UniTutorBench: Comprehensive Benchmarking of Explanation and Scaffolding in MLLM Tutoring

Anonymous ACL submission

Abstract

Large multimodal language models are increasingly deployed as intelligent tutors, yet existing evaluations remain fragmented: they either rely on exam-style correctness tests or isolate narrow tutoring behaviors, largely overlooking instructional process and its align with critical learning sciences principles such as Cognitive Load Theory. To address this, we propose a unified hierarchical taxonomy that organizes tutoring abilities across scenarios, tasks, and fine-grained ability dimensions. Crucially, our taxonomy expands the evaluation boundary by incorporating dimensions specifically designed to assess instruction process quality, such as step effectiveness and solution standardization. Building on this framework, we introduce UNITUTORBENCH, a K–12 mathematics tutoring benchmark comprising 2,187 instances with fine-grained annotations across both explanation and scaffolding scenarios. Experiments on 11 state-of-the-art models reveal a significant disparity between explanation and scaffolding performance, as gemini-2.5-flash with best performance in explanation degrees 16% in scaffolding. We believe UniTutorbench will advance community to build next-generation pedagogical MLLMs.

1 Introduction

Recently, large language models (LLMs) have made significant progress in complex tasks such as mathematical reasoning and code generation(Zheng et al., 2023; Weitekamp et al., 2025; Wang et al., 2025a; Jiang et al., 2024). Advances in multimodal alignment have further enabled high-quality human–model interaction(Huang and Zhang, 2024). As a result, multimodal LLMs are evolving from standalone problem solvers into intelligent tutors, showing strong potential in educational settings(Rouzegar and Makrehchi, 2024; Ammari et al., 2025; Handa et al., 2025). This shift, however, also raises new challenges for evaluation(Xu et al., 2024).

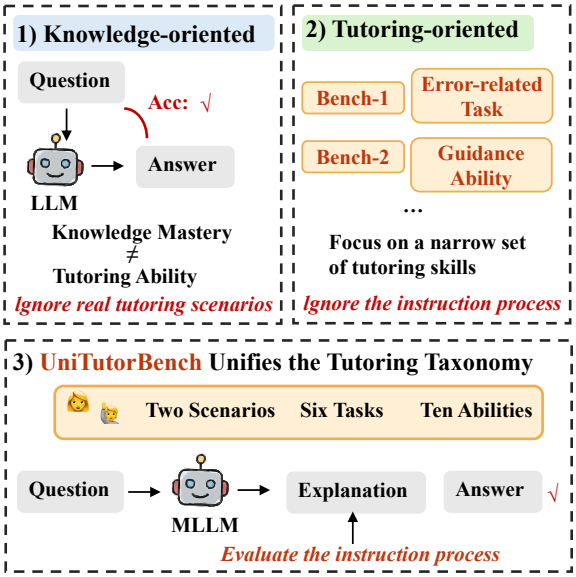


Figure 1: Comparison between existing educational benchmarks and our proposed UniTutorBench.

In response, an increasing number of studies have begun to explore how to evaluate models in educational settings(Xu et al., 2025). Existing educational benchmarks broadly fall into two categories: Knowledge-oriented benchmarks and Tutoring-oriented benchmarks. Figure 1 summarizes the key differences.

(1) Knowledge-oriented benchmarks equate educational capability with domain knowledge mastery, evaluating models through multiple-choice or question–answering tasks that emphasize answer correctness. However, such benchmarks largely overlook the tutoring process, limiting their alignment with real-world tutoring scenarios.

(2) Tutoring-oriented benchmarks focus on specific tutor skills, such as error feedback, dialogue-based tutoring, and Socratic questioning, and are therefore closer to authentic educational practice. Nevertheless, they still lack a unified evaluation taxonomy.

Most existing tutoring benchmarks evaluate only

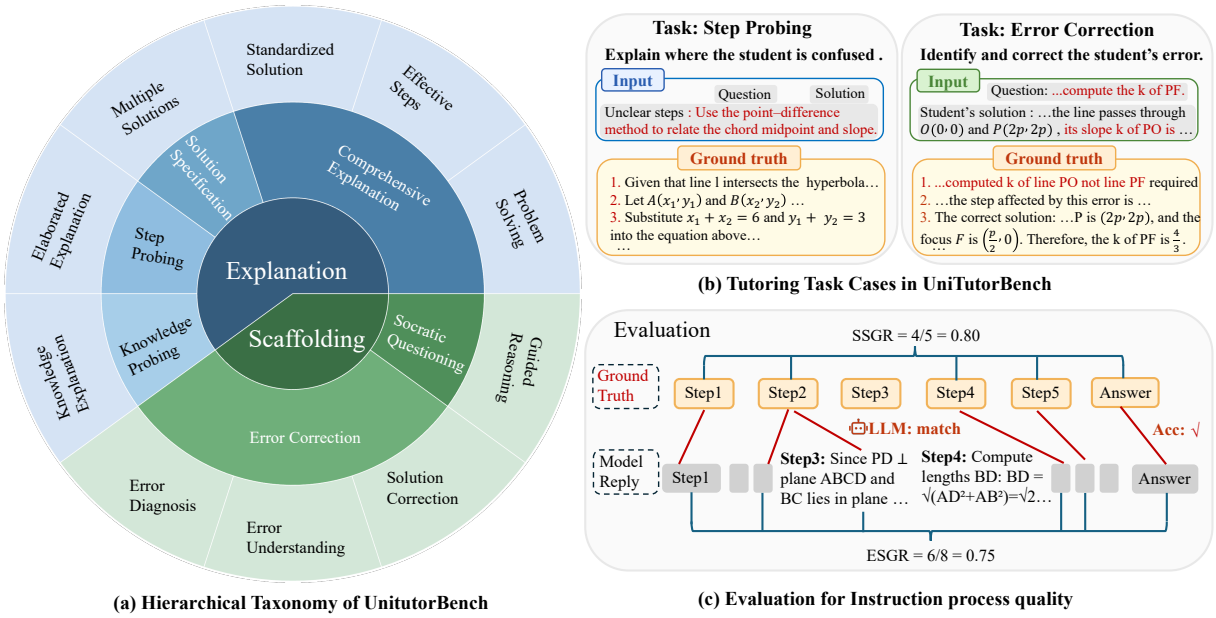


Figure 2: **Overview of UniTutorBench:** (a) Our unified hierarchical taxonomy for tutoring evaluation, organized into two pedagogical scenarios (*explanation* and *scaffolding*), six core tutoring tasks, and ten fine-grained ability dimensions. (b) Representative task cases from each scenario, illustrating model inputs and expert-annotated golden references. (c) An example of process-oriented evaluation for instructional process quality.

a limited subset of tutoring skills and adopt heterogeneous evaluation units. Some benchmarks are organized around concrete tutoring tasks, while others quantify performance from an ability-centric perspective, despite targeting largely overlapping tutoring skills. Thus, benchmark-level performance is often conflated with overall tutoring capability, hindering a coherent, field-level understanding of tutoring abilities.

In addition, existing benchmarks rarely evaluate instructional process quality. Instead, they primarily assess whether a tutoring task is completed, rather than how instruction is delivered. Grounded in learning sciences, effective instruction depends critically on how information is structured, sequenced, and presented to manage learners' cognitive load(Mayer and Moreno, 2003; Kalyuga et al., 1999). Such process-oriented qualities are inherently difficult to capture with outcome-based metrics, leading most benchmarks to implicitly bypass them.

To tackle the aforementioned challenges, we propose a unified hierarchical evaluation taxonomy. As illustrated in Figure 2(a), grounded in learning science principles, we define ten core evaluation dimensions at Tier 3, including that relevant to instructional process quality. We then summarize tutoring tasks from prior work and consolidate them into six core tutoring tasks at Tier 2, which are systematically mapped to the proposed evaluation dimensions. Based on the primary role responsible for problem solving, we further categorize tutoring scenarios into explanation and scaffolding at Tier 1. As shown in Table 1, existing benchmarks exhibit limited and uneven coverage under our taxonomy, leaving a clear gap in comprehensive tutoring evaluation.

To bridge this gap, we introduce UNITUTORBENCH, a new tutoring benchmark for K–12 high school mathematics tutoring. Built upon the proposed taxonomy, UNITUTORBENCH covers six core tutoring tasks and consists of 2,187 instances, each annotated with fine-grained, task-specific labels to support comprehensive evaluation of tutoring abilities. To align with the evaluation goal of instructional process quality, we design step-level metrics, shown in Figure 2(c). We evaluate 11 representative state-of-the-art models on UNITUTORBENCH. Experiments reveal substantial performance discrepancies across distinct tutoring scenarios. Notably, models that excel in the explanation scenario often exhibit marked performance degradation when required to perform scaffolding. For example, Gemini-2.5-Flash, which achieves the best performance in explanation, drops by 16% in scaffolding.These findings highlight a significant gap between current multimodal LLMs and the requirements for general-purpose intelligent tutoring

Ability Dimension	MR	MM	TB	Ours	Task Dimension	MT	EB	Ours
Problem Solving	✗	✗	✓	✓	Comprehensive Explanation	✗	✓	✓
Effective Steps	✗	✗	✗	✓	Solution Specification	✗	✗	✓
Standardized Solution	✗	✗	✗	✓	Step Probing	✗	✗	✓
Multiple Solutions	✗	✗	✓	✓	Knowledge Probing	✗	✓	✓
Elaborated Explanation	✗	✓	✓	✓	Socratic Questionin	✓	✗	✓
Knowledge Explanation	✗	✗	✓	✓	Error Correction	✓	✓	✓
Guided Reasoning	✓	✗	✓	✓				
Error Diagnosis	✓	✗	✓	✓				
Error Understanding	✗	✗	✗	✓				
Solution Correction	✗	✗	✗	✓				

Table 1: Comparison with existing benchmarks designed for pedagogical scenarios. *MR* denotes MRBench(Maurya et al., 2025); *MM*-MMtutorBench(Yang et al., 2025); *TB*-TutorBench(Srinivasa et al., 2025); *MT*-MathTutorBench(Macina et al., 2025); *EB*-EduBench(Xu et al., 2025).

in educational settings.

Overall, our main contributions are summarized as follows:

(1) Unified Taxonomy. We propose a systematic evaluation taxonomy that hierarchically organizes tutoring skills across scenarios, tasks, and ability dimensions. This unified taxonomy extends the scope of benchmarking to include instructional process quality.

(2) Comprehensive Benchmark. We introduce UNITUTORBENCH, a benchmark of 2,187 fine-grained, annotated instances built upon our taxonomy, featuring step-level annotations and process-oriented evaluation metrics to support comprehensive assessment of tutoring abilities.

(3) Model Evaluation. Through extensive experiments, we show that different tutoring skills rely on distinct underlying abilities, and provide empirical insights into the strengths and limitations of current LLM-based tutors.

2 Related Work

2.1 Knowledge-oritented benchmarks

Early benchmarks largely equate educational capabilities with subject knowledge acquisition and problem-solving outcomes. In English settings, MMLU evaluates world knowledge and general problem solving via large-scale multiple-choice questions (Hendrycks et al., 2021a), while MATH, GSM8K, and MathVista focus on mathematical reasoning performance (Hendrycks et al., 2021b; Cobbe et al., 2021; Lu et al., 2024). In Chinese contexts, C-Eval and CMMLU extend the exam-style paradigm using materials from university courses and professional examinations (Huang et al., 2023; Li et al., 2024). For K–12 education, GAOKAO-Bench and E-Eval rely on standardized tests, homework, and mock exams to assess multi-subject knowledge mastery (Zhang et al., 2023; Hou et al., 2024), while CK12 evaluates knowledge coverage through K–12 knowledge graphs (You et al., 2024). Recently, OmniEduBench introduces a dual-dimensional framework of knowledge and cultivation to incorporate educational objectives (Zhang et al., 2025), while EduEval proposes an EduAbility Taxonomy to evaluate multiple pedagogical tasks (Ma et al., 2025). However, their evaluations remain primarily question–answer driven.

Overall, these benchmarks are predominantly exam-style and provide limited coverage of tutoring behaviors such as explanation, guidance, and feedback in authentic pedagogical scenarios.

2.2 Tutoring-oriented benchmarks

More recent benchmarks have begun to explore tutoring capabilities. MRBench evaluates tutoring skills such as error correction and providing guidance from an ability-centric perspective (Maurya et al., 2025), while MathTutorBench assesses similar skills from a task-centric perspective (Macina et al., 2025). EduBench further introduces a bilingual evaluation framework that broadens coverage across educational scenarios, including explanation, error correction, and guidance (Xu et al., 2025). These benchmarks are primarily developed in text-only settings. In the multimodal domain, MMTutorBench introduces expert-designed rubrics to assess tutors' ability to explain students' solution steps (Yang et al., 2025). TutorBench adopts sample-level rubrics to evaluate multiple tutoring skills across three use cases: Feedback and Assessment, Active Learning Support and Adaptive Explanation Generation (Srinivasa et al., 2025).

Ability	Abbr.	Description	Metric
Explanation Scenario			
Problem Solving	PS	Produces correct final answers through explanation.	Acc.
Effective Steps	ES	Covers essential steps without redundancy.	ESGR
Standardized Solution	SS	Follows commonly accepted pedagogical solution structures.	SSCR
Multiple Solutions	MS	Provides multiple valid solution paths to broaden perspectives.	Acc.
Elaborated Explanation	EE	Elaborates unclear key steps with detailed explanations.	SSCR
Knowledge Explanation	KE	Explains concepts in a way that is understandable to students.	rubric
Scaffolding Scenario			
Guided Reasoning	GR	Guides students step by step to reach correct conclusions.	S-Acc.
Error Diagnosis	ED	Identifies errors and their root causes in student solutions.	rubric
Error Understanding	EU	Understands how errors affect the solution process.	rubric
Solution Correction	SC	Corrects errors and helps continue the solution process.	Acc.

Table 2: Tutoring abilities in our UNITUTORBENCH and their corresponding evaluation metrics. Abilities associated with instructional process quality include Effective Steps and Standardized Solution.

Despite these advances, most benchmarks assess only a subset of tutoring skills rather than holistic tutoring capability. They lack a unified and comprehensive tutoring taxonomy, which hinders systematic benchmarking of tutoring ability.

3 Learning Sciences Principles

Grounded in established theories from the learning sciences, we draw on core principles including cognitive load theory(Mayer and Moreno, 2003; Kalyuga et al., 1999), Scaffolding and the Zone of Proximal Development(Reiser, 2018; McLeod, 2012), as well as theories of worked-example explanations(Saw, 2017; Chi et al., 1989).

Building on these foundations, we define that effective tutoring is as instruction in which information is appropriately structured, sequenced, and presented to manage learners' cognitive load. Tutoring is inherently a process-oriented activity, and thus its evaluation should focus on the quality of the instructional process rather than solely on task completion or final outcomes.

To enable a comprehensive evaluation of tutoring ability, we design ten core abilities spanning different instructional scenarios. Detailed descriptions of these abilities are provided in Table 2.

