Phaedrus at BEA 2025 Shared Task: Assessment of Mathematical Tutoring Dialogues through Tutor Identity Classification and Actionability Evaluation

Rajneesh Tiwari
Independent Researcher, India
rajneesh.vish1@gmail.com

Pranshu Rastogi
Independent Researcher, India
rastogipranshu29@gmail.com

Abstract

As Large Language Models (LLMs) are increasingly deployed in educational environments, two critical challenges emerge: identifying the source of tutoring responses and evaluating their pedagogical effectiveness. This paper presents Phaedrus' comprehensive approach to the BEA 2025 Shared Task, addressing both tutor identity classification (Track 5) and actionability assessment (Track 4) in mathematical tutoring dialogues. For tutor identity classification, we distinguish between human tutors (expert/novice) and seven distinct LLMs using cross-response context augmentation and ensemble techniques. For actionability assessment, we evaluate whether responses provide clear guidance on student next steps using selective attention masking and instruction-guided training. Our multi-task approach combines transformer-based models with innovative contextual feature engineering, achieving state-of-the-art performance with a CV macro F1 score of 0.9596 (test set 0.9698) for identity classification and 0.655 (test set Strict F1 0.6906) for actionability assessment. We were able to score rank 5th in Track 4 and rank 1st in Track 5. Our analysis reveals that despite advances in human-like responses, LLMs maintain detectable fingerprints while showing varying levels of pedagogical actionability, with important implications for educational technology development and deployment. Our code and implementation details are publicly available at https://github.com/Rajneesh-Tiwari/BEA_2025_shared_task.

1 Introduction

The integration of Large Language Models (LLMs) into educational environments has created new opportunities and challenges for tutoring systems. As AI-powered tutors become increasingly prevalent, two fundamental questions emerge: (1) Can we reliably identify the source of tutoring responses to ensure transparency and accountability and (2) How effectively do these responses guide students toward learning objectives (Kochmar et al., 2022)

The BEA 2025 Shared Task (Kochmar et al., 2025) addresses these critical questions through two complementary tracks. Track 5 challenges participants to classify the source of mathematical tutoring responses, distinguishing between human tutors (expert and novice) and seven different LLMs: Gemini, GPT-4, Llama3-405B, Llama3-8B, Mistral, Phi3, and Claude Sonnet. Track 4 focuses on evaluating the actionability of these responses—whether they provide clear guidance on what students should do next, a crucial factor in effective pedagogical feedback (Daheim et al., 2024).

These tasks are inherently related: understanding who generated a response and how actionable it is provides a comprehensive view of educational dialogue quality. Our hypothesis is that different tutors (human or AI) not only leave distinctive linguistic fingerprints but also demonstrate varying capabilities in providing actionable guidance. This multi-dimensional analysis offers insights into the current state of AI tutoring systems and their pedagogical effectiveness compared to human tutors.

Our team (Phaedrus) approach leverages transformer-based models enhanced with task-specific innovations. For identity classification, we implement cross-response context augmentation, allowing models to compare different responses to the same question, and use specialized attention masking to focus on response characteristics. For actionability assessment, we develop instruction-guided training with selective attention mechanisms that focus on response-specific features indicating clear guidance. Both tasks benefit from sophisticated ensemble techniques and constraint satisfaction post-processing.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1098–1107

This paper presents our top-ranked systems for both tracks, describing our unified methodology, training strategies, and comprehensive analysis of results. Our findings demonstrate that while LLMs are becoming increasingly sophisticated in generating human-like responses, they still exhibit detectable patterns that distinguish them from human tutors, and they show varying capabilities in providing actionable pedagogical guidance.

2 Related Work

Recent research in educational dialogue assessment has focused on multiple dimensions of quality evaluation. Tack and Piech (2022) introduced a framework for evaluating LLM-based tutors across three dimensions: whether they speak like a teacher, understand a student, and help a student. Building on this work, Tack et al. (2023) organized shared task on generation of teacher responses in educational dialogues. The goal of the task was to benchmark the ability of generative language models to act as AI teachers, replying to a student in a teacher-student dialogue using existing automatic metrics (e.g., BERTScore (Zhang* et al., 2020), DialogRPT (Gao et al., 2020)) and manual evaluation aligned with the proposed dimensions, highlighting ongoing challenges in the reliable assessment of pedagogical dialogue quality.

