SmolLab_SEU at BEA 2025 Shared Task: A Transformer-Based Framework for Multi-Track Pedagogical Evaluation of AI-Powered Tutors

Md. Abdur Rahman[1] **MD AL AMIN**[2*] **Sabik Aftahee**[3*]
Muhammad Junayed[3] **Md Ashiqur Rahman**[1]
[1]Southeast University, Dhaka, Bangladesh
[2]St. Francis College, Brooklyn, New York, USA
[3]Chittagong University of Engineering & Technology (CUET), Chittagong, Bangladesh
{2021200000025@seu.edu.bd, alaminhossine@gmail.com,
u1904024@student.cuet.ac.bd, u2008023@student.cuet.ac.bd,
ashiqur.rahman@seu.edu.bd}

Abstract

The rapid adoption of AI in educational technology is changing learning settings, making the thorough evaluation of AI tutor pedagogical performance is quite important for promoting student success. This paper describes our solution for the BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-powered tutors, which assesses tutor replies over several pedagogical dimensions. We developed transformer-based approaches for five diverse tracks: mistake identification, mistake location, providing guidance, actionability, and tutor identity prediction using the MRBench dataset of mathematical dialogues. We evaluated several pre-trained models including DeBERTa-V3, RoBERTa-Large, SciBERT, and EduBERT. Our approach addressed class imbalance problems by incorporating strategic fine-tuning with weighted loss functions. The findings show that, for all tracks, DeBERTa architectures have higher performances than the others, and our models have obtained in the competitive positions, including 9th of Tutor Identity (Exact F1 of 0.8621), 16th of Actionability (Exact F1 of 0.6284), 19th of Providing Guidance (Exact F1 of 0.4933), 20th of Mistake Identification (Exact F1 of 0.6617) and 22nd of Mistake Location (Exact F1 of 0.4935). The difference in performance over tracks highlights the difficulty of automatic pedagogical evaluation, especially for tasks whose solutions require a deep understanding of educational contexts. This work contributes to ongoing efforts to develop robust automated tools for assessing.

1 Introduction

In the past few years, the combination of natural language processing (NLP) and education technology has become one of the most popular areas of study to improve learning, automate feedback, and assist educators and students. With the expansion of blended and fully online courses, there has been a marked increase in the need for scalable and sophisticated systems that can process learners' responses and tutors' comments. Such systems do not deal well with the subtle, context-sensitive characteristics of educational dialogues. So, assessing the pedagogical effectiveness and a standard evaluation taxonomy of such systems still remains a critical challenge.

An example of effective teaching is when an educator accurately pinpoints a student's misunderstanding, provides appropriate scaffolding towards clear concepts, and gives insightful feedback on desk-work that the students need to accomplish. Some automating aspects of this feedback loop, such as automated essay scoring (Phandi et al., 2015) and dialogic tutoring systems (Wang et al., 2024) have been given attention, but there is not much research that has been done to effectively capture the dynamics of the interplay student answers, tutor's engagement, and teaching style through feedback text's narrative structure.

While LLMs can generate coherent and contextually relevant responses, their ability to understand student misconceptions, provide actual guidance, and create meaningful learning experiences is not guaranteed. The general, area-independent metrics for natural language generation (NLG) (Liu et al., 2023; Gao et al., 2020) do not fit here as the majority of them lack consideration for pedagogical values and need gold references, which are seldom present in online interactions.

In this work, we tackle a comprehensive multi-track evaluation task designed for the evaluation of AI-tutor responses using a set of pedagogically motivated metrics. Building upon the foundations laid by the BEA 2023 Shared Task (Tack et al., 2023), which focused on generating AI teacher responses in educational dialogues, in the BEA 2025 Shared Task (Kochmar et al., 2025) iteration the focus shifted toward evaluating the quality of AI tutor responses. Specifically, it introduced a

*Authors contributed equally to this work.

taxonomy encompassing four pedagogically motivated dimensions: Mistake Identification, Mistake Location, Providing Guidance, and Actionability. Additionally, a fifth track challenged participants to identify the source of anonymized tutor responses, distinguishing between various LLMs and human tutors.

