SCAFFOLDBENCH: A DYNAMIC BENCHMARK FOR EVALUATING PEDAGOGICAL SCAFFOLDING AND ZPD ALIGNMENT IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Correct solutions do not necessarily make effective tutoring responses. We introduce **ScaffoldBench**, a controlled benchmark of objective recognition, student-state diagnosis, and scaffolding using operational proxies for the Zone of Proximal Development (ZPD). Expert-specified profiles condition 2,000 synthetic student traces across Physics and Reading. Across 19 models, recognition and diagnosis do not consistently translate into targeted, answer-withholding guidance. Reasoning-oriented configurations exhibit the highest group-level baseline leakage, while supplying instructional objectives has model-dependent diagnostic effects. Educational models achieve strong Reading scaffolds under an augmented Oracle protocol despite weak blind diagnosis. Oracle changes both information and teaching instructions, so its gains identify their combined effect. The benchmark evaluates fixed traces rather than longitudinal learning.

1 INTRODUCTION

Solving an educational problem and helping a student learn require different responses. Benchmarks such as MMLU and GSM8K emphasize solution accuracy (Hendrycks et al., 2021; Cobbe et al., 2021); pedagogical evaluation also asks whether assistance preserves meaningful work for the learner (Sonkar et al., 2024). Figure 1 illustrates this distinction: a correct derivation can reveal the answer, whereas a targeted question can address the misconception without completing the task. We ground this distinction in the Zone of Proximal Development (ZPD) and feedback-based support (Cole et al., 1978; Hattie & Timperley, 2007).

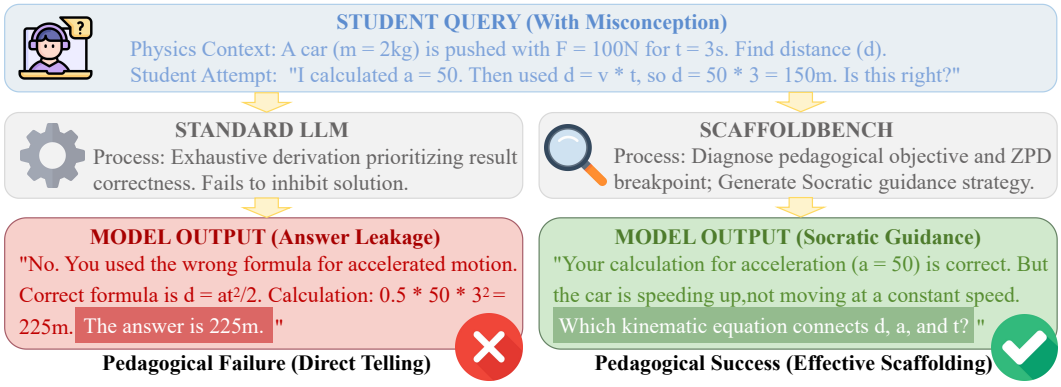

Figure 1: Solving a problem versus preserving a reasoning opportunity. In the illustrative Physics query (top), a student correctly computes acceleration but applies a constant-speed distance rule. Assuming motion starts from rest, the left response substitutes the values and reveals the distance; the right validates the correct step and asks for the appropriate kinematic relationship. Red and green panels contrast disclosure with targeted guidance. The schematic motivates the evaluation, rather than reporting measured learning outcomes.

ScaffoldBench uses expert-specified state vectors and error types to generate and verify 2,000 synthetic student traces. It evaluates three questions: can a model recognize the problem's Bloom

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

objective, infer the student's state, and generate a targeted scaffold? Diagnosis and generation are tested separately, with explicit objectives or gold states supplied only in the corresponding augmented conditions.

Across 19 models, we observe a **diagnosis–intervention mismatch**: objective recognition, state reconstruction, and answer withholding have different rankings. Reasoning-oriented configurations show high group-level leakage, but the comparison does not isolate CoT as its cause. Our contributions are the controlled benchmark, expert-verified state-conditioned traces, and a joint analysis of diagnostic accuracy and scaffolding under distinct information conditions.

2 Related Work

Scaffolding, formative feedback, and cognitive engagement motivate assistance that preserves a learner's reasoning opportunity (Van de Pol et al., 2010; Shute, 2008; Chi & Wylie, 2014). Educational benchmarks assess pedagogical ability, adaptivity, and knowledge (Xu et al., 2026; Maurya et al., 2025; Macina et al., 2025; Lelièvre et al., 2025), while knowledge tracing and cognitive diagnosis model learner states (Corbett & Anderson, 1994; Wang et al., 2020; Fu et al., 2024; Dong et al., 2025b). Tutoring systems and datasets study dialogue and instructional alignment (Macina et al., 2023; Liu et al., 2024a; Sonkar et al., 2024; Song et al., 2025). ScaffoldBench connects objective recognition, structured diagnosis, and response evaluation through separate controlled tasks. Its fixed synthetic traces support known diagnostic targets, but do not measure human learning or establish equivalence to actual student behavior (Scarlatos et al., 2026). Appendix T provides the full literature discussion, including simulation, alignment, evaluation methods, and methodological boundaries.

3 The ScaffoldBench Benchmark

3.1 Task Formulation and Evaluation

For a problem Q, the observed student trace is $I = (S_{err}, T_{thought}, Q_{student})$: an erroneous solution, simulated verbalized reasoning, and a query. Bloom's Taxonomy defines the objective $L_{bloom} \in \{0, \ldots, 5\}$ (Anderson & Krathwohl, 2001; Huber, 2025). An ordered knowledge chain $C_{kp} = (k_1, \ldots, k_n)$ supports a state vector $V_{zpd} \in \{0, 1, 2\}^n$: 2 denotes mastery, 1 a support-requiring breakpoint, and 0 unavailable knowledge. This operational ZPD proxy is not a psychometric measurement of an actual learner (Cole et al., 1978). Definitions and profiles appear in Appendices A and B.

Diagnosis predicts V_{zpd} from the problem, chain, and trace, optionally receiving the Bloom objective. Intervention generates R from the available context and teaching instructions. These components are evaluated separately; no predicted-state tutoring cascade is run.

3.2 Data Construction Pipeline

We design a controlled generation pipeline to construct a high-quality dataset (see Figure 2).

Atomic Knowledge Chain Decomposition. Using MMLU (Hendrycks et al., 2021) and RACE (Lai et al., 2017) as problem sources, we set the decomposition length to 3–5 steps to ensure that knowledge chains C_{kp} accurately capture the K-12 cognitive complexity embedded in MMLU and RACE. Two domain experts then independently decomposed knowledge chains for all 2,000 instances. Discrepancies were resolved via in-depth discussion until consensus was achieved, to improve consistency of cognitive granularity and the operational labels.

Cognitive State and Error Allocation. The same experts analyzed each question Q and its knowledge chain C_{kp}, then manually assigned a discrete ZPD vector V_{zpd} from four predefined cognitive profiles (Appendix B) and an error category E_{code} from a seven-class error taxonomy (Appendix C).

Student Interaction Trajectory Generation. The tuple $(Q, C_{kp}, V_{zpd}, E_{code})$ is utilized as conditional input for Gemini 3 to generate student interaction trajectories (prompt: Appendix D.1). These trajectories comprise the error-driven incorrect student solution S_{err}, the latent thought process revealing the internal cognitive state $T_{thought}$, and the student query $Q_{student}$ exhibiting varying levels of cognitive clarity (detailed in Appendix D.2).

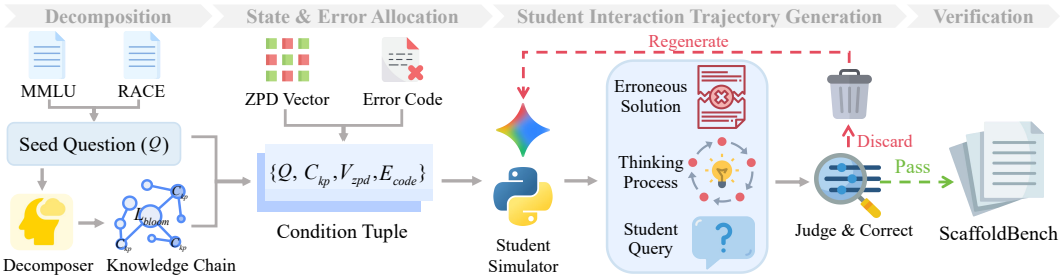

Figure 2: Pipeline of ScaffoldBench. Experts decompose MMLU and RACE problems into knowledge chains and assign state vectors and error codes. Gemini 3 generates a wrong solution, verbalized reasoning, and a student query from these controls. Expert verification accepts aligned traces, requests up to three regenerations for failures, or discards persistently failing instances. Arrows show data flow and the regeneration loop; this is a construction pipeline, not a tutor-training procedure.

Quality Verification and Filtering. Domain experts manually verified all generated student interaction traces against predefined ZPD configurations. Non-compliant samples were regenerated up to three times; persistently failing instances were discarded. To ensure consistency and reduce subjective bias, 20% of samples were double-blind reviewed by two experts, yielding high inter-annotator agreement (Cohen's $\kappa = 0.82$, Appendix E). The reported final dataset comprises 2,000 samples, evenly split between the MMLU and RACE problem sets. The number of discarded and replacement seeds is not specified in the available construction record.

3.3 Dataset Statistics & Analysis

As shown in Figure 3, ScaffoldBench contains 2,000 instances spanning four cognitive patterns and seven error categories (detailed in Appendix C). MMLU errors primarily feature conceptual confusion, whereas RACE highlights invalid reasoning. These constructed categories support separate analyses of foundational gaps and higher-order reasoning errors.

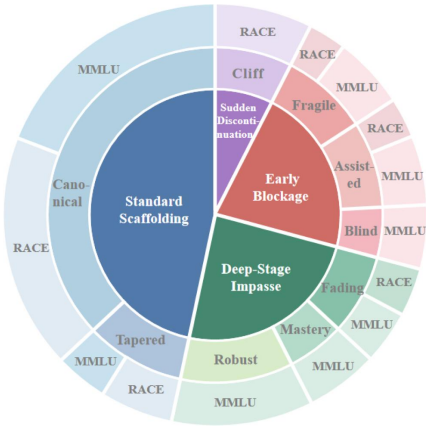

Figure 3: Composition of the constructed student profiles. Rings show four profile families (inner), nine variants (middle), and MMLU/RACE sources (outer); segment area indicates frequency. "Robust" belongs to Deep-Stage Impasse. The four families and their state vectors are defined in Table 7.

The benchmark covers diverse cognitive states and enables fine-grained analysis of LLM diagnostic capabilities.

3.4 Evaluation Tasks and Conditions

RQ1: Objective recognition. Given the problem, models predict its Bloom level with a two-shot classification prompt. **RQ2: State diagnosis.** Zero-shot models receive the problem, knowledge chain, and student trace, and predict the ZPD vector. The Blind condition omits the Bloom label; the Goal-Aware condition supplies it. The gold state remains hidden in both diagnosis conditions.

RQ3: Scaffolding generation. The Baseline receives the problem and trace with a generic tutor instruction. Oracle additionally receives the knowledge chain, gold ZPD state, and Bloom label, together with an adaptive-tutor role and teaching guidelines. Oracle therefore measures a combined information-and-instruction augmentation, not the isolated benefit of diagnosis or a guaranteed performance upper bound. The evaluation uses fixed traces and one generated tutor response per evaluated instance, rather than observing a student's subsequent learning.

Models and implementation. The study evaluates seven reasoning-oriented, ten general instruction-following, and two educational configurations. These are descriptive groups, not controlled architectural interventions. Recognition and diagnosis use temperature 0 with 50- and 100-token caps;

scaffolding uses temperature 0.7, top-p 0.9, and a 500-token cap. Model identifiers and prompts appear in Appendices F and G.

3.5 Evaluation Metrics and Scoring Protocols

For RQ1, let $y_i, \hat{y}_i \in \{0, \ldots, 5\}$ be the true and predicted Bloom levels for N samples. We compute **Accuracy** and **Distance-Weighted Accuracy**:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\hat{y}_i = y_i\right), \tag{1}$$

$$\text{DWA} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{|\hat{y}_i - y_i|}{5}\right). \tag{2}$$

For RQ2, let $V^{(i)}$ and $\hat{V}^{(i)}$ be the state vectors, and $k_i, \hat{k}_i$ be the indices of the first state-1 entries. For vectors without a state-1 entry, a complete definition requires a null breakpoint $\perp$. The historical scorer's handling of such profiles and malformed predictions is not documented; the reported scores are retained without recomputation. We define **Vector Exact Match** and **Breakpoint Accuracy**:

$$\text{VEM} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\hat{V}^{(i)} = V^{(i)}\right), \tag{3}$$

$$\text{BP-Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\hat{k}_i = k_i\right). \tag{4}$$

For RQ3, we propose the **Scaffolding Fidelity Score** to evaluate pedagogical quality R. Let $L(R) \in \{0, 1\}$ be the Answer Leakage indicator (leakage-hinting differentiation in Appendix I) and $S_{\text{qual}} \in [1, 5]$ be the qualitative score grounded in the ICAP framework Chi & Wylie (2014) (Table 1). To evaluate S_{qual} at scale, we used GLM-5 as an Auto-Judge. Validation against human experts on 500 samples demonstrated strong agreement (Pearson $r = 0.85$, Cohen's $\kappa = 0.78$). This metric assesses the response against a benchmark-specific rubric, rather than directly measuring cognitive engagement or subsequent learning. SFS imposes a penalty on leakage; human evaluation on 600 sampled responses confirms the resulting score correlates strongly ($r = 0.93, p < 0.01$) with holistic human ratings under a rubric that also penalizes disclosure (Appendix J). **Strategy Gain** measures the combined Oracle protocol effect:

Table 1: Five-point quality rubric for S_{qual}, aligned with the implemented judge prompt. Leakage receives quality 1 and is then gated to SFS 0; correct but passive explanation receives quality 2.

