problems lay a solid foundation for our benchmark.

Step Segmentation. Based on the fundamental logical structure of mathematical reasoning, we define each "because(:)-therefore(:)" pair as a complete reasoning step. Gemini-2.5-Pro is tasked with segmenting the pre-annotated standard solutions into such steps while explaining each segmentation decision to make sure the splitting follows our rules. This process generates standardized solution step length for every problem in our dataset, serving as crucial labels for subsequent analysis. To ensure the authenticity and accuracy of these step counts, we conducted a rigorous manual review of the model-segmented results. Figure 2 presents the step length distribution of all 2186 problems. The dataset includes a substantial number of long-step and ultra-long-step problems, with an average step of 6.51 and a maximum step of 24. This step segmentation process establishes an ideal testbed for evaluating long-step reasoning capabilities.

Auxiliary Line Extracting. Our dataset contains numerous problems requiring auxiliary lines to evaluate MLLMs' capability in auxiliary structure

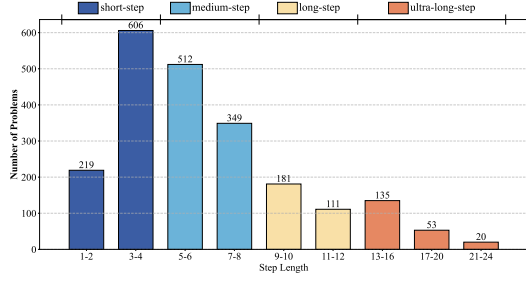


Figure 2: Problem quantity statistics across step lengths

construction. We employed Gemini-2.5-Pro to extract auxiliary line construction steps from the pre-annotated standard solutions. Considering the importance of visual input for MLLMs, we further manually collected corresponding diagrams with auxiliary lines from original exam papers, forming the visual-text pairs illustrated in Figure 3. The auxiliary lines are classified by difficulty into simple ones (involving only point connections) and complex ones (creating new geometric primitives like perpendiculars, angle bisectors, or inscribed circles). According to our statistics, GeoLaux includes 334 problems requiring complex auxiliary lines (15.3% of the total) and 580 problems needing simple auxiliary lines (26.5% of the total), laying the groundwork for comprehensive evaluation.

Comparison with Existing DataSets. As illustrated in Table 1, GeoLaux demonstrates three key advantages over other datasets: **(1) Long Steps:** GeoLaux surpasses existing benchmarks in both average and maximum solution step lengths. Notably, with 208 ultra-long-step problems (exceeding 13 steps), it poses significant challenges for models. **(2) Unique Auxiliary Line Annotation:** While auxiliary lines are essential for IMO-AG-30 (Trinh et al., 2024), which contains only 30 problems and provides no annotations for auxiliary lines. GeoLaux is the first large benchmark to provide complete, explicit and multi-modal annotations for auxiliary line construction methods. **(3) Integrated Calculation and Proving Problems:** GeoLaux enables fair cross-type performance comparisons of MLLMs in geometric problem solving.

4 Evaluation Strategy

Based on the dataset, we propose a 5-dimensional evaluation framework (Figure 4). In addition to general **answer correctness** (with metric ACS) and **error type**, we devise 3 novel evaluation dimensions: **process correctness** (with metric PCS),

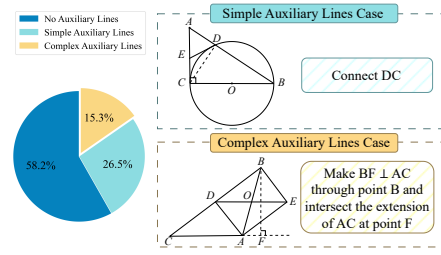


Figure 3: Distribution of auxiliary line types.

process quality (with metric PQS), and **auxiliary line impact**. The following four subsections detail these evaluation dimensions, with a concluding analysis of the evaluation framework’s reliability.

4.1 Correctness Evaluation (ACS & PCS)

The first two dimensions both evaluate MLLMs’ solution correctness, employing the Answer Correctness Score (ACS) alongside our novel and stricter Process Correctness Score (PCS). While ACS checks only the final answer, PCS mandates correct reasoning. To facilitate the independent evaluation of these two components, we require MLLMs to output step-by-step reasoning and the summarized final answer in JSON format, with the format validity ensured through manual review.

Answer Correctness Evaluation (ACS). The model’s self-summarized answer is then compared with the ground-truth answer through our evaluator model. Specifically, for problem q with ground-truth answer a and model answer \hat{a} , the Answer Correctness Score (ACS) is formally defined as:

$$ACS = \begin{cases} 1 & \text{if } \hat{a} = a \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Depending on the problem type, a can be a numerical value, an option, or a geometric condition.

Process Correctness Evaluation (PCS). We observe that MLLMs occasionally generate correct answers through flawed processes, a phenomenon we term **False Positives**. This necessitates a stricter process evaluation to assess true performance.

The structured step-by-step solutions enable our evaluator to score each individual reasoning step, assigning 1 for correct steps and 0 for incorrect ones. Given an n -step solution process, the evaluator assigns scores as follows:

$$\eta = (\eta_1, \eta_2, \dots, \eta_n), \quad \eta_i \in \{0, 1\}. \quad (2)$$

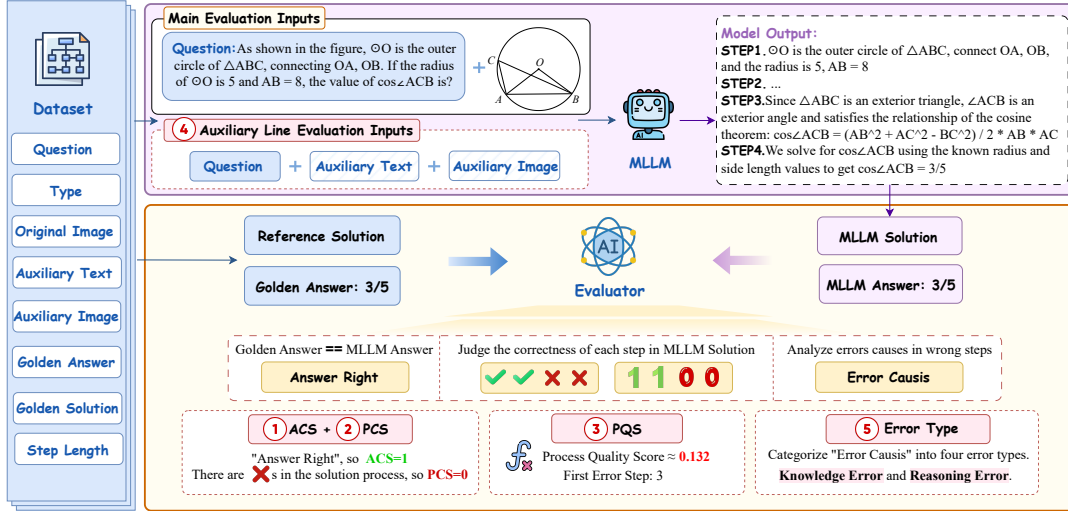


Figure 4: Five-dimension evaluation framework of GeoLaux. Given golden answer and solution from dataset, evaluator conducts a comprehensive assessment of MLLM outputs across the following dimensions: ① answer correctness, ② process correctness, ③ process quality, ④ auxiliary line impact, and ⑤ error type.

Building on this fine-grained scoring system, the Process Correctness Score (PCS) is defined as:

$$\text{PCS} = \begin{cases} 1 & \text{if } (\hat{a} = a) \wedge (0 \notin \eta) \\ 0 & \text{otherwise} \end{cases}. \quad (3)$$

This metric rigorously evaluates problem-solving correctness, requiring not only accurate final answers but also error-free reasoning processes.

4.2 Process Quality Evaluation (PQS)

To ensure a fair comparison of solution quality across different MLLMs, we design a step weight function that assigns specific weights to each step's score, ultimately computing a weighted overall process quality score. Our weighting function incorporates the following considerations:

- 1. Decreasing function:** Models that make errors in earlier steps exhibit weaker capability for accurate long-step reasoning, which implies that earlier steps should be assigned higher weights.
- 2. Convex function:** The importance gap is larger for earlier steps and smaller for later ones. For example, two models erring at steps 2 and 4 should show larger score differences than those erring at steps 12 and 14.
- 3. Moderate decreasing rate:** Weighting function should not decrease too rapidly. For long-step problems, performance in later steps remains critical and should retain significant weight.

Given these considerations, for a solution process with n steps, we define the weight function

for the i -th step as:

$$y_i = e^{-\frac{i}{n}}. \quad (4)$$

The initial process quality score, using Equation 2 for grading and Equation 4 for weighting, is defined as follows:

$$\text{PQS}' = \frac{\sum_{i=1}^n \eta_i \cdot y_i}{\sum_{j=0}^n y_j} = \frac{\sum_{i=1}^n \eta_i \cdot e^{-\frac{i}{n}}}{\sum_{j=1}^n e^{-\frac{j}{n}}}. \quad (5)$$

However, since model solutions always contain some correct steps, PQS' consistently falls between 0.6 and 1, failing to highlight models' differences in reasoning ability. Therefore, we apply the tanh activation function on PQS' .

$$\text{PQS} = \tanh(\alpha(\text{PQS}' - 1)) + 1, \quad (6)$$

where α is a hyperparameter set to 3.5. Through this approach, we obtain the final PQS normalized to $[0,1]$, which provides a more refined metric for evaluating MLLMs' reasoning capabilities. Appendix B records our other considerations on the weight functions and hyperparameters.

4.3 Auxiliary Line Evaluation

To rigorously evaluate the model's awareness of constructing auxiliary lines and verify their importance, we establish two distinct settings: one allows model to solve problems autonomously without additional prompts; the other, we provide MLLM with textual auxiliary line construction methods

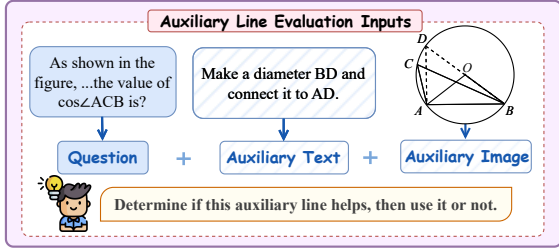


Figure 5: Illustration of auxiliary line evaluation inputs.

and the corresponding annotated diagrams from GeoLaux (as shown in Figure 5), heuristically prompting model to decide whether to adopt these lines to solve problems. Based on these two scenarios, our evaluation focuses on the following two aspects.

First, we examine model’s auxiliary construction awareness through calculating construction rate under both scenarios. Specifically, construction rate is obtained by having our evaluator determine the probability of MLLM employing specific auxiliary line constructions through keyword detection in its solutions. By comparing construction rates between autonomous and prompted scenarios, we can quantify model’s inherent awareness of utilizing auxiliary lines.

Second, we measure the impact of auxiliary lines through calculating model’s PCS after auxiliary construction under both scenarios. By comparing the PCS scores across these two settings, we can analyze how proper auxiliary construction influences the geometric problem-solving performance.

4.4 Error Type Evaluation.

Understanding error causes enables targeted improvements for MLLMs in GPS. Consequently, we additionally conduct detailed error analysis for each step in the model’s solution process, categorizing failure steps into four distinct types: **figure comprehension error, knowledge error, calculation error, and logical reasoning error.**

Figure comprehension error means model fails to correctly comprehend geometric elements and their relationships in the geometric figure. Knowledge error arises when the model applies incorrect formulas, theorems, or properties. Calculation error refers to mistakes in numerical calculations. And logical reasoning error encompasses flaws in the deductive process, including invalid causal relationships, over-skipping of reasoning steps, taking groundless assumptions as fact, etc.

Evaluator Model	Answer Acc.	Step Acc.
o4-mini	99.6%	97.2%
o4-mini (w/o Reference)	98.5%	94.9%
Gemini-2.5-Pro	99.2%	94.1%
Gemini-2.5-Pro (w/o Reference)	98.7%	91.7%
Claude-4.5	95.8%	89.7%
GPT-4o	93.8%	82.3%

Table 2: Performance of different evaluator models.

