Pedagogy-driven Evaluation of Generative AI-powered Intelligent Tutoring Systems

Kaushal Kumar Maurya and Ekaterina Kochmar

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

AIED 2025 (BlueSky Track)

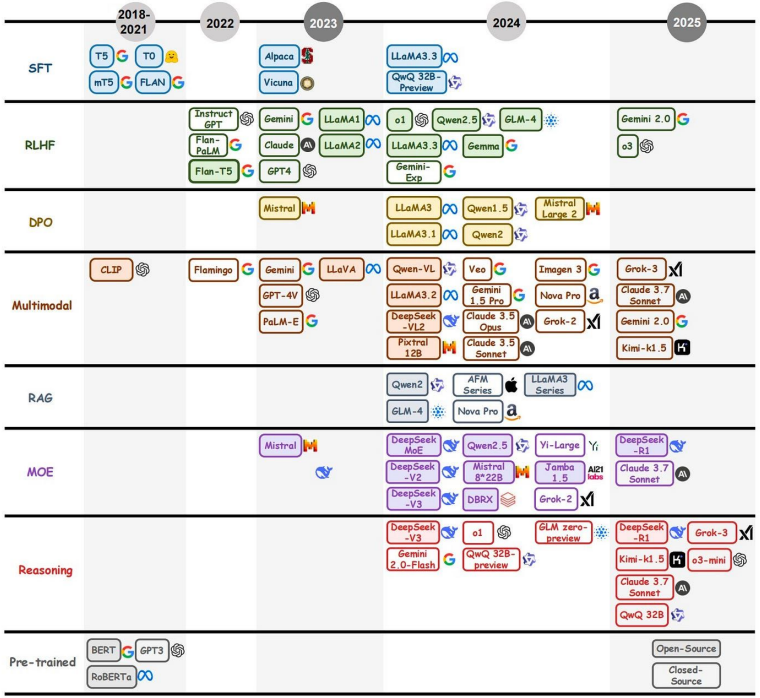
Outline

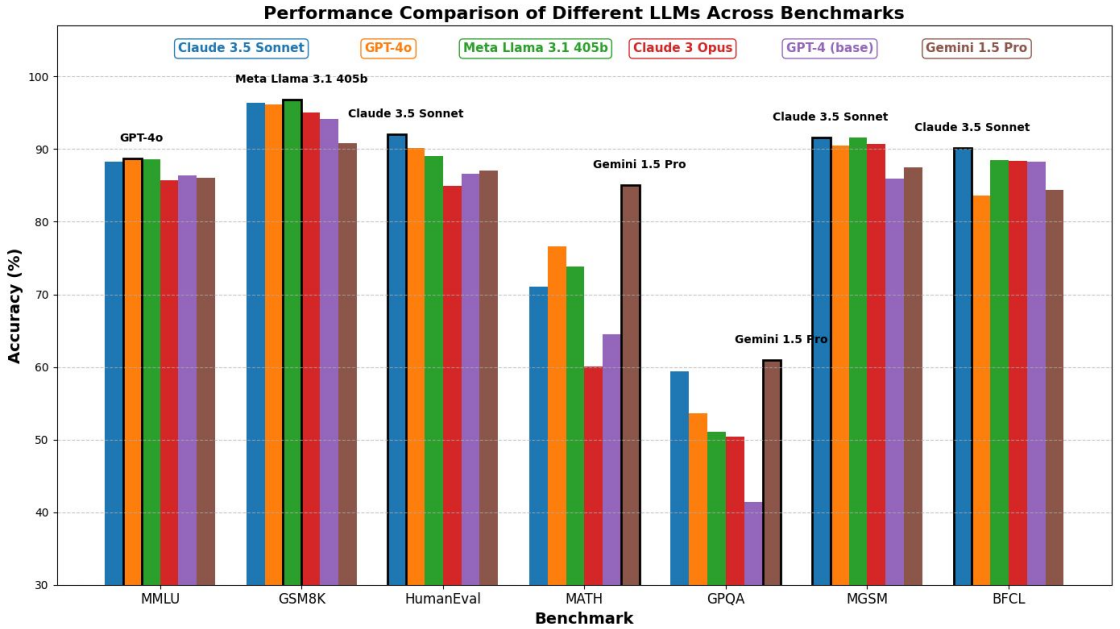
- ❏ Introduction
- ❏ Current State of Evaluation
- ❏ Key Challenges
- ❏ Path Forward
- ❏ Conclusion

