NAACL 2025

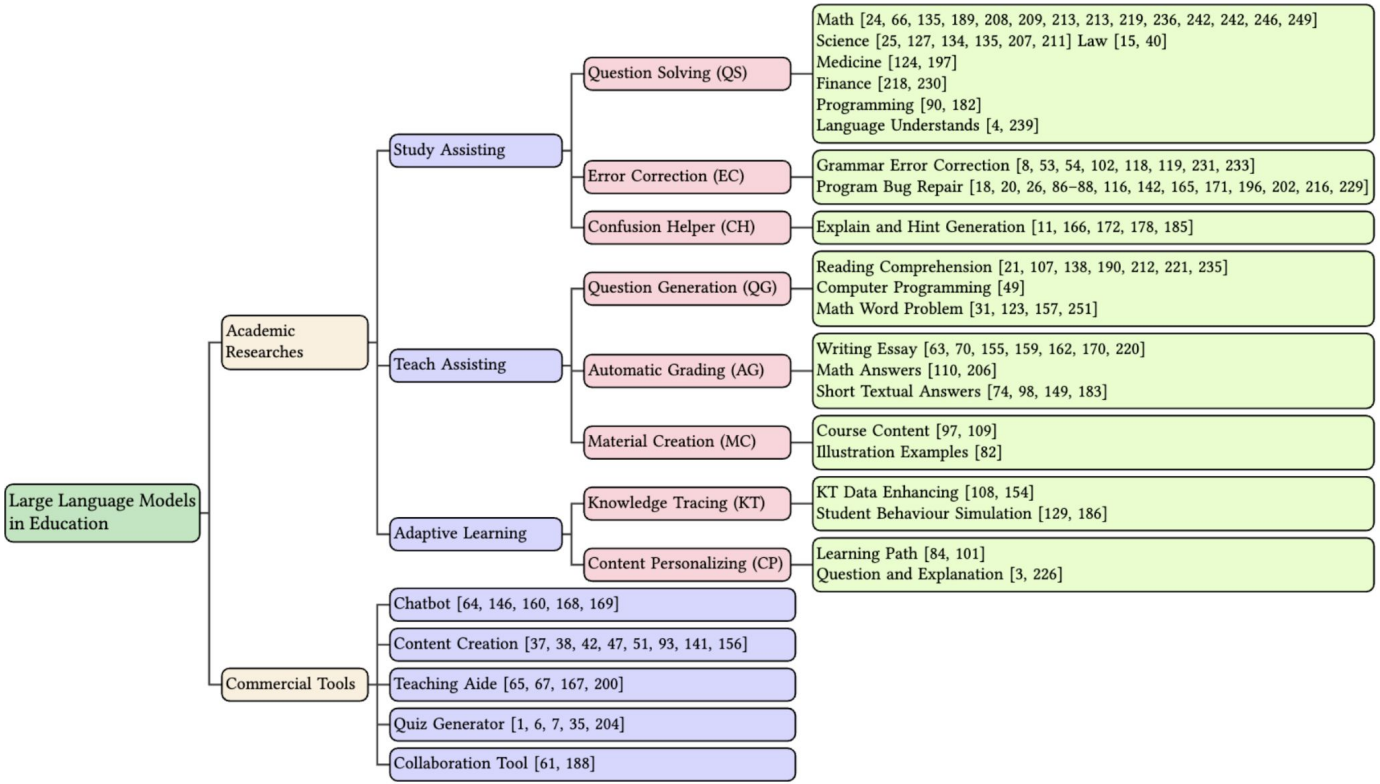
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

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

Kaushal Kumar Maurya and **KV Aditya Srivatsa**
Kseniia Petukhova and **Ekaterina Kochmar**

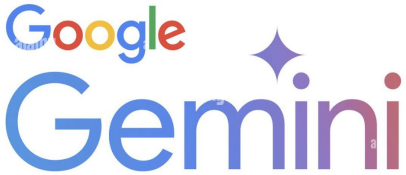
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Abu Dhabi, UAE

Opportunities with LLMs in Education



Ref: Wang et al., 2024. Large Language Models for Education: A Survey and Outlook. Arxiv.

Opportunities with LLMs in Education: AI Tutors



Imgae credit: internet

Open Research Questions

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?

Open Research Questions

RQ2: What are the *key pedagogical attributes* of an effective tutor?

Case Study: Student Mistakes Remediation Task (SMR)

Conversation topic: Simple Expressions

Conversation History:
Tutor: We have to solve the inner parentheses first.
Student: ok
Tutor: What is 5 times 6?
Student: 50

Tutor response: Are you sure?

Tutor response: That's correct, 5 multiplied by 6 equals 30.

Tutor response: Ah, not quite. 5 x 10 is 50. 5 x 6 is something else. Could you give it another try?

Ref: Jo Boaler. 2013. Ability and mathematics: The mindset revolution that is reshaping education.

SMR: Formal Definition

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.

Ref: Jo Boaler. 2013. Ability and mathematics: The mindset revolution that is reshaping education.