Institutions	MLLMs
Deepmind	Gemini-2.0-Flash-Thinking [†] (2024), Gemini-2.5-Pro [†] (2025)
OpenAI	GPT-4o (2024), GPT-4.1 (2025b), o1 [†] (2025c), o3-mini [†] (2025d), o3 [†] (2025a), o4-mini [†] (2025e)
Anthropic	Claude3.7 (2025a), Claude-4(2025b), Claude-4.5, Claude-4.5-Thinking [†] (2025c)
Alibaba	QvQ-72B [†] (2024), Qwen2.5-VL-7B, Qwen2.5-VL-72B (2025) Qwen3-VL-32B, Qwen3-VL-32B-Thinking [†] (2025)
Shanghai AI Lab	InternVL3-8B, InternVL3-78B (2025) InternVL3.5-8B, InternVL3.5-38B (2025c)
Zhipu	GLM-4.1V-9B-Base, GLM-4.1V-9B-Thinking [†] (2025)

Table 3: Evaluated models and corresponding institutions. MLLMs marked with [†] are thinking models.

4.5 Evaluation Framework Reliability

In our evaluation framework, after MLLM generates responses based on question and diagram, both its reasoning process and reference process from dataset are fed into the evaluator. Evaluator performs a relatively simple task where it scores each solution step by checking whether it contains errors (as defined in section 4.4), assigning 0 if an error is present and 1 otherwise. Notably, our evaluation framework does not employ a step-matching logic. The reference solution serves as an evaluation aid rather than the only correct solution path. Therefore, valid alternative approaches, such as adopting a coordinate geometry approach instead of the reference’s auxiliary line method or utilizing different yet effective auxiliary constructions, are not penalized. The application of key knowledge, computations, and handling of geometric primitives in reference solutions helps reduce error misjudgments, thereby enhancing evaluator’s reliability.

From solutions generated by o3, Qwen3-VL-32B-Thinking, and GPT-4.1, we randomly sampled 1000 instances and evaluated them using four advanced models. We conducted a manual review to verify the accuracy of these evaluators on both answer and step levels, as summarized in Table 2. The results indicate that o4-mini and Gemini-2.5-Pro achieve substantially high evaluation accuracy when aided by reference solutions, validating the

Model	Dataset	1 – 4 Steps		5 – 8 Steps		9 – 12 Steps		13 – 24 Steps		Step-wise	Overall AVG		
		ACS	PCS	ACS	PCS	ACS	PCS	ACS	PCS	Δ PCS(%)	ACS	PCS	PQS
Thinking MLLMs													
QvQ-72B	all	69.6	22.4	52.7	6.6	27.4	1.7	14.0	1.2	94.6	52.1	11.4	21.0
o3-mini	all	60.7	21.8	54.7	13.0	39.7	12.7	16.5	5.5	74.8	51.8	15.7	27.2
GLM-4.1V-9B-T*	all	91.1	71.0	78.6	44.3	68.3	31.7	35.7	10.7	84.9	75.8	42.8	57.5
o1	mini	86.3	64.5	80.9	57.3	82.9	61.0	42.9	35.7	44.7	79.7	58.8	68.6
Gemini-2.0-T*	all	89.7	72.2	81.6	53.4	64.7	34.9	40.2	17.7	75.5	78.7	54.9	72.9
Claude-4.5-T*	all	89.3	69.8	84.6	52.5	71.2	46.2	51.2	33.5	52.0	81.7	56.5	73.2
o4-mini	all	94.5	78.3	94.0	70.3	91.4	70.2	81.1	63.4	19.0	92.8	72.9	81.1
Qwen3-VL-32B-T*	all	94.9	86.2	92.8	74.4	84.9	66.4	64.0	48.8	43.4	90.3	75.7	84.2
o3	mini	94.4	83.9	93.9	80.9	92.7	73.2	78.6	53.6	36.1	92.4	78.5	86.0
Gemini-2.5-Pro	all	95.3	85.9	92.2	76.2	88.0	76.0	71.3	50.0	41.8	91.2	77.8	88.6
Non-Thinking MLLMs													
Qwen2.5-VL-7B	all	38.7	8.5	23.6	1.2	12.3	0.0	8.5	0.0	100.0	26.4	3.7	11.1
GPT-4o	all	57.7	14.7	49.4	2.8	28.1	0.3	10.4	0.6	95.9	46.1	6.7	20.4
InternVL3-8B	all	72.8	30.2	50.2	7.7	25.0	0.6	10.4	0.3	99.0	51.6	14.5	24.9
GLM-4.1V-9B	all	69.2	31.8	43.4	8.1	20.2	1.0	9.1	0.0	100.0	46.8	15.4	26.1
InternVL3.5-8B	all	76.2	33.0	51.5	8.2	27.7	0.7	9.8	0.0	100.0	54.0	15.8	27.7
InternVL3-78B	all	78.9	37.6	60.2	11.0	34.2	4.1	14.0	0.6	98.4	59.6	19.1	34.3
InternVL3.5-38B	all	79.4	40.7	56.7	12.0	27.1	0.7	10.4	0.0	100.0	57.0	20.2	34.7
Claude-3.7	all	68.5	21.0	55.4	6.6	30.1	2.1	14.0	1.8	91.4	53.1	10.9	35.4
Claude-4.5	all	70.1	21.0	61.2	10.8	40.4	6.5	17.1	3.0	85.7	57.6	13.3	36.4
Claude-4	all	67.9	22.5	60.6	10.7	39.7	6.2	17.7	2.4	89.3	56.6	13.8	36.6
GPT-4.1	all	70.2	22.2	61.0	9.9	40.1	5.8	18.9	4.3	80.6	57.8	13.4	36.3
Qwen2.5-VL-72B	all	77.5	39.5	59.5	14.1	30.5	0.7	16.5	1.2	97.0	58.4	20.6	37.3
Qwen3-VL-32B	all	92.0	69.9	84.9	48.5	74.3	31.8	55.5	20.1	71.2	83.7	51.9	73.6

Table 4: Model’s performance on GeoLaux. ACS = Answer Correctness Score, PCS = Process Correctness Score, PQS = Process Quality Score. Δ PCS = $(PCS_{1-4Steps} - PCS_{13-24Steps}) / PCS_{1-4Steps}$, measures performance drop as steps increase. The “T*” in the table is an abbreviation for “Thinking” in the full model name.

reliability of our evaluation framework. Prior research (Zhang et al., 2024a,b; Wang et al., 2025a) has also demonstrated the evaluator promising potential of advanced MLLMs. Comprehensive cases demonstrating the flexibility with alternative approaches and the specific performance of the four evaluators are detailed in Appendix D.

5 Experiments

Table 3 shows our evaluation covers 23 state-of-the-art MLLMs, with 10 thinking models and 13 non-thinking models. Among these, the 11 open-source models are executed on NVIDIA A100 GPUs. All models generate answers through one-shot method. o4-mini is selected as the final evaluator.

Due to prohibitive computational costs of o1 and o3 on the full dataset, we constructed GeoLaux-mini comprising 330 problems uniformly sampled from the original 2186 questions. This subset preserves the original distributions of both step lengths and auxiliary lines to ensure equitable assessment conditions, details in Appendix A.

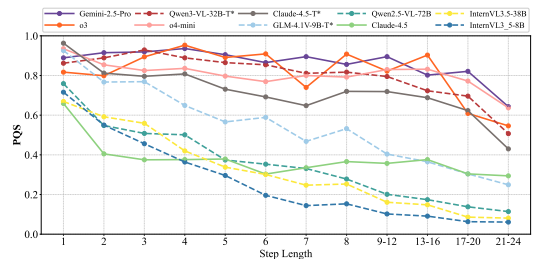


Figure 6: PQS across different step length problems.

5.1 Main Results

Table 4 presents the performance of 23 MLLMs on GeoLaux across different problem categories (short-step, medium-step, long-step, and ultra-long-step problems), ranked overall by PQS which represents reasoning capability. All models exhibit a certain gap between ACS and PCS, indicating the presence of False Positives and underscoring the unreliability of relying solely on answer correctness while ignoring the solution process. Therefore, PCS is adopted as the true measure of correctness.

Models’ Ranking. As demonstrated, thinking models significantly outperform non-thinking ones.

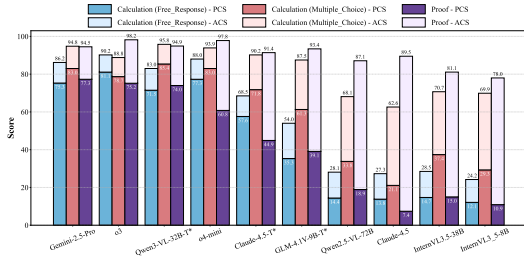


Figure 7: Comparison of ACS and PCS across free-response, multiple-choice, and proof problems.

Gemini-2.5-Pro, o3, and Qwen3VL-32B-Thinking achieve top three PQS, demonstrating superior reasoning capability. Meanwhile, o3 and o4-mini claim the top spots for PCS and ACS respectively. Among non-thinking models, Qwen3-VL-32B performs the best, leading across all metrics within its category, though a gap remains compared to the strongest thinking models. Notably, while current models achieve high ACS, the highest PCS on ultra-long-step problems is only 63.4 (o4-mini), highlighting the necessity of long-step datasets.

5.2 Long Step Analysis

" Δ PCS" of Table 4 measures performance degradation of models from steps 1-4 to steps 13-24, revealing 18 models exhibit a performance drop exceeding 50%, with some declines surpassing 90% or even reaching 100%. Throughout this transition, o4-mini demonstrates the most stable performance, despite still having a PCS decrease of 19.0%.

As problem step length increases, we also observe a universal decline in the reasoning quality of all models, as shown by the clear downward trend in PQS (Figure 6). This trend reflects the increasing difficulty of sustaining precise deduction as inference depth grows. Meanwhile, most thinking models achieve far higher PQS than non-thinking models, indicating non-thinking models not only make more errors in long-step problems but also tend to make errors at earlier steps of reasoning—since earlier steps carry greater weight in the PQS formula (Equation 4). This exposes their relatively weak capability in long-step reasoning.

5.3 False Positive Analysis

In our work, False Positives refer to cases where answer is correct but process contains errors, leading to discrepancies between ACS and PCS. We analyzed the distribution of False Positives across problem types. As shown in Figure 7, this issue is far more severe in multiple-choice and proof prob-

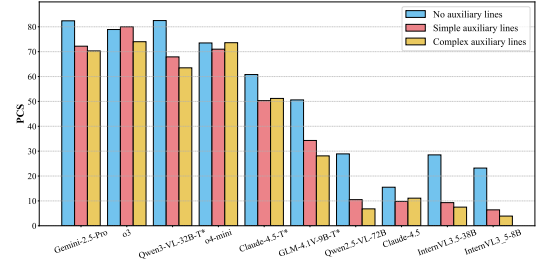


Figure 8: PCS under different auxiliary line complexity.

Model	Construction Rate(%)	Construction PCS
Claude-4.5-T*	6.4 / 56.4	33.3 / 44.3
o3	10.0 / 41.4	57.1 / 72.4
Gemini-2.5-Pro	28.6 / 83.6	72.5 / 82.1
GPT-4.1	15.0 / 17.9	4.7 / 8.0
Qwen3-VL-32B	38.6 / 56.4	35.6 / 44.3

Table 5: Comparison of self-attempt and heuristic prompting across construction rate and PCS.

lems than in free-response problems, attributed to chance guessing in multiple-choice tasks and the direct provision of target conclusions in proof tasks.

Limited hints can improve model performance, but clear answers make them lazy. Figure 7 also reveals an intriguing phenomenon: nearly all models' PCS exhibit a substantial increase for multiple-choice questions, yet suffer a precipitate decline in proof problems. This suggests that options may act as an incentive or a guide, assisting models in solving problems diligently to reach correct answers. However, when presented with correct answers directly, models become less rigorous about the solving process and resort to shortcuts (as cases in Appendix D). They pair correct final answers with flawed reasoning, essentially "deceiving" the users. This phenomenon warrants careful attention in future research.

