BJTU at BEA 2025 Shared Task: Task-Aware Prompt Tuning and Data Augmentation for Evaluating AI Math Tutors

Yuming Fan and **Chuangchuang Tan** * and **Wenyu Song**
20120300@bjtu.edu.cn, 21112002@bjtu.edu.cn,
20120313@bjtu.edu.cn
Beijing Jiaotong University

Abstract

We describe the BJTU submission to the BEA 2025 Shared Task on Evaluating the Pedagogical Ability of AI Tutors, which focuses on assessing AI-generated math tutoring responses across four dimensions: Mistake Identification, Mistake Location, Guidance, and Actionability. Our approach leverages a large language model (LLM) with task-specific prompt tuning and data augmentation techniques, including dialogue shuffling and class balancing. The system achieves strong results across all tracks, ranking first in Mistake Identification and performing competitively in the others. Our findings underscore the potential of prompt-based LLMs for pedagogically-aware response evaluation and offer insights into the design of AI tutors with improved educational feedback.

1 Introduction

Recent advances in large language models (LLMs) have opened up new possibilities in education, with AI-powered tutoring systems emerging as promising tools for personalized learning. These systems simulate teacher-like interactions through natural language dialogue, offering students real-time feedback and instructional support. However, evaluating the teaching capabilities of such AI tutors remains a significant challenge. On the one hand, existing evaluation frameworks lack standardization. Previous work adopts fragmented criteria, such as correctness, relevance, and actionability, making it difficult to compare model performance between studies.

However, conventional automatic metrics (e.g. ROUGE, BLEU) fail to capture key educational goals, such as effective knowledge delivery, error correction, and cognitive scaffolding. For example, Tack (Tack and Piech, 2022) focus on teacher language style, Macina (Macina et al., 2023) highlights the coherence of feedback, while Wang (Wang et al., 2024) emphasizes the empathetic tone of responses. This fragmented landscape hinders the development of standardized benchmarks for educational AI.

To address the above challenges, the BEA 2025 Shared Task(Kochmar et al., 2025) on Evaluating the Pedagogical Ability of AI Tutors introduces the first multidimensional benchmark centered on instructional competence. Our team, Team BJTU, focuses on the context of mathematics education, particularly the process of error remediation. We aim to develop automated models that systematically evaluate five core capabilities of AI tutoring systems: Mistake identification (identifying whether a student's response contains a mistake), mistake location (pointing to the exact location of the error), guidance (offering effective explanations or hints) and Actionability (providing responses that meaningfully guide the student's next learning steps).

Our team, BJTU, participated in Tracks 1, 2, 3, and 4 of the BEA 2025 Shared Task and achieved strong results, ranking 1st, 2nd, 4th, and 2nd respectively. Our approach leverages state-of-the-art language models that integrate textual cues to explore the instructional capabilities of AI tutors across multiple pedagogical dimensions. This paper outlines our methodology for tackling the task, discusses the challenges we encountered, and provides insight into how model design choices impact the effectiveness of AI-generated feedback in educational dialogues.

2 Related Work

Recent work has explored the use of large language models (LLMs) in educational dialogues, with the aim of assessing their pedagogical effectiveness. Tack and Piech (Tack and Piech, 2022) proposed the AI Teacher Test, evaluating models such as GPT-3 and Blender in three dimensions: speaking

*Corresponding Author.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1073–1077

like a teacher, understanding the student and providing helpful responses. Their findings showed that while LLMs produce fluent dialogue, they lack pedagogical ability.

Building on this, the BEA 2023 Shared Task (Tack et al., 2023) benchmarked teacher response generation using the TSCC dataset. Top systems used models such as GPT-3.5 and GPT-4, employing prompting and response reranking strategies. Although some systems achieved high scores, the task highlighted limitations in existing evaluation metrics for educational settings.

To address these gaps, Wang (Wang et al., 2024) introduced Bridge, a framework based on cognitive task analysis that models expert decision-making during remediation. Incorporating these decisions into LLM prompts significantly improved response quality, suggesting that structured pedagogical reasoning enhances LLM performance in tutoring contexts.

3 Method

3.1 Preprocessing

During the data preprocessing phase, we organized the historical dialogues between the Tutor and Student into a format suitable for fine-tuning. For each instance, we constructed prompts such as: *The following is a tutoring dialogue in the domain of mathematics. Based on the conversation history above, your task is to evaluate the following Tutor's Response and determine whether it successfully identifies the error in the student's reasoning*, as illustrated in Figure 1. Using this data, we fine-tuned a large language model (LLM) to perform the evaluation task.

In the testing phase, we applied the trained model to the test set for inference. The LLM was prompted to generate an evaluation of the given Tutor response and select one of three categorical labels—yes, some extent, or no—which was then recorded as the final output.

However, relying solely on the original training data risks overfitting the model to specific linguistic patterns, thereby limiting its generalization ability. To address this, we incorporated a series of data augmentation strategies aimed at improving the model's robustness and adaptability across diverse dialogue contexts.

