

GeoChallenge-90K: A Multi-Answer Multiple-Choice Benchmark for Diagram-Grounded Geometric Reasoning

Yushun Zhang^{1,2}, Weiping Fu^{1,2,3*}, Zesheng Yang^{1,2}, Bo Zhao^{1,2}, Lingling Zhang^{1,2}, Jian Zhang^{1,2}, Yumeng Fu^{1,2}, Jiaxing Huang^{1,2}, Jun Liu^{1,2,3},

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

²Shaanxi Province Key Laboratory of Big Data Knowledge Engineering,
Xi'an Jiaotong University, Xi'an, 710049, China

³National Engineering Research Center for Visual Information and Applications
meteoric@stu.xjtu.edu.cn, fuweiping1993@foxmail.com

Abstract

Evaluating the symbolic reasoning of large language models (LLMs) calls for geometry benchmarks that require multi-step proofs grounded in both text and diagrams. However, existing benchmarks are often limited in scale and rarely provide visually grounded multiple-choice questions, limiting reliable evaluation of complex reasoning. We introduce GeoChallenge, a dataset of 90K automatically generated multiple-choice geometry proof problems, each requiring multi-step reasoning over aligned textual descriptions and diagrams. GeoChallenge provides fine-grained complexity ratings and formal language annotations to enable controlled evaluation. Experiments on multiple advanced LLMs show a clear performance gap between models and humans (the best-performing model, GPT-5-nano, achieves 75.89 exact match vs. 94.74 for humans). Further analysis also reveals three common failure patterns of LLMs: (1) exact match failures under the multiple-choice setting; (2) weak visual reliance; and (3) overextended reasoning without convergence.

1 Introduction

Geometry problem solving is a fundamental task involving spatial reasoning and symbolic deduction (Trinh et al., 2024; Chervonyi et al., 2025; Pan et al., 2025; Dai et al., 2025), and is widely used to evaluate complex reasoning in large language models (LLMs) (Luo et al., 2025; Feng et al., 2025; Zhang et al., 2025c). Yet evaluating such capabilities of LLMs in a systematic and scalable manner remains an open challenge (Lei et al., 2024; Hariharan et al., 2025; Xu et al., 2025; White et al., 2024; Kazemi et al., 2023).

Geometry datasets and benchmarks have progressed along two complementary lines. Early efforts such as Geometry3K (Lu et al., 2021) and

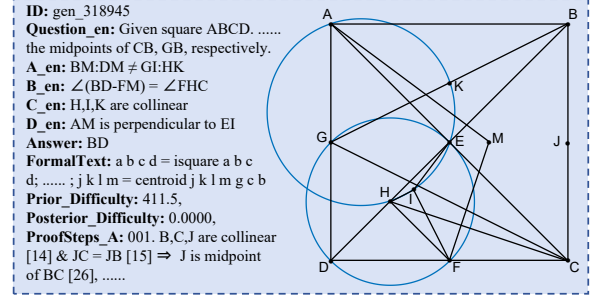


Figure 1: Examples in GeoChallenge-90K dataset.

GeoQA (Chen et al., 2021) translate diagram-text problems into machine-interpretable structures (e.g., symbolic relations or executable programs), enabling systematic evaluation of geometric reasoning. More recent benchmarks, including MathVerse (Zhang et al., 2024b) and OlympiadBench (He et al., 2024), target multimodal (Zhang et al., 2025b) and emphasize faithful diagram understanding in harder visual-math settings. However, reliable evaluation remains challenging: many datasets are manually curated from textbooks or contests, which limits coverage and scalability, and high-difficulty benchmarks stay small due to expert verification costs. Moreover, prevailing single-answer multiple-choice (Feng et al., 2025; Wang et al., 2025; Xing et al., 2024) or open-ended formats (Zhang et al., 2025a; Fu et al., 2025b) are either prone to guessing or difficult to grade at scale. Finally, text-diagram misalignment persists, weakening visually grounded evaluation, as highlighted by GeoGPT4V (Cai et al., 2024).

To address these issues, we propose a scalable automatic generation pipeline and introduce GeoChallenge-90K, a dataset of 90,279 challenging geometry proof questions. Each instance provides an aligned textual description, a rendered diagram, and four candidate options with possibly multiple correct answers, enabling option-level evaluation and discouraging elimination-style guessing.

*Corresponding author

GeoChallenge-90K covers a wide range of proof length (avg. 16.72 steps), diagram complexity, and difficulty, with fine-grained complexity ratings to support controlled stress tests.

Experiments on GeoChallenge-90K reveal a large gap between current models and human solvers. General-purpose models average 21.48% accuracy, reasoning-oriented models reach 56.07%, and humans achieve 94.74%. Hierarchical evaluation shows that general-purpose models degrade steeply with increasing complexity, whereas humans and reasoning-specialized models remain comparatively stable. Diagram ablations expose a grounding gap: removing diagrams substantially reduces human accuracy but only marginally affects models, suggesting that current LLMs do not reliably extract or calibrate diagram evidence. Error analysis further indicates recurring failures-logical fallacies, invalid outputs, and overextended reasoning with no verifiable conclusion-while human mistakes are mostly genuine reasoning slips rather than format-related failures.

In summary, our main contributions are:

- We introduce GeoChallenge-90K, a multi-answer multiple-choice benchmark for diagram-grounded geometric reasoning with text-diagram pairs, formal annotations, and fine-grained complexity control.
- Extensive experiments provide systematic empirical evidence that a substantial model-human gap persists in challenging, long-step geometric reasoning.
- Our diagnostic analysis reveals large gaps between exact-match and option-level metrics, weak grounding in diagram evidence, as well as unstable long-step reasoning that often leads to non-convergent answers.

2 Related Work

2.1 Geometry Problem Generation

Early geometry datasets were built via templates or manual collection and annotation (Prasetyanto et al., 2020; Kazemi et al., 2023; Zhang et al., 2024c; Chen et al., 2021; Lu et al., 2021; Chen et al., 2022), which produced valuable resources but faced scalability and coverage limits. Symbolic generation alleviates these issues by making problems machine-verifiable: Inter-GPS (Lu et al., 2021) provides parsing/representation, GeoGen (Bak, 2020) generates symmetric instances,

FormalGeo (Zhang et al., 2023b) formalizes verification, and R-CoT (Deng et al., 2024a) increases QA diversity. More recently, theorem-guided and verified pipelines further scale generation and alignment, including TR-CoT (Deng et al., 2024b), GenesisGeo (Zhu et al., 2025), and TrustGeoGen (Fu et al., 2025a).

LLMs/MMs introduce another paradigm. G-LLaVA (Gao et al., 2023) synthesizes Geo170K with text LLMs, while GeoGPT4V (Cai et al., 2024) leverages GPT-4V/Wolfram to improve difficulty and image-text alignment, demonstrating scalable, targeted generation (Chen et al., 2024; Li et al., 2024).

2.2 Existing geometry datasets

Geometry benchmarks range from early text-diagram datasets (e.g., GeoS (Seo et al., 2015)) to large-scale, richly annotated corpora such as Geometry3K (Lu et al., 2021), GeoQA/GeoQA+ (Chen et al., 2021; Cao and Xiao, 2022), UniGeo (Chen et al., 2022), and PGPS9K (Zhang et al., 2023a), which provide formal languages, programmatic solutions, and detailed diagram annotations. GeoLaux (Fu et al., 2025b) further targets long-step reasoning with auxiliary-line constructions.

With LLMs/MMs, benchmarks increasingly stress multimodal understanding, robustness, and perception, including MATH (Hendrycks et al., 2021), GeoEval (Zhang et al., 2024a), MM-MATH (Sun et al., 2024), GePBench (Xing et al., 2024), and FrontierMath (Glazer et al., 2024).

3 The GeoChallenge-90K dataset

3.1 Definitions

We first introduce the core concepts used in our automatic data generation procedure. Concrete examples are provided in Appendix A.

Clause. A clause is a statement that specifies a geometric relation among existing objects (e.g., points, lines): $f(X_1, \dots, X_n)$, where f is a predefined relation in table 11 over defined points X_i .

Construction. A construction is a geometric description that uses one or two clauses to uniquely define a new point x (where two indicate their intersection): $x = f(), g()$. Clause $g()$ is omitted if $f()$ alone suffices.

Premise. A premise is an ordered sequence of constructions C_i defining all points and relations in a problem: $P = (C_1, C_2, \dots, C_n)$.

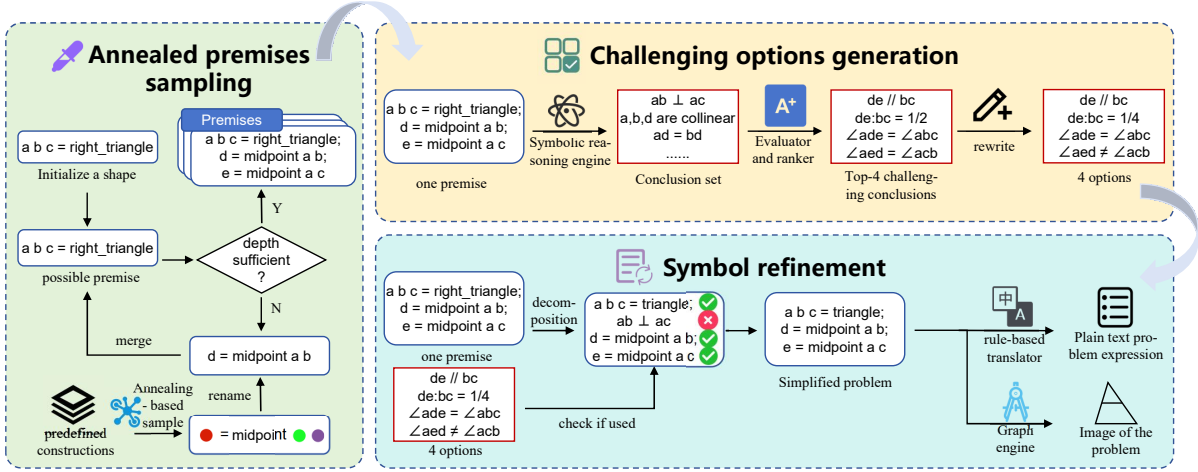


Figure 2: Pipeline of dataset generation

Problem. A problem is a structured task comprising a premise P and four options O_i allowing multiple correct answers: $(P, O_1, \dots, O_4)$. See Figure 1 for an example.

3.2 Problem Generation

3.2.1 Annealed premises sampling

As shown in Figure 2, we generate a large pool of premises by composing predefined clause templates into multi-layer geometric constructions. Starting from a minimal seed at depth 0, we iteratively add layers until a target depth is reached. At each layer, we sample one or two templates and instantiate them with valid arguments from currently defined entities, ensuring consistency, then append the resulting clauses to the premise.