5.4 Auxiliary Line Analysis

As shown in Figure 8, problems requiring auxiliary lines (especially complex ones) significantly challenge MLLMs, causing a noticeable drop in PCS compared to those without. o3 and o4-mini do not show a significant decline, because they often resort to coordinate-based methods when facing problems requiring auxiliary lines. Unfortunately, this approach escalates computational costs and solution complexity, undermining human interpretability.

Following Section 4.3, we compare the auxiliary line construction rates and the resulting PCS between autonomous solving and heuristic prompting,

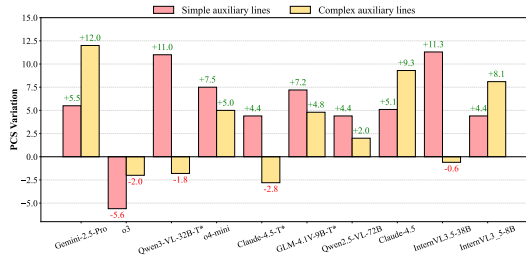


Figure 9: PCS delta after prompting auxiliary lines.

as summarized in Table 5. For both "Construction Rate (%)" and "Construction PCS", values to the left and right of "/" correspond to autonomous solving and heuristic auxiliary prompting, respectively.

This performance comparison reveals two key limitations: **(1) Weak awareness of autonomously constructing auxiliary lines:** The rate of MLLMs autonomously constructing auxiliary lines is extremely low. This sharply contrasts with the high construction rate when heuristically prompted with the correct lines, indicating that models often simply fail to realize they should try to construct an auxiliary line. **(2) Limited capability in generating correct auxiliary lines:** For problems where models proactively constructed auxiliary lines, their resulting PCS remains low and is significantly outpaced by the PCS achieved when using prompted standard auxiliary lines. This demonstrates that even when models do attempt to draw auxiliary lines, they frequently struggle to generate the most appropriate and mathematically correct auxiliary constructions to facilitate the solution.

Interestingly, Table 5 reveals that Non-Thinking MLLMs have a generally higher autonomously construction rate compared to Thinking models, but achieve a lower PCS. This divergence stems from differing reasoning strategies. Thinking models (e.g., Gemini-2.5-Pro) evaluate multiple paths and avoid unnecessary constructions that needlessly expand the search space if existing relationships suffice. Conversely, non-thinking models like Qwen3-VL-32B-Instruct tend to rely on mechanical rule application. They exhibit a heuristic tendency to attempt, frequently and blindly applying auxiliary lines to familiar geometric setups without verifying their strategic utility. Consequently, Qwen3-VL-32B-Instruct records the highest proactive construction rate but a substantially lower PCS.

While Table 5 evaluates PCS solely on problems where auxiliary lines are constructed, Figure 9 assesses PCS variations across all prompted in-

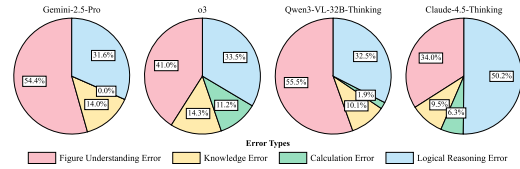


Figure 10: Error types distribution for MLLMs.

stances, irrespective of adoption. Although heuristic prompting generally improves PCS across most models, it reveals critical edge cases. For instance, o3 exhibits a performance drop because when suggested auxiliary lines are not adopted, the explicit hints likely disrupt its inherent coordinate-based strategies. Similarly, a few models benefiting from simple hints falter with complex ones, as they fail to comprehend complex constructions and are overwhelmed by the increased visual complexity. These negative impacts prove that relying solely on external auxiliary lines prompts is insufficient. Therefore, **advancing models' capabilities to autonomously construct optimal auxiliary lines** must be a primary focus for future research.

5.5 Error Type Analysis

We analyze the error causes according to the error types defined in section 4.4, and present representative error type distributions in Figure 10. The results demonstrate that incorrect geometric figure comprehension and flawed logical reasoning remain the fundamental bottlenecks limiting MLLMs' GPS capabilities. While knowledge errors and calculation errors persist across most models' solutions, these two error types prove relatively more addressable through external tools. Notably, Gemini-2.5-Pro exhibits virtually no calculation error during problem-solving, which likely contributes to its top performance on our benchmark. Appendix E contains some error analysis cases.

6 Conclusion

In this work, we present GeoLaux, a comprehensive geometric dataset with long-step problems and auxiliary line annotations. Based on this dataset, we evaluate 23 leading MLLMs through a five-dimensional framework, revealing: (1) a severe performance drop in long-step reasoning, (2) limited hints can improve model performance, and (3) the pivotal role of auxiliary line construction in GPS. These insights offer valuable guidance for advancing future MLLMs' geometric reasoning.

Acknowledgements

This work was supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM116), National Natural Science Foundation of China (No. 62137002, 62293550, 62293553, 62293554, 62450005, 62437002, 62477036, 62477037, 62577043, 62192781), the Shaanxi Provincial Social Science Foundation Project (No. 2024P041), the Youth Innovation Team of Shaanxi Universities "Multi-modal Data Mining and Fusion".

Limitations

While GeoLaux provides a robust benchmark for evaluating geometric reasoning, three key limitations warrant discussion. First, our dataset exclusively features problems at the Chinese High School Entrance Examination (Zhongkao) level, as we prioritize investigating foundational reasoning flaws relevant to real-world education. Although GeoLaux already poses a substantial challenge to existing models, incorporating higher-difficulty, Olympiad-level problems remains a direction for future research as models advance. Second, our evaluation framework relies on MLLM evaluator. While o4-mini demonstrates superior accuracy, employing it imposes substantial computational and financial costs when evaluating the full set of 2186 problems. This may constrain the feasibility of frequent, large-scale iterative testing. Finally, our current method for analyzing the impact of auxiliary lines involves heuristically prompting models with standard, ground-truth auxiliary lines. This approach simplifies the process and differs from realistic scenarios where models must autonomously construct auxiliary lines—a complex capability that remains distinct from merely utilizing provided hints. In future work, we will focus on optimizing evaluation efficiency and investigating the autonomous generation of geometric constructions to better reflect real-world problem-solving scenarios.

Ethical Considerations

In constructing GeoLaux, we strictly adhered to ethical guidelines and best practices. To ensure high-quality annotation, we employed two Master’s students, providing them with comprehensive operational guidelines and offering fair compensation for their contributions. The dataset sources

problems exclusively from publicly available examination materials, which subsequently underwent rigorous cleaning, deduplication, and standardization to guarantee data reliability while minimizing potential biases. Furthermore, we confirm that the dataset does not contain any personally identifiable information (PII) or private content. We are also committed to preventing data contamination; as detailed in Appendix F, we conducted extensive decontamination checks to ensure the validity of our evaluation. During writing, AI assistants were utilized solely for linguistic polishing to enhance readability. To promote open research while respecting copyright, we release the dataset and associated scripts under MIT and CC BY-NC-SA 4.0 licenses, strictly prohibiting commercial use.

References

- Avinash Anand, Raj Jaiswal, Abhishek Dharmadhikari, Atharva Marathe, Harsh Papat, Harshil Mital, Ashwin R Nair, Kritarth Prasad, Sidharth Kumar, Astha Verma, and 1 others. 2024a. Geovqa: A comprehensive multimodal geometry dataset for secondary education. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 102–108. IEEE.
- Avinash Anand, Raj Jaiswal, Abhishek Dharmadhikari, Atharva Marathe, Harsh Parimal Papat, Harshil Mital, Kritarth Prasad, Rajiv Ratn Shah, and Roger Zimmermann. 2024b. Improving multimodal llms ability in geometry problem solving, reasoning, and multistep scoring. *arXiv preprint arXiv:2412.00846*.
- Anthropic. 2025a. <https://www.anthropic.com/news/claude-3-7-sonnet>.
- Anthropic. 2025b. <https://www.anthropic.com/news/claude-4>.
- Anthropic. 2025c. <https://www.anthropic.com/news/claude-sonnet-4-5>.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 3313–3323.
- Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523.

- Jo-Ku Cheng, Zeren Zhang, Ran Chen, Jingyang Deng, Ziran Qin, and Jinwen Ma. 2025. Geouni: A unified model for generating geometry diagrams, problems and problem solutions. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3057–3066.
- Google Deepmind. 2024. <https://ai.google.dev/gemini-api/docs/thinking-mode>.
- Google Deepmind. 2025. <https://deepmind.google/models/gemini/pro/>.
- Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. *Palm-e: An embodied multimodal language model*. In *International Conference on Machine Learning*.
- Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, and 1 others. 2024. Omnimath: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*.
- Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Jingjing Jiang, Chao Ma, Xurui Song, Hanwang Zhang, and Jun Luo. 2025. Corvid: Improving multimodal large language models towards chain-of-thought reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3034–3046.
- Bert Jonsson, Julia Mossegård, Johan Lithner, and Linnea Karlsson Wirebring. 2022. Creative mathematical reasoning: Does need for cognition matter? *Frontiers in Psychology*, 12:797807.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- 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 mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv: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 (Volume 1: Long Papers)*, pages 6774–6786.
- Vesna Marinković. 2017. Argotrics—automated triangle construction solver. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(2):247–271.
- OpenAI. 2025a. <https://openai.com/index/introducing-o3-and-o4-mini/>.
- OpenAI. 2025b. <https://openai.com/index/gpt-4-1/>.
- OpenAI. 2025c. <https://openai.com/o1/>.
- OpenAI. 2025d. <https://openai.com/index/openai-o3-mini/>.
- OpenAI. 2025e. <https://openai.com/index/introducing-o3-and-o4-mini/>.
- Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, and 1 others. 2025. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20023–20070.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Qwen Team. 2024. Qvq: To see the world with wisdom. *Accessed on May, 5(2025):2*.
- 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.
- Peijie Wang, Zhong-Zhi Li, Fei Yin, Dekang Ran, and Cheng-Lin Liu. 2025a. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19541–19551.
- Peijie Wang, Chao Yang, Zhong-Zhi Li, Fei Yin, Dekang Ran, Mi Tian, Zhilong Ji, Jinfeng Bai, and Cheng-Lin Liu. 2025b. Solidgeo: Measuring multimodal spatial math reasoning in solid geometry. *arXiv preprint arXiv:2505.21177*.

- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025c. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.
- Jingxuan Wei, Caijun Jia, Qi Chen, Honghao He, Linzhuang Sun, Conghui He, Lijun Wu, Bihui Yu, and Cheng Tan. 2025. Geoint-r1: Formalizing multimodal geometric reasoning with dynamic auxiliary constructions. *arXiv preprint arXiv:2508.03173*.
- Wenjun Wu, Lingling Zhang, Jun Liu, Xi Tang, Yaxian Wang, Shaowei Wang, and Qianying Wang. 2024. E-gps: Explainable geometry problem solving via top-down solver and bottom-up generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13828–13837.
- Fangzhi Xu, Hang Yan, Qiushi Sun, Jinyang Wu, Zixian Huang, Muye Huang, Jingyang Gong, Zichen Ding, Kanzhi Cheng, Yian Wang, and 1 others. 2026. Odyssearena: Benchmarking large language models for long-horizon, active and inductive interactions. *arXiv preprint arXiv:2602.05843*.
- Liangyu Xu, Yingxiu Zhao, Jingyun Wang, Yingyao Wang, Bu Pi, Chen Wang, Mingliang Zhang, Jihao Gu, Xiang Li, Xiaoyong Zhu, and 1 others. 2025. Geosense: Evaluating identification and application of geometric principles in multimodal reasoning. *arXiv preprint arXiv:2504.12597*.
- Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.
- Hang Yan, Xinyu Che, Fangzhi Xu, Qiushi Sun, Zichen Ding, Kanzhi Cheng, Jian Zhang, Tao Qin, Jun Liu, and Qika Lin. 2026. Tide: Trajectory-based diagnostic evaluation of test-time improvement in llm agents. *arXiv preprint arXiv:2602.02196*.
- Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2025. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11798–11827.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.
- Jian Zhang, Zhangqi Wang, Haiping Zhu, Kangda Cheng, Kai He, Bo Li, Qika Lin, Jun Liu, and Erik Cambria. 2026a. Mars: Multi-agent adaptive reasoning with socratic guidance for automated prompt optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(19):16307–16315.
- Jian Zhang, Zhiyuan Wang, Zhangqi Wang, Fangzhi Xu, Qika Lin, Lingling Zhang, Rui Mao, Erik Cambria, and Jun Liu. 2026b. Maps: Multi-agent personality shaping for collaborative reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(19):16316–16324.
- Jiaxin Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Fei Yin, Cheng-Lin Liu, and Yashar Moshfeghi. 2024a. Geoeval: benchmark for evaluating llms and multimodal models on geometry problem-solving. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1258–1276.
- Kaiyan Zhang, Jiayuan Zhang, Haoxin Li, Xuekai Zhu, Ermo Hua, Xingtai Lv, Ning Ding, Biqing Qi, and Bowen Zhou. 2025. Openprm: Building open-domain process-based reward models with preference trees. In *The thirteenth international conference on learning representations*.
- Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. 2023. A multi-modal neural geometric solver with textual clauses parsed from diagram. *arXiv preprint arXiv:2302.11097*.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024b. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.
- Yushun Zhang, Weiping Fu, Zesheng Yang, Bo Zhao, Lingling Zhang, Jian Zhang, Yumeng Fu, Jiaying Huang, and Jun Liu. 2026c. Geochallenge: A multi-answer multiple-choice benchmark for geometric reasoning with diagrams. *arXiv preprint arXiv:2603.19252*.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Appendix Overview

- Section A: GeoLaux Details.
- Section B: PQS discussions.
- Section C: Prompts and Model Details.
- Section D: Process Evaluation Cases.
- Section E: Error Type Cases.
- Section F: Data Leakage Analysis.

A GeoLaux Details

A.1 GeoLaux-mini Details

We performed uniform sampling on GeoLaux to create GeoLaux-mini, a 330-problem subset specifically designed for testing computationally expensive models (o1, o3) and conducting supplementary auxiliary line heuristic evaluations.

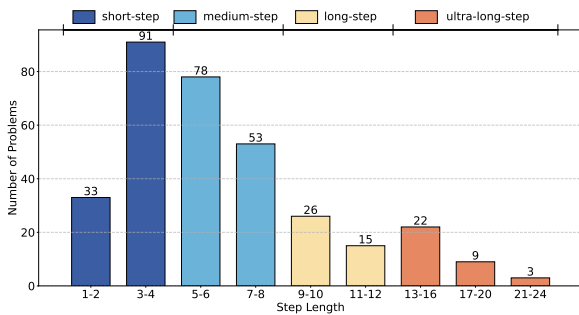


Figure 11: Problem quantity statistics across step lengths in GeoLaux-mini.

GeoLaux-mini maintains a similar step-length distribution to the original dataset (as Figure 11), containing a substantial number of medium-step, long-step, and ultra-long-step problems. The subset comprises 109 proof problems and 221 calculation problems, 190 problems that do not require auxiliary line and 140 problems that need auxiliary line. Its auxiliary line distribution illustrated in Figure 12.

A.2 Language robustness analysis

Due to the data source, the original language of the dataset is Chinese. However, the leading models we tested are inherently multilingual, having been pre-trained on massive Chinese corpora to achieve native-level reading comprehension. Additionally, middle school geometry problems feature highly concise text, standardized vocabulary, and formulaic syntactic structures (e.g., "As shown in the figure, given A, prove B"). As a result, the linguistic

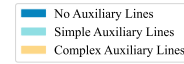


Figure 12: Distribution of auxiliary line types in GeoLaux-mini.

processing barrier is exceptionally low, minimizing the risk of tokenization artifacts. Therefore, the simple terminology used in these problems does not pose a meaningful challenge to the models' comprehension.

To further validate the linguistic robustness of our dataset, we translated the 330 problems from GeoLaux-mini into English (the translated dataset is also open-sourced on GitHub) and evaluated five representative large models. The performance of these models on GeoLaux-mini before and after translation is presented in Table 6.

As observed, the performance variations among different models before and after translation are inconsistent—some increase while others decrease—and the magnitude of these changes is consistently small, representing normal random fluctuation. This demonstrates that our experimental results are fundamentally unaffected by the language used, confirming their robustness.

A.3 GeoLaux Examples

The GeoLaux dataset encompasses a comprehensive collection of geometry problems that can be classified along three key dimensions: (1) by the presence of solvable answers in the questions, differentiating between calculation problems and proof problems; (2) by the necessity of auxiliary construction, distinguishing problems requiring auxiliary lines from those needing none; and (3) by solution step length, categorizing problems as short-step, medium-step, long-step, or ultra-long-step problems. Representative examples are illustrated in Figure 15.

B PQS discussions

PQS score comprises two components: the step weight function and the score activation function.

Model	1-4 Steps PCS	5-8 Steps PCS	9-12 Steps PCS	13-24 Steps PCS	AVG ACS	AVG PCS	AVG PQS
Claude-4.5-T*	71.0 / 68.1 (-2.9)	49.8 / 47.3 (-2.5)	46.5 / 48.0 (+1.5)	32.1 / 29.1 (-3.0)	81.2 / 81.2 (+0.0)	58.6 / 56.1 (-2.5)	74.7 / 73.8 (-0.9)
o3	83.9 / 83.1 (-0.8)	80.9 / 84.0 (+3.1)	73.2 / 75.9 (+2.7)	53.6 / 57.1 (+3.5)	92.4 / 89.1 (-3.3)	78.5 / 77.8 (-0.7)	86.0 / 84.9 (-1.1)
Gemini-2.5-Pro	85.5 / 87.1 (-1.6)	80.4 / 79.4 (-1.0)	73.0 / 75.7 (+2.7)	46.3 / 50.0 (+3.7)	89.4 / 85.8 (-3.6)	77.6 / 78.8 (+1.2)	88.8 / 89.9 (+1.1)
GPT-4.1	25.8 / 26.2 (+0.4)	9.2 / 12.5 (+3.3)	9.8 / 7.3 (-2.5)	5.3 / 7.1 (+1.8)	55.8 / 59.7 (+3.9)	14.5 / 16.4 (+1.9)	35.1 / 37.3 (+2.2)
Qwen3-VL-32B	77.4 / 74.2 (-3.2)	53.4 / 48.9 (-4.5)	35.3 / 39.9 (+4.6)	23.9 / 25.7 (+1.8)	85.2 / 86.1 (+0.9)	55.8 / 56.4 (+0.6)	73.3 / 74.2 (+0.9)

Table 6: Model performance on the GeoLaux dataset in Chinese (left of /) and English (right of /). Values in "()" represent the change in scores when transitioning from Chinese to English problems.

For an MLLM-generated solution process, evaluator scores each individual reasoning step, assigning 1 for correct steps and 0 for incorrect ones. Step weight function assigns specific weights to each step's score, then score activation function ultimately computing a weighted overall process quality score. These two functions are discussed separately below.

B.1 Step Weight Function

The design rationale for the step weight function is guided by the following principles:

(1) Decreasing function. In long-step problems, the later an error occurs, the stronger we consider the model's long-step reasoning capability to be. Consequently, models failing in the initial steps should incur heavier penalties, resulting in lower final scores. This logic implies that higher weights should be assigned to earlier steps, suggesting that the step weight function should be monotonically decreasing.

We propose several decreasing weight functions; their functional trends and the specific weights assigned to each step (for $n = 6$) are illustrated in Figure 13 (a). The functions shown in the figure are: the blue Function A:

$$y = \cos \frac{\pi i}{4n};$$

the red Function B:

$$y = e^{-\frac{i}{n}};$$

the purple Function C:

$$y = -\frac{i}{n(n-1)} + \frac{1}{2n} + \frac{1}{n-1};$$

and the green Function D:

$$y = \frac{1 + \frac{1}{n}}{i(i+1)}.$$

All are monotonically decreasing and together cover various types of functions—linear, concave, and convex.

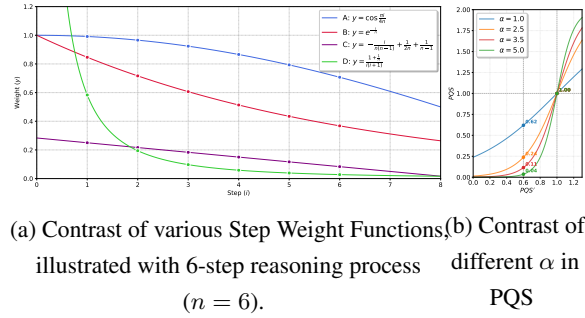


Figure 13: Diagram of the functional analysis in PQS.

(2) Convex function: The importance gap is larger for earlier steps and smaller for later ones. For example, if two models make mistakes at step 2 and step 4 respectively, it indicates that the difference in their capabilities is significantly greater compared to two models making errors at step 12 and step 14. Therefore, the weight function should be a convex function, meaning that the rate of decrease is faster in the early stage than in the later stage.

Based on this consideration, blue function A (concave function) and purple function C (linear function) are excluded.

(3) Moderate decreasing rate: The weighting function should not decrease too rapidly. For long-step problems, the performance in later steps remains critical for evaluation and should retain significant weight.

The green function D is a typical example of an excessively rapid decline. As can be seen, the weights at steps 5 and 6 are extremely small compared to weights at steps 1 and 2. Additionally, the difference between the weights at step 5 and step 6 is minimal, making it unsuitable as a weight function. On the other hand, the red function B, which corresponds to Equation 4 in the main text, perfectly satisfies all the requirements we have proposed and is therefore selected as the step weight function.

B.2 Score Activation Function

By multiplying the score assigned by the evaluator to each step with the corresponding weight of that step, summing up the results, and then normalizing by dividing by the sum of the weights across all steps, the initial process quality score can be obtained. The formula is as follows (which is Equation 5 in the main text):

$$\text{PQS}' = \frac{\sum_{i=1}^n \eta_i \cdot y_i}{\sum_{j=0}^n y_j} = \frac{\sum_{i=1}^n \eta_i \cdot e^{-\frac{i}{n}}}{\sum_{j=1}^n e^{-\frac{j}{n}}}.$$

For example, if a solution has 3 steps with scores [1,0,1], the initial process score is

$$\begin{aligned} \text{PQS}'_{\text{example}} &= \frac{1 \cdot e^{-\frac{1}{3}} + 0 \cdot e^{-\frac{2}{3}} + 1 \cdot e^{-\frac{3}{3}}}{e^{-\frac{1}{3}} + e^{-\frac{2}{3}} + e^{-\frac{3}{3}}} \\ &\approx 0.6832 \end{aligned}$$

However, we observed that since model solutions always contain some correct steps, PQS' consistently falls between 0.6 and 1, failing to highlight the differences in models' reasoning capabilities. To address this, we applied the \tanh activation function to PQS' .

$$\text{PQS} = \tanh(\alpha(\text{PQS}' - 1)) + 1,$$

Using the \tanh function for activation, unlike a linear function, can relatively amplify the differences between higher-scoring models to a greater extent, facilitating the selection of the best-performing model.

α is a hyperparameter, and the impact of its different values on the mapping of PQS is illustrated in Figure 13 (b) (focus on the PQS value range of 0.6–1). When α is set to 1, the mapped PQS values still fall within [0.61,1], failing to amplify the differences between models. When α is set to 2.5, the mapped range remains relatively narrow [0.24,1]. On the other hand, when α is set to 5, although it maps PQS' to a broader range, it significantly reduces the distinctions among lower-performing models, which is detrimental to overall ranking. Therefore, after balancing the objectives of selecting the optimal model and maintaining a comprehensive ranking of all 23 models, we chose $\alpha = 3.5$ for subsequent experiments. At this value, PQS is mapped to the interval [0.11,1].

