CU at BEA 2025 Shared Task: A BERT-Based Cross-Attention Approach for Evaluating Pedagogical Responses in Dialogue

Zhihao Lyu
University of Colorado Boulder
zhihao.lyu@colorado.edu

Abstract

Automatic evaluation of AI tutor responses in educational dialogues is a challenging task, requiring accurate identification of mistakes and provision of pedagogically effective guidance. In this paper, we propose a classification model based on BERT, enhanced with a cross-attention mechanism that explicitly models the interaction between the tutor's response and preceding dialogue turns. This design enables better alignment between context and response, supporting more accurate assessment along the educational dimensions defined in the BEA 2025 Shared Task. To address the substantial class imbalance in the dataset, we employ data augmentation techniques for minority classes. Our system consistently outperforms baseline models across all tracks. However, performance on underrepresented labels remains limited, particularly when distinguishing between semantically similar cases. This suggests room for improvement in both model expressiveness and data coverage, motivating future work with stronger decoder-only model and auxiliary information from systems like GPT-4.1. Overall, our findings offer insights into the potential and limitations of LLM-based approaches for pedagogical feedback evaluation.

1 Introduction

Recent progress in large language models (LLMs) like GPT-4, Gemini(Team et al., 2023), and LLaMA(Grattafiori et al., 2024) has rapidly improved AI conversational agents, especially in education. AI tutors, for example, can now offer students real-time, interactive feedback to boost engagement and learning(Lin et al., 2023). However, while these models generate fluent, human-like responses, evaluating the real educational value of their feedback remains challenging(Ou et al., 2023). Standard metrics such as BLEU and ROUGE fail to capture important aspects of educational dialogue—like identifying student mistakes or providing helpful guidance—which highlights the need for more fine-grained, pedagogically meaningful evaluation frameworks.

To address this gap, the BEA 2025 Shared Task goes a step further than previous tasks(Tack et al., 2023) by shifting the focus from dialogue generation to evaluating how LLMs assess educational dialogues. Evaluation is based on four key dimensions(Maurya et al., 2025): (1) Mistake Identification(Tack and Piech, 2022; Macina et al., 2023; Daheim et al., 2024), (2) Mistake Location(Daheim et al., 2024), (3) Providing Guidance(Tack and Piech, 2022; Liu et al., 2023), and (4) Actionability(Daheim et al., 2024). These dimensions capture what truly matters in educational feedback, moving beyond surface-level fluency. For more details on the task and evaluation setup, please refer to the official report(Kochmar et al., 2025).

In this paper, we present our submission to the BEA 2025 Shared Task, focusing on three evaluation tracks: Mistake Identification, Mistake Location, and Providing Guidance. Our approach enhances standard LLM classifiers with a cross-attention layer to better capture the relationship between student-tutor dialogue context and the tutor's response. Experimental results demonstrate that our method achieves strong performance across all tracks, validating the effectiveness of cross-attention for modeling educational feedback. Our team, CU, ranked 25th out of 44 in Track 1, 17th out of 31 in Track 2, and 20th out of 35 in Track 3.

2 Related Work

2.1 Early Work on Educational Feedback

Early research in educational psychology laid the theoretical foundation for understanding effective teaching practices. Hattie and Timperley (2007) proposed a widely adopted model of feedback focused on learning goals, progress monitoring, and

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1060–1072

actionable guidance, demonstrating its critical role in student achievement. The AutoTutor system (Graesser et al., 2005) formalized key tutoring strategies—such as identifying misconceptions and prompting elaboration—within an intelligent tutoring framework. Boyer et al. (2011) introduced a data-driven approach by modeling dialogue structures using hidden Markov models to predict learning gains. Wolfe et al. (2013) and Rus et al. (2017) analyzed tutor-student dialogues to assess the quality of instructional moves using semantic similarity and discourse act classification.

2.2 LLMs for Educational Dialogue Evaluation

Recent advances in LLMs have reshaped how we engage with language and text—transforming not only natural language processing(NLP) research but also the evaluation of educational dialogues. A growing body of research explores how LLMs can be used to assess or enhance educational feedback. For example, Balse et al. (2023) investigated the ability of GPT-3.5 to explain logical programming errors, finding that while explanations were often imperfect, they reliably identified key issues. Lee et al. (2024) improved LLM-based classification accuracy by structuring prompts to encode error relationships. Molina et al. (2024) showed that LLM tutors improve accessibility for non-native English speakers, while Xu et al. (2025) built a virtual AI tutor capable of analyzing student drafts and generating error-specific feedback. Reinforcement learning approaches such as that of Scarlatos et al. (2025) have further enhanced LLM tutors by optimizing pedagogical reward functions. Kakarla et al. (2024) demonstrated the potential of LLMs in evaluating human tutor responses, highlighting both strengths and limitations.

