Thapar Titan/s : Fine-Tuning Pretrained Language Models with Contextual Augmentation for Mistake Identification in Tutor–Student Dialogues

Harsh Dadwal[1], **Sparsh Rastogi**[1], **Jatin Bedi**[1]
[1]Thapar Institute of Engineering and Technology, Patiala, India
hdadwal_be22@thapar.edu, srastogi_be22@thapar.edu, jatin.bedi@thapar.edu,

Abstract

This paper presents Thapar Titan/s' submission to the BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-powered Tutors (Kochmar et al., 2025). The shared task consists of five subtasks; our team ranked 18th in Mistake Identification, 15th in Mistake Location, and 18th in Actionability. However, in this paper, we focus exclusively on presenting results for Task 1: Mistake Identification, which evaluates a system's ability to detect student mistakes.

Our approach employs contextual data augmentation using a RoBERTa based masked language model to mitigate class imbalance, supplemented by oversampling and weighted loss training. Subsequently, we fine-tune three separate classifiers: RoBERTa, BERT, and DeBERTa for three-way classification aligned with task-specific annotation schemas. This modular and scalable pipeline enables a comprehensive evaluation of tutor feedback quality in educational dialogues.

1 Introduction

With the rapid evolution of large language models (LLMs), their integration into the educational domain has expanded significantly. These models present a transformative opportunity to enhance equitable access to high-quality education, especially in remote or under-resourced areas where there is a persistent shortage of qualified educators. When implemented as AI-powered tutors, LLMs can facilitate interactive, human-like dialogues that potentially overcome the constraints of conventional educational tools and enable scalable, personalized learning experiences.

Nonetheless, despite their promise, current LLMs exhibit several notable limitations. They are susceptible to inherent biases derived from their training data, often display reduced reliability in solving mathematical problems requiring structured reasoning, and are prone to generating hallucinated or factually inaccurate responses. These deficiencies raise critical concerns about their dependability in educational settings where accuracy and clarity are paramount. Consequently, there is a growing imperative to establish rigorous and systematic frameworks for assessing the pedagogical efficacy of state-of-the-art generative models in the context of educational dialogues. Evaluating the pedagogical capabilities of generative models is crucial because AI tutors must do more than coherent dialogue generation, they need to provide accurate, constructive, and context-sensitive guidance that supports effective learning. This is especially important in mathematics and reasoning tasks, where precise problem-solving steps and logical explanations are essential. Without assessing these educational qualities, models may produce plausible but incorrect or misleading responses. Therefore, rigorous evaluation of pedagogical effectiveness is vital to ensure AI tutors genuinely enhance learning and meet educational standards.

Due to the absence of a unified evaluation framework, prior studies have adopted a variety of criteria to assess the effectiveness of AI tutoring systems. For instance, (Tack et al., 2023) and (Tack and Piech, 2022) focused on whether the model communicates like a teacher, understands student needs, and offers helpful guidance. (Macina et al., 2023) employed human evaluators to judge responses based on coherence, correctness, and fairness in tutoring. Meanwhile, (Wang et al., 2024) emphasized usefulness, empathy, and human-likeness, and (Daheim et al., 2024) assessed responses using targetedness, correctness, and actionability.

To address these challenges, this paper presents a classification approach based on fine tuning three pretrained language models RoBERTa (Zhuang et al., 2021), DeBERTa (He et al., 2021), and BERT (Devlin et al., 2019) designed to understand the