Therefore, for the solution with 3 steps scored as [1,0,1], the final activated PQS value is

$$\begin{aligned} \text{PQS}_{\text{example}} &\approx \tanh(3.5 \cdot (0.6832 - 1)) + 1 \\ &\approx 0.3399 \end{aligned}$$

C Prompts and Model Details

C.1 Prompt for Initial Solution Generation.

In the main experimental section, we employ one-shot prompt to guide MLLMs in generating responses in JSON format. The use of one-shot prompt ensure all models strictly adhere to our specified JSON format, thereby simultaneously obtaining both the step-by-step solution process (to facilitate subsequent evaluation) and numerical answers for calculation problems. Sample prompts for calculation problems and proof problems are shown in the Figure 16.

C.2 Prompt for Auxiliary Line Heuristic Solution Generation.

In the auxiliary line heuristic experiment, we provide the LLM with both the auxiliary line construction method from the reference solution and the corresponding diagram showing this auxiliary line. The model is then prompted to analyze why this particular auxiliary line was suggested and determine whether to incorporate it into its own solution approach. The specific prompting methodology is illustrated in the accompanying Figure 17.

C.3 Prompt for Solution Evaluation.

In our evaluation framework, we employ two distinct prompts to guide evaluators in assessing the generated solutions: one for step-by-step scoring and another for error type analysis, as illustrated in Figures 18 and 19 respectively. Both assessment components are conducted with reference to the standard solution provided in the reference answers, thereby enhancing the reliability of our evaluation.

C.4 Model Details.

For the nine closed-source models, we access them through API and perform inference using simple CPU computation. For the four open-source models, we conduct inference using a server equipped with two NVIDIA A100 GPUs. The detailed generation parameters are specified in Table 7.

D Process Evaluation Cases

D.1 Comparison of Different Evaluators

Figure 20 presents the process scores assigned by four models on the same problem, alongside scores given by human evaluators. It can be seen that o4-mini's scoring for model solutions aligns most closely with human judgment, followed by Gemini

Model	Hyperparameters
GPT-4o	model = gpt-4o-2024-08-06, temperature = 0.1, max_tokens = 4096
GPT-4.1	model = gpt-4.1-2025-04-14, temperature = 0.1, max_tokens = 4096
Claude-3.7	model = claude-3-7-sonnet-20250219, temperature = 0.1, max_tokens = 4096
Claude-4.0	model = claude-sonnet-4-20250514, temperature = 0.1, max_tokens = 4096
Claude-4.5	model = claude-sonnet-4-5-20250929, temperature = 0.1, max_tokens = 4096
Claude-4.5-Thinking	model = claude-sonnet-4-5-20250929-thinking, temperature = 0.1, max_tokens = 8192
Gemini-2.0-Thinking	model = gemini-2.0-flash-thinking-exp-01-21, temperature = 0.1, max_tokens = 8192
Gemini-2.5-Pro	model = gemini-2.5-pro-preview-03-25, temperature = 0.1, max_tokens = 10288
o1	model = o1, temperature = 0.1, max_tokens = 8192
o3	model = o3, temperature = 0.1, max_tokens = 10288
o3-mini	model = o3-mini-all, temperature = 0.1, max_tokens = 8192
o4-mini	model = o4-mini-2025-04-16, temperature = 0.1, max_tokens = 8192
Qwen2.5-VL-72B	model = Qwen/Qwen2.5-VL-72B-Instruct, temperature = 0.1, max_tokens = 10288
Qwen2.5-VL-7B	model = Qwen/Qwen2.5-VL-7B-Instruct, temperature = 0.1, max_tokens = 10288
QvQ-72B	model = Qwen/QvQ-72B-Preview, temperature = 0.1, max_tokens = 10288
Qwen3-VL-32B	model = Qwen/Qwen3-VL-32B-Instruct, temperature = 0.1, max_tokens = 10288
Qwen3-VL-32B-Thinking	model = Qwen/Qwen3-VL-32B-Thinking, temperature = 0.1, max_tokens = 10288
GLM-4.1V-9B-Base	model = ZhipuAI/GLM-4.1V-9B-Base, temperature = 0.1, max_tokens = 10288
GLM-4.1V-9B-Thinking	model = ZhipuAI/GLM-4.1V-9B-Thinking, temperature = 0.1, max_tokens = 10288
InternVL3-78B	model = OpenGVLab/InternVL3-78B, temperature = 0.1, max_tokens = 4096
InternVL3-8B	model = OpenGVLab/InternVL3-8B, temperature = 0.1, max_tokens = 4096
InternVL3.5-38B	model = OpenGVLab/InternVL3_5-38B, temperature = 0.1, max_tokens = 4096
InternVL3.5-8B	model = OpenGVLab/InternVL3_5-8B, temperature = 0.1, max_tokens = 4096

Table 7: Model Hyperparameters

2.5 Pro. Notably, in some cases—such as the second example in the figure—for problems solved via coordinate system methods, human evaluators require complex computational verification, whereas the MLLM-based evaluator holds a distinct advantage in this regard.

D.2 Different solution path evaluation cases

Figure 21 presents an evaluation case of an alternative solution path. In this example, while the reference solution employs auxiliary lines, the o3 model utilizes a coordinate geometry approach. Even though the o3 solution path diverges entirely from the reference, the evaluator (o4-mini) independently verifies the mathematical correctness and logic of each step. Since no conceptual, computational, or logical errors are found within the coordinate geometry framework, the evaluator identifies all steps as correct. Consequently, the Process Correctness Score (PCS) for this problem is 1.

Regarding the auxiliary lines construction step, the evaluator judges whether this step conforms to geometric principles—distinguishing between valid, compliant constructions and impractical, invalid ones (e.g., "Construct $AD \perp AB$ with foot

D " would be considered invalid because the notation " $AD \perp AB$ " implies that the perpendicular intersection is at point " A ", which contradicts the definition of " D " as the foot of the perpendicular). The evaluator marks auxiliary line construction that is geometrically sound as "correct" and subsequently assesses the validity of the steps that follow. Figure 22 presents two evaluation cases of alternative auxiliary lines: one involving a valid geometric relationship and another featuring an erroneous construction that violates geometric principles. The evaluator correctly assigns scores of 1 and 0, respectively.

D.3 Proof Process Evaluation Cases

Figure 23 shows two proof problems in which models use intermediate steps with obvious errors to fit the final conclusion to be proved. It can be observed that the solution process is often absurd, committing egregious mistakes, revealing how models tend to cut corners when solving proof-based problems.

Model	1-4 Steps ACS	1-4 Steps PCS	5-8 Steps ACS	5-8 Steps PCS	Orig.Ans.Rate
Claude-4.5-T*	89.7 / 85.4 (-4.3)	68.6 / 70.1 (+1.5)	82.3 / 79.8 (-2.5)	55.8 / 54.7 (-1.1)	1.3%
o3	94.5 / 95.2 (+0.7)	84.1 / 85.5 (+1.4)	94.7 / 94.2 (-0.5)	80.2 / 75.8 (-4.4)	1.3%
Gemini-2.5-Pro	97.6 / 97.6 (+0.0)	88.1 / 89.9 (+1.8)	95.4 / 94.4 (-1.0)	82.1 / 82.1 (+0.0)	0%
GPT-4.1	66.7 / 64.3 (-2.4)	23.8 / 25.2 (+1.4)	31.6 / 36.8 (+5.2)	12.4 / 13.2 (+0.8)	2.5%
Qwen3-VL-32B	95.2 / 95.2 (+0.0)	75.7 / 73.8 (-1.9)	84.5 / 86.8(+2.3)	50.5 / 51.2 (+0.7)	0%

Table 8: Comparison of model performance before (left of /) and after (right of /) numerical perturbation. Values in "()" represent the change in performance when introducing numerical perturbations. "Orig. Ans. Rate (%)" denotes the probability of the model incorrectly outputting the exact answer from the original, unperturbed problem.

Model	Avg. ROUGE-L ($\times 10^{-2}$)
Gemini-2.5-Pro	0.01
Qwen3-VL-32B	0.02
Gemini-2.0-Thinking	0.08
o3	0.12
o4-mini	0.12
o1	0.13

Table 9: ROUGE-L comparison across different models.

E Error Type Cases

The error types of the problems we used include four categories: (a) Figure Understanding Error, (b) Knowledge Error, (c) Calculation Error, and (d) Logical Reasoning Error. Their meanings are as follows:

- Figure comprehension error:** Failure to correctly understand the geometric primitives (points, lines, circles, etc.) implied by the diagram, such as misidentifying angle relationships, collinear relationships, etc.
- Knowledge Error:** While correctly understanding the point/line relationships, the solution employs incorrect formulas. This includes: using wrong formulas/theorems/properties, or selecting inappropriate formulas/theorems/properties for the given problem.
- Calculation Error:** While correctly understanding the geometric relationships and properly selecting/applying the relevant knowledge, the solution contains numerical calculation mistakes or unit conversion errors.
- Logical Reasoning Error:** The reasoning process contains logical fallacies, including but not limited to: invalid causal relationships between premises and conclusions (the "because-therefore" connection is unjustified),

AI making intuitive assumptions without basis, drawing conclusions by introducing irrelevant external information or incorrect assumptions, nonsensical responses, logically chaotic arguments, or inexplicable answers.

A MLLM evaluator is employed with prompt in Appendix C.3 to determine these error types and their corresponding steps. Some Error Cases are illustrated in Figure 24.

F Data Leakage Analysis

F.1 Textual Overlap Assessment

To evaluate potential data leakage in models while accounting for Chinese language characteristics, we conducted a similarity analysis using a two-step approach. We randomly selected 150 problems from our dataset and generated 3 distinct prompts for each problem to query the models, resulting in a total of 450 test cases. First, we generated responses using a 10-gram matching process, where each Chinese character was treated as a token. Subsequently, we employed the Chinese-adapted ROUGE-L metric to measure the overlap between these model outputs and ground-truth solutions. We observed uniformly low similarity metrics across all models (Table 9), with average ROUGE-L scores ranging from 0.01×10^{-2} to 0.13×10^{-2} (0.01% to 0.13%). In particular, these same models exhibited poor matching accuracy in our contamination detection experiments, despite their strong general reasoning capabilities. Similar to the approach in previous works (Xu et al., 2024; Gao et al., 2024), our results suggest that these models have not been exposed to or trained on our dataset. The consistently low ROUGE-L scores, all averaging below 0.14×10^{-2} (0.14%), indicate minimal overlap between our dataset and the training corpora of the tested models. These scores fall well within the expected range for clean,

Question: As shown in the figure, diameter AB intersects chord CD at point E, and point F is the midpoint of CD. Extend FO to intersect $\odot O$ at point G. If $AE=5BE=5$ and $\angle AEC=45^\circ$, then the length of FG is _____.

Choices: A. 4, B. $3\sqrt{2}$, C. $3+\sqrt{2}$, D. $5-\sqrt{2}$ **Answer:** $3+\sqrt{2}$

Reference Solution:

Step 1: \because point F is the midpoint of CD, extend FO to intersect $\odot O$ at point G, \therefore $FG \perp CD$
Step 2: $\because AE = 5BE = 5$, $\therefore BE = 1$, $AB = AE + BE = 6$
Step 3: $\because OG = 1/2 AB = 3$, $OB = AO = 3$, $OE = OB - BE = 2$
Step 4: $\because \angle AEC = 45^\circ$, $\therefore OF = OE \cdot \sin \angle OEF = 2 \cdot \sin 45^\circ = \sqrt{2}$
Step 5: $\therefore FG = OG + OF = 3 + \sqrt{2}$

(a) Problem before perturbation

Question: As shown in the figure, diameter AB intersects chord CD at point E, and point F is the midpoint of CD. Extend FO to intersect $\odot O$ at point G. If $AE=5BE=10$ and $\angle AEC=45^\circ$, then the length of FG is _____.

