Wonderland_EDU@HKU at BEA 2025 Shared Task: Fine-tuning Large Language Models to Evaluate the Pedagogical Ability of AI-powered Tutors

Deliang Wang and Chao Yang and Gaowei Chen
Faculty of Education, The University of Hong Kong, Hong Kong, China
wdeliang@connect.hku.hk

Abstract

The potential of large language models (LLMs) as AI tutors to facilitate student learning has garnered significant interest, with numerous studies exploring their efficacy in educational contexts. Notably, Wang and Chen (2025) suggests that the relationship between AI model performance and educational outcomes may not always be positively correlated; less accurate AI models can sometimes achieve similar educational impacts to their more accurate counterparts if designed into learning activities appropriately. This underscores the need to evaluate the pedagogical capabilities of LLMs across various dimensions, empowering educators to select appropriate dimensions and LLMs for specific analyses and instructional activities. Addressing this imperative, the BEA 2025 workshop initiated a shared task aimed at comprehensively assessing the pedagogical potential of AI-powered tutors.

In this task, our team employed parameter-efficient fine-tuning (PEFT) on Llama-3.2-3B to automatically assess the quality of feedback generated by LLMs in student-teacher dialogues, concentrating on mistake identification, mistake location, guidance provision, and guidance actionability. The results revealed that the fine-tuned Llama-3.2-3B demonstrated notable performance, especially in mistake identification, mistake location, and guidance actionability, securing a top-ten ranking across all tracks. These outcomes highlight the robustness and significant promise of the PEFT method in enhancing educational dialogue analysis.

1 Introduction

In sociocultural theory, Vygotsky and Cole (1978) posits that learning occurs through interactions within social contexts, where conversation and dialogue serve as the primary mediums. During dialogues, a series of verbal or text exchanges occur between individuals, leading to the co-construction and negotiation of meaning (Tao and Chen, 2023), which has been found to facilitate individuals' cognitive development (Mercer and Littleton, 2007). Consequently, learning scientists and educational psychologists advocate for educators to harness the power of dialogue to enhance student learning.

In educational settings, dialogue can take place between students and teachers, students and their peers, and students and machines, in both online and offline environments (Wang et al., 2024b). These interactions contain rich information pertinent to students' learning. Initially, to provide valuable feedback on specific aspects of students' dialogues to improve learning outcomes, researchers manually analyzed these dialogues using rubrics or coding schemes (Howe and Abedin, 2013). However, due to the substantial human and time costs, this manual approach is not feasible for large-scale contexts involving numerous dialogues. With the advent of artificial intelligence (AI), researchers have explored using conventional machine learning techniques to automate the analysis of educational dialogues. This method, however, remains semi-automatic, as it requires researchers to determine which linguistic or speech features should be included as input (Wang et al., 2025a). Furthermore, the performance of this method still has room for improvement. Subsequently, deep neural networks emerged, demonstrating exceptional performance in natural language processing tasks. Researchers have thus explored using deep learning techniques to automatically analyze educational dialogues (Wang et al., 2024a,b; Shan et al., 2023). Although this approach achieves remarkable performance, deep learning techniques struggle to generalize across various educational contexts (Wang et al., 2025c). When educators wish to analyze additional dimensions of dialogue information, another round of data collection, annotation, and model training is necessary (Wang et al., 2025b).

In the past two years, large language models (LLMs) have emerged, demonstrating impres-

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1040–1048

sive abilities to understand human language. Pretrained on vast corpora of texts, LLMs can perform various language-related tasks and respond in ways that align with human expectations. Researchers have accordingly utilized LLMs to analyze educational dialogues (e.g., Wang and Demszky, 2023; Wang et al., 2023; Moreau-Pernet et al., 2024). Additionally, owing to their interactive nature, LLMs have also been employed as AI tutors, directly engaging with students to facilitate their learning. For instance, GPT-3.5 and Llama3 have been used as AI tutors to detect and correct students' errors in their messages (Daheim et al., 2024) and provide scaffolding help (Phung et al., 2023). Despite their versatility, LLMs' performance varies across different dimensions. For example, they may excel in correcting grammatical errors but struggle with analyzing reasoning mistakes. Moreover, Wang and Chen (2025) suggests that more accurate AI models do not necessarily lead to more effective educational outcomes. Sometimes, incorrect answers from LLMs can prompt students to engage in deep reflection, thereby positively impacting their learning. This indicates that even less accurate LLMs may have practical applications. Therefore, it is essential to comprehensively evaluate LLMs' pedagogical capabilities when they function as tutors. Such evaluations can inform the design of AI tutors in educational practice for optimized outcomes.

