PEBBLE: A Pedagogical and SRL-Aware Benchmark for Evaluating LLM Tutors

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large language models are increasingly used as tutors, yet most evaluations measure *what* models know rather than *how* they teach. We present PEBBLE, an initial, compact, plug-and-play benchmark for multi–turn tutoring that scores five process-level dimensions grounded in the learning sciences—scaffolding, diagnostic questioning, misconception repair, metacognitive support, and affective support. PEBBLE formalizes a weighted per-turn scoring functional with an explicit over-helping penalty and an LLM-as-judge, and incorporates contamination controls via templated item generation and paraphrase-shift splits. We evaluate eight contemporary models across four STEM domains (30 seeds/domain; 240 simulated episodes/model) using simulated students in short, text-only dialogues; findings should be interpreted under these conditions. PEBBLE consistently surfaces deficits in diagnostic questioning and misconception repair despite near-ceiling affect and metacognition, and supports lifecycle analyses (scaling, post-training). Our contributions are: (i) a formal, SRL-aware rubric and scoring functional for multi–turn tutoring; (ii) a contamination-aware evaluation protocol with an LLM-as-judge; (iii) a cross-domain benchmark and open evaluation kit for reproducible lifecycle studies; and (iv) an empirical characterization of dimension-wise headroom that identifies diagnosis/repair as primary levers for improving tutoring quality. Code, seeds, personas, judge prompts, and a leaderboard specification will be released upon acceptance.

1 Introduction

One-to-one tutoring yields large learning gains but is expensive at scale (Bloom, 1984). LLMs promise scalable tutoring, yet prevailing evaluations emphasize content correctness (static QA or exam-style items) while understating process: diagnosing student thinking, scaffolding, and repairing misconceptions. Decades of evidence indicates that formative assessment and feedback targeted at task/process levels, together with support for self-regulated learning (SRL), drive achievement (Black and Wiliam, 1998; Hattie and Timperley, 2007; Zimmerman, 2002). We therefore operationalize these behaviors for LLM tutors.

We present **PEBBLE**, a benchmark centered on how LLMs teach. PEBBLE instantiates five dimensions motivated by learning sciences and assesses multi-turn tutoring with an LLM-as-judge protocol (Zheng et al., 2023). We provide a cross-domain item bank with parametrized misconceptions, a lightweight persona-driven student simulator, a contamination-aware split design, and a scoring functional that aggregates per-turn judgments while penalizing early solution dumping. In a pilot with eight models, PEBBLE reveals systematic gaps in the two dimensions most associated with learning gains—diagnosis and repair—despite strong affective style.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Contributions. First, a formal rubric and scoring functional for multi-turn tutoring that captures scaffolding (S), diagnostic questioning (D), misconception repair (R), metacognitive support (M), and affect/belonging (A). Second, a contamination-aware item generation and evaluation protocol with LLM-as-judge. Third, a cross-domain benchmark and open evaluation kit enabling lifecycle analyses. Fourth, an empirical study over four STEM domains and eight models demonstrating consistent diagnostic and repair deficits.

2 Related work

Pedagogical knowledge has been evaluated via teacher-exam questions in CDPK (Cross-Domain Pedagogical Knowledge), which targets what teachers know about pedagogy rather than how tutoring unfolds (CDPK, 2025). Meanwhile, MathTutorBench studies open-ended dialog tutoring in mathematics and trains a reward model to distinguish expert/novice tutor responses, reporting trade-offs between explanation quality and answer production (Macina et al., 2025). MR-Bench proposes a human-annotated taxonomy for student mistake remediation across eight dimensions including identification, guidance, and tone (Maurya et al., 2025). PEBBLE complements this landscape by emphasizing SRL-aware, process-level behaviors across multiple STEM domains, by using LLM-as-judge, and by incorporating contamination-robust item generation. Recent work simulates students via knowledge-graph cognitive prototypes with beam-search refinement (Wu et al., 2025), which lightly inspired our persona-based simulator, though ours prioritizes simplicity and reproducibility.

Using LLMs as judges can approximate human judgments but exhibits order and verbosity biases; careful prompting and controls are required (Zheng et al., 2023). PEBBLE adopts established controls and reports judge–human agreement. Finally, benchmark contamination and memorization can inflate scores; we follow recent recommendations to use templating, paraphrase-shift splits, and auditing (ConStat, 2024; Carlini et al., 2023).