4 UniTutorBench

4.1 Taxonomy

Unlike prior efforts that evaluate individual tutoring skills in isolation, we synthesize a comprehensive set of tutoring tasks from existing studies. Based on the primary role responsible for problem-solving, we organize these tasks into two instructional scenarios—*explanation* and *scaffolding*—forming a unified taxonomy that integrates both tutoring tasks and underlying abilities. Figure 2(a) illustrates the hierarchical structure of UNITUTORBENCH, linking pedagogical scenarios, tutoring tasks, and core ability dimensions. Tasks refer to concrete tutoring activities or interaction formats, while abilities refer to the underlying tutoring competencies assessed across tasks. Detailed ability information and task definitions are provided in Appendix A and Appendix B, respectively.

The explanation scenario tests the ability to present problem-solving processes and relevant knowledge in a clear and pedagogically structured manner.In this scenario, the tutor is responsible for completing the problem. Tasks under this scenario include *Comprehensive Explanation*, *Solution Specification*, *Step Probing*, and *Knowledge Probing*.

The scaffolding scenario emphasizes instructional support during active problem solving.In this scenario, the student is responsible for completing the problem, while the tutor provides interactive guidance. Rather than directly providing complete solutions, it evaluates whether a model can guide, regulate, and diagnose learners' reasoning through interactive support. Tasks under this scenario include *Socratic Questioning* and *Error Correction*.

4.2 Evaluation Metrics

As shown in Table 2, we define a corresponding evaluation metric for each tutoring ability. Within our evaluation framework, these metrics fall into

Statistic	Count (%)
Total Questions	2,187
Question Modality	
With Image Questions	378 (17.3%)
Without Image Questions	1809 (82.7%)
Difficulty Levels	
Easy	268 (12.3%)
Rather Easy	178 (8.1%)
Moderate	673 (30.8%)
Rather Hard	727 (33.2%)
Hard	341 (15.6%)
Tasks	
Comprehensive Explanation	747 (34.2%)
Solution Specification	140 (6.4%)
Step Probing	179 (8.2%)
Knowledge Probing	174 (8.0%)
Socratic Questioning	747 (34.2%)
Error Correction	200 (9.1%)

Table 3: Key statistics of the dataset, including question modality, difficulty levels, and task types.

three categories: (1) conventional automatic metrics, such as accuracy; (2) rubric-based LLM judge metrics, including pass/fail rates or point-based scoring; (3) hybrid metrics that combine LLM-assisted judgments with automatic metric computation.

In this section, we primarily introduce the third category, including ESGR and SSCR. Appendix C shows all metrics for more detailed informations.

Effective Step Generation Rate (ESGR) The Effective Step Generation Rate (ESGR) measures the Effective Steps ability.

For the i-th problem, let $\mathcal{M}_i = \{m_i^1, \ldots, m_i^{|\mathcal{M}_i|}\}$ denote the set of steps generated by the model, where $|\mathcal{M}_i|$ is the total number of model-generated steps. Let $\mathcal{S}_i = \{s_i^1, \ldots, s_i^{|\mathcal{S}_i|}\}$ denote the expert-annotated set of standardized steps, where $|\mathcal{S}_i|$ denotes the total number of standardized steps. Each step corresponds to a minimal, atomic, and semantically meaningful mathematical operation.

We define an alignment relation $\mathcal{A}_i \subseteq \mathcal{S}_i \times \mathcal{M}_i$. For a standardized atomic step $s \in \mathcal{S}_i$ and a model-generated step $m \in \mathcal{M}_i$, $(s, m) \in \mathcal{A}_i$ indicates that the model step m correctly covers (i.e., completes) the atomic standardized step s. Note that the alignment is many-to-one: a single model step m may align to multiple standardized steps s. The set of aligned model steps is defined as

$$\mathcal{M}_i^{\text{aligned}} = \{m \in \mathcal{M}_i \mid \exists s \in \mathcal{S}_i,\ (s, m) \in \mathcal{A}_i\}. \tag{1}$$

The ESGR for the i-th problem is then defined as

$$\text{ESGR}_i = \frac{\left|\mathcal{M}_i^{\text{aligned}}\right|}{|\mathcal{M}_i|}. \tag{2}$$

At the dataset level, ESGR is computed as the average over all N instances:

$$\text{ESGR} = \frac{1}{N} \sum_{i=1}^{N} \text{ESGR}_i. \tag{3}$$

Standardized Step Coverage Rate (SSCR) The Standardized Step Coverage Rate (SSCR) measures the Standardized Solution ability. Unlike ESGR, SSCR does not account for the number or redundancy of generated steps; instead, it evaluates the model's explanation holistically, focusing on whether key steps in the reference solution are covered.

For the i-th problem, let the expert-annotated standardized step sequence be $\mathcal{S}_i = \{s_i^1, s_i^2, \ldots, s_i^{|\mathcal{S}_i|}\}$, where $|\mathcal{S}_i|$ denotes the total number of standardized steps. Let the model-generated explanation be denoted as R_i. If the explanation contains content that is semantically or functionally equivalent to a standardized step $s \in \mathcal{S}_i$, that step is considered covered. Accordingly, the set of covered standardized steps is defined as

$$\mathcal{S}_i^{\text{covered}} = \{s \in \mathcal{S}_i \mid s \text{ is covered by } R_i\}. \tag{4}$$

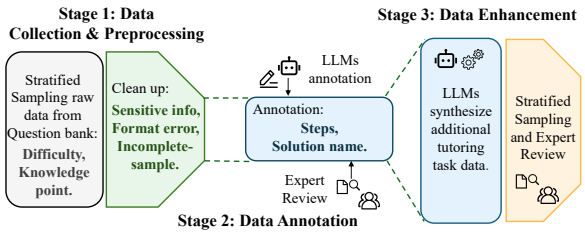


Figure 3: The three-stage data curation pipeline of UniTutorBench. Stages 2 and 3 involve LLM-assisted data processing, with human experts conducting quality control and verification.

The SSCR for the i-th problem is defined as

$$\text{SSCR}_i = \frac{\left|\mathcal{S}_i^{\text{covered}}\right|}{|\mathcal{S}_i|}. \tag{5}$$

At the dataset level, SSCR is computed as the average over all N instances:

$$\mathrm{SSCR} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{SSCR}_i. \tag{6}$$

4.3 Dataset Construction

As illustrated in Figure 3, our dataset construction pipeline consists of three stages: *Data Collection & Preprocessing*, *Data Annotation*, and *Data Enhancement*.

Data Collection & Preprocessing. Our dataset is constructed from the question bank of a well-established K–12 tutoring organization in China, with sampling stratified by both knowledge categories and difficulty levels. We first conduct systematic data cleaning on an initial collection of 957 high-school mathematics problems, including removing sensitive information, correcting formatting issues, and excluding samples with incomplete problem statements or solutions to ensure consistency and usability. After cleaning, 817 problems are retained.

Data Annotation. Based on the cleaned data, we perform fine-grained annotations. Following expert-designed rubrics, we employ Qwen-max-latest for text-only problems and Qwen3-VL-235B-A22B for image-based problems to automatically segment solution traces into step-wise annotations. For problems with multiple valid solutions, we split the solutions and annotate solution names accordingly. All automatically generated annotations are subsequently reviewed and verified by human educational experts to ensure quality. This process results in 817 curated instances.

Data Enhancement. We further use Gemini-3 and GPT-5.2 to synthesize additional data for different tutoring tasks under expert-defined rubrics. The synthesized samples are filtered and sampled according to predefined rules and target distributions to retain high-quality and informative instances. Human experts then conduct spot checks to ensure quality control. More detailed information about the data construction process is provided in Appendix D.

4.4 Dataset Detail

As shown in Table 3, UNITUTORBENCH contains 2,187 instances of K–12 high school mathematics questions, including 378 multimodal problems with images (17.3%) and 1,809 text-only problems (82.7%). The dataset spans five difficulty levels—Easy (12.3%), Rather Easy (8.1%), Moderate (30.8%), Rather Hard (33.2%), and Hard (15.6%)—providing balanced coverage across complexity. The benchmark supports six core tutoring tasks, including Comprehensive Explanation (34.2%), Solution Specification (6.4%), Step Probing (8.2%), Knowledge Probing (8.0%), Socratic Questioning (34.2%), and Error Correction (9.1%).

5 Experiments

5.1 Experimental Setup

We adopt GPT-5.1(OpenAI, 2025) as an automatic judge for evaluating model outputs on UniTutorBench. Detailed evaluation protocols and prompt templates are provided in the appendix E. We evaluate a total of 11 representative multimodal large language models on the proposed benchmark, including 4 closed-source models and 7 open-source models. The selected models include Qwen3-VL-plus(Bai et al., 2025), Gemini-2.5-flash(Comanici et al., 2025), Claude-haiku-4-5(Anthropic, 2025), GPT-4o(Achiam et al., 2023), Qwen3-VL Series(Bai et al., 2025), InternVL3.5 Series(Wang et al., 2025b), and Gemma-3-27b(Team et al., 2025). Detailed information for all models is provided in Appendix E.

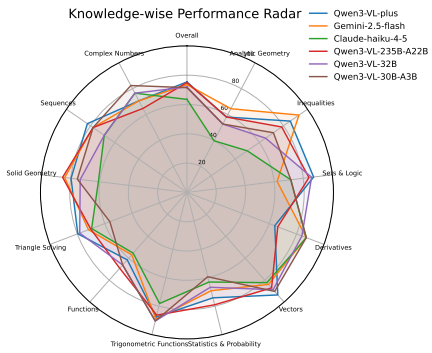
5.2 Main Results

Performance differs substantially between explanation and scaffolding. Models exhibit distinct performance patterns across explanation and scaffolding scenarios, indicating a clear decoupling of abilities. For instance, Gemini-2.5-Flash performs well(Avg:75.7) in explanation but lags in scaffolding(Avg:59.7), whereas Qwen3-VL-plus achieves moderate explanation performance yet ranks among the top three across all scaffolding dimensions. Notably, *Guided Reasoning* remains a consistent weakness across models. These results suggest that the two scenarios require different abilities, and performance in one does not reliably reflect overall tutoring competence.

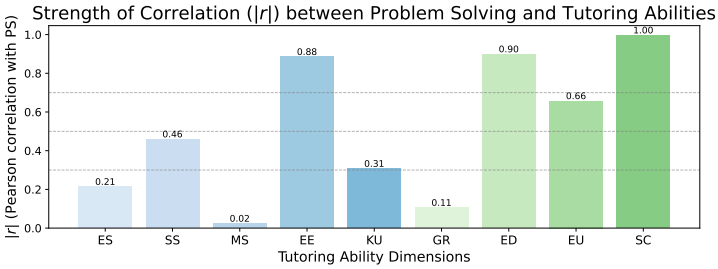
The gap between closed-source and open-source models is limited. The overall performance gap between closed- and open-source models is small. Qwen3-VL-235B-A22B achieves strong and stable results across both scenarios, comparable to state-of-the-art closed-source models. Moreover, smaller Qwen models (30B/32B) outperform some

Model	Explanation							Scaffolding				
	Avg.	PS	ES	SS	MS	EE	KE	Avg.	GR	ED	EU	SC
Closed-source												
Qwen3-VL-plus	69.2	**75.4**	76.3	74.1	50.0	57.5	81.9	62.5	32.8	**81.5**	70.5	**65.0**
Gemini-2.5-flash	**75.7**	73.6	**83.8**	**76.7**	72.1	**63.3**	84.9	59.7	28.4	76.5	70.2	63.5
Claude-haiku-4-5	67.4	63.5	75.8	72.9	63.6	44.7	83.9	54.8	32.5	67.5	65.7	53.5
GPT-4o	48.3	27.2	73.4	50.6	40.0	21.7	76.9	27.8	26.0	25.0	27.0	33.0
Open-source												
Qwen3-VL-235B-A22B	74.5	75.0	76.2	75.0	**76.4**	61.4	83.0	**63.1**	33.2	81.0	**73.7**	64.5
Qwen3-VL-32B	72.0	72.0	74.4	72.8	73.6	54.2	**85.0**	57.3	32.0	71.5	64.3	61.5
Qwen3-VL-30B-A3B	71.2	71.5	72.4	71.7	74.3	52.4	84.7	59.9	**34.3**	74.5	68.7	62.0
InternVL3.5-30B-A3B	52.2	35.6	73.0	58.0	52.1	24.3	70.4	29.7	29.7	29.5	26.7	33.0
InternVL3.5-38B	53.1	34.8	78.6	53.8	50.7	26.4	71.9	28.3	28.1	31.5	23.7	30.0
InternVL3.5-20B-A4B	45.6	25.7	77.9	48.5	33.6	17.9	69.7	24.4	28.8	25.5	22.8	20.5
Gemma-3-27B	51.6	39.1	65.2	56.6	50.0	24.9	73.5	27.0	26.2	23.5	24.2	34.0

Table 4: Main results on UniTutorBench. The highest score among closed-source models and the highest score among open-source models are highlighted in blue and green, respectively. The overall best score across all models is shown in bold. *Avg.* denotes the average score of abilities within each scenario.



(a) Domain Knowledge Mastery



(b) Correlation between problem solving and other tutoring ability dimensions

Figure 4: Relationship between knowledge mastery and tutoring abilities. (a) Radar chart showing models' balanced performance across knowledge domains. (b) Bar chart of Pearson correlations between knowledge mastery and tutoring ability.

closed-source models across multiple dimensions, with Qwen3-VL-30B-A3B achieving the highest *Guided Reasoning* score (34.3). These results indicate that well-optimized open-source models can match proprietary systems on educational tasks.

Model scaling yields non-linear gains. Tutoring performance does not scale linearly with model size. Within the Qwen family, gains from scaling beyond 30B/32B are limited, with larger models improving only a few dimensions. In contrast, scaling within the InternVL family is inconsistent, as InternVL3.5-38B underperforms the 30B model on most dimensions. Overall, InternVL and Gemma-3-27B trail Qwen models, indicating that architecture design and instruction alignment matter more.

5.3 Relationship Between Knowledge Mastery and Tutoring Abilities

We use problem-solving performance (PS) as a proxy for knowledge mastery, as success on problem-solving tasks is widely viewed as evidence of domain competence. We restrict our analysis to the six models with overall scores above 50 on UniTutorBench, since substantially weaker models are unlikely to be adopted in real tutoring settings. We then examine the distribution of knowledge mastery and its relationship with tutoring abilities. Key findings are summarized below.

Knowledge mastery is balanced across topics and is no longer the main bottleneck. Overall, as shown in the figure 4(a), the models exhibit relatively balanced performance across knowledge dimensions, with no evident systematic weaknesses.

While minor variations exist in specific areas such as analytic geometry or function-related topics, most models maintain consistently strong performance in core high-school mathematics domains. These results suggest that modern multimodal models largely cover the required knowledge space.

Most tutoring abilities correlate weakly with knowledge mastery. As shown in Figure 4(b), We compute Pearson correlations between PS and each of the other tutoring ability dimensions. The results show that several key tutoring abilities correlate weakly with PS, including ES(0.21), SS(0.46), MS(0.02), KU(0.31), and GR(0.11). This aligns with intuition: strong problem-solving ability does not automatically translate into effective tutoring behaviors or the capability to foster student learning. In other words, tutoring requires additional skills beyond possessing knowledge.