To balance structural diversity and computational tractability, we move beyond standard breadth-first expansion by introducing an annealed sampling strategy. By dynamically shrinking the branching factor, early layers aggressively explore diverse template and parameter combinations to broaden the search space, while deeper layers systematically prune candidates to suppress combinatorial explosion. We set the maximum depth to $N = 8$, producing 871,828 sampled premises.

3.2.2 Challenging options generation

Given a premise, we enumerate provable conclusions using a symbolic engine, score their difficulty, and select the top-scoring ones as the four options. We use AlphaGeometry (Trinh et al., 2024) for deduction with a rule set R . As summarized in Appendix B, the engine performs conclusion search via forward chaining, repeatedly applying rules in

R to newly derived facts until saturation or a preset depth limit.

Following GeoEval (Zhang et al., 2024a), we define difficulty as a weighted sum of five indicators: $\text{prior_difficulty} = \sum_{i=1}^5 w_i x_i$, where x_1 is description length, x_2 premise length, x_3 the number of points, x_4 proof-search depth, and x_5 proof length. To construct the four candidate options from the top-scoring conclusions, we employ three transformation strategies: Keep as True, Synonym Rewriting, and Conflict Rewriting, with detailed examples provided in Appendix C. Specifically, we apply Keep as True to at least one conclusion to guarantee a baseline correct answer. For the remaining conclusions, we independently apply either Synonym Rewriting (e.g., equivalence-preserving rewrites) or Conflict Rewriting (e.g., relation negation or ratio perturbation). This methodology naturally supports our Multi-Answer (MA) setting, as the application of Synonym Rewriting dynamically determines the number of valid options. We strictly avoid naive entity substitution (e.g., $AB \perp CD \rightarrow AB \perp CE$), which often yields degenerate or accidentally true options. The distribution of the correct options across the dataset is shown in Table 2.

3.2.3 Symbol refinement

We then refine the symbolic instances and render aligned diagrams. Text refinement includes (i) simplification by retaining only points/clauses used in both the premise and the proofs of candidate options, and (ii) bilingual rendering that maps the formal representation to English and Chinese with rule-based templates while preserving option semantics.

For diagrams, we render figures consistent with

Dataset	Size	AG	IF	CR	Avg PL	Avg DL	FA	LT	QT
GeoS (Seo et al., 2015)	186	✗	<i>T, I</i>	✗	NA	23.34	✗	EN	SA
MATH (Hendrycks et al., 2021)	12,500	✗	<i>T</i> only	✓	5.48	202.66	✗	EN	OE
Geometry3K (Lu et al., 2021)	3,002	✗	<i>T, I</i>	✗	NA	12.19	✓	EN	SA
GeoQA (Chen et al., 2021)	5,010	✗	<i>T, I</i>	✗	2.77	52.46	✗	ZH	SA
UniGeo (Chen et al., 2022)	9,543	✗	<i>T, I</i>	✗	2.57	16.23	✓	EN	SA, OE
GPSM8K (Cobbe et al., 2021)	8,790	✗	<i>T</i> only	✗	4.59	45.27	✗	EN	OE
PGDP5K (Hao et al., 2022)	5,000	✗	<i>I</i> only	✗	–	–	✓	–	–
PGPS9K (Zhang et al., 2023a)	9,022	✗	<i>T, I</i>	✗	2.43	9.46	✓	EN	OE
MathVista (Lu et al., 2024)	6,141	✗	<i>T, I</i>	✗	–	83.83	✗	EN, ZH	SA, OE
GeomVerse (Kazemi et al., 2023)	9,303	✓	<i>T, I</i>	✓	6.35	53.09	✓	EN	OE
Geo170K (Gao et al., 2023)	177,457	✓	<i>T, I</i>	✗	6.21	49.01	✗	EN	SA
GeoEval (Zhang et al., 2024a)	5,050	✗	<i>T, I</i>	✓	2.55	28.34	✓	EN	SA, OE
MM-MATH (Sun et al., 2024)	5,929	✗	<i>T, I</i>	✗	8.50	48.50	✗	EN	OE
AutoGeo-100K (Huang et al., 2025)	100,000	✓	<i>I</i> only	✓	–	22.41	✓	EN	–
MathVerse (Zhang et al., 2024b)	15,672	✗	<i>T, I</i>	✓	–	40.67	✗	EN	SA
GeoMM (Deng et al., 2024a)	33,595	✓	<i>T, I</i>	✓	1.51	32.47	✗	EN	OE
Average	–	–	–	–	3.92	44.98	–	–	–
GeoChallenge-90K	90,279	✓	<i>T, I</i>	✓	16.72	39.15	✓	EN, ZH	MA

Table 1: Comparison between GeoChallenge-90K benchmark and existing geometry problem solving benchmarks. AG: Automatic Generation. IF: Input Format, T for text and I for image. CR: Complexity Rating. Avg PL: Average Proof Length. Avg DL: Average Description Length. FA: Formal Annotation. LT: Language Type, EN for English and ZH for Chinese. QT: Question Type, SA for single-answer, MA for multiple-answer and OE for open-ended question

Correct Options	Frequency	Percentage
1 Option	49,452	54.78%
2 Options	32,862	36.40%
3 Options	5,081	5.63%
4 Options	2,884	3.19%
Total	90,279	100.00%

Table 2: Distribution of the number of correct options in GeoChallenge-90K

Model	MATH	Geom-etry3K	Geo-170K	Geo-Eval	Math Verse	Geo-Cha.
Gemini 1.5 Pro	82.19	73.88	56.3	–	37.3	24.45
Claude 3.5 Sonnet	58.05	71.65	51.2	–	38.9	21.81
GPT-4o	57.72	64.92	50.5	37.22	46.5	17.51

Table 3: Performance comparison across benchmarks.

the refined premise and explicitly annotate candidate options for visual grounding. To support stable batch rendering beyond the default AlphaGeometry module, we add rule-based option labeling and robustness fixes to prevent incomplete annotations and occasional non-termination.

3.2.4 Manual Verification

To ensure benchmark quality, we implement a manual verification process as a visual quality-control filter. Figures are rigorously audited based on legibility, geometric validity, and semantic alignment; comprehensive details regarding the verification protocols and checklist are provided in Appendix D.1.

3.3 Features of GeoChallenge-90K

GeoChallenge-90K is designed to evaluate diagram-grounded geometry theorem proving under rigorous, scalable, and controllable settings.

Built on the fully automatic symbolic pipeline in Section 3.2, all instances are machine-verifiable, enabling scalable construction without manual proof annotation. Table 1 compares GeoChallenge-90K with representative geometry and multimodal math benchmarks. The key characteristics of our dataset are summarized as follows.

Multi-answer MCQ evaluation. A distinctive feature of GeoChallenge-90K is its multi-answer MCQ format, where an instance may contain more than one correct option. Compared to single-answer MCQs, this setting substantially weakens elimination-style guessing and forces *per-option verification*: models must assess each candidate conclusion under the premise, often requiring different proof paths or relation checks. This format supports option-level metrics and diagnoses over-/under-selection behaviors.

Model	EMA $\uparrow$	EME $\uparrow$	EMM $\uparrow$	EMH $\uparrow$	P $\uparrow$	R $\uparrow$	F1 $\uparrow$	HA $\uparrow$	Avg #Sel
<i>Baselines</i>									
Random	6.67	6.67	6.67	6.67	39.31	53.33	45.25	49.29	2.13
<i>General-Purpose Models</i>									
GPT-4o	17.51	29.09	13.81	9.24	52.87	68.83	59.81	58.62	2.18
Claude 3.5 Sonnet	21.81	27.27	22.94	10.87	54.88	71.73	62.19	63.22	2.20
Gemini 1.5 Pro	24.45	38.18	22.27	9.24	62.31	76.79	68.79	69.74	2.01
llava-1.5-7B	16.96	37.82	10.69	1.09	48.31	31.36	38.03	57.87	1.03
Qwen2-VL-7B	26.65	59.64	17.37	0.0	61.36	47.44	53.51	65.69	1.19
<i>Reasoning-Oriented Models</i>									
GPT-o3	67.84	67.27	69.04	65.76	80.84	81.20	81.02	82.52	1.64
GPT-5-nano	75.88	70.59	79.31	75.00	84.82	83.78	84.30	85.49	1.61
Claude 4.5 Sonnet	41.96	50.00	41.38	30.00	72.02	77.90	74.85	76.56	1.75
Gemini 3 Pro	38.55	50.00	37.29	23.81	44.74	41.45	43.03	43.20	1.04
<i>Human Performance</i>									
Human	94.71	94.12	94.92	95.24	97.95	99.71	98.82	98.68	1.66

Table 4: Overall performance on GeoChallenge with both text and images provided. EMA/EME/EMM/EMH report Exact Match (EM) on the All/Easy/Medium/Hard splits, respectively.

Long-step Proofs under Concise Statements. GeoChallenge-90K targets long-step deduction. Proofs average 16.72 steps-over $4\times$ typical geometry benchmarks (3.92)-while statements remain similarly concise (39.15 vs. 44.98 words). This is by design that descriptions include only essential relations, increasing information density per token and thus reasoning difficulty. As shown in Table 3, models suffer a clear accuracy drop on GeoChallenge-90K relative to prior benchmarks, making it a more challenging and more diagnostic testbed.

Comprehensive geometric coverage and structural diversity. Beyond scale, GeoChallenge-90K is constructed to cover a broad range of geometric primitives and composite structures. Figure 3 summarizes the distribution of geometric elements. Importantly, less frequent shapes in many benchmarks, such as trapezoids and parallelograms, are still represented at scale over 10K instances each, which improves coverage of cross-element interactions and mitigates sparsity for rare configurations.

Dual-modality and bilingual alignment. Each instance includes both text and a rendered diagram, and the two modalities are semantically aligned: the textual statement fully specifies the geometric conditions reflected in the diagram. This alignment enables controlled comparisons among text+image, text-only, and a diagnostic image-only condition. We report text+image and text-only results in the main tables and the complementary image-only ablation in Appendix G. We further provide English and Chinese versions with strict semantic equivalence, reducing confounds from translation arti-

facts and enabling systematic analysis of language effects.

Rich annotations and controlled difficulty. GeoChallenge-90K includes structured formal representations for premises and options, along with two complementary difficulty signals: a prior difficulty estimated from complexity indicators, and posterior difficulty derived from tested models’ performance. Problems are stratified into three difficulty levels with a 3:5:2 split, enabling controlled stress testing and fine-grained performance analysis.