Our key contributions are as follows:

- Developing transformer-based approaches for comprehensive evaluation of AI-tutor responses using a set of pedagogically motivated metrics: mistake identification, mistake location, guidance provision, feedback actionability, and tutor identity prediction.
- Evaluated the performance of state-of-the-art transformer models across five key educational NLP tasks related to tutoring dialogues.

2 Related Works

Daheim et al. (2024) introduced a framework for stepwise solution verification for math reasoning, showing that grounding tutor responses in identified errors improves feedback accuracy where AI tutors are evaluated on their ability to identify and locate mistakes within student responses. Macina et al. (2023) presented *MathDial*, a large dataset of tutoring dialogues where LLMs often struggle with correct mistake spotting without targeted annotations. It includes annotations for mistake locations in math dialogues. This resource has been instrumental in training and evaluating models that can accurately identify and address specific errors in student solutions. Chen et al. (2024) proposed VATE, an AI-driven virtual teacher using prompt engineering and error pools for autonomous mistake analysis, achieving high accuracy in real-world deployment. Lastly, Macina et al. (2024) benchmarked pedagogical capabilities of LLM tutors, confirming that subject knowledge alone doesn't ensure effective error identification without specialized pedagogical training. Additionally, Yan et al. (2024) propose architectures designed to improve error localization in multimodal math tutoring, enhancing the clarity and usefulness of feedback. Recent work in intelligent tutoring systems (ITS) has emphasized the importance of scaffolding and adaptive feedback to enhance student learning outcomes. Liu et al. (2024) explored multimodal tutoring systems powered by large language models, demonstrating how pedagogical instructions can improve self-paced learning through structured scaffolding, evaluated via a seven-dimension rubric. Complementing this, Kochmar et al. (2020) showed that automated, personalized feedback using NLP and machine learning significantly boosts student performance, highlighting the need for tailoring feedback to individual learners. Similarly, Li et al. (2024) applied NLP-driven adaptive dialogs informed by the Knowledge Integration framework, illustrating how guided conversations help students integrate accurate scientific concepts during instruction. Together, these studies underline the potential of adaptive, pedagogically-aware NLP systems in delivering effective, personalized guidance within educational contexts. Maniktala et al. (2020) proposed "Assertions," an unsolicited hint mechanism delivering partially-worked example steps, which notably increased hint usage and improved learning outcomes, particularly for lower-proficiency learners. Blancas-Muñoz et al. Blancas-Muñoz et al. (2018) further emphasized the importance of actionable support by comparing task-relevant hints to distractions in robotic tutoring, finding that direct, task-specific guidance led to better learner performance. Extending this focus to virtual education settings, Liang Liang (2025) applied NLP-based Seq2Seq models for automated feedback generation, achieving high accuracy while enhancing personalization and actionability of feedback in online environments. Collectively, these studies highlight that actionable, timely, and context-aware feedback mechanisms are essential for effective ITS design.

3 Task and Dataset Description

We competed on the BEA 2025 Shared Task[1] (Kochmar et al., 2025) on Pedagogical Ability Assessment of AI-powered tutors. The goal of the work is to assess AI tutor responses in mathematical dialogues when students make errors or show uncertainty. The provided dataset, MRBench (Maurya et al., 2025), includes dialogue contexts, the final student utterance, and corresponding tutor responses from various LLMs (e.g., GPT-4, Llama-3.1) and human tutors. The aim is to find the tutor or predict pedagogical quality in many spheres, including mistake identification and guidance.

The organizers provided Development set, `mrbench_v3_devset.json`, split into 90% for training (2,228 Instances) and 10% for validation (248 Instances). The final evaluation came from

[1] https://sig-edu.org/sharedtask/2025

Split	Instances	Unique Words	Total Words
Train	2,228	8,134	512,392
Validation	248	3,833	52,981
Test	1,547	7,057	454,720

Table 1: Dataset statistics across different splits.

the Test set, `mrbench_v3_testset.json` (1,547 Instances). Table 1 shows the dataset statistics.