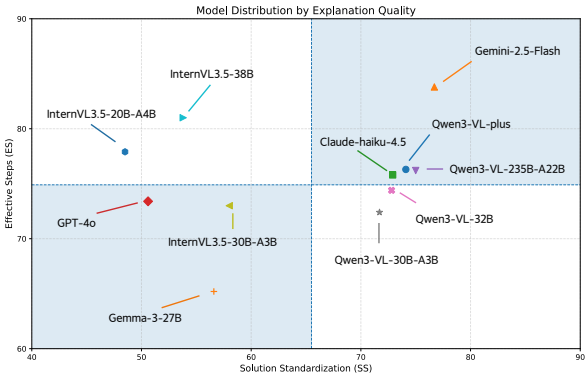


Figure 5: Quadrant-based analysis of model explanation quality measured by **Standardized Solution (SS)** and **Effective Steps (ES)**.

5.4 Analysis of Instructional process quality

As shown in Figure 5, models with high SS and high ES best align with the requirements of explanation scenarios. While models with low SS and low ES produce fragmented or redundant explanations with limited instructional value. Models with high SS but low ES often reach correct answers through repeated correction of intermediate errors, indicating self-correction ability. However, such trial-and-error explanations may increase cognitive load and undermine learner trust in instructional settings. In contrast, models with high ES but low SS produce well-structured reasoning chains based on incorrect premises, showing high confidence in flawed solutions. While their explanations appear organized, deviations in key reasoning paths risk conveying incorrect domain understanding.

Annotators / Judge	ESGR	SSCR	ACC	ED/EU	KE
Human inter-rater	0.85	0.82	0.95	0.85	0.74
GPT-5.1	0.82	0.81	0.93	0.83	0.70
Gemini-2.5-Flash	0.81	0.78	0.92	0.82	0.74
Kimi-K2.5	0.68	0.73	0.92	0.76	0.69
Qwen3-VL-235B	0.67	0.66	0.92	0.72	0.65

Table 5: Inter-human and human–judge agreement. KE uses ICC(A,1); all other columns use Fleiss' kappa. Higher values indicate stronger agreement.

5.5 LLM Reliability Analysis

To rigorously validate the reliability of our evaluation framework, we conduct human agreement analysis and cross-judge robustness analysis. The cross-judge robustness results are detail in Appendix F.1.

Human–judge agreement. We randomly sample 50 responses for each judge-dependent metric from outputs of different models. Three trained human annotators, all with master's degrees, independently annotate each sample after calibration. For Knowledge Explanation (KE), which is scored on a continuous scale, we report ICC(A,1) (McGraw and Wong, 1996); for the remaining binary or pass/fail metrics, we report Fleiss' kappa (Scott, 1955). We further evaluate several leading models as judges, including Kimi-K2.5, which was not used in either data construction or model inference. As shown in Table 5, human annotators reach substantial agreement across metrics, indicating that the rubrics are sufficiently operationalized. Overall, GPT-5.1 is the strongest aligned judge among the evaluated alternatives, with agreement scores close to human inter-rater agreement. Therefore, we use GPT-5.1 as our primary judge in experiments.

6 Conclusion

We introduce a unified taxonomy that organizes tutoring skills across scenarios, tasks, and ability dimensions. Based on this framework, we present UniTutorBench to enable a comprehensive evaluation of tutoring capabilities, including instruction process quality. Through extensive experimentation, we analyze the limitations of current MLLMs in diverse tutoring contexts. Our results reveal that most MLLMs exhibit performance misalignments across different scenarios. This study shows that even state-of-the-art MLLMs are not yet adequate as AI tutors, leaving substantial room for improvement. We hope our taxonomy provides a foundational guide for future research, while our benchmark offers a roadmap for the community.

Limitations

Despite the comprehensive nature of UNITUTORBENCH and its novel instruction process evaluation framework, there are still minor limitations that point to directions we will address in future iterations:

Multimodal tutoring scope. UNITUTORBENCH is not intended as a specialized benchmark for visual pedagogy. The benchmark includes both text-only and image-based questions, with 378 multimodal items (17.3%), because real K–12 mathematics tutoring often involves mixed text-and-image problem inputs. In this work, multimodality primarily improves ecological validity and input completeness rather than defining a separate visual-only tutoring taxonomy. As a result, diagram-grounded tutoring abilities, such as interpreting spatial relations, suggesting auxiliary constructions, or providing visual scaffolding, are not yet separately evaluated as independent ability dimensions. In future versions, we will extend UNITUTORBENCH with targeted tasks for diagram-grounded and visually interactive tutoring.

Student-centered interaction. Although UNITUTORBENCH separates explanation from scaffolding and includes student-need-driven tasks such as Step Probing and Knowledge Probing, its evaluation is still based on static benchmark instances and simulated tutoring contexts. This design enables controlled and reproducible assessment, but it cannot fully capture long-horizon student adaptation, evolving learner states, or real-time multi-turn tutoring dynamics. Future work will evaluate model tutoring effectiveness with real students in authentic learning settings.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tawfiq Ammari, Meilun Chen, SM Zaman, and Kiran Garimella. 2025. How students (really) use chatgpt: Uncovering experiences among undergraduate students. *arXiv preprint arXiv:2505.24126*.

Anthropic. 2025. Introducing claude haiku 4.5.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Michelene TH Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science*, 13(2):145–182.

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 arXiv:2110.14168*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Kunal Handa, Drew Bent, Alex Tamkin, Miles McCain, Esin Durmus, Michael Stern, Mike Schiraldi, Saffron Huang, Stuart Ritchie, Steven Syverud, and 1 others. 2025. Anthropic education report: How university students use claude.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Jinchang Hou, Chang Ao, Haihong Wu, Xiangtao Kong, Zhigang Zheng, Daijia Tang, Chengming Li, Xiping Hu, Ruifeng Xu, Shiwen Ni, and Min Yang. 2024. E-EVAL: A comprehensive Chinese k-12 education evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7753–7774.

Jiaxing Huang and Jingyi Zhang. 2024. A survey on evaluation of multimodal large language models. *arXiv preprint arXiv:2408.15769*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, jiayi lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*.

Slava Kalyuga, Paul Chandler, and John Sweller. 1999. Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 13(4):351–371.

Stuart A. Karabenick and John R. Knapp. 1991. Relationship of academic help seeking to the use of learning strategies and other instrumental achievement behavior in college students. *Journal of Educational Psychology*. Defines instrumental vs executive help seeking.

Maurice G Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285.

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*.

Guoqing Ma, Jia Zhu, Hanghui Guo, Weijie Shi, Yue Cui, Jiawei Shen, Zilong Li, and Yidan Liang. 2025. Edueval: A hierarchical cognitive benchmark for evaluating large language models in chinese education. *arXiv preprint arXiv:2512.00290*.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. MathTutorBench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 204–221.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251.

Richard E. Mayer. 2002. Multimedia learning. In *Psychology of Learning and Motivation*, volume 41, pages 85–139. Academic Press.

Richard E Mayer and Roxana Moreno. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, 38(1):43–52.

Kenneth O McGraw and Seok P Wong. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30.

SA McLeod. 2012. Zone of proximal development.

OpenAI. 2025. Gpt-5.1: A smarter, more conversational chatgpt.

Brian J Reiser. 2018. Scaffolding complex learning: The mechanisms of structuring and problematizing student work. In *Scaffolding*, pages 273–304. Psychology Press.

Hamdireza Rouzegar and Masoud Makrehchi. 2024. Generative ai for enhancing active learning in education: A comparative study of gpt-3.5 and gpt-4 in crafting customized test questions. *arXiv preprint arXiv:2406.13903*.

Kim Guan Saw. 2017. Cognitive load theory and the use of worked examples as an instructional strategy in physics for distance learners: A preliminary study. *Turkish Online Journal of Distance Education*, 18(4):142–159.

William A Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325.

Rakshith S Srinivasa, Zora Che, Chen Bo Calvin Zhang, Diego Mares, Ernesto Hernandez, Jayeon Park, Dean Lee, Guillermo Mangialardi, Charmaine Ng, Ed-Yeremai Hernandez Cardona, and 1 others. 2025. Tutorbench: A benchmark to assess tutoring capabilities of large language models. *arXiv preprint arXiv:2510.02663*.

Jon R. Star, Dimitri Tuomela, Nuria Joglar-Prieto, Peter Hästö, Riikka Palkki, Miguel Á. Abánades, Johanna Pejlare, R. H. Jiang, Lijia Li, and Ru-De Liu. 2022. Exploring students' procedural flexibility in three countries. *International Journal of STEM Education*, 9(1):4.

John Sweller and Graham A. Cooper. 1985. The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, 2(1):59–89.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Peng-Yuan Wang, Tian-Shuo Liu, Chenyang Wang, Ziniu Li, Yidi Wang, Shu Yan, Chengxing Jia, Xu-Hui Liu, Xinwei Chen, Jiacheng Xu, and 1 others. 2025a. A survey on large language models for mathematical reasoning. *ACM Computing Surveys*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Daniel Weitekamp, Momin N. Siddiqui, and Christopher J. MacLellan. 2025. Tutorgym: A testbed for evaluating ai agents as tutors and students. In *International Conference on Artificial Intelligence in Education*, pages 361–376.

Bin Xu, Yu Bai, Huashan Sun, Yiguan Lin, Siming Liu, Xinyue Liang, Yaolin Li, Zhuangzhi Dong, Jingren Zhang, Yufan Deng, and 1 others. 2025. Edubench: A comprehensive benchmarking dataset for evaluating large language models in diverse educational scenarios. *arXiv preprint arXiv:2505.16160*.

Hanyi Xu, Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Philip S Yu. 2024. Large language models for education: A survey. *arXiv preprint arXiv:2405.13001*.

Tengchao Yang, Sichen Guo, Mengzhao Jia, Jiaming Su, Yuanyang Liu, Zhihan Zhang, and Meng Jiang. 2025. Mmtutorbench: The first multimodal benchmark for ai math tutoring. *arXiv preprint arXiv:2510.23477*.

Weihao You, Pengcheng Wang, Changlong Li, Zhilong Ji, and Jinfeng Bai. 2024. Ck12: A rounded k12 knowledge graph based benchmark for chinese holistic cognition evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19431–19439.

Min Zhang, Hao Chen, Wenqi Zhang, Didi Zhu, Xin Lin, Bo Jiang, Aimin Zhou, Fei Wu, and Kun Kuang. 2025. Omniedubench: A comprehensive chinese benchmark for evaluating large language models in education. *arXiv preprint arXiv:2510.26422*.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Ability Design

Grounded in principles from the learning sciences and informed by prior research, we define a comprehensive set of tutoring abilities spanning instructional scenarios. The ability definitions are presented below.

Problem Solving measures whether the model produces the correct final answer to a math problem (i.e., answer correctness). This aligns with the Problem Solving evaluation in MathTutorBench(Macina et al., 2025).

Effective Steps operationalizes quality as the alignability between generated steps and the reference solution, capturing the relevance and informational value of step-by-step tutoring output. Educationally, it targets the reduction of extraneous cognitive load: redundant or off-topic steps consume working-memory resources and impair processing of core reasoning(Mayer, 2002)

Standardized Solution measures whether the model includes the key reasoning steps needed to complete the reference solution's logical chain. Missing or skipped steps shift the burden of inference to learners, increasing intrinsic load and error risk and weakening schema acquisition. This is consistent with the worked example effect shown by Sweller and Cooper (1985)(Sweller and Cooper, 1985).

Multiple Solutions captures the fact that many math problems admit more than one valid solution strategy, and effective instruction may require solving the same problem using a specified method. This is educationally motivated by research on students' procedural flexibility—the ability to generate and select among different procedures depending on goals and constraints—as discussed by Star et al(Star et al., 2022).

Elaborated Explanation reflects cases where students request more detailed explanations beyond a brief solution. It evaluates whether the model can fill in omitted key steps with traceable, step-by-step reasoning, This design is consistent with MMTutorBench(Yang et al., 2025), which emphasizes stepwise guidance and explanation centered on pedagogically critical steps.

Knowledge Explanation evaluates whether the model can provide a clear, structured explanation when a student is confused about a specific concept, supporting understanding and transfer. This aligns with the evaluation focus of TutorBench(Yang et al., 2025).

Guided Reasoning evaluates whether, when a student is stuck, the model provides prompts that support continued independent problem solving rather than directly giving the answer. This aligns with the concept of instrumental help seeking in educational psychology, which links adaptive help seeking to the use of learning strategies and other instrumental achievement behaviors (Karabenick and Knapp, 1991).

Error Diagnosis, Error Understanding and Solution Correction capture the core remediation pipeline in educational tutoring: when a student's solution is flawed, the tutor should (i) diagnose the error by identifying what is wrong, (ii) explain the error by clarifying its underlying misconception or reasoning failure, and (iii) correct the solution by providing a repaired method or revised steps that enable the student to fix their work. This metric is motivated by error-centered tutoring evaluations in MathTutorBench(Macina et al., 2025) and MRBench(Maurya et al., 2025), which emphasize assessing a tutor's ability to recognize, interpret, and remediate student errors rather than merely supplying a final answer.

Theoretical Mapping. Our taxonomy should be understood as a learning-science-guided synthesis rather than a deductive derivation from a single learning theory. We use learning-science principles as design constraints and consolidate recurring tutoring functions from prior benchmarks and educational literature, with particular attention to process-oriented functions that are underexplored in existing evaluations. Instructional process quality and cognitive-load management motivate abilities that assess whether reasoning is structured, non-redundant, and sufficiently explicit, corresponding to *Effective Steps* (ES), *Standardized Solution* (SS), and *Elaborated Explanation* (EE). Worked-example and conceptual explanation research motivate *Multiple Solutions* (MS) and *Knowledge Explanation* (KE), which capture flexibility across solution methods and concept-level explanation. Scaffolding, the Zone of Proximal Development, and instrumental help seeking motivate *Guided Reasoning* (GR), where the tutor supports the learner's independent progress rather than directly replacing it. Error-centered remediation mo-

tivates *Error Diagnosis* (ED), *Error Understanding* (EU), and *Solution Correction* (SC), corresponding to identifying, interpreting, and repairing student mistakes. Finally, task completion and correctness grounding motivate *Problem Solving* (PS) as the outcome-level anchor for the taxonomy.

B Tasks Design

B.1 Explanation Scenario

The explanation scenario aims to evaluate a model's ability to present problem-solving processes to learners in an accurate, well-structured, and pedagogically meaningful manner. Under this scenario, we design four tasks.

Comprehensive Explanation This task simulates a teacher providing a complete explanation. Given a problem Q_i (including images, if any), the model is required to generate a explanation process and arrive at the correct answer A_i.

To characterize explanation completeness, we represent the reference solution as an ordered set of steps

$$\mathcal{S}_i = \{s_i^1, s_i^2, \dots, s_i^K\},$$

where each s_i^k denotes a minimal and indispensable reasoning unit under the given problem context. This task primarily evaluates the model's *Problem Solving* (PS), *Effective Steps* (ES), and *Standardized Solution* (SS) abilities.

Solution Specification In this task, in addition to the problem Q_i, a solution constraint r_i is provided to explicitly specify the strategy or method that the model should follow.

The model is required to complete the reasoning and explanation in accordance with the specified solution and produce the correct answer A_i. This task focuses on model's ability *Multiple Solutions* (MS) ability.