2.1 Tutor Identity and AI Detection

The challenge of distinguishing between human and AI-generated text has established several foundations. Guo et al. (2023) demonstrated that transformer models effectively identify LLM-generated text through distinctive linguistic patterns, performing linguistic analysis to identify patterns between ChatGPT and human expert responses. In educational contexts specifically, Chen et al. (2024) introduced Dr.Academy, a benchmark for evaluating LLMs' questioning capabilities across general, humanities, science, and interdisciplinary educational domains, revealing that different models demonstrate varying strengths and distinctive patterns.

Our work extends these approaches by addressing a more complex classification problem: distinguishing not just between human and AI responses, but between multiple specific AI models and different types of human tutors in educational contexts.

2.2 Actionability and Pedagogical Effectiveness

The assessment of pedagogical effectiveness in tutoring responses has gained increasing attention. Macina et al. (2023) introduced MathDial, a dataset for mathematical tutoring dialogues, and evaluated tutor responses using coherence, correctness, and equitable tutoring criteria. Wang et al. (2024) assessed tutoring responses based on usefulness, care, and human-likeness, providing additional dimensions for evaluation.

Most relevant to our actionability assessment, Daheim et al. (2024) introduced a framework for evaluating tutoring responses that includes actionability as a key criterion, defining it as whether a response makes it clear what the student should do next. Their findings suggest that even state-of-the-art LLMs struggle to consistently provide actionable guidance in educational contexts.

2.3 Technical Approaches

For student response evaluation, Fateen and Mine (2023) compared in-context meta-learning and semantic score-based similarity approaches for automated short answer grading in Arabic, demonstrating different computational approaches to evaluating student responses. Additionally, Maurya et al. (2025) developed a comprehensive evaluation taxonomy for assessing LLM-powered AI tutors, highlighting distinctive features in AI-generated pedagogical interactions.

Our work builds upon these foundations while introducing novel techniques specifically tailored to both identity classification and actionability assessment in educational dialogues, including cross-response context augmentation, constraint satisfaction optimization, and instruction-guided training approaches.

3 Dataset and Task Overview

Both tracks utilize a unified dataset of mathematical tutoring dialogues (Maurya et al., 2025) combining MathDial (Macina et al., 2023) and Bridge (Wang et al., 2024) datasets with 300 development dialogues. Track 5 requires classifying responses into nine tutor categories (human expert/novice and seven LLMs), with each conversation containing unique tutor identifications. Track 4 evaluates response actionability using three categories (Yes/To some extent/No). Both tasks use exact macro F1 score as the primary evaluation metric.

4 Methodology

Our team (Phaedrus) approach combines multiple transformer-based models with task-specific architectural enhancements and ensemble techniques. We develop a unified framework that addresses both tutor identity classification and actionability assessment while leveraging shared components and complementary innovations.

4.1 Base Model Architecture

The core of our system utilizes several transformer variants, each selected for specific strengths in educational dialogue analysis:

- **DeBERTa-v3-large** (He et al., 2023): 24 layers, 1024 hidden size, 304M parameters
- **DeBERTa-v3-base** (He et al., 2023): 12 layers, 768 hidden size, 86M parameters
- **DeBERTa-v3-small** (He et al., 2023): 6 layers, 768 hidden size, 44M parameters
- **Longformer-base-4096** (Beltagy et al., 2020): 12 layers, 768 hidden size, 149M parameters, with efficient attention for long sequences
- **BigBird-RoBERTa-large** (Zaheer et al., 2021): 24 layers, 1024 hidden size, 340M parameters, with block sparse attention
- **Qwen-2.5-0.5B** (Qwen et al., 2025): 24 layers, 1024 hidden size, 0.5B parameters, featuring advanced positional embeddings and multi-query attention
- **Zephyr-7B-alpha** (Tunstall et al., 2023): 32 layers, 4096 hidden size, 7B parameters, based on Mistral architecture with sliding window attention

4.2 Shared Architectural Enhancements

Both tasks benefit from several common architectural innovations:

Response Tokenization and Selective Attention: We added special tokens `[R_START]` and `[R_END]` to explicitly mark tutor response boundaries. This enables custom attention masking that zeros out attention weights for tokens beyond the `[R_END]` marker, forcing models to focus specifically on response content rather than surrounding context.