4 Methodology

This section outlines our approaches utilized for Track 1 - Mistake Identification, Track 2 - Mistake Location, Track 3 - Providing Guidance, Track 4 - Actionability, and Track 5 - Guess the Tutor Identity. The study evaluated many transformer-based approaches using hyperparameter optimization to improve performance. The architectural frameworks used for all tasks is illustrated in Figure 1

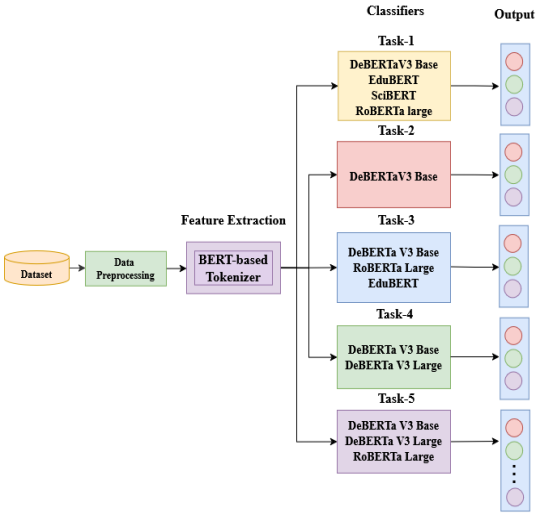

Figure 1: Overview of the Pedagogical Ability Assessment Process for AI-powered Tutors

4.1 Data Preprocessing and Feature Extraction

We processed `mrbench_v3_devset.json` and `mrbench_v3_testset.json` files for all five tracks. Every distinct tutor response in a conversation stood isolated. Using descriptive markers and newlines, concatenating the "Conversation History" and "Tutor Response" produced the input text for models; instances lacking tutor responses were removed. Relevant annotations (e.g., "Mistake_Identification") were retrieved for Tracks 1–4 (Mistake Identification, Mistake Location, Providing Guidance, Actionability); their "Yes," "To some extent," "No" labels were mapped to [0, 1, 2]. Development set tutor identities for Track 5 (Guess the Tutor Identity) were mapped to one of nine canonical tutor classes then to numerical labels [0–8]. Feature extraction used pre-trained Transformer models (DeBERTa-V3 base/large, RoBERTa-Large, EduBERT, SciBERT). each model's particular `AutoTokenizer` turned input texts into `input_ids`, `attention_mask`, and optionally `token_type_ids`, Padded or trimmed to 512 tokens.

4.2 Transformer-Based Models

The methodological foundation for all five tracks of BEA 2025 Shared Task focuses on the fine-tuning of pre-trained Transformer models (Vaswani et al., 2017). These architectures, with their well-known self-attention mechanisms, are proficient in capturing contextual relationships within text because of the highly sophisticated contexts and excel at capturing intricate contextual relationships within text makes them very suitable for a range of challenges in Natural Language Processing (NLP) (Devlin et al., 2019). Transformer's ability to model long-range dependencies is critical given the nuanced nature of assessing pedagogical abilities and identifying distinctive tutor characteristics from snippets of dialogues. A collection of models from the Hugging Face Transformers library[2] was chosen, including those pre-trained specifically on scientific or educational corpora as well as more general NLU models. SciBERT (Beltagy et al., 2019) and EduBERT (Clavié and Gal, 2019) are two, alongside RoBERTa-Large (Liu et al., 2019) and DeBERTa-V3 base and large configurations (He et al., 2021). For each task, these pretrained encoders were modified by adding a sequence classification head for each task. This head has a dropout layer and a linear layer that maps the output representation of the encoder associated with the special [CLS] token to the logits for the respective number of classes for each track. All models had the same input constructed by joining the "Tutor Response" and "Conversation History". Model-specific tokenizers were used according to each model's pretraining, with padding and truncation to 512 tokens.