4 Experiments

4.1 Experimental Setup

Prompting / Inference Settings. We adopt a unified prompting protocol across all evaluated models. A single prompt template (included in the supplementary material due to length) is used for every example to elicit a final multiple-choice answer along with brief, option-wise reasoning. We disable external tools and retrieval for all models. Inference is run with greedy decoding (temperature=0.0) and a maximum output length of 16,384 tokens; consequently, each instance is evaluated with one generation only, without additional sampling or self-consistency aggregation.

Models Evaluation. To establish a human baseline, we conducted a controlled evaluation with graduate-level mathematical experts. Comprehensive discussions are provided in Appendix D.2.

We evaluate prominent open-source and closed-source large models selected for strong mathematical reasoning, covering (i) general-purpose

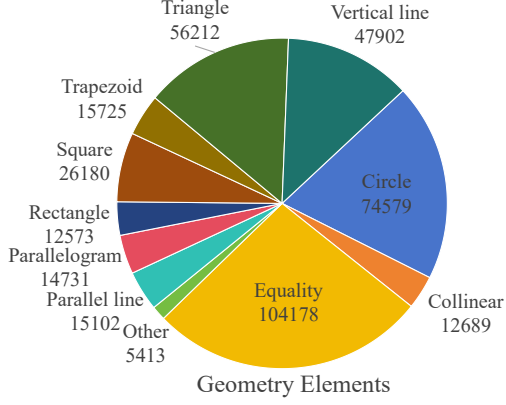


Figure 3: Geometry elements in GeoChallenge-90K

LMM/LLMs and (ii) reasoning-oriented models that allocate extra computation for deliberate reasoning. In the multimodal setting, we test GPT-4o (Hurst et al., 2024), Claude 3.5 Sonnet (Anthropic, 2024), Gemini 1.5 Pro (Team et al., 2024), LLaVA-1.5-7B (Liu et al., 2024), and Qwen2-VL-7B-Instruct (Yang et al., 2024), as well as reasoning-oriented models including GPT-o3-2025-04-16 (OpenAI, 2025a), GPT-o3-mini-2025-01-31 (OpenAI, 2025b), Claude 4.5 sonnet-20250929-thinking (Anthropic, 2025), and Gemini-3-pro-preview-11-2025 (Google, 2025). For text-only evaluations (LLM setting), we assess the same model families without visual inputs, and additionally include Deepseek-r1-0528 (DeepSeek-AI et al., 2025), Qwen3-235b-a22b-thinking-2507 (Yang et al., 2025), and WizardMath-7B (Luo et al., 2023).

For replicability and efficiency, we run most evaluations on GeoChallenge-small, a standardized subset of 908 problems sampled to match the difficulty profile of the full benchmark, enabling practical assessment of closed-source models.

Evaluation Protocol. We evaluate LLMs on a multi-select benchmark. For each problem i with K_i options, the model predicts a subset $\hat{S}_i \subseteq \{1, \dots, K_i\}$, compared against the ground-truth subset S_i .

We report Exact Match (EM), which is correct iff $\hat{S}_i = S_i$, both overall and by difficulty (Easy/Medium/Hard) determined by our difficulty scores. To measure partial correctness, we compute option-level precision/recall/F1 from the overlap between $\hat{S}_i$ and S_i and report macro-averages across problems. We also report Hamming Loss

(HL), the average per-option error rate:

$$\text{HL} = \frac{1}{N} \sum_{i=1}^N \frac{1}{K_i} \sum_{j=1}^{K_i} \mathbf{1}(\hat{y}_{i,j} \neq y_{i,j}),$$

where $y_{i,j}, \hat{y}_{i,j} \in \{0, 1\}$ are the ground-truth and predicted labels; we report Hamming Accuracy (HA) as $1 - \text{HL}$. Finally, we report the average number of selected options $\mathbb{E}[|\hat{S}_i|]$ to characterize selection behavior.

4.2 Main Results

We report overall results on GeoChallenge with both text and diagrams provided in Table 4. Since GeoChallenge is formulated as multiple-choice with long-step reasoning, we emphasize Exact Match as the primary metric; in contrast, option-level metrics (P/R/F1 and HA) may still reward partially correct selections even when the predicted answer is not fully correct.

No-guess multiple-choice reveals hidden difficulty. Under the no-guess protocol, the bottleneck shifts from partially identifying correct options to producing a single, answer-consistent prediction. A key observation is the large gap between strict EM and option-level metrics, especially for general-purpose models: e.g., Gemini 1.5 Pro attains high F1/HA but much lower EMA, with similar discrepancies for GPT-4o and Claude 4.5 Sonnet. This indicates that models often recover parts of the correct option answer yet fail to output the exact answer demanded by the no-guess protocol, making heuristics such as eliminating a few options or selecting multiple plausible answers ineffective. Consistently, Avg #Sel shows that general models select close to two options on average (near the random baseline), trading precision for recall, whereas reasoning-oriented systems are better calibrated and closer to human selection cardinality.

Human-model gap on long-step geometry. GeoChallenge exhibits a substantial human-model gap even with reasoning-oriented systems, and the gap remains on the Hard split. The best model in Table 4 (GPT-5-nano) remains obviously below human EMA, and the gap persists on the Hard split; meanwhile, general multimodal baselines lag far behind. This suggests the benchmark separates systems that sustain long-step, answer-consistent reasoning from those that succeed mainly via partial correctness or shallow

Model	EMA $\uparrow$	EME $\uparrow$	EMM $\uparrow$	EMH $\uparrow$	P $\uparrow$	R $\uparrow$	F1 $\uparrow$	HA $\uparrow$	Avg #Sel
<i>Baselines</i>									
Random	6.67	6.67	6.67	6.67	39.31	53.33	45.25	49.29	2.13
<i>General-Purpose Models</i>									
GPT-4o	13.99	32.00	8.69	0.0	43.75	55.89	49.08	48.62	2.14
Claude 3.5 Sonnet	17.62	32.73	15.37	0.54	51.43	63.48	56.83	60.21	2.02
Gemini 1.5 Pro	20.59	44.00	14.70	0.0	60.32	74.14	66.52	66.77	1.97
Qwen2-VL-7B	25.66	57.82	16.26	0.54	60.26	46.66	52.59	64.98	1.19
WizardMath-7B	8.15	9.45	7.80	7.07	38.48	59.32	46.68	49.78	2.40
<i>Reasoning-Oriented Models</i>									
GPT-o3-mini	50.88	50.00	53.45	45.00	59.52	59.60	59.56	59.82	1.64
GPT-o3	62.67	66.18	60.80	61.96	72.95	71.41	72.17	72.60	1.59
GPT-5-nano	58.15	64.00	58.35	48.91	67.02	64.96	65.97	66.93	1.55
Claude 4.5 Sonnet	36.23	51.64	34.30	17.93	66.58	67.36	66.97	71.34	1.62
Gemini 3 Pro	51.21	62.55	49.00	39.67	57.54	54.55	56.01	55.97	1.37
Deepseek r1	16.41	17.82	16.70	13.59	23.16	27.00	24.94	23.54	2.08
Qwen3-235b-thinking	75.55	81.09	75.95	66.30	82.18	82.30	82.24	82.74	1.58
<i>Human Performance</i>									
Human	42.86	38.24	46.55	40.00	66.96	68.75	67.85	72.77	1.69

Table 5: Overall performance on GeoChallenge with text-only provided. EMA/EME/EMM/EMH report Exact Match (EM) on the All/Easy/Medium/Hard splits, respectively.

heuristics. Reasoning-oriented models achieve relatively strong P/R/F1, implying they often locate relevant options, but their remaining deficit under strict EM points to unresolved challenges in multi-step consistency, diagram-grounded verification, and satisfying global geometric constraints.

Different degradation with increasing difficulty.

As difficulty increases, general-purpose models often degrade sharply, whereas reasoning-oriented systems decline more moderately and humans remain the most stable. General models typically drop steeply from Easy to Hard, indicating that success on simpler items does not reliably transfer to long-horizon, high-constraint geometric reasoning. In contrast, reasoning-oriented systems better preserve performance as complexity grows. These trends suggest that higher difficulty mainly amplifies failure modes tied to partial-cue reliance, while globally consistent reasoners are more robust under increasing constraints.

Intermediate trace analysis. To obtain a finer-grained view of how reasoning-oriented models combine visual evidence with formal deduction, we qualitatively inspect two representative reasoning traces. The cases illustrate two behavioral motifs. In **Spatial-to-Logical Mapping**, a spatial hypothesis is used to narrow the subsequent proof search. In **Algebraic Anchoring**, a coordinate system or algebraic representation is introduced to check an initial visual hypothesis. At the level of the generated rationale, the cases also motivate a descriptive distinction between *mapping errors*, in which an

unsupported or misread visual relation enters the symbolic state, and *computation errors*, in which a later algebraic or deductive step fails despite a plausible initialization. Appendix K presents the two cases in detail. Because this analysis is small-scale, qualitative, and based on model-generated rationales, it should be interpreted as exploratory behavioral evidence rather than a causal account or an exhaustive taxonomy of model errors.

5 Detailed Findings and Analysis

We next elaborate on the three findings highlighted in the Abstract and Introduction as below.

Finding 1: Exact-match fragility under no-guess multi-answer MCQ. Under the no-guess protocol, most general-purpose models fail by committing to an incorrect final option answer, suggesting that the primary bottleneck is the answer consistency rather than partial option identification. We categorize each prediction into four mutually exclusive outcomes: RIGHT_ANSWER (exact match), WRONG_ANSWER (committed but incorrect), NO_ANSWER (uncommitted answer), and OUT_OF_LENGTH (truncation). Figure 4 shows that general-purpose models are largely driven by WRONG_ANSWER, while NO_ANSWER varies widely across systems, reflecting different uncertainty-handling behaviors. In contrast, reasoning-oriented models exhibit higher RIGHT_ANSWER with mild abstention, motivating reporting error-type composition alongside EM.

Additionally, we observe that cross-lingual prompt reformulations heavily impact exact-match

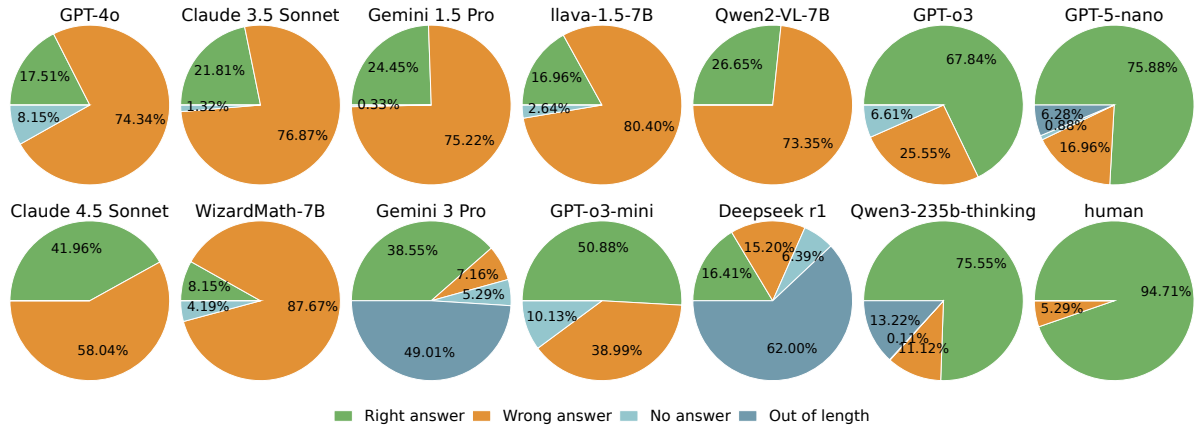


Figure 4: Error Type Distribution Across Model

consistency (see Appendix F for detailed analysis).

