SYSUpporter Team at BEA 2025 Shared Task: Class Compensation and Assignment Optimization for LLM-generated Tutor Identification

Longfeng Chen[1]*, Zeyu Huang[1]*, Zheng Xiao[2], Yawen Zeng[3]†, Jin Xu[1,4]
[1]South China University of Technology, Guangzhou, China
[2]Peking University, Beijing, China
[3]ByteDance, Beijing, China
[4]Pazhou Lab, Guangzhou, China
{ftclf_dh, fthzy2024, jinxu}@mail.scut.edu.cn
zhengxiao@stu.pku.edu.cn, yawenzeng11@gmail.com

Abstract

In this paper, we propose a novel framework for the tutor identification track of the BEA 2025 shared task (Track 5). Our framework integrates data-algorithm co-design, dynamic class compensation, and structured prediction optimization. Specifically, our approach employs noise augmentation, a fine-tuned DeBERTa-v3-small model with inverse-frequency weighted loss, and Hungarian algorithm-based label assignment to address key challenges, such as severe class imbalance and variable-length dialogue complexity. Our method achieved **0.969 Macro-F1 score** on the official test set, securing second place in this competition. Ablation studies revealed significant improvements: a 9.4% gain in robustness from data augmentation, a 5.3% boost in minority-class recall thanks to the weighted loss, and a 2.1% increase in Macro-F1 score through Hungarian optimization. This work advances the field of educational AI by providing a solution for tutor identification, with implications for quality control in LLM-assisted learning environments.

1 Introduction

The rapid advancement of large language models (LLMs) has opened new avenues for the development of AI-powered tutoring systems, enabling scalable and personalized learning support through intelligent conversational agents (Cai et al., 2025; Li et al., 2025). Contemporary studies demonstrate that AI-powered tutors can significantly enhance instructional efficiency (Tack et al., 2023), particularly in math education where adaptive feedback is crucial (Xu et al., 2025). Nevertheless, this technological progress introduces a critical challenge in educational practice: the growing difficulty in distinguishing LLM-generated tutor responses from those crafted by human educators. This tutor identification problem becomes particularly acute when examining nuanced pedagogical behaviors such as error correction strategies and instructional scaffolding (Macina et al., 2023).

The emergence of sophisticated LLM-based tutors has blurred the traditional boundaries between human and machine-generated educational content. While existing detection methods (Sanh et al., 2019; Liu et al., 2019) perform adequately in binary human-vs-LLM classification scenarios, they lack the granularity required for educational applications. Specifically, these approaches fail to differentiate between various state-of-the-art LLM architectures, distinguish expert versus novice human instructors, or identify the pedagogical strategies employed by different tutor types. This limitation becomes particularly problematic given the demonstrated variations in educational outcomes based on tutor quality.

The shared task (Kochmar et al., 2025) of "Pedagogical Ability Assessment of AI-powered Tutors" (Track 5: Tutor Identification) presents three main technical challenges: **1) Class Imbalance.** Severe class imbalance in the dataset's sample distribution across nine tutor categories (Maurya et al., 2025), **2) Complexity of Dialogue Sequences.** The complexity of variable-length dialogue sequences that complicate feature extraction, and **3) Subtle Linguistic Patterns.** Minimal lexical differences between expert humans and advanced LLMs that create subtle linguistic patterns. These characteristics render conventional classification approaches ineffective, particularly in maintaining performance across minority classes.

To address class imbalance and enhance classification performance, we employ a noise injection strategy for data augmentation, coupled with a two-stage class weight compensation mechanism. The model is fine-tuned using weighted cross-entropy loss with inverse-frequency class weighting to mitigate bias toward majority classes. For prediction, we implement an ensemble approach combining

*These authors contributed equally.
†Corresponding author.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1078–1083

multiple pre-trained models, followed by globally optimal label assignment via the Hungarian algorithm to ensure unique label distribution per dialogue group while maximizing prediction confidence (Kuhn, 1955). This comprehensive approach effectively handles class imbalance while maintaining prediction stability.