Outline

- ❏ Introduction
- ❏ Current State of Evaluation
- ❏ Key Challenges
- ❏ Path Forward
- ❏ Conclusion

Generative AI models (aka. LLMs) have been impressive!





LLMs Performance on Diverse Benchmarks

Ref: (1) Left Image Credit: bit.ly/44nr4zx (2) Right Image: Accuracy scores are obtained from https://blog.promptlayer.com/llm-benchmarks/?utm_source=chatgpt.com

Impact in Educational Domain: Raise of AI Tutors (aka. ITSs)

Khanmigo

GPT-4o

LLaMA

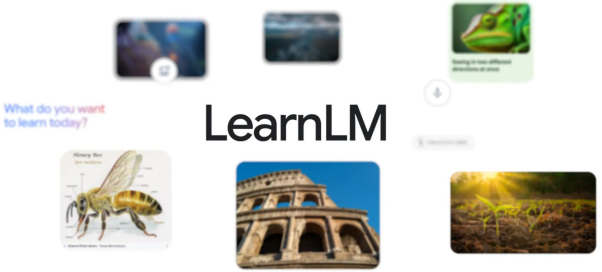
Gemini

Grok

Press Release

数学的性能で世界１位の評価を記録した、
数学特化の生成AIモデル「MathGPT」において、
Mathpresso, Inc.が日本語完全対応版の開発着手を発表

QANDA



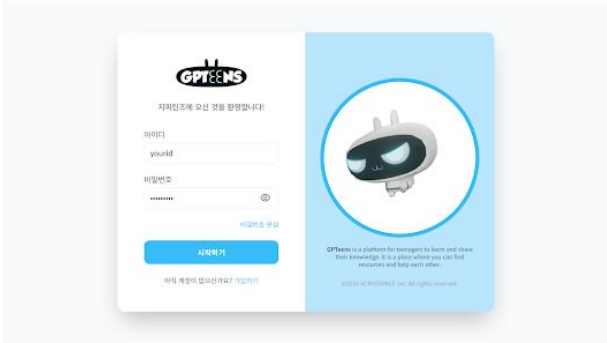


Image Credit: Google

Implications in Educational Domain: Can we trust them?

George W. Bush

Problem (In 2002): American students do not stand high in education globally
NCLM Act: Attempt to fulfill the expectations
Pros: Standardized test (equal opportunity), higher test scores
Cons: Standardized test scores led to shallow learning, reduced engagement, lots of expectations from teachers

Generative AI Can Harm Learning (Süngü, A., et al. 2024)

Setup:

- Pre-registered, Randomized Controlled Trial
- 9-11th grades in Turkey, 2023-2024 Fall semester
- Fifty, 90-minute classes
- Nearly 1000 students with 15% of the math curriculum
- Models: ChatGPT Base, GPT Tutor (ChatGPT + Prompts), GPT4

Outcome:

- With GPT4, 127% improvement over GPT Tutor
- 17% reduction in performance without tutor access