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1278–1282

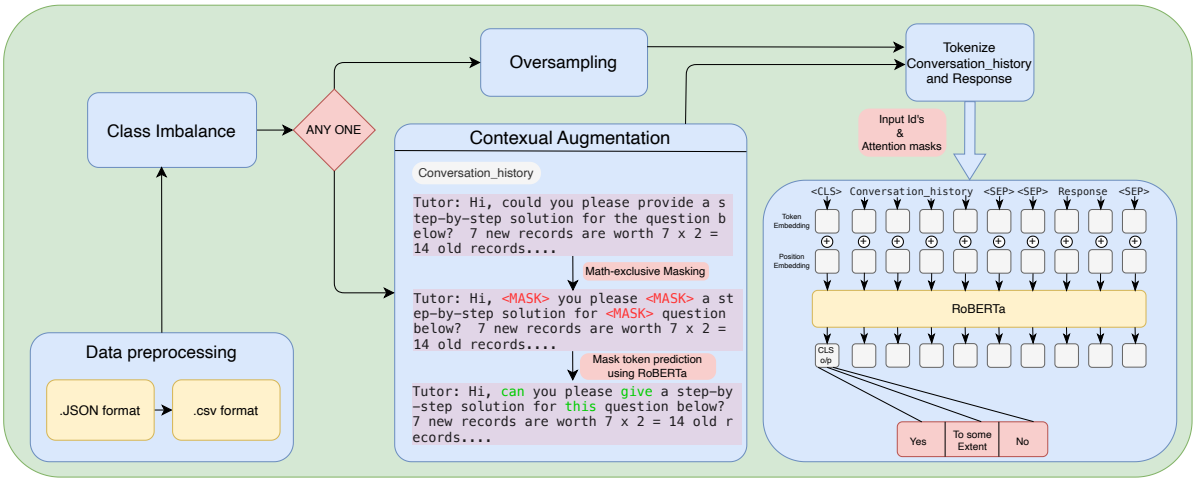

Figure 1: An schematic representation of the overall methodology

underlying context of educational dialogues and accurately identify student mistakes. To mitigate the inherent class imbalance in the dataset where certain types of errors are more frequent, our approach incorporates weighted training and contextual augmentation, ensuring the models do not develop internal biases toward specific mistake categories. Subsequent sections provide a detailed account of our methodology and findings.

2 Methodology

This work formulates the task of mistake classification in tutor–student dialogues as a multiclass classification problem. To address the pronounced class imbalance in the dataset, two complementary strategies were employed: conventional oversampling and contextual augmentation based on masked language modeling. The resulting balanced dataset was used to fine-tune transformer based models such as BERT, RoBERTa, and DeBERTa, with all layers unfrozen to facilitate effective weight optimization. The models were trained using categorical cross entropy loss and evaluated using macro F1 score and accuracy, with early stopping implemented based on macro F1. A detailed breakdown of this methodology is illustrated in Fig. 1 and further elaborated in the subsequent sections.

2.1 Dataset

We utilize the official dataset released as part of the BEA Shared Task 2025 (Maurya et al., 2025), comprising dialogues sourced from the MathDial (Macina et al., 2023) and Bridge (Wang et al., 2024)) datasets. The development set includes 300 dialogues, each consisting of several preceding tutor–student turns where the student either makes a mistake or expresses confusion, followed by the student's latest utterance and a set of tutor responses. These responses include those from human tutors extracted from the original datasets, as well as responses generated by seven LLMs-as-tutors, each identified by a unique model ID. In total, the development set contains over 2,480 tutor responses, each annotated for pedagogical quality. The annotations span three classes: yes, to some extent, and no, indicating whether the tutor successfully performs a given pedagogical function. However, the distribution is highly imbalanced, with approximately 78% of examples labeled as yes, 7% as to some extent, and only 14% as no. This skew poses a significant challenge, as it can lead to bias in model fine tuning if not properly addressed. The data is provided in JSON format with fields such as conversation id, conversation history, tutor responses, and annotations. The test set comprises 200 similarly structured dialogues from the same sources, containing unannotated responses from the same set of tutors, with tutor identities and pedagogical annotations withheld.

2.2 Data Augmentation

To address the severe class imbalance in the dataset, two complementary strategies were employed. The first involved conventional oversampling, in which the frequency of each example from the minority classes was increased by duplicating existing instances. Although this approach provided some improvement, it introduced a risk of overfitting due to repeated exposure to identical inputs. To mitigate this issue, contextual augmentation was also ap-

plied to generate diverse and meaningful examples for the underrepresented classes. A semantic masking approach was adopted, where selected words in the conversation history, which represents the student's input to the model, were masked while preserving domain specific terms and mathematical symbols. These key terms were excluded because they carry essential meaning and detail, which are critical for accurately assessing a tutoring scenario. Irrelevant stopwords were also omitted, as they do not contribute significant semantic content and would not enhance the quality of augmentation. After masking, we applied masked language modeling using a pretrained RoBERTa based model. These models predicted and replaced the masked tokens based on their surrounding context, generating fluent and semantically consistent variations of the input. By leveraging multiple models, we introduced a rich set of plausible alternatives while preserving the original intent of the student's question. Importantly, this augmentation was applied only to the input context and not to the tutor's response. Altering the responses could distort the assessment of the model's true predictive performance. This method allowed us to expand the dataset meaningfully, improve class balance, and maintain the authenticity of pedagogical evaluation.