Finding 2: Weak and inconsistent diagram grounding. Humans rely heavily on diagrams, whereas current LLMs under-use or inconsistently integrate visual evidence. Removing the diagram causes a 51.88-point EMA drop for humans, while option-level quality remains comparatively strong, suggesting that people can partially reconstruct the figure from text and preserve partial correctness even when exact selection becomes harder. However, the text+image versus text-only changes for most models are substantially smaller and are sometimes reversed. The diagnostic image-only ablation in Appendix G provides a complementary view: GPT-5-nano falls from 75.89 EMA with text+image input to 13.04 with an image-only premise, together with a 49.52-point decrease in F1. Overall, diagram usage emerges as a key capability dimension: humans treat diagrams as the primary substrate for geometry, whereas current LLMs do not reliably ground deductions in visual evidence.

Finding 3: Overextended reasoning without convergence. A non-trivial fraction of failures consists of generations that exhaust the decoding budget and are recorded as OUT_OF_LENGTH. Qualitative inspection suggests that the model often continues to expand, revisit, or revise its intermediate judgments without committing to a stable option set. To separate this behavior from delayed answer emission, we conduct a strict answer-first ablation on DeepSeek-R1(Appendix G). Requiring a parseable answer on the first output line reduces OUT_OF_LENGTH from 62.00% to 55.81%, but exact-match accuracy does not improve, changing from 16.41% to 16.27%; instead, the wrong-

answer and no-answer rates increase. The outcome is therefore sensitive to prompt design, yet the reduction in truncation does not yield more correct final answers. We consequently treat OUT_OF_LENGTH as a deployment-sensitive indicator of practical non-convergence under the evaluated prompting and decoding setup, rather than as an architecture-intrinsic measure of reasoning ability.

Implications for model design. Taken together, the preceding findings suggest three concrete directions for model development. First, the exact-match fragility and the mapping and computation errors observed in the trace analysis motivate **symbol-guided decoding**: a model could maintain an explicit geometric proof state, admit only deductions supported by the established premises and rule inventory, and verify the global consistency of the final option set before termination. Second, the weak and inconsistent use of diagram evidence motivates **visual-text/formal contrastive pre-training** that aligns diagram regions and geometric marks with their natural-language descriptions and formal predicates, using relation-level hard negatives to distinguish visually similar but logically different configurations. Third, the degradation on longer and more complex proofs, together with persistent non-convergence, motivates a **curriculum for complex reasoning processes** that progressively increases proof length, relation-graph density, and the number of interacting constraints, while supervising intermediate verification and explicit termination.

6 Conclusion

In this work, we presented GeoChallenge-90K, a large-scale benchmark for evaluating diagram-grounded geometry theorem proving in a rigorous and scalable setting. Our benchmark is constructed through a symbolic generation-and-verification pipeline, provides aligned text-diagram inputs with bilingual consistency, and exposes controllable complexity for fine-grained analysis. To better reflect real reasoning ability, we further adopt a no-guess, multi-answer multiple-choice protocol that enables strict exact-match evaluation while still allowing complementary option-level diagnostics.

Across comprehensive experiments and analyses, GeoChallenge-90K consistently separates shallow pattern matching from long-step, globally consistent reasoning: current general-purpose models struggle under strict evaluation, reasoning-oriented models improve substantially yet remain behind human solvers, and the gap widens as complexity increases. Our diagnostic studies suggest that the remaining challenges are not merely harder problems, but failures in answer consistency, convergence, and reliable integration of diagram evidence. We hope GeoChallenge-90K will serve as a practical testbed for developing and measuring future systems that reason over long proofs, verify global constraints, and ground deductions in diagrams more reliably.

7 Limitations

Our work has three main limitations. First, although Section 4.2 and Appendix K provide an exploratory intermediate-trace analysis, the analysis is limited to two representative cases and does not quantify how often the proposed Spatial-to-Logical Mapping, Algebraic Anchoring, mapping-error, and computation-error patterns occur across GeoChallenge-small. Second, the text-only and diagnostic image-only ablations characterize behavioral dependence on the available modalities but do not causally identify which visual cues or reasoning operations are decisive. In particular, the image-only condition is a diagnostic rather than a complete alternative formulation of the benchmark, because a rendered diagram need not uniquely encode all metric or relational information contained in the textual premise. Third, NO_ANSWER and OUT_OF_LENGTH are sensitive to prompting, answer-format constraints, decoding, and the maximum token budget. The DeepSeek-R1 answer-first

ablation reduces OUT_OF_LENGTH without improving exact-match accuracy. These categories should be interpreted as deployment-sensitive outcomes under the evaluated setup rather than architecture-intrinsic measures of reasoning ability.

8 Acknowledgements

The 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. 62293554, 62437002, 62450005).

References

- Anthropic. 2024. Claude 3.5 sonnet model card addendum. Technical report, Anthropic.
- Anthropic. 2025. Introducing claude sonnet 4.5. Technical report, Anthropic.
- Patrik Bak. 2020. Geogen. <https://github.com/PatrikBak/GeoGen>. Accessed April 22, 2025.
- Shihao Cai, Keqin Bao, Hangyu Guo, Jizhi Zhang, Jun Song, and Bo Zheng. 2024. Geogpt4v: Towards geometric multi-modal large language models with geometric image generation. *ArXiv preprint*.
- Jie Cao and Jing Xiao. 2022. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520.
- 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.
- Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. 2024. Premise order matters in reasoning with large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.
- Yuri Chervonyi, Trieu H. Trinh, Miroslav Olšák, Xiaomeng Yang, Hoang Nguyen, Marcelo Menegali, Junehyuk Jung, Junsu Kim, Vikas Verma, Quoc V. Le, and Thang Luong. 2025. Gold-medalist performance in solving olympiad geometry with alpageometry2.

- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *ArXiv preprint*.
- Weichen Dai, Rafael Medeiros Cabral, Ziyi Shou, Yan Cao, litingzhe, Hongfei Ai, Jing Yang, Pengcheng Tang, SHEN XIN, and Dongcai Lu. 2025. From symbolic perception to logical deduction: A framework for guiding language models in geometric reasoning.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.
- Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang Wu, Chengquan Zhang, Pengyuan Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, and 1 others. 2024a. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. *ArXiv preprint*.
- Linger Deng, Linghao Zhu, Yuliang Liu, Yu Wang, Qunyi Xie, Jingjing Wu, Gang Zhang, Yingying Zhu, and Xiang Bai. 2024b. Theorem-validated reverse chain-of-thought problem generation for geometric reasoning. *ArXiv preprint*.
- Yuan Feng, Yue Yang, Xiaohan He, Jiatong Zhao, Jianlong Chen, Zijun Chen, Daocheng Fu, Qi Liu, Renqiu Xia, Bo Zhang, and Junchi Yan. 2025. Geobench: Rethinking multimodal geometric problem-solving via hierarchical evaluation.
- Daocheng Fu, Zijun Chen, Renqiu Xia, Qi Liu, Yuan Feng, Hongbin Zhou, Renrui Zhang, Shiyang Feng, Peng Gao, Junchi Yan, and 1 others. 2025a. Trustgeogen: Scalable and formal-verified data engine for trustworthy multi-modal geometric problem solving. *ArXiv preprint*.
- Yumeng Fu, Jiayin Zhu, Lingling Zhang, Bo Zhao, Shaoxuan Ma, Yushun Zhang, Yanrui Wu, and Wenjun Wu. 2025b. Geolaux: A benchmark for evaluating mllms' geometry performance on long-step problems requiring auxiliary lines. *ArXiv preprint*.
- Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wan-jun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and 1 others. 2023. G-llava: Solving geometric problem with multi-modal large language model. *ArXiv preprint*.
- Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, and 1 others. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *ArXiv preprint*.
- Google. 2025. Gemini 3 developer guide. Technical report, Google.
- Yihan Hao, Mingliang Zhang, Fei Yin, and Lin-Lin Huang. 2022. Pgdp5k: A diagram parsing dataset for plane geometry problems. In *2022 26th international conference on pattern recognition (ICPR)*, pages 1763–1769. IEEE.
- Kaivalya Hariharan, Uzay Girit, Atticus Wang, and Jacob Andreas. 2025. Breakpoint: Scalable evaluation of system-level reasoning in llm code agents.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *ArXiv preprint*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *ArXiv preprint*.
- Zihan Huang, Tao Wu, Wang Lin, Shengyu Zhang, Jingyuan Chen, and Fei Wu. 2025. Autogeo: Automating geometric image dataset creation for enhanced geometry understanding. *IEEE Transactions on Multimedia*.
- 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*.
- Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. 2023. Geomverse: A systematic evaluation of large models for geometric reasoning. *ArXiv preprint*.
- Fangyu Lei, Qian Liu, Yiming Huang, Shizhu He, Jun Zhao, and Kang Liu. 2024. S3Eval: A synthetic, scalable, systematic evaluation suite for large language model. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1259–1286.
- Zhuowan Li, Bhavan Jasani, Peng Tang, and Shabnam Ghadar. 2024. Synthesize step-by-step: Tools, templates and llms as data generators for reasoning-based chart VQA. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13613–13623.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of

- foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.
- 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.
- Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *ArXiv preprint*.
- Shixian Luo, Zezhou Zhu, Yu Yuan, Yuncheng Yang, Lianlei Shan, and Yong Wu. 2025. Geogrambench: Benchmarking the geometric program reasoning in modern llms.
- OpenAI. 2025a. Openai o3 and o4-mini system card. Technical report, OpenAI.
- OpenAI. 2025b. Openai o3-mini system card. Technical report, OpenAI.
- Yicheng Pan, Zhenrong Zhang, Pengfei Hu, Jiefeng Ma, Jun Du, Jianshu Zhang, Quan Liu, Jianqing Gao, and Feng Ma. 2025. Enhancing the geometric problem-solving ability of multimodal llms via symbolic-neural integration.
- AAB Prasetyanto, TB Adji, and I Hidayah. 2020. Automatic question generator system conceptual model for mathematic and geometry parallel question replication. In *Journal of Physics: Conference Series*, 1, page 012023. IOP Publishing.
- Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1466–1476.
- Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. 2024. Mm-math: Advancing multimodal math evaluation with process evaluation and fine-grained classification. *ArXiv preprint*.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv preprint*.
- Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, (7995):476–482.
- Xiaofeng Wang, Yiming Wang, Wenhong Zhu, and Rui Wang. 2025. Do large language models truly understand geometric structures? *ArXiv preprint*.
- Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddhartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2024. Livebench: A challenging, contamination-limited llm benchmark.
- Shangyu Xing, Changhao Xiang, Yuteng Han, Yifan Yue, Zhen Wu, Xinyu Liu, Zhangtai Wu, Fei Zhao, and Xinyu Dai. 2024. Gepbench: Evaluating fundamental geometric perception for multimodal large language models. *ArXiv preprint*.
- Cheng Xu, Nan Yan, Shuhao Guan, Changhong Jin, Yuke Mei, Yibing Guo, and Tahar Kechadi. 2025. DCR: Quantifying data contamination in LLMs evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23002–23020.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report.
- Jian Zhang, Zhangqi Wang, Haiping Zhu, Kangda Cheng, Kai He, Bo Li, Qika Lin, Jun Liu, and Erik Cambria. 2025a. Mars: Multi-agent adaptive reasoning with socratic guidance for automated prompt optimization.
- Jian Zhang, Zhiyuan Wang, Zhangqi Wang, Fangzhi Xu, Qika Lin, Lingling Zhang, Rui Mao, Erik Cambria, and Jun Liu. 2025b. Maps: Multi-agent personality shaping for collaborative reasoning.
- Jian Zhang, Bifan Wei, Shihao Qi, haiping Zhu, Jun Liu, and Qika Lin. 2025c. Gkg-llm: A unified framework for generalized knowledge graph construction.
- Jiaxin Zhang, Zhongzhi Li, Mingliang Zhang, Fei Yin, Chenglin Liu, and Yashar Moshfeghi. 2024a. Geoeval: Benchmark for evaluating llms and multi-modal models on geometry problem-solving.
- Mingliang Zhang, Fei Yin, and Cheng-Lin Liu. 2023a. A multi-modal neural geometric solver with textual clauses parsed from diagram. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 3374–3382.

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.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, and 1 others. 2024c. Mavis: Mathematical visual instruction tuning. *arXiv e-prints*, pages arXiv–2407.

Xiaokai Zhang, Na Zhu, Yiming He, Jia Zou, Qike Huang, Xiaoxiao Jin, Yanjun Guo, Chenyang Mao, Yang Li, Zhe Zhu, and 1 others. 2023b. Formalgeo: An extensible formalized framework for olympiad geometric problem solving. *ArXiv preprint*.

Minfeng Zhu, Zi Wang, Sizhe Ji, Zhengtong Du, Junming Ke, Xiao Deng, Zanlang Yin, Xiuqi Huang, Heyu Wang, and Wei Chen. 2025. Genesisgeo: Technical report. *ArXiv preprint*.

A Examples of Clauses, Constructions, and Premises

Clauses *triangle*: This clause created three points to construct a triangle.

midpoint a b: This clause uniquely determines a point based on the positions of points A and B.

angle_bisector a b c: This clause describes the set of points that lie in the angle bisector of $\angle ABC$.

Constructions *a b c = triangle*: This construction created three points to construct a triangle.

x = angle_bisector a b c, on_line a c: This construction uniquely determines the intersection point between the angle bisector of $\angle ABC$ and the line AC.

x = midpoint a b, midpoint a c(invalid): This is an invalid construction, as the midpoints of AB and AC generally do not intersect at a single, well-defined point.

Premise

$a b c = \text{ieq_triangle}(a, b, c);$
 $d = \text{reflect}(d, c, a, b);$
 $e = \text{on_line}(e, d, a), \text{eqdistance}(e, d, a, b);$
 $f = \text{on_dia}(f, b, e), \text{on_circle}(f, a, c).$

This premise captures six point elements and their geometric relationships, which enables the construction of a diagram that matches the description (Figure 5)

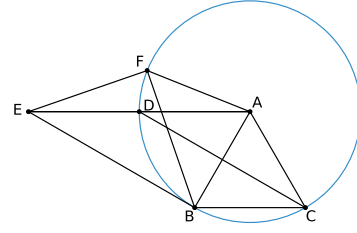


Figure 5: the diagram corresponding to the premise above

Algorithm 1 Conclusion search via forward chaining

Input: Premise

Parameter: max_level N

Output: All derived conclusions

```

1:  $C \leftarrow \text{Read(Premise)}$ 
2:  $level \leftarrow 0$ 
3: while  $level < N$  do
4:    $T \leftarrow \text{match\_theorems}(R)$ 
5:    $T \leftarrow T \cup \text{derive\_algebra}(R)$ 
6:   if  $\text{Not\_empty}(T)$  then
7:      $C \leftarrow C \cup T$ 
8:   else
9:     break
10:  end if
11: end while
12: return  $C$ 

```

B Forward Chaining Algorithm

This section details our forward-chaining conclusion search strategy: starting from the premise, we iteratively match theorems and derive algebraic facts under a maximum depth N , accumulating all newly derived conclusions until no further updates occur.

C Distractor Design

This appendix details the three methods (Keep as True, Synonym Rewriting, and Conflict Rewriting) used to transform correct conclusions derived by the symbolic reasoning engine into candidate options. Table 6 provides illustrative examples of these transformations applied to standard geometric relations.

name	Keep as True	Synonym Rewriting	Conflict Rewriting
midp a b c	midp a b c	nperp a b b c nperp a b a c nperp a c b c	perp a b b c ncoll a b c circle x a b c cyclic a b c x aconst a b c 75 midp b a c midp c a b
para a b c d	para a b c d	nperp a b c d	perp a b c d npara a b c d
circle o a b c	circle o a b c	cong o a o b cong o a o c cong o b o c ncoll a b c npara a b a c npara a b b c npara a c c b	ncong o a o b ncong o a o c ncong o b o c coll a b c eqangle o a o b c a c b cyclic o a b c para a b a c para a b b c para a c c b

Table 6: Examples of applying the three transformation methods to correct conclusions

D Details on Human Verification and Evaluation

D.1 Human Verification Checklist for Dataset Quality

Prior to the baseline evaluation, manual verification was conducted to guarantee data quality. Human annotators (graduate-level students with specialized mathematical backgrounds) audited the dataset using a strict three-point checklist. A figure was retained only if it passed all three criteria:

Readability. Is every character and label in the text clearly visible and legible within the image, without being obscured?

Geometric Validity. Do the points, lines, and surfaces drawn in the image strictly satisfy the preset geometric relationships?

Description alignment. Does the visual representation in the image accurately correspond to the semantic information provided in the description?

D.2 Evaluation Protocol and Objective

The human study was explicitly designed to establish a knowledgeable user baseline rather than an absolute, theoretical upper bound of human capability. To ensure a rigorous evaluation, we implemented two key protocols:

No-Guess Evaluation Protocol. We utilized a strict Exact Match (EM) metric. A participant’s response was scored as correct only if the predicted subset of options exactly matched the ground-truth answer. This strict setting effectively eliminates inflation of human scores through elimination-style guessing or over-selecting likely options.

The 3-Minute Time Limit. Participants were subjected to a strict 3-minute time limit per question. Post-evaluation feedback from our experts indicated that this constraint was rarely binding; most problems were successfully resolved within 90–120 seconds. In geometry problem-solving, if a solver fails to identify the critical auxiliary line or core theorem within the first minute, additional time typically yields diminishing returns.

D.3 Representativeness and Inter-Rater Consistency

To ensure the reliability of the human baseline, we addressed both sample representativeness and scorer consistency within a unified validation pipeline. First, the human evaluation was performed on a 908-question subset selected via stratified sampling from the complete 90K dataset. The final difficulty distribution of this subset consists of 275 easy, 449 medium, and 184 hard questions, yielding a ratio of 30.29% : 49.45% : 20.27% that perfectly aligns with the 3 : 5 : 2 difficulty distri-

Model	EMA↑	EME↑	EMM↑	EMH↑	P↑	R↑	F1↑	HA↑	Avg #Sel
<i>Both text and images provided</i>									
GPT-4o	11.89	19.64	9.13	7.07	40.27	54.66	46.38	45.84	2.25
Claude 3.5 Sonnet	17.62	30.18	13.59	8.70	52.37	64.86	57.95	60.85	2.05
Gemini 1.5 Pro	24.23	38.55	21.16	10.33	62.85	82.98	71.52	71.31	2.12
llava-1.5-7B	12.11	25.09	8.24	2.17	37.00	23.76	28.94	50.61	1.05
Qwen2-VL-7B	24.01	58.91	12.47	0.0	57.16	48.97	52.75	63.13	1.41
<i>Text-only provided</i>									
GPT-4o	9.58	19.27	7.57	0.0	43.45	68.81	53.26	46.72	2.71
Claude 3.5 Sonnet	12.78	23.64	11.36	0.0	48.39	69.37	57.02	57.74	2.34
Gemini 1.5 Pro	23.35	54.91	13.59	0.0	63.42	74.08	68.34	68.72	1.86
WizardMath-7B	7.71	10.18	7.13	5.43	38.34	48.55	42.85	47.58	2.07

Table 7: General-purpose models’ performance on GeoChallenge, where the same problems are presented with Chinese prompts rather than English.

bution of the entire dataset. Second, we observed a high inter-rater agreement rate of approximately 91% (which includes instances where both testers committed simultaneous errors). Qualitative analysis of the remaining 9% disagreement revealed that discrepancies primarily stemmed from intuitive diagram interpretation misjudgments under time pressure, rather than a misunderstanding of the question text or option semantics.

E GeoChallenge Details

The GeoChallenge-90K dataset comprises 90,279 automatically generated geometric proof problems. As detailed in Table 8 of the Appendix E, all problems presented in our dataset are offered with images and solutions of the problem, with both English and Chinese version of the problem provided. the dataset exhibits a stratified difficulty distribution with 27,083 easy-level (30.0%), 45,140 medium-level (50.0%), and 18,056 hard-level (20.0%) problems. This partitioning follows a 3:5:2 ratio based on a priori complexity scores derived from the weighted function shown in main text.

F Cross-lingual robustness under Chinese prompts

We evaluate general-purpose models under Chinese prompts and compare them with their English-prompt counterparts to assess cross-lingual robustness, reporting strict exact-match accuracy (EMA) alongside option-level metrics (F1, HA) and the average number of selected options (Avg #Sel). Across models, EMA is consistently more sensitive to language shifts than option-level metrics, indicating that changes in linguistic formulation can disrupt exact answer consistency even when partial option identification remains similar.