Literature Review: Diverse Evaluation Taxonomy

The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues

Anaïs Tack
Stanford University
atack@cs.stanford.edu

Chris Piech
Stanford University
piech@cs.stanford.edu

[1] Speak like a teacher, [2] Understand a student, and [3] Help a student

Stepwise Verification and Remediation of Student Reasoning Errors with Large Language Model Tutors

Nico Daheim*[1] Jakub Macina*[2,3]
Manu Kapur[4] Iryna Gurevych[1] Mrinmaya Sachan[2]

[1] Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science and Hessian Center for AI (hessian.AI), TU Darmstadt
[2]Department of Computer Science, ETH Zurich [3] ETH AI Center
[4]Professorship for Learning Sciences and Higher Education, ETH Zurich

[1] Targeted, [2] Correct, and [3] Actionable

Bridging the Novice-Expert Gap via Models of Decision-Making: A Case Study on Remediating Math Mistakes

Rose E. Wang Qingyang Zhang Carly Robinson
Susanna Loeb Dorottya Demszky
Stanford University
rewang@cs.stanford.edu, ddemszky@stanford.edu

[1] Prefer, [2] Useful, [4] Care, and [4] Not robot

MATHDIAL: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems

Jakub Macina* Nico Daheim* Sankalan Pal Chowdhury*
Tanmay Sinha Manu Kapur Iryna Gurevych Mrinmaya Sachan

ETH AI Center Department of Computer Science, ETH Zurich
Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science and Hessian Center for AI (hessian.AI), TU Darmstadt
National Institute of Education, Nanyang Technological University
Professorship for Learning Sciences and Higher Education, ETH Zurich
jakub.macina@ai.ethz.ch

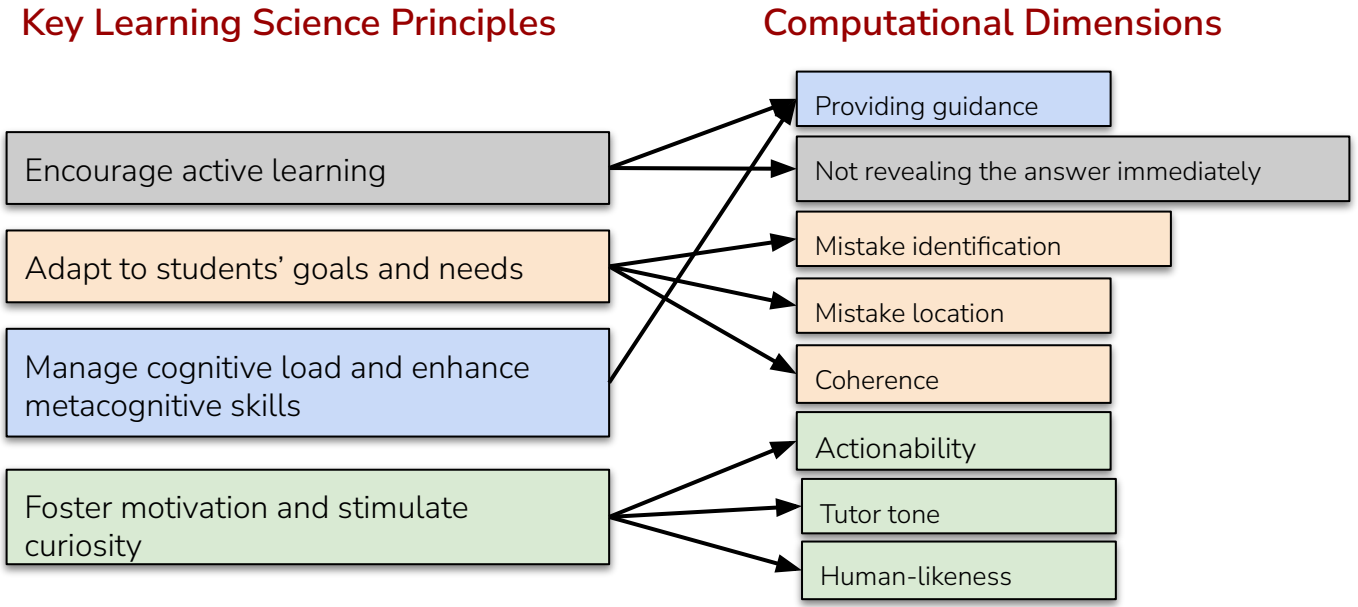
[1] Correctness, [2] Coherence, and [3] Equitable

Goal: Unification of AI Tutors Evaluation

Grounded on:

1. Previous research
2. Key learning science principles

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

Unified Evaluation Taxonomy

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	✓	✗	✓	✗	✓

TP'22: Anaïs Tack and Chris Piech. 2022. The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues. In Proceedings of the 15th International Conference on Educational Data Mining, EDM 2022, Durham, UK, July 24-27, 2022. International Educational Data Mining Society
MA'23: Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. arXiv preprint arXiv:2305.14536.
WA'24: Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024a. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2174–2199.
DA'24: Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. Stepwise Verification and Remediation of Student Reasoning Errors with Large Language Model Tutors. arXiv preprint arXiv:2407.09136

Validation of Evaluation Taxonomy

Are the proposed dimensions sufficient? || Are there redundancies among them?