2.3 Fine-Tuning for Classification

The overall problem was formulated as a multiclass classification task focused on identifying and localizing different types of mistakes within student-tutor dialogues. Three large language models, namely BERT large, RoBERTa, and DeBERTa, were chosen for fine-tuning due to their strong contextual understanding and performance in natural language tasks. The training was conducted using the final augmented dataset, which contained approximately 2,000 samples for each class to address the class imbalance and ensure balanced learning. To maximize performance, all layers of the models were unfrozen, allowing for comprehensive weight adjustment during training. The models were trained for up to 100 epochs on an Nvidia H100 GPU, with categorical cross entropy serving as the optimization loss function. Evaluation was performed using macro F1 score and accuracy metrics. Early stopping was applied based on the macro F1 score to prevent overfitting, and the best model weights were saved for subsequent evaluation.

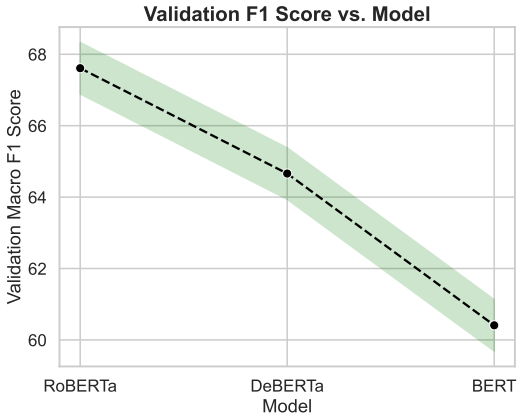

Figure 2: Model-wise Comparison of Validation Macro F1 Scores

3 Results and Discussion

Extensive experimentation was conducted across various hyperparameters and settings to assess their individual impact on model performance. RoBERTa was fixed as the baseline/default architecture for all experiments, and the mask ratio was set to a default of 15%, except where explicitly varied during the mask ratio ablation studies. The experiments focused on three key areas: evaluating different mask ratios during contextual masking (15%, 30%, and 50%), comparing transformer architectures (RoBERTa, BERT, and DeBERTa) at the default 15% mask ratio, and investigating two class imbalance handling techniques—contextual augmentation and conventional oversampling. We

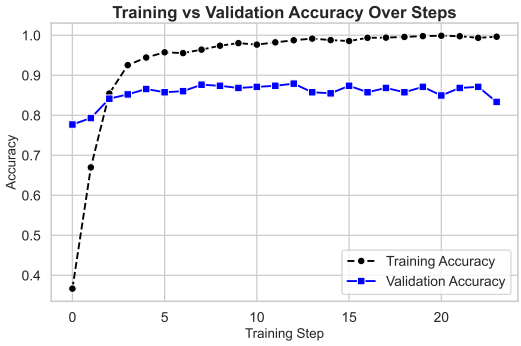

Figure 3: Training and Validation Accuracy over Optimization Steps

observed that the model achieved the highest performance with the default 15% mask ratio, yielding a training accuracy of 99.63%, validation accuracy of 87.9%, and validation F1 score of 67.61% (Fig. 3 and Fig. 4). Increasing the mask ratio to 30% and 50% led to a slight decrease in all performance metrics, with the lowest F1 scores observed at the

S. No.	Metric	Contextual Augmentation	Conventional Oversampling	Class Weights
1	Train Accuracy	99.63	100.00	99.77
2	Validation Accuracy	87.90	81.40	81.67
3	Validation F1 Score	67.61	63.75	63.48

Table 1: Performance comparison across different data augmentation and class imbalance handling techniques.

50% masking level (65.47%), as shown in Fig. 5. This indicates that excessive masking may hinder the model's ability to learn meaningful contextual representations, while the 15% mask ratio strikes an effective balance between regularization and information retention, enhancing generalization on the validation set.

Using the fixed baseline RoBERTa model at the default mask ratio, we compared the performance of different transformer architectures. RoBERTa and DeBERTa demonstrated superior results, with validation accuracies of 87.9% and 87.1%, respectively. RoBERTa slightly outperformed DeBERTa in validation F1 score (67.61% vs. 64.66%). BERT lagged with a validation accuracy of 81.4% and an F1 score of 60.41%. The stronger performance of RoBERTa and DeBERTa is attributable to their improved pre-training methods and architectural enhancements compared to BERT, facilitating better contextual understanding (Fig. 2).

