

LongTutor: Benchmarking Large Language Models for Long-term Personalized Tutoring

Ning Li¹, Zheng Zhang¹, Zhenya Huang^{1,2}, Rui Li¹, Yi Zhan¹, Yinbo Luo¹,
Qi Liu^{1,2}, Enhong Chen^{1*},

¹State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China,

²Institute of Artificial Intelligence, Hefei Comprehensive National Science Center,

{ningli03, zhangzheng}@mail.ustc.edu.cn, huangzhy@ustc.edu.cn,

{ruili2000, zy0119, lh_kanade}@mail.ustc.edu.cn, {qiliuq1, cheneh}@ustc.edu.cn

Abstract

The rapid advancement of large language models (LLMs) has driven the deployment of LLM-based AI tutors on online learning platforms. This widespread adoption highlights an urgent need for systematic benchmarks to evaluate their tutoring capabilities. However, existing evaluations predominantly focus on isolated, short-term interactions, overlooking the inherently long-term nature of learning. To bridge this gap, we introduce LongTutor, a benchmark for long-term personalized tutoring grounded in formative assessment theory. Built from expert-annotated real-world learning logs, LongTutor evaluates LLMs across three progressive tasks: historical evidence acquisition, knowledge state diagnosis, and adaptive teaching action. Our experiments reveal a critical capability mismatch: while LLMs excel at evidence acquisition, they struggle to effectively leverage long-term history for accurate diagnosis and adaptive teaching. To enable scalable benchmark expansion, we further propose an automated generator-verifier pipeline, paving the way toward truly long-term AI tutoring systems. We release our code and dataset at <https://github.com/liano3/LongTutor>.

1 Introduction

The rapid advancement of large language models (LLMs) has significantly empowered intelligent education (Kasneji et al., 2023; Xi et al., 2025; Chu et al., 2025), leading many online learning platforms, such as Khanmigo¹, Duolingo² and iFLY-TEK³, to deploy LLM-based AI tutors that guide students through the learning process rather than simply providing answers. Specifically, the AI tutor should provide targeted guidance to assist students when they provide incorrect answers. Along-

*Corresponding author.

¹<https://khanmigo.ai/>

²<https://www.duolingo.com/>

³<https://edu.iflytek.com/>

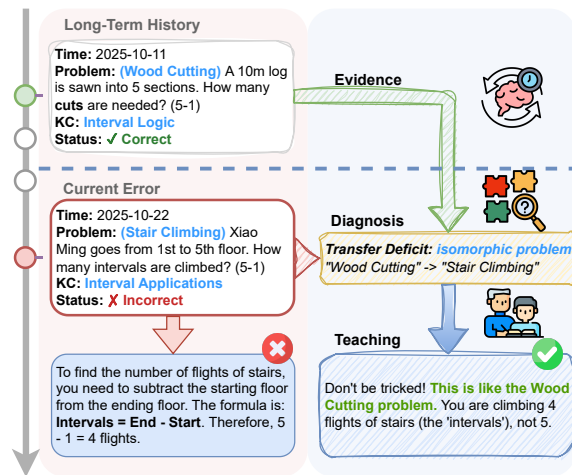


Figure 1: Illustration of long-term personalized tutoring. The system leverages the student’s mastery of the “Wood Cutting” problem (11 days prior) to teach an isomorphic “Stair Climbing” problem (current). Unlike the context-isolated response (Left) which offers generic advice, the personalized tutoring (Right) explicitly references the user’s specific history to facilitate effective analogical transfer.

side the growing adoption of these tutoring systems, there is a rising need to systematically evaluate their tutoring capabilities. To this end, the research community has proposed a variety of educational benchmarks. In particular, specialized datasets such as MathDial (Macina et al., 2023) emphasize multi-turn instructional scaffolding (Wood et al., 1976), while MathTutorBench (Macina et al., 2025) assess a broader spectrum of instructional competencies, ranging from error localization to the application of teaching strategies.

However, a critical gap exists in current benchmarks: they predominantly evaluate tutoring as isolated, short-term sessions, overlooking the vital role of long-term learning histories (Corbett and Anderson, 1994), which is defined as a chronological sequence of a student’s problem-solving

records. According to formative assessment theory (Black and Wiliam, 1998), effective tutoring should leverage students’ long-term learning history and follow an Evidence–Diagnosis–Teaching paradigm (VanLehn, 2006). In this work, we focus on structured, verifiable subjects such as mathematics, where these latent states and downstream teaching outcomes can be annotated with high reliability. As illustrated in Figure 1, rather than providing generic, context-isolated hints for a "Stair Climbing" problem, the tutor should first elicit evidence (a prior successful experience on the structurally analogous "Wood Cutting" problem, 11 days prior) from the student’s long-term learning history. Based on this evidence, the tutor diagnoses the student’s current knowledge state as Transfer Deficit, which means failure to recognize the shared structure between the two problems. Guided by this diagnosis, the tutor adapts its teaching by linking current difficulties to prior experiences, thereby scaffolding new understanding onto existing knowledge.

Aligned with the Evidence–Diagnosis–Teaching paradigm, we introduce LongTutor, the first benchmark dedicated to long-term personalized tutoring. It contains three progressive tasks: **(1) Evidence Acquisition** tests the tutor’s fundamental ability to acquire related evidences from long-term history as the basis for diagnosis, including information extraction (e.g., question text, knowledge points) from single records, multi-session reasoning across multiple interactions (Wei et al., 2022) (e.g., number of questions done in a specific day), and hallucination checks to prevent fabrication (Zhang et al., 2025a) (e.g., trap questions with false premises). **(2) State Diagnosis** evaluates the tutor’s analytical ability to perform cognitive inference, rooting student’s current error to specific knowledge states (e.g., conceptual gap), which aligned with the Knowledge Dimensions of the Revised Bloom’s Taxonomy (Anderson and Krathwohl, 2001). **(3) Teaching Action** assesses the tutor’s ability to scaffold the problem-solving process using adaptive teaching strategies (e.g., analogical transfer (Gick and Holyoak, 1980)) that are derived from the acquired evidence and diagnosed knowledge states.

Conducting effective evaluation on these tasks requires long-term student learning trajectories paired with high-quality diagnostic and instructional annotations. However, such data are extremely scarce. While existing datasets like MoocRadar (Yu et al., 2023) and XES3G5M (Liu

et al., 2023b) contain long-term histories (i.e., sequences of problem-solving records containing problem texts and timestamps), they lack the fine-grained expert labels necessary to assess state diagnosis and teaching actions.

To address the data scarcity challenge, we enriched real-world problem-solving logs (from XES3G5M (Liu et al., 2023b)) through human expert annotation. To ensure the reliability of the ground truth, we adhered to a rigorous quality control protocol involving independent dual-expert labeling followed by adjudication. This process yielded a high inter-annotator agreement (with Cohen’s Kappa = 0.90), confirming the trustworthiness of our labels. Consequently, this expert-annotated dataset serves as the core test set within our benchmark.

Using this benchmark to evaluate mainstream LLMs, we observe a distinct performance gap: while models demonstrate proficiency in basic evidence acquisition, they struggle significantly with the more complex tasks of state diagnosis and teaching action. This reveals that current LLMs fail to effectively leverage students’ learning histories for long-term personalized tutoring, indicating substantial room for improvement. Crucially, these findings underscore the significance of our benchmark in advancing personalized education.

Moreover, as the annotation process is labor-intensive and time-consuming, we propose an automated generator–verifier pipeline to support scalable expansion across larger datasets and multiple subject domains. The pipeline synthesizes high-quality data from raw logs and demonstrates strong label and ranking consistency against human annotations (Cohen’s Kappa = 0.81), providing a reliable and low-cost path for benchmark expansion.

2 Related Work

2.1 Benchmarks for LLMs in Education

The evaluation of LLMs in education has transitioned from assessing standardized problem-solving to evaluating pedagogical efficacy. Initially, benchmarks like GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021) and LILA (Mishra et al., 2022) focused on the model’s foundational reasoning capabilities. As LLMs demonstrated potential as tutors, the research focus expanded to their instructional roles. Datasets such as CIMA (Stasaski et al., 2020), MathDial (Macina et al., 2023) and SocraTeach (Liu et al., 2024b)

Benchmark/Dataset	Domain	Statistics		Data Capabilities		
		# Sess.	# Q	L.T. History	Diag. Label	Teach. Label
<i>General Long-Context</i>						
LOCOMO (Maharana et al., 2024)	Open-Domain	1k	7.5k	✓	✗	✗
LongMemEval (Wu et al., 2025)	Assistant	50k	500	✓	✗	✗
<i>Educational Dialogue</i>						
MathDial (Macina et al., 2023)	Math Tutoring	2.9k	-	✗	✗	✓
SocraTeach (Liu et al., 2024b)	Math Tutoring	35k	1.4k	✗	✗	✓
MathTutorBench (Macina et al., 2025)	Math Tutoring	4.5k	9.1k	✗	✓	✓
EduBench (Xu et al., 2025)	General Edu	4k	4k	✗	✗	✓
<i>Knowledge Tracing</i>						
MoocRadar (Yu et al., 2023)	University MOOCs	14k	-	✓	✗	✗
XES3G5M (Liu et al., 2023b)	Math Practice Logs	3.6k	-	✓	✓ [†]	✗
LongTutor (Ours)	Adaptive Tutoring	4.2k	21k	✓	✓	✓

Table 1: Comparison of LongTutor with existing datasets. **# Sess.**: Total number of dialogue sessions or student’s problem-solving sequences. **# Q**: Total number of questions or test samples. **L.T. History**: Whether the dataset provides long-term student interaction history (sequences of problem-solving records). **Diag. Label**: Availability of fine-grained diagnosis annotations (e.g., knowledge states, misconceptions). **Teach. Label**: Availability of adaptive teaching strategy or action annotations. [†]: Diagnosis is limited to binary correctness labels rather than semantic cognitive reasoning.

assess the ability to provide multi-turn scaffolding, while benchmarks like MathTutorBench (Macina et al., 2025) and EduBench (Xu et al., 2025) evaluate a wider spectrum of competencies, ranging from mistake location to pedagogical application.

However, a critical gap remains: these benchmarks predominantly treat tutoring as isolated snapshots. They overlook the long-term nature of learning, bypassing the complex requirement of mining a student’s history to infer latent knowledge states. While datasets like MoocRadar (Yu et al., 2023) and XES3G5M (Liu et al., 2023b) contain students’ long-term learning histories (i.e., sequences of problem-solving records with timestamps), they lack the fine-grained diagnostic labels (e.g., conceptual gap) and teaching action annotations necessary to assess the long-term personalized tutoring ability. This diagnostic step is also closely related to the broader literature on cognitive diagnosis, which studies how to infer latent learner states from response histories (Wang et al., 2024).

2.2 Long-Term Interaction Benchmarks

Beyond the educational domain, evaluating LLMs in long-term scenarios has become a critical research direction. In open-domain dialogue, benchmarks like MSC (Multi-Session Chat) (Xu et al., 2022) evaluate the model’s ability to maintain persona consistency across disparate sessions. Similarly, long-context benchmarks like LOCOMO (Maharana et al., 2024), InfiniteBench (Zhang et al., 2024) and Long-

MemEval (Wu et al., 2025) test the model’s capacity to process and reason over extended historical inputs, often utilizing "needle-in-a-haystack" retrieval tasks (Liu et al., 2024c).