Generalized Mean (GeM) Pooling: Instead of standard mean pooling, we implemented GeM pooling with a learnable parameter p to compute sequence-level representations. Given a sequence of hidden vectors $x = \{x_1, x_2, \ldots, x_n\}$, where each $x_i \in \mathbb{R}^d$ is the hidden representation of the i-th token and $n = |x|$ is the sequence length, GeM pooling is defined as:

$$\text{GeM}(x) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^p \right)^{1/p} \quad (1)$$

Here, the exponentiation and root are applied element-wise, and $p \in \mathbb{R}$ is a learnable parameter that controls the sharpness of the pooling operation.

Multi-Sample Dropout: Inspired by (Inoue, 2020), we implemented multi-sample dropout with varying rates (0.2 to 0.27) applied to the same representation, then averaged the results. This acts as an implicit ensemble, reducing variance without additional computational cost.

4.3 Task-Specific Innovations

4.3.1 Track 5: Identity Classification Enhancements

For tutor identity classification, we developed several specialized techniques:

Cross-Response Context Augmentation: Rather than treating each response in isolation, we concatenate all available responses to the same question from different tutors, creating rich comparative context. This allows models to learn distinctive patterns by seeing how different tutors address identical student queries.

Constraint Satisfaction Post-processing: We formulated the response classification task as a constraint satisfaction problem to ensure that each class is assigned at most once per conversation, reflecting the assumption that a tutor identity should not repeat in a single dialogue.

Let c denote a conversation with a set of responses $\mathcal{R}_c = \{r_1, r_2, \ldots, r_n\}$, and let $p_{r,j}$ represent the predicted probability that response $r \in \mathcal{R}_c$ belongs to class j, where $j \in \{0, 1, \ldots, 8\}$. Total there are 9 classes starting from 0 to 8 where class "0" is considered as "novice". We define binary decision variables $x_{r,j} \in \{0, 1\}$ indicating whether response r is assigned to class j. The objective is to maximize the total assignment confidence while satisfying the uniqueness constraint per class within each conversation:

$$\text{maximize} \quad \sum_{c} \sum_{r \in \mathcal{R}_c} \sum_{j=0}^{8} p_{r,j} \cdot x_{r,j} \tag{2}$$

$$\text{subject to} \quad \sum_{j=0}^{8} x_{r,j} = 1 \quad \forall r \in \mathcal{R}_c \tag{3}$$

$$\sum_{r \in \mathcal{R}_c} x_{r,j} \leq 1 \quad \forall j \in \{0, \ldots, 8\}, \forall c \tag{4}$$

$$x_{r,j} \in \{0, 1\} \tag{5}$$

We opted for a greedy algorithm due to its practical efficiency and implementation simplicity. By prioritizing responses with the highest prediction confidence and assigning them the most probable unassigned class, the method effectively resolves assignment conflicts with minimal computational cost. Empirical results show that this approach improves macro F1 scores by 2–3%, highlighting its effectiveness in enforcing consistent class assignments within conversations.