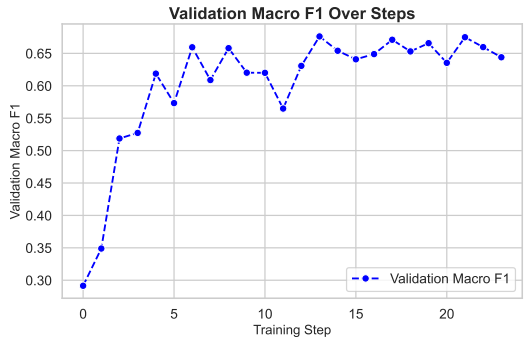

Figure 4: Validation Macro F1 Score Across Training Steps

For handling class imbalance, contextual augmentation and conventional oversampling were evaluated. While oversampling achieved perfect training accuracy (100%), it produced lower validation accuracy (81.4%) and F1 score (63.75%) compared to contextual augmentation (validation accuracy 87.9%, F1 67.61%), as shown in Table 1. This suggests that oversampling may lead to overfitting, whereas contextual augmentation, by generating semantically consistent synthetic samples, improves model generalization without overfitting.

Overall, these results emphasize the importance of choosing an appropriate mask ratio, selecting advanced transformer architectures, and using semantically informed augmentation techniques for robust model performance. Fixing RoBERTa as the baseline and adopting a 15% mask ratio proved effective across experiments. The findings highlight the necessity of careful hyperparameter tuning and data augmentation strategies, especially when addressing class imbalance. Future research may explore integrating these techniques further and evaluating them on larger, more diverse datasets.

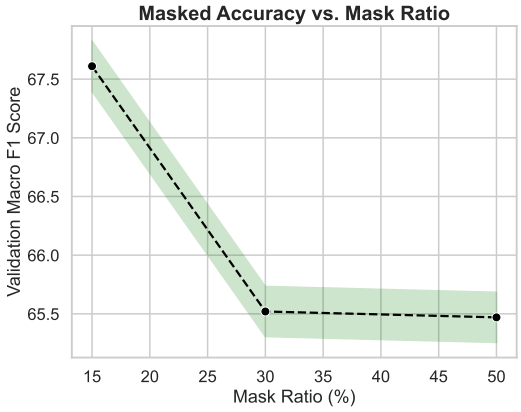

Figure 5: Effect of Masking Ratio on Validation Macro F1 Score

4 Conclusion

This study addresses the task of mistake classification in tutor–student dialogues by fine-tuning large pre-trained language models on a class-balanced dataset. To mitigate the issue of severe class imbalance, both conventional oversampling and contextual augmentation were employed, preserving the semantic integrity of student inputs. The use of BERT, RoBERTa, and DeBERTa enabled effective learning, and performance was evaluated using macro F1 and accuracy. Overall, the proposed framework enhances the reliability and generalizability of automated feedback systems. Future work may explore adaptive augmentation or dynamic feedback integration to further improve model robustness.

Limitations

This study is based on publicly available datasets, specifically MathDial and Bridge, which may not capture the full range of tutoring scenarios encountered in real-world educational settings. As a result, the model's performance and generalizability could be limited when applied to more diverse or complex dialogues beyond these datasets. Furthermore, while contextual augmentation was effective in mitigating class imbalance by generating additional examples for minority classes, this approach may inadvertently introduce subtle biases or produce variations that are not entirely representative of natural student language. Such synthetic alterations, although contextually coherent, might affect the model's robustness when faced with truly novel or unexpected inputs. Future studies could address these limitations by incorporating more diverse dialogue datasets and exploring augmentation strategies that more closely mimic real-world student behavior and language use.

References

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo Ponti. 2024. Elastic weight removal for faithful and abstractive dialogue generation. 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 7096–7112, Mexico City, Mexico. Association for Computational Linguistics.

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

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*, Online. Association for Computational Linguistics.

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*, pages 5602–5621, Singapore. Association for Computational Linguistics.

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.

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.

Anaïs Tack and Chris Piech. 2022. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues. *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, Mexico City, Mexico. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.