2.3 BERT for Dialogue and Tutoring Systems

Parallel to LLM advancements, BERT-based architectures have also proven effective for educational dialogue modeling and intelligent tutoring systems (ITS). In the domain of dialogue understanding, DialogueBERT (Zhang et al., 2021) and DialBERT (Li et al.) incorporate hierarchical context and speaker-role awareness to improve performance on tasks such as disentanglement, emotion recognition, and intent detection. CS-BERT (Wang et al., 2021), trained on domain-specific customer service dialogues, introduces masked speaker prediction and turn-level segment embeddings, yielding robust results in low-resource scenarios. Within ITS applications, BERT has been adapted for various pedagogical tasks. LBKT (Li et al., 2024) combines BERT and LSTM with Rasch-based embeddings for long-sequence knowledge tracing, improving interpretability and accuracy. Tutor-KD (Kim et al., 2022) introduces tutor-guided difficulty adaptation in knowledge distillation, enhancing BERT's generalization. Wang et al. (2024) compare BERT with ChatGPT for dialogic pedagogy support and note that, while BERT performs well in structured analysis, it lacks the interactive fluency teachers often prefer.

3 Research Gap

Despite progress in educational theory and NLP, evaluating the pedagogical quality of AI tutor responses remains difficult. Traditional methods emphasize structured feedback but rely on manual annotation and lack scalability, while LLMs offer fluency yet often miss deeper educational goals like mistake identification and guidance. Although some work proposes education-driven metrics, most automated approaches fail to effectively model dialogue context. BERT-based models show potential in educational settings but are still underused for evaluating tutor responses within full dialogue history.

To address this, we introduce a BERT-based classifier with a cross-attention mechanism that explicitly models tutor–dialogue interactions, enabling more accurate and context-aware evaluation across multiple pedagogical dimensions.

4 Methodology

In this section, we present the model architecture, including the data processing, BERT-based representation generation, and the cross-attention and classification layers. An overview of the model is illustrated in Figure 1. First, The conversation history and tutor response are preprocessed separately, with special tokens inserted at the beginning of each utterance to indicate their order. These inputs are then encoded using a pretrained BERT model to obtain high-dimensional representation. They are passed into a cross-attention layer, where the response serves as the query and the conversation history as the key and value. Finally, the cross-attended representation is fed into a classification layer that predicts one of three labels: *Yes*, *No*, or *To some extent*.

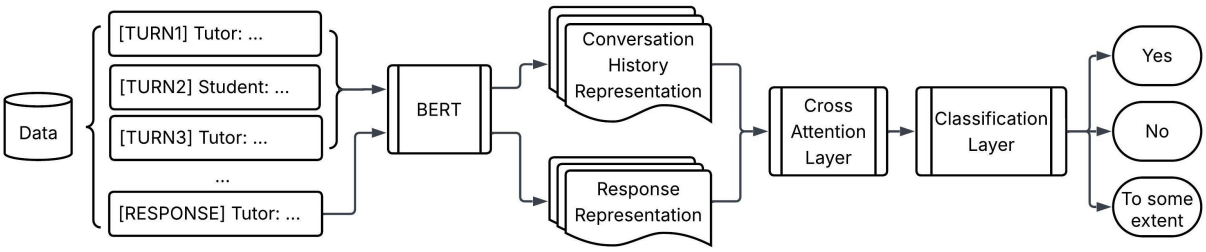


Figure 1: Overview of the proposed model architecture.

4.1 Data Preprocessing and Representation Generation

Data Augmentation. By examining the label distribution of the shared task dataset, we observed a significant imbalance between the *Yes*, *No*, and *To some extent* labels (see Table 1). Also, the *Yes* and *To some extent* labels are semantically similar, which may require the model to make finer distinctions. To address this issue without substantially altering the data distribution, we applied data augmentation only to the training set. Specifically, we used GPT-4.1 to rephrase all instances with minority labels once, thereby augmenting the dataset across all three tracks. This results in a simple 2:1 ratio between augmented and original samples for the minority classes. The ratio was determined heuristically rather than through systematic tuning, with the aim of increasing class diversity while preserving the overall label distribution. This augmentation strategy led to improved F1 scores in our subsequent experiments. The prompt used for rephrasing is provided in Appendix A.

Track 1 Label	Before Aug	After Aug
Yes	1932	1932
No	370	666
To some extent	174	313

Track 2 Label	Before Aug	After Aug
Yes	1543	1543
No	713	1283
To some extent	220	396

Track 3 Label	Before Aug	After Aug
Yes	1407	1932
No	566	1018
To some extent	503	905

Table 1: Comparison of label counts before and after data augmentation across the three tracks.

Input Labeling. To preserve the contextual meaning and sequential order of the conversation history, we manually insert order indicators (e.g., *[TURNx]*) at the beginning of each utterance and mark the tutor's response with a *[RESPONSE]* token. Compared to the insertion of turn and role embeddings in DialogBERT (Zhang et al., 2021), this simple modification is easier to implement while still demonstrating effectiveness.