Statistics	
Total Problems	90,279
Coverage	100% / 100% /
(Img / Sol / EN / ZH)	100% / 100%
Difficulty	90,279
* <i>Easy</i>	27,083 (30.0%)
* <i>Medium</i>	45,140 (50.0%)
* <i>Hard</i>	18,056 (20.0%)
Geometry Relations	385,284
* <i>Equality</i>	104,178
* <i>Parallel</i>	15,102
* <i>Vertical</i>	47,902
* <i>Collinear</i>	12,689
* <i>Circle-related</i>	74,579
* <i>Triangle</i>	56,212
* <i>Square</i>	26,180
* <i>Trapezoid</i>	15,725
* <i>Parallelogram</i>	14,731
* <i>Rectangle</i>	12,573
* <i>Other</i>	5,413

Table 8: Detailed statistics of the GeoChallenge dataset. EN and ZH denote English and Chinese statements, respectively.

As shown in table 7, in the multimodal setting (text + images), GPT-4o degrades substantially from English to Chinese (EMA -5.62, F1 -13.43, HA -12.78), and Claude 3.5 Sonnet also drops consistently (EMA -4.19, F1 -4.24, HA -2.37), whereas Gemini 1.5 Pro remains largely stable (EMA -0.22) with slightly improved F1/HA, suggesting stronger cross-lingual robustness. We also observe a recurring over-selection pattern under Chinese prompts: Qwen2-VL-7B shows a smaller EMA decrease (26.65 to 24.01) but selects more options (+0.22 Avg #Sel), and under text-only evaluation GPT-4o exhibits the same mechanism more clearly (EMA -4.41 vs. F1 +4.18 with Avg #Sel +0.57).

Taken together, language-induced over-selection can preserve or even improve F1/HA by increasing coverage of correct options, while reducing EMA because any extra or missing option breaks the exact match.

G Additional Diagnostic Ablations

G.1 Image-only diagnostic

To directly assess how much task-relevant information can be recovered from the rendered diagram without the natural language geometric premise, we evaluate GPT-5-nano under the image-only condition defined in Section 4.1. GPT-5-nano is selected because it is the strongest model in the text+image setting. Its matched text+image result is reported as the primary control, and Qwen2-VL-7B under text+image input is included as a general-purpose multimodal reference.

The results in Table 9 show a marked degradation under image-only input: GPT-5-nano achieves an EM of 13.04% and an F1 score of 34.78%, only marginally above the random baseline (6.67%) and far below its text+image performance (75.89% EM, 84.30% F1). Its average number of selected options is also slightly lower (1.43 vs. 1.61), indicating that the model becomes less reliable rather than more selective when deprived of textual cues. Notably, even this image-only performance remains below the text+image performance of General-Purpose Models like Qwen2-VL-7B (26.65% EM, 53.51% F1).

These results provide further evidence that current models extract little usable geometric information from diagrams alone, and that their visual reasoning remains weak and unstable without textual guidance. This additional ablation therefore reinforces our main conclusion that model performance on GeoChallenge-90K is driven primarily by textual reasoning, while visual evidence is not yet exploited in a robust manner.

G.2 Strict answer-first ablation for DeepSeek-R1

We compare the standard prompt with a stricter answer-first variant. Under the strict variant, the first generated line must contain a parseable final option set and no text may precede it. The remaining reasoning is generated after the answer line.

The experiment in Table 10 shows that forcing the model to answer first reduces the timeout rate only from 62.00% to 55.81%, while accuracy does

not improve (16.41% $\rightarrow$ 16.27%); instead, the error rate and refusal rate increase slightly. This supports our interpretation that OUT_OF_LENGTH is not just a length-budget issue, but a sign that the model still fails to converge to a stable answer even when output order is controlled.

H Predefined Clause Templates

This appendix summarizes the predefined clause templates used in our automatic geometry-problem generation pipeline; the complete list is provided in Table 11.

I Predefined Rules

This appendix lists the predefined inference rules employed by the symbolic reasoning engine for forward deduction; the full rule set is presented in Table 12.

J LLM Prompt Templates

All model evaluations were conducted using carefully designed and standardized system prompts to ensure full reproducibility across different experimental settings. The prompt templates remained consistent throughout all evaluation scenarios, with minor adaptations made only for modality-specific requirements.

For text-only model evaluations, we maintained strict prompt consistency by using identical textual templates across all comparable experiments. Notably, despite the visual nature of some tasks, we intentionally omitted any special instructions regarding missing images in text-only settings. This design choice was based on two key considerations: (1) our problem descriptions maintain high text-image consistency, allowing models to theoretically reconstruct the visual information from textual descriptions alone (though this capability may be challenging for current models), and (2) we observed that models typically wouldn’t refuse to answer due to absent images when provided with sufficiently detailed textual descriptions.

For text+image evaluations, we employed the structured prompt templates illustrated in Figure 6 and 7 for English and Chinese versions respectively. These multimodal prompts systematically incorporated both the textual instructions and visual content, with clear markers distinguishing between image inputs and textual components. The visual examples accompanying each prompt were carefully selected to be representative of the task

Model and input	EMA↑	P↑	R↑	F1↑	HA↑	Avg. #Sel
GPT-5-nano (text+image)	75.89	84.82	83.78	84.30	85.49	1.61
GPT-5-nano (image-only)	13.04	36.36	33.33	34.78	51.09	1.43
Qwen2-VL-7B (text+image)	26.65	61.36	47.44	53.51	65.69	1.19

Table 9: Diagnostic image-only results

Prompt	RIGHT_ANSWER	WRONG_ANSWER	NO_ANSWER	OUT_OF_LENGTH
Standard prompt	16.41	15.20	6.39	62.00
Strict first-line answer	16.27	18.60	9.30	55.81
Change	-0.14	+3.40	+2.91	-6.19

Table 10: DeepSeek-R1 error composition under the standard and strict answer-first prompts

requirements while avoiding potential biases in image content or composition.

K Detailed Case Studies on Intermediate Reasoning Traces

To illustrate the behaviors discussed in Section 4.2, we present two representative intermediate traces from reasoning-oriented models. The cases are selected for explanatory value: they show how a spatial hypothesis may guide proof search and how an algebraic construction may be used to check that hypothesis.

Case Study 1: Spatial-to-Logical Mapping via Symmetry. This case evaluates the model’s ability to perform spatial transformations as an active logical pruning tool.

Problem Context: Given an equilateral triangle ABC with a point D such that $\angle ACD = \angle DBA$, the goal is to prove specific geometric relationships involving point D .

Model Reasoning Trace: Instead of resorting to immediate algebraic manipulation, the model’s intermediate trace shows an immediate invocation of a spatial reflection. The model reasons: Since $\angle ACD = \angle DBA$ and $\triangle ABC$ is equilateral ($AB = AC$), reflecting point D across the median AM (A -altitude) swaps vertices B and C . Hence, we have $\angle ACD = \angle ABD'$, where D' is the reflection of D in AM .

Diagnostic Insight: This trace proves that the model uses high-level visual intuition to dynamically prune the logical search space. It identifies the symmetry that swaps B and C visually before executing formal calculations. However, our fine-grained analysis also reveals a critical limitation: if this initial visual intuition is slightly skewed, the model commits to a flawed symmetry-based path early on, which we identify as a primary driver of spatial hallucination errors.

Case Study 2: Algebraic Anchoring and Error Mitigation. This case examines how the model uses coordinate tracking to cross-check and anchor its visual intuition.

Problem Context: A right isosceles triangle where tangents are drawn to a circle centered at the right-angle vertex A .

Model Reasoning Trace: Rather than treating the image descriptions as static text, the model proactively establishes a Cartesian frame to verify its visual perception. The trace shows a conscious initialization step: Let us place the triangle in a rectangular coordinate system with $A(0, 0)$, $B(1, 0)$, and $C(0, 1)$. The locus $y = x$ strictly gives the angle bisector AD ...

Diagnostic Insight: While traditional MLLMs suffer because they decouple the visual diagram from the symbolic proof, reasoning-oriented architectures use this coordinate system as an internal feedback loop to “double-check” geometric locations. Through this, we can separate failures into Mapping Errors (disconnect between image and coordinate initialization) versus Computation Errors (algebraic slips within the correct coordinate frame). Error propagation typically occurs when the model fails to resolve the explicit contradictions between its rigid coordinate calculations and its initial visual guess.

L More Problems in GeoChallenge Dataset

Figures 8, 9, and 10 illustrate representative examples of easy, medium, and hard problems, respectively. A clear progression in complexity is observable across these difficulty levels: the textual descriptions grow significantly longer, incorporating more intricate logical constraints and nuanced phrasing, while the accompanying images exhibit a marked increase in structural sophistication, evidenced by the rising number of geometric