Despite these advancements, a significant misalignment exists between these open-domain benchmarks and the requirements of educational tutoring. These benchmarks typically focus on information retention (e.g., recalling a user’s past statement), but tutoring requires inferring latent knowledge states and take teaching action.

3 LongTutor Benchmark

In this section, we formally define the task of long-term personalized tutoring and detail the construction of LongTutor. As illustrated in Figure 2, our benchmark structures the evaluation around three capabilities: evidence acquisition, state diagnosis, and teaching action.

3.1 Task Formulation

We formulate the long-term personalized tutoring as a sequential process grounded in a student’s long-term history. Let $\mathcal{H}_t = \{h_1, h_2, \dots, h_{t-1}\}$ denote the student’s learning history, where each record $h_i = (q_i, k_i, \tau_i, c_i)$ consists of the question text q_i , associated knowledge points k_i , a timestamp τ_i , and the binary correctness label $c_i \in \{0, 1\}$. At the current timestamp t , the student encounters a new problem q_t and commits an error. The model \mathcal{M} is evaluated with three progressive tasks: produce evidence query answers \mathcal{A}_{evi} , diagnostic knowledge

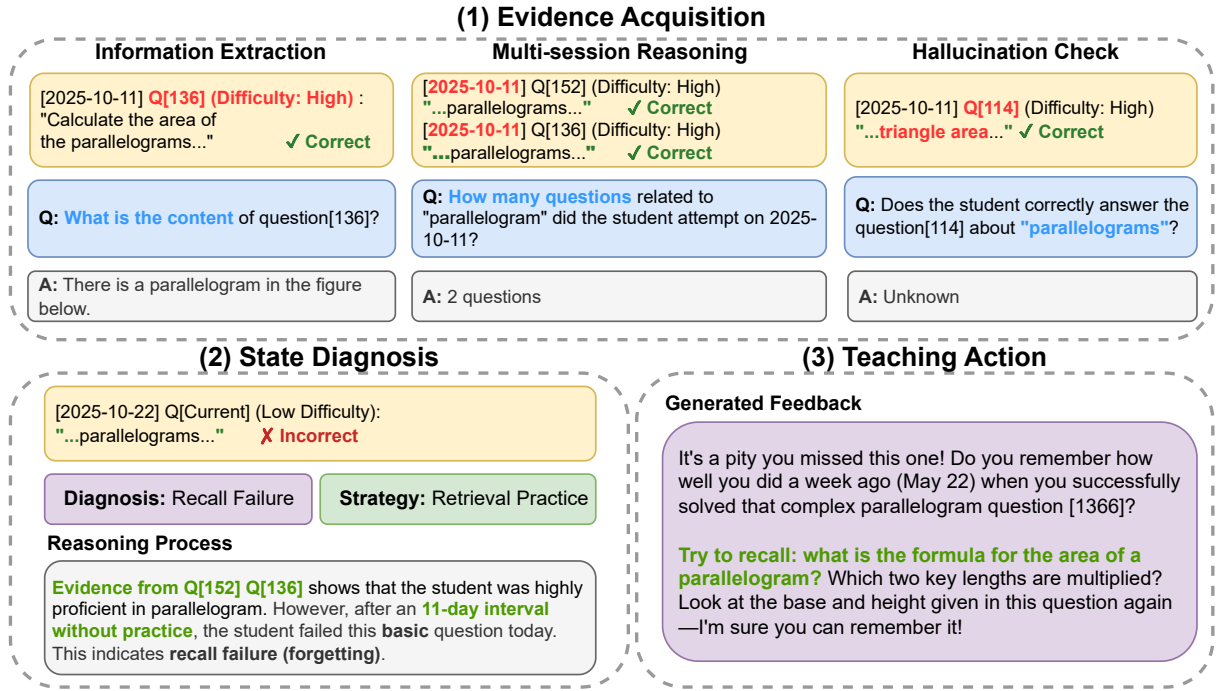


Figure 2: Overview of the three-level evaluation tasks in LongTutor. (1) Evidence Acquisition tests the model’s ability to extract facts, reason across multiple sessions, and detect hallucinations. (2) State Diagnosis evaluates the inference of student states by analyzing the learner’s history. (3) Teaching Action assesses the generation of personalized feedback that explicitly leverages these diagnostic insights.

state label y_{diag} , and teaching response R_{int} .

Task 1: Evidence Acquisition. The prerequisite for personalized tutoring is a robust understanding of the student’s long-term history. To ensure subsequent diagnoses are grounded in fact rather than hallucination, this task evaluates the ability to acquire historical information across three dimensions: (1) **Information Extraction** targets specific details within a single record, (2) **Multi-session Reasoning** requires aggregating patterns across multiple interactions, and (3) **Hallucination Check** tests the robustness against false premises. As shown in Figure 2, the model must extract the problem text from Q[136], count the "parallelogram" questions on 2025-10-11 by combining records Q[152] and Q[136], and correctly output "Unknown" for Q[114] when queried with a mismatched topic (parallelograms vs. triangles). These probes operationalize trace-level grounding under information overload, rather than serving as standalone IR exercises.

Task 2: State Diagnosis. Once the evidence is acquired, the model diagnoses the student’s latent knowledge state to identify why the error occurred to support adaptive teaching action. We formulate this as a multi-class classification task. To

ensure pedagogical rigor, we ground our diagnostic framework in the Revised Bloom’s Taxonomy (Anderson and Krathwohl, 2001), which categorizes knowledge into four dimensions: Factual, Conceptual, Procedural, and Metacognitive. We map these dimensions to a specific label set: $\mathcal{Y} = \{Recall Failure, Conceptual Gap, Procedural Error, Transfer Deficit\}$. This task requires the model to predict the correct diagnosis label $y_{diag} \in \mathcal{Y}$ based on historical evidence, thereby dictating the subsequent teaching action. Detailed definitions and symptomatic descriptions of this taxonomy are provided in Appendix D.

Task 3: Teaching Action. Finally, the ultimate objective is to translate the diagnostic insight into an effective teaching action. This task evaluates whether the model can generate a teaching response R_{int} explicitly leveraging the student’s long-term learning history and executing history-aware teaching strategies. For instance, as shown in Figure 2, the model observes that the student has a Recall Failure on a basic parallelogram question after an 11-day interval of successful practice on complex ones. The generated response directly references the student’s past success on a complex parallelogram problem (Q[136]) from 11 days prior,

employs a Retrieval Practice (Roediger III and Karpicke, 2006) strategy by asking the student to recall the formula.

3.2 Evaluation Protocols

We propose a multi-dimensional evaluation protocol suite corresponding to the three subtasks. Detailed definitions and calculation formulas are provided in Appendix G.

Evidence Acquisition. We evaluate retrieval precision using Semantic Accuracy (SA). This metric extends standard Exact Match (EM) by incorporating an LLM-based semantic equivalence check to account for correct answers that are paraphrased.

State Diagnosis. We treat state diagnosis as a multi-class classification task. To address potential class imbalances among cognitive states, we primarily report the Macro-F1 Score. We also provide Accuracy for a comprehensive performance view.

Teaching Action. Due to the open-ended nature of teaching, we adopt a dual-view evaluation. We report ROUGE-L (Lin, 2004) only as a supplementary lexical-overlap reference, since pedagogically strong responses need not match the reference wording. Our primary assessment uses an LLM-as-a-Judge framework (Liu et al., 2023a; Zheng et al., 2023) with an instance-specific pedagogical rubric. Specifically, the judge evaluates each response based on four dimensions: (i) History Utilization measures how effectively the model leverages the student’s learning history; (ii) Strategy Alignment assesses whether the feedback adheres to the diagnosed teaching strategy (Shute, 2008); (iii) Coherence evaluates the logical structure and linguistic clarity of the response; and (iv) Appropriateness examines whether the difficulty are appropriate (Vygotsky, 1978).

3.3 Data Annotation

To construct reliable evaluation for long-term personalized tutoring, we curate a human-annotated dataset with ground-truth labels for the three tasks.

Data Preparation. The proposed tasks necessitate detailed problem texts for semantic evidence acquisition and timestamps for tracking temporal dynamics (e.g., forgetting behaviors). Consequently, we build upon XES3G5M (Liu et al., 2023b), a dataset of real-world primary school mathematics problem-solving sequences. To ensure long-term history, we filter for sequences with

length $L > 200$ and a span exceeding 7 days. This yields 3,437 structured histories, from which 1,000 samples were randomly selected for manual annotation as our gold-standard test set.

Annotation Protocol. To ensure the highest standard of pedagogical validity, we recruited three highly qualified domain experts, all holding advanced degrees in Education and possessing profound theoretical knowledge alongside practical teaching experience. Following established annotation practices (Klie et al., 2024; Macina et al., 2023), we employ a rigorous protocol involving independent dual expert annotation followed by adjudication to ensure label reliability. Specifically, two experts independently labeled each sample to produce evidence Q&A, diagnostic labels, and teaching actions. Any discrepancies were subsequently reviewed and resolved by the third senior expert to establish the final gold standard. Please refer to Appendix E for the detailed procedures.

Quality Validation. To quantitatively assess label reliability, we computed Cohen’s Kappa (κ) between the two primary experts over the entire dataset. The resulting agreement score for diagnostic labels was $\kappa = 0.90$, indicating an exceptional level of consistency. The third senior expert was involved solely for adjudication, ensuring that the final ground truth strictly aligns with pedagogical standards. Illustrative samples covering all four diagnostic categories are provided in Appendix F.

3.4 Dataset Analysis

Figure 3 summarizes the statistics of the human-annotated set. As shown in Figure 3(a), diagnostic labels are imbalanced, with Procedural Error and Recall Failure being the most frequent. This indicates that most students’ errors come from the application of knowledge and forgetting. Figure 3(b) shows the distribution of students’ historical accuracy, covering a variety of proficiency range.

4 Experiments

4.1 Experimental Setup

Models. We evaluate a diverse set of LLMs spanning three categories: (1) Closed-source models (e.g., GPT and Gemini families); (2) Open-source General models, covering leading architectures like Llama (Grattafiori et al., 2024), Qwen (Yang et al., 2025), GLM (Du et al., 2022), and DeepSeek (Liu et al., 2024a); (3) Education-Specialized models,

Model	Task 1: Evidence				Task 2: Diagnosis			Task 3: Teaching				
	IE	MR	HC	Avg.	Acc	F1	R-L	Hist.	Strat.	Coher.	Appr.	Avg.
<i>Closed-Source General</i>												
Gemini-2.5-Pro	97.0	62.8	87.0	82.3	42.0	39.7	38.1	2.8	4.0	5.0	4.9	4.2
Gemini-2.5-Flash	95.6	57.8	84.4	79.3	46.2	28.1	34.8	2.3	3.6	4.9	4.7	3.9
GPT-5	93.4	31.2	79.0	67.9	40.0	32.2	29.4	1.4	2.4	4.5	3.9	3.1
GPT-5-Mini	77.0	27.8	82.2	62.3	47.4	29.4	29.3	1.2	2.1	4.4	3.7	2.9
<i>Open-Source General</i>												
Llama-3.1-8B-Instruct	51.6	12.2	21.0	28.3	19.0	8.0	1.9	1.0	1.0	1.2	1.2	1.1
Llama-4-Scout	52.0	10.6	37.2	33.3	40.0	15.1	25.0	1.0	1.4	3.3	3.0	2.2
Qwen2.5-7B-Instruct	65.4	13.2	33.2	37.3	40.8	16.6	26.2	1.0	1.6	3.1	3.0	2.2
Qwen3-8B	78.6	16.8	35.0	43.5	34.8	16.0	25.6	1.2	1.7	3.5	3.2	2.4
DeepSeek-V3.2	93.2	34.8	78.6	68.9	26.2	24.7	33.1	2.9	3.0	4.7	4.6	3.8
GLM-4.6	91.4	31.2	79.6	67.4	36.2	27.5	36.5	1.3	2.3	4.7	4.4	3.2
<i>Education-Specialized</i>												
EDU-Qwen2.5-7B	59.6	9.3	17.7	28.9	26.6	13.1	25.3	1.0	1.4	2.4	2.3	1.8
EduChat-R1-8B	71.6	13.2	14.8	33.2	23.8	11.6	8.9	1.1	1.3	2.4	2.3	1.8
EduChat-7B	49.4	7.2	17.4	24.7	49.4	13.8	11.7	1.0	1.2	2.3	2.3	1.7

Table 2: Main evaluation results on the LongTutor Golden Set. **Bold** denotes the best performance within each category. Metrics are organized by task: (i) Evidence Acquisition (**IE**: Information Extraction, **MR**: Multi-session Reasoning, **HC**: Hallucination Check); (ii) State Diagnosis (**Acc**, **F1**); and (iii) Teaching Action (supplementary **R-L**: ROUGE-L, and primary Judge scores: **Hist.**, **Strat.**, **Coher.**, **Appr.**). **Avg.** columns denote the mean scores for Evidence and Teaching tasks, respectively.