Representation Generation. Given BERT's strong performance and widespread success across various NLP tasks (Devlin et al., 2019), we retain its original architecture and use its encoder only as a representation generator. Specifically, BERT first generates three types of embeddings from the input: token embeddings, segment embeddings, and position embeddings. These embeddings are added and then fed into the Transformer's self-attention layers(Vaswani et al., 2017), which consist of multiple attention heads and stacked layers that compute contextualized representations for each token. After processing through these layers, the BERT encoder produces high-dimensional vectors as the final hidden states for both the conversation history and the tutor response. The input representation process is illustrated in Figure 2.

4.2 Cross-Attention and Classification Layer

After obtaining token-level representations of the tutor's response and the conversation history using a BERT encoder, we combine them using a cross-attention mechanism to model the relationship between the two. Inspired by the decoder structure in the Transformer architecture, we treat the response as the **query** and the conversation history as the **key** and **value**. This allows each token in the response to selectively attend to relevant parts of the dialogue history (Figure 3). Formally, let

- $R \in \mathbb{R}^{l \times d}$ be the representation matrix of the tutor's response, where l is the number of

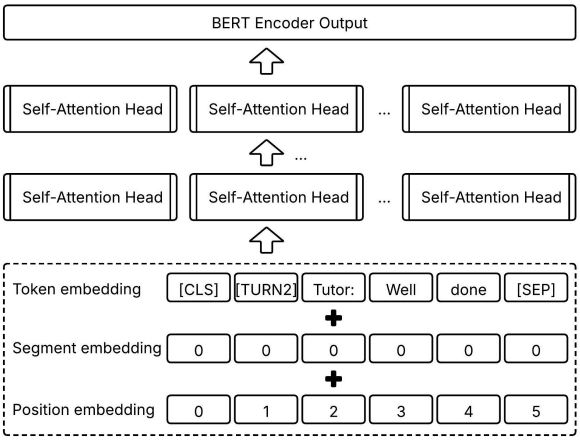


Figure 2: Illustration of the input representation process in BERT, including embedding generation and self-attention encoding.

tokens in responses, and d is the hidden size;

- $H \in \mathbb{R}^{n \times d}$ be the representation matrix of the conversation history, where n is the number of tokens in the history.

$$Weight = softmax(\frac{R \cdot W_Q (H \cdot W_K)^\top}{\sqrt{d}}) \quad (1)$$

$$Attention(R, H, H) = Weight \cdot H \cdot W_V \quad (2)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$ are trainable projection matrices. Here, $Weight \in \mathbb{R}^{l \times n}$ represents the attention weights between each token in the response (R) and each token in the conversation history (H), where l and n are the number of tokens in the response and history, respectively. This mechanism enables the response to selectively attend to relevant segments of the dialogue context, producing a contextualized representation that informs the final classification.

After obtaining the cross-attended response representations, we extract the hidden state corresponding to the *[CLS]* token (the first token position) to serve as the aggregate representation of the response. This vector is then passed through a dropout layer for regularization, followed by a linear classification layer that maps the hidden representation to a logits vector of dimension $\mathbb{R}^{d \times m}$, where d is the hidden size and $m = 3$ is the number of classification labels used across all tasks. The resulting logits are used to compute the weighted cross-entropy loss during training.

Algorithm 1: Dialogue-level Split with Label Distribution Balancing

Input: Training and validation dialogues
Output: Training and validation splits with similar label distributions

```
Compute global label distribution ratio from all dialogues
Split dialogues into initial training (80%) and validation (20%) sets
Compute label distribution in both sets
for iteration = 1 to max_iters do
    foreach train dialogue d_i (sampled subset) do
        foreach val dialogue d_j (sampled subset) do
            Swap d_i and d_j between training and validation sets
            Compute new label distributions
            Compute ratio error in both sets
            if new error < old error then
                Accept the swap
                Update label counts
                break both loops
            end
        end
    end
end
```

5 Experiments

5.1 Dataset

As the shared task required, we use a development set and a test set from the MathDial(Macina et al., 2023) and Bridge(Wang et al., 2023) datasets.

Development Set: Contains 300 dialogues where students make mistakes or show confusion. Each dialogue includes the student's last question and responses from multiple tutors (LLMs and humans), with over 2,480 responses labeled for pedagogical quality.

Test Set: Contains 200 similar dialogues. Tutor responses are not labeled and tutor identities are hidden. The set is intended only for official evaluation and is not available for model development.