June 2025

Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task△

Nataliya Kosmyna [1]
MIT Media Lab
Cambridge, MA

Eugene Hauptmann
MIT
Cambridge, MA

Ye Tong Yuan
Wellesley College
Wellesley, MA

Jessica Situ
MIT
Cambridge, MA

Xian-Hao Liao
Mass. College of Art and Design (MassArt)
Boston, MA

Ashly Vivian Beresnitzky
MIT
Cambridge, MA

Iris Braunstein
MIT
Cambridge, MA

Pattie Maes
MIT Media Lab
Cambridge, MA

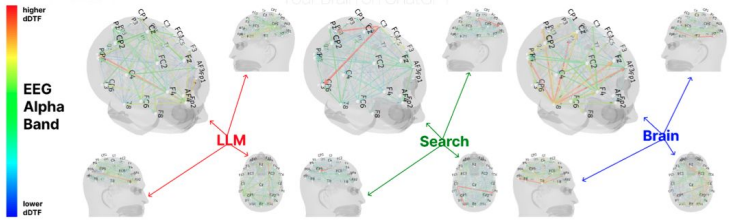
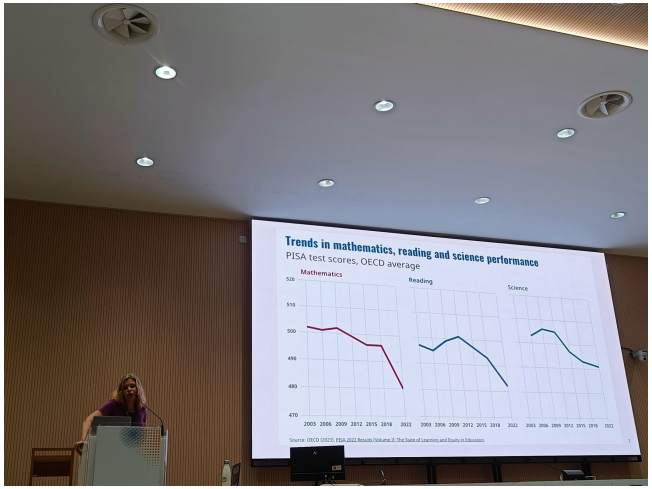
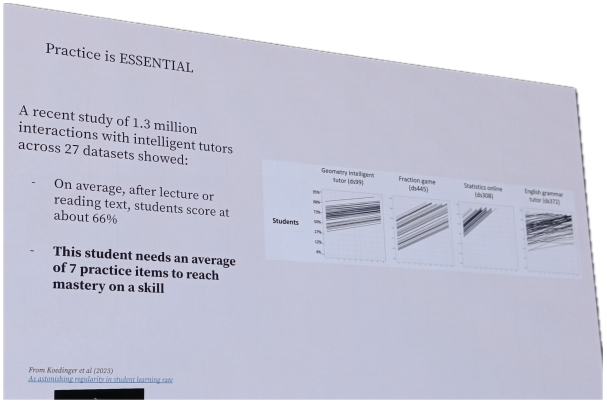


Image Credit: https://www.edweek.org/policy-politics/no-child-left-behind-an-overview/2015/04 https://www.slideshare.net/slideshow/no-child-left-behind-ppt/7766134#2

Implications in Educational Domain: Can we trust them?





Root Causes: Unreliable and Inconsistent Evaluation Practices

Large Exploration Space:

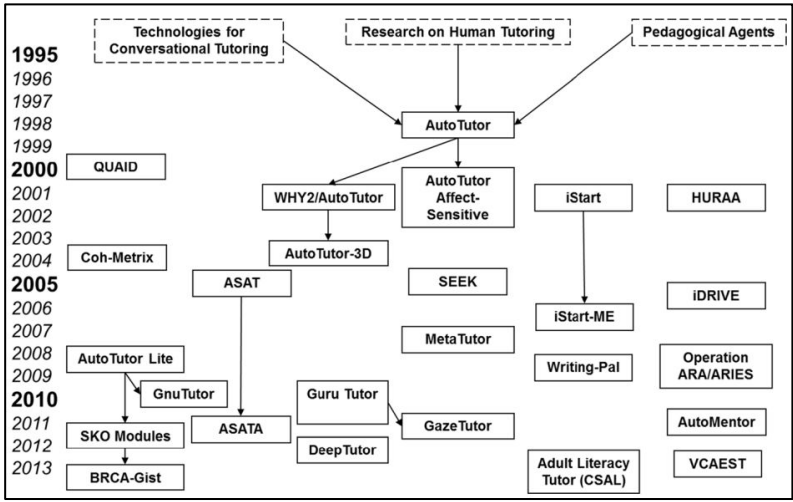
- Educational research is *multidisciplinary*
- Many learning theories, *limited RCTs*
- Many tutor moves, *little clarity* on which lead to learning gains
- *Personalized* protocols and benchmarks with *homogeneous* populations
- Evaluation and guidelines are *subjective* or *inconsistent*
- *Generic metrics* used to evaluate pedagogical capabilities
- Models and research are centered in fewer (or WEIRD) countries — l*ess inclusive* and *less adaptive*
- Many more