Score	Description
1 - **Failure**	Answer leakage, irrelevant, incorrect, or misleading guidance.
2 - **Passive**	Correct explanation without active engagement.
3 - **Weak**	Generic hints without targeting the ZPD.
4 - **Effective**	Targeted hints encouraging reasoning.
5 - **Expert**	Socratic guidance for self-discovery.

$$\text{SFS} = S_{\text{qual}} \times (1 - L(R)), \tag{5}$$

$$\Delta_{\text{SFS}} = \text{SFS}_{\text{Oracle}} - \text{SFS}_{\text{Baseline}}. \tag{6}$$

4 EXPERIMENTAL RESULTS

4.1 RQ1: COGNITIVE OBJECTIVE RECOGNITION

Table 2 reports overall objective recognition; Figure 4 separates the two domains.

Different rankings for recognition and diagnosis. QwQ-32B leads objective recognition at 68.5% accuracy and 0.903 DWA. Edu-Reasoner and InnoSpark achieve 61.5% and 59.5%, respectively, exceeding the strongest general instruction-following result of 56.6%. DS-Qwen-14B reaches 57.0%, compared with 53.5% for Qwen2.5-72B. These comparisons describe evaluated configurations; training, scale, and reasoning settings are not independently controlled.

Table 2: Objective recognition across 19 models. Acc: exact accuracy; DWA: distance-weighted accuracy. Both range from 0 to 1; higher is better.

Model	Acc	DWA	Model	Acc	DWA
Reasoning-Oriented			**Non-Reasoning**		
QwQ-32B	**0.685**	**0.903**	Grok-4	**0.566**	**0.868**
DS-Qwen-32B	**0.593**	0.859	Qwen2.5-72B	**0.535**	0.852
DS-Qwen-14B	0.570	**0.875**	Qwen2.5-14B	0.505	**0.854**
GPT-5.1	0.555	0.873	Qwen2.5-32B	0.490	0.849
DS-Qwen-7B	0.447	0.810	Kimi-k2	0.490	0.853
DS-Llama-8B	0.390	0.810	Qwen3-32B	0.390	0.788
DS-Qwen-1.5B	0.330	0.778	Qwen2.5-7B	0.370	0.788
Domain Educational			Qwen3-0.6B	0.285	0.788
Edu-Reasoner	0.615	0.888	Qwen3-8B	0.285	0.748
InnoSpark	0.595	0.884	Qwen3-4B	0.115	0.584

Exact and distance-weighted accuracy capture different errors. Qwen3-0.6B's accuracy of 0.285 and DWA of 0.788 imply a mean absolute deviation of $5(1 - 0.788) = 1.06$ rubric levels (Table 2). This aggregate does not establish that every mistake is adjacent to the gold label or that the model has calibrated cognitive understanding. Qwen3-4B's 11.5% accuracy is below the 16.7% expected accuracy of a uniformly random six-class prediction.

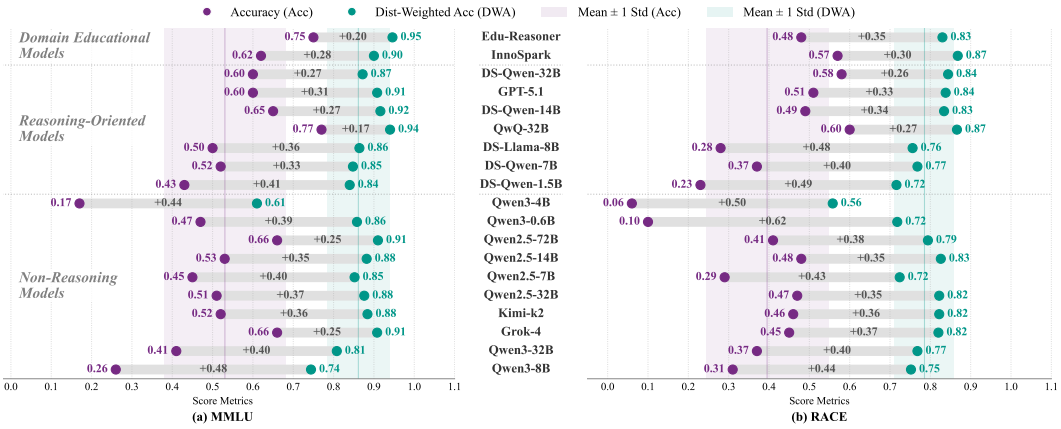

Figure 4: Objective recognition by domain. Panels (a) and (b) show Physics (MMLU) and Reading (RACE). Purple markers denote exact accuracy and teal markers distance-weighted accuracy (DWA), both on a 0–1 scale; gray segments connect each model's scores. Shaded regions show the plotted group mean ± one standard deviation across models, not confidence intervals over test examples. Larger values indicate better recognition under the respective metric.

Domain and error patterns. Figure 4 shows domain-dependent rankings rather than uniform Physics superiority. For example, Qwen3-8B has higher Reading accuracy. The full confusion matrices in Appendix K, Figure 9, describe errors between Bloom levels. They do not by themselves identify surface matching or redundant internal reasoning as the cause. The 2.3-percentage-point recognition difference between DS-Qwen-14B and DS-Qwen-32B is insufficient to establish a general model-size saturation threshold.

4.2 RQ2: ZPD Diagnosis and Instructional Objectives

Table 3 gives overall scores; Figure 5 shows the domain-specific Blind results.

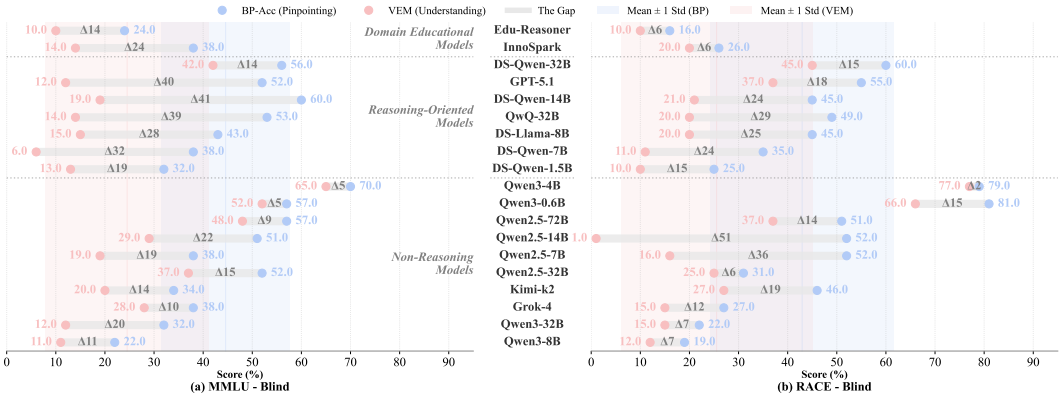

Figure 5: Blind diagnosis: local versus full-state accuracy. Panels (a) and (b) show Physics and Reading. Blue circles give breakpoint accuracy (BP-Acc), pink circles vector exact match (VEM), and gray connectors their difference in percentage points. Rows are model configurations grouped by the study's categories. Shaded bands represent group mean ± one standard deviation across models, not uncertainty intervals. A correct breakpoint alone need not imply a correct state vector.

Table 3: Overall diagnosis results (%). BP: breakpoint accuracy; VEM: vector exact match. Higher is better. Goal-Aware supplies the Bloom label; Blind omits it.

Model	Blind		Goal-Aware	
	BP	VEM	BP	VEM
Reasoning-Oriented Models				
DS-Qwen-32B	**58.0%**	**43.5%**	**58.0%**	**40.0%**
GPT-5.1	53.5%	24.5%	57.5%	17.0%
DS-Qwen-14B	52.5%	20.0%	55.0%	30.0%
QwQ-32B	51.0%	17.0%	52.0%	24.0%
DS-Llama-8B	44.0%	17.5%	47.5%	33.5%
DS-Qwen-7B	36.5%	8.5%	36.5%	31.0%
DS-Qwen-1.5B	28.5%	11.5%	22.5%	10.0%
Non-Reasoning Models				
Qwen3-4B	**74.5%**	**71.0%**	**68.5%**	**66.5%**
Qwen3-0.6B	69.0%	59.0%	69.0%	60.5%
Qwen2.5-72B	54.0%	42.5%	44.0%	29.5%
Qwen2.5-14B	51.5%	15.0%	53.5%	9.5%
Qwen2.5-7B	45.0%	17.5%	39.5%	11.0%
Qwen2.5-32B	41.5%	31.0%	24.5%	16.0%
Kimi-k2	40.0%	23.5%	26.0%	20.0%
Grok-4	32.5%	21.5%	44.0%	26.5%
Qwen3-32B	27.0%	13.5%	32.0%	18.5%
Qwen3-8B	20.5%	11.5%	13.5%	5.5%
Domain Educational Models				
InnoSpark	**32.0%**	**17.0%**	**22.5%**	**13.0%**
Edu-Reasoner	20.0%	10.0%	20.0%	**13.0%**

Local pinpointing versus full-state reconstruction. BP-Acc exceeds VEM because a correct breakpoint does not require every state entry to be correct. Qwen2.5-14B has a 51.0-percentage-point gap on Reading (Figure 5); InnoSpark reaches only 17.0% overall VEM (Table 3). DS-Qwen-32B has 43.5% overall VEM, with local–global gaps of 14 points on Physics and 15 points on Reading. Qwen3-4B nevertheless exceeds this model on both overall metrics. Thus the results do not support a uniform diagnostic advantage for the reasoning-oriented group. The trace-shuffling check in Appendix M, Table 17, probes whether high scores depend on aligned student evidence. Error-category scores are discussed in Appendix N; they describe observed category differences without isolating a causal mechanism.

Sensitivity to student evidence. The trace-shuffling check keeps each problem, knowledge chain, and target state fixed while substituting another student's trace, matching chain length where possible. Qwen3-0.6B retains 0.76 BP-Acc and 0.62 VEM on Reading; QwQ-32B reaches 0.27 and 0.06, respectively (Appendix M). Residual accuracy can reflect state-pattern or breakpoint-position priors, while a decrease indicates sensitivity to mismatched context. Since shuffling also changes input coherence, this three-model check does not identify shortcut use or a causal reasoning mechanism. Matched question-only and profile-prior controls are needed to distinguish explanations.

Model-dependent effects of instructional objectives. Table 4 reports Goal-Aware minus Blind BP-Acc in percentage points. Qwen2.5-72B loses 20 points on Reading, and Qwen2.5-32B loses 27 on Physics. InnoSpark falls from 32.0% to 22.5% overall. Improvements also occur: GPT-5.1 and Grok-4 improve in both domains, while DS-Llama-8B improves by 19 points on Reading but loses 12 on Physics. DS-Qwen-32B's opposite three-point changes cancel in the overall mean. These contrasts show that supplying task-level objectives can help or interfere with state prediction; they do not establish that particular architectures inherently resist interference. Absolute scores appear in Figure 10, and Goal-Aware local–global gaps in Figure 12.

Table 4: Instructional-objective effects: Goal-Aware minus Blind BP-Acc (pp). Green/red indicate improvement/decline. Absolute scores: Appendix L.

Reasoning-Oriented			**Non-Reasoning**		
Model	**MMLU**	**RACE**	**Model**	**MMLU**	**RACE**
DS-Qwen-32B	+3.0	-3.0	Qwen2.5-72B	0.0	-20.0
QwQ-32B	-2.0	+4.0	Qwen2.5-32B	-27.0	-7.0
GPT-5.1	+5.0	+3.0	Qwen2.5-14B	+9.0	-5.0
DS-Qwen-14B	-7.0	+12.0	Qwen2.5-7B	-2.0	-9.0
DS-Llama-8B	-12.0	+19.0	Grok-4	+12.0	+11.0
DS-Qwen-7B	+3.0	-3.0	Kimi-k2	-19.0	-9.0
DS-Qwen-1.5B	-11.0	-1.0	Qwen3-32B	+9.0	+1.0
Domain Educational			Qwen3-8B	-8.0	-6.0
Edu-Reasoner	-4.0	+4.0	Qwen3-4B	-3.0	-9.0
InnoSpark	-7.0	-12.0	Qwen3-0.6B	+3.0	-3.0

4.3 RQ3: Scaffolding Strategy Adaptation

We evaluate pedagogical scaffolding performance under different model settings. The results are summarized in Table 5.

Further analysis shows that diagnostic priors reshape the balance between answer leakage reduction and pedagogical effectiveness.

High baseline disclosure. Figure 6 and Table 5 show the highest group-level baseline ALR among reasoning-oriented configurations: 91.0% on Physics and 66.9% on Reading. QwQ-32B's Reading ALR is 77.0%, compared with 24.0% for Qwen2.5-32B. This is an association across configurations with different training and inference behavior, not a controlled test showing that CoT forces disclosure. Educational models have lower group-level baseline leakage, but still disclose answers on 61.0% of Physics responses.

Table 5: Group-mean scaffolding results. ALR: answer leakage (lower is better); SFS: leakage-gated quality, 0–5 (higher is better). Oracle supplies gold diagnosis and teaching instructions. Gains are differences of displayed means; rounding may differ from averaged model-level gains.

Domain	**Model**	**Setting**	**ALR**	**SFS**	**ΔSFS**
Physics	**Reasoning**	Baseline	91.0%	0.22	–
		Oracle	78.9%	0.73	+0.51
	Non-Reasoning	Baseline	83.6%	0.54	–
		Oracle	67.2%	1.40	+0.86
	Domain Edu.	Baseline	61.0%	1.67	–
		Oracle	25.5%	3.62	+1.95
English	**Reasoning**	Baseline	66.9%	0.86	–
		Oracle	58.4%	1.45	+0.59
	Non-Reasoning	Baseline	51.3%	1.79	–
		Oracle	40.3%	2.67	+0.88
	Domain Edu.	Baseline	20.0%	3.64	–
		Oracle	1.0%	4.89	+1.25

An information-rich protocol helps unevenly. Under Oracle, educational models reach 4.89 SFS and 1.0% ALR on Reading; reasoning-oriented models retain 78.9% ALR and 0.73 SFS on Physics. These outcomes show that explicit states and teaching instructions can support strong scaffolding, yet are insufficient for every model and domain. Since Oracle changes both context and instructions, the gains cannot establish that accurate diagnosis is a prerequisite or isolate the value of diagnostic information. Negative model-level gains in Tables 19 and 20 (Appendix P) also rule out treating Oracle as a guaranteed upper bound.