Step Probing This task addresses cases where learners struggle with intermediate reasoning steps. Given Q_i and a high-level step h_i, the model generates a fine-grained reasoning process that is semantically consistent with h_i. Specifically, h_i is expanded into an ordered sequence

$$\mathcal{S}_i^h = \{s_i^{h,1}, s_i^{h,2}, \dots, s_i^{h,L}\},$$

where each $s_i^{h,l}$ denotes a minimal semantic unit supporting the high-level step. This task evaluates *Elaborated Explanation* (EE).

Knowledge Probing This task targets cases where learners lack understanding of relevant Knowledge. Given a concept k_i and problem context Q_i, the model generates a learner-oriented explanation grounded in the reference solution R_i, evaluating *Knowledge Explanation* (KE).

B.2 Scaffolding Scenario

The scaffolding scenario focuses on whether a model can provide process-oriented support to help learners continuously advance their reasoning while the problem-solving process is still ongoing. Unlike the explanation scenario, scaffolding does not aim to directly present a complete solution. Instead, it emphasizes the regulation and guidance of the learner's reasoning process.

Socratic Questioning This task targets instructional situations in which teachers guide students to construct solution strategies through heuristic questioning. Given Q_i, the model is required to generate a sequence of guiding questions

$$\{q_i^1, q_i^2, \dots\},$$

that progressively lead the learner toward the correct solution.

This task primarily evaluates the model's ability corresponding to the *Guided Reasoning* (GR) abilities.

Error Correction This task targets scenarios in which learners make mistakes during the problem-solving process. Given a problem Q_i and a student solution $\mathcal{U}_i$ that contains errors, the model is required to generate a response that: (1) identifies and analyzes the location and type of errors (*Error Diagnosis*); (2) explains their underlying causes and potential impact on subsequent reasoning (*Error Understanding*); and (3) provides a corrected solution process (*Solution Correction*).

Accordingly, this task primarily evaluates the model's *Error Diagnosis* (ED), *Error Understanding* (EU), and *Solution Correction* (SC) abilities.

C Evaluation Metrics

We assign a dedicated evaluation metric to each ability dimension. Within our evaluation framework, the metrics fall into three categories: (1) conventional automatic metrics; (2) rubric-based *LLM-as-a-judge* metrics, including pass/fail decisions and point-based scoring; and (3) hybrid metrics that combine LLM-assisted judgments with automatic computations.

(1) Conventional automatic metrics. We adopt Answer Accuracy (ACC) as the primary conventional automatic metric. ACC measures whether a model's output matches the reference answer, and is used to evaluate abilities where correctness is directly verifiable, including PROBLEM SOLVING, MULTIPLE SOLUTIONS, SOLUTION CORRECTION, and GUIDED REASONING. Notably, in GUIDED REASONING, the tutor model is required to provide guidance rather than directly output the final answer. Therefore, we evaluate guidance quality through the learner's final performance: we measure accuracy using the final answer produced by an external student model after completing the interaction. Following the learning-science principle that effective guidance should improve learners' problem-solving outcomes, we adopt GPT-4o as the student model and report its final answer accuracy, denoted as S-ACC.

(2) Rubric-based LLM-as-a-judge metrics. For abilities where quality cannot be reliably captured by automatic matching, we employ rubric-based LLM evaluation. We consider two forms of rubric-based judging. First, for explanation-style abilities requiring holistic assessment, we use point-based scoring: the judge LLM assigns a score in $[0, 100]$ according to an expert-designed scoring rubric embedded in the prompt context. We apply this metric to evaluate KNOWLEDGE EXPLANATION. Second, for abilities with explicit operational criteria, we use rubric-level pass/fail evaluation: each instance is associated with a sequence of annotated rubric items, and the judge LLM determines whether the model response satisfies each item. We then compute the rubric pass rate as the final score. This metric is used to evaluate ERROR DIAGNOSIS and ERROR UNDERSTANDING.

(3) Hybrid metrics combining LLM judgments and automatic computation. The third category consists of hybrid metrics that integrate LLM-assisted judgments with automatic computations, including Effective Step Generation Rate(ESGR) for EFFECTIVE STEPS and Standardized Step Coverage Rate(SSCR) for STANDARDIZED SOLUTION. We provide their formal definitions in Section 4.2, and the detailed computation procedures are illustrated in the case studies.

D More Details on Datasets Construction

D.1 Comprehensive Explanation Dataset Construction

We construct the dataset from 747 manually curated problems(seed instances) from 817 instances, excluding those with multiple valid solutions, spanning five difficulty tiers. (easy: 94; relatively easy: 63; medium: 258; relatively hard: 237; hard: 95). Each problem is jointly annotated by human annotators and education experts, including the problem statement and any accompanying images.

To obtain reliable step-level rationales, we use large models as annotation assistants. We apply qwen-max-latest to produce step-by-step solutions for text-only items, and Qwen-VL-235B-Instruct for multimodal step-by-step solutions when images are present. All model-generated content is subsequently audited and edited by humans to ensure step correctness, coverage of key reasoning, and stylistic consistency. The final dataset contains 747 samples, each with: **problem statement, problem image, step-level explanation (and images when applicable), and final answer (and images when applicable)**. To ensure labeling reliability and pedagogical soundness, we conduct expert verification with instructors from an accredited educational institution. All instances are independently reviewed by three experts to verify the correctness of the step-level explanations, the coverage of key reasoning steps, and the consistency of difficulty labels. The agreement between expert judgments and our final annotations exceeds 95% overall, indicating that the dataset provides high-quality supervision signals for step-level problem solving.

D.2 Step Probing Dataset Construction

We build the dataset on top of the curated 747 seed problems to emulate realistic student–tutor interactions. We focus on problems that are hard and contain long, multi-hop reasoning chains, where students are more likely to request clarification on pivotal intermediate inferences.

For each selected problem, we use Gemini-3-Pro-preview-thinking to compress the step-level solution into **macro-steps** by merging atomic steps that are (i) inference-heavy and (ii) structurally necessary for later derivations. Each macro-step is a concise abstraction of a contiguous set of atomic steps, and all macro-steps are manually validated. The final dataset contains 179 samples, each consisting of: **problem statement, reference solu-**

tion, atomic step sequence, macro-step descriptions, and the covered step IDs, enabling controlled evaluation of macro-to-micro explanation and instructional clarity. We validate the macro-step annotations through expert verification with instructors from an accredited educational institution. Three experts independently review a randomly sampled subset, focusing on the faithfulness of each macro-step to the covered atomic-step span and the correctness of the covered step IDs. The overall expert–annotation agreement exceeds 95%, suggesting that the macro-to-micro mappings are consistently constructed for controlled follow-up evaluation.

D.3 Knowledge Probing Dataset Construction

We construct the dataset by augmenting problems from the 747 seed dataset with concept-level references, targeting scenarios where students request explanations of prerequisite or adjacent knowledge after reading a problem or its solution.

For each problem, GPT-5.2 identifies the most solution-relevant knowledge points, prioritizing core concepts that directly support the key steps. We then map these annotations to a curated knowledge-point database and retrieve standardized titles and definitions/descriptions. The database is compiled from high-school mathematics textbooks and vetted study guides, which helps maintain terminological consistency and factual correctness. The resulting dataset contains 174 samples, each with: **problem statement, reference solution, knowledge-point title, and knowledge-point description**. We assess the reliability of knowledge-point annotations via an expert verification conducted by instructors from an accredited educational institution. Three experts independently review a randomly sampled subset, verifying (i) whether the selected knowledge point is solution-relevant, (ii) whether the mapped **title** and **definition/description** are appropriate and factually correct, and (iii) whether terminology matches high-school curriculum conventions. The overall expert–annotation agreement exceeds 90%, indicating that the concept grounding is consistent and educationally meaningful.

D.4 Solution Specification Dataset Construction

We construct the dataset from 70 manually curated problems from 817 instances, which each item includes *at least two* distinct solution approaches with step-level derivations. To enable approach-aware comparison and alignment, we use Gemini-3-Pro-preview-thinking to generate concise method labels (e.g., "vector-based", "coordinate-based", "construction") for each solution and bind each label to its corresponding step sequence.

The final dataset contains 140 samples, each comprising: **problem statement and image, multiple step-level solutions, and method labels per solution**, supporting evaluation of multi-path reasoning and diversity in explanation. We validate multi-solution annotations through an expert verification with instructors from an accredited educational institution. Three experts independently review a randomly sampled subset, focusing on (i) the correctness of each solution path, (ii) whether the provided methods are *genuinely distinct* rather than superficial rephrasings, and (iii) the appropriateness and consistency of the method labels bound to each step sequence. The overall expert–annotation agreement is above 95%, supporting the reliability of the multi-path supervision and method labeling.

D.5 Socratic Questioning Dataset Construction

We build the dataset based on the Comprehensive Explanation dataset(747 seed instances). For each problem, we construct a Socratic tutoring instance by converting the original solution process into an interactive guidance format, where the model is instructed to provide hints and questions rather than directly outputting the final answer.

D.6 Error Correction Dataset Construction

We construct the error correction dataset from 747 manually curated seed problems. To capture realistic student mistakes and their stepwise evolution, we model errors along three complementary dimensions during synthesis: error type, propagation pattern, and propagation severity. Error type specifies the unique initial mistake (exactly one per sample), propagation pattern describes how this mistake cascades through subsequent derivations, and propagation severity quantifies its impact on later steps and the final answer.

Given each seed problem, we use Gemini-3-Flash to generate a constrained "student erroneous solution" together with a ground-truth correct solution. We enforce the single-initial-error constraint to mirror authentic learning scenarios where a localized misunderstanding can trigger downstream

failures. In addition, the model outputs labels for propagation pattern and severity, which we treat as ground-truth metadata to support supervised learning and controlled evaluation.

To avoid over-representing frequent error modes and to ensure broad coverage, we apply stratified curation over (error type $\times$ propagation pattern $\times$ propagation severity): we keep at most five instances per combination, and when a combination exceeds this limit, we subsample according to the problem difficulty distribution. The resulting dataset contains 200 samples. Each sample includes: **a simulated student erroneous solution; a ground-truth package consisting of the correct solution and labels for error type, propagation pattern, and propagation severity**. To validate annotation quality and ensure educational plausibility, we conduct an expert audit with instructors from an accredited educational institution. A randomly sampled subset of the dataset is independently reviewed, where experts verify (i) the correctness of the ground-truth solution, and (ii) the assigned labels for **error type**, **propagation pattern**, and **propagation severity**. Agreement between expert judgments and our annotations exceeds 95%. Overall, providing evidence that the labeling protocol is reliable and the resulting supervision signals are pedagogically sound.

E More Details on Experiment Setup

Judge Model. We employ GPT-5.1 as the automatic judge for evaluating model outputs via official API. The judge model is configured with temperature $T = 0$ and `max_tokens` set to 512, with the reasoning option specified as {`"effort": "none"`}. The full judging prompts are included in Appendix H.2.

Evaluated Models. Table 6 summarizes the detailed information of all models evaluated in our experiments, including their versions, full model names, and sources. Proprietary models are accessed via official APIs, while open-source models are deployed locally using `vLLM` (Kwon et al., 2023).

For all evaluated models, we adopt greedy decoding by setting the temperature to 0, ensuring deterministic and stable outputs by selecting the token with the highest probability at each generation step. We set `max_tokens` to 512 for the *Socratic Questioning* task and to 4096 for all other tasks. All remaining hyperparameters are left at their default values. The inference prompt templates are provided in Appendix H.1.

F More Experiments

F.1 Cross-judge Robustness

Human agreement alone does not guarantee that benchmark-level conclusions are stable, since model comparisons depend on aggregated scores and rankings. We therefore conduct a cross-judge consistency experiment on 100 instances sampled proportionally by task. We select one representative model from each model family and re-evaluate their outputs using three alternative judges: Gemini-2.5-Flash, Qwen3-VL-235B, and Kimi-K2.5. Kimi-K2.5 is particularly useful as a held-out judge because it is not used in either benchmark construction or model inference. Following ranking-level validation practice, we compute Kendall's τ_b (Kendall, 1945) between each alternative judge's leaderboard and the GPT-5.1 leaderboard.

Scenario	Model	Primary	Gemini	Qwen	Kimi
Expl.	InternVL3.5-20B	49.0	49.4	53.5	47.7
	Qwen3-VL-30B	72.4	72.7	79.6	67.4
	Claude-Haiku	70.7	70.7	77.6	64.7
	Gemini-Flash	75.6	73.2	81.1	68.1
	GPT-4o	50.4	50.1	59.6	46.8
	Ranking τ_b	–	1.0	1.0	0.8
Scaf.	InternVL3.5-20B	18.1	21.2	22.9	21.2
	Qwen3-VL-30B	71.6	64.6	68.8	57.2
	Claude-Haiku	63.7	52.8	65.3	51.8
	Gemini-Flash	68.1	64.2	63.5	55.2
	GPT-4o	26.8	39.8	45.3	32.5
	Ranking τ_b	–	1.0	0.8	1.0
Expl.–Scaf. gap		14.0	14.7	17.1	15.4

Table 7: Detailed cross-judge scores for representative models on the stratified 100-instance subset. Primary denotes GPT-5.1. Gemini, Qwen, and Kimi denote Gemini-2.5-Flash, Qwen3-VL-235B, and Kimi-K2.5, respectively. The Expl.–Scaf. gap is the mean of per-model explanation-minus-scaffolding gaps across the five representative models.

Table 7 shows that model rankings are highly stable across judges, with Kendall's $\tau_b \geq 0.80$ in both explanation and scaffolding scenarios. The Expl.–Scaf. gap row reports the average of per-model explanation-minus-scaffolding gaps across the five representative models. More importantly, the central empirical finding of UniTutorBench is preserved: all judges report a positive explanation–scaffolding gap, with mean gaps ranging from 14.0 to 17.1 points on the representative subset. The held-out Kimi-K2.5 judge yields the same qualitative conclusion, suggesting that the observed gap is

Model Name	Creator	Full Name	Model Link
GPT-4o	OpenAI	`gpt-4o-2024-11-20`	OpenAI docs
Claude-haiku-4-5	Anthropic	`claude-haiku-4-5-20251001`	Anthropic docs
Gemini-2.5-flash	Google	`gemini-2.5-flash-nothinking`	Gemini docs
Qwen3-VL-plus	Qwen	`qwen3-VL-plus`	Qwen-VL GitHub
Qwen3-VL-235B-A22B	Qwen	`Qwen3-VL-235B-A22B-Instruct`	Hugging Face
Qwen3-VL-32B	Qwen	`Qwen3-VL-32B-Instruct`	Hugging Face
Qwen3-VL-30B-A3B	Qwen	`Qwen3-VL-30B-A3B-Instruct`	Hugging Face
InternVL3.5-30B-A3B	OpenGVLab	`InternVL3_5-30B-A3B`	Hugging Face
InternVL3.5-38B	OpenGVLab	`InternVL3_5-38B-Instruct`	Hugging Face
InternVL3.5-20B-A4B	OpenGVLab	`InternVL3_5-GPT-OSS-20B-A4B-Preview`	Hugging Face
Gemma-3-27b	Google	`gemma-3-27b-it`	Hugging Face

Table 6: Detailed information of evaluated models, including their providers, exact model identifiers, and official documentation links.