Outline

- ❑ Introduction
- ❑ Current State of Evaluation
- ❑ Key Challenges
- ❑ Path Forward
- ❑ Conclusion

Current State of Evaluation: Traditional Approaches

- Assessing teachers' practices [2]
  - Analysis of classroom artifacts (teacher' assignments and student work)
  - Teaching portfolios, self-reports, logs, interviews, etc.
- Role-simulation reports & extrinsic studies [2]
- Self-assessment with self-regulatory learning theories [3]



AutoTutor and family [1]

Limitations [4]:

- Not applicable to AI responses due to static nature of evaluation
- Not-scalable
- Hard to handle ethical and biased AI responses

Ref: [1] Nye, Benjamin D., Arthur C. Graesser, and Xiangen Hu. "AutoTutor and family: A review of 17 years of natural language tutoring." International Journal of Artificial Intelligence in Education 24.4 (2014): 427-469.
[2] Goe, L., Bell, C., Little, O.: Approaches to Evaluating Teacher Effectiveness: A Research Synthesis. National Comprehensive Center for Teacher Quality (2008)
[3] Crossley, S.A., Varner, L.K., Roscoe, R.D., McNamara, D.S.: Using Automated Indices of Cohesion to Evaluate an Intelligent Tutoring System and an Automated Writing Evaluation System. AIED 2013.
[4] Macina, J., Daheim, N., Wang, L., Sinha, T., Kapur, M., Gurevych, I., Sachan, M.: Opportunities and Challenges in Neural Dialog Tutoring. In: Vlachos, A., Augenstein, I. (eds.) EACL 2024.

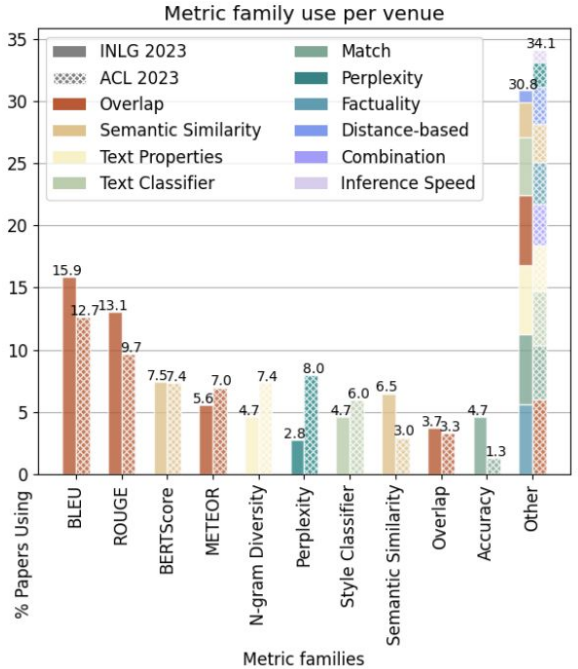
Current State of Evaluation: NLG-based Evaluation

NLG-based Metrics for Gen-AI ITS [1,2]

- Lexical and semantic match
- Validate coherence, fluency, human-likeness, etc.
- Ex: BLEU, BERTScores, etc.

Limitations [3]:

- Do not capture the pedagogical values
- Require a gold reference (often unavailable or non-unique)
- Can be hacked with superficial responses [4]



Ref: [1] Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. Automatic Metrics in Natural Language Generation: A survey of Current Evaluation Practices. INLG 2024

[2] Tack, A., Kochmar, E., Yuan, Z., Bibauw, S., Piech, C.: The BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues. In: Proceedings of BEA 2023. pp. 785–795 (2023)

[3] Jurenka, I., Kunesch, M., McKee, K.R., Gillick, D., Zhu, S., Wiltberger, S., Phal, S.M., Hermann, K., et al.: Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach. arXiv. 2024.