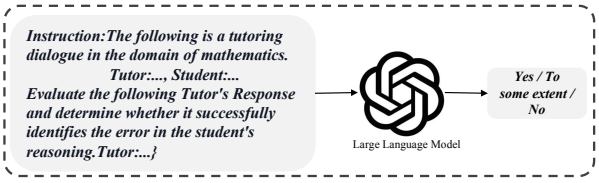

Figure 1: Prompt Construction.

3.2 Data Augmentation

In the shared task, our team BJTU demonstrated strong performance across all four tracks, as shown in table 2. We used the data set released for the BEA 2025 Shared Task (Maurya et al., 2025), which is based on a unified taxonomy for assessing pedagogical ability.

To mitigate the model's reliance on fixed option positions and enhance its ability to generalize in ranking tasks, we adopted a dialogue-shuffling augmentation strategy. Concretely, we randomly permuted the sequence of tutor-student interaction pairs within each dialogue instance. This allows the model to better learn from the full instructional process provided by the tutor, rather than becoming overly dependent on a particular response order. By disrupting positional regularities, the model is encouraged to attend to the actual content of the tutor's guidance. Moreover, since the dataset comprises tutoring interactions from multiple distinct AI tutors, shuffling further reduces the risk of overfitting by limiting memorization of stylistic patterns.

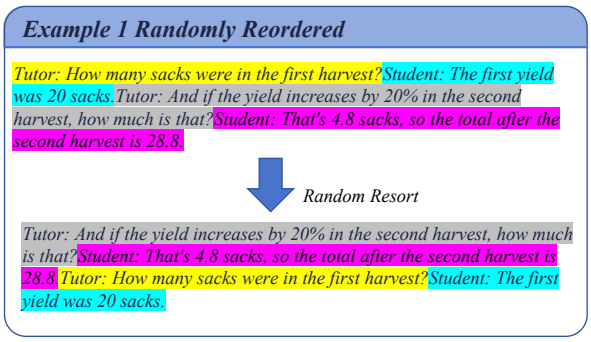

Figure 2: Randomly reordered method.

To address the issue of class imbalance observed in the training data, we applied targeted data augmentation strategies to improve model generalization. As shown in Table 1, all four subtasks exhibit a significant skew toward the "Yes" class, with notably fewer examples labeled as "To Some Extent" or "No." This imbalance can lead the model to

overfit to the majority class and hinder its ability to accurately recognize minority class instances.

Task	Yes	To Some Extent	No
Mistake Identification	1932	174	370
Mistake Location	1543	220	713
Provide Guidance	1407	503	556
Actionability	1310	369	797

Table 1: Label distribution across the four subtasks.

To mitigate this, we implemented random downsampling for the "Yes" instances. Specifically, we randomly sampled half of the 'Yes' instances, while all "No" and "To Some Extent" instances were preserved. This simple yet effective strategy reduced the dominance of the majority class and encouraged the model to better capture the characteristics of less frequent classes.

In addition, we introduced a lightweight prompt engineering strategy to improve the model's awareness of the task objective. Taking the Mistake Identification task as an example, where the objective is to determine whether the tutor's response successfully identifies an error in the student's reasoning, we attached an explicit task instruction to the input. Specifically, the complete prompt template as follows: *The student's last utterance contains a mistake. The AI tutor responds to this mistake. Your task is to assess whether the tutor's response successfully identifies the mistake made by the student..... Your task is to evaluate the following tutor responsend determine whether it successfully identifies the error in the student's reasoning.* This additional context helps guide the momodel's attention toelevant reasoning errors in the dialogue. Although the modification is simple, empirical results suggest that such task-aware prompts can improve model performance, highlighting the importance of clear task framing in multi-choice dialogue understanding tasks.

4 Experiment Results

We employed the Qwen2.5 (Bai et al., 2023) model series as the backbone and trained our models using the dataset constructed in the Method section. Specifically, we conducted training and inference using four Ascend-910B nodes, each equipped with eight GPUs. The learning rate was set to 5e-6, the gradient accumulation steps were configured as 8, and the models were trained for a total of five epochs.

For Mistake Identification, BJTU secured 1st place with an exact macro F1 score (Ex. F1) of 0.7181, indicating its effectiveness in accurately identifying errors in student responses. In the Mistake Location track, BJTU ranked 2nd with an Ex. F1 score of 0.5940, demonstrating its ability to locate errors in student reasoning. For Providing Guidance, BJTU placed 4th with an Ex. F1 score of 0.5725, reflecting its solid performance in selecting appropriate guidance responses from multiple options. In the Actionability track, BJTU again showed strong results, ranking 2nd with an Ex. F1 score of 0.6992, demonstrating its capability to determine the practical applicability of the responses. These results highlight the consistency and versatility of BJTU's system across different task domains, proving its robustness in handling various aspects of educational dialogue systems.

We adopted a unified strategy across all four tracks, as the tasks share similar objectives centered on evaluating and improving AI tutor responses. Instead of building separate models, we applied the same framework and prompt design to each task, which simplified our approach and proved effective across different evaluation aspects.