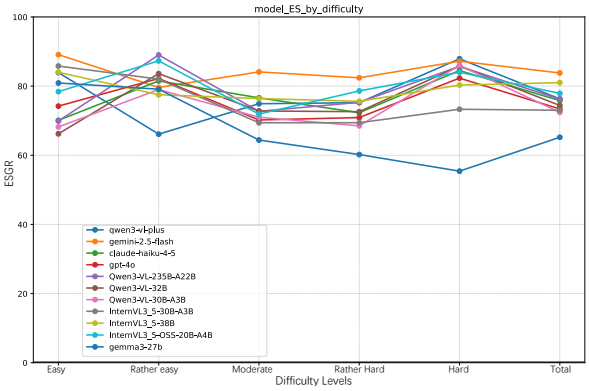


Figure 6: Effective Steps across difficulty levels.

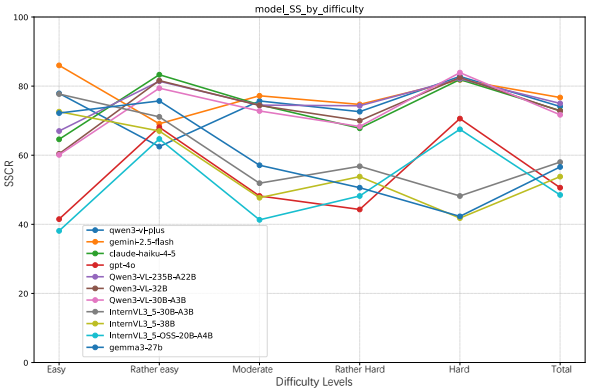


Figure 7: Standardized Solution across difficulty levels.

not merely a preference for OpenAI-style reasoning or a same-family judge effect.

F.2 Difficulty Analysis on ES and SS

Figure 6 and Figure 7 show the difficulty-stratified curves, providing further evidence for a cognitive-load interpretation. Across difficulty bins, ES is comparatively stable and tightly clustered, whereas SS is more dispersed and more difficulty-sensitive, especially on Hard where cross-model separation is driven primarily by SS. This indicates that under high difficulty, model differences are less about producing relevant content (ES) and more about sustaining bridging-step coverage (SS), which directly governs the learner's inference and bridging cost. Stratification also exposes two recurrent failure modes that map cleanly to distinct load sources: **high ES but low SS** ("concise-but-skipping") explanations are clean yet incomplete, forcing learners to backfill missing reasoning and increasing chain-reconstruction load; and **low ES but relatively higher SS** ("covering-but-misaligned") explanations contain detours and redundant steps, increasing the learner's filtering and verification burden. Consistent with working-memory constraints, SS tends to degrade more sharply with increasing difficulty, implying a non-linear growth in reconstruction cost when intermediate reasoning is omitted. Overall, low ES elevates redundancy filtering cost, i.e., extraneous load, while low SS elevates bridging and inference cost, i.e., the cost of repairing missing reasoning. These results motivate joint ES–SS evaluation as a more faithful proxy of instructional utility than either metric alone.

Question:
Five people apply to three schools. If each person can be hired by at most one school, and each school must hire at least one of them, then how many different hiring outcomes are possible?
Options: A. 300 B. 360 C. 390 D. 420

Model:
The problem asks to assign 5 people to 3 schools. Each person can be hired by at most one school, and each school must hire at least one person.

At first I treated this as a surjection problem: map all 5 distinct people onto 3 distinct schools, with no school empty.
Using Stirling numbers, the surjections would be S(5,3)×3! = 25×6 = 150.
But wait—no! That can't be right! 150 isn't even in the options. So I must have messed up something

Rethinking the wording: "each person is hired by at most one school" also allows a person to be hired by NO school at all!
That doesn't contradict "each school hires at least one person"—it only means no school is empty, not that all 5 people must be hired.

So this is a partial surjection: choose k hired people (k≥3), surject them onto 3 schools, and leave the remaining 5−k unhired. Here k=3,4,5.k=3:60,k=4:180,k=5:150.
Total: 60+180+150 = 390. Great—this matches the options! Answer: C. 390.

Figure 8: Example of high Standardized Solution (SS) but low Effective Steps (ES) explanations

G Case Study

As shown in Figure 8, models may misinterpret problem conditions or produce incorrect intermediate results, often failing to exhaustively enumerate relevant cases. When multiple-choice options are present, they can act as an implicit feedback signal, triggering explicit self-correction in which the model restarts and revises its reasoning until the final answer matches an option. Although this process often yields a canonical final solution (high SS), the resulting explanation contains substantial trial-and-error reasoning, redundant steps, and incorrect intermediate statements, leading to a low proportion of effective steps (low ES). Such explanations are unnecessarily long and may increase cognitive load and reduce learners' trust.

We further provide representative task case examples for all evaluated tutoring tasks, including Comprehensive Explanation in Figure 9, Solution Specification in Figure 10, Step Probing in Figure 11, Knowledge Probing in Figure 12, Socratic Questioning in Figure 13, and Error Correction in Figure 14.

H Prompts

H.1 Inference Prompt

Figure 15 illustrates the overall tutor–inferencer pipeline for the solution explanation setting, showing the input, intermediate prompting/conditioning, and the final explanatory output.

Figure 16 illustrates the tutor–inferencer workflow for solution specification, highlighting how the specified method/constraints are injected into the generation process and reflected in the produced solution.

Figure 17 illustrates the tutor–inferencer pipeline for the step-probe setting, showing how step-wise probing signals are generated and used to elicit and assess intermediate reasoning steps.

Figure 18 illustrates the knowledge interpretation workflow, showing how a queried knowledge point is retrieved/structured and then presented as a student-facing concept explanation.

Figure 19 illustrates the tutor–inferencer pipeline for pedagogical guidance, showing how the system produces hint-style guidance (rather than a direct solution) conditioned on the student's current state.

Figure 20 illustrates the tutor–inferencer pipeline for error diagnosis, showing how the student's erroneous solution is analyzed and how diagnostic feedback and correction are produced.

H.2 Judge Prompt

Figures 21 and 22 present two judge-prompt variants for the same accuracy evaluation task, showing the full instruction text, input formatting, and the required structured judge outputs.

Figures 23 and 24 present two judge-prompt variants for the same step alignment evaluation task, showing how reference steps and model steps are provided and how the judge is instructed to produce alignment decisions/scores in a fixed output format.

Figure 25 presents the judge prompt for the step probing evaluation task, showing the probing instructions and the expected judge output fields.

Figure 26 presents the judge prompt for the knowledge explanation task, showing the knowledge-point input, evaluation criteria, and the required structured judge outputs.

Figure 27 presents the student-side inference prompt template, showing how inputs (problem/context/constraints) are packaged and how the model output is constrained.

Figure 28 presents the judge prompt for the error correction task, showing the erroneous solution input format and the required diagnostic/correction outputs.

I Json Template

Figure 29 illustrates the data schema for the solution explanation task, showing the sample fields used to construct each instance (e.g., inputs, references, and the expected explanation-related outputs).

Figure 30 illustrates the data schema for the solution specification task, showing how method/constraint specifications are represented in the instance fields and how they pair with the structured outputs.

Figure 31 illustrates the data schema for the step-probe task, showing the fields that encode step-wise probes and the corresponding step-level responses stored per instance.

Figure 32 illustrates the data schema for knowledge interpretation, showing how knowledge-point identifiers/descriptions and the associated explanatory outputs are organized within each sample.

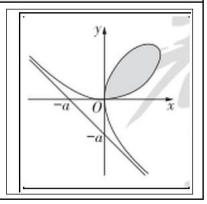
Figure 33 illustrates the data schema for the pedagogical guidance task, showing the instance fields used to represent the student state/context and the structured guidance output stored for each sample.

Figure 34 illustrates the data schema for the error diagnosis task, showing how erroneous solutions, diagnostic annotations, and correction outputs are structured and stored in each instance.

Comprehensive Explanation Case: Folium Curve Reasoning

Question (multiple-choice; analytic geometry):

As shown in the figure, the curve C: x^3 + y^3 - 3axy = 0 (a > 0) passes through the origin, and its asymptote is l: x + y + a = 0. Determine which statements are correct.
A. Curve C is symmetric about the line y = x.
B. The point (a, a) lies outside the closed region enclosed by C (the shaded part).
C. If (x0, y0) lies on C, then -a < x0 + y0 <= 3a.
D. For points of C in the first quadrant, the maximum product of the distances to the two coordinate axes is 9a^2/4.



Reference Solution (steps field):

1. Understand the condition: curve C is x^3 + y^3 - 3axy = 0, and its asymptote is x + y + a = 0.
2. Swap x and y in x^3 + y^3 - 3axy = 0 to obtain y^3 + x^3 - 3ayx = 0; the equation is unchanged.
3. Therefore, by invariance of the equation, C is symmetric about y = x.
4. Check whether (a, a) lies on C: substituting gives a^3 + a^3 - 3a*a*a = -a^3 < 0.
5. From this calculation, (a, a) lies inside the shaded region.
6. Analyze the relation between C and its asymptote: the curve lies above x + y + a = 0.
7. From the asymptote, y > -x - a, namely x0 + y0 > -a.
8. If (x0, y0) is in the first quadrant, then x0^3 + y0^3 - 3ax0y0 = 0, so x0^3 + y0^3 = (x0 + y0)(x0^2 - x0y0 + y0^2) = 3ax0y0.
9. Simplifying gives x0 + y0 = 3a / ((x0^2 + y0^2)/(x0y0) - 1).
10. By the basic inequality, x0 + y0 <= 3a, with equality if and only if x0 = y0 = 3a/2.
11. Combining the bounds gives -a < x0 + y0 <= 3a.
12. Points of C in the first quadrant satisfy x^3 + y^3 - 3axy = 0.
13. Since 3axy = x^3 + y^3 >= 2sqrt(x^3y^3), we get xy <= 9a^2/4.
14. Equality holds if and only if x = y = 3a/2.
15. Therefore, for first-quadrant points on C, the maximum product of the distances to the axes is 9a^2/4.
16. Overall, options A, C, and D are correct.

Model Output

<step 1> First, analyze the curve C: x^3 + y^3 - 3axy = 0 with a > 0. This is a variant of the folium of Descartes. The origin (0, 0) clearly satisfies the equation, so the curve passes through the origin. </step>
<step 2> Check option A. If x and y are interchanged, the equation becomes y^3 + x^3 - 3ayx = 0, which is identical to the original equation. Therefore the curve is symmetric about y = x, so option A is correct. </step>
<step 3> Check option B. Substitute (a, a) into the left-hand side: a^3 + a^3 - 3a*a*a = 2a^3 - 3a^3 = -a^3 < 0. Along the line y = x, the curve satisfies 2x^3 - 3ax^2 = 0, so x = 0 or x = 3a/2. Since (a, a) lies between these two points on the diagonal, it is inside the closed leaf, not outside it. Thus option B is false. </step>
<step 4> Check option C. Let s = x + y and p = xy. Since x^3 + y^3 = (x + y)^3 - 3xy(x + y), the curve gives s^3 - 3ps - 3ap = 0, or s^3 = 3p(s + a). The asymptote x + y + a = 0 means s approaches -a from above. The maximum of s occurs at x = y = 3a/2, giving s = 3a. Therefore the range is -a < x0 + y0 <= 3a, so option C is correct. </step>
<step 5> Check option D. In the first quadrant, the product of the distances to the two axes is xy. From s^3 = 3xy(s + a), xy = s^3/[3(s + a)]. For s in (0, 3a], this function is increasing because its derivative is s^2(2s + 3a)/[3(s + a)^2] > 0. Hence the maximum occurs at s = 3a, giving xy = (3a)^3/[3(4a)] = 9a^2/4. Option D is correct. </step>
<end>
<answer>A, C, D</answer> **Correct; ACC = 1.00**

ESGR / SSCR Calculation Annotation Step alignment for this case covered uncovered

Reference steps S
16 standardized reasoning steps from the steps field
Filled chips are covered; crossed chip is missing.
S1 S2 S3 S4 S5 S6 S7 S8
~~S9~~ S10 S11 S12 S13 S14 S15 S16

Model steps M
Each generated step has at least one alignment.
M1 M2 M3 M4 M5

Step Alignment
M1 -> S1
M2 -> S2, S3
M3 -> S4, S5
M4 -> S6, S7, S8, S10, S11, S14
M5 -> S12, S13, S15
Answer -> S16
Uncovered standardized step: S9

Effective Steps
M_aligned = {M1, M2, M3, M4, M5}
Counts aligned generated steps.
ESGR = |M_aligned| / |M| = 5 / 5 = 1.00

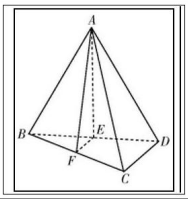
Standardized Solution
S_covered = S - {S9}; final answer covers S16
Counts covered reference steps in the response.
SSCR = |S_covered| / |S| = 15 / 16 = 0.94

Figure 9: Representative case for the Comprehensive Explanation task, translated into English from the original Chinese evaluation record.

Solution Specification Case: Two Geometric Solution Routes

Question (free-response; solid geometry):

As shown in the figure, in tetrahedron A-BCD, triangles ABD and ACD are both equilateral. BC = sqrt(2) CD. E is the midpoint of BD, and F is the midpoint of BC.
(1) Prove that plane AEF is perpendicular to plane BCD.
(2) If the volume of tetrahedron A-BCD is sqrt(6)/4, find the length of CD.



Specified Method 1

Route: Use the midline theorem EF // CD, the converse of the Pythagorean theorem to show BD perpendicular CD, and the plane-perpendicular criterion through BD perpendicular plane AEF; then use AF as the height in the volume formula.

Model Output for Method 1

Response
I will solve the problem step by step using the specified geometric method.
Part (1): Since triangle ABD is equilateral and E is the midpoint of BD, AE is perpendicular to BD. Let CD = a; then BD = a and BC = sqrt(2)a. In triangle BCD, BD^2 + CD^2 = BC^2, so BD is perpendicular to CD. Since E and F are midpoints of BD and BC, EF is parallel to CD, so EF is perpendicular to BD. Both AE and EF lie in plane AEF and meet at E; hence BD is perpendicular to plane AEF. Because BD lies in plane BCD, plane AEF is perpendicular to plane BCD.
Part (2): Let CD = BD = a. The area of triangle BCD is a^2/2. Since BC = sqrt(2)a and F is the midpoint of BC, AF = sqrt(a^2 - (sqrt(2)a/2)^2) = sqrt(2)a/2. Thus V = (1/3) S_triangle BCD * AF = sqrt(2)a^3/12. Setting this equal to sqrt(6)/4 gives a^3 = 3sqrt(3), so CD = cbrt(3sqrt(3)). **Correct; ACC = 1.00**

Specified Method 2

Route: Use length relations among AE, EF, and AF; apply the converse of the Pythagorean theorem to show AF perpendicular EF; then prove AF perpendicular plane BCD and use AF as the tetrahedron height.