Given that all 2,480 responses are associated with only 300 dialogues, we perform dialogue-level splitting to ensure that no dialogue appears in both training and validation sets. This prevents data leakage and ensures a fair evaluation. Combined with the label imbalance issue noted in Section 4,

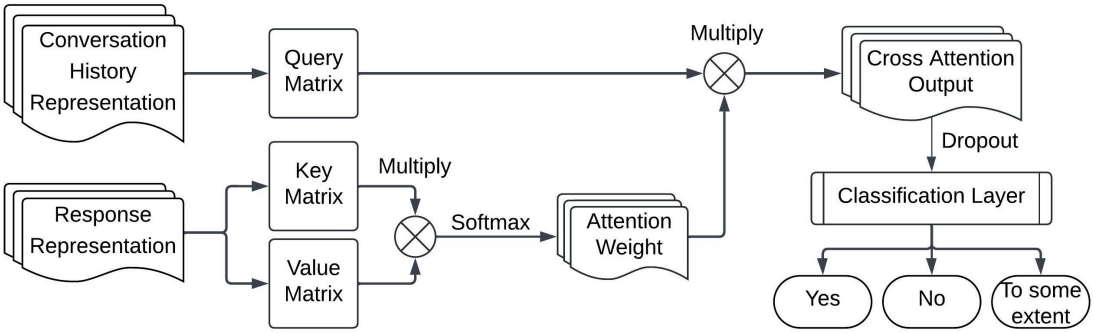


Figure 3: Overview of the cross attention mechanism and classification.

this requires careful design of the data split. First, we perform an initial 80/20 split of the data into training and validation sets. Then, to ensure that the label distributions in both sets are similar, we iteratively swap samples between them. The pseudocode is shown in Algorithm 1.

5.2 Experimental Setup

We fine-tuned all models using the BERT base uncased(110M) architecture. Inputs combined the tutor's response and dialogue context, with cross-attention as described in Methodology. The *[CLS]* token was used for classification, followed by dropout and a linear layer. All inputs were tokenized using the BERT tokenizer with a maximum sequence length of 512. Sequences longer than 512 tokens were truncated, and shorter sequences were padded accordingly.

Training was conducted on a single NVIDIA RTX 4060 Ti GPU. We used the AdamW optimizer with a learning rate of 2e-5 and a batch size of 5. Models were trained for up to 5 epochs with early stopping based on Macro-F1 score on the validation set. A cosine learning rate scheduler was used, and a dropout rate of 0.1 was applied before the final classification layer. To address class imbalance, we adopted a log-weighted cross-entropy loss, where the weight for each class i was computed as $w_i = \log\left(\frac{N}{n_i}\right)$, with N the total number of samples and n_i the number of samples in class i. The overall training procedure is summarized in Algorithm 2.

5.3 Baselines

To provide a reference for zero-shot performance, we included two LLM baselines: GPT-4.1 and LLaMA 3.2 1B. For GPT-4.1, we used the OpenAI API and designed a custom prompt to elicit pedagogical labels (*Yes*, *No*, or *To some extent*) for each tutor response, given the tutor's response and dialogue context. This model was not fine-tuned on our dataset and operates purely in a zero-shot setting. The full prompt example is included in Appendix B. For LLaMA 3.2 1B, we used the open-source model and ran it locally. Similar to GPT-4.1, we applied a handcrafted prompt to guide the model in classifying tutor responses. The LLaMA model was also evaluated in a zero-shot. The prompt used is provided in Appendix C. These baselines allow us to assess the effectiveness of our fine-tuned BERT models against general-purpose LLMs without task-specific adaptation.

5.4 Main Results

We now present the results of our fine-tuned BERT-based models, comparing variants with and without the proposed cross-attention mechanism(CA), as well as the impact of data augmentation(Aug). These models are evaluated on all three shared task tracks and compared with the zero-shot baselines (Section 5.3). Our team, CU, participated in three tracks and ranked 25th/44 in Track 1, 17th/31 in Track 2, and 20th/35 in Track 4. The results are shown in Table 2.

5.5 Discussion

5.5.1 Improving Track 1 Performance with Cross-Attention

As shown in Table 2 and Figure 4, incorporating the cross-attention mechanism substantially improved the model's performance on Track 1. The Macro-F1 score increased from 0.578 to 0.687, and accuracy improved from 0.849 to 0.867. While the baseline BERT model performed reasonably well on the majority class *Yes*, it failed to identify any instances of the minority class *To some extent*, as shown by a complete absence of predictions for that label in the confusion matrix. This resulted in a biased classifier with high accuracy but limited

Algorithm 2: Training procedure for BERT with Cross-Attention

Input: Training set $\mathcal{D}_{\text{train}}$, Validation set $\mathcal{D}_{\text{val}}$, number of epochs N, batch size B

Output: Best model parameters θ^*

Initialize BERT-based model with cross-attention, parameters θ ;
Initialize AdamW optimizer and cosine learning rate scheduler ;
$\theta^* \leftarrow \theta$, best_val_f1 $\leftarrow 0$;
for *epoch* $= 1$ **to** N **do**
  for *each batch* (x_{dial}, x_{resp}, y) *in* $\mathcal{D}_{train}$ **do**
    Forward pass: $z \leftarrow \text{Model}(x_{\text{dial}}, x_{\text{resp}})$;
    Compute loss: $L \leftarrow \text{CrossEntropy}(z, y)$;
    Backward pass: update θ via optimizer ;
    Update learning rate scheduler ;
  end
  Evaluate model on $\mathcal{D}_{\text{val}}$ to obtain F1 score;
  if *F1* $>$ *best_val_f1* **then**
    best_val_f1 $\leftarrow$ F1;
    $\theta^* \leftarrow \theta$ `// Save best model`
  end
end
return θ^*