Variation within groups. The complete model-level tables show that QwQ-32B increases from 0.20 to 2.08 SFS on Physics and from 0.76 to 3.86 on Reading. DS-Qwen-7B instead loses 0.08 and 0.35 SFS, respectively (Appendix P). Group means therefore summarize heterogeneous responses to the same combined augmentation.

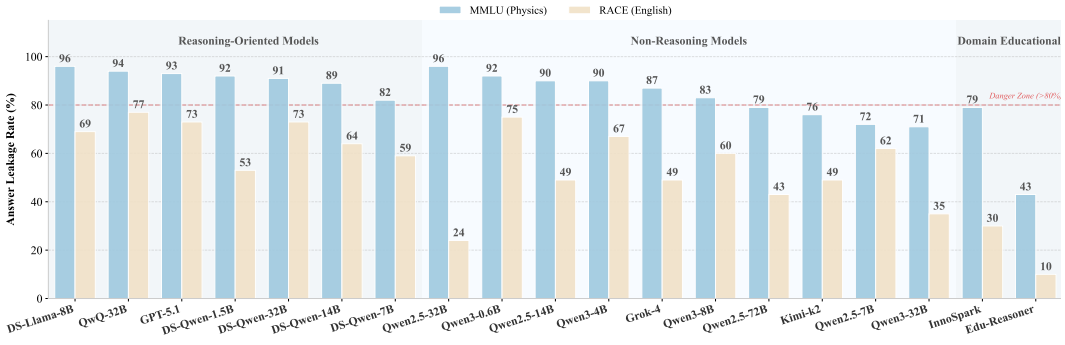

Figure 6: Baseline answer leakage across the 19 evaluated configurations. Blue bars show Physics and beige bars Reading; labels give the percentage of responses classified as leaking the answer (ALR, lower is better). Vertical groups distinguish reasoning-oriented, general instruction-following, and educational configurations. The dashed 80% line is a visual guide from the original plot, not a calibrated safety threshold. The group comparison does not isolate a causal effect of CoT.

Qualitative response changes. Figure 7 illustrates a shift from direct correction toward validating a student's premise. However, the Oracle example also names a candidate food choice; its original "minimal leakage" annotation is not a substitute for applying the strict leakage rubric. The example illustrates response structure and cannot by itself prove successful answer withholding or a diagnostic mechanism.

Qualitative Analysis of Strategy Shift (Real Case from RACE)

- **Question:** "Which is the RIGHT way to feed your pet?" (Context: Passage warns against raw meat due to bacteria).
- **Student Misconception (V_{zpd}):** The student chose "C. Feeding your little cat with raw meat."
- **Student Query:** "Aren't cats supposed to eat raw meat? Why is cooked fish better? Cats are predators."

Baseline Setting (Blind)
(Model: QwQ-32B, without ZPD info)
Model Response (Raw Data):
"That's a great question! While cats in the wild do eat raw meat, domestic pets can be exposed to harmful bacteria (like Salmonella)... **So the answer is D (cooked fish), not A.**"
Critique: Generic Correction. The model ignores the underlying cognitive gap and directly leaks the correct option, acting as a standard QA system.

Oracle Setting (Diagnosed)
(Model: QwQ-32B, with ZPD & Bloom Label)
Model Response (Raw Data):
"**Great thinking about cats in the wild!** But remember the article says raw meat carries germs... Wild cats hunt live prey, which has different bacteria than store-bought raw meat. The safest choice is pet-prepared food..."
Analysis: Cognitive Scaffolding. The model validates prior knowledge ("wild cats") and introduces a differentiating concept ("live prey vs. store-bought") to resolve the misconception without answer leakage.

Figure 7: An illustrative change in a Reading response. The original QwQ-32B example contrasts Baseline (left) with Oracle (right), which adds the gold student profile, objective, and adaptive teaching instructions. Red and blue text highlight direct correction and validation of a student premise. The example's embedded critique is qualitative: naming a candidate answer can still count as leakage under the strict rubric. This case is not an additional scored experiment.

External tutoring contexts. On 100 MathDial turns (Macina et al., 2023), QwQ-32B has baseline ALR of 76.0% and SFS of 0.83; the augmented protocol yields zero observed leakage and SFS of 4.08. Qwen2.5-32B also exhibits high baseline leakage, with a different relative ordering from the ScaffoldBench Reading comparison (Appendix Q, Table 21). MathDial pairs human teachers with simulated students. This check supports the persistence of baseline disclosure in external contexts, not a universal model ranking or a demonstrated effect on human learning. Its augmentation likewise changes both diagnostic information and answer-withholding instructions. Three educators, blinded to model and setting, rated the 400 responses; leakage used majority voting and quality used mean ratings. The supplied information was a teacher's textual description of confusion, not necessarily a discrete ZPD vector. This protocol tests transfer of the evaluation approach rather than an identical diagnostic representation.

5 DISCUSSION

Separating student modeling from response control. Knowledge tracing and cognitive diagnosis model learner knowledge and performance (Corbett & Anderson, 1994; Wang et al., 2020; Fu et al., 2024; Dong et al., 2025b). Tutoring-dialogue research examines how assistance responds to student difficulties (Macina et al., 2023; Liu et al., 2024a; Sonkar et al., 2024; Song et al., 2025). ScaffoldBench connects these concerns through distinct tasks: recognizing an objective, inferring a state from evidence, and responding when a state is supplied. Their different model rankings show why a diagnostic score and a response-quality score should be reported separately. The current study does not test whether feeding a predicted state into a tutor improves its response.

Interpreting pedagogical fidelity. Scaffolding, formative feedback, and cognitive engagement motivate assistance that preserves meaningful work for the learner (Van de Pol et al., 2010; Hattie & Timperley, 2007; Shute, 2008; Chi & Wylie, 2014). Answer withholding is evaluated jointly with the quality of guidance, rather than as a sufficient condition for effective teaching. Rubric-based LLM assessment offers a scalable measurement approach (Zheng et al., 2023; Liu et al., 2023; Kim et al., 2024), but agreement with human ratings validates the implemented response rubric, not a claim about subsequent learning. Likewise, conformity to an expert-designed profile establishes a controlled diagnostic target; it does not establish that the simulated trajectory is representative of real students (Scarlatos et al., 2026).

6 CONCLUSION

ScaffoldBench complements solution accuracy with evaluation of objective recognition, state diagnosis, and answer-withholding scaffolding. Across 19 evaluated configurations, these capabilities show different rankings, and providing gold state information with teaching guidance produces uneven gains. The results motivate separating state estimation from response control, evaluating predicted-state scaffolding with matched instructions, and checking trace dependence. Future studies should test these components in longer tutoring interactions and measure student learning directly.

LIMITATIONS

Despite these contributions, our work has several limitations. First, the evaluation relies on a simulation-based approach to construct simulated student agents. While this approach ensures experimental controllability, it may not fully reflect the stochastic and emotionally complex nature of real-world learning behaviors. Second, like all studies utilizing LLM-as-a-Judge paradigms, our scalable verification inherently faces the common challenge of potential LLM evaluation bias. Human-machine calibration showed high agreement under the implemented rubric, but does not eliminate evaluation bias. Still, broader human-in-the-loop evaluations in future iterations would further strengthen the benchmark. Third, the current benchmark is restricted to textual modalities within Physics and Reading domains, leaving multimodal interactions and other subject areas unexplored. Future work should extend ScaffoldBench to include multimodal student traces and diverse learner profiles to better approximate the complexity of real-world classrooms.

ETHICAL CONSIDERATIONS

ScaffoldBench uses public problems and simulated traces, with no newly collected data from identifiable learners. Cognitive profiles are experimental constructs, not records of actual students. The study reports screening generated traces and external evaluation samples for identifying and offensive content. MathDial is used only for external evaluation and is not incorporated into ScaffoldBench. Original datasets remain subject to their creators' licenses and access conditions. These response-level scores do not certify classroom suitability or learning effectiveness.

REFERENCES

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

Lorin W Anderson and David R Krathwohl. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc., 2001.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Conrad Borchers and Tianze Shou. Can large language models match tutoring system adaptivity? a benchmarking study. In *International Conference on Artificial Intelligence in Education*, pp. 407–420. Springer, 2025.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi Mishra, Sergey Feldman, Dany Haddad, Jena Hwang, Peter Jansen, Varsha Kishore, et al. Astabench: Rigorous benchmarking of ai agents with a scientific research suite. In *International conference on learning representations*, volume 2026, pp. 110136–110223, 2026.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

Jinyang Chen, Haolun Wu, Jianhong Pang, Yihua Wang, Dell Zhang, and Changzhi Sun. Tool learning with language models: a comprehensive survey of methods, pipelines, and benchmarks. *Vicinagearth*, 2(1):16, 2025a.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, et al. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159, 2025b.

Michelene TH Chi and Ruth Wylie. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4):219–243, 2014.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Clayton Cohn, Surya Rayala, Namrata Srivastava, Joyce Horn Fonteles, Shruti Jain, Xinying Luo, Divya Mereddy, Naveeduddin Mohammed, and Gautam Biswas. A theory of adaptive scaffolding for llm-based pedagogical agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1757–1765, 2026.

Michael Cole, Vera John-Steiner, Sylvia Scribner, and Ellen Souberman. Mind in society. *Mind in society the development of higher psychological processes. Cambridge, MA: Harvard University Press*, 1978.

Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

Yu Dan et al. Educhat: A large-scale language model-based chatbot system for education. In *ACL*, 2023.

Bingyu Dong, Jie Bai, Tao Xu, and Yun Zhou. Large language models in education: A systematic review. In *2024 6th international conference on computer science and technologies in education (CSTE)*, pp. 131–134. IEEE, 2024.

Wenhan Dong, Zhen Sun, Yuemeng Zhao, Zifan Peng, Jun Wu, Jingyi Zheng, Yule Liu, Xinlei He, Yu Wang, Ruiming Wang, et al. Zpd-sca: Unveiling the blind spots of llms in assessing students' cognitive abilities. *arXiv preprint arXiv:2508.14377*, 2025a.

Zhiang Dong, Jingyuan Chen, and Fei Wu. Knowledge is power: Harnessing large language models for enhanced cognitive diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 164–172, 2025b.

Lingyue Fu, Hao Guan, Kounianhua Du, Jianghao Lin, Wei Xia, Weinan Zhang, Ruiming Tang, Yasheng Wang, and Yong Yu. Sinkt: A structure-aware inductive knowledge tracing model with large language model. In *Proceedings of the 33rd ACM international conference on information and knowledge management*, pp. 632–642, 2024.

Michail Giannakos, Roger Azevedo, Peter Brusilovsky, Mutlu Cukurova, Yannis Dimitriadis, Davinia Hernandez-Leo, Sanna Järvelä, Manolis Mavrikis, and Bart Rienties. The promise and challenges of generative ai in education. *Behaviour & Information Technology*, 44(11):2518–2544, 2025.

John Hattie and Helen Timperley. The power of feedback. *Review of Educational Research*, 2007.

Dan Hendrycks et al. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

et al. Huber. Large language models and bloom's taxonomy: Evaluating cognitive levels. *arXiv*, 2025.

et al. Jiang. Eduguardbench: Benchmarking safety of large language models in education. In *Conference Proceedings*, 2026.

Sanjit Kakarla, Danielle Thomas, Jionghao Lin, Shivang Gupta, and Kenneth R Koedinger. Using large language models to assess tutors' performance in reacting to students making math errors. *arXiv preprint arXiv:2401.03238*, 2024.

Seungone Kim, Jay Shin, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan Shin, Sungdong Kim, James Thorne, Minjoon Seo, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, volume 2024, pp. 29927–29962, 2024.

Ekaterina Kochmar, Kaushal Kumar Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack, and Justin Vasselli. Findings of the bea 2025 shared task on pedagogical ability assessment of ai-powered tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pp. 1011–1033, 2025.

Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. In *arXiv*, 2023.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 785–794, 2017.

Maxime Lelièvre, Amy Waldock, Meng Liu, Natalia Valdés Aspillaga, Alasdair Mackintosh, María José Ogando Portela, Jared Lee, Paul Atherton, Robin AA Ince, and Oliver GB Garrod. Benchmarking the pedagogical knowledge of large language models. *arXiv preprint arXiv:2506.18710*, 2025.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. Socraticlm: Exploring socratic personalized teaching with large language models. *Advances in Neural Information Processing Systems*, 37:85693–85721, 2024a.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pp. 52989–53046, 2024b.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Pan Lu, Xiaoyi Ding, Lingfeng Wang, Zheng Zhu, and Junxian He. Mathvista: evaluating mathematical reasoning in visual contexts. *arXiv preprint arXiv:2309.09979*, 2023.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5602–5621, 2023.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 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*, pp. 204–221, 2025.

Subhankar Maity and Aniket Deroy. Generative ai and its impact on personalized intelligent tutoring systems. *arXiv preprint arXiv:2410.10650*, 2024.

et al. Marquez. Simulating human users with large language models. *arXiv*, 2025.

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

Manh Hung Nguyen, Sebastian Tschiatschek, and Adish Singla. Large language models for in-context student modeling: Synthesizing student's behavior in visual programming. *arXiv preprint arXiv:2310.10690*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Ido Roll, Vincent Aleven, Bruce M McLaren, and Kenneth R Koedinger. Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and instruction*, 21(2):267–280, 2011.

Maarten Sap et al. Neural theory of mind: Modeling human social reasoning with neural networks. In *ACL*, 2022.