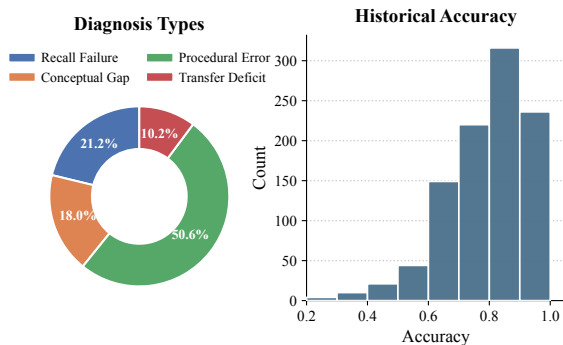


Figure 3: Statistics of LongTutor. (a) Distribution of diagnosis categories, dominated by procedural errors and transfer deficits. (b) Distribution of historical accuracy. The skew towards higher accuracy reflects the selection bias inherent in long-term active learners.

including EduChat (Dan et al., 2023) and education-tuned variants. All evaluations are conducted in a zero-shot setting (Brown et al., 2020). We employ Gemini-3-Flash as the judge for the teaching action task. Please refer to Appendix H for detailed implementation specifications, hardware configurations, and inference hyper-parameters.

4.2 Main Results

Table 2 presents the comparative results on the LongTutor benchmark. Aligned with our three evaluation dimensions, we highlight the following key findings. For supplementary analysis (e.g., instruction following issues) and performance met-

rics of additional open-source models, please refer to Appendix I and Table 7.

Evidence Acquisition: The Reasoning Gap.

While models generally exhibit proficiency in surface-level retrieval, they struggle significantly with reasoning over long contexts. As shown in Table 2, top-tier models like Gemini-2.5-Pro achieve near-perfect scores in information extraction (IE: 97.0%), yet their performance drops precipitously in multi-session reasoning (MR: 62.8%). This gap is even more pronounced in open-source models (e.g., Llama-3.1 drops from 51.6% to 12.2%), revealing that current LLMs rely heavily on local pattern matching rather than synthesizing global information. Furthermore, the varying performance on hallucination check (HC) indicates that distinct memory boundaries remain a challenge, with models often fabricating details when evidence is absent.

State Diagnosis: A Universal Challenge.

The task of inferring latent knowledge states proves to be exceptionally difficult. We observe a universal underperformance across all models, with even the strongest closed-source model (Gemini-2.5-Pro) achieving a Macro-F1 of only 39.7%. Most open-source models struggle to surpass 20% F1. This low baseline underscores the high research value of the state diagnosis task: unlike standard QA, correctly identifying specific knowledge states (e.g.,

differentiating a conceptual gap from a procedural error) requires a depth of cognitive modeling that current general-purpose LLMs have yet to master.

Teaching Action: The History Deficit and Specialization Tax. In the final teaching action stage, most models fail to effectively leverage historical evidence to personalize teaching. The History Utilization (Hist.) scores are consistently low (mostly < 2.0 on a 5-point scale), indicating that generated feedback is often generic rather than tailored to the student’s long-term learning history. Notably, we observe a "Specialization Tax": education-specific models (e.g., EduChat-7B, EDU-Qwen-2.5-7B) consistently lag behind their base model (e.g., Qwen2.5-7B) in all tasks. For instance, EDU-Qwen2.5-7B scores significantly lower in evidence acquisition (28.9 vs. 37.3) and teaching Action (1.8 vs. 2.2) compared to its base model. This suggests that current domain-specific fine-tuning (e.g., short-form or instruction-dense datasets) may inadvertently erode the models’ inherent long-context robustness.

Diagnosis Failure Modes: A Fine-grained View. Beyond the overall F1 scores, we identify three distinct failure patterns that hinder effective state diagnosis: Category Collapse, where models (e.g., GLM-4.6) completely fail to identify specific knowledge states like Recall Failure; Diagnostic Hallucination, characterized by aggressive over-prediction of complex states like Transfer Deficits; and a pervasive Majority Class Bias. These patterns suggest that current models lack the nuance required for high-stakes pedagogical reasoning. A comprehensive breakdown of these model-specific issues is provided in Appendix I.2.

Qualitative Analysis. Appendix I.3 presents a case comparison between GPT-5 and GPT-5-mini, illustrating how identifying a Transfer Deficit enables significantly lower cognitive load for the student.

4.3 Evaluation Validity and Task Dependency

We conducted four additional analyses to test the benchmark. First, on a stratified sample of 100 teaching responses, the rubric-based judge shows strong agreement with human experts (overall Spearman’s $\rho = 0.87$, 92% within ± 1 point; Table 9), supporting the reliability of LLM-as-a-Judge for constrained pedagogical verification. Second, in an agent-based interactive simulation where

Configuration	Evidence (T1)		Diagnosis (T2)		Teaching (T3)	
	Acc.	Δ	F1	Δ	Score	Δ
Gemini-2.5-Flash (Base)	79.3	-	28.1	-	3.88	-
<i>History Type</i>						
w/ Short History	47.3	-32.0	27.0	-1.1	4.00	+0.12
w/ Filtered History	87.9	+8.6	32.3	+4.2	3.90	+0.02
<i>Feature Augmentation</i>						
w/ Statistical Features	79.2	-0.1	49.2	+21.1	3.90	+0.02
<i>Instruction Strategies</i>						
w/ Teaching Instructions	78.0	-1.3	28.7	+0.6	4.63	+0.75

Table 3: Sensitivity analysis of input using Gemini-2.5-Flash. We report performance on three tasks (T1, T2, T3) and changes (Δ) relative to the baseline. **Bold** indicates the best performance in each column; colors denote **improvement** or **decline**.

a fixed GPT-5 student model is initialized with the gold cognitive state, our static Teaching Score strongly correlates with the resulting Teaching Success Rate (TSR) achieved within three turns (Spearman’s $\rho = 0.90$; Table 8), indicating that the static benchmark captures more than stylistic imitation.

Third, we quantify the dependency between Diagnosis and Teaching by replacing the gold diagnosis with model-predicted diagnoses before generation. Teaching quality drops from 4.6 to 3.9, and model-level Diagnosis F1 correlates with Teaching Score at $\rho = 0.88$ (Table 10), confirming that downstream teaching depends on accurate upstream diagnosis. Finally, a Top-20 semantic RAG baseline underperforms full-history conditioning on all three tasks (Evidence: 74.2 vs. 79.3; Diagnosis: 21.4 vs. 28.1; Teaching: 3.3 vs. 3.9; Table 11), suggesting that long-term tutoring is not a pure retrieval problem: temporal continuity and error trajectories matter.

4.4 Analysis of Long-term History

To understand how long-term history influences tutoring effectiveness, we conducted a sensitivity analysis on Gemini-2.5-Flash across three dimensions: history type, feature augmentation, and instruction strategy. The detailed configuration definitions are provided in Appendix J. The quantitative results, summarized in Table 3, lead to the following key findings.

The Trade-off Between Recall and Reasoning. Truncating to short history ($L = 20$) reveals a significant trade-off: while it causes a severe degradation in evidence acquisition capabilities (dropping from 79.3 to 47.3), the teaching score surprisingly improves ($\Delta + 0.12$). This suggests that while long

context is indispensable for evidence acquisition, it may introduce noise that distracts the model during the generation of teaching response. In contrast, Filtered History (only contains historical records sharing the same knowledge points as the target problem) achieves the best, boosting evidence acquisition to 87.9 ($\Delta + 8.6$) and diagnosis to 32.3 ($\Delta + 4.2$) by filtering out irrelevant noise while retaining semantic continuity.

The Critical Role of Statistical Features. The most striking observation for the state diagnosis task is the massive gain from incorporating statistical features. This configuration achieves a Macro-F1 of 49.2, outperforming the baseline by +21.1 points. This demonstrates that models struggle to implicitly infer knowledge states from raw textual logs, but explicitly prepending statistical summaries acts as a powerful scaffold, bridging the reasoning gap.

Impact of Instructional Detail. Incorporating detailed teaching instructions proves to be the decisive factor for the teaching action task. Explicit guidance on teaching strategies boosts the teaching score from 3.88 to 4.63 ($\Delta + 0.75$). This indicates that fine-grained instructions are necessary to align the model’s output with high-quality educational standards, ensuring the response is not just correct but also pedagogically effective.

5 Scalable Expansion

5.1 Automated Pipeline

Reliance on manual expert annotation imposes a severe bottleneck on both data scale and domain adaptability. Recent work has explored using generative agents to synthesize learner responses for intelligent education systems (Gao et al., 2025). To overcome this bottleneck in benchmark construction, we propose an automated generator-verifier pipeline that mirrors the rigorous human expert annotation protocol (Park et al., 2023). The overall architecture and workflow of this pipeline are illustrated in Figure 4.

Generator (The Annotator). Simulating the role of an expert annotator, the Generator transforms raw interaction logs into candidate benchmarks. To enhance diagnostic precision and mitigate interference from irrelevant noise, we refine the long-context input by selecting only historical records with overlapping knowledge points

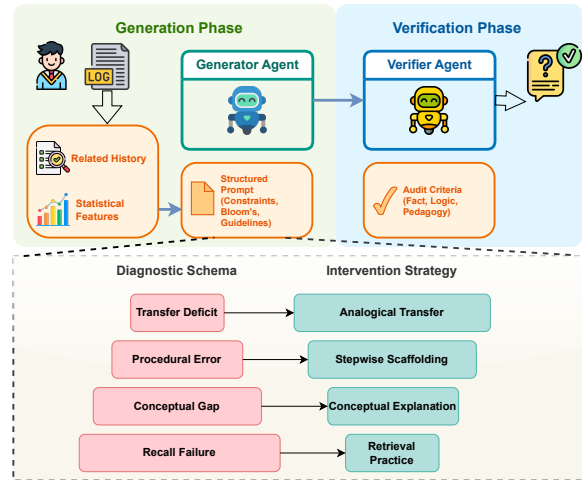


Figure 4: Overview of the scalable annotation pipeline. The upper panel illustrates the generator-verifier architecture, where the Generator synthesizes data from related history and statistical features, followed by a rigorous audit by the Verifier. The lower panel details the Diagnostic Schema, which maps knowledge states to targeted teaching strategies.

and augmenting them with statistical features (e.g., question error rates and recent accuracy). Both refinements are grounded in the empirical findings from Section 4.4. This setting is materially easier than the benchmark’s raw zero-shot diagnosis task: the Generator receives filtered histories and explicit statistical features, while the Verifier only needs to audit and repair a structured draft instead of generating labels from scratch.