Model	Acc.	Macro-F1
zero-shot GPT-4.1	0.807	0.557
zero-shot LLaMA 3.2 1B	0.758	0.440
BERT (no CA)	0.849	0.578
BERT + CA	**0.870**	**0.651**

(a) Track 1: Mistake Identification

Model	Acc.	Macro-F1
zero-shot GPT-4.1	0.548	0.472
zero-shot LLaMA 3.2 1B	0.619	0.371
BERT base(no CA)	0.678	0.429
BERT base + CA	**0.689**	0.504
BERT base + CA + Aug	0.681	**0.515**

(b) Track 2: Mistake Location

Model	Acc.	Macro-F1
zero-shot GPT-4.1	0.549	0.403
zero-shot LLaMA 3.2 1B	**0.591**	0.363
BERT base(no CA)	0.587	0.476
BERT base + CA	0.589	0.484
BERT base + CA + Aug	0.585	**0.493**

(c) Track 3: Providing Guidance

Table 2: Model performance across three task tracks.

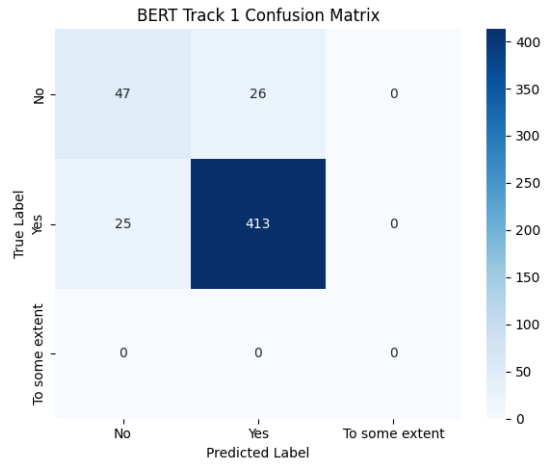
generalization.

In contrast, the BERT+Cross Attention model significantly reduced this bias. It not only improved the recall for the *No* class (from 47 to 55 true positives), but also successfully predicted 10 instances of *To some extent*, a class that the baseline model could not recognize at all. Although the number of correct predictions for *Yes* slightly decreased (from 413 to 374), this reflects a more balanced and context-sensitive classification behavior. These findings suggest that cross-attention enables the model to better align the tutor's response with subtle errors in the student's utterance, resulting in more robust performance across all categories.

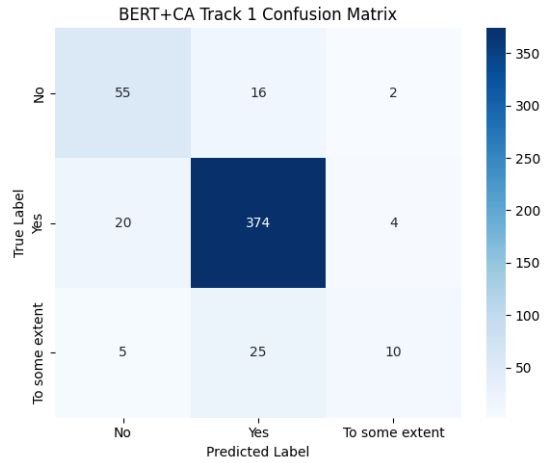
5.5.2 Benefits and Limitations of Data Augmentation in Track 2 and 3

To quantitatively assess the effect of data augmentation on class-wise performance for Track 2, we compare the classification reports of the cross-attention model before and after augmentation (see Table 3). Without augmentation, the model achieves high performance on the majority class *Yes* (F1 = 0.77), but almost entirely fails to recognize the minority class *To some extent* (F1 = 0.00). With data augmentation, the model's ability to identify *No* and *To some extent* is significantly improved, with F1-scores rising from 0.48 to 0.63 and from 0.00 to 0.20, respectively. Although the recall for *Yes* decreases slightly (from 0.92 to 0.85), the overall classification results become more balanced, as indicated by the higher Macro-F1 score. These results highlight the utility of data augmentation in mitigating class imbalance and promoting fairer evaluation across all categories.

A similar pattern is observed for Track 3: after applying data augmentation, overall accuracy decreases slightly, while Macro-F1 improves only marginally. This suggests that the benefit of augmentation is consistent but limited when class imbalance is severe.



(a) BERT



(b) BERT+CA

Figure 4: Confusion matrix comparison on Track 1.