Algorithm 1 Constraint Satisfaction Algorithm

for all conversations c **do**
  $A_c \leftarrow \emptyset$ ▷ Set of already assigned classes in conversation c
  Sort responses $r \in \mathcal{R}_c$ by $\max_j p_{r,j}$ in descending order
  for all response r in sorted order **do**
    $\hat{j} \leftarrow \arg\max_{j \notin A_c} p_{r,j}$ ▷ Best unassigned class
    Assign class $\hat{j}$ to response r
    $A_c \leftarrow A_c \cup \{\hat{j}\}$
  end for
end for

4.3.2 Track 5: Meta-Model Ensemble with Pseudolabeling

Our Track 5 ensemble combines six transformer models through a sophisticated meta-modeling pipeline and was able to achieve 1st position Table 1:

1. **Base Model Predictions**: We collect class probability outputs from all six transformer models (54 features total)

2. **Feature Enhancement**: We augment with TF-IDF vectors, count vectors, linguistic features, and math-specific markers

3. **Gradient Boosting**: We train LightGBM, XGBoost, and CatBoost models on combined features

4. **Pseudolabeling**: High-confidence test predictions (probability > 0.85) are added to training data with constraint satisfaction

5. **Voting**: Final predictions use weighted voting across all meta-models

4.3.3 Track 4: Actionability Assessment Enhancements

For actionability assessment, we implemented instruction-guided training:

Actionability Criteria Instruction: We incorporated explicit actionability assessment criteria directly into model input:

```
Instruction: Analyze the tutor's
    response and determine if it
    provides actionable guidance to the
    student.
Classification Rules:
- Label as "Yes" if the response gives
    specific, clear instructions on what
     to do next
- Label as "To some extent" if the
    response hints at needed action but
    lacks specificity
- Label as "No" if the response only
    provides the answer without guidance
Remember: Focus on whether the response
    guides the student's next steps, not
     just whether it's correct.
```

4.3.4 Track 4: Optimized Weighted Ensemble

For Track 4, we developed a streamlined ensemble approach:

1. **Model-Level Weighting**: Global weights for each model applied to all class probabilities

2. **Model-Class Weighting**: Individual weights for each model-class combination (12 weights total)

3. **Threshold Optimization**: Class-specific probability thresholds to address class imbalance

4. **Hyperparameter Optimization**: Optuna-based optimization using macro F1 as the target metric

Table 1: Task 5 Leaderboard: Identity Classification

Rank	Team	Ex. F1	Ex. Acc
1	**Phaedrus**	0.9698	0.9664
2	SYSUpporter	0.9692	0.9657
3	Two Outliers	0.9172	0.9412
4	JInan_Smart Education	0.8965	0.8940
5	BLCU-ICALL	0.8930	0.8908

4.4 Training Strategy

Our team (Phaedrus) training strategy incorporated several techniques to maximize performance across both tasks:

Cross-Validation: We employed 5-fold Stratified Group K-Fold cross-validation, ensuring dialogues from the same conversation ID remained in the same fold to prevent data leakage while maintaining class distribution.

Hyperparameter Configuration: We used AdamW optimizer with weight decay of 0.003, learning rates ranging from 1e-5 to 3e-5 depending on model size, and OneCycleLR scheduler with maximum learning rate reached at 30% of training steps. For larger models, we implemented gradient accumulation with effective batch sizes of 16-32.

Task-Specific Input Formatting: To maximize ensemble diversity, we designed distinct input templates for different model architectures, each optimized for their specific attention mechanisms and training paradigms:

BERT-family Models (DeBERTa, Longformer, BigBird):

```
[Question] + [SEP] + [R_START] + [
    Response] + [R_END] + [SEP] + [
    Context]
```

Structured format with explicit token boundaries for enhanced attention control

Qwen-2.5 Model:

```
Track 5: Question: [Question]; Answer: [
    Response]; Context: [Context]
Track 4: Question: [Question]; Response:
     [Response]
```

Natural language format optimized for instruction-following capabilities

Zephyr-7B Model:

```
Question: [Question]; Answer: [Response]
```

Parameter-Efficient Fine-tuning: For larger models, we utilized Low-Rank Adaptation (LoRA) with model-specific configurations: Qwen-2.5 used rank=256/alpha=512 (Track 4) or rank=64/alpha=128 (Track 5), while Zephyr used rank=16/alpha=32. Models were quantized to 4-bit or bfloat16 precision to reduce memory requirements.

Early Stopping and Regularization: We implemented early stopping based on validation macro F1 score with patience of 3 epochs. Dropout rates were set to 0.1 for base models, with multi-sample dropout providing additional regularization through ensemble-like averaging.