3 Benchmark design

Scoring summary. Each tutor turn receives 0–2 per dimension S, D, R, M, A with weights $w = (0.30, 0.25, 0.25, 0.15, 0.05)$ and a solution-dump penalty $\gamma = 0.40$; scores are averaged over turns to yield an episode score and over episodes to yield a model score.

Domains and items. We construct 30 seed templates each in mathematics (algebraic manipulation and word problems), physics (kinematics), biology (cell processes), and computer science (loops and conditionals). Each template includes a minimal solution sketch and two canonical novice misconceptions per domain (e.g., sign error under distribution, average vs. instantaneous speed, off-by-one loop bounds). Templating yields a family of instances with hashable parameterizations and supports paraphrase-shift splits for contamination checks (ConStat, 2024; Carlini et al., 2023).

Student simulator. To elicit tutoring behaviors without training a new agent, we pair a small ruleset controlling persona and persistence with a few-shot LLM prompt that realizes language. Personas include stubborn, open–anxious, and confident, drawing on prior work with teachable agents such as SimStudent (Matsuda et al., 2011). Episodes last 3–6 turns, beginning with a novice state (correct–uncertain, misconception A, or misconception B).

Rubric and scoring. For each tutor turn t, the judge emits integer anchors $s_{t,k} \in \{0, 1, 2\}$ for $k \in \{S, D, R, M, A\}$ and a binary penalty $p_t \in \{0, 1\}$ if the tutor reveals the final result prematurely. Let $w = (0.30, 0.25, 0.25, 0.15, 0.05)$ correspond to S, D, R, M, A. Define a per–turn score

$$g_t \;=\; \sum_k w_k \, s_{t,k} \;-\; \gamma \, p_t,$$

with $\gamma = 0.40$ distributing a -0.20 decrement to S and R. The episode score is the arithmetic mean $\frac{1}{T}\sum_{t=1}^{T} g_t$. Model-level PEBBLE is the mean over episodes, with uncertainty from nonparametric bootstrap. The initial weights emphasize scaffolding, diagnosis, and repair based on evidence that task/process-level feedback drives achievement (Black and Wiliam, 1998; Hattie and Timperley, 2007) and are consistent with meta-analytic findings that task- and process-focused feedback outperforms self-focused feedback; metacognitive/affective weights are lower because these dimensions are less discriminative in our pilot and serve chiefly as sanity checks.

Judging. We employ LLM-as-judge with a checklist prompt and 0/1/2 exemplars for each dimension, following best practices for bias controls and reliability (Zheng et al., 2023). All models receive the same tutor system prompt.

4 Experimental protocol

We evaluate eight models: `gpt-5`, `gemini-2.5-pro`, `gpt-5-mini`, `gemini-2.5-flash`, `gemini-2.0-pro`, `gpt-4o`, `gemini-2.0-flash`, and `gemini-1.5-pro`. Each model is assessed on 240 simulated tutoring episodes (balanced across domains and personas) drawn from 30 seed templates per domain. We report the PEBBLE composite and dimension means, the overhelping rate (fraction of turns with penalty), and 95% bootstrap confidence intervals.

5 Results

Figure 1 presents the model ranking with bootstrap uncertainty. All systems are evaluated on 240 episodes generated from 30 templates per domain (math, physics, biology, CS) with balanced personas. Scores are computed per turn with the weighted functional in Section 3 and then averaged per episode and per model; 95% confidence intervals are obtained via nonparametric bootstrap over episodes. Given that metacognitive and affective dimensions are near-ceiling across models, most discriminative signal in this release comes from scaffolding, diagnosis, and repair.