[4] asselli, J., Vasselli, C., Nohejl, A., Watanabe, T.: NAISTeacher: A Prompt and Rerank Approach to Generating Teacher Utterances in Educational Dialogues. In: Proceedings of BEA 2023. pp. 772–784 (2023)

Current State of Evaluation: Pedagogically Oriented

Human experts to evaluate pedagogical performance [1,2]:

- Better correlation with user satisfaction

Limitations [1,2,3]:

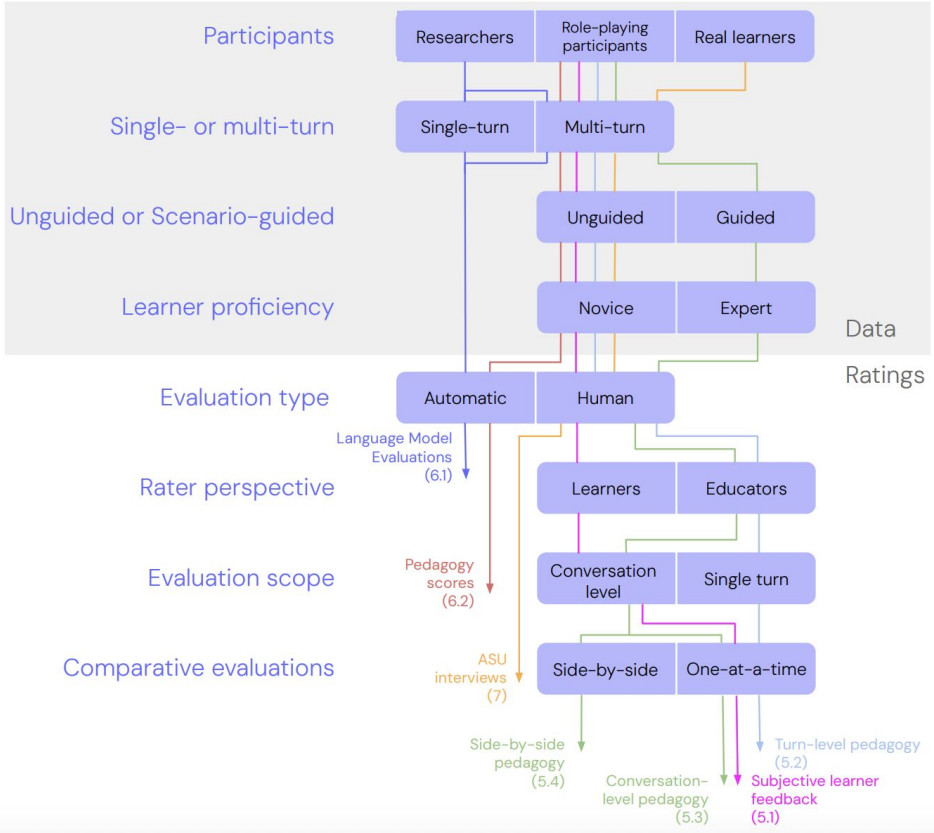
- Limited access to pedagogical experts
- Small number of annotators, leading to biases
- No unified protocol for evaluation
- Paid raters act as learners
- Evaluations with real students done in small number of participants - not scalable

Ref: [1] Tack, A., & Piech, C. (2022). The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues. arXiv preprint arXiv:2205.07540.
[2] Jurenka, I., Kunesch, M., McKee, K.R., Gillick, D., Zhu, S., Wiltberger, S., Phal, S.M., Hermann, K., et al.: Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach. arXiv. 2024.
[3] Alhafni, Bashar, Sowmya Vajjala, Stefano Bannò, Kaushal Kumar Maurya, and Ekaterina Kochmar. "LIms in education: Novel perspectives, challenges, and opportunities." COLING 2025.