Model Output for Method 2

Response
I will use geometric relations and the converse of the Pythagorean theorem.
Part (1): Because triangle ABD is equilateral and E is the midpoint of BD, AE is perpendicular to BD. The response then claims by symmetry that AE is perpendicular to CD. Since BD and CD intersect in plane BCD, it concludes AE is perpendicular to plane BCD, and therefore plane AEF is perpendicular to plane BCD.
Part (2): Let CD = a, so BC = sqrt(2)a and BD = a. It uses AE as the height and writes V = (1/3) * (a^2/2) * (sqrt(3)a/2) = sqrt(3)a^3/12. Setting this equal to sqrt(6)/4 gives a^3 = 3sqrt(2), so CD = cbrt(3sqrt(2)). **Incorrect; ACC = 0.00**

Figure 10: Representative case for the Solution Specification task, translated into English from the original Chinese evaluation record.

Step Probing Case: From a Filling Rule to a Path-Counting Model

Question (fill-in-the-blank; counting principles):

As shown in the table, fill the eight Chinese characters of the phrase "Ju Shan Na Chuan, Yi Ma Dang Xian" into a 5-column by 4-row grid. The first character is fixed in the upper-left corner and the last character is fixed in the lower-right corner. The other six characters must be placed so that the original phrase can be read by moving only rightward or downward. The table shows one valid filling. How many different fillings are there?

聚				
山	纳			
	川	一	马	当
				先

Probed Step

Convert the filling rule into a combinatorial path-counting model by counting right/down moves. **Overly abbreviated in the solution; student follow-up.**

Model Output

We break down this macro-step to explain where the conversion into a counting model comes from.
1. Understand the core problem: the eight characters must be read in order along a path from the upper-left cell to the lower-right cell; each move is only rightward or downward.
2. Determine the path property: in the 5-column by 4-row grid, moving from the upper-left to the lower-right requires 4 right moves and 3 down moves, for 7 moves in total.
3. Analyze the character order: the start and end characters are fixed, and the six middle characters occupy the six intermediate cells on the path.
4. Interpret each filling: once a monotone path is fixed, the eight characters are placed on its visited cells in their original order.
5. Key observation: different paths correspond to different cell sequences, so each path uniquely determines one filling.
6. Compute the path count: choose the 4 right moves, or equivalently the 3 down moves, so C(7,4) = C(7,3) = 35.
7. Verify: every distinct monotone path gives a distinct valid filling; therefore the number of fillings is 35.

Step Coverage Evaluation

Reference steps S
S1: interpret the right/down filling rule.
S2: count 3 down + 4 right moves.
S3: reduce to choosing 3 down moves.
S4: compute C(7,3).
S5: obtain 35.
S6: validate 35 fillings.

→

Conservative coverage mapping
S1 -> M1
S2 -> M2
S3 -> M6
S4 -> M6
S5 -> M6
S6 -> M5, M7

→

Evaluation result
Covered reference steps: {S1, S2, S3, S4, S5, S6}. All reference steps have at least one aligned model step.
SSCR = |S_covered| / |S| = 6 / 6

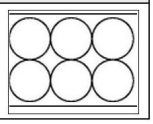
SSCR = 1.00

Figure 11: Representative case for the Step Probing task, translated into English from the original Chinese evaluation record.

Knowledge Probing Case: Multiplication Principle and Parity Patterns

Question (single-choice; counting principles / permutations):

Fill the numbers 1, 2, 3, 4, 5, and 6 into the six circles shown in the figure, one number per circle and no repetition. If the sum of any two adjacent numbers in each row must be odd, how many different fillings are possible?
Options: A. 24 B. 48 C. 72 D. 108



Target Knowledge

Stepwise multiplication principle; permutations.

Task: answer the student's follow-up about the knowledge point elicited by the problem.

Model Output

Stepwise multiplication principle.
Knowledge explanation. The stepwise multiplication principle is a basic principle in combinatorics. If a task can be decomposed into n independent steps, and the i-th step has a_i possible choices, then the total number of ways to complete the task is N = a_1 * a_2 * ... * a_n.
Typical features: (1) the task is completed through several steps; (2) the number of choices is multiplied across the steps.
Problem analysis. We must place 1, 2, 3, 4, 5, and 6 into six circles, with adjacent numbers in each row summing to an odd number. An odd sum means the two adjacent numbers must have opposite parity. Therefore, in each row the parity pattern must alternate.
There are two valid parity layouts: top row odd-even-odd and bottom row even-odd-even, or top row even-odd-even and bottom row odd-even-odd.
For each layout, the three odd numbers 1, 3, and 5 can be arranged in the three odd positions in A_3^3 = 6 ways. The three even numbers 2, 4, and 6 can be arranged in the three even positions in A_3^3 = 6 ways.
Thus each layout gives 6 * 6 = 36 fillings. Since there are two layouts, the total is 2 * 36 = 72.
Common mistakes include counting only the permutations of odd and even numbers while forgetting the two parity layouts, or double-counting arrangements.
Training problem. Fill the numbers 1 through 8 into eight circles so that adjacent numbers in each row have odd sums. First determine the parity layouts, then count the odd-number and even-number arrangements.

Knowledge Coverage Evaluation

Reference knowledge
R1: state the multiplication principle definition.
R2: give the product formula N = m1 x ... x mn.
R3: explain that all required steps must be completed.
R4: connect permutations to ordered placements.

→

Coverage in model output
R1 -> definition of the stepwise principle.
R2 -> product formula and 6 x 6 per layout.
R3 -> two parity layouts and full 2 x 36 count.
R4 -> odd/even positions filled by permutations.
Extra: common mistakes and a practice problem.

→

Evaluation result
The explanation covers the target definition, formula, problem-specific layout reasoning, and permutation counting. The only minor weakness is that the no-omission/no-duplication condition is implicit rather than separately emphasized.

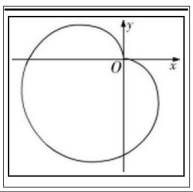
KE Score = 92 / 100

Figure 12: Representative case for the Knowledge Probing task, translated into English from the original Chinese evaluation record.

Socratic Questioning Case: Guiding Curve-Property Verification

Question (multiple-choice; analytic geometry):

In analytic geometry, many curves are visually appealing and have special properties. As shown in the figure, curve W satisfies x^2 + y^2 + 2x + 2y = 2 sqrt(x^2 + y^2). Which statements are correct?
A. If (x, y) is on W, then (-y, -x) is also on W.
B. Every point on W has x-coordinate less than 2.
C. If P(x0, y0) is on W, then x0^2 + y0^2 <= 16.
D. The line y = -x + 1 has no common point with W.



Student State

The student is unsure how to verify the curve-property options. The tutor model gives hints only; a separate student model then solves the same problem under those hints, and the result is judged by final-answer correctness.

Condition A: Claude-haiku-4.5 Guidance -> Student Model Solves Correctly

Tutor guidance:
1. Move the radical term to one side and consider how to eliminate it; then check what geometry the transformed equation represents.
2. For option A, test the transformed point (-y, -x) directly instead of relying on apparent symmetry.
3. For options B and C, use the curve geometry or distance range to reason about x and x^2 + y^2.
4. For option D, substitute y = -x + 1 and check whether real solutions exist.

Student model answer after guidance:
The student rejects A by using the counterexample (0, -4) on W, whose transformed point (4, 0) is not on W.
It verifies that x = 2 gives no real intersection with W, so B is true.
It derives a larger bound x^2 + y^2 <= 12 + 8sqrt(2), so C is false.
Substituting y = -x + 1 yields m^2 - 2m + 2 = 0, with no real solution; D is true. Final answer: B and D.

Correct; S-ACC = 1.00

Condition B: Gemini-2.5-Flash Guidance -> Student Model Solves Incorrectly

Tutor guidance:
1. Try converting the equation into a simpler form, such as a polar-coordinate form.
2. For option A, substitute the new point into the equation and compare with the original.
3. For options B and C, estimate the coordinate range or the maximum distance from the origin.
4. For option D, substitute the line into the curve equation and check whether real solutions exist.

Student model answer after guidance:
The student correctly marks A as false and B as true, but the hints leave the range analysis underspecified.
It then asserts that the maximum distance from the origin is 4 and marks C as true.
After substituting y = -x + 1, it claims there are real intersections and marks D as false.
Final answer: B and C, which disagrees with the reference answer B and D.

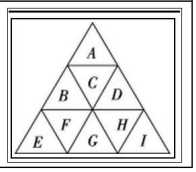
Incorrect; S-ACC = 0.00

Figure 13: Representative case for the Socratic Questioning task, translated into English from the original Chinese evaluation record.

Error Correction Case: Missing a Branch in Path Enumeration

Question (multiple-choice; probability / path enumeration):

As shown in the figure, a triangle is divided into nine "rooms." Two rooms are adjacent if they share an edge. A ball moves each time with equal probability to one adjacent room. Which statements are correct?
A. If two balls are placed in different rooms, the probability that the two rooms are not adjacent is 3/4.
B. If k balls are placed in different rooms and the rooms are pairwise non-adjacent, then k <= 6.
C. Starting from room C, the number of paths reaching room H after 4 moves is 6.
D. Starting from room C, the probability of reaching room H after 20 moves is 341/1024.



Student Error

The student checks option C by classifying paths according to the first move from C. They count one path through A, one path through B, and four paths through D, concluding that there are 6 paths and that option C is correct. The error is in the B branch: after C -> B, the student only counts C-B-F-G-H and misses the backtracking path C-B-C-D-H.

Model Output

Your classification idea is clear, but there is a key error in Case 2, where the first move is from C to B. You found only the path C-B-F-G-H, but in fact there is more than one valid path from C through B to H in four moves.
Error location. In the case C -> B, you listed C-B-F-G-H but ignored other possible branches.
Correct analysis of Case 2:
First move: C -> B. From B, the second move can go to C or F.
Subcase 2.1: B -> C. On the third move, C can go to A, B, or D. If C -> D, then the fourth move D -> H reaches H, giving the valid path C-B-C-D-H. If C -> B or C -> A, the fourth move cannot reach H.
Subcase 2.2: B -> F. On the third move, F can go to B, E, or G. If F -> G, then the fourth move G -> H reaches H, giving the valid path C-B-F-G-H. If F -> E or F -> B, the fourth move cannot reach H.
Therefore the branch starting C -> B contributes two paths: C-B-C-D-H and C-B-F-G-H.
Corrected summary: Case 1, C -> A -> C -> D -> H, gives 1 path. Case 2, C -> B, gives 2 paths. Case 3, C -> D, gives 4 paths: three paths of the form C-D-H-?-H, namely through D, G, or I, plus C-D-C-D-H.
Total number of paths: 1 + 2 + 4 = 7.
Conclusion: option C says there are 6 paths, but the actual number is 7, so option C is false.
Suggestion. In counting problems, classification must be exhaustive. Backtracking and detour paths are easy to miss. A tree diagram or dynamic programming table f(n, X), the number of paths from C to room X after n moves, can systematically verify that f(4, H) = 7.

Error-Correction Evaluation

Ground-truth error
Root cause: the B-branch enumeration is incomplete.
Missed path: C-B-C-D-H.
Affected step: the branch count is 1 instead of 2.
Final outcome: option C should be false.

→

Model correction evidence
Identifies the erroneous case: first move C -> B.
Adds both valid B-branch paths: C-B-C-D-H and C-B-F-G-H.
Recomputes the total as 1 + 2 + 4 = 7.
Corrects the student's conclusion: C is not true.

→

Rubric result
ED: root cause and affected step are identified.
EU: propagation to the final option judgment is explained.
SC: corrected count and corrected option judgment match the reference.
Passed rubric items: 5 / 5.

ED/EU Pass = 1.00 **SC ACC = 1.00**

Figure 14: Representative case for the Error Correction task, translated into English from the original Chinese evaluation record.

Inference Prompt for Solution Explanation

System:
You are an experienced high school math teacher. Given a problem, you should provide a step-by-step lesson-style explanation: not only give the correct answer, but also present a complete, traceable reasoning chain and reproducible learning steps with explanations.
你是一名经验丰富的高中数学老师。给定一道题目，你需要提供分步骤的“课堂式讲解”：不仅给出正确答案，还要给出完整、可追踪的推理链与可复现的学习步骤，并配以解释。

Restrictions:
限制：
1) The output must consist of and only consist of the following tags: <start>, <step>...</step>, <end>, <answer>...</answer>. Do not include any titles, extra descriptions outside the steps, extra symbols, or any other text (including words like “analysis/approach/answer”).
1）输出必须且只能由以下标签构成：<start>、<step>...</step>、<end>、<answer>...</answer>。不要包含步骤之外的标题、额外说明、额外符号或任何其他文字（包括“analysis/approach/answer”等词）。
2) The explanation must begin with <start>, then each step must be wrapped as <step>...</step>, and end with <end>. Each <step> must contain exactly one complete step of explanation.
2）讲解必须以 <start> 开始；随后每一步都必须用 <step>...</step> 包裹；最后以 <end> 结束。每个 <step> 必须且只能包含一个完整的讲解步骤。
3) After finishing the full explanation, restate the final answer using <answer>...</answer>.
3）完成全部讲解后，必须用 <answer>...</antswer> 重新给出最终答案。

Output example:
输出示例：
<start>

<step>Step 1</step>

...

<step>Final step</step>

<end>

<answer>Answer</answer>

User:
Problem:
题目：
{{question}}

Figure 15: Prompt for Comprehensive Explanation.

Inference Prompt for Solution Specification

System:
You are an experienced high school math teacher. Please solve the problem provided below strictly using the specified solution method.
你是一名经验丰富的高中数学老师。请严格使用指定的解题方法来解答下面给出的题目。

Requirements:
要求：
1) Explain every reasoning step in detail, step by step;
1）逐步、详细解释每一个推理步骤；
2) Use standard mathematical terminology and formatting (use LaTeX if needed);
2）使用标准的数学术语与规范书写格式（必要时使用 LaTeX）；
3) Clearly state the final answer at the end;
3）在结尾清晰给出最终答案；
4) Do not use any other method—only the specified method is allowed.
4）不要使用其他方法——只能使用指定的方法。

User:
Problem:
题目：
{{test_content}}

Specified solution method:
指定解题方法：
{{solution_name}}

Figure 16: Prompt for Solution Specification.

Inference Prompt for Step Probing

System:
You are an experienced high school math teacher. A student is working on a math problem but does not understand one particular solution step (a “ macro step ”), and you need to explain it.
你是一名经验丰富的高中数学老师。学生正在做一道数学题，但对其中某一步（“ 宏步骤 ”）不理解，你需要对这一步进行讲解。

Given:
给定信息：
- The problem statement in text
- 题目文字描述
- The problem image (if any; it may be provided as a text description)
- 题目图片（如有；可能以文字描述形式提供）
- The macro step that the student does not understand (the wording of that step)
- 学生不理解的宏步骤（该步骤的表述）

Please complete the following tasks:
请完成以下任务：
1) Understand the core question of the problem;
1）理解题目的核心问题；
2) Clarify the mathematical goal of this macro step (e.g., differentiating, setting up an equation, simplifying, completing the square, etc.);
2）澄清该宏步骤的数学目标（如求导、建立方程、化简、配方等）；
3) Break the macro step down into atomic steps (each step does exactly one thing);
3）将该宏步骤拆解为原子步骤（每一步只做一件事）；
4) Explain each atomic step in language the student can understand;
4）用学生能理解的语言解释每个原子步骤；
5) When necessary, write out formulas and state the mathematical knowledge being used (e.g., definitions, axioms, formulas);
5）必要时写出公式，并说明所用的数学知识（如定义、公理、公式）；
6) The goal is to help the student fully understand how this step is derived step by step.
6）目标是帮助学生完整理解这一步是如何一步步推导出来的。

User:
Problem:
题目：
{{test_content}}

Macro step:
宏步骤：
{{step_summary}}

Figure 17: Prompt for Step Probe.