Choices: A. 8, B. $6\sqrt{2}$, C. $6+2\sqrt{2}$, D. $10-2\sqrt{2}$ **Answer:** $6+2\sqrt{2}$

Reference Solution:

Step 1: \because point F is the midpoint of CD, extend FO to intersect $\odot O$ at point G, \therefore $FG \perp CD$
Step 2: $\because AE = 5BE = 10$, $\therefore BE = 2$, $AB = AE + BE = 12$
Step 3: $\because OG = 1/2 AB = 6$, $OB = AO = 6$, $OE = OB - BE = 4$
Step 4: $\because \angle AEC = 45^\circ$, $\therefore OF = OE \cdot \sin \angle OEF = 4 \cdot \sin 45^\circ = 2\sqrt{2}$
Step 5: $\therefore FG = OG + OF = 6 + 2\sqrt{2}$

(b) Problem after perturbation

Figure 14: Example of numerical perturbation.

unleaked data, further confirming the novelty and integrity of our dataset.


F.2 Numerical Perturbation Test

Beyond n-gram analysis, to more reliably and definitively verify whether models are genuinely reasoning rather than memorizing, we designed a numerical perturbation experiment.

We focused on 159 short- and medium-step calculation problems in the GeoLaux-mini set, as models generally perform better on these shorter problems, any reliance on memorization (contamination) would be much more obvious. Specifically, we utilized the advanced Gemini-3-Pro to apply random numerical perturbations to these 159 problems, followed by strict human verification. We ensured that only the specific numerical values are altered, while all geometric relationships and the core questions remained completely identical.

As shown in Table 8, the models' performance shows no significant change after introducing numerical perturbations, and they almost never output the original, pre-perturbation answers (any rare occurrences are likely due to general solving errors). This provides strong evidence that the models are engaging in genuine reasoning rather than simply outputting memorized data from pre-training. Furthermore, it underscores the integrity and practical significance of our dataset, confirming it is free from data leakage and ensures a fair evaluation. Figure 14 is an example before and after the disturbance. This perturbed dataset has also been made publicly available on GitHub.

Question: As shown in the figure, given that the side length of square ABCD is 4, points E and F lie on AD and DC respectively, with AE = DF = 1. BE and AF intersect at point G, and point H is the midpoint of BF. Connect GH. Find the length of GH.



Answer: 5/2 **Type:** Calculation **Step Length:** 6 [Original Image](#)

Reference Solution:

Step 1: Since quadrilateral ABCD is a square, $\angle BAE = \angle D = 90^\circ$, and $AB = AD$.

Step 2: In triangles $\triangle ABE$ and $\triangle DAF$, $\angle BAE = \angle D$, $AB = AD$, $AE = DF$, $\therefore \triangle ABE \cong \triangle DAF$ (by SAS congruence criterion), $\therefore \angle ABE = \angle DAF$.

Step 3: Since $\angle ABE + \angle BEA = 90^\circ$, $\therefore \angle DAF + \angle BEA = 90^\circ$, $\therefore \angle AGE = \angle BGF = 90^\circ$ (the angle at the intersection of BE and AF is a right angle).

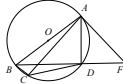
Step 4: Since point H is the midpoint of BF, and $\angle BGF = 90^\circ$, triangle BGF is a right triangle with hypotenuse BF, $\therefore GH = (1/2) \times BF$.

Step 5: Given $BC = 4$, and $CF = CD - DF = 4 - 1 = 3$, $\therefore BF = \sqrt{BC^2 + CF^2} = \sqrt{4^2 + 3^2} = \sqrt{16 + 9} = \sqrt{25} = 5$.

Step 6: Therefore, $GH = (1/2) \times BF = (1/2) \times 5 = 5/2$.

(a) Calculation problem

Question: As shown in the figure, quadrilateral ABCD is inscribed in circle O, with AB being the diameter of circle O. The diagonals AC and BD intersect at point E. The tangent line AF to circle O intersects the extension of BD at point F. If BD bisects $\angle ABC$, prove that AD bisects $\angle CAF$.



Answer: Null **Type:** Proof **Step Length:** 8 [Original Image](#)

Reference Solution:

Step 1: $\because AF$ is a tangent to $\odot O$, $\therefore BA \perp AF$, $\therefore \angle BAF = 90^\circ$, $\therefore \angle FAD + \angle BAD = 90^\circ$.

Step 2: $\because AB$ is the diameter of $\odot O$, $\therefore \angle ADB = 90^\circ$, $\therefore \angle ABD + \angle BAD = 90^\circ$.

Step 3: $\because BD$ bisects $\angle ABC$, $\therefore \angle ABD = \angle CBD$.

Step 4: Substituting $\angle ABD = \angle CBD$ into the equation from Step 2, $\angle CBD + \angle BAD = 90^\circ$.

Step 5: $\because \angle CBD$ and $\angle CAD$ subtend the same arc CD, $\therefore \angle CBD = \angle CAD$ (angles in the same segment).

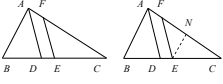
Step 6: Substituting $\angle CBD = \angle CAD$ into the equation from Step 4, we get $\angle CAD + \angle BAD = 90^\circ$.

Step 7: $\because \angle FAD + \angle BAD = 90^\circ$, and from Step 6, $\angle CAD + \angle BAD = 90^\circ$. Therefore, $\angle FAD = \angle CAD$.

Step 8: $\therefore AD$ bisects $\angle CAF$.

(b) Proof problem

Question: As shown in the figure, in triangle ABC, $AB = 5$, $AC = 9$, AD is the angle bisector of $\angle BAC$, point E is the midpoint of BC, and EF is parallel to AD. Find the length of AF.



Answer: 2 **Type:** Calculation **Step Length:** 8 [Original Image](#) [Auxiliary Image](#)

Auxiliary Text: Set point N as the midpoint of AC, connect EN.

Reference Solution:

Step 1: By the Midsegment Theorem, $EN \parallel AB$ and $EN = (1/2)AB = 2.5$.

Step 2: $\because EN \parallel AB$, $\therefore \angle CNE = \angle BAC = \angle BAD + \angle CAD$ (corresponding angles).

Step 3: $\because EF \parallel AD$, $\therefore \angle DAC = \angle EFN$ (corresponding angles). $\therefore \angle DAC = \angle EFN$.

Step 4: $\because AD$ is the angle bisector of $\angle BAC$, $\therefore \angle BAD = \angle CAD$, $\angle EFN = \angle BAD$.

Step 5: $\because \angle CNE$ is an exterior angle of $\triangle FEN$, $\therefore \angle CNE = \angle EFN + \angle FEN$.

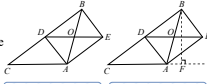
Step 6: $\because \angle EFN = \angle FEN$, $\therefore \triangle FEN$ is isosceles with $FN = EN = 2.5$.

Step 7: $\because N$ is the midpoint of AC, $\therefore NC = (1/2)AC = 4.5$. Then, $FC = FN + NC = 7$.

Step 8: $\therefore AF = AC - FC = 9 - 7 = 2$.

(c) Auxiliary lines construction problem

Question: As shown in the figure, in $\triangle ABC$, $AB = AC$, AD is the median to side BC. A line parallel to BC is drawn through point A, and a line parallel to AD is drawn through point B. These two lines intersect at point E. (2) Connect DE, intersecting AB at point O. If $AC = 10$ and $BE = 6$, find the value of $\sin \angle AOD$.



Answer: 24/25 **Type:** Calculation **Step Length:** 15 [Original Image](#) [Auxiliary Image](#)

Auxiliary Text: Make $BF \perp AC$ through point B and intersect the extension of AC at point F.

Reference Solution:

Step 1: Make the auxiliary line upon.

Step 2: $\because BA = AC$, and AD is the median to side BC, $\therefore AD \perp BD$.

Step 3: $\because AE \parallel BD$ and $BE \parallel AD$, \therefore Quadrilateral ADBE is a rectangle.

Step 4: $\because AC = AB = 10$, and $BE = 6$, $\therefore AE = BD = \sqrt{AB^2 - BE^2} = \sqrt{10^2 - 6^2} = 8$.

Step 5: $\because BC = 2BD = 16$.

Step 6: $\because S_{\triangle ABC} = 1/2 * BC * AD$, and $AD = BE = 6$ (since ADBE is a rectangle), $\therefore S_{\triangle ABC} = 1/2 * 16 * 6 = 48$.

Step 7: $\because S_{\triangle ABC} = 1/2 * AC * BF$, $\therefore 1/2 * 10 * BF = 48$.

Step 8: $\therefore BF = 48/5$.

Step 9: Quadrilateral ABDE is a rectangle, $\therefore OD = OB$.

Step 10: $\therefore \angle ODB = \angle OBD$.

Step 11: $\because AB = AC$, $\therefore \angle C = \angle ABC = \angle OBD$.

Step 12: $\therefore \angle C = \angle ODB$.

Step 13: $\therefore DE \parallel AC$.

Step 14: $\therefore \angle AOD = \angle BAF$.

Step 15: $\therefore \sin \angle AOD = \sin \angle BAF = BF/AB = (48/5) / 10 = 48/50 = 24/25$.

(d) Ultra-long step problem

Figure 15: Examples from the GeoLaux dataset.

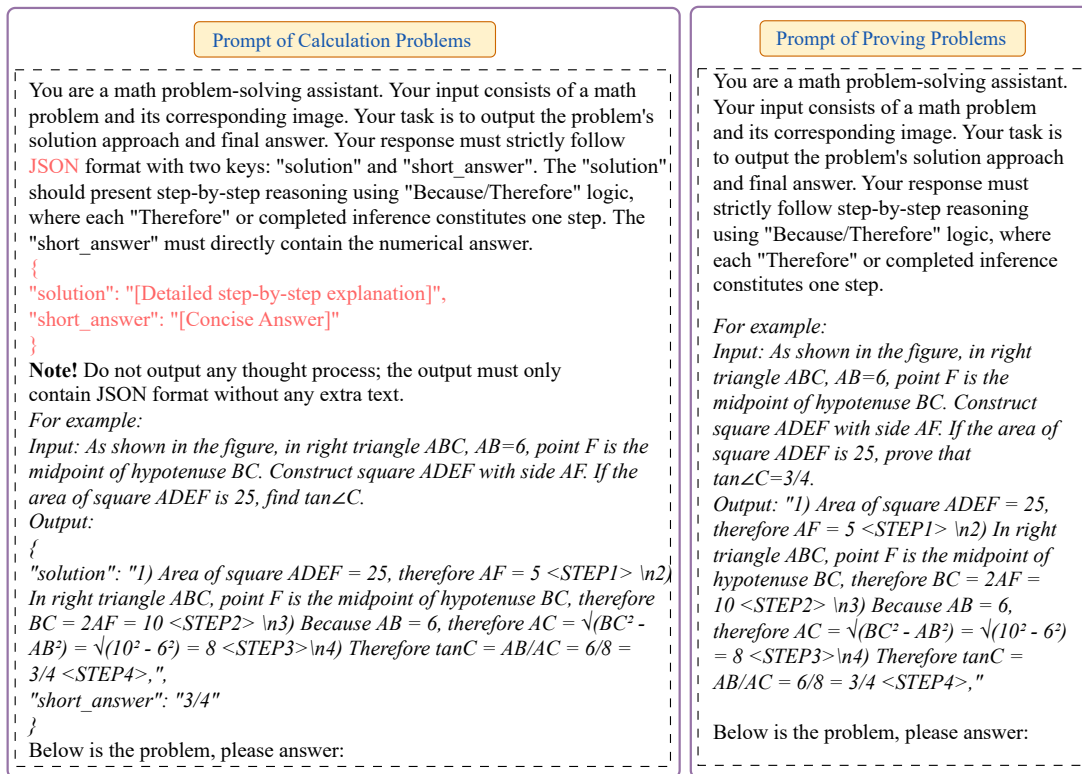


Figure 16: One-shot solution generation prompt for main evaluation.

