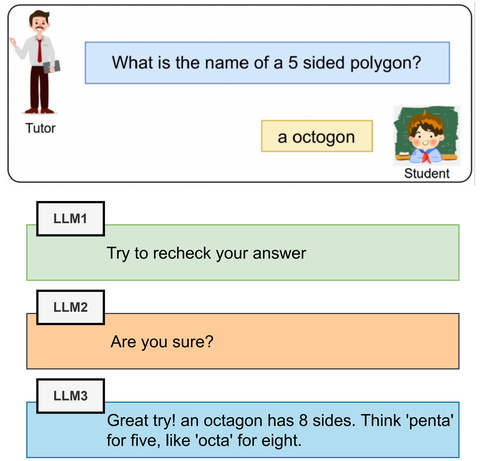
Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors

Kaushal Kumar Maurya KV Aditya Srivatsa Kseniia Petukhova Ekaterina Kochmar

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) Abu Dhabi, UAE

NAACL 2025

Introduction



Research Objectives

- **RQ1:** To what extent do LLM-powered AI tutors exhibit the pedagogical competencies essential for effective AI tutoring?
- **RQ2:** What are the key pedagogical attributes of an effective tutor?

Research Space: Student Mistake Remediation Task

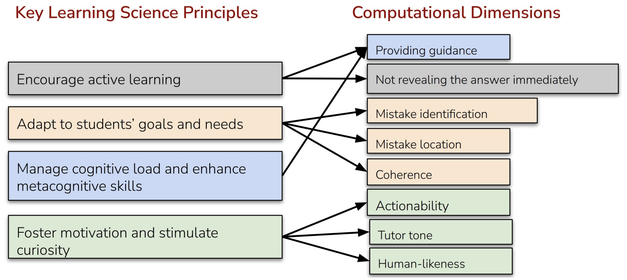
Consider the conversation history between a tutor and a student:

$$H = \{(T_1, S_1), (T_2, S_2), \ldots, (T_t, S_t)\}$$

where T_i and S_i denote the i-th responses from the tutor and student, respectively. Let S_k represent the student's most recent k utterances, where $k \in [1, \ldots, t]$, containing an error or misconception.

The objective is to assess the pedagogical appropriateness of the human/AI tutor's response T_{t+1}, which aims to address and rectify the issue in S_k.

Proposed Unified Evaluation Taxonomy

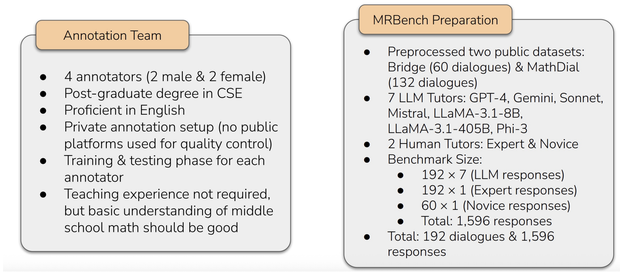


Dimension	Definition	Desiderata
Mistake identification	Has the tutor identified/recognized a mistake in a student's response?	Yes
Mistake location	Does the tutor's response accurately point to a genuine mistake and its location?	Yes
Revealing of the answer	Does the tutor reveal the final answer (whether correct or not)?	No
Providing guidance	Does the tutor offer correct and relevant guidance, such as an explanation, elaboration, hint, examples, and so on?	Yes
Actionability	Is it clear from the tutor's feedback what the student should do next?	Yes
Coherence	Is the tutor's response logically consistent with the student's previous responses?	Yes
Tutor tone	Is the tutor's response encouraging, neutral, or offensive?	Encouraging
Human-likeness	Does the tutor's response sound natural rather than robotic or artificial?	Yes

Dimension	TP'22	MA'23	WA'24	DA'24	Ours
Mistake identification	✓	✓	✗	✓	✓
Mistake location	✗	✗	✗	✓	✓
Revealing of the answer	✗	✓	✗	✗	✓
Providing guidance	✓	✗	✓	✗	✓
Actionability	✗	✗	✗	✓	✓
Coherence	✗	✓	✗	✗	✓
Tutor tone	✓	✗	✓	✗	✓
Human-likeness	✓	✗	✓	✗	✓

Evaluation dimensions considered in previous research on AI tutoring for student mistake remediation. *TP'22* refer to [3], *MA'23* – [2], *WA'24* – [4], and *DA'24* – [1].

Setup: Annotation and MRBench Data Preparation



- Conducted validation pilot Study
- Metrics: DAMR & AC || LLM as Judge: `Prometheus2` & `Llama-3.1-8B`

Summary of the Result

Tutor	Mistake Identification	Mistake Location	Revealing of the Answer	Providing Guidance	Actionability	Coherence	Tutor Tone	Human-likeness
`*Novice`	43.33	16.67	80.00	11.67	1.67	50.00	90.00	35.00
`Expert`	76.04	63.02	90.62	67.19	**76.04**	79.17	92.19	87.50
`Llama-3.1-8B`	80.21	54.69	73.96	45.31	42.71	80.73	19.79	93.75
`Phi3`	28.65	26.04	73.96	17.71	11.98	39.58	45.31	52.08
`Gemini`	63.02	39.58	67.71	37.50	42.71	56.77	21.88	68.23
`Sonnet`	85.42	69.79	**94.79**	59.38	60.94	88.54	**54.69**	**96.35**
`Mistral`	93.23	73.44	86.46	63.54	70.31	86.98	15.10	95.31
`GPT-4`	**94.27**	**84.38**	53.12	76.04	46.35	90.17	37.50	89.62
`Llama-3.1-405B`	**94.27**	**84.38**	80.73	**77.08**	74.48	**91.67**	16.15	90.62

Tutor	Observation
`GPT-4`	Reveals the answer too quickly
`Sonnet`	Focuses on human-likeness and an encouraging tone
`Gemini`	Delivers less coherent and accurate responses
`Phi3`	Fails to understand the context, performing the worst
`Llama-3.1-405B`	Achieves the best performance but lacks high scores along many dimensions
`Novice (Human)`	Provides ambiguous and short responses
`Expert (Human)`	Focuses more on actionability and less on other dimensions

Contributions and Take-aways

- Unified evaluation taxonomy based on learning science principles (8 dimensions)
- Released MRBench: 192 conversations, 1,596 responses from 7 LLM-based and 2 human tutors + human annotations
- Investigated pedagogical abilities of LLMs as AI tutors from human perspective – there is a long way to go
- LLM as evaluator judge* – so far, unreliable

Want to read full paper?

Paper

BEA Shared Task 2025

References

[1] Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. Stepwise verification and remediation of student reasoning errors with large language model tutors. In *EMNLP*, Miami, Florida, USA, November 2024.

[2] 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 EMNLP 2023*, Singapore, December 2023.

[3] Anaïs Tack and Chris Piech. The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues. In *Proceedings Educational Data Mining, EDM 2022, Durham, UK, July 24-27, 2022*, 2022.

[4] Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *NAACL*, pages 2174–2199, 2024.