In Track 1, Mistake Identification, the goal was a 3-way classification problem determining if a tutor's response acknowledged a student's mistake,

[2] https://huggingface.co/transformers

with labels "Yes", "To some extent", and "No". For this track, we experimented with SciBERT, EduBERT, RoBERTa-Large and DeBERTa-V3 base. SciBERT, which was pretrained on a large corpus of scientific literature, was selected because it could be expected to perform well with the formal and technical language of mathematics. EduBERT was chosen because it was trained on educational data, which may enhance understanding in teaching cases. RoBERTa-Large, a robustly optimized model, served as a strong general-purpose baseline, while DeBERTa-V3 base offered a more recent architecture known for its efficiency and strong performance. As highlighted, class imbalance was tackled by fine-tuning SciBERT models with weighted CrossEntropy Loss, where the greater class imbalance was compensated by inversely modifying class weights to their occurrence within the training data, and also Focal Loss (Lin et al., 2017) (with $\gamma = 2.0$ and $\alpha = 2.0$ in some configurations of SciBERT) that diminishes the emphasis on well-classified examples. For estimation smoothing SciBERT's label smoothing was set to 0.1 which was designed to counterbalance overconfidence. EduBERT and RoBERTa-Large models under this track predominantly applied weighted CrossEntropy Loss while the DeBERTa-V3 base model for this track used the standard Cross Entropy Loss provided by the Hugging Face sequence classification framework. These models were trained with the goal of detecting nuanced indications of mistake recognition in tutor responses.

Track 2, Mistake Location, was developed with similar 3-way classification where response "Yes", "To some extent" and "No" were used to capture if a tutor is precise to the location of the student's error. For this track, we primarily utilized the DeBERTa-V3 base model. With disentangled attention and the new pre-training objective (ELECTRA-style) DeBERTa-V3 architecture enhances understanding for relations between tokens and the context which we believed could prove useful in determining whether certain parts of the student's solution were referred to. Cross Entropy Loss with weights was implemented for fine-tuning for this track. This was important considering that the label distribution for "Mistake Location" was often skewed, and weighting is known to address underrepresented classes effectively trying to achieve understanding if the understanding was indeed accurate and crystal clear.

For Track 3, Providing Guidance, the focus was on assessing the tutor's evaluation on whether the answer provided to the student was useful, relevant, correct, and helpful, once again using 3-class schema ("Yes," "To some extent," "No"). In this track, we experimented with DeBERTa-V3 base, RoBERTa-Large, and EduBERT. The selection of DeBERTa-V3 and RoBERTa-Large was driven by their proficiency in NLU which is vital when evaluating the guidance provided on whether it is correct and relevant. EduBERT was included because his domain-specific pre-training could help identify pedagogically sound explanations, hints, or supporting questions. As in the last tracks, all these models were first fine-tuned using weighted Cross Entropy Loss. This was important for illustrating how the models adapted to differentiate effective and partially effective guidance along with ineffectual guidance, all distinct components of instructional prowess.

Track 4, Actionability, checked if the tutor's commentary offered unambiguous next steps by employing the same 3-way classification labels. For this track, we trained both DeBERTa-V3 'base' and 'large' models. The justification for the 'large' variant is to test if additional model size could capture the more acute interpretative reasoning necessary to assess if a tutor's remark was adequately sharp and instructive to enable responsive movement from the student. The larger model, with more parameters, better at understanding implicit suggestions or clues regarding the clarity of the anticipated student answer. During training, we uniformly used a weighted Cross Entropy Loss for all layers to constrain the label distribution along this dimension, hoping that the models could reliably distinguish non-constructive or minimal responses for a given prompt from non-informative utterances and conversational dead ends.

Finally, Track 5, Guess the Tutor Identity, posed a challenge of 9 classes: who among the tutors (Expert, Novice, or one of seven LLMs) gave the response in the anonymized form. For this exercise, we used DeBERTa-V3 base, DeBERTa-V3 large, and RoBERTa-Large. These techniques were chosen because of their past performance in capturing sophisticated stylistic differences, preferences, and idiosyncratic features like distinct 'fingerprints' for human tutors and LLM systems. The problem is complex at its core because varying forms or expressions, such as style fusion, which exist in different domains like various LLMs or between a novice human and some LLMs, might deeply over-

lap. Addressing the multi-class setup with nine distinct tutor identities was particularly challenging due to the imbalance in the number of available examples for each tutor. To mitigate this situation, we relied heavily on weighted Cross Entropy Loss which disproportionate class representation mitigates imbalance with a more prevailing class in the data. In turn, this prevented the models from specializing excessively to the most common types of tutors.