5 Experiments and Results

5.1 Experimental Setup

We trained our models using 5-fold cross-validation with early stopping based on validation macro F1 score. Each model was trained for 25 epochs using AdamW optimizer with weight decay of 0.003 and OneCycleLR scheduler.

5.2 Track 5: Tutor Identity Classification Results

Table 2 presents the performance of our identity classification system.

Our Track 5 system achieved a macro F1 score of 0.9596, securing rank 1st 1 in the competition leaderboard. The results demonstrate several key findings:

1. **Cross-Response Context** provides the largest individual contribution, confirming that comparative information between different tutor responses is highly valuable for distinguishing tutor identities.

2. **Pseudolabeling** adds consistent improvement across all classes, with the largest gains for classes with fewer training examples.

3. **Ensemble Diversity** proves crucial, as each model contributes uniquely to final performance.

Model	Val Macro F1	Val Accuracy	LB Macro F1	LB Accuracy
DeBERTa-v3-base	0.8971	0.8901	NA	NA
DeBERTa-v3-large	0.8995	0.8914	NA	NA
Longformer-base	0.8945	0.8865	NA	NA
BigBird-RoBERTa-large	0.8761	0.8671	NA	NA
Qwen-2.5-0.5B	0.8938	0.8869	NA	NA
Zephyr-7B-alpha	0.8811	0.8740	NA	NA
LightGBM meta-model	0.9226	0.9172	0.9250	0.9263
+ Pseudolabeling	0.9585	0.9547	0.9604	0.9619
Final Ensemble	**0.9596**	**0.9560**	**0.9698**	**0.9664**

Table 2: Track 5 performance on validation set using 5-fold cross-validation and leaderboard results

Table 3: Task 4 Results: Actionability Assessment

Rank	Team	Ex. F1	Ex. Acc	Len. F1	Len. Acc
1	bea-jh	0.7085	0.7298	0.8527	0.8837
2	BJTU	0.6992	0.7363	0.8633	0.8940
3	MSA	0.6984	0.7537	0.8659	0.8908
4	lexiLogic	0.6930	0.7162	0.8393	0.8675
5	**Phaedrus**	0.6907	0.7298	0.8346	0.8656

5.3 Track 4: Actionability Assessment Results

Table 4 presents the performance of our actionability assessment system.

Our Track 4 system achieved a macro F1 score of 0.655, securing 5th Table 3 place on the competition leaderboard. The results reveal:

1. **Model-Class Weighting** outperforms simple model-level weighting, suggesting different models have strengths for different actionability categories.

2. **Instruction Guidance** significantly improves model understanding of actionability criteria.

3. **Middle Category Challenge**: The "To some extent" category shows lower performance, reflecting inherent ambiguity in partial actionability.

5.4 Feature Importance Analysis

Figure 1 shows feature importance from our Track 5 LightGBM meta-model, revealing model-class specialization patterns.

The analysis reveals that different architectures excel at detecting specific tutor identities, validating our multi-model ensemble approach. Each LLM leaves distinct "fingerprints" detectable by specialized transformer architectures.

6 Discussion

Our comprehensive approach to both tutor identity classification and actionability assessment provides valuable insights into the current state of AI tutoring systems and their relationship to human tutoring effectiveness.

6.1 Cross-Task Insights

The combination of both tasks reveals important patterns:

1. **Identity-Actionability Correlation**: Our analysis suggests that human expert tutors consistently receive higher actionability ratings than most LLMs, indicating that the source of a response correlates with its pedagogical effectiveness.

2. **LLM Differentiation**: Different LLMs show distinct patterns not only in linguistic fingerprints but also in their ability to provide actionable guidance. This suggests that model architecture and training approaches influence pedagogical capabilities.