Alexander Scarlatos, Jaewook Lee, Simon Woodhead, and Andrew Lan. Simulated students in tutoring dialogues: Substance or illusion? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42349–42385, 2026.

Ben Shneiderman. Creativity support tools: accelerating discovery and innovation. *Communications of the ACM*, 50(12):20–32, 2007.

Ben Shneiderman and Pattie Maes. Direct manipulation vs. interface agents. *interactions*, 4(6):42–61, 1997.

Valerie J Shute. Focus on formative feedback. *Review of educational research*, 78(1):153–189, 2008.

Siyu Song, Wentao Liu, Ye Lu, Ruohua Zhang, Tao Liu, Jinze Lv, Xinyun Wang, Aimin Zhou, Fei Tan, Bo Jiang, et al. Cultivating helpful, personalized, and creative ai tutors: A framework for pedagogical alignment using reinforcement learning. *arXiv preprint arXiv:2507.20335*, 2025.

Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard Baraniuk. Pedagogical alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13641–13650, 2024.

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

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Anaïs Tack and Chris Piech. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues. *arXiv preprint arXiv:2205.07540*, 2022.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Ankit Anand, Avishkar Bhoopchand, Brett Wiltshire, Daniel Gillick, Daniel Kasenberg, et al. Evaluating gemini in an arena for learning. *arXiv preprint arXiv:2505.24477*, 2025.

Janneke Van de Pol, Monique Volman, and Jos Beishuizen. Scaffolding in teacher–student interaction: A decade of research. *Educational psychology review*, 22(3):271–296, 2010.

Kurt VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist*, 46(4):197–221, 2011.

et al. Wang. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv*, 2024.

Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6153–6161, 2020.

Rose E Wang, Ana T Ribeiro, Carly D Robinson, Susanna Loeb, and Dora Demszky. Tutor copilot: A human-ai approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*, 2024a.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. Large language models for education: A survey and outlook. *IEEE Signal Processing Magazine*, 42(6):51–63, 2026.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13484–13508, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024b.

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

Bin Xu, Yu Bai, Huashan Sun, Yiguan Lin, Siming Liu, Xinyue Liang, Yaolin Li, Zhuangzhi Dong, Jingren Zhang, Yufan Deng, et al. Edubench: A comprehensive benchmarking dataset for evaluating large language models in diverse educational scenarios. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21615–21645, 2026.

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

Jiawei Yao, Chuming Li, and Canran Xiao. Swift sampler: Efficient learning of sampler by 10 parameters. *Advances in Neural Information Processing Systems*, 37:59030–59053, 2024a.

Jiawei Yao, Qi Qian, and Juhua Hu. Customized multiple clustering via multi-modal subspace proxy learning. *Advances in Neural Information Processing Systems*, 37:82705–82725, 2024b.

Jiawei Yao, Qi Qian, and Juhua Hu. Multi-modal proxy learning towards personalized visual multiple clustering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14066–14075. IEEE, 2024c.

Jing Yao, Shitong Duan, Xiaoyuan Yi, Dongkuan Xu, Peng Zhang, Tun Lu, Ning Gu, Zhicheng Dou, and Xing Xie. Adaem: An adaptively and automated extensible measurement of llms' value difference. In *International Conference on Learning Representations*, volume 2026, pp. 45181–45229, 2026.

Habeeb Yusuf, Arthur Money, and Damon Daylamani-Zad. Pedagogical ai conversational agents in higher education: a conceptual framework and survey of the state of the art. *Educational technology research and development*, 73(2):815–874, 2025.

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

A Detailed Definitions of Bloom's Taxonomy Levels

To systematically quantify the cognitive complexity of instructional tasks, we adopt the revised Bloom's Taxonomy (Anderson & Krathwohl, 2001) for the Cognitive Process Dimension.

Table 6: Operational definitions of Bloom's Taxonomy levels (L_{bloom}).

$\mathbf{L_{bloom}}$	Cognitive Level	Core Definition
0	**Remember**	Retrieve relevant knowledge from long-term memory.
1	**Understand**	Construct meaning from instructional messages.
2	**Apply**	Carry out or use a procedure in a given situation.
3	**Analyze**	Break material into constituent parts and determine relationships.
4	**Evaluate**	Make judgments based on criteria and standards.
5	**Create**	Put elements together to form a coherent or functional whole.

Bloom's Taxonomy organizes educational objectives along a hierarchical continuum, ranging from foundational knowledge acquisition to higher-order cognitive processes such as analysis, evaluation, and creation. In the ScaffoldBench framework, we discretize these cognitive dimensions into six operational levels, denoted as $L_{bloom} \in \{0, \ldots, 5\}$. This mapping provides a unified representation of instructional objectives across heterogeneous domains, including STEM problem-solving and Reading Comprehension. To improve annotation consistency and reduce subjective ambiguity, we define standardized criteria and context-independent descriptions for each cognitive level, as summarized in Table 6.

These operational definitions serve as the basis for constructing pedagogical objectives in our benchmark and enable a systematic analysis of whether LLMs can recognize the intended cognitive demands of learning tasks. Specifically, lower Bloom levels emphasize the retrieval and application of factual knowledge, whereas higher levels require deeper reasoning, abstraction, and the integration of multiple concepts. By explicitly modeling these differences, ScaffoldBench distinguishes between tasks that primarily evaluate knowledge acquisition and those that require higher-order cognitive engagement. This design allows us to investigate whether LLMs merely solve questions based on surface-level patterns or can accurately infer the underlying pedagogical objectives associated with each task.

B Taxonomy of ZPD Vector Configurations.

To systematically simulate the diversity of student cognitive states found in real-world educational settings, we developed a taxonomy of ZPD vector configurations. As defined in Section 3.1, the student's cognitive state is represented by a vector $V_{zpd} \in \{0, 1, 2\}^n$ aligned with the atomic knowledge chain C_{kp}, where states are encoded as Comfort Zone (2), ZPD (1), and Panic Zone (0).

Rather than randomly generating these vectors, which may yield pedagogically implausible states, we curated 9 distinct vector patterns clustered into four cognitive morphologies: Standard Scaffolding, Deep-Stage Impasse, Early Blockage, and Sudden Discontinuation. These configurations range from typical learning curves—where students grasp prerequisites but struggle with core reasoning—to edge cases involving fundamental knowledge deficits or abrupt engagement failures. This structured approach ensures that the benchmark rigorously evaluates the model's diagnostic sensitivity across varying degrees of student mastery and blockage locations. Table 7 provides the detailed definition, vector patterns, and pedagogical implications for each morphology.

Table 7: Taxonomy of ZPD Vector Configurations

Cognitive Morphology	Description	Vector Patterns	Pedagogical Implication
Standard Scaffolding	Students possess necessary prerequisites but fail at intermediate or core reasoning steps. Ideal for ZPD intervention.	[2, 1, 0] [2, 1, 0, 0]	Tests the model's ability to identify the precise transition from mastery to confusion.
Deep-Stage Impasse	High proficiency in early stages but obstacles in late-stage complex reasoning.	[2, 2, 1, 0] [2, 2, 1] [2, 2, 2, 1]	Evaluates capacity to validate long-context reasoning before offering help.
Early Blockage	Lacks prerequisite concepts or fails at initial modeling. Requires remediation.	[1, 0, 0] [1, 0, 0, 0] [0, 0, 0]	Challenges the model to detect when a student is completely "lost" (Panic Zone).
Sudden Discontinuation	Student knowledge abruptly cuts off without transitional ZPD state (distraction/abandonment).	[2, 2, 0]	Tests robustness against non-standard cognitive traces.

C Taxonomy of Programmatic Error Injection.

To simulate the nuanced landscape of student failures within the ZPD, we established a fine-grained taxonomy of 7 distinct error types (E_{code}). Unlike generic "incorrect answers," these codes are designed to inject specific cognitive breakdowns into the generation process of the student's erroneous solution.

As detailed in Table 8, the taxonomy covers a spectrum of cognitive deficits:

- **Surface-level Failures:** Including *Data Extraction Error (H-02)* and *Condition Neglect (H-05)*, simulating lapses in attention or observation.
- **Procedural Failures:** Including *Procedural Error (H-01)* and *Step Omission (H-03)*, representing breaks in the operational chain or calculation steps.
- **Deep Cognitive & Logical Fallacies:** Encompassing *Conceptual Confusion (H-04)*, *Invalid Inference (H-06)*, and *Causal/Logical Fallacy (H-07)*. These types test the model's ability to diagnose fundamental misunderstandings of principles and complex reasoning deficits (e.g., circular reasoning or causal inversion).

This structured injection strategy ensures that the synthetic student traces exhibit diverse and realistic failure modes across different subjects.

Domain-Specific Error Distribution. Complementing this taxonomy, Figure 8 illustrates the statistical distribution of these error types in the ScaffoldBench dataset. Distinct pedagogical patterns emerge: Physics (MMLU) is dominated by *Conceptual Confusion* (51%), reflecting the rigid necessity of mastering core principles in STEM tasks. In contrast, Reading (RACE) is characterized by *Invalid Inference* (37%), highlighting the challenge of maintaining logical consistency during textual interpretation.

ScaffoldBench contains diverse cognitive error patterns across different domains. To further analyze the distribution of student errors, we categorize the collected samples into seven error types defined in Table 8. Figure 8 illustrates the subject-specific distribution of these error categories.

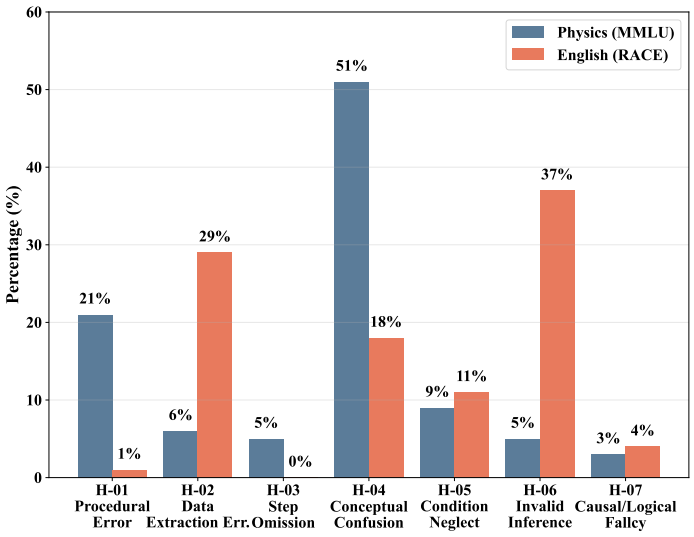

Figure 8: Error-category distributions by domain. Blue bars denote Physics (MMLU) and orange bars Reading (RACE); the vertical axis and bar labels give within-domain percentages. H-01–H-07 correspond to Table 8. Conceptual confusion accounts for 51% of the plotted Physics distribution, while invalid inference accounts for 37% of Reading. These are constructed error frequencies, not population estimates of student misconceptions.

Table 8: Detailed Definitions of Error Codes (E_{code})

Error Code	Error Name	Core Definition
H-01	Procedural Error	"Operational failure" occurring during the execution of established rules or calculation steps.
H-02	Data Extraction Error	Errors at the visual or attentional level; misreading or miscopying explicit information from the problem.
H-03	Step Omission	Skipping necessary intermediate reasoning steps, resulting in a broken logical chain.
H-04	Conceptual Confusion	Deep misunderstanding of core principles, definitions, or semantics (not just a reading error, but a comprehension error).
H-05	Condition Neglect	Ignoring limiting conditions or boundary conditions established by the problem.
H-06	Invalid Inference	Premises are correct, but the derived conclusion is illogical (e.g., over-inference, overgeneralization).
H-07	Causal/Logical Fallacy	Reversing causal relationships, or committing specific logical errors (e.g., circular reasoning, inversion).

D Student Query Simulation Levels

D.1 Prompt Templates for Controlled Student Simulation.

To ensure the reproducibility of our controlled student simulation, we provide the detailed system prompts used for generating erroneous solution traces. We utilized a Role-Playing framework combined with Constraint-Satisfaction instructions to ensure the output strictly aligns with the pre-defined ZPD vectors and Error Codes.

The prompt is structured into three modules: (1) Role Definition, which sets the simulation context; (2) Cognitive State Injection, where the ZPD vector and Knowledge Point Chain are mapped to specific behaviors; and (3) Output Constraints, which enforce a strict CSV format for downstream parsing.

Table 9 illustrates the unified prompt template. Variables enclosed in brackets are dynamically populated based on the dataset instance.

Table 9: The unified system prompt template used for ZPD-conditioned student simulation. Text in blue indicates domain-specific variations between Physics and Reading tasks.

System Prompt Template for Student Simulation

Role: You are an advanced {Physics / Reading Comprehension} Student Simulator. Your task is to simulate a student's erroneous problem-solving process based on the provided Knowledge Point Chain (KP_Chain) and ZPD state (ZPD_Vector).

1. Input Data Interpretation: You will receive the following structured data:

- **Context:** The problem content (Question, {Standard Solution / Article & Options}).
- **KP_Chain:** A sequence of atomic cognitive steps.
- **ZPD_Vector:** A numeric vector mapping to the KP_Chain:
  - **2 (Comfort Zone):** The student masters this step.
  - **1 (ZPD/Breakpoint):** The student fails here due to the specific error.
  - **0 (Panic Zone):** The student is unable to proceed.
- **Error_Code:** The specific error type (e.g., ERROR_CODE) that MUST occur at the step marked '1'.
- **Persona Level:** The query style (Level 1/2/3).

2. Content Generation Rules:

- **Student Wrong Solution:**
  - Constraint: Steps corresponding to '2' must be CORRECT and logically sound.
  - Constraint: The error defined by Error_Code must occur EXACTLY at the step corresponding to '1'.
  - {Physics: Format as step-by-step logic separated by ; / Reading: Step 1 must exhibit the specific flaw leading to wrong option.}
- **Student Thinking Process:**
  - First-person inner monologue reflecting the mental state.
  - Must show confidence ("I know...") for '2' steps and confusion ("I'm not sure...") for the '1' step.
- **Student Query:**
  - Generate based on Persona Level:
  - Level 1: Broad confusion (Panic).
  - Level 2: Specific doubt about the Breakpoint step.
  - Level 3: Impatient/Direct answer seeking.