To further evaluate the effectiveness of different augmentation strategies, we conducted an ablation study comparing several variants of the model on the Codabench. The results are summarized in Table 3. Among all configurations, the combination of task description and dialogue shuffling achieved the best strict macro F1 score (0.7181), suggesting that explicitly describing the task helps the model better align its generation with the intended objective.

When applying shuffling alone, the model obtained the highest strict accuracy (0.8694), indicating improved precision in certain classes. However, its slightly lower F1 score suggests a trade-off in class coverage. Introducing class balancing on top of shuffling led to a modest increase in strict F1 (0.7104), but did not produce consistent improvements across all metrics. This aligns with our hypothesis that label distribution reweighting offers limited benefit when the test set closely mirrors the training set.

The base model, which only uses prompt construction without augmentation, performed slightly worse overall but still maintained reasonable robustness. These findings highlight that prompt design alone plays an important role and that combining shuffling with task description provides the most

Track	Team	Ex. F1	Ex. Acc	Len. F1	Len. Acc
1. Mistake Identification	BJTU	0.7181	0.8623	0.8957	0.9457
	TutorMind	0.7163	0.8759	0.9108	0.9528
	Averroes	0.7155	0.8675	0.8997	0.9425
	MSA	0.7154	0.8759	0.9152	0.9535
	BD	0.7110	0.8772	0.8966	0.9412
2. Mistake Location	BLCU-ICALL	0.5983	0.7679	0.8386	0.8630
	BJTU	0.5940	0.7330	0.7848	0.8261
	K-NLPers	0.5880	0.7641	0.8404	0.8610
	MSA	0.5743	0.6975	0.7848	0.8209
	SG	0.5692	0.7602	0.8118	0.8400
3. Providing Guidance	MSA	0.5834	0.6613	0.7798	0.8190
	SG	0.5785	0.7052	0.7860	0.8216
	BLCU-ICALL	0.5741	0.6716	0.7487	0.8061
	BJTU	0.5725	0.6490	0.7445	0.8100
	K-NLPers	0.5606	0.6270	0.7446	0.8000
4. Actionability	bea-jh	0.7085	0.7298	0.8527	0.8837
	BJTU	0.6992	0.7363	0.8633	0.8940
	MSA	0.6984	0.7537	0.8659	0.8908
	lexiLogic	0.6930	0.7162	0.8393	0.8675
	Phaedrus	0.6907	0.7298	0.8346	0.8650

Table 2: Top-5 system performances for each subtask, ranked by exact macro F1 (Ex. F1). Secondary metrics include exact accuracy (Ex. Acc), lenient macro F1 (Len. F1), and lenient accuracy (Len. Acc).

Strategy	Strict Acc.	Lenient Acc.	Strict F1	Lenient F1
Base(Only prompt construction)	0.8604	0.9476	0.7030	**0.9026**
Shuffling + Class Balance	0.8565	**0.9483**	0.7104	0.9017
Shuffling only	**0.8694**	0.9444	0.6957	0.8984
Shuffling + Task describe	0.8623	0.9457	**0.7181**	0.8957

Table 3: Performance of different development runs under strict and lenient matching criteria.

notable gains across evaluation metrics.

These findings suggest that among the augmentation techniques we explored, randomizing the dialogue order is particularly effective in improving the robustness of the model in unseen examples. However, the benefit of class balancing appears to depend on whether there is a label distribution mismatch between training and test sets.

5 Conclusion

In this paper, we presented BJTU's approach to the BEA 2025 Shared Task on Evaluating the Ability of AI Tutors. Focusing on mathematics education, we designed a system that effectively evaluates tutor responses along four instructional dimensions: mistake identification, mistake location, guidance, and actionability. Our method leveraged large language models, prompt engineering, and targeted data augmentation techniques, including dialogue shuffling and class balancing, to enhance model generalization and robustness.

Our system achieved strong overall results, ranking within the top four in all tracks, including first place in Mistake Identification. These outcomes demonstrate the potential of well-structured prompting and augmentation strategies to improve the pedagogical evaluation capabilities of LLMs.

Looking forward, we aim to explore more fine-grained annotation schemes, incorporate multimodal feedback, and develop more interpretable evaluation models. We hope our findings contribute to the advancement of standardized and scalable benchmarks for AI-assisted education.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, and et al. 2023. Qwen technical report. *arXiv:2309.16609*.

Ekaterina Kochmar, Kaushal Kumar Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack, and Justin Vasselli. 2025. Findings of the bea 2025 shared task on pedagogical ability assessment of ai-powered tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv preprint arXiv:2305.14536.

Kaushal Kumar Maurya, KV Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors. In *Proceedings of the 2025 Conference of the North, Central and South American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251.

Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. The bea 2023 shared task on generating ai teacher responses in educational dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 785–795, Toronto, Canada. Association for Computational Linguistics.

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 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. ArXiv preprint arXiv:2205.07540.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024. 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.