Annotation Team

- 4 annotators (2 male & 2 female)
- Post-graduate degree in CSE
- Proficient in English
- Private annotation setup (no public platforms used for quality control)
- Training & testing phase for each annotator
- Teaching experience not required, but basic understanding of middle school math should be good

Validation Pilot Study

- Started with 8+ dimensions
- MRBench built on Bridge and MathDial public datasets
- 8 sampled dialogues from MRBench (4 from each dataset)
- *544 annotations per annotator:* each dialogue has multiple responses and is annotated across 8 dimensions
- All annotators converged on 8 dimensions
- Fleiss' kappa: 0.65

MRBench Sample Annotated Example


```
{
   "conversation_id": "29640377",
   "conversation history":
    "Tutor: Here is our first question.\n
    Tutor: What is the value of 10/5?\n
    Student: 50",
   "tutor_responses": {
     "Llama318B": {
       "response": "I think there might be a small mistake there, let's take a look at the problem
together – when you divide 10 by 5, we actually get 2, not 50.",
       "annotation": {
         "Mistake_Identification": "Yes",
         "Mistake_Location": "Yes",
         "Revealing_of_the_Answer": "Yes",
         "Providing_Guidance": "To some extent",
         "Coherence": "Yes",
         "Actionability": "No"
         "Tutor_Tone": "Neutral",
         "Human-likeness": "Yes",
       }
}
```

Open Research Questions

RQ1: To what extent do LLM-powered AI tutors exhibit the *pedagogical competencies* essential for effective AI tutoring?

Experimental Setup

MRBench Preparation

- Preprocessed two public datasets: Bridge (60 dialogues) & MathDial (132 dialogues)
- 7 LLM Tutors: GPT-4, Gemini, Sonnet, Mistral, LLaMA-3.1-8B, LLaMA-3.1-405B, Phi-3
- 2 Human Tutors: Expert & Novice
- Benchmark Size:
  - 192 × 7 (LLM responses)
  - 192 × 1 (Expert responses)
  - 60 × 1 (Novice responses)
  - Total: 1,596 responses
- Total: 192 dialogues & 1,596 responses

Annotation

- Human annotation with 4 annotators
- LLM as judge:
  - Prometheus2
  - Llama-3.1-8B

Assessment Metrics

- Desired Annotation Match Rate (DAMR)
- Annotation Correlation (AC) based on Pearson's correlation

Results and Discussions

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

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

BEA shared task at ACL 2025:
Towards development of sophisticated automated evaluation methods for each dimension

Contributions and Take-aways

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

* Our explorations were limited to Prometheus2 and LLaMA 3.1-8B LLMs, and a few prompts - detailed in the paper.

Call for the Community

Towards a Unified Evaluation Ecosystem

- Let's collaboratively develop a comprehensive evaluation taxonomy and benchmark for diverse AI tutor use cases.

We've presented the first footprint—let's expand it together.

Empowering Open-Source Evaluation

- MRBench is public!
- Join us in building scalable, automated evaluation metrics to accelerate research in AI tutoring.

Collective Progress Through Collaboration

- It's time for the community to unite and harmonize evaluation practices.

Let's build smarter, fairer, and more impactful AI tutors—together.

References

[1]. Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., ... & Wen, Q. (2024). Large language models for education: A survey and outlook. arXiv preprint arXiv:2403.18105.

[2]. Boaler, J. (2013, March). Ability and mathematics: The mindset revolution that is reshaping education. Forum.

[3]. Anaïs Tack and Chris Piech. 2022. The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues. In Proceedings of the 15th International Conference on Educational Data Mining, EDM 2022, Durham, UK, July 24-27, 2022. International Educational Data Mining Society

[4]. Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. arXiv preprint arXiv:2305.14536.

[5]. Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024a. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2174–2199.

[6]. Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. Stepwise Verification and Remediation of Student Reasoning Errors with Large Language Model Tutors. arXiv preprint arXiv:2407.09136

[7]. Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. arXiv preprint arXiv:2405.01535.

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

[9]. Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530.

[10]. Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. In https://api.semanticscholar.org/CorpusID:268232499.

[11]. Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. arXiv preprint arXiv:2310.06825

[12]. Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The LLaMA 3 herd of models. arXiv preprint arXiv:2407.21783.

[13]. Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219.

[14]. Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, et al. 2024. Towards responsible development of generative AI for education: An evaluation-driven approach. arXiv preprint arXiv:2407.12687.

NAACL 2025

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Thank you!!

Paper

BEA shared task @ ACL 2025: To develop ***sophisticated*** automated evaluation methods for each dimension.

GitHub

Acknowledgment: We thank Google for supporting this research through *Google Academic Research Award (GARA) 2024*.

Correspondence:
kaushal.maurya@mbzuai.ac.ae