Are Agents Ready to Teach?

A Multi-Stage Benchmark for Real-World Teaching Workflows

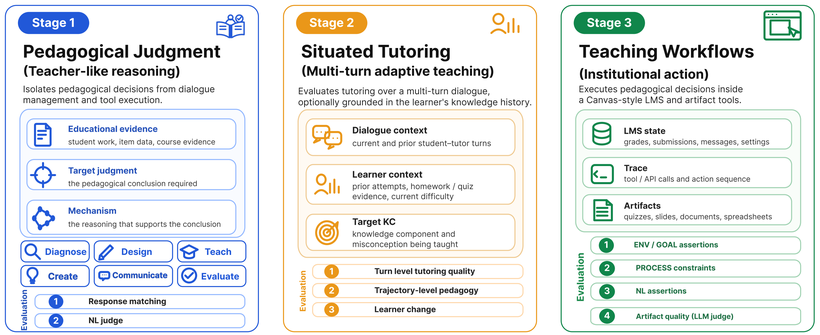
Zixin Chen[1,2], Peng Liu[2], Rui Sheng[1], Haobo Li[1], Jianhong Tu[2], Xiaodong Deng[2], Kashun Shum[1,2], Dayiheng Liu[2], Huamin Qu[1]

[1] The Hong Kong University of Science and Technology
[2] Qwen Team, Alibaba Group

Abstract

TeachArena uses **354 source-grounded tasks** to evaluate evidence-grounded teaching decisions, adaptive tutoring, and verified course interventions. Across **17 models**, bounded judgment scores are generally high, while tutoring and workflows expose gaps in eliciting learner reasoning and carrying a diagnosis into action.

Motivation and benchmark design



Teaching insight › Grounded evidence › Observable trace › Matched verifier

A **teaching-system episode** links the evidence, a teaching decision, and the resulting learner behavior or course state.

Stage	Where tasks come from	What the verifier checks
117 judgment tasks	MathDial and MathTutorBench, open resources, education research, and authored cases.	Checks the decision and its rationale, with programmatic response matching where applicable.
100 tutoring episodes	Scenarios based on open resources or MathDial, plus adapted educational problems.	Scores individual tutor turns, the full tutoring trajectory, and learner-produced change.
137 LMS workflows	Instructor requests and seeded records in a synthetic learning management system.	Checks course state, action order, and semantic alignment. Uses artifact rubrics where applicable.

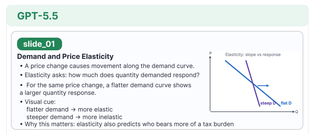
Diagnostic case studies

Stage 2 · Modeling the learner

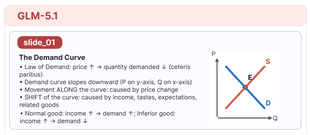
Profile	Prior knowledge, misconceptions, self-regulation, motivation, and affect.
Reply and fidelity	Generate from profile and dialogue. Screen replies for fidelity before the tutor sees them.
State tracking	Update learner state after an accepted reply, using evidence in the dialogue.

What the three scores capture

Tutor turn Illustrative move	Credits spotting the reversed conditional. Deducts for missing immediate reassurance.
Trajectory Observed episode	Credits fading support after the correct calculation. The handoff is incomplete without a new-context check.
Learner change Independent evidence	Credits independent reasoning, self-monitoring, and engagement. Discounts mastery after tutor-supplied reasoning.



Assigned deck revised



Assigned deck left unchanged

Stage 3 · From midterm evidence to a course update

Fall 2023 midterm 120 attempts › Identify four weak concepts › Edit Week 5 deck Build visual quiz › Then notify advisor and class

GPT-5.5 aligns the deck, quiz, and outreach to the diagnosis. **GLM-5.1** builds a separate deck and leaves Week 5 unchanged.

Evaluation results

#	Model	Judgment	Tutoring	Workflows	Overall
1	Claude Opus 4.8	**0.930**	**0.773**	**0.704**	**0.803**
2	Claude Opus 4.6	**0.947**	0.763	**0.661**	**0.791**
3	GPT-5.5-pro	0.916	0.739	**0.693**	**0.783**
4	GPT-5.5	0.916	0.717	0.659	**0.764**
5	Qwen3.7-Max	0.914	**0.766**	0.580	**0.753**
6	GLM-5.2	**0.932**	0.763	0.562	**0.752**
7	Qwen3.6-Plus	0.905	0.756	0.585	**0.749**
8	GPT-5.4	0.874	0.729	0.596	**0.733**
9	DeepSeek-V4-Pro	0.885	0.738	0.572	**0.731**
10	GLM-5.1	0.871	0.762	0.547	**0.727**
11	Gemini-3.1-pro	0.892	0.763	0.514	**0.723**
12	Gemini-2.5-pro	0.867	**0.769**	0.477	**0.704**
13	Gemini-3.5-flash	0.906	0.737	0.432	**0.692**
14	Kimi K2.6	0.901	0.761	0.395	**0.686**
15	Qwen3-Max	0.900	0.683	0.449	**0.677**
16	Gemini-3.1-flash-lite	0.819	0.748	0.452	**0.673**
17	GPT-4.1	0.799	0.732	0.447	**0.659**

Overall weights stages equally. Bold stage scores mark each stage's top three.

Discussion of findings

Teaching capabilities can diverge

Tutoring ranks correlate weakly with judgment (0.24) and workflows (0.21). Judgment and workflows correlate more (0.58).

Fast answers can crowd out reasoning

When learners ask for shortcuts, elicit their next reasoning step and check what they can do independently.

Tool success can miss the teaching goal

Verify that the diagnosis reaches the intended course object, the change persists, and outreach reflects the intervention.

Paper · arXiv:2605.14322

Dataset and evaluators · CinderD / TeachArena