To address these issues, BEA 2025 is organizing a shared task to assess whether LLMs can serve as effective AI tutors based on four metrics: mistake identification, mistake location, guidance provision, and actionability clarity (Kochmar et al., 2025). Given educational dialogues between students and tutors in the mathematical domain, which are grounded in student mistakes or confusion, seven LLMs are tasked with providing feedback on students' utterances across these four dimensions. Following meticulous annotation of LLMs' pedagogical capabilities in these metrics, BEA 2025 provides a development set with annotations and invites participants to complete the remaining annotations in the test set. To accomplish this task, we selected a parameter-efficient fine-tuning (PEFT) technique, namely LoRA (Low-Rank Adaptation), and fine-tuned an LLM, specifically Llama-3.2-3B. This decision was motivated by two considerations. First, PEFT techniques have been shown to enable LLMs to outperform BERT and RoBERTa in educational dialogue analysis tasks (e.g., Wang and Chen, 2025; Wang et al., 2025b). Second, unlike the well-trained BERT and RoBERTa models specialized for specific tasks and fully fine-tuned LLMs with performance degradation on other tasks, parameter-efficiently fine-tuned LLMs not only excel in these tasks but also retain the ability to perform other tasks akin to the original LLMs.

2 Method

2.1 Task description

The dataset for the BEA 2025 shared task comprises 500 educational dialogues between students and tutors within the mathematical domains (Maurya et al., 2025), specifically from MathDial (Macina et al., 2023) and Bridge (Wang et al., 2024c). Each dialogue includes multiple prior exchanges from both the tutor and the student, in which in the student's final utterance contains a mistake. In response to these dialogues, seven LLMs—namely, GPT-4, Gemini, Sonnet, Mistral, Llama3.1-8B, Llama3.1-405B, and Phi-3—generate feedback aimed at identifying and addressing the mistake (Kochmar et al., 2025). The responses generated by the LLMs are annotated according to their quality across the following pedagogically motivated dimensions:

- *Mistake identification:* Since all dialogues in the dataset contain a mistake made by the student, a good quality response from the tutor should include the relevant mistake identification.
- *Mistake location:* A good tutor response should not only notify the student of the committed error, but also point to its location in the answer and outline what the error is to help student remediate it in their next response.
- *Providing guidance:* A good tutor response should provide the student with relevant and helpful guidance, such as a hint, an explanation, a supporting question, and the like.
- *Actionability:* Once the guidance is provided to a student, it should be clear from a good tutor response what the student should do next; in other words, the tutor response should not be vague, unclear or a conversation stopper.

The BEA 2025 shared task provides a development set consisting of 300 educational dialogues with annotations on these four dimensions and a

Conversation History:
Student: okey
Tutor: Now we have the same dominators so we can subtract the numbers directly.
Tutor: What is 25 minus 18?
Student: 8.

Llama3.1-405B:
Response: That's close, but let's double-check. If we have 25 and subtract 18, wouldn't we actually get 7?
Mistake Identification: Yes,
Mistake Location: Yes,
Providing Guidance: Yes,
Actionability: Yes

.....

Conversation History:
Tutor: What is the 12 subtracted by 1?
Student: 12.

Tutor 1:
Response: Actually, if we subtract 1 from 12, we get 11, not 12. Let's try it again together.
Mistake Identification: ?
Mistake Location: ?
Providing Guidance: ?
Actionability: ?

.....

Figure 1: Example of dialogues in the development set (blue) and test set (green).

test set containing 200 educational dialogues without annotations on these four dimensions. For each dimension, the annotations have three labels, respectively *Yes*, *To some extent*, and *No*. Their description and number can be seen in Table 1. Participants are tasked with predicting annotations for the test set based on the development set, which involves a three-way classification task for each metric. An example of a well-annotated dialogue from the development set and an unannotated dialogue from the test set is illustrated in Figure 1. More detailed description of the BEA 2025 shared task can be seen in Kochmar et al. (2025).

In addition to the four tracks previously mentioned, the BEA 2025 shared task also included anonymized responses generated by LLMs, as well as responses produced by both expert and novice tutors (i.e., **Identifying Tutors**). We were invited to predict the source of each response in the test set. Consequently, this track constitutes a nine-way classification task.

2.2 PEFT method

We opted for LoRA (Low-Rank Adaptation), a well-regarded PEFT method, to assess the quality of LLMs' pedagogical responses in the test set. LoRA effectively maintains the pre-existing weights within LLMs, incorporating trainable low-rank decomposition matrices into the internal Transformer framework (Hu et al., 2021). This technique substantially minimizes the number of parameters that need training, which is essential for fine-tuning LLMs in downstream tasks.