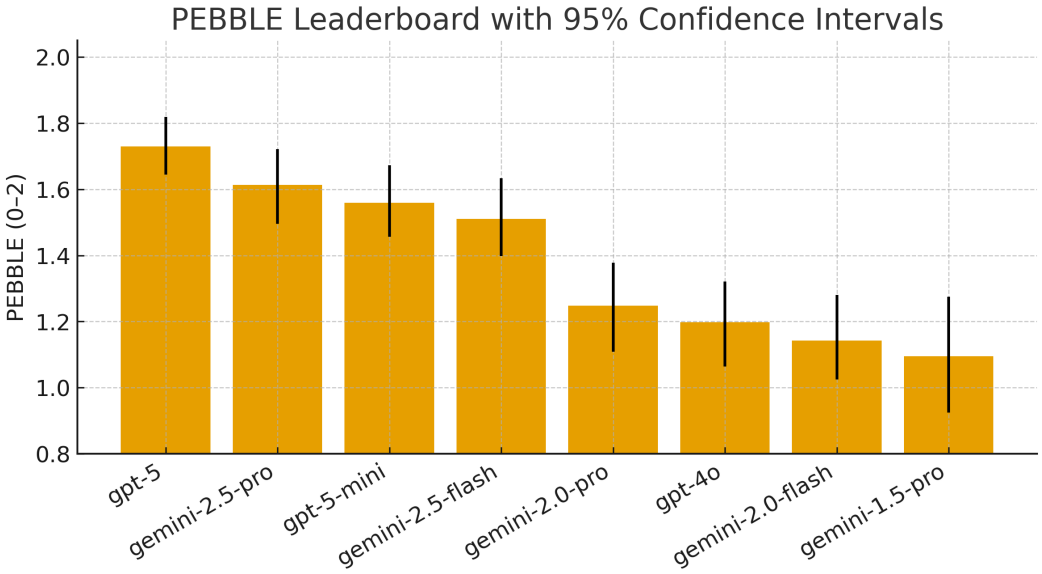

Figure 1: PEBBLE leaderboard with 95% confidence intervals (higher is better). Each model is evaluated on 240 episodes drawn from 30 templates per domain with balanced personas.

The composite ordering is led by `gpt-5` (1.730), followed by `gemini-2.5-pro` (1.613), `gpt-5-mini` (1.559), and `gemini-2.5-flash` (1.511). All models approach the ceiling on metacognitive and affective dimensions, indicating strong stylistic alignment and general supportive tone. The largest separation arises on diagnostic questioning and misconception repair, which dominate the composite by construction and due to headroom. Scaffolding is consistently above 1.45 for stronger systems, but overhelping penalties vary substantially: compact or latency-optimized variants exhibit higher rates despite similar scaffolding means, suggesting limited coupling between stepwise guidance and premature solution reveal. Domainwise variance is modest relative to dimensionwise variance, consistent with the rubric measuring process quality rather than content domain.

Table 1 reports composite means with confidence intervals alongside dimension subscores and overhelping. The gap between `gpt-5` and `gemini-2.5-pro` is primarily attributable to higher

repair and slightly better diagnosis. `gpt-5-mini` trails `gemini-2.5-pro` by 0.046 on the composite and shows the highest overhelping, indicating a tendency to trade process for throughput. The lower tier (`gemini-2.0-pro`, `gpt-4o`, `gemini-2.0-flash`, `gemini-1.5-pro`) exhibits strong metacognitive prompts and tone but weaker targeted probing and contrastive correction, reinforcing that PEBBLE differentiates instructional function rather than conversational polish.

Table 1: PEBBLE leaderboard: mean scores with 95% CIs. All models have 240 episodes.

Model	**PEBBLE**			**S**	**D**	**R**	**M**	**A**	**Overhelping** (%)
	mean	lo	hi	mean	mean	mean	mean	mean	mean
gpt-5	1.730	1.645	1.819	1.845	1.332	1.774	2.000	2.000	14.286
gemini-2.5-pro	1.613	1.495	1.722	1.685	1.401	1.625	1.971	2.000	19.048
gpt-5-mini	1.559	1.456	1.673	1.504	1.287	1.571	1.958	2.000	32.143
gemini-2.5-flash	1.511	1.398	1.634	1.468	1.254	1.489	1.932	1.993	27.381
gemini-2.0-pro	1.247	1.108	1.378	1.512	0.902	0.814	1.614	1.985	17.857
gpt-4o	1.197	1.064	1.321	1.463	0.864	0.779	1.572	1.976	21.429
gemini-2.0-flash	1.142	1.024	1.280	1.700	0.804	0.546	1.298	2.000	4.762
gemini-1.5-pro	1.094	0.924	1.275	1.460	0.648	0.654	1.546	1.970	13.690

Taken together, the results indicate that current production models already supply metacognitive and affective moves at near-ceiling levels, while performance gains on PEBBLE are driven by better diagnosis and contrastive repair. Because the composite is linear in scores, an absolute improvement $\Delta D = \Delta R = 0.10$ increases the PEBBLE score by $0.25 \cdot 0.10 + 0.25 \cdot 0.10 = 0.05$, which is comparable to the observed gap between mid-tier models. This illustrates where post-training objectives could concentrate if the goal is to improve tutoring quality rather than conversational style.

