Jinan Smart Education at BEA 2025 Shared Task: Dual Encoder Architecture for Tutor Identification via Semantic Understanding of Pedagogical Conversations

Lei Chen
Guangdong Institute of Smart Education, Jinan University
Guangzhou, 510632, Guangdong, China
leibnizchen@foxmail.com

Abstract

With the rapid development of smart education, educational conversation systems have become an important means to support personalized learning. Identifying tutors and understanding their unique teaching style are crucial to optimizing teaching quality. However, accurately identifying tutors from multi-round educational conversation faces great challenges due to complex contextual semantics, long-term dependencies, and implicit pragmatic relationships. This paper proposes a dual-tower encoding architecture to model the conversation history and tutor responses separately, and enhances semantic fusion through four feature interaction mechanisms. To further improve the robustness, this paper adopts a model ensemble voting strategy based on five-fold cross-validation. Experiments on the BEA 2025 shared task dataset show that our method achieves 89.65% Marco-F1 in tutor identification, ranks fourth among all teams(4/20), demonstrating its effectiveness and potential in educational AI applications. We have made the corresponding code publicly accessible at `https://github.com/leibnizchen/Dual-Encoder`.

1 Introduction

This paper will introduce in detail the methods and experiments on mentor identification in the BEA 2025 shared task (Ekaterina et al., 2025).

Different teachers show unique language characteristics and guidance preferences in practice, including dimensions such as expression methods, guidance techniques, and feedback patterns. These differences exist not only in the surface language form, but also in the information architecture and semantic logic of the feedback content. If the tutor's identity can be accurately recognized and their teaching quality evaluated, it would not only help analyze and optimize teaching styles but also provide strong support for improving teaching quality and instructional methods (Gan et al., 2023).

However, teaching dialogues are highly temporal dynamics. Semantic evolution, problem progression, and students' cognitive trajectories will have a profound impact on the generation of feedback in the current round. There are often complex pragmatic connections between teacher responses and contexts, which are difficult to model through explicit rules, which poses a great challenge to identity recognition. In recent years, natural language processing technology has shown great potential in semantic understanding and generation, providing new ideas for teaching context modeling and personalized feedback generation. However, to accurately portray teacher style, there are still problems such as data scarcity, identity generalization, and style transfer (Liu et al., 2019; He et al., 2023).

To address the above problems, this paper proposes a dual-tower encoding structure that integrates identity perception and context modeling capabilities for tutor identity recognition based on the characteristics of teaching conversation. This method extracts semantic features from the conversation context and tutor responses respectively, and designs four feature interaction mechanisms to enhance semantic fusion capability. Furthermore, we propose a voting strategy based on 5-fold cross-validation, in which the best-performing model from each fold is selected, and final identity recognition is completed through ensemble voting to improve the stability and robustness of the model.

The main contributions of this paper are as follows:

- A dual-tower encoding architecture is proposed to separate the semantic modeling processes of conversation context and tutor response, enhancing the recognition ability of personalized teaching styles.
- A Feature Interaction Modeling is designed, to overcome the limitations of traditional dual-tower models that rely solely on concatenation

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1034–1039

or similarity measures.

- A model ensemble voting strategy based on the optimal models from 5-fold cross-validation is introduced to effectively improve tutor identification accuracy and the generalization ability of the model.

Experimental results on the BEA 2025 Shared Task [1]dataset (Maurya et al., 2025) show that the proposed method achieves 89.65% Macro-F1 in the tutor identification task, verifying its effectiveness and potential for application in smart education.