Verifier (The Adjudicator). To ensure alignment with human standards, the Verifier acts as the senior adjudicator. It rigorously audits the Generator’s output against the raw history, checking for three dimensions defined in our manual protocol: factual consistency, diagnostic logic, and teaching validity. Upon detecting discrepancies, the Verifier refines the data to meet the criteria. This separation of roles explains why the pipeline can produce reliable silver data even though zero-shot diagnosis remains challenging for individual models.

5.2 Data Quality Validation

To validate the reliability of this pipeline, we conducted an agreement study on a random sample of 100 instances against human ground truth. Our selected configuration (GPT-5 + Gemini-3-Flash) achieved a Cohen’s Kappa of 0.81 and an accuracy of 87%, matching the qualitative reliability tier of human experts ("Almost Perfect" agreement). As detailed in Table 12, this pipeline significantly out-

Task	Evidence	Diagnosis	Teaching
Spearman’s ρ	1.00	0.90	1.00
Kendall’s τ	1.00	0.80	1.00

Table 4: Rank-order consistency between Gold and Synthetic sets across tasks. High correlation indicates that model rankings are preserved across data types.

performs single-model baselines (e.g., GPT-5 alone achieved only $\kappa = 0.61$), demonstrating the necessity of the cross-verification.

5.3 Evaluation Consistency Validation

The utility of our automated pipeline lies in its ability to serve as a reliable and low-cost proxy for human-annotated benchmark. To validate this, we generated a synthetic set (LongTutor-Synth, 2,437 instances) and compared model rankings against the human-annotated gold standard (LongTutor-Gold). As shown in Table 4, the pipeline demonstrates exceptional fidelity, achieving perfect rank alignment (Spearman’s $\rho = 1.00$) for evidence acquisition and teaching action tasks, and strong consistency ($\rho = 0.90$) for state diagnosis. This confirms that the synthetic evaluations preserve the relative performance of models observed in human annotations. Detailed per-model scores are provided in Appendix K.

Finally, to mitigate concerns regarding domain specificity and enrich community resources, we applied the pipeline to a completely different data source, MoocRadar (Yu et al., 2023), covering university-level Mooc scenarios. This resulted in the release of 729 high-quality instances, validating the pipeline’s adaptability to varied educational contexts and providing a more diverse benchmark for future research.

6 Conclusion

We presented LongTutor, the first benchmark evaluating LLMs on long-term personalized tutoring via an Evidence–Diagnosis–Teaching framework. Unlike existing single-session evaluations, LongTutor assesses LLMs on utilizing learning histories for personalized tutoring. Our experiments reveal a significant gap: while LLMs excel at historical evidence acquisition, they struggle to translate evidence into accurate diagnoses and adaptive teaching, indicating that retrieval capability alone is insufficient. Additionally, we validated a generator–verifier pipeline to address data scarcity. We hope LongTutor fosters the development of AI tu-

tors that can harness the full potential of long-term learning history to provide truly personalized and pedagogically intelligent guidance.

Limitations

While LongTutor provides a robust framework for long-term personalized tutoring evaluation, several limitations remain. First, our current focus is primarily on mathematics, where errors follow clear logical patterns; the applicability of our framework to subjective disciplines like creative writing remains to be explored. Second, despite our reciprocal verification mechanism, the synthetic data generated by frontier models may still harbor subtle inductive biases or specific pedagogical preferences that differ from diverse human teaching styles. Finally, our benchmark evaluates static historical snapshots rather than dynamic, real-time interactions, and the long-term scope is currently constrained by model context windows. Future work will aim to address these issues by incorporating interactive simulators and more efficient memory architectures for extensive student histories.

Acknowledgments

This research was supported by grants from the National Natural Science Foundation of China (Grants No.U23A20319, 62477044), the Key Technologies R & D Program of Anhui Province (No. 202423k09020039), the Young Elite Scientists Sponsorship Program by CAST (No. 2024QNRC001), the Fundamental Research Funds for the Central Universities (No.WK2150110038), and the Open Research Fund of Zhejiang Key Laboratory of Intelligent Education Technology and Application (No.2025ZNJYKF006).

References

- Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom’s taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.
- Paul Black and Dylan Wiliam. 1998. Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, 5(1):7–74.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child,

- Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jingheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S. Yu, and Qingsong Wen. 2025. [LLM agents for education: Advances and applications](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13782–13810, Suzhou, China. Association for Computational Linguistics.
- 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*.
- Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278.
- Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, and 1 others. 2023. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773*.
- Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. [GLM: General language model pretraining with autoregressive blank infilling](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.
- Weibo Gao, Qi Liu, Linan Yue, Fangzhou Yao, Rui Lv, Zheng Zhang, Hao Wang, and Zhenya Huang. 2025. Agent4edu: Generating learner response data by generative agents for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23923–23932.
- Mary L Gick and Keith J Holyoak. 1980. Analogical problem solving. *Cognitive psychology*, 12(3):306–355.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the MATH dataset](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.
- 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.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024b. Socraticlm: Exploring socratic personalized teaching with large language models. *Advances in Neural Information Processing Systems*, 37:85693–85721.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024c. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Zitao Liu, Qiongqiong Liu, Teng Guo, Jiahao Chen, Shuyan Huang, Xiangyu Zhao, Jiliang Tang, Weiqi Luo, and Jian Weng. 2023b. Xes3g5m: A knowledge tracing benchmark dataset with auxiliary information. *Advances in Neural Information Processing Systems*, 36:32958–32970.
- Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. [MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

- 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, Suzhou, China. Association for Computational Linguistics.
- Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. [Evaluating very long-term conversational memory of LLM agents](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.
- Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. 2022. [LILA: A unified benchmark for mathematical reasoning](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5807–5832. Association for Computational Linguistics.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.
- Henry L Roediger III and Jeffrey D Karpicke. 2006. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, 17(3):249–255.
- Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research*, 78(1):153–189.
- Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. 2020. [CIMA: A large open access dialogue dataset for tutoring](#). In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64, Seattle, WA, USA → Online. Association for Computational Linguistics.
- Kurt VanLehn. 2006. The behavior of tutoring systems. *International journal of artificial intelligence in education*, 16(3):227–265.
- Lev S Vygotsky. 1978. *Mind in society: The development of higher psychological processes*, volume 86. Harvard university press.
- Fei Wang, Weibo Gao, Qi Liu, Jiatong Li, Guanhao Zhao, Zheng Zhang, Zhenya Huang, Mengxiao Zhu, Shijin Wang, Wei Tong, and 1 others. 2024. A survey of models for cognitive diagnosis: New developments and future directions. *arXiv preprint arXiv:2407.05458*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- David Wood, Jerome S Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2):89–100.
- Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. [Longmemeval: Benchmarking chat assistants on long-term interactive memory](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.
- Bin Xu, Yu Bai, Huashan Sun, Yiguan Lin, Siming Liu, Xinyue Liang, Yaolin Li, Yang Gao, and Heyan Huang. 2025. Edubench: A comprehensive benchmarking dataset for evaluating large language models in diverse educational scenarios. *arXiv preprint arXiv:2505.16160*.
- Jing Xu, Arthur Szlam, and Jason Weston. 2022. [Beyond goldfish memory: Long-term open-domain conversation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Jifan Yu, Mengying Lu, Qingyang Zhong, Zijun Yao, Shangqing Tu, Zhengshan Liao, Xiaoya Li, Manli Li, Lei Hou, Hai-Tao Zheng, and 1 others. 2023. Moocradar: A fine-grained and multi-aspect knowledge repository for improving cognitive student modeling in moocs. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2924–2934.
- Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024. [∞Bench: Extending long context evaluation beyond 100K tokens](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277, Bangkok, Thailand. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2025a. Siren’s song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.

Zheng Zhang, Ning Li, Qi Liu, Rui Li, Weibo Gao, Qingyang Mao, Zhenya Huang, Baosheng Yu, and Dacheng Tao. 2025b. The other side of the coin: Exploring fairness in retrieval-augmented generation. *arXiv preprint arXiv:2504.12323*.

Zheng Zhang, Qi Liu, Hao Jiang, Fei Wang, Yan Zhuang, Le Wu, Weibo Gao, and Enhong Chen. Fairlisa: Fair user modeling with limited sensitive attributes information. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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 Dataset Card & License

We explicitly state that all artifacts are released for research purposes. The dataset content, derived from XES3G5M (Liu et al., 2023b) and MocoRadar (Yu et al., 2023), is compatible with the CC-BY 4.0 license as it involves significant human annotation and refinement. Details are provided in Table 5.

B Potential Risks

First, although we adopt expert annotation and a generator–verifier pipeline, synthetic data may still reflect biases or pedagogical preferences of the underlying LLMs. Second, improper use of diagnostic labels without human oversight could lead to overconfident or misleading educational feedback. Third, future deployments should also monitor subgroup fairness and the trade-offs between fairness and overall utility, rather than focusing solely on average performance (Zhang et al., 2025b).

C Use of AI Assistants

We disclose the use of large language models and AI assistants in the preparation of this work as follows:

Writing and Polishing. AI assistants (e.g., Gemini 3 Flash, ChatGPT) were used to refine the linguistic quality and clarity of the manuscript. The authors provided original drafts of the text, and the AI models suggested improvements in grammar,

style. All AI-generated suggestions were manually reviewed, edited, and verified by the authors to ensure technical accuracy and to prevent the introduction of hallucinations.

Coding and Infrastructure. GitHub Copilot and similar AI-driven coding assistants were utilized during the development of the data processing pipeline and evaluation scripts. All code logic was rigorously tested by the authors through unit tests and manual debugging to ensure the integrity of the experimental results.

D State Taxonomy Details

To ensure the theoretical validity of our diagnostic framework, we align our knowledge state labels (Error Types) with the Knowledge Dimensions of the Revised Bloom’s Taxonomy (Anderson and Krathwohl, 2001). This alignment serves a dual purpose: it provides a rigorous standard for classifying student errors and directly guides the selection of the optimal teaching strategy.

Table 6 presents the comprehensive taxonomy, detailing the definition, symptoms, and targeted teaching strategy for each diagnostic category.

E Annotation Protocol Details

To ensure the high quality and consistency of the LongTutor benchmark, we implemented a structured three-step annotation pipeline. As mentioned in the main text, all steps were conducted under independent dual expert annotation followed by adjudication to ensure label reliability. The specific procedures for each task are as follows:

Step 1: Evidence Query Construction. Annotators designed probing queries to rigorously evaluate history retrieval and reasoning capabilities. These queries were constructed to cover three distinct categories: (1) specific record queries targeting factual details from a single interaction record; (2) cross-session reasoning queries requiring the aggregation of information or identification of patterns across multiple historical sessions; and (3) hallucination traps asking about non-existent records or absent details, which require the model to correctly abstain rather than fabricating information.