Throughout all five tracks, there were several components that the fine-tuning procedure shared homogeneous components. We utilized the AdamW optimizer (Loshchilov and Hutter, 2017), which integrates weight decay more effectively than traditional Adam, helping to prevent overfitting. As well as a blended learning rate scheduler which warms up for the first 10% of the total training steps. Doing so reinforces training stability during the early epochs. For example, many SciBERT and RoBERTa-L configurations achieve effective batch sizes of 16 with a device batch size of 8 and 2 accumulation steps. This technique of accumulation helps in training large models on memory-restricted GPUs while also allowing for smoother gradient estimates and enhanced model performance. As shown in Table 2, training continued until reaching the set maximum number of epochs. The model for each task was finalized based on the validation set with the highest macro F1 score for Tracks 1-4 and accuracy on Track 5. This selection process acts as an implicit early stopping mechanism. The class weights for the Cross Entropy Loss were determined by the label's corresponding training portion frequency across the development set, meaning classes who were less present in the dataset had a greater impact on loss and as such received more focus from the model. All experiments were carried out with a fixed random seed (SEED = 42) in order to ensure the reproducibility of our results.

5 Result Analysis

This section presents an analysis of the performance of various Transformer-based models across the five tracks of the BEA 2025 Shared Task. The evaluation metrics, as defined by the shared task organizers, include exact and lenient accuracy and macro F1-score for Tracks 1-4, and exact macro F1-score for Track 5. The performance of our submitted models is detailed in Table 3.

Model	LR	WD	BS	GA	EP
Track 1: Mistake Identification					
SciBERT	1e-5	0.01	8	2	12
EduBERT	1.5e-5	0.01	8	2	12
RoBERTa-Large	1.5e-5	0.01	8	2	12
DeBERTa-V3-Base	2e-5	0.01	8	1	8
Track 2: Mistake Location					
DeBERTa-V3-Base	1.5e-5	0.01	8	2	12
Track 3: Providing Guidance					
DeBERTa-V3-Base	1.5e-5	0.01	8	2	12
RoBERTa-Large	1.5e-5	0.01	8	2	12
EduBERT	1.5e-5	0.01	8	2	12
Track 4: Actionability					
DeBERTa-V3-Base	1.5e-5	0.01	2	2	12
DeBERTa-V3-Large	1.5e-5	0.01	2	2	12
Track 5: Tutor Identity					
DeBERTa-V3-Base	2e-5	0.01	8	2	15
DeBERTa-V3-Large	1.8e-5	0.01	2	2	10
RoBERTa-Large	2e-5	0.01	8	2	15

Table 2: Hyperparameters used across the five tracks. LR: Learning Rate, WD: Weight Decay, BS: Batch Size, GA: Gradient Accumulation, EP: Epochs.

Model	E-F1	E-Acc	L-F1	L-Acc
Track 1: Mistake Identification				
RoBERTa Large	0.6339	0.7938	0.8395	0.9043
SciBERT	0.6393	**0.8500**	0.8545	0.9121
EduBERT	0.6597	0.8429	0.8665	0.9205
DeBERTaV3 Base	**0.6617**	0.8397	**0.8782**	**0.9315**
Track 2: Mistake Location				
DeBERTaV3 Base	**0.4935**	**0.6057**	**0.7051**	**0.7401**
Track 3: Providing Guidance				
RoBERTa Large	0.4758	**0.5863**	**0.6997**	**0.7750**
EduBERT	0.4918	0.5785	0.6885	0.7395
DeBERTaV3 Base	**0.4933**	0.5695	0.6990	0.7608
Track 4: Actionability				
DeBERTaV3 Base	0.6117	0.6781	0.8170	0.8500
DeBERTaV3 Large	**0.6284**	**0.6955**	**0.8223**	**0.8565**
Track 5: Tutor Identity				
RoBERTa Large	0.8237	0.8151	-	-
DeBERTaV3 Base	0.8618	0.8597	-	-
DeBERTaV3 Large	**0.8621**	**0.8621**	-	-

Table 3: Performance of all models across five tracks. E-F1: Exact macro F1 score, E-Acc: Exact Accuracy, L-F1: Lenient macro F1 score, L-Acc: Lenient Accuracy