3. **Detectability vs. Quality**: Despite LLMs' increasing sophistication in generating

Model	Val Macro F1	Val Accuracy	LB Macro F1	LB Accuracy
DeBERTa-v3-small	0.6169	0.7124	NA	NA
DeBERTa-v3-base	0.6262	0.7161	NA	NA
DeBERTa-v3-large	0.6360	0.7112	NA	NA
Qwen-2.5-0.5B	0.6387	0.7205	NA	NA
Model-level weights opt.	0.6536	0.7387	NA	NA
Model-class weights opt.	0.6548	0.7346	0.6836	0.7292
Final Ensemble	**0.6551**	**0.7350**	**0.6907**	**0.7298**

Table 4: Track 4 performance on validation set using 5-fold cross-validation and leaderboard results

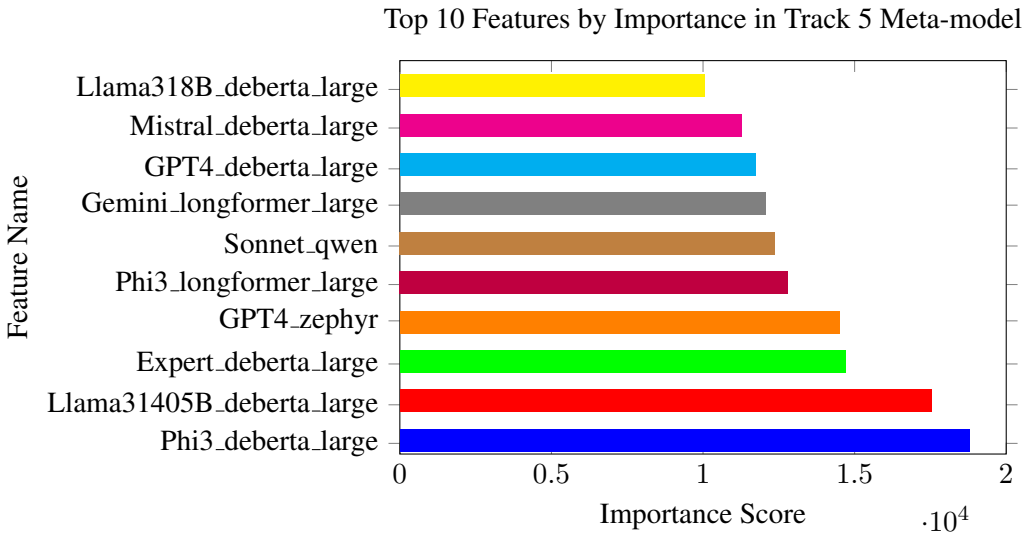

Figure 1: Top 10 feature importance scores showing model-class specialization in tutor identity. Each feature represents how confident a specific transformer architecture is in predicting a particular tutor identity. For example, "Phi3_deberta_large" indicates the DeBERTa-large model's probability output for the Phi3 LLM class classification.

human-like responses, they remain detectable through subtle patterns while showing varying quality in educational effectiveness.

6.2 Technical Contributions

Our methodology contributes several innovations to educational dialogue assessment:

1. **Cross-Response Context Augmentation**: This technique significantly improves identity classification by providing comparative information, suggesting that tutor identity is best understood in relation to alternative responses.

2. **Constraint Satisfaction Integration**: The post-processing approach for enforcing unique class assignments demonstrates how task-specific constraints can be integrated into neural classification systems.

3. **Instruction-Guided Training**: The explicit incorporation of assessment criteria into model input proves effective for actionability evaluation, suggesting broader applications for criterion-based classification tasks.

4. **Multi-Model Specialization**: Our feature importance analysis confirms that different transformer architectures capture complementary aspects of educational dialogues, supporting diverse ensemble approaches.

6.3 Educational Implications

The findings have significant implications for educational technology:

1. **Transparency and Accountability**: The ability to reliably identify AI vs. human tutoring responses enables better transparency in educational settings where students may not be aware of AI involvement.

2. **Quality Assurance**: Automated actionability assessment can provide real-time feedback to improve both human and AI tutoring responses, potentially enhancing educational outcomes.

3. **AI Development Guidance**: The identification of specific areas where LLMs fall short in actionability provides clear targets for improving AI tutoring systems.