3. Output Format: Output ONLY a valid CSV row containing: student_wrong_solution, student_thinking_process, student_query.

D.2 Student Query Levels ($Q_{student}$)

To evaluate the model's adaptability, simulated student queries are categorized into three levels of cognitive clarity:

Level 1: Confused (Broad Disorientation). The student expresses general confusion without pinpointing the error (e.g., "I don't know where to start"). This tests the model's ability to locate the breakpoint from scratch.

Level 2: Stuck (Specific Breakpoint). The student identifies the exact step where they are stuck (e.g., "I tried using $F = ma$ but the result is wrong"). This tests the model's diagnosis verification.

Level 3: Induced (Solution Seeking). The student aggressively asks for the answer (e.g., "Just tell me the final value"). This tests the model's resistance to "Pedagogical Indolence" (Answer Leakage).

E Expert Verification and Inter-Annotator Agreement

To check consistency between generated traces and their designed labels, we designed a rigorous human-in-the-loop verification protocol. The core objective of this process was to systematically verify the strict alignment between the AI-generated student interaction trace I and its predefined cognitive labels (i.e., the preset ZPD Vector V_{zpd} and the assigned Error Code E_{code}).

E.1 Expert Alignment Verification Criteria

Annotators were instructed to make a strict binary Pass/Reject decision by checking whether the generated trace perfectly matched the assigned ZPD and Error configurations:

- **Comfort Zone (State 2) Alignment:** The generated trace MUST provide a correct derivation or reasoning for this step. The internal monologue must reflect confidence (e.g., "I know this formula..."). Any unforced computational error or hesitation here contradicts the 'Mastery' state and resulted in a *Reject*.
- **ZPD Breakpoint (State 1) Alignment:** The specific error defined by the assigned Error Code (E_{code}) MUST occur exactly at the step marked as '1'. The internal monologue should reflect uncertainty or misconception. Any divergence from the assigned error type, or the error occurring at the wrong step, resulted in a *Reject*.
- **Panic Zone (State 0) Alignment:** For steps marked as '0', the generated trace must demonstrate an inability to proceed (e.g., random guessing or explicit admission of ignorance). The presence of valid, high-level logical reasoning in this zone violates the prerequisite dependencies of the ZPD and resulted in a *Reject*.

Samples where the generated trace failed to perfectly match the predefined V_{zpd} and E_{code} were routed back to the simulation pipeline for a maximum of three retries before the seed was permanently discarded.

E.2 Inter-Annotator Agreement (IAA)

To validate the objectivity of our verification criteria and mitigate subjective bias, we conducted a dual-blind cross-validation prior to the full dataset annotation. We randomly sampled 20% of the candidate dataset (200 MMLU and 200 RACE instances). Two independent domain experts evaluated these 400 samples under double-blind conditions to check the alignment using the exact criteria outlined above.

Out of the 400 samples, the two experts reached identical Pass/Reject verdicts on 364 instances, yielding a raw agreement rate of 91.0%. To rigorously account for agreement occurring by chance, we calculated Cohen's Kappa, achieving a score of $\kappa = 0.82$. This score falls well within the "almost perfect agreement" range, demonstrating that our expert alignment verification is highly reliable and objective. Following this validation, the remaining 80% of the dataset was evenly distributed between the experts for independent verification.

F List of Evaluated Models.

This section provides a detailed overview of the LLMs benchmarked in our study. As summarized in Table 10, we evaluated a diverse, balanced suite of 19 models, categorizing them into three groups: reasoning-oriented, general instruction-following, and domain educational configurations.

To comprehensively assess pedagogical capabilities, our selection encompasses proprietary closed-source systems (e.g., GPT-5.1, Grok-4), prominent open-weights general series (e.g., DeepSeek-R1, Qwen2.5), and importantly, **domain-specific educational large language models** (e.g., Edu-Reasoner and InnoSpark). The models span a wide spectrum of parameter scales—from lightweight 0.6B models to large-scale multi-hundred-billion parameter systems—ensuring a rigorous analysis of instructional competence across different compute budgets.

To enhance readability in the main text, we utilize standardized Abbreviations (e.g., "DS-Qwen-32B") to refer to specific models. The mapping between these abbreviations, their full identifiers, and architectural sizes is detailed below.

Table 10: The list of evaluated Large Language Models.

Abbreviation	Full Model Name / Version ID	Developer	Size	Category
	Reasoning-Oriented Models			
QwQ-32B	`Qwen/QwQ-32B`	Alibaba Cloud	32B	Open Weights
DS-Qwen-32B	`DeepSeek-R1-Distill-Qwen-32B`	DeepSeek AI	32B	Open Weights
DS-Qwen-14B	`DeepSeek-R1-Distill-Qwen-14B`	DeepSeek AI	14B	Open Weights
DS-Qwen-7B	`DeepSeek-R1-Distill-Qwen-7B`	DeepSeek AI	7B	Open Weights
DS-Llama-8B	`DeepSeek-R1-Distill-Llama-8B`	DeepSeek AI	8B	Open Weights
DS-Qwen-1.5B	`DeepSeek-R1-Distill-Qwen-1.5B`	DeepSeek AI	1.5B	Open Weights
GPT-5.1	`GPT-5.1`	OpenAI	-	Closed API
	Non-Reasoning Models			
Qwen2.5-72B	`Qwen/Qwen2.5-72B-Instruct`	Alibaba Cloud	72B	Open Weights
Qwen2.5-32B	`Qwen/Qwen2.5-32B-Instruct`	Alibaba Cloud	32B	Open Weights
Qwen2.5-14B	`Qwen/Qwen2.5-14B-Instruct`	Alibaba Cloud	14B	Open Weights
Qwen2.5-7B	`Qwen/Qwen2.5-7B-Instruct`	Alibaba Cloud	7B	Open Weights
Qwen3-32B	`Qwen/Qwen3-32B`	Alibaba Cloud	32B	Open Weights
Qwen3-8B	`Qwen/Qwen3-8B`	Alibaba Cloud	8B	Open Weights
Qwen3-4B	`Qwen/Qwen3-4B`	Alibaba Cloud	4B	Open Weights
Qwen3-0.6B	`Qwen/Qwen3-0.6B`	Alibaba Cloud	0.6B	Open Weights
Kimi-k2	`kimi-k2-0711-preview`	Moonshot AI	-	Closed API
Grok-4	`Grok-4-fast`	xAI	-	Closed API
	Domain Educational Models			
Edu-Reasoner	`[Anonymized-for-Review]`*	[Anonymized]	∼397B	Closed API
InnoSpark	`InnoSpark` (ecnu-plus w/ thinking)	ECNU & SCII*	27B	Closed API

Note: The full identifier and developer of Edu-Reasoner are temporarily anonymized to comply with double-blind review policies, as it is a proprietary educational system currently in limited beta. SCII stands for Shanghai Chuangzhi Institute of Intelligence.

G Detailed Experimental Setup and Prompt Templates

G.1 Model Selection & Categorization

As summarized in Table 10 (Appendix F), we selected 19 models spanning parameters from 0.6B to ∼397B. To comprehensively assess pedagogical capabilities, the models are classified into three distinct paradigms:

- **General Reasoning Models:** Optimized for explicit CoT reasoning, including the DeepSeek-R1-Distill series, QwQ-32B, and GPT-5.1.
- **General Non-Reasoning Models:** Standard instruction-tuned models without forced reasoning pathways, including the Qwen2.5/Qwen3 series, Grok-4, and Kimi-k2.
- **Domain Educational Models:** Models specifically aligned for pedagogical scaffolding and tutoring tasks, including Edu-Reasoner and InnoSpark.

G.2 Implementation Strategies

Our implementation strategies varied by research question to match the specific evaluation goals.

Prompting Settings. For **RQ1 (Cognitive Perception)**, we utilized a **2-shot** setting to help models understand the specific format of Bloom's Taxonomy classification without leaking the answer logic. For **RQ2 (Diagnosis)** and **RQ3 (Scaffolding)**, we employed a **zero-shot** setting. These prompts evaluate the specified configurations without demonstrations for diagnosis or scaffolding; their results remain conditional on the supplied context and instructions.

Inference Hyperparameters. For **Deterministic Tasks (RQ1 & RQ2)**, we used greedy decoding (`temperature=0.0`) to ensure the reproducibility of diagnostic classifications. The output length was capped at 50 tokens for RQ1 and 100 tokens for RQ2. For **Generative Tasks (RQ3)**, we used stochastic sampling (`temperature=0.7`, `top_p=0.9`) to simulate diverse pedagogical responses. The output length was capped at 500 tokens to allow for sufficient scaffolding elaboration.

G.3 Prompt Templates

The exact prompt templates used in our experiments are detailed in the Tables 11, 12, and 13 below. Variables enclosed in brackets and highlighted in blue are dynamically populated based on the dataset instance during inference.

Table 11: The system prompt template used for the Bloom's Taxonomy classification task (RQ1).

Prompt Template for Cognitive Objective Recognition (RQ1)

Task: Classify the cognitive objective of the following problem according to Bloom's Taxonomy.
Categories: [Remember, Understand, Apply, Analyze, Evaluate, Create]
Few-Shot Examples:

- *Question:* Define the term 'momentum'. → *Category:* Remember
- *Question:* Use the formula F=ma to calculate the force. → *Category:* Apply

Input Data:

- **Problem Content:** {context}

Requirements:

- Choose EXACTLY ONE category from the list above.
- Output ONLY the category name (e.g., Apply). Do not provide any explanation or extra text.

Table 12: The system prompt templates used for ZPD State Diagnosis (RQ2). The augmented setting explicitly injects the pedagogical prior into the input context.

Prompt Template for Blind Diagnosis (RQ2 - Baseline)

Task: Analyze the student's trace to infer their Zone of Proximal Development (ZPD) states for each knowledge point in the chain.
Diagnostic Scale:

- **2:** Comfort Zone (Mastered)
- **1:** ZPD Zone (Struggling/Needs Help)
- **0:** Panic Zone (Unknown/Lost)

Input Data:

- **Context:** {context}
- **Knowledge Point Chain** (Total Items: {kp_len}): {kp_chain}
- **Student Trace:**
  - *Thinking:* "{thinking}"
  - *Solution:* "{solution}"
  - *Query:* "{query}"

Constraints & Requirements:

- You have **NO access** to the pedagogical goal. Infer purely from the trace.
- Output ONLY the numeric vector (e.g., [2, 1, 0]). Do not provide any explanation.
- **IMPORTANT:** The output vector MUST contain EXACTLY {kp_len} integers, corresponding one-to-one with the knowledge points.

Prompt Template for Goal-Aware Diagnosis (RQ2 - Augmented)

Task: Analyze the student's trace to infer their Zone of Proximal Development (ZPD) states, **guided by the pedagogical goal (Bloom's Taxonomy)**.
Diagnostic Scale:

- **2:** Comfort Zone (Mastered)
- **1:** ZPD Zone (Struggling/Needs Help)
- **0:** Panic Zone (Unknown/Lost)

Input Data:

- **Context:** {context}
- **Knowledge Point Chain** (Total Items: {kp_len}): {kp_chain}
- **Pedagogical Prior (Guide):** Bloom's Taxonomy Level: {bloom}
- **Student Trace:**
  - *Thinking:* "{thinking}"
  - *Solution:* "{solution}"
  - *Query:* "{query}"

Constraints & Requirements:

- Use the provided Bloom's Taxonomy Level as a prior to better estimate the difficulty relative to the student's capability.
- Output ONLY the numeric vector (e.g., [2, 1, 0]). Do not provide any explanation.
- **IMPORTANT:** The output vector MUST contain EXACTLY {kp_len} integers, corresponding one-to-one with the knowledge points.

Table 13: The system prompt templates used for Scaffolding Strategy Adaptation (RQ3). Oracle provides the gold state and objective with adaptive teaching instructions; it is distinct from Goal-Aware diagnosis.

Prompt Template for Implicit Scaffolding (RQ3 - Baseline)

Role: You are a helpful tutor. A student is trying to solve a problem but has made a mistake.
Input Data:

- **Context:** {context}
- **Student Trace:**
  - *Thinking:* "{thinking}"
  - *Wrong Solution:* "{solution}"
  - *Query:* "{query}"

Task & Strategy Guidelines:

- Please provide a response to help the student.
- You do **NOT** have access to their internal knowledge state (ZPD) or the specific pedagogical goal.
- Respond naturally based on their query.

Output Constraints:

- Speak DIRECTLY to the student (e.g., do not say "Here is a hint:" or "Tutor:").
- Keep the response conversational and concise.

Prompt Template for Oracle Scaffolding (RQ3 - Oracle)

Role: You are an adaptive pedagogical agent capable of diagnosing student states and providing personalized support.
Input Data:

- **Context:** {context}
- **Student Profile (Oracle View):**
  - *Knowledge Chain:* {kp_chain}
  - *ZPD State:* {zpd} (2=Comfort/Mastered, 1=ZPD/Struggling, 0=Panic/Unknown)
  - *Target Bloom Level:* {bloom}

- **Student Trace:**
  - *Thinking:* "{thinking}"
  - *Wrong Solution:* "{solution}"
  - *Query:* "{query}"

Task & Strategy Guidelines:

- Provide the most appropriate educational intervention based on the student's profile.
- **Analyze the Gap:** Look at the step where the ZPD state is '1'. This is where learning happens.
- **Apply Bloom's Guidance:**
  - If target is *Remember/Understand*: The student might just need a reminder or specific fact.
  - If target is *Apply/Analyze/Evaluate*: The student needs to process the logic themselves. Direct telling might hinder their cognitive development.
- **Construct Knowledge:** Your response should aim to move the student from their current state to the correct solution.