Step 2: Diagnostic State Labeling. Given the student’s current incorrectly answered problem and the full learning history \mathcal{H}_t , annotators inferred the underlying cognitive cause of the error. They assigned a diagnostic label based on our predefined

Property	Details
Dataset Name	LongTutor
Task Domain	Education, Personalized Tutoring, State Diagnosis, Teaching Action
Language	Simplified Chinese (zh-CN), English
Data Statistics	Total Size: 4,166 Student Interaction Sequences <ul style="list-style-type: none"> • XES3G5M Source: 3,437 sequences (Math) • MoocRadar Source: 729 sequences (MOOC/STEM)
Annotation Method	<ul style="list-style-type: none"> • Human-Annotated (Gold): 1,000 sequences from XES3G5M. • Model-Synthesized (Silver): 2,437 sequences from XES3G5M + 729 from MoocRadar (Generated by LLM with strict schema constraints).
Content Type	Student exercise history, Statistical features, Diagnostic labels (Recall Failure, Conceptual Gap, etc.), Teaching Strategies, and Natural Language Feedback.
License (Data)	CC-BY 4.0 (Creative Commons Attribution 4.0 International)
License (Code)	MIT License (for evaluation scripts)

Table 5: Overview and Datasheet for LongTutor. This table summarizes the dataset statistics, sources, annotation methods, and licensing terms.

Error Type (Our Label)	Bloom’s Knowledge Dimension (Anderson and Krathwohl, 2001)	Definition & Symptoms	Targeted Strategy
Recall Failure	<i>Factual Knowledge</i>	The student has previously mastered the specific knowledge but fails to retrieve basic elements or terminology due to memory decay.	Retrieval Practice
Conceptual Gap	<i>Conceptual Knowledge</i>	The student demonstrates a fundamental misunderstanding of the interrelationships among basic elements, theories, or classifications.	Conceptual Explanation
Procedural Error	<i>Procedural Knowledge</i>	The student understands the underlying concept but commits errors in the execution of algorithms, specific techniques, or calculation steps.	Stepwise Scaffolding
Transfer Deficit	<i>Metacognitive Knowledge</i>	The student possesses the knowledge in isolation but lacks the conditional knowledge to apply it in novel contexts or variations.	Analogical Transfer

Table 6: The diagnostic taxonomy of LongTutor. We map our educational error types to the Knowledge Dimensions of the Revised Bloom’s Taxonomy. This mapping guides the selection of the optimal teaching strategy. In particular, we associate transfer deficits with metacognitive knowledge, as they reflect a lack of conditional knowledge regarding when and how to apply known concepts across contexts.

taxonomy (see Table 6), ensuring the diagnosis reflects the student’s knowledge state (e.g., conceptual gap versus recall failure) rather than just the surface error.

Step 3: Teaching Response Drafting. Annotators authored a reference golden response conditioned on the assigned diagnosis. To serve as a valid ground truth for personalized tutoring, each response was required to explicitly leverage the student’s learning history (e.g., referencing a similar past mistake or a mastered concept), adhere strictly to the prescribed teaching strategy associated with the diagnostic label, and maintain coherence and appropriateness with respect to the student’s proficiency level.

E.1 Instructions Given to Annotators

To ensure annotation quality, we provided a comprehensive instruction manual to all participants. The guidelines included: (1) **Task Definitions**, which clearly outlined the three-step pipeline (Evidence Query, Diagnostic Labeling, and Teaching Response Drafting); (2) **Example Gallery**, showcasing high-quality reference annotations versus common pitfalls; and (3) **Safety and Privacy Protocols**, instructing annotators to flag any inappropriate content and avoid introducing Personally Identifiable Information (PII) during drafting. A dedicated training session was conducted to clarify the nuances of our diagnostic taxonomy and the "scaffolding" nature of the required pedagogical responses.

E.2 Annotator Recruitment and Payment

We recruited three expert annotators holding advanced degrees (Master’s or Ph.D.) in Education, ensuring they possess both profound theoretical knowledge and practical teaching experience. Participants were compensated at a rate of 50 CNY per hour. This rate significantly exceeds the local minimum wage.

E.3 Data Consent

All annotators provided explicit written consent for their contributions to be used for research purposes and the public release of the LongTutor benchmark. We informed all participants about the project’s goals, the intended use of the data, and their right to withdraw from the study at any time. All data used for annotation were sourced from public datasets.

E.4 Ethical Considerations.

All participation was voluntary, and annotators were informed that they could withdraw from the project at any stage without penalty. No personal sensitive data was collected from the annotators during the recruitment or compensation process, and the tasks involved no exposure to harmful or offensive content.

F Data Examples

We present category-wise qualitative examples from the LongTutor test set to illustrate the annotation of the three hierarchical tasks. Each table corresponds to one diagnostic category defined in Table 6. For clarity and readability, the input context shown in each table is a truncated excerpt of the full learning history, containing only interactions relevant to the corresponding evaluation.

Recall Failure. Table 15 shows an illustrative example diagnosed as Recall Failure, highlighting cases where students fail to retrieve previously mastered factual knowledge. The example demonstrates how historical interactions are used for evidence acquisition queries, diagnosis, and history-aware teaching.

Conceptual Gap. Table 16 presents an example labeled as Conceptual Gap, where errors stem from misunderstandings of conceptual relationships. The case illustrates how diagnostic reasoning and targeted conceptual explanations are grounded in the student’s learning history.

Procedural Error. Table 17 provides an example of Procedural Error, in which students understand the underlying concepts but fail during procedural execution. The example highlights stepwise, history-aware scaffolding strategies.

Transfer Deficit. Table 18 shows an example diagnosed as Transfer Deficit, illustrating failures to apply known knowledge in novel contexts. The case demonstrates the use of analogical transfer grounded in prior learning experiences.

G Metric Calculation Details

Semantic Accuracy (SA) Given a model-generated answer \hat{a} and the ground-truth evidence a , SA is defined as:

$$\text{SA}(\hat{a}, a) = \begin{cases} 1, & \text{if } \text{EM}(\hat{a}, a) = 1 \\ & \vee G(\hat{a}, a) = \text{True} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where G is an LLM-based evaluator (prompted zero-shot) that determines if \hat{a} preserves the core semantic components of a .

Macro-F1 Score For the cognitive state classification, Macro-F1 is calculated as the arithmetic mean of per-class F1 scores:

$$\text{Macro-F1} = \frac{1}{|\mathcal{Y}|} \sum_{i \in \mathcal{Y}} \frac{2 \cdot P_i \cdot R_i}{P_i + R_i} \quad (2)$$

where P_i and R_i denote the precision and recall for category $i \in \mathcal{Y}$.

ROUGE-L Since the text is in Simplified Chinese, we utilized jieba for word-level tokenization before calculating these metrics to ensure accurate overlap matching. We used the standard rouge-score Python package for calculation. We treat ROUGE-L as a supplementary lexical-overlap reference rather than a primary pedagogical metric.

LLM-as-a-Judge Scoring The holistic score S for the Teaching Action task is derived from four pedagogical dimensions \mathcal{D} (History Utilization, Strategy Alignment, Appropriateness, Coherence):

$$S = \frac{1}{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} s_j, \quad s_j \in [1, 5] \quad (3)$$

where s_j represents the score assigned by the judge for dimension j . In the main comparison table, we report both the four dimension scores and their

mean. Appendix I.1 further validates this rubric against human expert scoring and dynamic tutoring outcomes.

H Experimental Setup Details

Models. We utilized a combination of local deployment and API access to evaluate the wide range of models: (1) Open-Source Models: Sourced from Hugging Face⁴. Models fitting within local memory were deployed using the vLLM library (Kwon et al., 2023). (2) Large-Scale Models: Models exceeding local memory limits (e.g., GLM-4.6) were accessed via third-party cloud APIs, including OpenRouter⁵ and SiliconFlow⁶. (3) Proprietary Models: Evaluated directly via their official provider APIs.

Hardware Configuration. Local experiments were conducted on 2×NVIDIA A100 (40GB) GPUs, and a 16-core Intel Xeon Gold 6426Y CPU.

Inference Hyperparameters. To ensure reproducibility and deterministic outputs, we applied a uniform decoding strategy across all models: (1) Temperature: Set to $T = 0.0$ to minimize randomness (greedy decoding). (2) Generation Limit: Maximum generation length set to 3,000 tokens to accommodate long teaching responses. (3) Judge Configuration: For the LLM-as-a-Judge evaluation in Task 3, we utilized Gemini-3-Flash with $T = 0.0$ to ensure consistent scoring standards. Evaluation is performed on each sample individually, and the final results are reported as the average across the entire dataset.

Input & Prompting Strategy. To simulate a realistic long-term tutoring scenario, we adhere to a standardized input protocol for all tasks: (1) Context Window (\mathcal{H}_t): We construct a long-term context consisting of the student’s most recent $L = 100$ practice records. Each historical record explicitly includes the question text, associated knowledge points, interaction timestamp, and the correctness label (Correct/Incorrect). (2) Target Problem: The sequence concludes with the current incorrectly answered problem, which serves as the target for diagnosis and teaching. (3) Prompt Format: We employ a unified zero-shot prompt template to wrap this context, instructing the model

to act as an educational expert. The exact prompt structure is illustrated in Figure 7.

I Additional Results

Table 7 reports the evaluation results of additional models that are not included in the main comparison table due to space constraints. These models primarily consist of alternative open-source baselines. All models are evaluated under the same experimental protocol, inference configuration, and evaluation metrics as described in the main paper.

Beyond the core findings presented in Section 4.2, we provide further granular insights into the model behaviors observed during the evaluation.

Open-Source: Strong Pedagogy, Limited Context. Chinese open-source models (e.g., DeepSeek-V3.2, GLM-4.6) show impressive teaching quality, with Coherence and Appropriateness scores competitive with closed-source leaders. However, a significant gap exists in long-context evidence acquisition (Task 1), where they trail Gemini-2.5-Pro (Avg 68.9 vs. 82.3). This indicates that while their instructional generation is high-quality, maintaining precision over long-term history remains a challenge.

Instruction Following Collapse in English-Centric Models. We observe a catastrophic breakdown in instruction following for Western-centric models (e.g., Llama-3.1-8B) when processing long-context Chinese educational histories. Unlike native Chinese models, they struggle to adhere to complex formatting constraints (e.g., specific JSON schemas) under this linguistic distribution shift, leading to near-random performance (e.g., Llama-3.1’s extremely low F1 of 8.0 and Teaching Avg of 1.1).

The Accuracy-F1 Discrepancy in Diagnosis. A significant divergence between Accuracy and Macro-F1 is observed in weaker models. For instance, EduChat-7B achieves a relatively high Accuracy of 49.4% but a low F1 of 13.8%. This discrepancy reveals that these models tend to predict "safe" majority-class labels rather than accurately identifying diverse knowledge states. In contrast, stronger models like Gemini-2.5-Pro maintain a more balanced performance (42.0% Acc vs. 39.7% F1), indicating a more genuine capability.