Our method achieved 0.969 Macro-F1 on the official test set, securing **second place in this competition**. Ablation studies[1] demonstrate component-wise improvements: a 5.3% boost in minority-class recall thanks to the weighted loss, and a 2.1% increase in Macro-F1 score through Hungarian optimization. The remainder of this paper is structured as follows: Section 2 reviews relevant literature in LLMs and text detection. Section 3 formally defines the tutor identification problem and introduces dataset. Section 4 formalizes our technical approach. Section 5 presents empirical results and case analyses. Finally, we conclude with broader implications and future directions in Section 6.

2 Related Work

LLM-generated Text Detection. The proliferation of large language models (LLMs) has spurred interest in detecting LLM-generated text. Following the emergence of the GPT-2 Output Detector (Solaiman et al., 2019), which is based on the RoBERTa pretrained model (Liu et al., 2019) and achieves up to 88% accuracy on GPT-2 text, numerous detectors have been developed. **?** employs statistical analysis of word probabilities and ranks for GPT-2 detection. Habibzadeh (2023) initially used perplexity and burstiness, claiming 88% accuracy for human and 72% for AI text. OpenAI's Text Classifier[2], fine-tuned on diverse models, provides probabilistic categories for distinguishing human and AI text, requiring at least 1000 characters. GPTKit[3] sets up multiple models (including (Sanh et al., 2019; Liu et al., 2019)). CheckForAI[4] combines GPT-2 Output Detector with custom models. CopyLeaks[5] claims 99.12% accuracy across languages.

In contrast to general LLM-generated text detectors, our work focuses on the more nuanced task of identifying the specific origin of text within a defined set of tutors and LLMs. To achieve this, we leverage DeBERTa (He et al., 2020), which features disentangled attention and an enhanced mask decoder. DeBERTa has demonstrated superior performance in NLP tasks, achieving an accuracy of 91.1% on the MNLI benchmark, compared to RoBERTa-Large's 90.2%. These results make DeBERTa a promising approach for our classification task.

3 Dataset Analysis

The dataset provided for this shared task (Kochmar et al., 2025) is sourced from the MathDial (Macina et al., 2023) and Bridge (Wang et al., 2023) datasets. The dataset, including instructional annotations developed by Maurya et al. (2025), was provided by the shared task organizers in accordance with the established annotation protocol and guidelines. Out of 300 dialogues, 200 responses were annotated by four annotators. The average Fleiss' Kappa among the four annotators reached 0.65, indicating substantial agreement and demonstrating the reliability of this annotation task. Each dialogue includes the prior multi-turn interactions between a tutor and a student, the student's final utterance containing an error, and a collection of responses generated by both seven large language model LLM-based tutors and human tutors in response to that utterance. The LLM tutors include: GPT-4 (Achiam et al., 2023), Gemini (Team et al., 2024), Sonnet (Anthropic, 2023), Mistral (Jiang et al., 2023), Llama-3.1-8B and Llama-3.1-405B (Grattafiori et al., 2024), and Phi-3 (Abdin et al., 2024). Human tutors are categorized into two groups: Expert and Novice.

The test set consists of 191 dialogues. These dialogues include the prior conversational context, the final incorrect student utterance, and a set of unannotated tutor responses from the same group of tutors used in the development set.

For Track 5: tutor identification task, the required data include the tutor responses and their corresponding identities. Table 1 presents the distribution of the dataset.

4 Methodology

As shown in Figure 1, we propose a unified approach to address class imbalance and enhance classification performance. It combines noise injection for data augmentation, a two-stage class weight compensation mechanism. During infer-

[1] Ablation studies conducted with simplified validation due to submission constraints

[2] https://platform.openai.com/ai-text-classifier

[3] https://gptkit.ai/

[4] https://checkforai.com/

[5] https://copyleaks.com/

Class	Train Set Count	Test Set Count
Expert	300	191
Novice	76	19
Sonnet	300	191
Llama3.1-8B	300	191
Llama3.1-405B	300	191
GPT4	300	191
Mistral	300	191
Gemini	300	191
Phi3	300	191
Total	**2,476**	**1,547**

Table 1: The statistics of the dataset in track 5.

ence, we employ an ensemble of pre-trained models and apply the Hungarian algorithm for globally optimal and unique label assignment within each dialogue group. This ensures both robustness and stable prediction under imbalanced conditions.