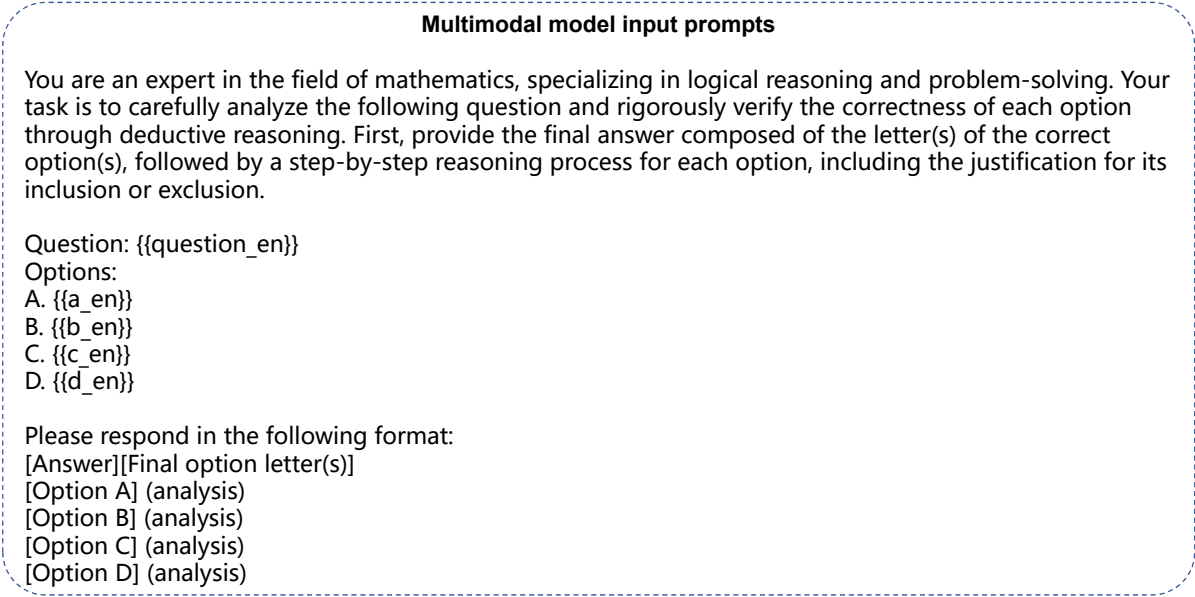


Figure 6: English version of prompt to LLMs

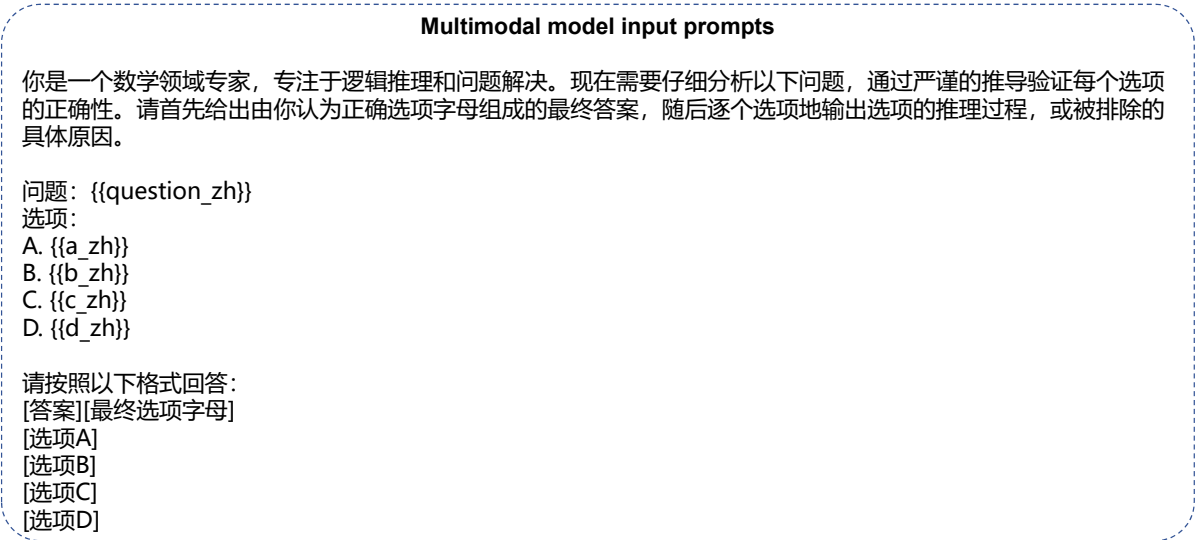


Figure 7: Chinese version of prompt to LLMs

points, connecting lines, and layered annotations. Notably, the easy-level problems typically involve straightforward deductions with minimal intermediate reasoning steps, whereas medium and hard problems demand deeper theorem applications, often requiring multi-hop inference chains and careful consideration of implicit spatial relationships. This escalation in cognitive demand aligns closely with human intuition—the harder problems not only present more visual clutter but also necessitate greater mental effort in parsing, planning, and executing solutions. The deliberate stratification of difficulty ensures that the benchmark captures a wide spectrum of reasoning capabilities, from basic

pattern recognition to advanced geometric theorem synthesis, mirroring the gradual skill development observed in human problem-solving. Furthermore, the consistency between objective complexity metrics (e.g., token count, graph density) and subjective human assessment underscores the validity of our difficulty calibration methodology.

"Question_en": "Given a right trapezoid ABCD with $AB \parallel CD$, and $\angle A$ as the right angle. G is a free point. Through point B, two tangents BI and BJ are drawn to the circle $\odot A$ with radius AG, with the point of tangency at I and J, respectively.",

"A_en": "AD is not perpendicular to CD",

"B_en": " $AI:BI = AJ:BJ$ ",

"C_en": "AD is parallel to IJ",

"D_en": "A is the circumcenter of $\triangle GIJ$ ",

"Answer": "BCD",

"Prior_Difficulty": 59.0,

"Posterior_Difficulty": 0.75,

"SimplifiedFormalText": "a b c d = r trapezoid a b c d; g = free g; i j = tangent i j b a g",

"UsedPoints": ["a", "b", "c", "d", "g", "i", "j"],

"premises": ["AB $\parallel$ CD [00]", "AB $\perp$ AD [01]", "AI = AG [02]", "BI $\perp$ AI [03]", "AJ = AG [04]", "BJ $\perp$ AJ [05]"],

"ProofSteps_A": "001. AB $\perp$ AD [00] & AB $\parallel$ CD [01] $\Rightarrow$ CD $\perp$ AD\n",

"ProofSteps_B": "001. AI = AG [00] & AJ = AG [02] $\Rightarrow$ A is the circumcenter of $\triangle JGI$ [04]\n002. A is the circumcenter of $\triangle JGI$ [04] & BI $\perp$ AI [01] $\Rightarrow$ $\angle JIB = \angle JGI$ [05]\n003. A is the circumcenter of $\triangle JGI$ [04] & BJ $\perp$ AJ [03] $\Rightarrow$ $\angle BJI = \angle JGI$ [06]\n004. $\angle JIB = \angle JGI$ [05] & $\angle BJI = \angle JGI$ [06] $\Rightarrow$ $\angle BJI = \angle JIB$ [07]\n005. $\angle BJI = \angle JIB$ [07] $\Rightarrow$ IB = JB [08]\n006. AI = AG [00] & AJ = AG [02] & IB = JB [08] $\Rightarrow$ IA:IB = JA:JB\n",

"ProofSteps_C": "001. AJ = AG [04] & AI = AG [02] $\Rightarrow$ A is the circumcenter of $\triangle JGI$ [06]\n002. A is the circumcenter of $\triangle JGI$ [06] & BI $\perp$ AI [03] $\Rightarrow$ $\angle JIB = \angle JGI$ [07]\n003. A is the circumcenter of $\triangle JGI$ [06] & BJ $\perp$ AJ [05] $\Rightarrow$ $\angle BJI = \angle JGI$ [08]\n004. BI $\perp$ AI [03] & BJ $\perp$ AJ [05] $\Rightarrow$ $\angle BJA = \angle BIA$ [09]\n005. $\angle BJA = \angle BIA$ [09] $\Rightarrow$ A, J, B, I are concyclic [10]\n006. A, J, B, I are concyclic [10] $\Rightarrow$ $\angle BJI = \angle BAI$ [11]\n007. AB $\perp$ AD [01] & BI $\perp$ AI [03] $\Rightarrow$ $\angle BAI = \angle(AD-BI)$ [12]\n008. $\angle JIB = \angle JGI$ [07] & $\angle BJI = \angle JGI$ [08] & $\angle BJI = \angle BAI$ [11] & AB $\parallel$ CD [00] & $\angle BAI = \angle(AD-BI)$ [12] $\Rightarrow$ $\angle(AD-BI) = \angle JIB$ [13]\n009. $\angle(AD-BI) = \angle JIB$ [13] $\Rightarrow$ AD $\parallel$ IJ\n",

"ProofSteps_D": "001. AJ = AG [01] & AI = AG [00] $\Rightarrow$ A is the circumcenter of $\triangle JGI$ \n"

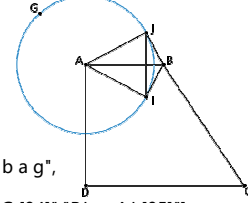


Figure 8: Example of easy level problem in GeoChallenge

"Question_en": "Given square ABCD. Construct the nine-point circle $\odot H$ of $\triangle CBA$, intersecting BA, CA, CB at E, F, G, respectively. $\angle GFI = 45^\circ$, IC=GA. Line FI intersects the perpendicular bisector of EF at K.",

"A_en": " $\angle(AH-BK) \neq \angle GDE$ ",

"B_en": " $AE:AK = DG:DK$ ",

"C_en": "BK is perpendicular to CH",

"D_en": "EH is not perpendicular to EK",

"Answer": "BC",

"Prior_Difficulty": 176.0,

"Posterior_Difficulty": 0,

"SimplifiedFormalText": "a b c d = isquare a b c d; e f g h = ninepoints

e f g h c b a; i = s_angle g f i 45, eqdistance i c g a; k = on_line k f i, on_bline k e f",

"UsedPoints": ["a", "b", "c", "d", "e", "f", "g", "h", "i", "k"],

"premises": ["AB = BC [00]", "AB $\perp$ BC [01]", "E, A, B are collinear [02]", "EB = EA [03]", "C, F, A are collinear [04]", "FA = FC [05]", "C, G, B are collinear [06]", "GC = GB [07]", "HE = HF [08]", "HF = HG [09]", " $\angle IFG = 1$ Pi/4 [10]", "KE = KF [11]", "F, I, K are collinear [12]", " $\angle KEF = \angle EFK$ [13]"],

"ProofSteps_A": "001. AB $\perp$ BC [00] & BC $\parallel$ AD [02] & AB $\parallel$ CD [03] $\Rightarrow$ $\angle ADC = \angle ABC$ [19]\n002. AB $\perp$ BC [00] & BC $\parallel$ AD [02] & AB $\parallel$ CD [03] $\Rightarrow$ AD $\perp$ DC [20]\n003. $\angle ADC = \angle ABC$ [19] $\Rightarrow$ C, D, A, B are concyclic [21]\n004. AB $\parallel$ CD [03] & BC $\parallel$ AD [02] $\Rightarrow$ $\angle CDA = \angle DCB$ [22]\n005. AB $\parallel$ CD [03] & BC $\parallel$ AD [02] $\Rightarrow$ $\angle ADC = \angle CBA$ [23]\n006. C, D, A, B are concyclic [21] & $\angle CDA = \angle DCB$ [22] $\Rightarrow$ CA = DB [24]\n

.....
112. $\angle BKG = \angle EDG$ [129] & $\angle EHA = \angle HGK$ [127] & H, G, E are collinear [103] $\Rightarrow$ $\angle(BK-AH) = \angle EDG$ \n",

"ProofSteps_B": ".....",

"ProofSteps_C": ".....",

"ProofSteps_D": "....."