Outline

- ❑ Introduction
- ❑ Current State of Evaluation
- ❑ Key Challenges
- ❑ Path Forward
- ❑ Conclusion

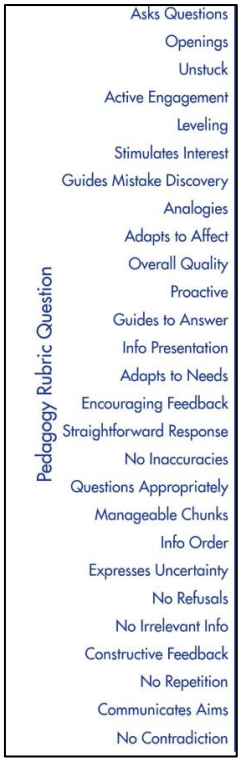
Key Challenges: Diverse Pragmatics of Evaluation



Ref: Jurenka, I., Kunesch, M., McKee, K.R., Gillick, D., Zhu, S., Wiltberger, S., Phal, S.M., Hermann, K., et al.: Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach. arXiv. 2024.

Key Challenges: Lack of Unified Pedagogical Practices

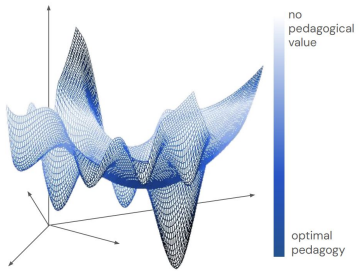
- Koedinger et al. (2013): Synthesized a list of 30 independent instructional principles
- Tack and Piech (2022): Proposed 3 strategies for good tutor
- Jurenka et al. (2024): Collected 28 strategies from various educational stakeholders



Teacher	Yes, good! And to charge it up, you need to __ it ___
Student	...
Teacher	connect to the source of electricity
Student	i understand
Teacher	plug it __?
Student	in

Responses

		likely said by a teacher	understanding the student	helping the student
Teacher	yes, good. And when the battery is full, you need to ____ (disconnect it)	1.769	1.632	1.632
GPT-3	yes!	0.116	-0.071	-0.071
Blender	do you have any pets? I have a dog and a cat. I love them	-1.829	-1.554	-1.554



Ref: [1] Koedinger, Kenneth R., Julie L. Booth, and David Klahr. "Instructional complexity and the science to constrain it." *Science* 342.6161 (2013): 935-937.
[2] Jurenka, I., Kunesch, M., McKee, K.R., Gillick, D., Zhu, S., Wiltberger, S., Phal, S.M., Hermann, K., et al.: Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach. arXiv. 2024.
[3] Tack, A., & Piech, C. (2022). The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues. arXiv preprint arXiv:2205.07540.

Key Challenges: Lack of Unified Evaluation Benchmarks

Dataset	Synthetic?	Domain	#Dialogues	#Moves	Grounding	Setting
CIMA	No	Language	391	5	Image, answer	1:1
Bridge	No	Math	700	10	Image, Confusion	1:1
MathDial	Yes	Math	2861	4	Confusion, answer	1:1
TSCC	No	Language	102	5	None	1:1
TalkMoves	No	Science	567	10	None	Classroom
NCTE	No	Math	1660	None	None	Classroom
MRBench	Mixed	Math	500	10	Confusion, answer	1:1

Outline

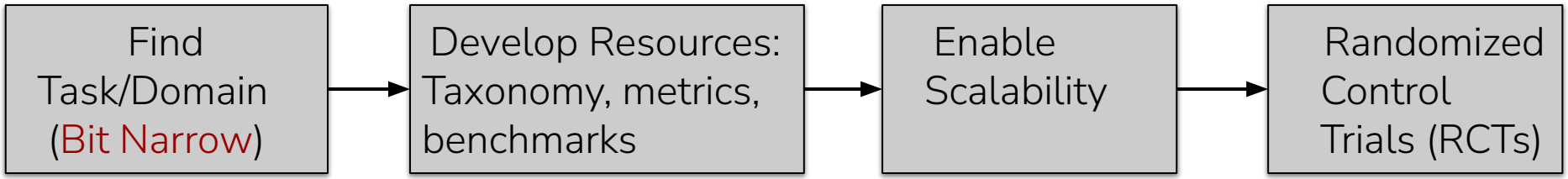
- ❑ Introduction
- ❑ Current State of Evaluation
- ❑ Key Challenges
- ❑ Path Forward
- ❑ Conclusion

Path Forward I : Evaluation Unification (Our Vision)

Research Hypothesis: Lack of task/domain-specific unified evaluation limits progress in ITSs.