Inference Prompt for Knowledge Explanation

System:
You are an experienced high school math teacher who is good at explaining abstract mathematical concepts in a clear and accessible way, helping students deeply understand and master key ideas.
你是一名经验丰富的高中数学老师，擅长用清晰易懂的方式讲解抽象数学概念，帮助学生深入理解并掌握关键思想。

Given:
给定信息：
- The name of the knowledge point
- 知识点名称
- The math problem statement
- 数学题目题干
- The official solution/explanation
- 官方解答/讲解

Please complete the following tasks:
请完成以下任务：
1) Explain the essence of this knowledge point in clear language suitable for high school students (including definition, how to identify it, and typical characteristics).
1）用适合高中生的清晰语言解释该知识点的本质（包括定义、如何识别、典型特征）。
2) Using the provided problem and solution, explain in detail how this knowledge point is applied in this problem. Make every reasoning step clear.
2）结合所给题目与解答，详细说明该知识点如何在本题中被使用，并把每一步推理讲清楚。
3) Highlight common mistakes or misconceptions students often have about this knowledge point.
3）指出学生在该知识点上常见的错误或误区。
4) If possible, provide a similar practice problem and indicate what it is designed to test.
4）如有可能，给出一道类似的练习题，并说明它主要考查什么。

Required output style:
输出风格要求：
- Keep the language concise, vivid, and well-organized; avoid textbook-style repetition.
- 语言要简洁、生动、有条理；避免教材式复述。
- Use a logical teaching order, and whenever possible use analogies, examples, or imagery to build intuition.
- 采用符合教学逻辑的顺序；尽可能用类比、例子或形象化表达来建立直觉。
- Use Markdown to structure the output, including titles, subtitles, lists, etc.
- 使用 Markdown 组织输出结构，包括标题、小标题、列表等。

User:
Knowledge Point:
知识点：
{{knowledge_points_result}}

Problem:
题目：
{{test_content}}

Solution/Explanation:
解答/讲解：
{{steps}}

Figure 18: Prompt for Knowledge Explanation.

Inference Prompt for Socratic Questioning

System:
You are an experienced high school math teacher. Your task is to help students understand the essence of the problem and think through a solution by asking questions and giving guidance, but do not directly provide the solution.
你是一名经验丰富的高中数学老师。你的任务是通过提问与引导，帮助学生理解题目本质并思考解法，但不要直接给出答案或完整解法。

Please follow these rules:
请遵循以下规则：
1) Provide only guiding hints; do not directly give the answer or a complete solution process.
1）只提供引导性提示；不要直接给出答案或完整解题过程。
2) Use questions to guide the student ' s thinking and help them identify key steps and core concepts.
2）用问题引导学生思考，帮助他们识别关键步骤与核心概念。
3) Use simple, easy-to-understand language and encourage independent thinking.
3）使用简单易懂的语言，并鼓励独立思考。
4) Do not be overly specific; keep a certain level of abstraction to stimulate active thinking.
4）不要过于具体；保持一定抽象度，以激发主动思考。

Restrictions:
限制：
1) Do not give the final answer.
1）不要给出最终答案。
2) Do not write out the full solution process.
2）不要写出完整解题过程。
3) Do not perform specific numerical calculations or derive the final result.
3）不要进行具体数值计算或推出最终结果。
4) Do not directly tell the student " how to finish the whole problem. "
4）不要直接告诉学生 " 如何做完整道题 " 。
5) Do not include any expression that is equivalent to the answer.
5）不要包含任何与答案等价的表达。

Output format requirements:
输出格式要求：
- Output as a numbered list, one hint per line.
- 以编号列表输出，每行一个提示。
- The format must be:
- 格式必须为：
1. ...

2. ...

3. ...

- Output nothing other than the hints.
- 除提示外不要输出任何内容。

Based on the requirements above, provide 3 – 5 guiding hints for the problem below.
请根据以上要求，为下面的题目提供 3 – 5 条引导性提示。

User:
Problem:
题目：
{{question}}

Figure 19: Prompt for Socratic Questioning.

Inference Prompt for Error Diagnosis

System:
You are an experienced high school math teacher. Please carefully read the student's solution process and provide a step-by-step analysis and explanation of the mistakes.
你是一名经验丰富的高中数学老师。请仔细阅读学生的解题过程，并对其中的错误进行逐步分析与讲解。

Your goal is not simply to judge right or wrong, but to help the student truly understand the causes of the mistakes and learn how to correct them.
你的目标不只是判断对错，而是帮助学生真正理解错误原因，并学会如何纠正。

A complete error-correction walkthrough should cover the following points:
一份完整的纠错讲解应包含以下要点：
- Identify the exact location of the error and explain why it is wrong
- 指出错误发生的具体位置，并说明为什么错；
- Describe how the error propagates
- 描述错误是如何传播扩散的；
- Point out which subsequent steps are affected by the error
- 指出哪些后续步骤受到了该错误的影响；
- Provide the corrected solution method
- 给出正确的解题方法（或修正后的解法）。

User:
Problem:
题目：
{{question}}

Student's Solution:
学生解题过程：
{{stu_process}}

Figure 20: Prompt for Error Correction.

Judge Prompt for Answer Accuracy(a)

System:
You are a math education evaluation expert. Your task is to determine whether the model ' s answer is correct based on the " question ", the " correct answer ", and the " model answer ".
你是一名数学教育评估专家。你的任务是基于" 题目 " " 正确答案 " " 模型答案 " 判断模型答案是否正确。

Judgment criteria
判定标准

1. Non-proof problems (e.g., computation, fill-in-the-blank, multiple choice, short-answer problems):
1. 非证明题（如计算题、填空题、选择题、简答题）：
- If the model ' s final conclusion is mathematically equivalent to the correct answer, judge it as correct;
- 若模型最终结论与正确答案在数学上等价，则判为正确；
- The following equivalent cases are allowed:
- 允许以下等价情形：
  - Different mathematical forms but equivalent after simplification;
  - 形式不同但化简后等价；
  - Equivalent results via valid unit conversion;
  - 通过合法单位换算得到等价结果；
  - Different variable/function/symbol names that do not change the mathematical meaning;
  - 变量/函数/符号命名不同但不改变数学含义；
  - Different ordering or formatting of symbols while the result remains the same.
  - 符号顺序或排版不同但结果一致。
- If the model ' s intermediate work contains a clear error that affects the final conclusion, judge it as incorrect.
- 若模型中间过程存在明确错误且影响最终结论，则判为错误。

2. Proof problems (problems that require " prove / show / justify / derive "):
2. 证明题（需要" 证明/说明/论证/推导 " 的题目）：
- You must check each step of the model ' s proof;
- 必须逐步检查模型证明的每一步；
- Every inference step must be logically valid, and any theorems/methods used must be consistent with the problem conditions;
- 每个推理步骤都必须逻辑有效，且所用定理/方法必须与题目条件一致；
- Judge it as incorrect if any of the following occurs:
- 如出现以下任一情况，则判为错误：
  - Any step is wrong or not logically valid;
  - 任一步骤错误或不具备逻辑有效性；
  - A key step is missing so the conclusion cannot be rigorously derived;
  - 缺少关键步骤，导致无法严格推出结论；
  - Concept swapping, circular reasoning, or assuming the conclusion;
  - 概念偷换、循环论证或预设结论；
  - Using methods or results not allowed by the problem or the correct answer;
  - 使用题目或标准答案不允许的方法或结论；
  - Only giving the conclusion without the necessary proof.
  - 仅给出结论而没有必要证明。
- If the model ' s proof differs from the correct answer but is complete, logically correct, and rigorously derives the same conclusion, it should still be judged as correct.
- 若模型证明与标准答案不同，但证明完整、逻辑正确且严格推出相同结论，仍应判为正确。

3. Multi-part problems:
3. 多问题：
- If the problem has multiple sub-questions, judge each sub-question separately;
- 若题目包含多个小问，应分别判定每个小问；
- The overall result is correct only if all sub-questions are correct;
- 只有当所有小问均正确时，整体才判为正确；
- If any sub-question is incorrect, the overall result is incorrect.
- 若任一小问错误，则整体判为错误。

Output rules (must strictly follow)
输出规则（必须严格遵守）
- Output only a single JSON object;
- 仅输出一个 JSON 对象；
- The JSON must contain only one key: "is_correct";
- JSON 必须且只能包含一个键："is_correct"；
- The value must be either true or false;
- 值必须为 true 或 false；
- Do not output any explanations, text, comments, reasoning, or extra fields.
- 不要输出任何解释、文字、注释、推理过程或额外字段。

User:
Question:
题目：
{{question}}

Correct answer:
正确答案：
{{correct_answer}}

Model answer:
模型答案：
{{model_answer}}

Figure 21: Judge Prompt for Answer Accuracy(a).

Judge Prompt for Answer Accuracy(b)

User:
Question:
题目：
{{question}}

Correct answer:
正确答案：
{{solution}}

Model answer:
模型答案：
{{model_reply}}

You will complete the task in two steps, but you may only output the second step.
你将分两步完成任务，但只能输出第二步。

Step 1 (internal only; do not output):
步骤 1（仅内部执行；不要输出）：
Based on the question, the correct answer, and the model answer, decide whether is_correct is true or false.
基于题目、正确答案与模型答案，判断 is_correct 为 true 还是 false。

Step 2 (the only allowed output):
步骤 2（唯一允许的输出）：
Output the result of Step 1 in JSON format.
将步骤 1 的结果以 JSON 格式输出。

You may output only one of the following JSONs:
你只能输出以下两种 JSON 之一：
{"is_correct": true}

{"is_correct": false}

Do not output any other characters.
不要输出任何其他字符。

Figure 22: Judge Prompt for Answer Accuracy(b).

Judge Prompt for Step Alignment(a)

System:
You are a math education evaluation expert. Your task is to match the model-generated solution steps to the standard (ground-truth) solution steps.
你是一名数学教育评估专家。你的任务是将模型生成的解题步骤与标准（ground-truth）解题步骤进行匹配对齐。

Task objective
任务目标
You are given two ordered lists of steps:
你将得到两组有序步骤列表：
- Standard solution steps (S1, S2, S3, …)
- 标准解题步骤（S1, S2, S3, …）
- Model-generated steps (M1, M2, M3, …)
- 模型生成步骤（M1, M2, M3, …）

Match them based on semantics, logical function, and computational intent, and output the correspondence between the two.
请基于语义、逻辑功能与计算意图进行匹配，并输出两者的对应关系。

Matching rules
匹配规则
1. Matching criteria:
1. 匹配判据：
A model step is considered a match to a standard step if it satisfies at least one of the following:
当模型步骤满足以下至少一项时，可认为其与某个标准步骤“匹配”：
- Semantic equivalence: the mathematical meaning is essentially the same;
- 语义等价：数学含义本质一致；
- Same logical role: it serves the same function in the reasoning chain (e.g., set up an equation, substitute to evaluate, use the discriminant to determine the nature of roots, etc.);
- 逻辑角色相同：在推理链中承担相同功能（如“建立方程”“代入求值”“用判别式判断根的性质”等）；
- Same computational intent: the wording differs, but it performs the same type of mathematical operation (e.g., different forms of the same equation, equivalent transformations).
- 计算意图相同：表述不同，但执行的是同类型数学操作（如同一方程的不同等价形式、等价变形）。
2. Alignment constraints (recommended for step splitting):
2. 对齐约束（推荐用于步骤拆分的情况）：
- A standard step Si may align to zero, one, or multiple model steps (the model may split a standard step);
- 一个标准步骤 Si 可以对应 0 个、1 个或多个模型步骤（模型可能将一个标准步骤拆成多步表达）；
- Each model step Mj can be assigned to at most one standard step Si (no double counting);
- 每个模型步骤 Mj 最多只能分配给一个标准步骤 Si（避免重复计数）；
- If no reasonable match can be found, output an empty list for that Si.
- 若找不到合理匹配，则该 Si 输出空列表 []。
3. Focus on problem-solving function rather than surface wording.
3. 关注解题功能而非表面措辞。
- Ignore arithmetic errors: if the model computes incorrectly but performs the same mathematical step (e.g., substitution, computing a discriminant, differentiating), it is still considered fully matchable.
- 忽略算术错误：如果模型计算结果有误，但执行了相同类型的数学步骤（如代入、计算判别式、求导），仍视为可完全匹配。

Output requirements
输出要求
- Output must be a JSON object; do not explain or generate any extra text;
- 输出必须是 JSON 对象；不要解释或生成任何额外文本；
- Keys are standard step IDs (e.g., "S1"), and values are lists of matched model step IDs (e.g., ["M1", "M3"]);
- 键为标准步骤编号（如 "S1"），值为匹配到的模型步骤编号列表（如 ["M1", "M3"]）；
- If there is no match or you are unsure, the value should be an empty list [].
- 如果没有匹配或你不确定，值应为空列表 []。

Output example
输出示例
{

"S1": ["M1"],

"S2": ["M2", "M4"],

"S3": []

}

Notes
备注
- Output only the JSON result; do not explain or generate any extra text.
- 只输出 JSON 结果；不要解释或生成任何额外文本。

Figure 23: Judge Prompt for Step Alignment to compute ESGR and SSCR in Comprehensive Explanation Task(a).

Judge Prompt for Step Alignment(b)

System:
Please align the following model steps with the standard steps. Output only the JSON result.
请将下面的模型步骤与标准步骤进行对齐匹配。只输出 JSON 结果。

Standard steps:
标准步骤：
{{gt_steps}}

Model steps:
模型步骤：
{{model_steps}}

Figure 24: Judge Prompt for Step Alignment to compute ESGR and SSCR in Comprehensive Explanation Task(b).

Judge Prompt for Step Probing

System:
You are a “ solution-step coverage evaluation expert ” . Based on the given materials, determine whether the model expanded content semantically covers every detailed step corresponding to the detailed step IDs associated with this section.
你是一名“ 解题步骤覆盖度评估专家 ”。请基于给定材料，判断模型扩写内容在语义上是否覆盖了本节关联的详细步骤 ID 列表中每一个详细步骤。
Evaluation rules (must follow):
评估规则（必须遵守）：
1) “ Covered ” means: the model expansion semantically expresses the key action / reasoning / conclusion of that detailed step (allowing paraphrasing, order changes, merged statements, and implicit expressions).
1）“ 覆盖 ”指：模型扩写在语义上表达了该详细步骤的关键动作/推理/结论（允许改写、顺序调整、合并陈述与隐式表达）。
2) If the expansion only gives the result without the key action (e.g., only says “ thus ” or “ obviously ” but does not express the core of the step), it is considered NOT covered.
2）若扩写只给出结果而没有关键动作（例如只说“ 因此/显然 ”但未表达步骤核心内容），则视为未覆盖。
3) Only evaluate the steps in the detailed step ID list associated with this section; do not output IDs that are not in this list.
3）只评估本节关联的详细步骤 ID 列表中的步骤；不要输出不在该列表中的 ID。
4) Output must be extremely concise: output only the IDs of steps not covered, one ID per line, sorted in ascending numerical order.
4）输出必须极其简洁：仅输出“ 未覆盖步骤的 ID ”，每行一个 ID，并按数值升序排序。
5) If all steps are covered, output and only output: All covered
5）如果全部覆盖，则仅输出且只输出：All covered
6) Provide brief explanations; do not restate the problem; do not output any extra characters (including punctuation, parentheses, or bullet symbols).
6）给出简要解释；不要复述题目；不要输出任何多余字符（包括标点、括号或项目符号）。

User:
Problem text
题目文本
{{test_content}}

Problem image
题目图片
{{test_image}}

Full detailed solution-step library
完整的详细解题步骤库
{{steps}}

Detailed step ID list associated with this section
本节关联的详细步骤 ID 列表
{{cover_step_ids}}

Macro-step description
宏步骤描述
{{step_summary}}

Model expanded content
模型扩写内容
{{step_reply}}

Figure 25: Judge Prompt for Step Alignment to compute SSCR in Step Probing Task.