2 Related Work

2.1 LLM-Powered AI Tutors

Educational conversation teaching systems have made significant progress in the field of natural language processing (NLP). Qiang (2025) proposed key technologies based on recurrent neural networks (Transformers) (Vaswani et al., 2017), reinforcement learning, and multimodal learning analysis, demonstrating the application potential of these technologies in personalized learning path recommendation and adaptive content generation. (Mansur et al., 2019)proposed a personalized learning model based on deep learning algorithms to explore the most suitable learning strategies for students. The model fully considered the key factors of personalized learning during the construction and testing process, including adaptability, personalization, differentiation, and ability-oriented learning. (Gan et al., 2023)proposed an intelligent tutoring system based on a large language model (LLM) to improve students' performance. (Cain, 2024; Makharia et al., 2024) used advanced prompt engineering techniques to deploy language models as intelligent tutors to improve the personalization and interactivity of teaching.

2.2 Contextual Content Understanding

Context understanding is the core challenge of effectively modeling long-range dependencies and capturing subtle semantic relationships in the context. Early methods such as recurrent neural networks (RNNs) laid the foundation for sequence modeling, but often suffered from problems such as gradient vanishing and limited context preservation capabilities. The emergence of the Transformer architecture (Vaswani et al., 2017) introduced the self-attention mechanism, which significantly improved the ability to capture global context information. On this basis, pre-trained language models such as BERT (Devlin et al., 2019) and its variants (RoBERTa) (Liu et al., 2019), DebertaV3 (He et al., 2023)) have become standard tools for deep semantic understanding in a wide range of tasks. To more effectively handle longer contexts, models such as Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2021) adopt sparse attention mechanisms. To further enhance context modeling, researchers have incorporated external knowledge through models like K-BERT, integrated memory mechanisms such as those used in Memory Networks and Transformer-XL (Dai et al., 2019), and improved coreference resolution with models like SpanBERT (Joshi et al., 2020). Despite these advances, several challenges remain, including handling semantic ambiguity, preserving long-range dependencies, mitigating context truncation, and enabling complex multi-hop reasoning.

3 Methods

The dual encoder architecture is widely used in long text information matching. Base on the work of (Wang et al., 2023; Guo et al., 2024), we proposed a dual encoder architecture for tutor identification via Semantic Understanding of Pedagogical Conversations model, the core structure of which is shown in Figure 1. The model captures the deep semantic representation of the conversation history and tutor responses through independent bidirectional encoders, and adopts a multimodal feature fusion strategy to achieve fine-grained semantic interaction modeling.

3.1 Dual encoder architecture

The model uses a dual Transformer encoder structure with independent parameters, which are defined as history encoder $E_h(\cdot)$ and response encoder $E_h(\cdot)$. Given the input sequence (conversation history) $\{h_i\}_{i=1}^{L}$ and (tutor response) $\{r_j\}_{j=1}^{M}$, the context-aware semantic representation is obtained through the pre-trained language model:

$$H = E_h(\mathrm{Emb}(h_1, ..., h_L)) \in \mathbb{R}^{L \times d} \quad (1)$$

$$R = E_r(\mathrm{Emb}(r_1, ..., r_L)) \in \mathbb{R}^{M \times d} \quad (2)$$

Where $d = 768$ is the hidden layer dimension, L and M represent the length of the conversation history and the tutor response, respectively, and $Emb(\cdot)$ represents the word embedding layer. To

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

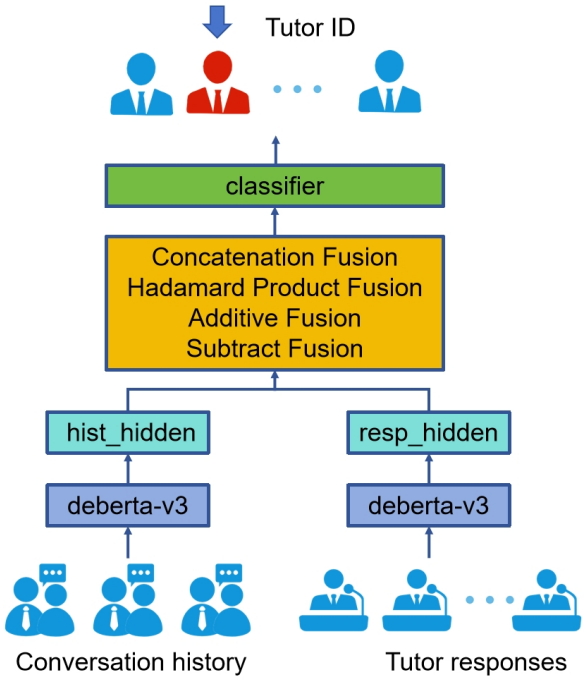

Figure 1: The core structure of Dual Encoder Architecture.