4.1 Noise Injection for Data Augmentation

We selected several commonly used machine learning models along with the DeBERTa series for evaluation. The original training dataset was partitioned into training and validation subsets with an 8:2 ratio to facilitate comparable performance assessment. We adopted both Macro-F1 score and accuracy (ACC) as evaluation metrics. Considering that Macro-F1 demonstrates greater robustness to class imbalance, it was designated as our primary evaluation criterion. The comparative results for both metrics are presented in Table 2. Based on these experimental findings, we selected DeBERTa-v3-small for further fine-tuning to enhance its classification performance.

Model	Validation Set	
	Macro-F1 Score	Accuracy
Logistic Regression	0.796	0.811
Random Forest	0.778	0.789
Extra Trees	0.786	0.798
XGBoost	0.736	0.757
DeBERTa-v3-base	0.806	0.821
DeBERTa-v3-small	**0.812**	**0.834**

Table 2: Performance comparison of baselins on the validation set (Macro-F1 Score and Accuracy).

Subsequent analysis of the validation set predictions revealed a notable discrepancy between the model's accuracy and Macro-F1 scores. While achieving high accuracy, the model exhibited relatively poor performance in terms of Macro-F1, suggesting inadequate handling of class imbalance. This observation indicates that the current model

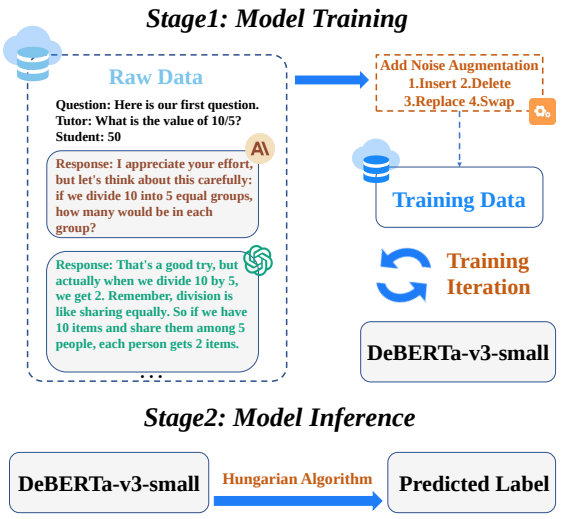

Figure 1: Overview of our proposed method.

architecture may require modification to better address the imbalanced nature of our dataset

Therefore, the original dataset is expanded through multimodal noise injection to mitigate overfitting in small-sample scenarios. For each text sample x_i, we generate its noisy variant $\tilde{x}_i$ as follows:

$$\begin{aligned} &\tilde{x}_i = T(x_i), \\ &T \in \{\text{insert}, \text{delete}, \text{replace}, \text{swap}\}, \end{aligned} \tag{1}$$

where the noise transformation T is randomly selected with uniform probability from four operations, with a noise ratio $\alpha = 10\%$. This augmentation strategy doubles the dataset size from original N samples to $2N$. Crucially, the original labels remains unaltered during augmentation, preserving consistency in label distribution.

To address potential amplification of original class distribution disparities, we implement a two-stage class weight compensation mechanism:

4.2 Fine-tuning DeBERTa with Weighted Cross-Entropy Loss Function

To address class imbalance in the training set, we adopt an inverse-frequency weighting scheme to compute balanced class weights. Let the training set consist of C classes, with N_c denoting the number of samples in class c, and let $N_{\text{total}} = \sum_{c=1}^{C} N_c$ be the total number of training samples. The weight for class c is defined as:

$$w_c = \frac{N_{\text{total}}}{C \cdot N_c}, \quad c = 1, \ldots, C. \tag{2}$$

This weighting strategy assigns higher importance to underrepresented classes, thereby mitigating the bias toward majority classes during model training.