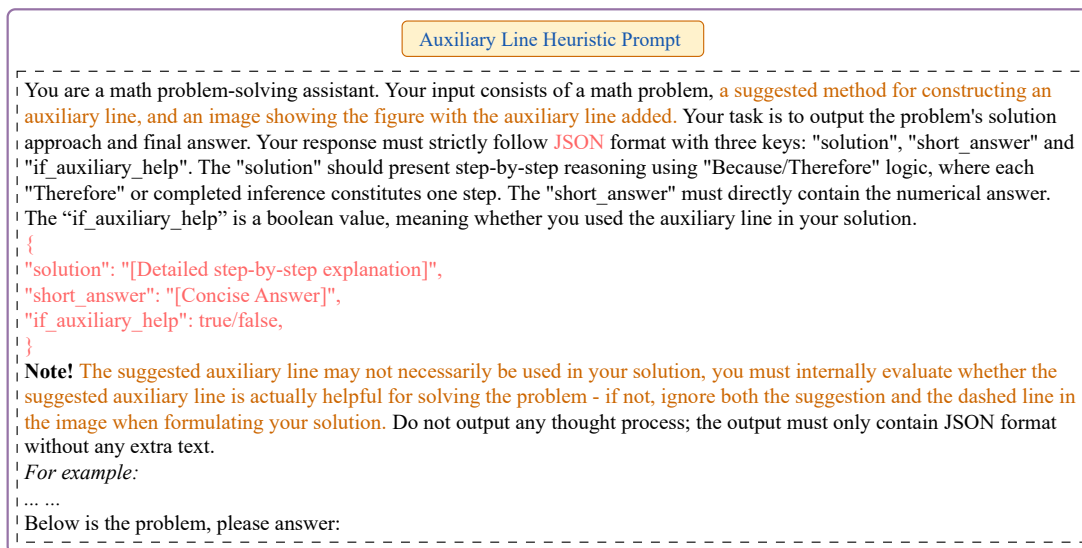


Figure 17: One-shot solution generation prompt for auxiliary line heuristic evaluation.

Step-by-Step Evaluation Prompt

You are a mathematics teacher. You will be given a problem, its corresponding diagram, reference answers, and a student's solution. Please evaluate the student's answer by referencing the solution steps in the reference answers. Your task consists of two parts:

1. Step-by-step evaluation: Carefully examine the student's solution and assess the correctness of each individual step. Score each step based solely on its own validity (1 for correct, 0 for incorrect).
2. Final answer judgment: Determine whether the student's final answer matches the correct answer provided for the problem.

Your response must follow the specified JSON format containing two keys: 'Step_by_step_evaluation' and 'Final_judgment'. For 'Step_by_step_evaluation', provide a list (e.g., [1, 0, ...]) indicating whether each step is correct (1) or incorrect (0). For 'Final_judgment', input either 1 (correct) or 0 (incorrect). The required format is as follows:

```
{
  "Step_by_step_evaluation": [x, x, x, x],
  "Final_judgment": 1/0,
}
```

Note! Do not output any thought process; the output must only contain JSON format without any extra text. Below is the problem and student' solution, please assess:

Figure 18: Zero-shot Step-by-Step Evaluation prompt.

Error Type Evaluation Prompt

You are a mathematics teacher capable of accurately identifying and categorizing students' error types. You will be provided with a mathematical problem, its corresponding diagram, the standard solution, and an incorrect response generated by an AI model. I will explicitly indicate which steps in the AI's solution are incorrect. Your task, with reference to the standard answer, is to analyze the root causes of these errors and precisely classify their types.

Your responsibilities consist of two components:

1. Error Cause Analysis: For each step marked as incorrect (score=0), determine why it's wrong and provide a detailed explanation of the fundamental error.
2. Error Type Classification: Based on your analysis, categorize each error into one of the following types: Figure Understanding Error, Knowledge Error, Calculation Error and Logical Reasoning Error.

Please select from these error types and output the corresponding error category for each incorrect step in sequence. For steps without errors, output "N/A". The specific definitions of each error type are as follows:

- 【Graphical Interpretation Error】** : Failure to correctly understand the geometric primitives (points, lines, circles, etc.) implied by the diagram, such as misidentifying angle relationships, collinear relationships, etc.
- 【Knowledge Error】** : While correctly understanding the point/line relationships, the solution employs incorrect formulas. This includes: using wrong formulas/theorems/properties, or selecting inappropriate formulas/theorems/properties for the given problem.
- 【Calculation Error】** : While correctly understanding the geometric relationships and properly selecting/applying the relevant knowledge, the solution contains numerical calculation mistakes or unit conversion errors.
- 【Logical Reasoning Error】** : The reasoning process contains logical fallacies, including but not limited to: invalid causal relationships between premises and conclusions (the "because-therefore" connection is unjustified), AI making intuitive assumptions without basis, drawing conclusions by introducing irrelevant external information or incorrect assumptions, nonsensical responses, logically chaotic arguments, or inexplicable answers.

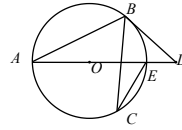
Your response must strictly adhere to the following JSON format containing two keys: "Error_analysis" and "Error_type". For the "Error_analysis" field: analyze the root cause of each incorrect step and explain why it is classified as that particular error type. For the "Error_type" field: output a list with length equal to the total number of steps, sequentially specifying the error type for each step or inserting N/A. The required format is as follows:

```
{
  "Error_analysis": "",
  "Error_type": [x, x, x, x],
}
```

Below is the problem and student' solution, please assess:

Figure 19: Zero-shot Error Type Evaluation prompt.

Question: As shown in the figure, the extension of diameter AE of circle $\odot O$ intersects the tangent line BD passing through point B at point D. Let point C be a point on circle $\odot O$ such that $\angle BCE = 25^\circ$. Then, $\angle ABD =$



[Original Image](#)

Answer: 115 Type: Calculation Step Count: 5



Evaluator: o4-mini Gemini-Claude-2.5-Pro 4.5 gpt-4o human

o3 Solution Process:

Step 1: Given that $\angle BCE = 25^\circ$, \therefore within a circle, the central angle subtended by chords CB and CE equals half the arc BE they subtend, \therefore we have arc BE = $2 \times 25^\circ = 50^\circ$.

Step 2: \because AE is the diameter, $\therefore \angle ABE$ (the semicircle) = 180° , $\therefore \angle AB = 180^\circ - 50^\circ = 130^\circ$.

Step 3: \because Circular Angle Theorem, \therefore For any point C on the circle (not lying on arc AB), we have $\angle ACB = \frac{1}{2} \cdot \text{arc AB} = \frac{1}{2} \cdot 130^\circ = 65^\circ$.

Step 4: \because Tangent-Chord Theorem, \therefore The angle $\angle ABD$ formed by the tangent BD and the chord BA is equal to the central angle $\angle ACB$ of the arc AB inscribed in the circle.

Step 5: $\because \angle ACB = 65^\circ, \therefore \angle ABD = \angle ACB = 65^\circ$

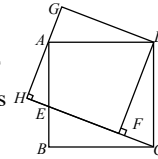
✓	✓	✓	✓	✓
✓	✓	✓	✓	✓
✓	✓	✓	✓	✓
✗	✗	✓	✓	✗
✗	✓	✓	✗	✗

o3 Answer: 65

Answer Judgement:

✗	✗	✗	✗	✗
---	---	---	---	---

Question: As shown in the figure, in square ABCD, E is a point on side AB. DF is perpendicular to CE at point F, AH is perpendicular to CE at point H, and GD is perpendicular to DF, intersecting AH at point G. Prove that $FH = AH + CF$.



[Original Image](#)

Type: Proving Step Count: 8



Evaluator: o4-mini Gemini-Claude-2.5-Pro 4.5 gpt-4o human

o3 Solution Process:

Step 1: Let the side length of square ABCD be 1. Establish a coordinate system: $A(0,1), B(0,0), C(1,0), D(1,1)$. Let point E lie on AB. Let $AE = t$ ($0 < t < 1$), then $E(0,t)$.

Step 2: The line CE passes through points $C(1,0)$ and $E(0,t)$, and its equation is $y = -tx + t$.

Step 3: Draw the foot H of the perpendicular from A to CE. ① The slope of the perpendicular line is $1/t$; ② Let $H(x_H, y_H)$. From $y_H - 1 = (1/t)(x_H - 0)$ and $y_H = -tx_H + t$, we solve for $x_H = t(t-1)/(1+t^2)$ and $y_H = t(t+1)/(1+t^2)$.

Step 4: Draw the foot F of the perpendicular from D to CE. ① The perpendicular line retains slope $1/t$ and passes through $D(1,1)$; ② Let $F(x_F, y_F)$. From $y_F - 1 = (1/t)(x_F - 1)$ and $y_F = -tx_F + t$, we solve for $x_F = (1 + t^2 - t)/(1 + t^2)$ and $y_F = t^2/(1 + t^2)$.

Step 5: Calculate the lengths of the three segments from coordinates: (1) $AH^2 = (x_H)^2 + (y_H - 1)^2 = (1 - t)^2 / (1 + t^2)$, $\therefore AH = (1 - t) / \sqrt{1 + t^2}$; (2) $CF^2 = (x_F - 1)^2 + (y_F)^2 = t^2 / (1 + t^2)$, $\therefore CF = t / \sqrt{1 + t^2}$.

Step 6: Find FH: $FH^2 = (x_H - x_F)^2 + (y_H - y_F)^2 = [-1/(1+t^2)]^2 + [t/(1+t^2)]^2 = 1/(1+t^2)$, $\therefore FH = 1/\sqrt{1+t^2}$.

Step 7: Compare Steps 5 and 6: $AH + CF = (1-t)/\sqrt{1+t^2} + t/\sqrt{1+t^2} = 1/\sqrt{1+t^2} = FH$.

Step 8: Therefore, $FH = AH + CF$, and the proposition is proven.

✗	✗	✓	✓	✗
✓	✓	✓	✓	✓
✓	✓	✓	✓	✓
✓	✓	✓	✓	✓
✓	✓	✓	✓	✓
✓	✓	✓	✓	✓
✓	✓	✓	✓	✓
✓	✓	✓	✓	✓

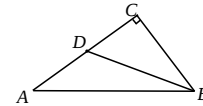
o3 Answer: $FH = AH + CF$.

Answer Judgement:

✓	✓	✓	✓	✓
---	---	---	---	---

Figure 20: Comparative case of Evaluation Processes across Different MLLMs

Question: As shown in the figure, in $\triangle ABC$, D is the midpoint of AC. Draw a line through D such that $DE \perp AC$, intersecting BC at point E. Draw a line through A such that $AF \parallel BC$, intersecting DE at point F. Connect AE and CF. If $CF = 2$, $\angle FAC = 30^\circ$, and $\angle B = 45^\circ$, find the length of AB.