Class	Precision	Recall	F1
No	0.64	0.39	0.48
Yes	0.67	0.92	0.77
To some extent	0.00	0.00	0.00

(a) BERT+CA

Class	Precision	Recall	F1
No	0.67	0.60	0.63
Yes	0.75	0.85	0.79
To some extent	0.27	0.15	0.20

(b) BERT+CA+Aug

Table 3: Class-wise precision, recall, and F1 score for Track 2 before and after data augmentation. Each class contains the same number of validation samples (No: 144, Yes: 301, To some extent: 59).

6 Conclusion

In this paper, we presented a system for evaluating tutor responses in educational dialogues, with a particular focus on three pedagogical dimensions as outlined in the BEA 2025 Shared Task. Leveraging a BERT-based architecture augmented with a cross-attention layer, our approach aimed to improve the model's ability to capture context and provide more accurate multi-label classification. Experimental results demonstrate that our system achieves strong performance on Track 1, while also revealing challenges in distinguishing between semantically similar categories, such as *Yes* and *To some extent* in Track 2 and 3. Data augmentation techniques were employed to mitigate class imbalance, resulting in modest improvements, particularly in minority classes. Despite these advances, our findings indicate that substantial gaps remain before such systems can be reliably deployed in real-world educational settings. Overall, our work contributes new insights into the application of LLMs for pedagogical evaluation and highlights key challenges for future research.

7 Future Work

During the training process, we observed that the number of cross-attention layers may influence classification accuracy. In future work, we plan to conduct further experiments with more advanced, higher-capacity decoder-only models, and systematically explore the effect of varying the number of cross-attention layers. In addition, the current cross-attention layer still struggles to recognize minority classes in dialogue. To address this, we aim to leverage state-of-the-art models as an auxiliary information. For example, we could use GPT-4.1 to first estimate the probability that each utterance in the conversation contains a mistake, and then pass these probabilities as initial attention weights to the cross-attention layer. This approach may enable the model to more precisely identify errors within the dialogue. Furthermore, GPT-4.1 could be used to perform more sophisticated data preprocessing, such as extracting all potential errors, so that the classification model only needs to determine whether the tutor's response correctly identifies and addresses those errors.

Our current approach is inherently pedagogy-specific: it is trained on dialogue data annotated with educational dimensions, and designed to model the relationship between student language and tutor feedback. Both the training objective and

model architecture reflect the goal of evaluating responses in a pedagogically meaningful way. In the future, further gains might be achieved by incorporating explicit pedagogical constructs, such as known error types or feedback taxonomies, into the modeling process. We see this as a promising direction for enhancing both model performance and educational relevance.

8 Limitations

Despite the promising results demonstrated by our system, several limitations remain. First, while the model achieves strong performance on Track 1, its accuracy on Track 2 and Track 3 remains below 70%, with Macro-F1 scores falling short of 60%. This gap suggests that the system is not yet robust enough for real-world educational deployment. As shown in Table 3, the model tends to favor the majority class (*Yes*) and continues to struggle with the *No* and *To some extent* categories. Notably, *To some extent* is semantically close to *Yes*, and despite our data augmentation efforts, its precision in Track 2 remains below 30%, indicating substantial room for improvement in recognizing minority classes.

Second, although BERT has long been a strong performer in NLP tasks, its encoder-decoder architecture is increasingly surpassed by newer, decoder-only models such as LLaMA and Qwen(Yang et al., 2025). These models are rapidly becoming the mainstream in LLM research. However, their substantially larger parameter sizes make them less accessible for users with limited computational resources. Furthermore, the additional cross-attention layer proposed in this work increases computational demands even further. After the shared task deadline, we experimented with LLaMA 3.2 1B augmented with our cross-attention mechanism and conducted full fine-tuning. Compared to BERT, LLaMA 3.2 1B has nearly ten times more parameters, making local training on personal computers nearly infeasible. This poses an even greater barrier for educators or practitioners who may lack expertise in machine learning or access to high-performance hardware.

References

Rishabh Balse, Viraj Kumar, Prajish Prasad, and Jayakrishnan Madathil Warriem. 2023. Evaluating the quality of llm-generated explanations for logical errors in cs1 student programs. In *Proceedings of the 16th Annual ACM India Compute Conference*, pages 49–54.

Kristy Elizabeth Boyer, Robert Phillips, Amy Ingram, Eun Young Ha, Michael Wallis, Mladen Vouk, and James Lester. 2011. Investigating the relationship between dialogue structure and tutoring effectiveness: a hidden markov modeling approach. *International Journal of Artificial Intelligence in Education*, 21(1-2):65–81.

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. *arXiv preprint arXiv:2407.09136*.

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.

Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48(4):612–618.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research*, 77(1):81–112.

Sanjit Kakarla, Danielle Thomas, Jionghao Lin, Shivang Gupta, and Kenneth R Koedinger. 2024. Using large language models to assess tutors' performance in reacting to students making math errors. *arXiv preprint arXiv:2401.03238*.