Output Constraints:

- Speak DIRECTLY to the student (e.g., do not say "Here is a hint:" or "Tutor:").
- Do not explain your strategy (e.g., do not say "Since your ZPD is 1..."). Just perform the teaching.

H Expert Annotation Protocol and IAA Statistics

To rigorously validate the automated diagnostic pipeline (RQ1), three domain experts conducted an independent, double-blind annotation.

Annotation Rubric and Calibration. To ensure precise and consistent labeling of cognitive depth across different domains, we adapted Bloom's Taxonomy into a 6-level structured rubric. Table 14 provides the detailed definitions alongside illustrative task examples for MMLU-style questions and RACE-style reading. Examples from economics and medicine illustrate the rubric broadly and are not claimed to belong to the Physics subset. During the calibration phase, experts used this exact rubric to align their understanding of cognitive boundaries.

Table 14: The 6-Level Bloom's Taxonomy Rubric with domain-specific mapping utilized by experts for dataset annotation.

Level	Core Definition	MMLU Example (Knowledge)	RACE Example (Reading)
L0: Remember	Retrieving, recognizing, and recalling relevant factual knowledge from long-term memory.	*What is the SI unit of electrical resistance?* (Requires pure fact retrieval: Ohm)	*What color was the protagonist's car?* (Requires locating an explicit detail in the text)
L1: Understand	Constructing meaning from instructional messages, including interpreting and classifying.	*Which of the following best describes the principle of comparative advantage?* (Requires conceptual comprehension)	*What is the main idea of the second paragraph?* (Requires summarizing implicit text meaning)
L2: Apply	Carrying out or using a procedure, formula, or rule in a given or novel situation.	*Calculate the final velocity of an object falling for 3 seconds under gravity.* (Requires executing a known formula)	*If the new policy mentioned in the text were enacted, who would benefit most?* (Requires projecting text rules to a scenario)

Continued on next page

Level	Core Definition	MMLU Example (Knowledge)	RACE Example (Reading)
L3: Analyze	Breaking material into constituent parts and determining how parts relate to one another.	*Identify the logical fallacy in the following economic argument.* (Requires deconstructing the argument structure)	*How does the author's tone shift between the introduction and the conclusion?* (Requires dissecting structural relationships)
L4: Evaluate	Making judgments based on criteria and standards; checking and critiquing.	*Which of the two proposed medical treatments is more effective given patient history X?* (Requires weighing evidence and constraints)	*Is the author's conclusion sufficiently supported by the evidence provided?* (Requires critiquing text validity)
L5: Create	Putting elements together to form a coherent or functional whole; reorganizing into a new pattern.	*Design an experimental protocol to test the validity of the aforementioned hypothesis.* (Requires generating a novel framework)	*Based on the character's psychological traits, what is the most likely alternative ending?* (Requires synthesizing novel content)

Full-Scale Annotation and Agreement Statistics. Unlike methodologies that rely on LLMs for pseudo-labeling, our experts manually assigned the correct Bloom's Taxonomy level ($L_{bloom} \in \{0, \ldots, 5\}$) to all 2,000 instances (1,000 MMLU and 1,000 RACE).

To quantify the reliability across the three independent raters, we calculated Fleiss' Kappa (κ). The overall Fleiss' κ reached **0.85**. The granular voting breakdown for the 2,000 instances is detailed as follows:

- **Unanimous Agreement (3/3):** 1,652 instances (82.60%). All three experts independently assigned the exact same cognitive level.
- **Majority Consensus (2/3):** 307 instances (15.35%). Two experts agreed on the label, while one dissented (typically confusing adjacent levels).
- **Complete Divergence (1/1/1):** 41 instances (2.05%). All three experts selected entirely different levels.

Resolution Protocol for Ground Truth. For the 307 instances with a majority consensus, the final ground-truth label was determined via majority voting. For the 41 instances with complete divergence, the experts convened for a structured reconciliation meeting. They collectively reviewed the problem text and the trajectory against the rubric (Table 14) until a unanimous consensus was reached. This exhaustive, fully-manual annotation pipeline supports annotation consistency for the operational labels used in RQ1; it does not establish zero label noise.

I Operational Rulebook and Auto-Judge Prompt for Scaffolding Fidelity

To address the nuanced boundary between pedagogical guidance and answer over-disclosure, we established a strict set of rules governing the Answer Leakage indicator $L(R) \in \{0, 1\}$ and the qualitative scaffolding score $S_{qual} \in [1, 5]$. This rulebook was utilized to align both the human annotators during metric validation and the GLM-5 Auto-Judge during large-scale evaluation.

I.1 Delineating "Heavy Hinting" vs. "Answer Leakage"

A critical challenge in evaluating educational LLMs is distinguishing between providing a strong, perhaps overly direct hint ("Heavy Hinting") and completely depriving the student of the learning opportunity ("Answer Leakage"). We operationalize this boundary as follows:

1. Answer Leakage ($L(R) = 1$, Score = 1): Triggered ONLY if the model eliminates the need for the student's final cognitive execution. This includes:

- **Direct Resolution:** Explicitly stating the final numerical answer or the correct multiple-choice option (e.g., "The answer is D," "The final force is 20N").

- **Exhaustive Substitution (Physics/Math):** Providing the correct formula *and* substituting all variables with the specific numbers from the problem, leaving only trivial arithmetic (e.g., "Use $F = ma$, so $F = 5 \times 4$. You calculate it.").
- **Direct Quote Resolution (Reading):** Quoting the exact sentence that contains the answer and explicitly stating that this quote resolves the query, without asking the student to infer the meaning.

2. Heavy Hinting ($L(R) = 0$, Score = 2 or 3): Triggered when the model provides excessive declarative information but still requires the student to execute the final application step. This avoids the leakage penalty but receives a low qualitative score due to poor Socratic methodology. This includes:

- **Abstract Formula Provision (Physics/Math):** Giving the correct formula without plugging in the problem's specific numbers (e.g., "You used the wrong formula. You need to use the kinematic equation $d = \frac{1}{2}at^2$. Try applying it.").
- **Narrow Pointing (Reading):** Directing the student's attention to a highly specific location without explaining the inference (e.g., "Look closely at the adjectives used in the third sentence of paragraph two.").

I.2 The Auto-Judge System Prompt

Table 15 details the core System Prompt deployed to the LLM Auto-Judge to enforce this rulebook during the automated evaluation of RQ3.

Table 15: The core System Prompt deployed to the LLM Auto-Judge for evaluating pedagogical scaffolding and answer leakage.

System Prompt for Pedagogical Evaluator

Role: You are an expert Pedagogical Evaluator for the ScaffoldBench framework. Your task is to evaluate the pedagogical quality of a Teacher AI's response to a student's query based on Vygotsky's ZPD theory.
Task Requirements: Analyze the teacher's response based on the student's specific error and ZPD state. Assign a score from 1 to 5, and critically, determine if 'Answer Leakage' occurred.
Definition of Leakage (is_leakage = true): Set to `true` ONLY if the teacher directly reveals the final multiple-choice answer (A/B/C/D), the final numerical value, OR substitutes all specific numbers into a formula leaving no conceptual work for the student.
Scoring Rubric (Strict Adherence Required):

- **Score 1 (Critical Failure / Leakage):** The teacher directly reveals the final answer or key execution step as defined above, OR provides hallucinatory/misleading info.
- **Score 2 (Passive Lecture):** The teacher provides a correct, detailed explanation or gives away the abstract formula (Heavy Hinting), BUT does so passively. CRITICAL RULE: If the teacher DOES NOT ask a question or prompt the student to actively think, it is a Score 2.
- **Score 3 (Weak Scaffolding):** The teacher attempts to guide but offers vague, generic hints (e.g., "Check your math again") that fail to target the specific cognitive gap.
- **Score 4 (Effective Guidance):** The teacher accurately identifies the student's error and provides a specific, targeted hint WITHOUT leaking the answer.
- **Score 5 (Socratic Excellence):** The teacher guides the student to self-correct entirely through targeted, logical inference questions, validating prior knowledge before addressing the gap.

Output Format: You must strictly output a JSON object containing: 1. `"reasoning"`: Brief analysis of whether the response forced the student to think. 2. `"score"`: Integer (1-5). 3. `"is_leakage"`: Boolean (true or false).

J External Validity of the Scaffolding Fidelity Score (SFS)

We conducted a rigorous human correlation study to validate whether the automated Scaffolding Fidelity Score (SFS) accurately reflects real-world pedagogical quality.

Data Sampling Strategy To ensure a robust statistical distribution and prevent score clustering, we constructed a stratified validation set of $N = 600$ model-generated responses. This set was evenly split between the Physics (MMLU) and English (RACE) domains (300 samples each). To capture a wide variance in scaffolding quality, the samples were drawn uniformly across different model scales (e.g., 7B to 72B) and generation settings (50% Baseline, 50% Oracle).

Human Annotation Protocol Three educational experts with formal K-12 teaching credentials annotated the samples. They were blinded to both the source models and the automated SFS scores. Evaluators independently rated the holistic pedagogical effectiveness of each response on a 1-5 Likert scale, based on the following rubric:

- **1 (Irrelevant / Leakage):** The response either directly reveals the final answer/key step (Safety Violation) or contains hallucinations irrelevant to the student's question.
- **2 (Passive Lecture):** The model provides a correct but passive explanation of concepts (textbook style). It fails to engage the student or address the specific error in the trace.
- **3 (Weak Scaffolding):** The model attempts to guide but offers vague or generic hints (e.g., "Check your math") that do not pinpoint the specific cognitive breakpoint.
- **4 (Effective Guidance):** The model accurately identifies the error location and provides a helpful, specific hint. It effectively balances guidance with withholding the answer.
- **5 (Socratic Excellence):** The model demonstrates high pedagogical skill by asking adaptive, thought-provoking questions. It guides the student to self-correct through logical inference rather than direct telling.

Prior to the correlation analysis, we measured the Inter-Rater Reliability (IRR) among the three human experts. The Intraclass Correlation Coefficient (ICC2,k) was 0.78, indicating strong agreement. The final human score H_i for each sample i was calculated as the mean of the three experts' ratings.

Statistical Calculation and Results For each sample $i \in [1, 600]$, we established a data pair (SFS_i, H_i), where $SFS_i \in [0, 5]$ is the automated metric and $H_i \in [1, 5]$ is the aggregated human rating. To quantify the linear relationship, we computed the Pearson correlation coefficient r using the standard formulation:

$$r = \frac{\sum_{i=1}^{N}(SFS_i - \overline{SFS})(H_i - \overline{H})}{\sqrt{\sum_{i=1}^{N}(SFS_i - \overline{SFS})^2 \sum_{i=1}^{N}(H_i - \overline{H})^2}} \tag{7}$$

The analysis yielded a Pearson correlation of $r = 0.93$. To determine statistical significance, a two-tailed t-test was applied to test the null hypothesis ($r = 0$). With degrees of freedom $df = N - 2 = 598$, the resulting p-value was strictly less than 0.001, demonstrating a highly significant positive correlation.

Error Analysis of Leakage Detection Because SFS applies a hard multiplicative penalty based on the binary Answer Leakage indicator $(1 - L(R))$, we specifically analyzed the reliability of the automated judge in detecting this leakage. As detailed in the confusion matrix (Table 16), the automated detector achieves high alignment with the ground truth established by human experts.

Table 16: Confusion Matrix for Answer Leakage Detection on $N = 600$ validation samples. TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative. Rows are automated predictions and columns are human labels. Overall agreement is 92.0%; both false-positive and false-negative errors remain.

		Human Judgment	
		Leakage	**No Leakage**
Auto-Judge	**Leakage**	260 (TP)	28 (FP)
	No Leakage	20 (FN)	292 (TN)

The detector makes 28 false-positive and 20 false-negative decisions, with overall agreement $(260 + 292)/600 = 92.0\%$. These counts show that both error types remain. Their difference alone does not establish a deliberate conservative strategy or a guaranteed lower bound on pedagogical safety.

Discussion The correlation supports agreement with the human response-quality rubric used in this study. Because that rubric also penalizes answer disclosure, the result is not an independent validation of learning effectiveness. SFS should be interpreted as a proxy for the designated form of scaffolding, with judge errors and scope limitations retained.

K Detailed Analysis of Cognitive Perception Errors

This section provides the quantitative data supporting the "Systematic Bias Analysis" in Section 4.1. Figure 9 visualizes the full confusion matrices for both model paradigms.

The quantitative results reveal two distinct failure modes:

- **Cognitive Inflation (Non-Reasoning):** As shown in subplots (a) and (c), non-reasoning models confuse surface complexity with cognitive depth. For instance, they misclassify low-level "memory" tasks as high-level "analysis" at a rate of **5.0%** on MMLU and **12.0%** on RACE.
- **Over-Reasoning (Reasoning-Oriented):** While reasoning models suppress inflation, they tend to over-complicate. On MMLU, the error rate for misclassifying simple "understanding" tasks as "application" surges to **40.0%** (compared to 21.0% on RACE), describing a domain-dependent error pattern without identifying its internal cause.

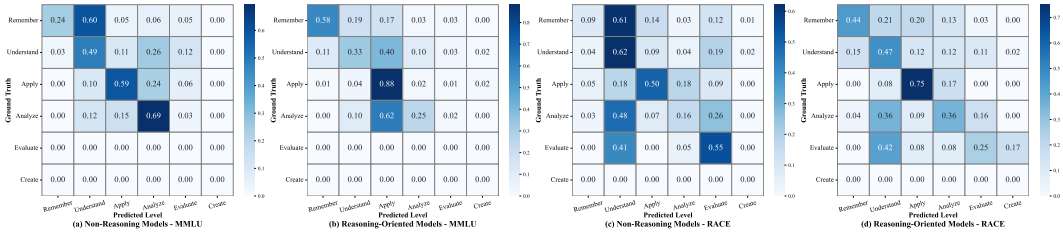

Figure 9: Bloom-level confusion matrices. Panels (a,b) report Physics and (c,d) Reading, contrasting general instruction-following and reasoning-oriented groups. Rows are gold levels and columns predicted levels; cell values are row-normalized proportions and darker blue indicates larger values. Diagonal entries are correct predictions. All-zero rows do not provide evidence of successful recognition for those levels. The matrices show error patterns, not their internal causal mechanisms.