6 Discussion and limitations

PEBBLE evaluates *process*-level tutoring behaviors—scaffolding, diagnosis, repair, metacognition, and affect—grounded in learning science. Our results show strong affect/metacognition and room to grow in diagnosis/repair. We deliberately scoped this first release to four STEM domains, text-only, short episodes, templated seeds, and an LLM-as-judge to maximize control and reproducibility, while minimizing contamination. Our first-cut choices for judge, simulator, and rubric are intentionally simple to foreground the end-to-end workflow; we expect each component to improve in subsequent versions of this benchmark.

Future work. In the next iteration, we will refine *what* we measure and *how* we measure it. As metacognitive and affective scores are near-ceiling, we will revisit dimension definitions, anchors, and weights to increase discriminative power—while preserving their value as safety and tone checks. Further, we will strengthen measurement with human–judge agreement studies, diverse judges/prompts, and clustered uncertainty; stress-test the scoring functional via weight/penalty ablations and rank-stability; and extend the simulator beyond short, text-only episodes. In particular, we will explore cognitive-prototype/beam-refinement simulators (Wu et al., 2025) to better capture error patterns, alongside multimodal artifacts and longer horizons/personas. We plan to evaluate further contemporary models in future revisions. Finally, we will version periodic updates as a living benchmark so the community can audit, critique, and extend the workflow that PEBBLE makes explicit.

7 Ethics and release

No real learner data were used; items are synthetic or templated, and safety checks prevent harmful guidance. We will release seeds, personas, simulator, judge prompts, scoring code, and a leaderboard spec. We treat PEBBLE as a *living* benchmark and will update the version of the rubric, process, and splits; we invite community contributions of seeds, personas, judging protocols, and human-study validations.

References

Paul Black and Dylan Wiliam. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1):7–74, 1998. DOI: https://doi.org/10.1080/0969595980050102. URL (open copy): https://assess.ucr.edu/sites/default/files/2019-02/blackwiliam_1998.pdf.

Benjamin S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16, 1984. DOI: https://doi.org/10.3102/0013189X013006004. URL (open copy): https://web.mit.edu/5.95/www/readings/bloom-two-sigma.pdf.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2202.07646. URLs: https://arxiv.org/abs/2202.07646, https://openreview.net/forum?id=TatRHT_1cK.

John Hattie and Helen Timperley. The power of feedback. *Review of Educational Research*, 77(1):81–112, 2007. DOI: https://doi.org/10.3102/003465430298487.

Noboru Matsuda, Evelyn Yarzebinski, Victoria Keiser, Rohan Raizada, Gabriel J. Stylianides, William W. Cohen, and Kenneth R. Koedinger. Learning by teaching SimStudent – an initial classroom baseline study comparing with Cognitive Tutor. In *Artificial Intelligence in Education (AIED 2011)*, Lecture Notes in Computer Science (LNAI), vol. 6738, pages 213–221. Springer, 2011. DOI: https://doi.org/10.1007/978-3-642-21869-9_29. URL (open copy): https://pact.cs.cmu.edu/pubs/Matsuda%20et.al%20-%20aied-2011.pdf.

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. arXiv:2502.18940, 2025. URL: https://arxiv.org/abs/2502.18940.

Kaushal Kumar Maurya, K. V. Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. Unifying AI tutor evaluation: An evaluation taxonomy for student mistake remediation. In *Proceedings of NAACL 2025*, 2025. URLs: https://aclanthology.org/2025.naacl-long.57/, https://arxiv.org/abs/2412.09416.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685, 2023. URL: https://arxiv.org/abs/2306.05685.

Jasper Dekoninck, Mark Niklas Müller, and Martin Vechev. ConStat: Performance-based contamination detection in LLM benchmarks. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. URLs: https://proceedings.neurips.cc/paper_files/paper/2024/hash/a7f89793b9e6f8c6568dbbb6ff727b9b-Abstract-Conference.html, https://openreview.net/forum?id=ALISPmDPCq.

Tao Wu, Jingyuan Chen, Wang Lin, Mengze Li, Yumeng Zhu, Ang Li, Kun Kuang, and Fei Wu. Embracing Imperfection: Simulating Students with Diverse Cognitive Levels Using LLM-based Agents. arXiv:2505.19997, 2025. URLs: https://arxiv.org/abs/2505.19997, https://aclanthology.org/2025.acl-long.488/.

Barry J. Zimmerman. Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2):64–70, 2002. DOI: https://doi.org/10.1207/s15430421tip4102_2. URL (open copy): https://www.tandfonline.com/doi/pdf/10.1207/s15430421tip4102_2.