Subsequently, the standard cross-entropy loss is modified by incorporating the computed class weights. Given a training batch of size B, the weighted cross-entropy loss is formulated as:

$$\mathcal{L}(\theta) = -\frac{1}{B}\sum_{i=1}^{B} w_{y_i} \log p_\theta(y_i|x_i), \tag{3}$$

where y_i denotes the true label of sample x_i, and $p_\theta(y_i|x_i)$ represents the predicted probability output by the model parameterized by θ.

By scaling the loss contribution of each sample according to its class weight, this approach enhances the gradient contributions from minority classes while preserving the overall optimization direction. As a result, the classification boundary becomes more sensitive to underrepresented classes, leading to improved generalization performance on imbalanced datasets.

To ensure the training effectiveness of the model, we adopt K-fold cross-validation, a robust model evaluation technique that not only maximizes the utilization of limited datasets but also reduces the dependency of evaluation results on data partitioning methods (Kohavi et al., 1995), to assess and optimize the detection model's performance. The original training set is randomly divided into K subsets of approximately equal size. For each iteration, one subset is selected as the validation set, while the remaining K-1 subsets are used as the training set. The model's performance is ultimately assessed by aggregating the results from the K training and validation cycles.

4.3 Prediction via Hungarian Algorithm

Given an input text set $X = \{x_1, x_2, \ldots, x_n\}$, we employ k pre-trained models for prediction and average their output probabilities to mitigate the limitations of individual models, enhance generalizability, reduce prediction variance while preventing overfitting. Each model outputs a probability distribution matrix $P_i \in \mathbb{R}^{n\times c}$, with c denoting the number of classes. During the ensemble phase, we compute the average probability across all models:

$$\bar{P} = \frac{1}{k}\sum_{i=1}^{k} P_i. \tag{4}$$

This strategy effectively reduces model bias and enhances prediction stability. Through further analysis, we observe that each dialogue group consistently contains 7 AI responses, 1 Expert response, and randomly features 1 Novice response. Based on this pattern, we design a Hungarian algorithm-based prediction method to ensure globally optimal unique label assignment for each dialogue group. The detailed procedure is as follows:

Step 1: Cost Matrix Construction For each dialogue group $G \subseteq X$, extract its average probability matrix $\bar{P}_G \in \mathbb{R}^{m\times c}$, where $m \in \{8, 9\}$ represents the number of responses in the group. When $m = 8$, we exclude the `Novice` label (class 9) and adjust the probability matrix to $\bar{P}'_G \in \mathbb{R}^{8\times 8}$. The cost matrix is defined as:

$$C = -\log(\bar{P}_G). \tag{5}$$

This transformation converts the probability maximization problem into a linear assignment problem that minimizes negative log probabilities.

Step 2: Optimal Matching Solution The Kuhn-Munkres (Hungarian) algorithm is applied to solve:

$$\min \sum_{i=1}^{m}\sum_{j=1}^{c'} C_{i,j} \cdot Z_{i,j}, \tag{6}$$

subject to the constraints:

$$\sum_i Z_{i,j} \le 1, \quad \sum_j Z_{i,j} = 1, \quad Z_{i,j} \in \{0, 1\}, \tag{7}$$

where Z denotes the assignment matrix, and $c' = c$ (when $m = 9$) or $c' = c - 1$ (when $m = 8$).

Step 3: Label Mapping and Confidence Calculation The algorithm returns optimal matching indices (i, j), mapping column index j back to the original label (when $m = 8$, adjustment is needed to skip the `Novice` label). The final predicted label and confidence score are:

$$\text{label} = \arg\max_j \bar{P}_{i,j}, \quad \text{confidence} = \max_j \bar{P}_{i,j} \tag{8}$$

This strategy achieves global optimal assignment with polynomial time complexity $O(m^3)$, ensuring label uniqueness while maximizing prediction confidence.

