Findings of the BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-powered Tutors

Ekaterina Kochmar,[1] Kaushal Kumar Maurya,[1] Kseniia Petukhova,[1]
KV Aditya Srivatsa,[1] Anaïs Tack,[2] and Justin Vasselli[3]

[1]MBZUAI, [2]KU Leuven, [3]Nara Institute of Science and Technology

Mohamed bin Zayed University of Artificial Intelligence

ACL 2025 VIENNA

Introduction

How can we test whether state-of-the-art generative models are good AI tutors, capable of replying appropriately to a student in an educational dialogue?

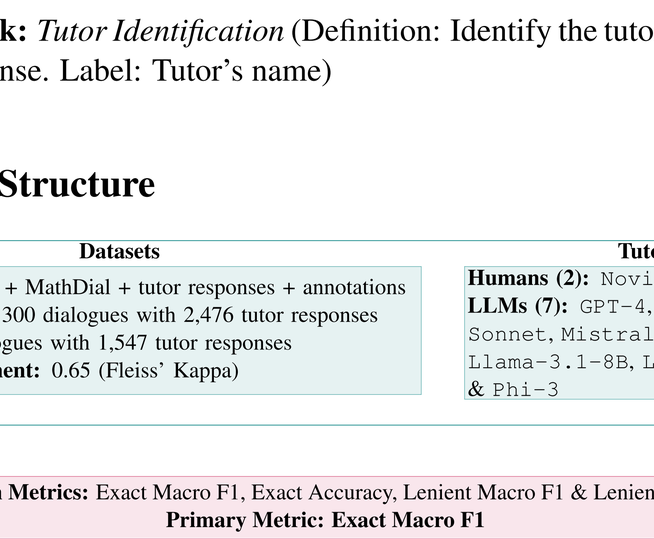
Built upon the evaluation framework introduced by Maurya et al. [1]:

Evaluation Dimension	TP'22	MA'23	WA'24	DA'24	Ours	Learning Science Principle
Mistake Identification	✓	✓	✗	✓	✓	Adapt to students' goals
Mistake Location	✗	✗	✗	✓	✓	Adapt to students' goals
Revealing of the Answer	✗	✓	✗	✗	✓	Encourage active learning
Providing Guidance	✓	✗	✓	✗	✓	Manage cognitive load
Actionability	✗	✗	✗	✓	✓	Foster motivation and curiosity
Coherence	✗	✓	✗	✗	✓	Adapt to students' goals
Tutor Tone	✓	✗	✓	✗	✓	Foster motivation and curiosity
Human-likeness	✓	✗	✓	✗	✓	Foster motivation and curiosity

Limitation: Hard to *scale* and *adapt*, since the evaluation depends on human annotation for each dimension.

Open Question: Can we develop reliable automated metrics for each dimension?

Shared Task Tracks



Teams were invited to evaluate the pedagogical appropriateness of the tutor's current response by developing systems across the following dimensions:

Track	Definition	Label
Mistake Identification	Has the tutor identified/recognized a mistake?	Yes To some extent No
Mistake Location	Does the tutor point to a genuine mistake and its location?	
Providing Guidance	Does the tutor offer correct and relevant guidance, such as an explanation, elaboration, hint, examples, etc.?	
Actionability	Is it clear what the student should do next?	

Additional Track: *Tutor Identification* (Definition: Identify the tutor who originated the current response. Label: Tutor's name)

Shared Task Structure

Datasets

MRBench: Bridge + MathDial + tutor responses + annotations
Development Set: 300 dialogues with 2,476 tutor responses
Test Set: 200 dialogues with 1,547 tutor responses
Annotator Agreement: 0.65 (Fleiss' Kappa)

Tutors

Humans (2): `Novice` & `Expert`
LLMs (7): `GPT-4`, `Gemini`, `Sonnet`, `Mistral`, `Llama-3.1-8B`, `Llama-3.1-405B` & `Phi-3`

Evaluation Metrics: Exact Macro F1, Exact Accuracy, Lenient Macro F1 & Lenient Accuracy
Primary Metric: Exact Macro F1