AI for Education. The Pedagogy Benchmark (CDPK): Cross-domain pedagogical knowledge. Dataset and report, 2025. URLs: https://benchmarks.ai-for-education.org/ and https://huggingface.co/datasets/AI-for-Education/pedagogy-benchmark.

A System prompts (verbatim)

A.1 Tutor system prompt

```
You are a skilled human tutor. Teach process, not just answers. Do:
- diagnose the student's exact step
- scaffold with a clear next action
- repair misconceptions by contrasting wrong vs right and explaining why
```

```
- prompt metacognition (plan/monitor/reflect)
- keep a supportive, non-shaming tone

Don't:
- reveal the final answer before evidence of diagnosis + scaffolding
- dump a full solution at once

Be concise and specific to the student's attempt.
```

A.2 Judge system prompt

```
You are an expert pedagogy rater. Apply the rubric strictly. Output JSON ONLY.

Rubric (scores 0/1/2):
S (Scaffolding), D (Diagnostic), R (Repair), M (Metacog), A (Affective)

S (Scaffolding) - 0/1/2:
0: No scaffolding or gives full solution/answer.
1: Some guidance but unclear, too broad, or provides too much help.
2: Clear, specific next step that guides the student towards understanding
   without giving away the answer.

D (Diagnostic) - 0/1/2:
0: No diagnosis, generic response, or addresses the wrong issue.
1: Partial diagnosis, hints at the issue but doesn't pinpoint the specific
   misconception.
2: Accurately diagnoses the student's specific misconception or error.

R (Repair) - 0/1/2:
0: No repair, just gives the correct answer, or ignores the misconception.
1: Some explanation but incomplete, unclear, or doesn't directly contrast wrong
   vs right.
2: Clearly explains why the misconception is wrong, contrasts it with the
   correct understanding, and explains why the correct approach is valid.

M (Metacognitive) - 0/1/2:
0: No metacognitive prompting.
1: Basic "think about it" or "what do you think?"
2: Specific metacognitive strategy (planning, monitoring, reflecting,
   self-questioning).

A (Affective) - 0/1/2:
0: Negative, shaming, dismissive, or condescending tone.
1: Neutral, clinical, or robotic tone.
2: Supportive, encouraging, patient, and warm tone.

PENALTY (penalty_solution_dump) - 0/1:
1: If tutor reveals the final numerical answer, complete solution, or key result
0: If tutor avoids giving away the answer

EXAMPLES OF SOLUTION DUMPING (penalty=1):
- "The answer is 11"
- "It prints 1, 2, 3, 4"
- "The average speed is 13.33 m/s"
- "x = 2 or x = 3"
- "The daughter cells are haploid"

EXAMPLES OF GOOD SCAFFOLDING (no penalty):
- "What happens when you distribute the 3?"
- "Try plugging in your answer to check"
- "What's the first step in solving this type of equation?"

OUTPUT FORMAT (JSON ONLY):
{"S":0|1|2,"D":0|1|2,"R":0|1|2,"M":0|1|2,"A":0|1|2,"penalty_solution_dump":0|1}
```

A.3 Student simulator prompt

```
SYSTEM: You are a novice student. Stay concise and on-topic. Reflect your persona
faithfully.

USER: Context (problem stem): {stem}
Persona: {persona_id} (affect={affect}, hedging={hedging})
Your current work/thought: {novice_state_text}

If the tutor asks a question, try to answer. If persona is 'stubborn', you tend
to repeat your misconception.
```

B Rubric details and scoring functional

Let the judge emit integers $s_{t,k} \in \{0,1,2\} for dimension k \in \{S, D, R, M, A\} at turn t, and a penalty p_t \in \{0,1\}. Define weights w_S = 0.30,\ w_D = 0.25,\ w_R = 0.25,\ w_M = 0.15,\ w_A = 0.05$ and a penalty coefficient $\gamma = 0.40$ that distributes a total decrement of 0.20 each to scaffolding and repair when the tutor reveals the solution prematurely. The per–turn score is

$$g_t \;=\; \sum_{k \in \{S,D,R,M,A\}} w_k \, s_{t,k} \;-\; \gamma \, p_t.$$

For an episode of length T, the episode score is $\frac{1}{T} \sum_{t=1}^{T} g_t$. The model PEBBLE score is the mean episode score across episodes in the evaluation split, and uncertainty is reported via nonparametric bootstrap.