obtain a global semantic representation, we extract the hidden states from the first layer of DeBertaV3. This step provides a lightweight yet informative semantic encoding, which will serve as the foundation for downstream tasks,the formula is as follows:

$$\mathbf{h} = H[:, 0, :] \in \mathbb{R}^d \tag{3}$$

$$\mathbf{r} = R[:, 0, :] \in \mathbb{R}^d \tag{4}$$

3.2 Feature Interaction Modeling

In order to effectively model the deep semantic association between the conversation history and the tutor's response, we designed a multi-dimensional feature fusion mechanism. This mechanism aims to integrate the conversation context information and the response representation from multiple semantic perspectives. We found that relying on a single feature fusion strategy, such as concatenation or addition, has limited performance when dealing with complex semantic relationships and is difficult to fully capture potential semantic interaction information. The ablation experiment section provides proof. To overcome this problem, we constructed the following four complementary fusion strategies from the perspective of information redundancy control and semantic complementarity enhancement:

- Concatenation Fusion: Concatenation fusion is a basic and widely used feature integration method that directly splices the conversation history vector h with the tutor response vector r, retaining all the semantic information in the original representation:

$$f_c = [h; r] \in \mathbb{R}^{2d} \tag{5}$$

- Hadamard Product: The Hadamard product is an effective method for modeling nonlinear interactions between features. The fusion result retains strong activation only when the corresponding dimensions of the two feature vectors have high values:

$$f_m = h \odot r \in \mathbb{R}^d \tag{6}$$

- Additive Fusion: Its main function is to capture semantic commonality and consistency. Unlike concatenation and fusion, the addition operation emphasizes the relative direction and consistency of two vectors in the semantic space:

$$f_a = h + r \in \mathbb{R}^d \tag{7}$$

- Subtract Fusion: It is used to characterize the semantic difference between two vectors. In conversation modeling, difference features often carry key information to distinguish valid and invalid responses:

$$f_s = abs(h - r) \in \mathbb{R}^d \tag{8}$$

The final joint representation is:

$$f = [f_c; f_m; f_a; f_s] \in \mathbb{R}^{5d} \tag{9}$$

3.3 Classifier Design

The feature vector is mapped to dimension reduction through a cascade of processing modules:

$$y = W_2(LayerNorm(ReLu(W_1 f + b_1))) + b_2 \tag{10}$$

Where $W_1 \in \mathbb{R}^{5d \times 256}$,$W_2 \in \mathbb{R}^{C \times 256}$, C is the number of categories. The processing flow is implemented through a three-layer cascaded architecture.

4 Experiment

This section verifies the effectiveness of the model through systematic experiments, adopts a five-fold cross-validation strategy to ensure the reliability of the evaluation, and analyzes the contribution of key components through ablation experiments.

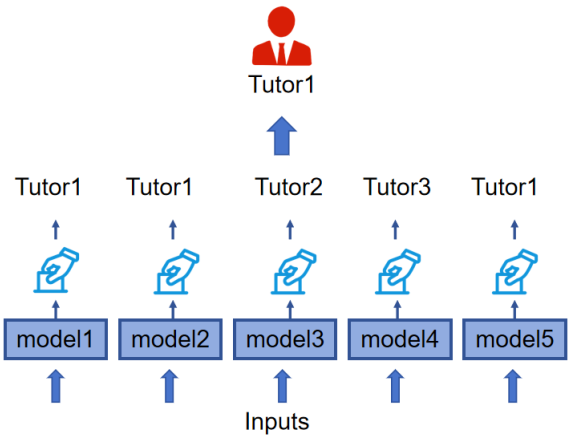

Figure 2: 5-fold cross validation model ensemble.