Development Phase

Data: Development set
Duration: 12 March – 10 April 2025

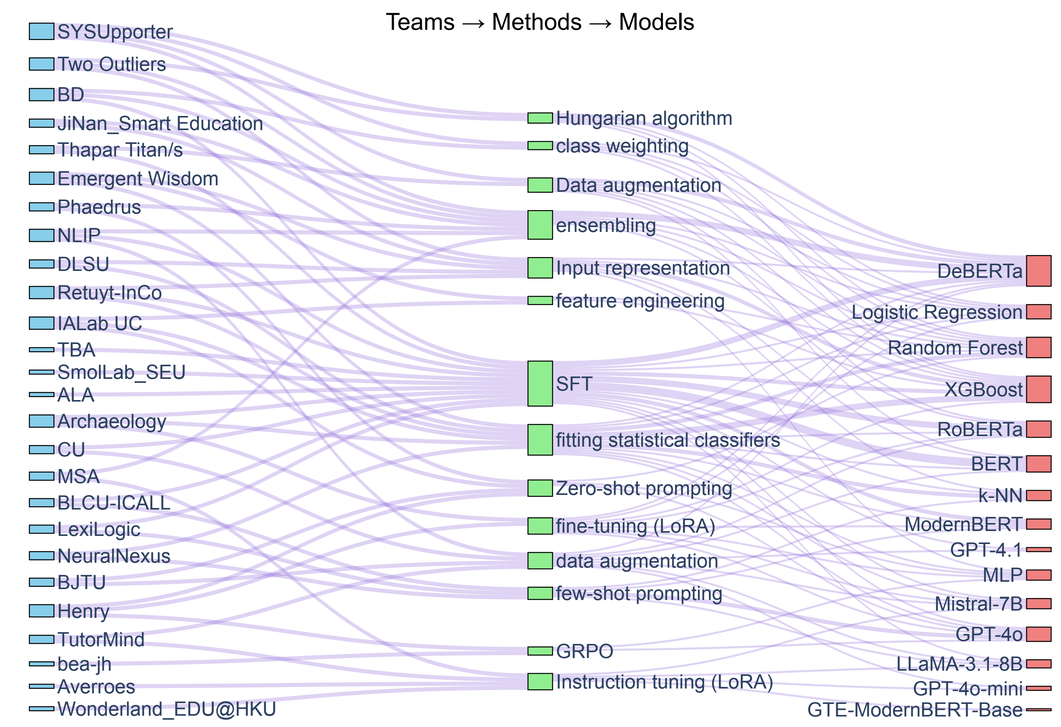
Test Phase

Data: Test set
Duration: 10 April – 24 April 2025

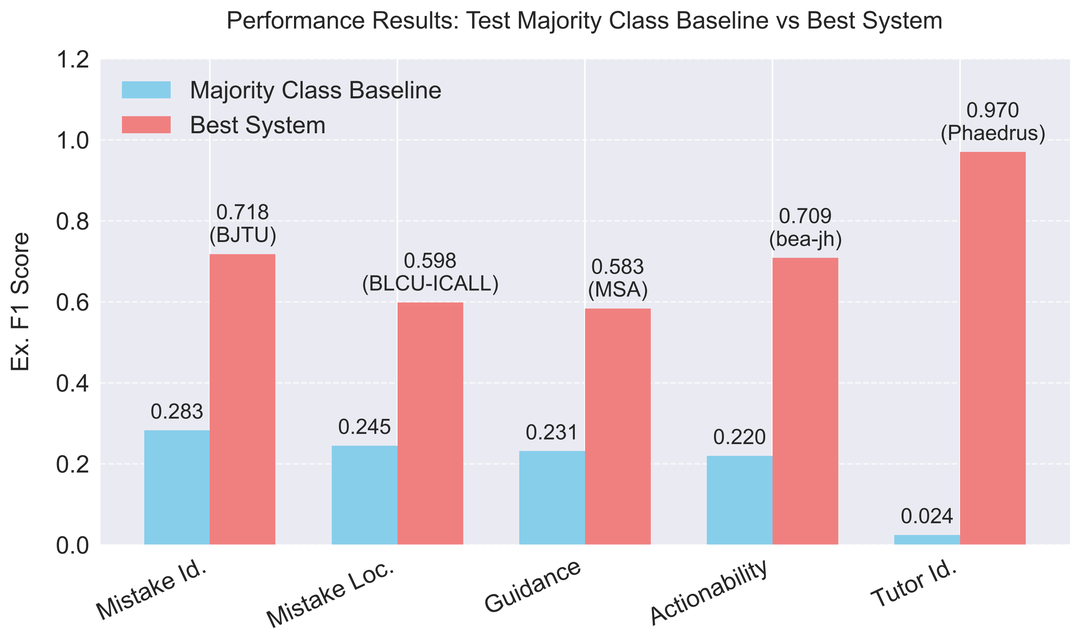
Submission: Via CodaBench for Final Test Phase

Track	# Submissions	# Teams
Mistake Identification	153	44
Mistake Location	86	32
Providing Guidance	105	36
Actionability	87	30
Tutor Identification	54	20

Teams and Methodologies



Results and Observations



Track	Team (Rank)	Approach
MI	BJTU (1)	Zero-shot prompts with data augmentation
	TutorMind (2)	LoRA-tuned LLMs with synthetic data augmentation
	Averroes (3)	Fine-tune eight instruction-tuned LLMs
ML	BLCU-ICALL (1)	ICL with `Gemini-2.5-pro`
	BJTU (2)	Zero-shot prompts with data augmentation
	K-NLPers (3)	GPT-4.1 with reflective prompting
PG	MSA (1)	LoRA-tuned Mathstral-7B with ensemble disagreement
	SG (2)	`Gemma-3-27B-IT` with multi-step prompting
	BLCU-ICALL (3)	ICL with `Gemini-2.5-pro`
AC	bea-jh (1)	GRPO-trained `GLM-4-9B` outputs tagged rationales and answers
	BJTU (2)	Zero-shot prompts with data augmentation
	MSA (3)	LoRA-tuned Mathstral-7B with ensemble disagreement
TI	Phaedrus (1)	Ensembled LLMs with token cues and greedy constraint-based post-processing
	SYSUpporter (2)	Synthetic noise, class-weighted loss, and Hungarian algorithm
	Two Outliers (3)	DiReC separates content/style with contrastive learning, using CatBoost and Hungarian matching for tutor ID

Future Directions

- Distinguishing subtle pedagogical quality
- Handling diverse tutor styles
- Adapting to new domains

Acknowledgment

This research is partially supported by Google through the Google Academic Research Award (GARA) 2024.

References

[1] Kaushal Kumar Maurya, KV Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *NAACL 2025*, pages 1234–1251.