Junho Kim, Jun-Hyung Park, Mingyu Lee, Wing-Lam Mok, Joon-Young Choi, and SangKeun Lee. 2022. Tutoring helps students learn better: Improving knowledge distillation for bert with tutor network. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7382.

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*. Association for Computational Linguistics.

Yanggyu Lee, Suchae Jeong, and Jihie Kim. 2024. Improving llm classification of logical errors by integrating error relationship into prompts. In *International Conference on Intelligent Tutoring Systems*, pages 91–103. Springer.

T Li, JC Gu, X Zhu, Q Liu, ZH Ling, Z Su, and S Wei. Dialbert: A hierarchical pre-trained model for conversation disentanglement. arxiv 2020. *arXiv preprint arXiv:2004.03760*.

Zhaoxing Li, Jujie Yang, Jindi Wang, Lei Shi, and Sebastian Stein. 2024. Integrating lstm and bert for long-sequence data analysis in intelligent tutoring systems. *arXiv preprint arXiv:2405.05136*.

Chien-Chang Lin, Anna YQ Huang, and Owen HT Lu. 2023. Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. *Smart Learning Environments*, 10(1):41.

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*.

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. *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 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.

Ismael Villegas Molina, Audria Montalvo, Benjamin Ochoa, Paul Denny, and Leo Porter. 2024. Leveraging llm tutoring systems for non-native english speakers in introductory cs courses. *arXiv preprint arXiv:2411.02725*.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2023. Dialogbench: Evaluating llms as human-like dialogue systems. *arXiv preprint arXiv:2311.01677*.

Vasile Rus, Nabin Maharjan, Lasang Jimba Tamang, Michael Yudelson, Susan R Berman, Stephen E Fancsali, and Steven Ritter. 2017. An analysis of human tutors' actions in tutorial dialogues. In *FLAIRS*, pages 122–127.

Alexander Scarlatos, Naiming Liu, Jaewook Lee, Richard Baraniuk, and Andrew Lan. 2025. Training llm-based tutors to improve student learning outcomes in dialogues. *arXiv preprint arXiv:2503.06424*.

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. *arXiv preprint arXiv:2306.06941*.

Anaïs Tack and Chris Piech. 2022. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues. *arXiv preprint arXiv:2205.07540*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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, Yaqian Zheng, and Gaowei Chen. 2024. Chatgpt or bert? exploring the potential of chatgpt to facilitate preservice teachers' learning of dialogic pedagogy. *Educational Technology & Society*, 27(3):390–406.

Peiyao Wang, Joyce Fang, and Julia Reinspach. 2021. Cs-bert: a pretrained model for customer service dialogues. In *Proceedings of the 3rd workshop on natural language processing for conversational AI*, pages 130–142.

Rose E Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2023. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. *arXiv preprint arXiv:2310.10648*.

Christopher R Wolfe, Colin L Widmer, Valerie F Reyna, Xiangen Hu, Elizabeth M Cedillos, Christopher R Fisher, Priscilla G Brust-Renck, Triana C Williams, Isabella Damas Vannucchi, and Audrey M Weil. 2013. The development and analysis of tutorial dialogues in autotutor lite. *Behavior research methods*, 45:623–636.

Tianlong Xu, YiFan Zhang, Zhendong Chu, Shen Wang, and Qingsong Wen. 2025. 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*, volume 39, pages 28801–28809.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhenyu Zhang, Tao Guo, and Meng Chen. 2021. Dialoguebert: A self-supervised learning based dialogue pre-training encoder. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3647–3651.

A Prompt Used for Rephrasing

```
<task>
    <instruction>
        You are a helpful assistant for paraphrasing a target utterance in a dialogue.
        Your goal is to rewrite the <target> utterance in a different way while preserving its
     original meaning.
        The paraphrased version must be natural, fluent, and semantically equivalent.
        Make sure the paraphrase fits well within the conversation context, both before and after the
      target.

        Guidelines:
        1. Do NOT simply repeat the original sentence.
        2. Maintain the same intention, tone, and meaning.
        3. Ensure coherence with <previous_context> and <post_context>.
        4. Output only the paraphrased version of the <target>.

        Avoid repeating the same phrasing or word order.
    </instruction>

    <previous_context>
        {previous_context}
    </previous_context>

    <target>
        {target}
    </target>

    <post_context>
        {post_context}
    </post_context>
</task>
```

Appendix 1: XML prompt used for paraphrasing minority-class responses

B GPT-4.1 Prompt

We queried GPT-4.1 via the OpenAI API using a two-part prompt. The system message defined the instruction for each task, and the user message provided the specific conversation and tutor response to be evaluated.