For Track 1, Mistake Identification, DeBERTa-V3 Base has achieved the best exact macro F1 score of 0.6617 achieving the highest exact macro F1 score. This model also demonstrated strong performance with a lenient macro F1 score of 0.8782 and lenient accuracy of 0.9315. EduBERT's performance on exact macro F1 score was just slightly weaker at 0.6597 while SciBERT had the best exact accuracy of 0.8500. The best exact macro F1 score with DeBERTa-V3 Base seems to suggest that, even with a standard Cross Entropy Loss, there are greater architectural advantages in the model that allow it to grasp the intricacies of 3-way classification better than other models. The results from

SciBERT and EduBERT indicate that the use of weighted loss functions was beneficial in achieving competitive exact scores and were likely instrumental in achieving class imbalance resolution.

For Track 2, Mistake Location, our sole entry was the DeBERTa-V3 Base model which was tuned with weighted Cross Entropy Loss and achieved an exact macro F1 score of 0.4935, lenient macro F1 score of 0.7051. There were no other entries. Scores suggest that identifying the precise origins of a mistake's location is strictly harder than simply identifying an error. The considerable gap between the exact macro F1 score and lenient macro F1 scores highlights that while the model could often recognize some level of mistake location awareness ("To some extent"), achieving definitive localization ("Yes") was less frequent.

For Track 3, Providing Guidance, DeBERTa-V3 Base reached the highest exact macro F1 score of 0.4933. EduBERT was a strong contender with exact macro F1 score of 0.4918, while RoBERTa-Large scored 0.4758 in exact macro F1 score. The lenient macro F1 scores were approximately 0.69 for all three models, with RoBERTa-Large and DeBERTaV3-Base at 0.6997 and 0.6990 respectively. The exact macro F1 scores, marginally surpassing 0.50, highlight the challenge posed in automatically evaluating the correctness and relevance of pedagogical guidance. The imbalances among the "Yes", "To some extent", and "No" categories for this particular dimension is what prompted the use of weighted Cross Entropy Loss for this model causing all of the categories to blend in with the aim of unifying the discrepancies.

In Track 4, Actionability, DeBERTa-V3 Large showed the best performance with an exact macro F1 score of 0.6284 and exact Accuracy of 0.6955. The DeBERTa-V3 Base model was slightly behind with an exact macro F1 score of 0.6117. It seems that the larger sized DeBERTa model boosts with added features helped with classifying the actionability of tutor responses. Both models had employed Cross Entropy Loss that was helpful for the other model in achieving such classifiers.

For Track 5, Guess the Tutor Identity, where exact macro F1-score is the primary metric, DeBERTa-V3 Large achieved the best exact macro F1 score of 0.8621. The corresponding exact accuracy for this model was also 0.8621. DeBERTa-V3 Base also performed robustly with an exact macro F1 score of 0.8618 and exact accuracy of 0.8597, followed by RoBERTa-Large at 0.8237 (exact macro F1 score) and 0.8151 (exact accuracy). The strong performance, particularly of the DeBERTa architectures, indicates their capability to discern subtle stylistic and content based patterns distinguishing the nine different tutor identities. This multi-class problem for which the weighted Cross Entropy Loss was quite useful for dealing with was clearly non-trivial.

To summarize, DeBERTa-V3 base and large architectures achieved the best results considering the most important evaluation metric is exact macro F1 score for most tracks. The large showed some advantages in Tracks 4 and 5 where increased model complexity might be helpful. The low exact macro F1 scores, especially for Tracks 2 and 3, suggest difficulties automatically capturing the intricacies of assessment within teaching highlight the intricacies involved in evaluating pedagogical features.