Research Statement: Develop unified evaluation taxonomies, metrics, and benchmarks tailored to specific tasks/domains to evaluate ITSs.

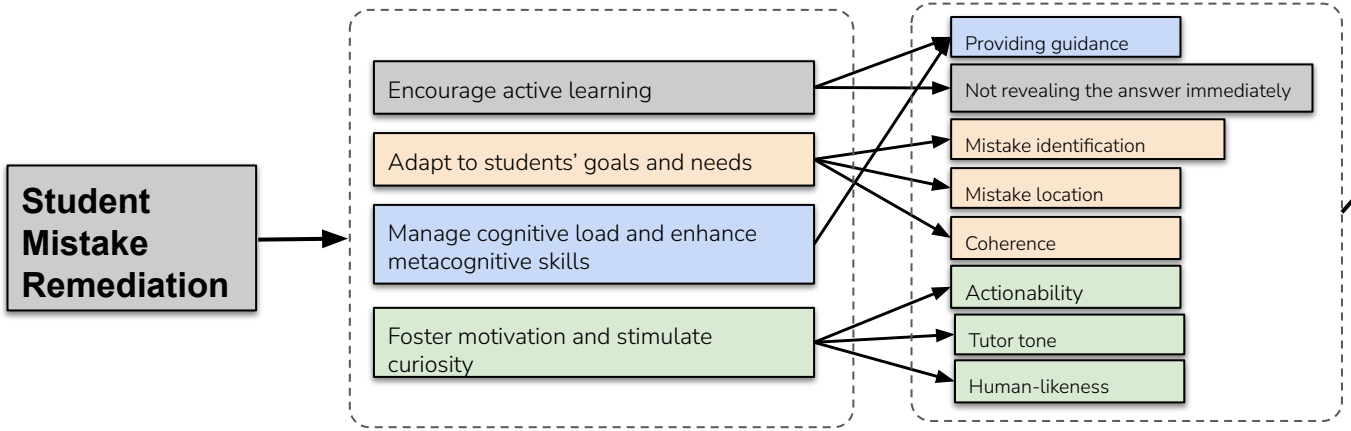
Potential Approach:



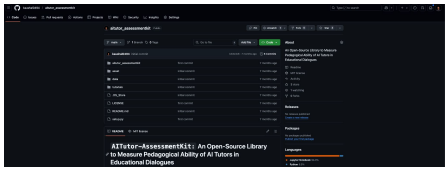
Extensions: Multi-Agent Systems

Evaluation Unification: Case Study I

Task | Learning Science Principles | Computational Dimensions | Outcome(s)



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	✓	✗	✓	✗	✓

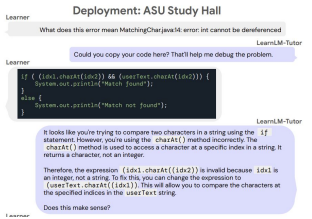


Not Addressed: RCTs

Ref: Maurya, Kaushal Kumar, K. V. Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. "Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors." NAACL 2025.

Evaluation Unification: Case Study II

Task	Learning Science Principles	Outcome(s)
Stay on topic	Manage cognitive load	Task-specific evaluation benchmarks annotated by humans + automated evaluation directions
Do not reveal the answer and guide towards the answer	Encourage active learning	
Adapt to the learner's level	Adaptivity	
⋮	⋮	



Not Addressed: Scalability

Ref: Jurenka, I., Kunesch, M., McKee, K.R., Gillick, D., Zhu, S., Wiltberger, S., Phal, S.M., Hermann, K., et al.: Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach. arXiv. 2024.