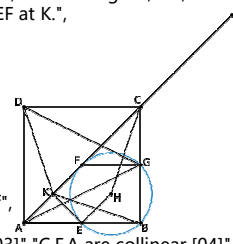


Figure 9: Example of medium level problem in GeoChallenge

name	meaning
$X = \text{angle_bisector}(A,B,C)$	Construct point X on the angle bisector of $\angle ABC$
$X = \text{angle_mirror}(A,B,C)$	Construct point X such that BC is the bisector of $\angle ABX$
$X,Y,Z,I = \text{centroid}(A,B,C)$	Construct point I as the centroid of $\triangle ABC$ with midpoints X, Y, Z
$X = \text{circle}(A,B,C)$	Construct point X as the circumcenter of $\triangle ABC$
$A,B,C,D = \text{eq_quadrangle}()$	Construct quadrilateral ABCD with $AD = BC$
$A,B,C,D = \text{eq_trapezoid}()$	Construct trapezoid ABCD with $AD = BC$
$X = \text{eq_triangle}(B,C)$	Construct point X such that $\triangle XBC$ is an equilateral triangle
$X = \text{eqangle2}(A,B,C)$	Construct point X such that $\angle BAX = \angle XCB$
$X = \text{eqangle3}(A,B,D,E,F)$	Construct point X such that $\angle AXB = \angle EDF$
$A,B,C,D = \text{eqdia_quadrangle}()$	Construct quadrilateral ABCD with $AC = BD$
$X = \text{eqdistance}(A,B,C)$	Construct point X such that $XA = BC$
$X = \text{excenter}(A,B,C)$	Construct point X as the excenter of $\triangle ABC$
$X = \text{foot}(A,B,C)$	Construct point X as the foot of A on BC
$A,B,C = \text{ieq_triangle}()$	Construct equilateral triangle $\triangle ABC$
$X = \text{incenter}(A,B,C)$	Construct point X as the incenter of $\triangle ABC$
$X = \text{intersect}(f,g)$	Construct point X as function $f() \cap g()$
$X = \text{intersection_lc}(A,O,B)$	Construct point X as circle O $\cap$ line AB
$X = \text{intersection_ll}(A,B,C,D)$	Construct point X as line AB $\cap$ line CD
$X = \text{intersection_lp}(A,B,C,M,N)$	Construct point X as line AB $\cap$ line through C parallel to line MN
$X = \text{intersection_lt}(A,B,C,D,E)$	Construct point X as line AB $\cap$ line through C perpendicular to line DE
$X = \text{intersection_pp}(A,B,C,D,E,F)$	Construct point X such that line XA $\parallel$ line BC and line XD $\parallel$ line EF
$X = \text{intersection_tt}(A,B,C,D,E,F)$	Construct point X such that line XA $\perp$ line BC and line XD $\perp$ line EF
$A,B,C = \text{iso_triangle}()$	Construct A, B, C such that $AB = AC$
$A,B,C,D = \text{isquare}()$	Construct square ABCD
$X = \text{lc_tangent}(A,O)$	Construct point X such that OA is perpendicular to AX
$X = \text{midpoint}(A,B)$	Construct point X as the midpoint of AB
$X = \text{mirror}(A,B)$	Construct point X such that B is the midpoint of AX
$X,Y,Z,I = \text{ninepoints}(A,B,C)$	Construct midpoints X, Y, Z and point I as the circumcenter of $\triangle XYZ$
$X = \text{nsquare}(A,B)$	Construct point X such that XB is the right isosceles triangle
$X = \text{on_aline}(A,B,C,D,E)$	Construct point X such that $\angle XAB = \angle CDE$
$X = \text{on_bline}(X,A,B)$	Construct point X on the perpendicular bisector of AB
$X = \text{on_circle}(O,A)$	Construct point X such that $OA = OX$
$X = \text{on_circum}(A,B,C)$	Construct point X on the circumcircle of A, B, C
$X = \text{on_dia}(A,B)$	Construct point X such that AX is perpendicular to BX
$X = \text{on_line}(A,B)$	Construct point X on line AB
$X = \text{on_pline}(A,B,C)$	Construct point X such that XA is parallel to BC
$X = \text{on_tline}(A,B,C)$	Construct point X such that XA is perpendicular to BC
$X = \text{orthocenter}(A,B,C)$	Construct point X as the orthocenter of ABC
$X = \text{parallelogram}(A,B,C)$	Construct point X such that ABCX is a parallelogram
$A,B,C,D = \text{r_trapezoid}()$	Construct right trapezoid ABCD
$A,B,C = \text{r_triangle}()$	Construct right triangle ABC
$A,B,C,D = \text{rectangle}()$	Construct rectangle ABCD
$X = \text{reflect}(A,B,C)$	Construct point X as the reflection of A about BC
$A,B,C = \text{risos}()$	Construct point X as the isosceles triangle ABC
$X = \text{s_angle}(A,B,\alpha)$	Construct point X such that $\angle ABX = \alpha$
$A,B = \text{segment}()$	Construct segment AB
$X = \text{shift}(B,C,D)$	Construct point X such that $XB=CD$ and $XC=BD$
$X,Y = \text{square}(A,B)$	Construct point X, Y such that XYAB is a square
$X,Y = \text{tangent}(A,O,B)$	Construct point X,Y as the tangent touch points from A to circle (O,B)
$A,B,C,D = \text{trapezoid}()$	Construct trapezoid ABCD
$A,B,C = \text{triangle}()$	Construct triangle ABC

Table 11: Predefined Clause Templates for Generating Geometric Problems

premise(s)	conclusion
perp A B C D, perp C D E F, ncoll A B E	para A B E F
cong O A O B, cong O B O C, cong O C O D	cyclic A B C D
eqangle A B P Q C D P Q	para A B C D
cyclic A B P Q	eqangle P A P B Q A Q B
eqangle6 P A P B Q A Q B, ncoll P Q A B	cyclic A B P Q
cyclic A B C P Q R, eqangle C A C B R P R Q	cong A B P Q
midp E A B, midp F A C	para E F B C
midp E A B, midp F A C, midp G B C	cong E F G B
midp E A B, midp F A C, midp G B C	cong E F G C
para A B C D, coll O A C, coll O B D	eqratio3 A B C D O O
perp A B C D, perp E F G H, npara A B E F	eqangle A B E F C D G H
eqangle A B C D M N P Q, eqangle C D E F P Q R U	eqangle A B E F M N R U
eqratio A B C D M N P Q, eqratio C D E F P Q R U	eqratio A B E F M N R U
eqratio6 D B D C A B A C, Coll D B C, ncoll A B C	eqangle6 A B A D A D A C
eqangle6 A B A D A D A C, Coll D B C, ncoll A B C	eqratio6 D B D C A B A C
cong O A O B, ncoll O A B	eqangle O A A B A B O B
eqangle6 A O A B B A B O, ncoll O A B	cong O A O B
circle O A B C, perp O A A X	eqangle A X A B C A C B
circle O A B C, eqangle A X A B C A C B	perp O A A X
circle O A B C, midp M B C	eqangle A B A C O B O M
circle O A B C, coll M B C,	midp M B C
eqangle A B A C O B O M	
perp A B B C, midp M A C	cong A M B M
circle O A B C, coll O A C	perp A B B C
cyclic A B C D, para A B C D	eqangle A D C D C D C B
midp M A B, perp O M A B	cong O A O B
cong A P B P, cong A Q B Q	perp A B P Q
cong A P B P, cong A Q B Q, cyclic A B P Q	perp P A A Q
midp M A B, midp M C D	para A C B D
midp M A B, para A C B D, para A D B C	midp M C D
eqratio O A A C O B B D, coll O A C,	
coll O B D, ncoll A B C, sameside A O C B O D	para A B C D
para A B A C	coll A B C
midp M A B, midp N C D	eqratio M A A B N C C D
eqangle A B P Q C D U V, perp P Q U V	perp A B C D
eqratio A B P Q C D U V, cong P Q U V	cong A B C D

Table 12: Rules Used by the Symbolic Reasoning Engine for Deriving New Conclusions

"Question_en": "Given square ABCD. Construct the nine-point circle $\odot H$ of $\triangle CDB$, intersecting DB, CB, CD at E, F, G, respectively. Through point F, draw a line parallel to BH, and this line intersects circle determined by points F, D, and C at J. M is the midpoint of DJ. N is the midpoint of AJ.",

"A_en": "CN $\neq$ JN",

"B_en": " $\angle AMF \neq \angle GND$ ",

"C_en": "AM:HN $\neq$ HM:GM",

"D_en": "AM is parallel to GN",

"Answer": "D",

"Prior_Difficulty": 413.5,

"Posterior_Difficulty": 0,

"SimplifiedFormalText": "a b c d = isquare a b c d; e f g h = ninepoints e f g h c d b; j = on_circum j f d c, on_pline j f b h; m = midpoint m d j; n = midpoint n a j",

"UsedPoints": ["a", "b", "c", "d", "e", "f", "g", "h", "j", "m", "n"],

"premises": ["AB = BC [00]", "AB $\perp$ BC [01]", "AB $\parallel$ CD [02]", "AD $\parallel$ BC [03]", "E, D, B are collinear [04]", "ED = EB [05]", "F, B, C are collinear [06]", "FB = FC [07]", "G, D, C are collinear [08]", "GC = GD [09]", "HE = HF [10]", "HF = HG [11]", "J, F, D, C are concyclic [12]", "JF $\parallel$ BH [13]", "J, M, D are collinear [14]", "MD = MJ [15]", "A, J, N are collinear [16]", "NJ = NA [17]"],

"ProofSteps_A": "001. FC = FB [07] & OD = OA [21] $\Rightarrow$ FC:FB = OD:OA [22]\n002. AB $\parallel$ CD [02] & F, B, C are collinear [06] & A, O, D are collinear [20] & FC:FB = OD:OA [22] $\Rightarrow$ FO $\parallel$ CD [23]\n003. F, B, C are collinear [06] & A, O, D are collinear [20] & AB $\perp$ BC [00] & AB $\parallel$ CD [02] & FO $\parallel$ CD [23] & BC $\parallel$ AD [03] $\Rightarrow$ $\angle FCD = \angle FOD$ [24]\n.....247. $\angle (JO-ED) = \angle (JC-OG)$ [54] & E, D, B are collinear [04] & GO $\perp$ HG [67] & BD $\perp$ AC [50] & GO $\parallel$ AC [52] & OG $\parallel$ JO [143] & BE $\perp$ GO [80] & H, F, G are collinear [92] & BD $\parallel$ FG [63] & $\angle (CO-EI) = \angle (AC-GI)$ [118] & CO $\parallel$ EI [116] & $\angle (CO-EI) = \angle GOI$ [156] & $\angle BDG = \angle NEO$ [263] & G, D, C are collinear [08] & E, F, O are collinear [36] & EF $\parallel$ CD [34] & $\angle OGM = \angle AEN$ [266] & A, E, C are collinear [61] & GM $\parallel$ CJ [267] $\Rightarrow$ JC $\perp$ CA [268]\n248. JC $\perp$ CA [268] & N is midpoint of AJ [37] $\Rightarrow$ JN = CN\n",

"ProofSteps_B": ".....",

"ProofSteps_C": ".....",

"ProofSteps_D": "....."

Figure 10: Example of hard level problem in GeoChallenge