4.1 Dataset

Train set: 300 teaching scenario conversations provided by BEA 2025 Shared Task(Maurya et al., 2025). Each teaching scenario Conversation has at most 9 tutor responses, with a total of 2476 responses. We randomly divide tutor replies into training set/validation set at a ratio of 4:1. Test set: 191 teaching scenario conversations, including 1547 responses with unknown tutor identity information.

4.2 Five-fold Cross Validation Strategy

In order to systematically evaluate the generalization performance of the model and effectively suppress overfitting, this study adopts a stratified five-fold cross-validation framework. Cross-validation process:

- Iterative validation: Each subset is designated as the validation set in turn, and the remaining four subsets are merged into the training set to complete five rounds of independent training and validation processes.
- Model selection: Continuously monitor the performance of the validation set during each round of training, and the model weight parameters corresponding to the highest Macro-F1 score are retained.
- Cyclic validation: Through five complete iterations, it is ensured that each sample participates in the validation process exactly once.

This scheme obtains robust model parameters through cross-validation, and effectively improves the accuracy and stability of model prediction by combining ensemble learning strategies.

fold	Macro-F1(%)	Accuracy(%)
1	0.8903	0.8891
2	0.9103	0.9023
3	0.9012	0.8993
4	0.8890	0.8922
5	0.8997	0.9013
Average	0.8981± 0.87%	0.8968± 0.59%

Table 1: Results of five-fold cross validation on the training set/validation set.

4.3 Five-fold Cross Validation Experimental Results

Our method is stable across training/validation and final test sets. Table 1 shows the detailed performance of the model in the five-fold cross validation. The experimental results show that the model exhibits strong stability and robustness under different data partitions In the five-fold cross validation, the mean of Macro-F1 reached 0.8981, the standard deviation was only ±0.87%, and the fluctuation range was controlled within 2.13 percentage points; the standard deviation of Accuracy was ±0.59%, which further verified the robustness of the model in dealing with changes in data distribution. This provides a feasibility basis for the model integration method.

4.4 Feature Interaction Ablation Experiment

Table 2 shows the ablation experiment results of the model fusion mechanism, which shows the impact of different fusion strategies on model performance (Macro-F1 and Accuracy). It includes both individual usage and removal of four fundamental fusion operations: concatenation, Hadamard product, addition, and subtraction. As can be seen from the table: using a single fusion method leads to slightly lower performance compared to the full model. Among them, Subtract-only achieved relatively high performance (Macro-F1 0.8849, Accuracy 0.8845), showing its effectiveness in capturing differences. Removing individual fusion methods also results in performance drops. Among them, the performance decrease caused by removing the Hadamard fusion (w/o Hadamard) is more obvious (Macro-F1 0.8832, Accuracy 0.8815), indicating that Hadamard plays an important role in capturing feature interactions. The full model performs best in all indicators, with Macro-F1 reaching 0.8981, Accuracy 0.8968, and a small standard deviation, which verifies that the synergy of each fusion oper-

fusion methods	Dimension	Macro-F1(%)	Accuracy(%)
Concatenation-only	2d	0.8811 ± 0.67%	0.8891 ± 0.57%
Additive-only	1d	0.8823± 0.83%	0.9003± 0.85%
Subtract-only	1d	0.8849± 0.84%	0.8845± 0.87%
Hadamard-only	1d	0.8822± 0.76%	0.8823± 0.67%
w/o Concatenation	3d	0.8901± 0.84%	0.8812± 0.80%
w/o Additive	4d	0.8873± 0.57%	0.8843± 0.83%
w/o Subtract	4d	0.8845± 0.81%	0.8839± 0.89%
w/o Hadamard	4d	0.8832 ± 0.82%	0.8815± 0.77%
Full model (Proposed)	5d	0.8981± 0.87%	0.8968± 0.59%

Table 2: Ablation Experiment Results of Feature Fusion.

ation has a positive contribution to improving the robustness and predictive ability of the model.