5 Main Results

The evaluation results of all models are summarized in Table 3. It is worth noting that the De-

Model	Macro-F1 Score	Accuracy
DeBERTa-v3-small	0.812	0.834
+ Augmentation, k=1	0.888	0.888
+ Augmentation, k=5	0.901	0.901
+ Augmentation + Weighted, k=5	0.949	0.963
+ Augmentation + Weighted + Hungarian, k=5	**0.969**	**0.966**

Table 3: Performance comparison of DeBERTa-v3-small under different training strategies and ensemble settings. Among them, *Augmentation* refers to noise injection for data augmentation techniques, *k* denotes the number of candidates averaged during inference, *Weighted* indicates the use of weighted cross-entropy loss, and *Hungarian* refers to prediction via Hungarian algorithm.

BERTa model without noise injection for data augmentation was not submitted to the CodaLab platform. Instead, its performance was evaluated using a validation set composed of 20% of the original training data, as described in Section 4.1 on data pre-processing. The results of the other four models were obtained using the official test set via the CodaLab evaluation platform.

The DeBERTa model without any noise injection for data augmentation reflects the baseline performance of the model under the original imbalanced data distribution. After introducing noise injection for data augmentation strategies, the Macro-F1 score improved to 0.888, indicating the initial effectiveness in mitigating the impact of class imbalance. Subsequently, we applied 5-fold cross-validation to the DeBERTa model and selected the best-performing model across the folds, which further increased the Macro-F1 score to 0.901, demonstrating improved stability and generalization capability.

Building upon this, the incorporation of a weighted cross-entropy loss function led to an additional improvement in performance, with the Macro-F1 score reaching 0.949. Finally, by integrating the Hungarian algorithm for prediction optimization, the overall Macro-F1 score achieved a significant improvement, reaching 0.969. This result confirms the effectiveness of the proposed approach in addressing complex classification tasks. Our best-performing model ranked second on the official leaderboard.

6 Conclusion

In this work, we propose an effective framework for distinguishing between human-written and LLM-generated responses in mentor-style answers. Our method is based on the DeBERTa model and incorporates various techniques to enhance its generalization and robustness, including data augmentation strategies, a weighted cross-entropy loss function design, and a prediction optimization mechanism based on the Hungarian algorithm. This proposed approach effectively addresses the challenges posed by the rapid development of generative artificial intelligence in content authentication.

Experiments are conducted on the test set provided by the Codabench platform, and the results validate the superior performance of the framework. Furthermore, this study presents a component analysis that explores the contribution of each module to the overall performance, offering valuable insights and directions for future research and improvements in related fields.

Limitations

We still have the following limitations: 1) In terms of generalization, our method is tailored to the tutor identification task, raising questions about its generalizability to similar tasks. We plan to address this issue of generalization in future work. 2) Furthermore, although our method has demonstrated excellent performance on this competition's test set, it has not yet been tested in real-world scenarios. We plan to apply and evaluate our method in the educational field and will share our findings when appropriate.

Acknowledgments

The authors sincerely appreciate the event organizers for their hard work. , and the reviewers for their careful reading and insightful comments.

This work is supported in part by the National Natural Science Foundation of China (62372187), in part by the National Key Research and Development Program of China (2022YFC3601005) and in part by the Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004).

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2023. The claude 3 model family: Opus, sonnet, haiku. Artificial Intelligence Model.

Leng Cai, Junxuan He, Yikai Li, Junjie Liang, Yuanping Lin, Ziming Quan, Yawen Zeng, and Jin Xu. 2025. Rtbagent: A llm-based agent system for real-time bidding.

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*.

Farrokh Habibzadeh. 2023. Gptzero performance in identifying artificial intelligence-generated medical texts: a preliminary study. *Journal of Korean medical science*, 38(38).

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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 (BEA)*.

Ron Kohavi and 1 others. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Xiangyu Li, Yawen Zeng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. 2025. Hedgeagents: A balanced-aware multi-agent financial trading system.

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

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, and 1 others. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

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*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

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*.

Liangyu Xu, Yingxiu Zhao, Jingyun Wang, Yingyao Wang, Bu Pi, Chen Wang, Mingliang Zhang, Jihao Gu, Xiang Li, Xiaoyong Zhu, Jun Song, and Bo Zheng. 2025. Geosense: Evaluating identification and application of geometric principles in multimodal reasoning.