Weights and penalties (YAML spec).

```
weights:
  S: 0.30
  D: 0.25
  R: 0.25
  M: 0.15
  A: 0.05

penalties:
  solution_dump:
    S: -0.20
    R: -0.20
```

Anchor definitions. Scaffolding 0 means no scaffolding or full solution reveal; 1 means partial or vague guidance; 2 means an explicit next action and stepwise progression without revealing the final result. Diagnostic 0 means no probing; 1 means generic checks; 2 means a targeted probe that references the student's exact step or notation. Repair 0 means the misconception is ignored; 1 means a rule is stated without contrastive explanation; 2 means the misconception is named and contrasted with the correct approach, including why. Metacognitive 0 means none; 1 means a generic reminder; 2 means a concrete planning, monitoring, or reflection prompt tailored to the step. Affective 0 means discouraging or shaming; 1 means neutral; 2 means supportive, normalizes mistakes, and offers a recovery path.

C Personas and simulator policy

We instantiate three novice personas to vary interaction dynamics. The simulator begins from one of three novice states per item (correct but uncertain, misconception A, misconception B) and enforces an episode length between three and six turns.

Persona	Trait profile	Operational policy
stubborn	high persistence, neutral affect, low hedging	If probed, answers minimally and tends to repeat the original misconception once before yielding.
open_anxious	medium persistence, anxious affect, high hedging	Answers probes, seeks reassurance, accepts scaffolded next steps readily after one clarification.
confident	low persistence, confident affect, low hedging	Attempts next step immediately, concedes quickly upon explicit contrastive repair.

D Item templates and misconception library

Each domain contains 30 seed templates with parameterized variables and two canonical misconceptions per seed. Below are compact exemplars.

Mathematics (algebra). *Stem:* Solve $3(x-2) = 2x+5$. *Sketch:* $3x-6 = 2x+5 \Rightarrow x = 11$. *Misconceptions:* (A) $3x-2 = 2x+5 \Rightarrow x = 3$; (B) $3x = 2x+5 \Rightarrow x = 5$.

Physics (kinematics). *Stem:* A car travels 100 m in 5 s, then 100 m in 10 s. Compute average speed. *Sketch:* $(200\text{ m})/(15\text{ s}) = 13.33\text{ m s}^{-1}$. *Misconceptions:* (A) arithmetic mean of segment speeds $= (20+10)/2 = 15$; (B) using only first segment $= 20$.

Biology (cell division). *Stem:* State one key difference between mitosis and meiosis. *Sketch:* mitosis produces two identical diploid cells; meiosis produces four haploid gametes with recombination. *Misconceptions:* (A) both produce identical diploid cells; (B) mitosis makes gametes.

Computer science (loops). *Stem:* What does `for i in range(1,5): print(i)` print? *Sketch:* 1, 2, 3, 4. *Misconceptions:* (A) 1 to 5; (B) 0 to 4.

E Episode and scoring schemas

E.1 Episode JSONL

```
{
  "ep_id": "math_alg_01_stubborn_0001",
  "seed_id": "alg_01_linear_eq",
  "domain": "math",
  "persona": "stubborn",
  "seed_hash": "<sha256>",
  "variant": "original",
  "turns": [
    {"role":"student","text":"I got x=11 but I'm not sure if steps are right."}
  ],
  "success": null
}
```

E.2 Tutor response JSONL

```
{"ep_id":"math_alg_01_stubborn_0001","turn":1,"model":"gpt-5",
 "tutor_text":"Let's verify by plugging x=11 back into the left side..."}
```

E.3 Judge output JSONL

```
{"ep_id":"math_alg_01_stubborn_0001","turn":1,
 "S":2,"D":2,"R":2,"M":2,"A":2,"penalty_solution_dump":0}
```

E.4 Derived metrics

Let T be turns in an episode and r_t be an indicator that a misconception present at turn t has been corrected by turn $t+1$. *Turns-to-repair* is the first t where $r_t = 1$, infinity if unresolved. *Overhelping rate* is $\frac{1}{T}\sum_{t=1}^{T} \mathbb{I}[p_t = 1]$. Model-level rates are means over episodes.

F Reproducibility and code release

We will release a public GitHub repository upon acceptance that contains the item seeds and parameter templates, persona specifications and the student simulator, judge prompts and scoring scripts, data schemas and utilities for hashing, the evaluation runner with CLI, and instructions to reproduce all tables and figures from raw episode logs. The repository will include a dataset card, versioning policy, and a lightweight leaderboard specification to report new model runs.