Overall, the ablation study confirms the effectiveness and necessity of the proposed multi-fusion mechanism.

Conclusion

This study solves the problem of tutor identification in educational conversation systems by introducing a dual encoding framework to effectively model conversation history and tutor response. By combining advanced feature interaction mechanisms and integrated voting strategies, the method demonstrates strong performance and robustness, achieving 89.65% Macro-F1 on the BEA 2025 shared task dataset. These results confirm the value of our approach in capturing personalized teaching styles and improving semantic consistency in feedback generation.

Limitations

Although our proposed dual-encoder framework performs well on the tutor identification task, it still has some limitations. First, the effectiveness of the model depends on the availability of labeled data, which may be limited in real-world educational settings. Second, the current approach assumes the existence of clear conversational turns and well-structured dialogues. Third, while the model captures personalized teaching styles to some extent, it does not explicitly incorporate speaker-specific historical profiles, which may further improve the recognition accuracy. Finally, the generalizability of the model across different educational domains and languages remains to be explored.

Ethical Considerations

This study uses de-identified educational conversation data provided by the BEA 2025 Shared Task organizers. No personally identifiable information is included. The task of tutor identification is aimed at supporting pedagogical analysis and improving educational tools, not at surveilling or ranking human educators. All model outputs are intended for research use only, and ethical guidelines for educational data processing have been followed.

References

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

William Cain. 2024. Prompting change: Exploring prompt engineering in large language model ai and its potential to transform education. *TechTrends*, 68(1):47–57.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *Preprint*, arXiv:1901.02860.

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.

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

Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large language models in ed-

ucation: Vision and opportunities. In *2023 IEEE international conference on big data (BigData)*, pages 4776–4785. IEEE.

Tan Guo, Baojiang Zhou, Fulin Luo, Lei Zhang, and Xinbo Gao. 2024. Dmfnet: Dual-encoder multi-stage feature fusion network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*.

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

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Preprint*, arXiv:1907.10529.

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. *Preprint*, arXiv:1907.11692.

Radhika Makharia, Yeoun Chan Kim, Su Bin Jo, Min Ah Kim, Aagam Jain, Piyush Agarwal, Anish Srivastava, Anant Vikram Agarwal, and Pankaj Agarwal. 2024. Ai tutor enhanced with prompt engineering and deep knowledge tracing. In *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, volume 2, pages 1–6. IEEE.

Andi Besse Firdausiah Mansur, Norazah Yusof, and Ahmad Hoirul Basori. 2019. Personalized learning model based on deep learning algorithm for student behaviour analytic. *Procedia Computer Science*, 163:125–133.

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.

SUN Qiang. 2025. Deep learning-based modeling methods in personalized education. *Artificial Intelligence Education Studies*, 1(1):23–47.

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.

Zhongchen Wang, Min Xia, Liguo Weng, Kai Hu, and Haifeng Lin. 2023. Dual encoder–decoder network for land cover segmentation of remote sensing image. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:2372–2385.

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.

Appendix

.1 Other Shared Tasks

We also participated in 4 tasks beyond the tutor identification task, and achieved the following rankings:

Mistake Identification task: 14/44

Mistake Identification task: 13/32

Providing Guidance task: 9/35

Actionability task: 11/35

The above data is from the official statistics of BEA workshop at ACL 2025.

.2 Method Details

The above four tasks all adopt a unified method framework. Specifically, we construct a dual tower encoder architecture based on the DeBERTaV3 pre trained model. Unlike the feature interaction modeling strategy introduced in the tutor identification task, this study did not adopt complex interaction mechanisms for these four tasks, but simply concatenated the feature vectors output by the twin towers. Subsequently, a routing selection module is introduced to screen and optimize the concatenated features, and finally the final category prediction is completed through a linear layer.