System Instruction For Track 1

```
You are given a conversation between a tutor and a student. The last utterance is from the student
     and contains a mistake. The tutor then responds to it.

Your task is to evaluate whether the tutor's response recognizes the mistake in the student's
     utterance.

Use the following guidelines:

- "Yes": The mistake is clearly identified or recognized in the tutor's response.
- "To some extent": The tutor implies there may be a mistake, but does not state it clearly or seems
     uncertain.
- "No": The tutor does not acknowledge the mistake (e.g., simply answers the question without
     referencing the error).

Respond with exactly one of the following labels:
Yes
To some extent
No

Do not include any explanation or extra text.
```

Appendix 2: Track 1 system instruction used with GPT-4.1.

System Instruction For Track 2

```
You are given a conversation between a tutor and a student. The last utterance is from the student
    and contains a mistake. The tutor then responds to it.

Your task is to evaluate whether the tutor's response clearly identifies the mistake and where it
    occurs in the student's response.

Use the following guidelines:

- "Yes": The tutor clearly points to the exact location of a genuine mistake in the student's
    response.
- "To some extent": The tutor shows some awareness of the mistake, but the reference is vague,
    unclear, or easy to misunderstand.
- "No": The tutor does not provide any detail about the mistake or its location.

Respond with exactly one of the following labels:
Yes
To some extent
No

Do not include any explanation or extra text.
```

Appendix 3: Track 2 system instruction used with GPT-4.1.

System Instruction For Track 3

```
You are given a conversation between a tutor and a student. The last utterance is from the student
    and contains a mistake. The tutor then responds to it.

Your task is to evaluate whether the tutor's response provides correct and relevant guidance in
    response to the student's mistake.

Use the following guidelines:

- "Yes": The tutor provides guidance that is correct and directly relevant to the student's mistake (
    e.g., explanation, elaboration, hint, or examples).
- "To some extent": Some guidance is given, but it is partially incorrect, incomplete, or somewhat
    misleading.
- "No": No guidance is provided, or the guidance is irrelevant or factually incorrect.

Respond with exactly one of the following labels:
Yes
To some extent
No

Do not include any explanation or extra text.
```

Appendix 4: Track 3 system instruction used with GPT-4.1.

User Input Template

```
Conversation history:
{conversation}
Tutor Response:
{response}
```

Appendix 5: User input format for GPT-4.1 prompting.

C LLaMA 3.2 1B Prompt

We used a locally hosted version of LLaMA 3.2 1B in a zero-shot setting. The full prompt sent to the model followed the expected chat-style format, including system, user, and assistant messages, as shown below.

System Instruction For Track 1

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
Evaluate whether the tutor's response recognizes the student's mistake in the conversation.

Classification guidelines:
- "Yes": The tutor clearly identifies or acknowledges the mistake in the student's utterance.
- "To some extent": The tutor implies there may be a mistake, but the identification is vague or
    uncertain.
- "No": The tutor does not recognize the mistake (e.g., simply answers the question without
    acknowledging any error).
<|eot_id|>

<|start_header_id|>user<|end_header_id|>
Dialogue transcript:
{all_history}
Final tutor response:
{response_text_full}

Does the final tutor response recognize the student's mistake?

Only respond with one of the following labels:
Yes
To some extent
No
<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>
Classification:
```

Appendix 6: Track 1 full prompt used with LLaMA 3.2 1B for zero-shot classification.

System Instruction For Track 2

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
Evaluate whether the tutor's response clearly identifies a genuine mistake and its location in the
    student's utterance.

Classification guidelines:
- "Yes": The tutor clearly points to the exact location of a genuine mistake in the student's
    response.
- "To some extent": The response shows some awareness of the mistake, but the reference is vague,
    unclear, or potentially confusing.
- "No": The response does not mention the mistake or provide any detail about it.
<|eot_id|>

<|start_header_id|>user<|end_header_id|>
Dialogue transcript:
{all_history}
Final tutor response:
{response_text_full}

Does the tutor's response clearly identify the mistake and where it occurs?

Only respond with one of the following labels:
Yes
To some extent
No
<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>
Classification:
```

Appendix 7: Track 2 full prompt used with LLaMA 3.2 1B for zero-shot classification.

System Instruction For Track 3

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
Evaluate whether the tutor's response provides correct and relevant guidance in response to the
    student's mistake.

Classification guidelines:
- "Yes": The tutor provides guidance that is correct and directly relevant to the student's mistake (
    e.g., explanation, elaboration, hint, or example).
- "To some extent": Guidance is provided, but it is partially incorrect, incomplete, or somewhat
    misleading.
- "No": The response lacks guidance, or the guidance is irrelevant or factually incorrect.
<|eot_id|>

<|start_header_id|>user<|end_header_id|>
Dialogue transcript:
{all_history}
Final tutor response:
{response_text_full}

Does the tutor's response provide correct and relevant guidance?

Only respond with one of the following labels:
Yes
To some extent
No
<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>
Classification:
```

Appendix 8: Track 3 full prompt used with LLaMA 3.2 1B for zero-shot classification.