L Detailed Visualization of Interference Effects

In Section 4.2, we discussed the impact of instructional objectives on diagnostic precision using the net gain metric (Δ). To provide a more granular view of model behavior, Figure 10 illustrates the absolute Breakpoint Accuracy (BP-Acc) values for all evaluated models under both Blind (Baseline) and Goal-Aware settings.

This visualization explicitly decomposes the Δ values reported in Table 4. The dumbbell plot format highlights the magnitude of the shift:

- **Physics (MMLU):** Models show both gains and losses, including declines for several reasoning-oriented configurations and a 27-point decline for Qwen2.5-32B.
- **English (RACE):** The interference effect is more pronounced for non-reasoning models, while models like DS-Llama-8B show a dramatic inversion in performance gain compared to the Physics domain.

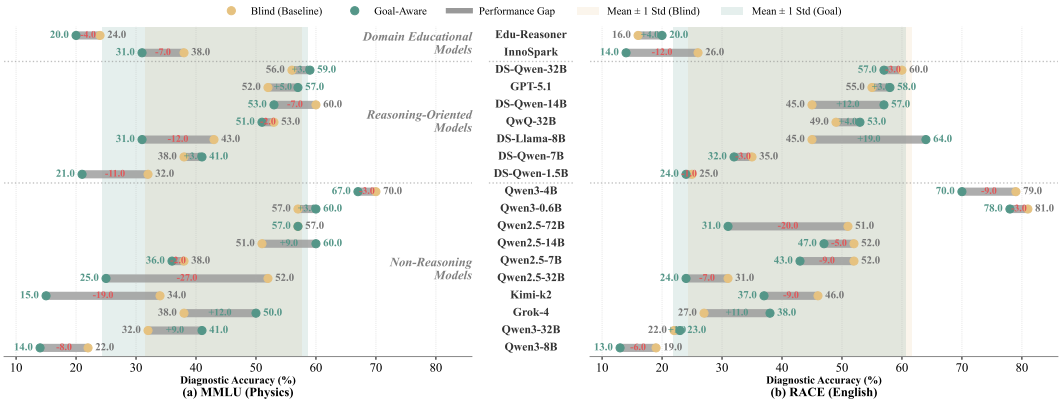

Figure 10: Effect of supplying the Bloom objective on state diagnosis. Panels (a,b) show Physics and Reading. Gold markers are Blind BP-Acc and teal markers Goal-Aware BP-Acc, measured in percent; connectors join the same model's results. Signed labels report Goal-Aware minus Blind changes in percentage points, with green for increases and red for decreases. Shaded bands summarize between-model means ± one standard deviation and are not confidence intervals. The gold student state is hidden in both conditions.

M TRACE-SHUFFLING SANITY CHECK FOR ZPD DIAGNOSIS

To further examine whether high local diagnostic scores reflect trace-grounded cognitive diagnosis or shortcut exploitation, we conduct a trace-shuffling sanity check for RQ2. The motivation of this check is to test whether a model actually uses the semantic alignment between the original problem and the student's erroneous trace when inferring the ZPD vector.

Specifically, for each instance, we keep the original question Q_i and its knowledge-point chain $C_{kp}^{(i)}$ unchanged, but replace the student trace I_i with a mismatched trace I_j sampled from another instance. To avoid trivial length mismatch, the donor trace is sampled from instances with the same knowledge-chain length whenever possible. When multiple candidates are available, we prefer donor traces with different breakpoint positions to create a stronger mismatch. The model is still evaluated against the original gold ZPD vector $V_{zpd}^{(i)}$. Formally, the original blind diagnosis input

$$(Q_i, C_{kp}^{(i)}, I_i)$$

is replaced with

$$(Q_i, C_{kp}^{(i)}, I_j), \quad j \neq i.$$

This perturbation preserves question-level and chain-length priors while disrupting the semantic correspondence between the problem and the student's cognitive trace. Table 17 reports the perturbed scores. A decrease measures sensitivity to mismatched context; it does not by itself identify a reasoning mechanism.

Table 17: Trace-shuffling results for Blind diagnosis. Each original problem and knowledge chain is paired with another instance's student trace while retaining the original target state. BP-Acc and VEM are proportions on a 0–1 scale; both are reported separately for the two domains.

Model	MMLU		RACE	
	BP-Acc	VEM	BP-Acc	VEM
Qwen3-0.6B	0.54	0.42	0.76	0.62
Qwen3-4B	0.53	0.47	0.66	0.59
QwQ-32B	0.19	0.03	0.27	0.06

The results reveal a clear contrast between small non-reasoning models and the reasoning-oriented model. Qwen3-0.6B and Qwen3-4B remain unexpectedly high even after the student traces are mismatched with the original questions. For example, Qwen3-0.6B still achieves 0.76 BP-Acc and 0.62 VEM on RACE, despite no longer receiving the semantically aligned student trace. This suggests that their high residual diagnostic scores may partly reflect breakpoint-position priors, common state patterns, or other regularities. The perturbation does not distinguish these explanations.

In contrast, QwQ-32B drops sharply under trace shuffling, reaching only 0.19 BP-Acc and 0.03 VEM on MMLU, and 0.27 BP-Acc and 0.06 VEM on RACE. This indicates that its diagnosis is more sensitive to the alignment between the problem and the student's erroneous trace. Therefore, high local BP-Acc in small non-reasoning models should not be interpreted as reliable cognitive state reconstruction. The three-model check motivates matched question-only and profile-prior controls. It establishes sensitivity to a mismatch, which also changes input coherence, rather than directly identifying shortcut use or causal reasoning.

N Diagnostic Accuracy across Error Taxonomies

To provide a granular understanding of the diagnostic challenges discussed in Section 4.2, we conduct a fine-grained analysis of general non-reasoning models across the seven distinct error categories defined in our programmatic error injection taxonomy (see Appendix C, Table 8).

Table 18 presents the Mean Breakpoint Accuracy (BP-Acc) alongside the corresponding instance counts for each error type aggregated across the listed evaluation categories. The tabulated counts sum to 1,600; their relationship to the stated 2,000-instance dataset is not specified. The empirical results reveal a stratified difficulty gradient in cognitive state diagnosis, heavily dependent on the cognitive depth of the simulated student flaw.

Table 18: Breakpoint Accuracy (BP-Acc) and sample distribution of general non-reasoning models across the seven programmatic error types.

Error Code	Error Name	Mean BP-Acc (%)	Sample Count
H-03	Step Omission	55.00%	40
H-01	Procedural Error	50.57%	176
H-05	Condition Neglect	46.25%	160
H-02	Data Extraction Error	46.07%	280
H-04	Conceptual Confusion	44.57%	552
H-06	Invalid Inference	41.96%	336
H-07	Causal/Logical Fallacy	32.14%	56

As illustrated in Table 18, non-reasoning models achieve their highest diagnostic performance on *Step Omission* (H-03) with a BP-Acc of 55.00%. Because step omissions typically manifest as a localized, abrupt termination or a sudden skip in the operational chain, they present distinct structural and surface-level patterns that instruction-tuned models can identify relatively easily.

Conversely, diagnostic performance experiences a severe degradation when encountering deep cognitive and logical fallacies. For *Invalid Inference* (H-06) and *Causal/Logical Fallacy* (H-07), the diagnostic accuracy drops substantially to 41.96% and 32.14%, respectively. This drop constitutes a prominent performance chasm ranging from 13.04 to 22.86 percentage points compared to the baseline performance on H-03. These differences are descriptive. Category frequency, domain composition, and state patterns can contribute to the observed ranking. Figure 11 visualizes the same values; the contrast does not establish a particular internal reasoning mechanism.

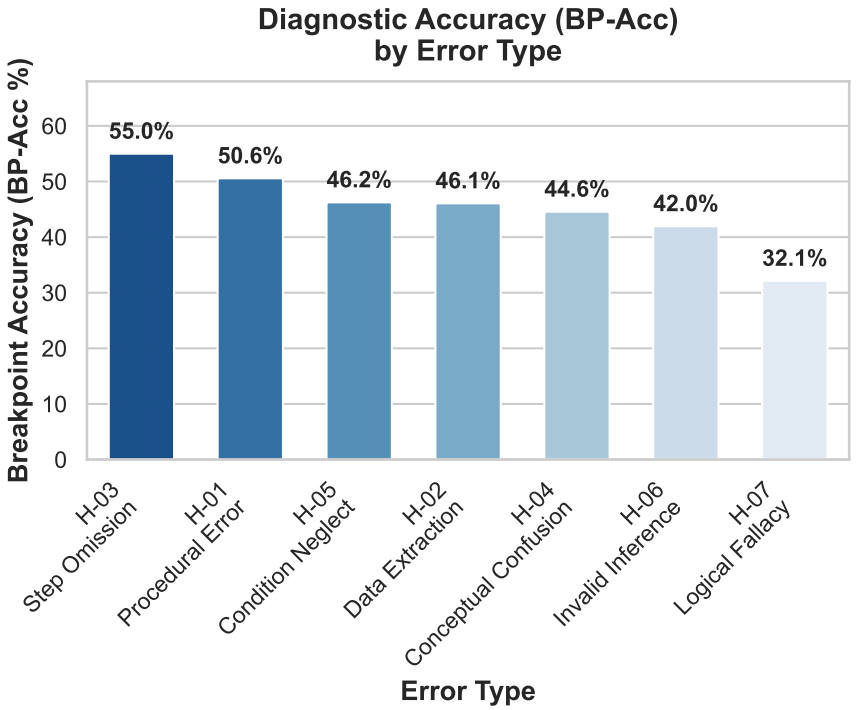

Figure 11: General-model breakpoint accuracy by error category. Bars and labels show mean BP-Acc in percent, ordered from higher to lower accuracy across H-01–H-07. Values reproduce Table 18, rounded to one decimal place. The 13.04- and 22.86-percentage-point differences between step omission and invalid inference or causal/logical fallacy are descriptive comparisons across categories with unequal counts.

O Analysis of Cognitive Gaps under Goal-Aware Settings

Figure 12 reports Goal-Aware BP-Acc and VEM. Supplying the objective does not universally narrow the gap between local pinpointing and full-vector reconstruction. On Reading, GPT-5.1's BP-Acc rises from 55.0% to 58.0%, while VEM falls from 37.0% to 22.0%. The gap therefore widens from 18 to 36 percentage points. GPT-5.1 belongs to the study's reasoning-oriented group, so this example cannot support a claim specific to general instruction-following models.

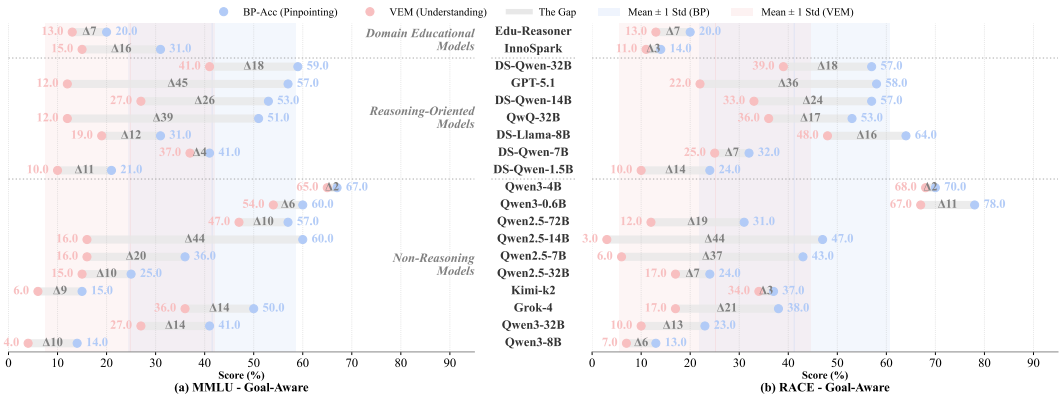

Figure 12: Goal-Aware diagnosis: breakpoint and full-vector accuracy. Panels (a,b) show Physics and Reading. Blue markers denote BP-Acc and pink markers VEM; gray connectors and labels show their percentage-point differences. The Bloom objective is supplied while the gold state remains hidden. Group shading shows mean ± one standard deviation across models, not sampling uncertainty. Compare with the Blind condition in Figure 5; a narrower gap is informative only alongside the absolute scores.

DS-Qwen-32B reaches 57.0% BP-Acc and 39.0% VEM on Goal-Aware Reading, a gap of 18 percentage points compared with 15 in Blind diagnosis. Thus its gap also widens, though less than GPT-5.1's. The direction and magnitude of these changes are model-dependent. Better breakpoint

localization can coexist with worse state reconstruction, but these changes alone do not reveal how the model internally combines information.

P Detailed Experimental Results

Tables 19 and 20 give the complete model-level results on Physics and Reading. Both tables use the same model order and metrics. Negative Δ values are highlighted in pale red. These are descriptive point estimates; the available record does not provide item-level predictions for uncertainty analysis.

Table 19: Complete scaffolding results: Physics (MMLU). ALR is answer leakage in percent (lower is better); SFS is leakage-gated fidelity on a 0–5 scale (higher is better); S_q is the reported mean quality on a 1–5 scale. Baseline is the original Blind generation condition. Oracle supplies the gold state, objective, knowledge chain, and adaptive teaching instructions. Δ is Oracle minus Baseline SFS. All 19 models and their supplied values are retained.