[Original Image](#)

Reference Solution Process:

- Step 1:** $\because AE \perp BC, BD \perp AC \therefore \angle AED = \angle BDA = 90^\circ$.
- Step 2:** $\because \angle BAD$ and $\angle BED$ are complementary $\therefore \angle BAD + \angle BED = 90^\circ$.
- Step 3:** $\therefore \triangle ADF \cong \triangle CDE$.
- Step 4:** $\because AF = CE \therefore$ Quadrilateral AECF is a parallelogram.
- Step 5:** $\because DE \perp AC, \therefore$ Parallelogram AECF is a rhombus.
- Step 6:** \because AECF is a rhombus $\therefore AF = CF = 2$
- Step 7:** $\because AD = AF \cdot \cos 30^\circ = \sqrt{3}$
- Step 8:** $\therefore AC = 2AD = 2\sqrt{3}$.
- Step 9: Auxiliary Line: Draw $AM \perp BC$ through point A.**
- Step 10:** $\because AM = AC \cdot \sin 30^\circ = \sqrt{3}$
- Step 11:** $\therefore AB = AM / \sin 45^\circ = \sqrt{6}$

o3 Solution Process:

- Step 1:** $\because AF \parallel BC, \angle FAC = \angle ACB = 30^\circ$.
- Step 2:** Given $\angle B = 45^\circ$ and $\angle C = 30^\circ, \therefore \angle A = 180^\circ - 45^\circ - 30^\circ = 105^\circ$
- Step 3:** Let $AC = c$, **Set up a coordinate system: $A(0,0), C(c,0)$, then $D(c/2, 0)$**
- Step 4:** $\because \angle C = 30^\circ \therefore$ the angle between CB (and thus AF) and AC is $30^\circ, \therefore |\tan \theta| = \tan 30^\circ = 1/\sqrt{3}$. Let the slope of AF be $m = -1/\sqrt{3}$ (taking the negative sign does not affect the length).
- Step 5:** $\because DE \perp AC$, and passes through D, so the equation of DE is $x = c/2$, F is the intersection of AF and DE, $\therefore F(c/2, m \cdot c/2)$.
- Step 6:** Calculate CF: $CF^2 = (c - c/2)^2 + (0 - m \cdot c/2)^2 = c^2/4 \cdot (1 + m^2)$
- Step 7:** $m^2 = (1/\sqrt{3})^2 = 1/3, \Rightarrow CF^2 = c^2/4 \cdot (1 + 1/3) = c^2/4 \cdot 4/3 = c^2/3$
- Step 8:** Given $CF = 2 \Rightarrow 4 = c^2/3 \Rightarrow c = 2\sqrt{3}$, so $AC = 2\sqrt{3}$
- Step 9:** In $\triangle ABC$, by the Law of Sines: $AB / \sin C = AC / \sin B \Rightarrow AB = AC \cdot \sin 30^\circ / \sin 45^\circ = 2\sqrt{3} \cdot (1/2) / (\sqrt{2}/2) = \sqrt{6}$
- Step 10:** $\therefore AB = \sqrt{6}$

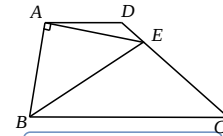


Every step is correct

Even though the o3 solution path diverges entirely from the reference, according to o4-mini's check, the Process Correctness Score (PCS) for this problem is 1.

Figure 21: Evaluation case of Solutions Different from the Reference Answer

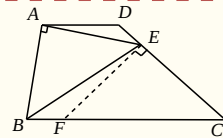
Question: As shown in the figure, in quadrilateral ABCD, $AD \parallel BC$ and $\angle C = 45^\circ$. An isosceles right triangle $\triangle BAE$ is constructed with AB as a leg ($AB = AE$, $\angle BAE = 90^\circ$), such that vertex E lies on side CD. If $CE = 6$, find the length of AD.



Original Image

Reference Auxiliary Line:

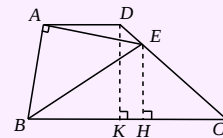
Construct $EF \perp CD$ through point E, intersecting BC at point F.



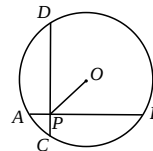
Auxiliary Line from Gemini-2.5-Pro:

Construct $EH \perp BC$ at point H, and $DK \perp BC$ at point K.

Valid and Correct Auxiliary Line



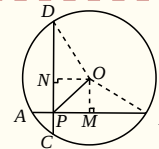
Question: As shown in the figure, in a circle O with radius 5, AB and CD are two chords that are perpendicular to each other, intersecting at point P. Given that $AB = CD = 6$, find the length of OP.



Original Image

Reference Auxiliary Line:

Construct $OM \perp AB$ at M, $ON \perp CD$ at N, and connect OB, OD.



Auxiliary Line form GPT 4.1:

Construct a line through O perpendicular to AB at P, so the distance from OP to AB is x.

Invalid and Incorrect Auxiliary Line

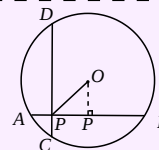
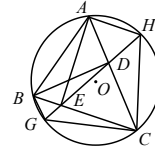


Figure 22: Evaluation case of Alternative Auxiliary Line Construction Methods

Question: As shown in the figure, triangle ABC is inscribed in circle O. AE and BD are the altitudes from vertices B and C to sides BC and AC, respectively. Chord GH of circle O passes through points D and E. Connect CG, CH, BG, and AH. Given that $\angle BAD$ and $\angle BED$ are supplementary angles. (2) Prove that $CG = CH$.



Original Image

Type: Proving Step Count: 12

Evaluator: o4-mini

o3 Solution Process:

Step 1: $\because AE \perp BC, BD \perp AC \therefore \angle AED = \angle BDA = 90^\circ$.

Step 2: $\because \angle BAD$ and $\angle BED$ are complementary $\therefore \angle BAD + \angle BED = 90^\circ$.

Step 3: \because In quadrilateral ABED, we have $\angle AED = 90^\circ$ (<STEP1>) and $\angle BAD + \angle BED = 90^\circ$ (<STEP2>) $\therefore \angle AED + \angle ABD = 180^\circ \therefore$ Quadrilateral ABED is a cyclic quadrilateral.

Step 4: Theorems of a Circle-Inserted Quadrilateral \Rightarrow Circumscribed angles subtended by the same arc are equal $\therefore \angle BED = \angle BAD$ (same arc BD)

Step 5: From <STEP2> and <STEP4>, we have $2\angle BAD = 90^\circ \therefore \angle BAD = \angle BED = 45^\circ$

Step 6: In $\triangle BAC$, $\angle BAD = 45^\circ$ and $D \in AC \therefore \angle CAB = 45^\circ$

Step 7: Similarly, we obtain $\angle CBA = 45^\circ \therefore \triangle ABC$ is an isosceles right triangle, with BC as its hypotenuse and O as the midpoint of BC.

Step 8: Let $BC = 2R$ (where R is the radius of circle $\odot O$). Then $BO = CO = R \therefore$ Points B, O, and C are collinear, and O is the midpoint of BC.

Step 9: Take points E and D on BC and AC, respectively. Let GH be the chord passing through D and E \therefore E, D, G, and H are collinear.

Step 10: Draw $OC \perp GH$ at M \therefore M is the midpoint of GH (Chord Perpendicular Bisector Theorem)

Step 11: \because O is the midpoint of BC, and $OC \perp GH$ (<STEP10>) \therefore C and M lie on the perpendicular bisector of GH.

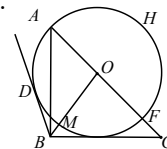
Step 12: \because C lies on the perpendicular bisector of GH $\therefore CG = CH$



Model adopt flawed intermediate steps to justify the conclusion

Question: As shown in the figure, in right triangle ABC, $\angle ABC = 90^\circ$.

Let O be a point on AC. Draw a circle centered at O with radius OA, intersecting AC at point F. Let BD be the tangent line to circle $\odot O$ at point D. Connect OB, intersecting circle $\odot O$ at point M. Given that $\angle DBO = \angle CBO$, (1) Prove that BC is a tangent line to circle $\odot O$.



Original Image

Type: Proving Step Count: 5

Evaluator: o4-mini

o3 Solution Process:

Step 1: \because BD is a tangent to $\odot O$, $\therefore OD \perp BD$ (a tangent is perpendicular to the radius).

Step 2: Given that $\angle DBO = \angle CBO$, \therefore Line BC is axisymmetric with respect to OB relative to line BD.

Step 3: Reflect point D on circle $\odot O$ across axis OB to obtain point M. Clearly, $OM = OD = r$. Therefore, M remains on $\odot O$, and points O, B, and M are collinear.

Step 4: By the property of axial symmetry, since OD is perpendicular to BD, OM is also perpendicular to BC, that is, $OM \perp BC$.

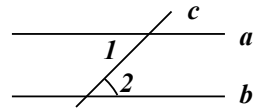
Step 5: Since M belongs to $\odot O$ and OM is a radius, and since OM is perpendicular to BC, BC and $\odot O$ share only one common point M. This satisfies the tangent line criterion for a radius perpendicular to a line. Therefore, BC is a tangent line to $\odot O$ (with M as the point of tangency).



Model adopt flawed intermediate steps to justify the conclusion

Figure 23: Examples of shortcuts in Proof Problems

Question 1: Lines a and b are intersected by line c . If $a \parallel b$ and $\angle 1 = 63^\circ$, then the measure of $\angle 2$ is ?



o3 solution process:

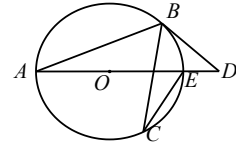
- (1) Given that $a \parallel b$ and c is the transversal, $\angle 1$ and $\angle 2$ are both located between the two parallel lines and are on the same side (**consecutive interior angles**).
- (2) ... , $\angle 1 + \angle 2 = 180^\circ$. (3) ... , $\angle 2 = 180^\circ - 63^\circ = 117^\circ$.



Figure Understanding Error



Question 2: The extension of the diameter AE of circle $\odot O$ intersects with the tangent BD at point B at point D . C is a point on the circle $\odot O$ such that $\angle BCE = 25^\circ$. Find $\angle ABD$.



o3 solution process:

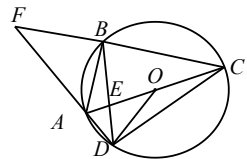
- (1)-(2) Given that $\angle BCE = 25^\circ$, ... , $\text{arc } BE = 2 \times 25^\circ = 50^\circ$... therefore $\text{arc } AB = 130^\circ$
- (3) By the inscribed angle theorem: ... , $\angle ACB = (1/2)\text{arc } AB = 65^\circ$.
- (4) By the tangent-chord angle theorem: **the angle between tangent BD and chord BA , $\angle ABD$, is equal to the inscribed angle $\angle ACB$ subtended by arc AB .**
- (5) Therefore, $\angle ABD = 65^\circ$.



Knowledge Error



Question 3: Quadrilateral $ABCD$ is inscribed in circle $\odot O$. The diagonals AC and BD intersect at point E , and AC is the diameter of the circle. Extend DA and CB to meet at point F . Connect OD . It is given that $\angle AOD = \angle ACB$ and $AE / CE = 3 / 8$. Find DE / BE .



Gemini-2.5-pro solution process:

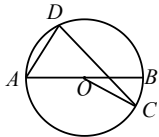
- (1)-(2) Since ... , $\angle BAE = \angle CDE$ (they subtend the same arc BC), $\angle ABE = \angle DCE$
- (3)-(4) Therefore, $\triangle ABE \sim \triangle DCE$ (by AA similarity). Hence, ... , $AE / DE = BE / CE$.
- (5) Rearranging this proportion gives: **$DE / BE = CE / AE$.**
- (6) Since it is given that $AE / CE = 3 / 8$, it follows that $CE / AE = 8 / 3$. Therefore, $DE / BE = 8 / 3$.



Calculation Error



Question 4: AB is the diameter of circle O , and points C and D on either side of AB both lie on circle O . If $\angle BOC = 30^\circ$, then $\angle ADC = ?$



Gemini-2.5-pro solution process:

- (1) ... Given that $\angle BOC = 30^\circ$, therefore $\angle AOC = \angle AOB - \angle BOC = 150^\circ$.
- (2) Since $\angle AOC$... the measure of minor arc AC is 150° .
- (3) $\angle ADC$ is an inscribed angle subtended by major arc AC (**alternatively, in cyclic quadrilateral $ADCB$, $\angle ADC$ and $\angle ABC$ are opposite angles and therefore supplementary**), so $\angle ADC + \angle ABC = 180^\circ$.
- (4) Since ... Therefore, $\angle ABC = (1/2) \times \angle AOC = 75^\circ$.
- (5) Thus, $\angle ADC = 180^\circ - \angle ABC = 180^\circ - 75^\circ = 105^\circ$.



Logical Reasoning Error



Figure 24: Examples of different error types.