⁴<https://huggingface.co/>

⁵<https://openrouter.ai/>

⁶<https://cloud.siliconflow.cn/>

Model	Task 1: Evidence				Task 2: Diagnosis			Task 3: Teaching				
	IE	MR	HC	Avg.	Acc	F1	R-L	Hist.	Strat.	Coher.	Appr.	Avg.
<i>Additional Open-Source Models</i>												
Qwen3-14B	75.6	15.2	68.4	53.1	23.2	17.7	30.1	1.9	2.0	3.9	3.7	2.9
Qwen3-32B	81.8	21.2	68.4	57.1	36.2	23.6	34.4	1.6	2.2	4.5	4.2	3.1
GLM-4-9B-Chat	68.8	8.2	39.6	38.9	45.4	17.8	7.6	1.0	1.3	2.3	2.1	1.7
DeepSeek-V3	92.4	32.6	77.0	67.3	23.8	17.4	36.9	1.2	2.1	4.6	4.1	3.0
InternLM3-8B-Instruct	22.6	4.2	9.0	11.9	27.6	13.5	13.8	1.2	1.2	2.1	2.1	1.7
Minstral-3-8B-Instruct	58.4	7.4	5.8	23.9	34.8	18.1	5.8	1.1	1.1	1.7	1.7	1.4

Table 7: Additional evaluation results for other open-source models. The formatting and metrics follow the conventions of Table 2. **Avg.** for Teaching is calculated as the mean of the four judge scores (History, Strategy, Coherence, Appropriateness).

Model	Static Score	TSR
Gemini-2.5-Pro	4.2	52.0%
GPT-5	3.1	45.5%
DeepSeek-V3.2	3.8	41.0%
Llama-4-Scout	2.2	36.2%
EduChat-7B	1.7	18.5%

Table 8: Correlation between the static Teaching Score and dynamic Teaching Success Rate (TSR). The model ranking yields Spearman’s $\rho = 0.90$.

I.1 Evaluation Validity and Task Dependency

We report four supplementary analyses added in the camera-ready revision to address evaluation validity and pipeline rationality concerns.

Dynamic Simulation. To complement static evaluation, we build an agent-based tutoring setup. A fixed GPT-5 student simulator is injected with the gold diagnosis label and interacts with the tutor model for up to three turns. We measure Teaching Success Rate (TSR), defined as the percentage of sessions in which the student solves the target problem within three turns. Table 8 shows a strong rank correlation between TSR and the static teaching score.

Human-LLM Agreement. We additionally compare the LLM judge against two expert annotators on a stratified sample of 100 teaching responses, using the same rubric for both sides. Table 9 shows that agreement is strongest on the pedagogically central dimensions of history grounding and strategy alignment.

Task Dependency. To verify that the three tasks form a functional pipeline rather than isolated probes, we compare teaching quality under ideal versus predicted diagnosis inputs. Table 10 shows that misdiagnosis leads to clearly worse teaching,

Dimension	Spearman’s ρ	Within ± 1 pt
History Utilization	0.91	94%
Strategy Alignment	0.86	90%
Coherence	0.78	87%
Appropriateness	0.75	83%
Overall Score	0.87	92%

Table 9: Human-LLM agreement on teaching evaluation. Strong alignment supports rubric-based LLM judging for this constrained pedagogical verification task.

Diagnostic Input	Teaching Score	Observation
Gold Diagnosis	4.6	Targeted and pedagogically aligned scaffolding.
Predicted Diagnosis	3.9	Misdiagnosis leads to irrelevant instructional moves.

Table 10: Impact of diagnosis accuracy on teaching quality. Accurate upstream diagnosis is a prerequisite for high-quality downstream teaching.

while the model-level Diagnosis F1 and Teaching Score exhibit a strong Spearman correlation of $\rho = 0.88$.

RAG Baseline. Finally, we compare full-history conditioning with a strong semantic retrieval baseline that retrieves the Top-20 most similar historical interactions before generation. Table 11 shows that RAG underperforms full-history conditioning on all three tasks, suggesting that longitudinal tutoring requires temporal continuity in addition to semantic relevance.

I.2 Fine-grained Diagnosis Analysis

To investigate the discrepancy between Accuracy and F1 scores (e.g., GLM-4.6: 36.2% Acc vs. 27.5 F1), we analyze error distributions in Figure 5.

Context Strategy	Evidence (T1)	Diagnosis (T2)	Teaching (T3)
Long Context	79.3%	28.1	3.9
RAG (Top-20)	74.2%	21.4	3.3

Table 11: RAG versus full-history conditioning. Long-Tutor benefits from preserving chronological continuity rather than retrieving only semantically similar snippets.

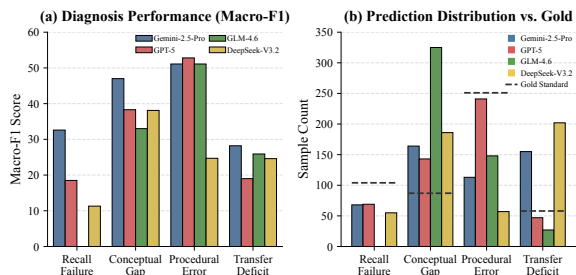


Figure 5: Fine-grained diagnosis analysis: (a) Macro-F1 across knowledge states, highlighting GLM-4.6’s collapse on *Recall Failure*; (b) Prediction counts vs. Gold Standard (dashed). Models like DeepSeek-V3 and GLM-4.6 exhibit significant over-prediction bias in specific categories.

Three failure patterns emerge:

Category Collapse. GLM-4.6 exhibits a total collapse on Recall Failures, achieving 0.0 F1 (Figure 5a). Analysis reveals zero predictions against 104 gold samples, effectively treating the 4-way task as 3-way. This suggests a failure to align linguistic cues of memory retrieval errors with the appropriate pedagogical labels.

Diagnostic Hallucination. Open-source models show severe calibration issues (Figure 5b). DeepSeek-V3 aggressively over-diagnoses Transfer Deficits (202 predicted vs. 58 gold), conflating general confusion with cross-context failures. Similarly, GLM-4.6 defaults to Conceptual Gaps (325 predicted vs. 87 gold), over-attributing errors to fundamental understanding gaps rather than procedural slips.

Majority Class Bias. Several strong proprietary models still favor a "safe bet" strategy, peaking at 52.8% F1 on the majority class (Procedural Error, $n = 251$) but failing on minority states like Recall Failure (18.5% F1). In contrast, Gemini-2.5-Pro maintains a balanced distribution closest to the Gold Standard, explaining its superior Macro-F1 (39.7).

I.3 Qualitative Case Study

To demonstrate the concrete impact of model capability on utilizing long-term student history, we present a detailed comparison in Table 14.

Scenario Overview. The student failed to solve a new path-counting problem (Q8285) that introduced a specific constraint ("move from smaller to larger numbers"). Crucially, the input history included successful attempts at "Shortest Path Labeling" problems (e.g., Q779–Q784), indicating the student possesses the underlying skill but failed to adapt it.

Analysis of Model Behaviors. The performance of different models illustrates a clear divide in how historical data is utilized to guide pedagogical decisions. Model A (GPT-5-mini), despite having access to the student history, failed to leverage specific prior knowledge and defaulted to a generic "Procedural Error" diagnosis. This led to a high-cognitive-load teaching involving technical concepts like "Directed Graphs" and "Dynamic Programming," which likely exceeds the student’s zone of proximal development and ignores their actual learning trajectory. In contrast, Model B (GPT-5) accurately identified a "Transfer Deficit" by recognizing that the student possesses the underlying skill ("Labeling Method") but failed to adapt it to the new context. Consequently, Model B generated an Analogical Transfer cue that explicitly linked the current problem to the student’s past successes. This approach effectively facilitates knowledge transfer and reduces cognitive load, showcasing a more sophisticated level of true personalization.

J Sensitive Study Settings

We define the baseline configuration consistent with the standard protocol in Appendix H, utilizing the full interaction history (\mathcal{H}_t) with $L = 100$ records. Against this baseline, we introduced three variations to isolate specific factors:

Context Selection Strategies. We first evaluated a short history ($L = 20$) approach, where the input is truncated to retain only the most recent 20 interaction records to test the model’s dependency on long-term versus short-term memory. Alternatively, we implemented Filtered History, which replaces the rigid chronological window with a dynamic retrieval of up to 100 historical records sharing the

same knowledge points (KPs) as the target problem, thereby prioritizing semantic relevance over temporal proximity.

Feature Augmentation. We investigated the impact of statistical features by explicitly prepending a quantitative summary block to the textual input. This block incorporates metrics derived from the full history, such as the student’s overall accuracy, error rates on related knowledge points, and recent performance trends.

Instruction Strategy. We enhance the system instruction with detailed, step-by-step teaching guidelines. For instance, "If the error is a Transfer Deficit, you must provide a real-world analogy..."

K Pipeline Validation Details

Data Quality. Detailed breakdown of our pipeline agreement study highlights the critical role of the Generator–Verifier architecture in achieving expert-level reliability. We compared two pipeline configurations against single-model baselines using a random sample of 100 instances, with results summarized in Table 12. The pairing of GPT-5 as the generator and Gemini-3-Flash as the verifier achieved the highest performance ($\kappa = 0.81$, "Almost Perfect"), whereas removing the verifier caused a significant performance decline across all models. For instance, GPT-5 alone dropped to $\kappa = 0.61$, sliding into the "Substantial" tier. This ablation study confirms that cross-verification is necessary to reach the "Almost Perfect" threshold, significantly outperforming single-model labeling.

Evaluation Consistency. Table 13 provides the performance scores for various models across the Evidence Acquisition, State Diagnosis, and Teaching Action sub-tasks, comparing results on both Human-Annotated (Gold) and Synthetic (Synth) datasets. Gemini-2.5-Pro leads the evaluation with the highest scores in all categories, including a Evidence Acquisition accuracy of 82.3% and a Diagnosis F1 of 39.7% on the gold set. Notably, the performance rankings of the models are perfectly preserved between the Gold and Synthetic datasets for both the Evidence Acquisition and Teaching tasks. This consistency demonstrates the high quality of the synthetic data produced by our pipeline and its reliability as a proxy for human-standard evaluation.

Configuration (Generator / Verifier)	Kappa (κ)	Acc (%)	Interpretation
Human vs. Human (Experts)	0.90	93	Almost Perfect
GPT-5 / Gemini-3-Flash	0.81	87	Almost Perfect
Gemini-3-Flash / GPT-5	0.70	79	Substantial
Gemini-3-Flash (w/o Verifier)	0.65	75	Substantial
GPT-5 (w/o Verifier)	0.61	74	Substantial

Table 12: Detailed Agreement study for the Diagnostic Labeling task ($N = 100$). The Generator / Verifier pipeline significantly outperforms single-model baselines.

Model	Evidence (Acc)		Diagnosis (F1)		Teaching (Score)	
	Gold	Synth	Gold	Synth	Gold	Synth
Gemini-2.5-Pro	82.3	78.9	39.7	37.0	4.18	4.18
Gemini-2.5-Flash	79.3	78.0	28.1	26.3	3.88	3.85
Qwen3-32B	57.1	54.7	23.6	21.9	3.13	3.10
Qwen3-14B	53.1	53.3	17.7	16.3	2.88	2.80
Qwen3-8B	43.5	47.6	16.0	17.8	2.40	2.40

Table 13: Performance consistency between Human-Annotated (Gold) and Synthetic (Synth) datasets. The rankings are perfectly preserved for Evidence and Teaching tasks.

L Experimental Prompts

We provide the detailed prompt templates utilized in our experiments.

Figure 6 details the prompts for our LLM-as-a-Judge framework. These include the Semantic Accuracy evaluator for evidence retrieval and the LLM-as-a-Judge for teaching actions. The latter uses a structured rubric to score responses based on history utilization, strategy alignment, coherence, and appropriateness.

Figure 7 presents the inference prompt used to evaluate the AI Tutors (Models-Under-Test). This prompt features a modular design to facilitate ablation studies (e.g., toggling diagnostic rules or statistical features). Notably, it incorporates a "Constraints Reminder" block at the end of the user input to mitigate instruction forgetting caused by long interaction histories.