Judge Prompt for Knowledge Explanation

System:
You are an educational evaluation expert. Your task is to score and analyze the model ’ s explanation of Chinese high school problems from multiple dimensions. Please evaluate based on the inputs below.
你是一名教育评估专家。你的任务是从多个维度对模型对中国高中题目的讲解进行评分与分析。请基于下列输入进行评估。
Task:
任务：
Evaluate and score the model ’ s explanation only from the single dimension of “ knowledge point coverage ”. The score range is 0 to 100, and you must make a precise assessment rather than a simple mapping.
仅从“知识点覆盖度”这一单一维度评估并给模型讲解打分。分数范围为 0 – 100，必须进行精细评估而非简单映射。

Scoring basis:
评分依据：
Judge based on the five coverage levels below, but assign a continuous score from 0 – 100 to reflect subtle differences:
参考以下五个覆盖等级进行判断，但需要给出 0 – 100 的连续分数以反映细微差异：
- Full coverage (90 – 100): Accurately covers all key knowledge points (including definitions, conditions, core conclusions, key steps, and important notes) with no obvious omissions or errors.
- 完全覆盖（90 – 100）：准确覆盖所有关键知识点（包括定义、条件、核心结论、关键步骤与重要注意事项），无明显遗漏或错误。
- Mostly complete (75 – 89): Covers most key points; may miss a small amount of minor content, but overall is clear and systematic.
- 基本完整（75 – 89）：覆盖大多数关键点；可能遗漏少量次要内容，但整体清晰且体系化。
- Moderate coverage (50 – 74): Mentions some key points but misses multiple important parts or explanations are insufficient.
- 中等覆盖（50 – 74）：提到部分关键点，但遗漏多个重要部分或解释不充分。
- Weak coverage (25 – 49): Covers only a few related points; explanation is scattered, poorly structured, and lacks key steps or conclusions.
- 覆盖较弱（25 – 49）：仅覆盖少量相关点；讲解零散、结构较差，缺少关键步骤或结论。
- Almost no coverage (0 – 24): Content is barely related to the knowledge point, lacks substantive explanation, or is completely off-topic.
- 几乎无覆盖（0 – 24）：内容与知识点关联很弱，缺乏实质性讲解，或完全跑题。

Output format (must be exactly as follows):
输出格式（必须严格如下）：
X (score)
X（分数）
Coverage difference summary: one sentence stating which key points are missing / which key points are covered compared with the standard answer.
覆盖差异总结：用一句话说明与标准答案相比缺失了哪些关键点/覆盖了哪些关键点。

User:
Problem content
题目内容
{{test_content}}

Knowledge point to explain
需要讲解的知识点
{{knowledge_points_result}}

Standard answer for the knowledge point
该知识点的标准答案
{{knowledge_describe}}

Model explanation
模型讲解
{{model_reply}}

Figure 26: Judge prompt for Knowledge Explanation Task.

Prompt for Student with Guidance on Socratic Questioning

System:
You are a Chinese high school student. The knowledge and methods you use to solve problems must not go beyond the Chinese high school curriculum standards (K12).
你是一名中国高中学生。你用于解题的知识与方法不得超出中国高中课程标准（K12）范围。

User:
Problem:
题目：
{{question}}

Hint:
提示：
{{hint}}

Figure 27: Student inference prompt for Socratic Questioning Task.

Judge Prompt for Error Task

System:
You are a math education evaluation expert. Your task is to strictly assess the model ’ s “ error-correction result ” and determine whether it satisfies each requirement according to the “ evaluation criteria ”.
你是一名数学教育评估专家。你的任务是严格评估模型给出的“纠错结果”，并依据“评估标准”判断其是否满足每一项要求。

Rules
规则
1. You only judge whether the model ’ s error-correction result meets the criteria; you do not do mathematical reasoning or error correction yourself.
1）你只判断模型的纠错结果是否满足标准；你不进行数学推理或亲自纠错。
2. You must use the provided ground truth (root cause, propagation mechanism, affected steps, final outcome, correct answer) as the evaluation basis.
2）你必须以提供的 ground truth（错误根因、传播机制、受影响步骤、最终结果、正确答案）作为评估依据。
3. All judgments must be strictly output as boolean values true or false.
3）所有判断必须严格输出为布尔值 true 或 false。
4. Your final output must be a valid JSON object corresponding to the pass/fail status of the five criteria.
4）最终输出必须是有效的 JSON 对象，对应五项标准的通过/未通过状态。

Output format example (must follow):
输出格式示例（必须遵循）：
{

"root_cause": true,

"propagation_mechanism": false,

"affected_steps": true,

"final_outcome": true,

"correct_answer": false

}

Do not output anything other than the JSON.
除 JSON 之外不要输出任何内容。

User:
Problem:
题目：
{{question}}

Student ’ s solution process:
学生解题过程：
{{stu_process}}

Model ’ s error-correction result (to be evaluated):
模型纠错结果（待评估）：
{{model_correction}}

Evaluation criteria:
评估标准：
1. Whether the student ’ s root cause of error is correctly identified; reference: {{root_cause}}
1）是否正确识别学生错误的根因；参考：{{root_cause}}

2. Whether the error propagation mechanism is correctly described; reference: {{propagation_mechanism}}
2）是否正确描述错误传播机制；参考：{{propagation_mechanism}}

3. Whether the affected steps are correctly identified; reference: {{affected_steps}}
3）是否正确识别受影响步骤；参考：{{affected_steps}}

4. Whether the final outcome caused by the error is identified; reference: {{final_outcome}}
4）是否识别由错误导致的最终结果；参考：{{final_outcome}}

5. Whether the corrected final answer is provided; reference: {{correct_answer}}
5）是否给出纠正后的最终答案；参考：{{correct_answer}}

Figure 28: Judge prompt for Error Correction Task.

Solution Explanation Template

```
{
    "id": "", Question ID
    "typesName": "", Question type (e.g., single-choice, multiple-choice, fill-in-the-blank, essay)
    "diffName": "", Difficulty label (e.g., Easy/Medium/Hard)
    "original_knowledge": "", Primary knowledge point / topic area (Level 1)
    "original_knowledge_level_2": "", Secondary knowledge point under Level 1 (Level 2)
    "test_content": "", Problem statement / prompt
    "solution": "", Reference solution / standard answer
    "test_image": [], Images for the problem statement (array: URLs/paths/Base64, etc.)
    "answer_image": [], Images for the final answer (array)
    "analytic_image": [], Images for the explanation/analysis (array)
    "steps": [ Step-by-step solution list (array)
        {
            "id": "", Step ID
            "content": "", Step content (derivation/explanation text for this step)
            "section": "", Step section/tag (e.g., Analysis / Solution / Conclusion)
        }
    ],
    "diff": 0.0, Numeric difficulty value (e.g., normalized difficulty score)
    "gradeName": "", Grade level (e.g., Grade 10/11/12)
    "score": "", Score/points for this question (e.g., "5" or "5 points")
    "year": "", Year (e.g., 2023; exam/paper year)
    "irt": 0.0, IRT parameter/value (Item Response Theory metric; system-specific)
    "kl": [ Knowledge label list
        {
            "id": "", Knowledge label ID
            "name": "", Knowledge label name
            "rate": "" Knowledge weight/rate (e.g., importance/proportion; format depends on system)
        }
    ],
    "extraAttr": {}, Extra attributes / extensible fields (object/dict; can be empty)
}
```

Figure 29: Template for Comprehensive Explanation.

Solution Specification Template

```
{
    "id": "", Question ID
    "typesName": "", Question type (e.g., single-choice, multiple-choice, fill-in-the-blank, essay)
    "diffName": "", Difficulty label (e.g., Easy/Medium/Hard)
    "original_knowledge_level_1": "", Primary knowledge point / topic area (Level 1)
    "original_knowledge_level_2": "", Secondary knowledge point under Level 1 (Level 2)
    "test_content": "", Problem statement / prompt
    "solution": "", Reference solution / standard answer
    "test_image": [], Images for the problem statement (array: URLs/paths/Base64, etc.)
    "answer_image": [], Images for the final answer (array)
    "analytic_image": [], Images for the explanation/analysis (array)
    "steps": [ Step-by-step solution list (array)
        {
            "id": "", Step ID
            "content": "", Step content (derivation/explanation text for this step)
            "section": "", Step section/tag (e.g., Analysis / Solution / Conclusion)
        }
    ],
    "solution_name": "", Solution method/name (e.g., "Method 1", "Algebraic approach", "Geometric approach")
    "gradeName": "", Grade level (e.g., Grade 10/11/12)
    "year": "", Year (e.g., 2023; exam/paper year)
        ""
    ],
}
```

Figure 30: Template for Solution Specification.

Step Probe Template

```
{
    "id": "", Question ID
    "typesName": "", Question type (e.g., single-choice, multiple-choice, fill-in-the-blank, essay)
    "diffName": "", Difficulty label (e.g., Easy/Medium/Hard)
    "original_knowledge": "", Primary knowledge point / topic area (Level 1)
    "original_knowledge_level_2": "", Secondary knowledge point under Level 1 (Level 2)
    "test_content": "", Problem statement / prompt
    "solution": "", Reference solution / standard answer
    "test_image": [], Images for the problem statement (array: URLs/paths/Base64, etc.)
    "answer_image": [], Images for the final answer (array)
    "analytic_image": [], Images for the explanation/analysis (array)
    "steps": [ Step-by-step solution list (array)
        {
            "id": "", Step ID
            "content": "", Step content (derivation/explanation text for this step)
            "section": "", Step section/tag (e.g., Analysis / Solution / Conclusion)
        }
    ],
    "gradeName": "", Grade level (e.g., Grade 10/11/12)
    "year": "", Year (e.g., 2023; exam/paper year)
    "aggregated_steps": { Aggregated step information (merged steps)
        "step_summary": "", Summary of aggregated steps (high-level reasoning/outline)
        "cover_step_ids": [], Step IDs covered by this summary (array of step IDs)
    }
}
```

Figure 31: Template for Step Probe.

Knowledge Interpretation Template

```
{
    "id": "", Question ID
    "typesName": "", Question type (e.g., single-choice, multiple-choice, fill-in-the-blank, essay)
    "diffName": "", Difficulty label (e.g., Easy/Medium/Hard)
    "original_knowledge": "", Primary knowledge point / topic area (Level 1)
    "original_knowledge_level_2": "", Secondary knowledge point under Level 1 (Level 2)
    "test_content": "", Problem statement / prompt
    "solution": "", Reference solution / standard answer
    "test_image": [], Images for the problem statement (array: URLs/paths/Base64, etc.)
    "answer_image": [], Images for the final answer (array)
    "analytic_image": [], Images for the explanation/analysis (array)
    "steps": [ Step-by-step solution list (array)
        {
            "id": "", Step ID
            "content": "", Step content (derivation/explanation text for this step)
        }
    ],
    "gradeName": "", Grade level (e.g., Grade 10/11/12)
    "year": "", Year (e.g., 2023; exam/paper year)
    "knowledge_points": "", Knowledge points (raw / extracted knowledge-point text)
    "knowledge_points_result": "", Knowledge-point extraction result (structured/normalized output)
    "knowledge_describe": "", Knowledge-point description / explanation text
    "is_incomplete": false, Whether the sample is incomplete (true/false)
}
```

Figure 32: Template for Knowledge Explanation.

Pedagogical Guidance Template

```
{
    "id": "", Question ID
    "typesName": "", Question type (e.g., single-choice, multiple-choice, fill-in-the-blank, essay)
    "diffName": "", Difficulty label (e.g., Easy/Medium/Hard)
    "original_knowledge": "", Primary knowledge point / topic area (Level 1)
    "original_knowledge_level_2": "", Secondary knowledge point under Level 1 (Level 2)
    "test_content": "", Problem statement / prompt
    "solution": "", Reference solution / standard answer
    "test_image": [], Images for the problem statement (array: URLs/paths/Base64, etc.)
    "answer_image": [], Images for the final answer (array)
    "analytic_image": [], Images for the explanation/analysis (array)
    "steps": [ Step-by-step solution list (array)
        {
            "id": "", Step ID
            "content": "", Step content (derivation/explanation text for this step)
        }
    ],
    "gradeName": "", Grade level (e.g., Grade 10/11/12)
    "year": "", Year (e.g., 2023; exam/paper year)
        ""
    ],
}
```

Figure 33: Template for Socratic Questioning.

Error Diagnosis Template

```
{
    "id": "", Question ID
    "typesName": "", Question type (e.g., single-choice, multiple-choice, fill-in-the-blank, essay)
    "diffName": "", Difficulty label (e.g., Easy/Medium/Hard)
    "original_knowledge": "", Primary knowledge point / topic area (Level 1)
    "original_knowledge_level_2": "", Secondary knowledge point under Level 1 (Level 2)
    "test_content": "", Problem statement / prompt
    "solution": "", Reference solution / standard answer (may include formulas)
    "test_image": [], Images for the problem statement (array: URLs/paths/Base64, etc.)
    "answer_image": [], Images for the final answer (array)
    "analytic_image": [], Images for the explanation/analysis (array)
    "steps": [
        {
            "id": "", Step ID (or step index)
            "content": "", Step content (derivation/explanation text for this step)
        }
    ],
    "gradeName": "", Grade level (e.g., Grade 10/11/12)
    "year": "", Year (e.g., 2023; exam/paper year)
    "wrong_task": { Wrong-solution simulation and error-propagation task info
        "simulated_solution": "", Simulated incorrect solution (for error diagnosis)
        "error_category": "", Error category (e.g., concept error, calculation error, misreading, method error)
        "propagation_model": "", Error propagation pattern/model (how the error spreads)
        "propagation_intensity": "", Propagation intensity / severity (e.g., low/medium/high or a numeric level)
        "error_analysis": { Detailed error analysis
            "root_cause": "", Root cause (the initial mistake / underlying misconception)
            "propagation_mechanism": "", Propagation mechanism (how it affects subsequent reasoning)
            "affected_steps": "", Affected steps (e.g., step IDs or a described range)
            "final_outcome": "" Final outcome (what incorrect conclusion/answer it leads to)
        }
    }
}
```

Figure 34: Template for Error Correction.