4. **Hybrid Systems**: Understanding the complementary strengths of human and AI tutors can inform the design of hybrid systems that leverage the best aspects of both.

6.4 Methodological Insights

Our approach reveals several important methodological considerations:

1. **Task Complementarity**: The combination of identity classification and quality assessment provides a more comprehensive evaluation framework than either task alone.

2. **Context Importance**: Both tasks benefit significantly from contextual information, whether through cross-response comparison or instruction guidance.

3. **Ensemble Effectiveness**: Different ensemble strategies (meta-learning vs. weighted voting) prove optimal for different tasks, suggesting that ensemble design should be tailored to specific problem characteristics.

4. **Constraint Integration**: The successful integration of domain constraints (uniqueness) into neural models demonstrates the value of combining symbolic and connectionist approaches.

These findings collectively demonstrate that effective educational dialogue assessment requires sophisticated approaches that consider both the source and quality of responses, with important implications for the development and deployment of AI tutoring systems.

7 Acknowledgments

We would like to thank the organizers of the Pedagogical Ability Assessment of AI-powered Tutors shared task and the Building Educational Applications (BEA) 2025 workshop for running this competition. We also thank the anonymous reviewers for their insightful and constructive comments, which helped raise the standard of this manuscript considerably.

8 Limitations

While our multi-task approach achieved strong performance on both BEA 2025 Shared Task (Kochmar et al., 2025) tracks, several limitations should be noted:

1. **Domain Specificity**: Our models were trained and evaluated specifically on mathematical tutoring dialogues. Performance may not generalize to other educational domains with different discourse patterns or pedagogical requirements.

2. **Language and Cultural Constraints**: The dataset primarily consisted of English-language dialogues reflecting specific educational contexts. Performance on multilingual or cross-cultural tutoring scenarios remains untested.

3. **Temporal Limitations**: As LLMs continue to evolve rapidly, the distinctive patterns identified by our models may change. Future versions of the same LLMs might exhibit different characteristics, potentially reducing classification effectiveness.

4. **Computational Requirements**: Our approach relies on large transformer models and sophisticated ensemble techniques, requiring significant computational resources that may limit practical deployment in resource-constrained educational environments.

5. **Interpretability Challenges**: While our models achieve high classification accuracy, they provide limited insights into the specific linguistic or pedagogical features that drive classification decisions, making it difficult to extract actionable guidance for improving tutoring responses.

6. **Category Granularity**: The discrete categorization schemes may oversimplify complex phenomena—tutor identity includes many

sub-variations within categories, and actionability might be better represented as a continuum rather than discrete classes.

Future work could address these limitations by expanding to multiple educational domains, developing more efficient architectures, incorporating explainable AI techniques, and exploring the explicit modeling of cross-task relationships. Additionally, longitudinal studies tracking LLM evolution and cross-cultural validation would strengthen the generalizability of these approaches.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Yuyan Chen, Chenwei Wu, Songzhou Yan, Panjun Liu, and Yanghua Xiao. 2024. Dr.Academy: A benchmark for evaluating questioning capability in education for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3138–3167, Bangkok, Thailand. Association for Computational Linguistics.

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*, pages 8386–8411, Miami, Florida, USA. Association for Computational Linguistics.

Menna Fateen and Tsunenori Mine. 2023. In-context meta-learning vs. semantic score-based similarity: A comparative study in Arabic short answer grading. In *Proceedings of ArabicNLP 2023*, pages 350–358, Singapore (Hybrid). Association for Computational Linguistics.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response rankingtraining with large-scale human feedback data. In *EMNLP*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *Preprint*, arXiv:2301.07597.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Hiroshi Inoue. 2020. Multi-sample dropout for accelerated training and better generalization. *Preprint*, arXiv:1905.09788.

Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors. 2022. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. Association for Computational Linguistics, Seattle, Washington.

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

Jakub Macina, Nico Daheim, Sankalan 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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 522–529, Durham, United Kingdom. International Educational Data Mining Society.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.

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.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big bird: Transformers for longer sequences. *Preprint*, arXiv:2007.14062.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.