Figure 8 illustrates the data annotation prompt used by the automated pipeline to generate annotated data. It enforces strict logical consistency between the student’s history, the state diagnosis, and the teaching action.

Category	Content (Chinese / English Translation)
<i>Student Context (Input for Both Models)</i>	
History Log	Events: Q779–Q784 (Multiple Entries) Result: Correct Skill: Labeling Method / Shortest Path Counting (标数法/最短路径) <i>Evidence:</i> Student consistently solved path counting problems using the addition principle correctly.
Current Problem	ID: Q8285 Result: Error Task: Count paths from Room 1 to 10. Constraint: Must move from a smaller number to an adjacent larger number .
<i>Model A: GPT-5-mini</i>	
Diagnosis	Procedural Error (Failed to link specific history)
Reasoning	Student mastered concepts but made errors in enumeration steps. (Ignores the specific method used in history).
Teaching	[Stepwise Scaffolding / High Cognitive Load] 先复述题意... 步骤一：把图看成有向图，列出邻接表。步骤二：从终点10开始向回推，用动态规划：设 $f(n)$ 为从房间 n 到10的走法数... (Step 1: Treat the map as a Directed Graph and list the Adjacency Table . Step 2: Use Dynamic Programming backwards from node 10: Let $f(n)$ be the number of ways from room n to 10...)
<i>Model B: GPT-5</i>	
Diagnosis	Transfer Deficit (Inferred from History Q779-Q784)
Reasoning	The current problem is similar to previous "Labeling Method" problems (Q779-Q784), but introduces a direction constraint. The student failed to <i>transfer</i> the existing method to the new constraint.
Teaching	[Analogical Transfer / Personalized] 先把这题转化为和你做对的“最短路径标数法”类似... 规则：只能从相邻且编号更小的房间走到更大的房间。步骤：把1号房间标1。按编号从小到大，每个房间的走法数=所有与它相邻且编号更小的房间的数之和。... 你先试着标到5号？ (Let's transform this into the " Shortest Path Labeling Method " you solved correctly before... Rule: Only move from smaller to larger. Step: Label room 1 as "1". For each room, the number of ways = sum of numbers in adjacent smaller rooms... Try labeling up to room 5 first?)

Table 14: Case study comparing diagnosis and teaching quality. Although both models received the history log, GPT-5-mini failed to leverage the student’s prior mastery of the “Labeling Method,” resulting in a complex, technical explanation. GPT-5 correctly diagnosed a “Transfer Deficit” and guided the student to adapt their existing knowledge to the new problem constraint.

PART I: Input Context (History & Current Question)		
History	Item 1359	Time: 2021-05-05 12:57:25 Result: Correct KC: 菱形对角线(Rhombus Diagonals) Text: 已知菱形的对角线AC和BD, 求其长度。 Translation: Given the diagonals AC and BD of a rhombus, determine their lengths.
	Daily Summary	Date: 2021-05-05 Correct Count: 14
Current	Meta Question	Time: – Result: Wrong KC: 平行四边形面积(Parallelogram Area) Text: 计算下列平行四边形的面积。
PART II: Output Tasks (Ground Truth)		
Task	Type	Annotation (Chinese / English)
1. Evidence	Information Extraction	Q: 在2021年5月5日12:57:25, 学生回答的问题1359中, 菱形的对角线AC和BD的长度分别是多少? Q: What were the lengths of diagonals AC and BD in question [1359] answered on 2021-05-05 12:57:25? A: AC=18, BD=6。 A: AC = 18, BD = 6.
	Multi-session Reasoning	Q: 学生在2021年5月5日当天总共正确回答了多少道题目? Q: How many questions did the student answer correctly on 2021-05-05? A: 14道。 A: 14 questions.
	Hallucination Check	Q: 学生在什么时候正确回答过关于“圆形面积计算”的问题? Q: When did the student correctly answer a question about “Circle Area Calculation”? A: Unknown A: Unknown.
2. Diagnosis	Label	Recall Failure (遗忘导致的提取失败/ Recall Failure)
	Reason	学生在相关几何概念上的历史正确率极高 (related_concept_acc=1.0), 表明其已掌握该知识。然而, 距离上次正确回答同类问题已过去16天 (>4天), 且当前题目整体错误率较低 (error_rate=0.15) 却仍然作答错误, 符合遗忘曲线下的知识提取困难特征。 Reason: Despite perfect historical accuracy on related concepts (related_concept_acc=1.0), the student failed after a 16-day gap since the last correct attempt (>4 days). Given the low difficulty of the current problem (error_rate=0.15), this error is consistent with recall failure caused by forgetting.
3. Teaching	Strategy	Retrieval Practice (提取练习)
	Response	小同学, 我发现你之前在面积计算上可是“百分百正确”的小专家呢! 你在5月5号那天一口气做对了十几道不同的面积题。这次出错, 更像是因为隔了一段时间没练习, 公式一时没被成功想起来。先来帮它“找回来”吧: 1) 你还记得平行四边形的面积公式吗? 它需要用到哪两个量? 2) 看看图中带有直角标记的线段, 它分别对应的是“底”还是“高”? 试着先写出公式, 再代入数值计算一次。 Response: You previously showed perfect performance on area calculations, but after a long gap, the formula may not be readily retrievable. Let us recall it step by step by first writing down the formula and identifying the base and height indicated in the figure.

Table 15: An illustrative example of the *Recall Failure* category in the LongTutor benchmark. The student exhibits retrieval difficulty despite strong prior mastery, highlighting the impact of forgetting over time.

PART I: Input Context (History & Current Question)		
(... Previous interactions omitted ...)		
History	Item 151	Time: 2021-05-22 10:07:35 ID: 152 Result: Wrong KC: 平行四边形(Parallelogram) Text: 下图中有个平行四边形, 请判断... Translation: There is a parallelogram in the figure below, please determine...
	Item 152	Time: 2021-05-22 10:42:10 ID: 1366 Result: Correct KC: 平行四边形面积(Parallelogram Area) Text: 一个平行四边形底是20, 高是15, 求面积。 Translation: A parallelogram has a base of 20 and height of 15. Calculate the area.
Current	Meta Question	Time: 2021-05-29 ID: 450 Result: Wrong KC: 平行四边形面积(Parallelogram Area) Text: 计算下面图形的面积。 Translation: Calculate the area of the figure below.
PART II: Output Tasks (Ground Truth)		
Task	Type	Annotation (Chinese / English)
1. Evidence	Information Extraction	Q: 学生在2021-05-22 10:07:35回答的题目[152]的具体内容是什么? Q: What was the specific content of question [152] answered on 2021-05-22 10:07:35? A: 下图中有个平行四边形。 A: There is a parallelogram in the figure below.
	Multi-session Reasoning	Q: 学生在2021-05-22这一天, 一共尝试了几道关于“平行四边形”相关知识的题目? Q: How many questions related to “Parallelogram” did the student attempt on 2021-05-22? A: 2道 A: 2 questions.
	Hallucination Check	Q: 学生在2021-05-24回答的有关“三角形面积计算”的题目编号是多少? Q: What is the ID of the question regarding “Triangle Area Calculation” answered on 2021-05-24? A: Unknown A: Unknown.
2. Diagnosis	Label	Conceptual Gap (概念性缺失/ Conceptual Gap)
	Reason	学生在平行四边形相关概念上的历史正确率仅为0.29。在题目[152]和当前题目[450]中均未能正确识别“底”和“对应的高”, 尽管其曾在[1366]中成功应用面积公式, 说明其对核心概念关系的理解仍不牢固。 Reason: The student shows low accuracy (0.29) on related conceptual knowledge. Although they previously applied the area formula correctly in [1366], repeated failures in identifying the base and corresponding height in [152] and the current problem indicate a fundamental conceptual misunderstanding.
3. Teaching	Strategy	Conceptual Explanation (概念性讲解)
	Response	小同学, 我们又见面啦! 这道题其实藏着一个秘密: 平行四边形的面积就像是把长方形“推歪了”得到的, 所以它的面积公式还是“底×高”。不过要注意, 这里的“高”必须和底边垂直。回想一下你在[1366]中是如何找到那条垂直线段的, 这次也试着先在图中找出互相垂直的两条线吧! Response: This problem hides a small secret: a parallelogram can be viewed as a slanted rectangle, so its area is still base × height. However, the height must be perpendicular to the base. Recall how you identified this perpendicular segment in [1366], and try to locate it in the current figure.

Table 16: An illustrative example of the *Conceptual Gap* category in the LongTutor benchmark. **Part I** shows the historical context and the current incorrectly answered problem. **Part II** presents expert annotations for Evidence acquisition, State Diagnosis, and Teaching Action.

PART I: Input Context (History & Current Question)		
History	Item 5021	Time: 2021-05-21 04:54:10 Result: Correct KC: 长方形性质(Rectangle Properties) Text: 判断题: 长方形是特殊的平行四边形。() Translation: True or False: A rectangle is a special type of parallelogram.
	Daily Summary	Date: 2020-10-31 KC: 面积/周长 (长方形、正方形) Correct Count: 4
Current	Meta Question	Item: 2271 Result: Wrong KC: 周长变化 (图形切割) Text: 比较图形切割前后周长的变化。
PART II: Output Tasks (Ground Truth)		
Task	Type	Annotation (Chinese / English)
1. Evidence	Information Extraction	Q: 学生在2021年5月21日04:54:10回答正确的那道判断题的具体内容是什么? Q: What was the content of the judgment question answered correctly on 2021-05-21 04:54:10? A: 判断题: 长方形是特殊的平行四边形。() A: True/False: A rectangle is a special type of parallelogram.
	Multi-session Reasoning	Q: 在2020年10月31日这一天, 学生一共正确回答了几道与长方形或正方形面积/周长相关的题目? Q: How many area/perimeter questions related to rectangles or squares did the student answer correctly on 2020-10-31? A: 4道。 A: 4 questions.
	Hallucination Check	Q: 学生在2021年1月22日回答关于“正方形面积计算”的填空题时填写的具体数值是多少? Q: What numerical value did the student fill in for a square area calculation on 2021-01-22? A: Unknown A: Unknown.
2. Diagnosis	Label	Procedural Error (过程性错误/ Procedural Error)
	Reason	学生在相关知识点上的历史正确率为1.0, 且在同一天刚刚正确完成了长方形性质判断题, 说明其对概念本身并不存在缺失。当前题目涉及图形切割后的周长变化比较, 属于多步骤推理问题, 且整体错误率较高(0.62)。学生的失误更可能源于未能系统追踪“切割后新增边”的计算步骤, 而非对周长概念的误解。 Reason: Given perfect historical accuracy and a recent correct judgment on rectangle properties, conceptual misunderstanding is unlikely. The current multi-step perimeter comparison task (error rate 0.62) suggests a failure in tracking newly introduced edges after cutting, indicating a procedural rather than conceptual error.
3. Teaching	Strategy	Stepwise Scaffolding (分步支架式引导)
	Response	小同学, 这道题我们需要一步一步“跟着图形走”。你之前在处理长方形周长和图形变化时表现得很稳定。我们慢慢拆解: 1) 沿中点剪开后, 剪出的线段长度等于原来的宽(8厘米)。 2) 想一想, 这条“剪痕”在两个新图形中分别变成了什么? 3) 当新的边出现时, 周长是如何发生变化的? 你先不用急着算总周长, 先告诉老师: 一共新增了几条8厘米的边。 Response: Let us follow the geometric transformation step by step. Focus first on identifying the new edges introduced by the cut before computing the final perimeter.