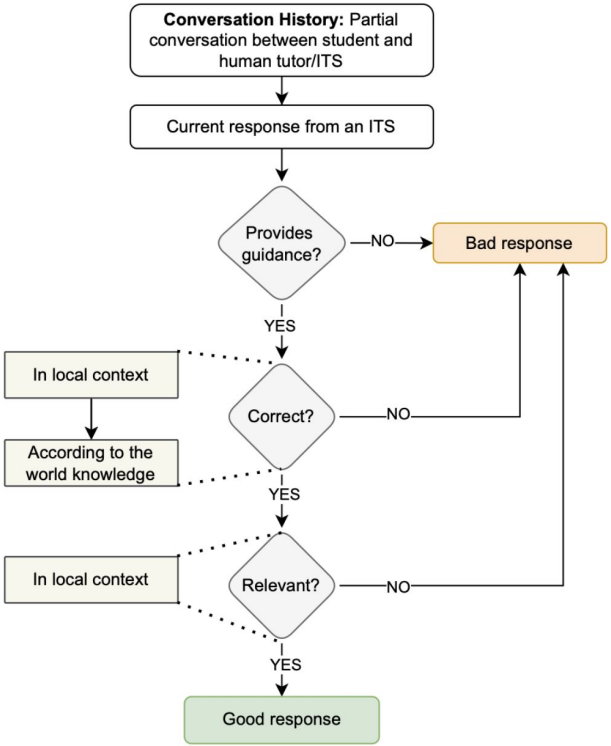
Path Forward II: Measuring Pedagogical Guidance

Research Hypothesis: By offering timely and context-aware appropriate guidance (e.g., hints, explanations, etc.), tutors can help students navigate their learning journey.

Research Statement: Develop a robust quantitative framework to assess the appropriateness and richness of the pedagogical guidance provided by GenAI-powered ITSs.

Grounded on: Metacognition

Potential Approach:



Ref: Dehaene, S.: How we learn: Why brains learn better than any machine... for now. Penguin (2021)

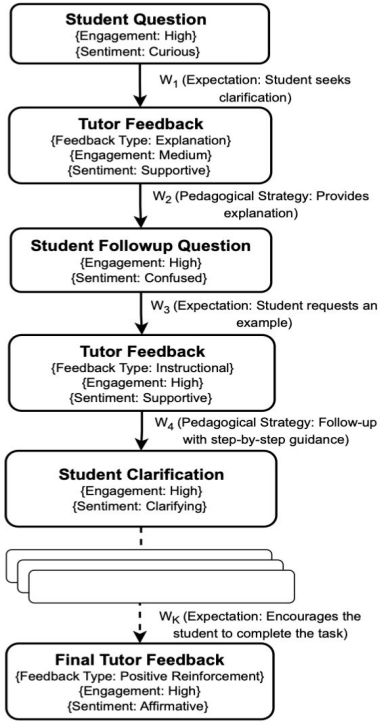
Path Forward III: Measuring Active Learning

Research Hypothesis: Active learning enables students to engage more deeply in the learning process through critical thinking and reflection.

Research Statement: Develop a robust quantitative framework to measure the active learning capabilities of GenAI-powered ITSs.

Grounded on: Constructivism and inquiry-based learning

Potential Approach:



Ref: [1] Piaget, J.: The theory of stages in cognitive development. McGraw-Hill (1972) [2] Vygotsky, L.S.: Mind in society: The development of higher psychological processes. 1978

Outline

- ❑ Introduction
- ❑ Current State of Evaluation
- ❑ Key Challenges
- ❑ Path Forward
- ❑ Conclusion

Conclusion

- ITS evaluation is challenging due to the diverse pragmatics of evaluation.
- Current state-of-the-art evaluation approaches are unreliable, inconsistent, or subjective.
- The community should work towards a pedagogy-oriented evaluation framework that is unified and grounded in learning science principles.

Paper

Let us know your thoughts!!!

Correspondence:
kaushal.maurya@mbzuai.ac.ae