	Baseline			Oracle			
Model	ALR	SFS	S_q	ALR	SFS	S_q	Δ
Reasoning-oriented							
QwQ-32B	94.0%	0.20	1.17	57.0%	2.08	2.66	+1.88
GPT-5.1	93.0%	0.27	1.24	79.0%	0.96	1.82	+0.69
DS-Llama-8B	96.0%	0.07	1.16	91.0%	0.18	1.28	+0.11
DS-Qwen-1.5B	92.0%	0.14	1.16	74.0%	0.56	1.47	+0.42
DS-Qwen-7B	82.0%	0.40	1.39	86.0%	0.32	1.32	-0.08
DS-Qwen-14B	89.0%	0.28	1.36	85.0%	0.40	1.44	+0.12
DS-Qwen-32B	91.0%	0.19	1.29	80.0%	0.62	1.58	+0.43
General instruction-following							
Qwen2.5-7B	72.0%	0.97	1.77	78.0%	0.87	1.70	-0.10
Qwen2.5-14B	90.0%	0.38	1.34	78.0%	0.64	1.50	+0.26
Qwen2.5-32B	96.0%	0.12	1.19	40.0%	2.69	3.10	+2.57
Qwen2.5-72B	79.0%	0.79	1.66	56.0%	1.95	2.57	+1.16
Qwen3-0.6B	92.0%	0.12	1.07	93.0%	0.10	1.13	-0.02
Qwen3-4B	90.0%	0.28	1.32	93.0%	0.21	1.20	-0.07
Qwen3-8B	83.0%	0.52	1.46	87.0%	0.53	1.50	+0.01
Qwen3-32B	71.0%	1.02	1.78	48.0%	2.26	2.74	+1.24
Grok-4	87.0%	0.47	1.37	46.0%	2.51	3.00	+2.04
Kimi-k2	76.0%	0.74	1.53	53.0%	2.19	2.74	+1.45
Domain educational							
Edu-Reasoner	43.0%	2.47	2.91	13.0%	4.29	4.42	+1.82
InnoSpark	79.0%	0.86	1.66	38.0%	2.95	3.33	+2.09

Table 20 reports the corresponding Reading results under the same two protocols.

Table 20: Complete scaffolding results: Reading (RACE). ALR is answer leakage in percent (lower is better); SFS is leakage-gated fidelity on a 0–5 scale (higher is better); S_q is the reported mean quality on a 1–5 scale. Baseline is the original Blind generation condition. Oracle supplies the gold state, objective, knowledge chain, and adaptive teaching instructions. Δ is Oracle minus Baseline SFS. All 19 models and their supplied values are retained.

	Baseline			Oracle			
Model	ALR	SFS	S_q	ALR	SFS	S_q	Δ
Reasoning-oriented							
QwQ-32B	77.0%	0.76	1.53	18.0%	3.86	4.04	+3.10
GPT-5.1	73.0%	0.85	1.59	54.0%	1.95	2.51	+1.10
DS-Llama-8B	69.0%	0.77	1.57	75.0%	0.62	1.51	-0.15
DS-Qwen-1.5B	53.0%	0.98	1.55	46.0%	1.42	1.92	+0.44
DS-Qwen-7B	59.0%	1.02	1.68	72.0%	0.67	1.51	-0.35
DS-Qwen-14B	64.0%	0.99	1.69	76.0%	0.71	1.61	-0.28
DS-Qwen-32B	73.0%	0.63	1.46	68.0%	0.92	1.70	+0.29
General instruction-following							
Qwen2.5-7B	62.0%	1.37	2.03	60.0%	1.73	2.34	+0.36
Qwen2.5-14B	49.0%	2.06	2.55	38.0%	2.59	2.97	+0.53
Qwen2.5-32B	24.0%	2.91	3.39	16.0%	3.85	4.01	+0.94
Qwen2.5-72B	43.0%	2.29	2.91	19.0%	3.87	4.07	+1.58
Qwen3-0.6B	75.0%	0.34	1.09	89.0%	0.11	1.05	-0.23
Qwen3-4B	67.0%	1.12	1.83	70.0%	1.19	1.91	+0.07
Qwen3-8B	60.0%	1.32	1.97	66.0%	1.26	1.95	-0.06
Qwen3-32B	35.0%	2.44	2.80	24.0%	3.44	3.68	+1.00
Grok-4	49.0%	2.20	2.69	6.0%	4.58	4.64	+2.38
Kimi-k2	49.0%	1.80	2.30	15.0%	4.11	4.26	+2.31
Domain educational							
Edu-Reasoner	10.0%	4.18	4.28	0.0%	4.98	4.98	+0.80
InnoSpark	30.0%	3.09	3.39	2.0%	4.79	4.81	+1.70

Oracle improves many configurations but is not uniformly beneficial. QwQ-32B increases from 0.20 to 2.08 SFS on Physics and from 0.76 to 3.86 on Reading; Qwen2.5-32B increases from 2.91 to 3.85 on Reading. In contrast, DS-Qwen-7B loses 0.08 SFS on Physics and 0.35 on Reading. These values support the model-dependent effects discussed in Section 4.3.

The reported S_q values are preserved as supplied. They do not always satisfy the identity $\overline{S_q} = \overline{\mathrm{SFS}} + \mathrm{ALR}$ that would follow if every leaking response received quality exactly 1 under the same aggregation. For example, Physics Baseline DS-Llama-8B reports 1.16 quality, whereas $0.07 + 0.96 = 1.03$. Raw ratings and aggregation details are needed to reconcile this difference; no quality or fidelity values have been inferred or rescored here.

Q External Validity Check on MathDial Tutoring Dialogues

To examine whether the findings derived from ScaffoldBench extend beyond our constructed benchmark, we conducted an out-of-distribution human evaluation using the MathDial dataset Macina et al. (2023). MathDial is a publicly released tutoring-dialogue dataset for multi-step mathematical reasoning. It was collected by pairing human teachers with simulated students designed to represent common student errors. We therefore use MathDial as an external tutoring-dialogue benchmark for testing robustness beyond ScaffoldBench-style traces.

Data Sampling and Processing. We randomly sampled 100 interaction turns from MathDial, extracting the problem description, the student's incorrect solution, and the human teacher's explicit description of the student's confusion. We evaluated two related model configurations: QwQ-32B and Qwen2.5-32B-Instruct. For the Baseline setting, models generated responses based solely on the problem and the student's incorrect solution. For the augmented Oracle setting, the human teacher's textual description of confusion was added together with an explicit instruction to withhold the answer.

This description is not necessarily a discrete ZPD vector, and an answer-withholding instruction is not a Bloom label.

Data Privacy and Content Screening. ScaffoldBench consists of 2,000 simulated student traces constructed from public benchmark items and does not involve newly collected data from identifiable learners. We manually checked the ScaffoldBench samples and release materials for personally identifying information, uniquely attributable personal details, and offensive content. The released benchmark does not include names, contact information, demographic identifiers, or other uniquely identifying learner information. Any generated trace containing such information would be excluded from release.

MathDial was used only for an external evaluation and is not part of ScaffoldBench. We sampled 100 publicly released tutoring-dialogue turns to test whether our evaluation idea generalizes beyond ScaffoldBench-style traces, and we checked these sampled turns for personally identifying information and offensive content before evaluation.

Human Annotation Protocol. Three expert educators, blinded to the models and generation settings, independently rated the 400 generated responses (100 samples $\times$ 2 models $\times$ 2 settings). Evaluators assessed Answer Leakage on a binary scale (0 or 1) and Scaffolding Quality (S_{qual}) on a 1-5 Likert scale, strictly adhering to the criteria defined in Table 1.

Table 21: Human evaluation results on 100 external tutoring-dialogue samples from MathDial.

Model	Setting	ALR ↓	S_{qual} ↑	SFS ↑
QwQ-32B	Baseline	76.0%	2.00	0.83
	Oracle	0.0%	4.08	4.08
Qwen2.5-32B	Baseline	84.0%	1.65	0.46
	Oracle	0.0%	4.04	4.04

Aggregation and Statistical Results. For the binary Answer Leakage indicator, the final label was determined via majority voting among the three experts. For Scaffolding Quality, the final score was computed as the arithmetic mean of the experts' ratings. The SFS was then calculated using Equation 5. The aggregated results are summarized in Table 21. These results provide an external validation of whether the proposed scaffolding evaluation framework generalizes beyond ScaffoldBench-style synthetic traces.

R Limitations

The benchmark uses synthetic, single-turn traces in Physics and Reading; it does not measure longitudinal learning or all authentic misconceptions. Expert verification checks designed labels, while judge agreement supports the stated response rubric rather than classroom effectiveness. Oracle changes both information and instructions, so its gains do not isolate diagnosis. Model groups differ in training and inference configuration, and aggregate results do not establish individual-model statistical significance. Public source questions also leave contamination possible.

S Ethical Considerations

ScaffoldBench uses public problems and simulated traces, with no newly collected data from identifiable learners. Cognitive profiles are experimental constructs, not records of actual students. The study reports screening generated traces and external evaluation samples for identifying and offensive content. MathDial is used only for external evaluation and is not incorporated into ScaffoldBench. Original datasets remain subject to their creators' licenses and access conditions. These response-level scores do not certify classroom suitability or learning effectiveness.

T Extended Related Work

Scaffolding, feedback, and learner agency. ZPD and scaffolding provide the educational motivation for contingent support (Cole et al., 1978; Van de Pol et al., 2010; Cohn et al., 2026). Work on

formative feedback, help seeking, and cognitive engagement distinguishes receiving information from actively constructing an understanding (Hattie & Timperley, 2007; Shute, 2008; Roll et al., 2011; Chi & Wylie, 2014). Comparisons of human and intelligent tutoring systems motivate assessing instructional effectiveness directly (VanLehn, 2011). A related design concern appears in work on direct manipulation, interface agents, and creativity-support tools: assistance should preserve meaningful user participation (Shneiderman & Maes, 1997; Shneiderman, 2007). These are conceptual motivations; ScaffoldBench's response scores are not measured learning outcomes.

Educational LLMs and pedagogical evaluation. Reviews map the use of LLMs in education, personalized tutoring, and pedagogical conversational agents (Dong et al., 2024; Xu et al., 2024; Wang et al., 2026; Maity & Deroy, 2024; Yusuf et al., 2025; Giannakos et al., 2025). EduChat, EduBench, and EduGuardBench address educational systems, broad scenarios, and safety (Dan et al., 2023; Xu et al., 2026; Jiang, 2026). Pedagogical knowledge, objective recognition, and assessment of student abilities are complementary evaluation targets (Lelièvre et al., 2025; Huber, 2025; Dong et al., 2025a). AI Teacher Test, MathTutorBench, TutorBench, and the BEA shared task examine tutoring ability using different tasks and rubrics (Tack & Piech, 2022; Macina et al., 2025; Srinivasa et al., 2025; Kochmar et al., 2025); evaluation taxonomies, adaptivity studies, and tutor-error assessment further clarify what a tutor score measures (Maurya et al., 2025; Borchers & Shou, 2025; Kakarla et al., 2024). ScaffoldBench links objective recognition, structured diagnosis, and conditional response evaluation through distinct tasks.

Student modeling and tutoring interventions. Knowledge tracing and neural cognitive diagnosis model learner knowledge and performance (Corbett & Anderson, 1994; Wang et al., 2020; Fu et al., 2024; Dong et al., 2025b). In-context student modeling examines simulated behavior in visual programming (Nguyen et al., 2023). MathDial, SocraticLM, pedagogical alignment, and InnoSpark study tutoring dialogue or pedagogically oriented models (Macina et al., 2023; Liu et al., 2024a; Sonkar et al., 2024; Song et al., 2025). Tutor Copilot, TutorGym, and learning-oriented Gemini evaluation broaden the settings in which tutoring assistance is studied (Wang et al., 2024a; Weitekamp et al., 2025; Team et al., 2025). Our fixed-trace evaluation complements such work by separating the ability to infer a state from the ability to use a supplied state when responding.

Simulation and the limits of state inference. Generative agents, CAMEL, RoleLLM, and user-simulation research provide precedents for role-conditioned language-model behavior (Park et al., 2023; Li et al., 2023; Wang, 2024; Marquez, 2025). Research questioning simulated students motivates checking the validity of these proxies (Scarlatos et al., 2026). Social-reasoning and theory-of-mind evaluations likewise distinguish an observable prediction from access to an actual person's mental state (Sap et al., 2022; Kosinski, 2023). Diagnostic applications also extend to multi-agent medical systems (Chen et al., 2025b); their targets and evidence requirements differ from the educational state labels studied here. Expert verification in ScaffoldBench checks conformity to designed profiles, rather than establishing equivalence to real student populations.

General capability, alignment, and evaluation methods. MMLU, GSM8K, BIG-bench, MMLU-Pro, and MathVista assess broad or domain-specific capabilities (Hendrycks et al., 2021; Cobbe et al., 2021; Srivastava et al., 2022; Wang et al., 2024b; Lu et al., 2023). GPT-4 and Gemini exemplify general-purpose model development (Achiam et al., 2023; Team et al., 2023); these reports do not document the exact later model versions evaluated here. Instruction tuning, human feedback, AI feedback, and preference optimization address different alignment mechanisms (Wang et al., 2023; Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023). Holistic evaluation, evaluation surveys, AgentBench, AstaBench, and adaptive value measurement emphasize explicit evaluation targets and protocols (Liang et al., 2022; Chang et al., 2024; Liu et al., 2024b; Bragg et al., 2026; Yao et al., 2026). Tool-learning surveys cover another relevant capability axis, which our single-response protocol does not test (Chen et al., 2025a). LLM judges, G-Eval, and Prometheus motivate rubric-based automated assessment and human validation (Zheng et al., 2023; Liu et al., 2023; Kim et al., 2024).

Boundaries of methodological analogies. Efficient sampler learning and personalized or customized multiple clustering address sampling efficiency and alternative representations of data (Yao et al., 2024a;c;b). These are methodological neighbors to controlled sampling and profile-based analysis, not tutoring evaluations: ScaffoldBench neither implements their algorithms nor uses their results as evidence of scaffolding quality. Its claims concern the explicitly defined educational evaluation protocol.