Table 17: An illustrative example of the *Procedural Error* category in LongTutor. The student demonstrates intact conceptual understanding but fails to correctly execute the multi-step reasoning required after a geometric transformation.

PART I: Input Context (History & Current Question)		
History	Item 1832	<p>Result: Correct</p> <p>KC: 奇数与偶数的应用(Parity Reasoning)</p> <p>Text: 找出两个整数，使得它们的和为264，差为57，这样的两个数是否存在？</p> <p><i>Translation: Find two integers whose sum is 264 and difference is 57. Do such numbers exist?</i></p>
	Daily Summary	<p>Date: 2021-02-27</p> <p>KC: 奇数与偶数的应用</p> <p>Correct Count: 4</p>
Current	Meta Question	<p>Item: 1835 Result: Wrong</p> <p>KC: 奇偶性不变量（翻动问题）</p> <p>Text: 根据翻动规则判断给定情形是否可以实现。</p>
PART II: Output Tasks (Ground Truth)		
Task	Type	Annotation (Chinese / English)
1. Evidence	Information Extraction	<p>Q: 学生在记录[1832]中回答的问题内容是什么？</p> <p><i>Q: What was the content of the question answered in record [1832]?</i></p> <p>A: 找出两个整数，使得它们的和为264，差为57，这样的两个数是否存在？</p> <p><i>A: Find two integers whose sum is 264 and difference is 57.</i></p>
	Multi-session Reasoning	<p>Q: 在2021-02-27当天的历史记录中，学生总共正确回答了多少道知识点为“奇数与偶数的应用”的题目？</p> <p><i>Q: How many questions related to parity reasoning did the student answer correctly on 2021-02-27?</i></p> <p>A: 4道。</p> <p><i>A: 4 questions.</i></p>
	Hallucination Check	<p>Q: 学生在回答[1838]这道填空题时，具体的思考耗时是多少秒？</p> <p><i>Q: How many seconds did the student spend thinking when answering item [1838]?</i></p> <p>A: Unknown</p> <p><i>A: Unknown.</i></p>
2. Diagnosis	Label	Transfer Deficit (迁移困难/ Transfer Deficit)
	Reason	<p>学生在“奇数与偶数的应用”相关题目上的历史正确率达到100%，并且不久前刚刚正确解决了同类的翻动或开关问题。然而，当前题目仅在规则形式上做出变化（翻动次数从固定变为递增），其核心逻辑仍然依赖奇偶性不变量。学生未能将先前成功的问题建模方式迁移到该变式情境中，体现出典型的跨情境迁移困难。</p> <p><i>Reason: Despite perfect historical accuracy and recent success on analogous flipping problems, the student fails to transfer the invariant-based parity reasoning to a structurally modified scenario, indicating a clear transfer deficit.</i></p>
3. Teaching	Strategy	Analogical Transfer (类比迁移引导)
	Response	<p>小同学，这道题其实和你之前做过的“翻杯子”“开关灯”问题是同一类哦！我们来做个对照：</p> <p>1) 每个状态从初始到目标，必须被翻动奇数次；</p> <p>2) 多个对象同时满足条件时，总翻动次数的奇偶性是关键不变量；</p> <p>3) 比较题目给出的总翻动次数与理论要求的奇偶性。</p> <p>试着用这个思路，再判断一次当前情形是否可行。</p> <p><i>Response: Try mapping this problem back to previously solved flipping or switching tasks by identifying the invariant parity condition before evaluating feasibility.</i></p>

Table 18: An illustrative example of the *Transfer Deficit* category in LongTutor, where the student possesses intact domain knowledge but fails to transfer an invariant-based reasoning strategy to a structurally modified problem.

Evaluation Prompts for Automated Benchmarking

– METRIC 1: LLM-BASED SEMANTIC ACCURACY (*G*) –

Evaluate whether the MODEL response is factually consistent with the GOLD reference for the QUERY.

CRITERIA FOR A MATCH

- **Semantic Equivalence:** The MODEL answer must convey the same core meaning as the GOLD reference.
- **Correctness:** The MODEL answer must not contain contradictions or factual errors relative to the GOLD.
- **Style Ignored:** Ignore differences in length, phrasing, or punctuation.

OUTPUT FORMAT

Output strictly valid JSON with no markdown formatting:
{*"match"*: true} if it matches, or {*"match"*: false} otherwise.

TASK: {task} **QUERY:** {query}
GOLD_ANSWER: {gold_answer}
MODEL_ANSWER: {pred_answer}

– METRIC 2: LLM-AS-A-JUDGE (TEACHING ACTION) –

ROLE: Strict Educational Teaching Evaluator.

TASK: Compare the MODEL response to the GOLD reference (representing one possible high-quality answer). Rate on a 1–5 scale (1: Very Poor, 5: Excellent).

SCORING RUBRIC

1. **History Utilization** (s_1): Does the model use student's learning history to transition and connect new knowledge? Is the usage appropriate and supportive of the current step?
2. **Strategy Alignment** (s_2): Is the approach precise (targets the diagnosed misconception/goal) rather than providing generic help?
3. **Coherence** (s_4): Logical consistency (no contradictions, sequential steps) and language fluency (natural, helpful tutoring tone).
4. **Appropriateness** (s_3): Is it concise and aligned with the learner's **Zone of Proximal Development (ZPD)**? Is it accessible without excessive jargon?

FINAL OUTPUT FORMAT

Return strictly a JSON object:

```
{ "history_score": 1-5, "strategy_score": 1-5, "coherence_score": 1-5, "appropriateness_score": 1-5, "overall_score": 1-5, "reason": "Detailed justification..." }
```

Figure 6: LLM-based evaluation prompts. The Semantic Accuracy evaluator (*G*) performs factual consistency checks for Task 1, while the Multi-dimensional Judge assesses the pedagogical quality of Teaching Actions across four key dimensions.

Prompt Template for Model Inference (AI Tutor Evaluation)

[System Instruction]

ROLE

You are an expert AI Tutor. Analyze student data to answer queries, diagnose state and provide targeted teaching.

CONFIGURATION

[Diagnosis & Strategy Schema] (Always Present)

- **Recall Failure**: Forgot previously mastered knowledge... → Strategy: **Retrieval Practice**...
- **Conceptual Gap**: Fundamental misunderstanding... → Strategy: **Conceptual Explanation**...
- (...Definitions for Procedural Error and Transfer Deficit...)

[Optional Module: **Statistical Features Definitions**] (Condition: use_feature=True)

- related_concept_acc: The student's historical accuracy on related concepts.
- related_last_correct_interval_days: Time interval since last correct answer.
- cur_question_error_rate: Global Item Difficulty...

[Optional Module: **Diagnostic Heuristics**] (Condition: use_feature=True)

[Decision Matrix (Evaluate from Top to Bottom)]

Priority	Condition (History & Feature)	Question Type	Diagnosis
1	High Prior Acc + Long Gap (>4d)	Any	Recall Failure
2	Low Prior Acc OR New KC	Any	Conceptual Gap
3	Mid Acc + Short Gap	Novel Context	Transfer Deficit
4	Mid Acc + Short Gap	Complex/Multi-step	Procedural Error

[Optional Module: **Teaching Guidelines**] (Condition: use_teach=True)

Adhere to these 4 dimensions:

1. **History Utilization**: DO NOT use generic phrases. MUST explicitly reference specific evidence.
2. **Strategy Alignment**: Strict mapping (e.g., Recall Failure → Ask guiding question).
3. **Human-like Tone**: Avoid robotic headers. Use conversational markers.
4. **Appropriateness**: Keep language simple (K-6 level).

TASKS

1. **Evidence**: Answer queries based on History...
2. **Diagnosis**: Infer the student's state. Select ONE diagnosis from the Schema.
3. **Teaching**: Generate teaching response. Optimize against dimensions (if enabled).

OUTPUT FORMAT

Return strictly a JSON object:

```
{ "evidence": [...], "diagnosis": "...", "reason": "...", "strategy": "...", "teaching": "..." }
```

[User Input]

INTERACTION HISTORY

[...Chronological sequence of student exercises...]

[Optional Input: **Statistical Features**] (Condition: use_feature=True)

- related_concept_acc: 0.85
- related_last_correct_interval_days: 5.0
- cur_question_error_rate: 0.65
- (...calculated feature values...)

CURRENT QUESTION

[...Question Text and Metadata...]

EVIDENCE QUERIES

(Optional: Included if test_obj has queries)

- Q1: What concept did the student fail on 3 days ago?
- Analyze the data and generate the JSON response.

Figure 7: The inference prompt template fed to the AI Tutor models. The **Core Schema** is always present, while the **Statistical Features** module (definitions and heuristic logic) and **Teaching Guidelines** module are injected conditionally based on the experimental setting.

System Prompts for the LongTutor Annotation Pipeline

– STAGE 1: INITIAL GENERATOR PROMPT –

ROLE

Expert Educational Data Scientist. Generate a "Gold Standard" benchmark by analyzing student logs to create test cases and ideal teaching.

EVIDENCE CHECK TASK (For AI Tutor Evaluation) Generate 3 Q&A pairs (Simplified Chinese) with specific constraints:

- **Q1 (Information Extraction):** Verbatim detail from a single interaction record.
- **Q2 (Multi-session Reasoning):** Summary/pattern (count, accuracy, temporal logic) across multiple records. Use scope constraints (e.g., "within current concepts") to ensure uniqueness.
- **Q3 (Hallucination Check):** Ask about non-existent records/details. Answer must be "Unknown". No "Yes/No" or "Is there" phrasing.

DIAGNOSIS & STRATEGY DECISION MATRIX

Evaluate strictly from Top to Bottom based on Student History and Features:

Priority	Condition: History & Features	Condition: Question Type	Diagnosis	Strategy
1	High Prior Acc (>80%) & Long Gap (>4 days)	Any Question	Recall Failure	Retrieval Practice
2	Low Prior Acc (<50%) OR New Knowledge Point	Any Question	Conceptual Gap	Concept Explanation
3	Mid/High Acc (>50%) & Short Gap (<4 days)	Novel Context / Word Problem	Transfer Deficit	Analogical Transfer
4	Mid/High Acc (>50%) & Short Gap (<4 days)	Complex / Multi-step	Procedural Error	Stepwise Scaffolding

– STAGE 2: VERIFIER & OPTIMIZATION PROMPT –

VERIFICATION GOAL

Verify and refine the draft JSON. Principle: Minimal Intervention—only fix quality or logic issues.

QUALITY CHECKLIST

1. **Unambiguous Q&A:** Ensure Q2 is uniquely computable from the provided history slice alone.
2. **Grounding:** Q1 answer must be a verbatim extraction; Q3 must not use forbidden "Yes/No" patterns.
3. **Logic Alignment:** Re-check if the draft diagnosis correctly matches the Decision Matrix priorities.
4. **Teaching Humanization:** Ensure teaching content references specific past evidence (e.g., "I remember you solved [ID] correctly..."). Replace robotic transitions with conversational markers (e.g., "Wait a second...", "Notice how...").

FINAL OUTPUT FORMAT

Return strictly a JSON object:

```
{ "evidence": [...], "diagnosis": "...", "reason": "...", "strategy": "...", "teaching": "..." }
```

Figure 8: Complete system prompt pipeline. The Generator establishes initial diagnostic labels and evidence probes, while the Verifier enforces factual grounding, logical consistency across the decision matrix, and human-like instructional tone.