6 Conclusion

This paper introduces a system developed for the UNLP This paper detailed our participation in the BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-powered Tutors, presenting systems built upon fine-tuned Transformer models. We evaluated multiple architectures including DeBERTa-V3, RoBERTa-Large, SciBERT, and EduBERT for the five distinct tracks of Mistake Identification, Mistake Location, Providing Guidance, Actionability, and Tutor Identity. Throughout the investigations, DeBERTa-V3 was the top performer across all tracks based on the primary Exact macro F1 score metric. For Mistake Identification, DeBERTa-V3 Base achieved the Exact macro F1 score of 0.6617, and for Providing Guidance, 0.4933. For the other tracks of Actionability and Tutor Identity, DeBERTa-V3 Large excelled with 0.6284 and 0.8621 Exact macro F1 score respectively. For Mistake Location, DeBERTa-V3 Base scored an Exact macro F1 score of 0.4935. These findings support the assertion that sophisticated Transformer models are capable of intricate pedagogical assessments. The Exact macro F1 scores obtained for Providing Guidance and Mistake Location depict the challenges associated with higher-degree classification, demonstrating the inherent difficulty of the tasks. Methodological choices, such as strategic hyperparameter tuning and the application of appropriate loss functions (e.g., weighted Cross Entropy Loss or Focal Loss) to manage class imbalances, were important for optimizing performance. This work

contributes to the ongoing efforts to develop robust automated tools for assessing and improving AI tutor effectiveness in educational dialogues.

Limitations

Our study, while demonstrating the effectiveness of Transformer models for assessing pedagogical abilities, has several limitations. First of all, the performance, especially on exact macro F1-scores for challenging tasks like Mistake Location and Providing Guidance, indicates that current models still find it difficult to have the fine-grained semantic knowledge needed for these demanding tests. Second, our method depends on the particular annotations and definitions given in the MRbench dataset; model performance may change depending on alternative educational taxonomies or data from other fields outside mathematics. Moreover, although weighted loss functions helped us to solve class imbalance, significant imbalances for some labels or tutor identities could still influence generalization. Finally, the computational resources needed for fine-tuning and experimenting with several big Transformer models can be significant, therefore perhaps restricting more general architectural research or more comprehensive hyperparameter searches.

Acknowledgments

This work was supported by Southeast University, Bangladesh.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Maria Blancas-Muñoz and 1 others. 2018. Hints vs distractions in intelligent tutoring systems: Looking for the proper type of help. *arXiv preprint arXiv:1806.07806*.

Hao Chen and 1 others. 2024. Ai-driven virtual teacher for enhanced educational efficiency: Leveraging large pretrain models for autonomous error analysis and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press.

Benjamin Clavié and Kobi Gal. 2019. Edubert: Pretrained deep language models for learning analytics. *arXiv preprint arXiv:1912.00690*.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. *arXiv preprint arXiv:2009.06978*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

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*.

Ekaterina Kochmar and 1 others. 2020. Automated personalized feedback improves learning gains in an intelligent tutoring system. *arXiv preprint arXiv:2005.02431*.

Chen Li and 1 others. 2024. Applying natural language processing adaptive dialogs to promote knowledge integration during instruction. *Education Sciences*, 15(2):207.

Meng Liang. 2025. Leveraging natural language processing for automated assessment and feedback production in virtual education settings. *Journal of Educational Computing Research*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengyuan Liu and 1 others. 2024. Scaffolding language learning via multi-modal tutoring systems with pedagogical instructions. *arXiv preprint arXiv:2404.03429*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jakub Macina, Nico Daheim, Sankalan 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 *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. *arXiv preprint arXiv:2405.12240*.

Mehak Maniktala and 1 others. 2020. Avoiding help avoidance: Using interface design changes to promote unsolicited hint usage in an intelligent tutor. *arXiv preprint arXiv:2009.13371*.

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 Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.

Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Lisbon, Portugal. Association for Computational Linguistics.

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 2023)*, pages 785–795, Toronto, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Deliang Wang, Dapeng Shan, Ran Ju, Ben Kao, Chenwei Zhang, and Gaowei Chen. 2024. Investigating dialogic interaction in k12 online one-on-one mathematics tutoring using ai and sequence mining techniques. *Education and Information Technologies*, pages 1–26.

Yibo Yan, Shen Wang, Jiahao Huo, Philip S. Yu, Xuming Hu, and Qingsong Wen. 2024. Mathagent: Leveraging a mixture-of-math-agent framework for real-world multimodal mathematical error detection. *arXiv preprint arXiv:2405.12284*.