In deep neural networks, weight matrices generally exhibit full rank, indicating they possess the maximum number of linearly independent rows or columns. Nonetheless, pre-trained models frequently exhibit low intrinsic dimensionality, signifying that a low-dimensional reparameterization can be as effective for fine-tuning as utilizing the full parameter space (Aghajanyan et al., 2021). Therefore, it may not be necessary to adjust all parameters during fine-tuning for a particular downstream task. Instead, a lower-dimensional reparameterization can be employed to fine-tune LLMs (Xu et al., 2023). LoRA accomplishes this by utilizing two trainable low-rank matrices for the purpose of weight updates (Hu et al., 2021).

Formally, in the context of full fine-tuning, the update of an LLM's weight matrix (denoted as $W_0 \in \mathbb{R}^{d\times k}$) can be described by the expression $W_0 + \triangle W$. LoRA represents $\triangle W$ with two lower-rank trainable weight matrices, $W_{up} \in \mathbb{R}^{d\times r}$ and $W_{down} \in \mathbb{R}^{r\times k}$, as shown in Equation 1, where the rank r is much smaller than $min(d, k)$. As a result,, the original weight matrix W_0 remains unchanged during fine-tuning, thereby conserving memory, while only W_{up} and W_{down} are subject to updates. Given that r is much smaller than the minimum value between d and k, the computational demands of LoRA are markedly lower compared to full fine-tuning.

Table 1: The description of labels in each task in the development set.

	Labels	Description	Number
Mistake Identification	Yes	The mistake is clearly identified/ recognized in the tutor's response.	1932
	No	The tutor does not recognize the mistake (e.g., they proceed to simply provide the answer to the asked question).	370
	To some extent	The tutor's response suggests that there may be a mistake, but it sounds as if the tutor is not certain.	174
Mistake Location	Yes	The tutor clearly points to the exact location of a genuine mistake in the student's solution.	1543
	No	The response does not provide any details related to the mistake.	220
	To some extent	The response demonstrates some awareness of the exact mistake, but is vague, unclear, or easy to misunderstand.	713
Providing Guidance	Yes	The tutor provides guidance that is correct and relevant to the student's mistake.	1407
	No	The tutor's response does not include any guidance, or the guidance provided is irrelevant to the question or factually incorrect.	503
	To some extent	Guidance is provided but it is fully or partially incorrect, incomplete, or somewhat misleading.	566
Actionability	Yes	The response provides clear suggestions on what the student should do next.	1310
	No	The response does not suggest any action on the part of the student (e.g., it simply reveals the final answer).	369
	To some extent	The response indicates that something needs to be done, but it is not clear what exactly that is.	797

$$W_0 + \triangle W = W_0 + W_{up} W_{down} \qquad (1)$$

2.3 Fing-tuning

We selected an open-source LLM, Llama-3.2-3B-Instruct[1], to conduct PEFT. Llama-3.2-3B-Instruct, released by Google in September 2024, is an instruction-tuned, text-only large model optimized for multilingual dialogue applications, supporting eight languages. It surpasses many existing open-source and proprietary chat models on standard industry benchmarks.

A critical component of PEFT is the preparation of a well-annotated dataset. Accordingly, we meticulously designed a prompt to evaluate the pedagogical abilities of LLMs based on their responses in the development set, incorporating both instructional and contextual elements. The instruction was framed as follows: *The following dialogue is a discussion between a student and a tutor in the mathematical domain, where the last utterance from the student contains a mistake. Then, an AI tutor provides a response attempting to remediate such mistakes.* ***THEN TASK DESCRIPTION***. The description for each task is as follows:

> **Mistake Identification:** *Please analyze whether the AI tutor's response identifies the student' mistake and classify it as Yes, No, or To some extent.*
>
> **Mistake Location:** *Please analyze whether tutors' responses accurately point to a genuine mistake and its location in the students' responses and classify it as Yes, No, or To some extent.*
>
> **Providing Guidance:** *Please analyze whether tutors' responses offer correct and relevant guidance, such as an explanation, elaboration, hint, examples, and so on and classify it as Yes, No, or To some extent.*
>
> **Actionability:** *Please analyze whether tutors' feedback is actionable, i.e., it makes it clear what the student should do next, and classify it as Yes, No, or To some extent.*

The definitions of **Yes**, **No**, and **To some extent** were provided as context. Subsequently, we provided the educational dialogue and the LLMs' response as input. The expected output was the corresponding label accompanied by an explanation. Figure 2 illustrates an example of the prompt for the task of mistake identification and the expected answer used to fine-tune Llama-3.2-3B-Instruct. Examples of prompts for the other three tasks are available in Figures 3, 4, and 5.

For the ***Identifying Tutors*** track, the instruction and task description are as follows:

[1]https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

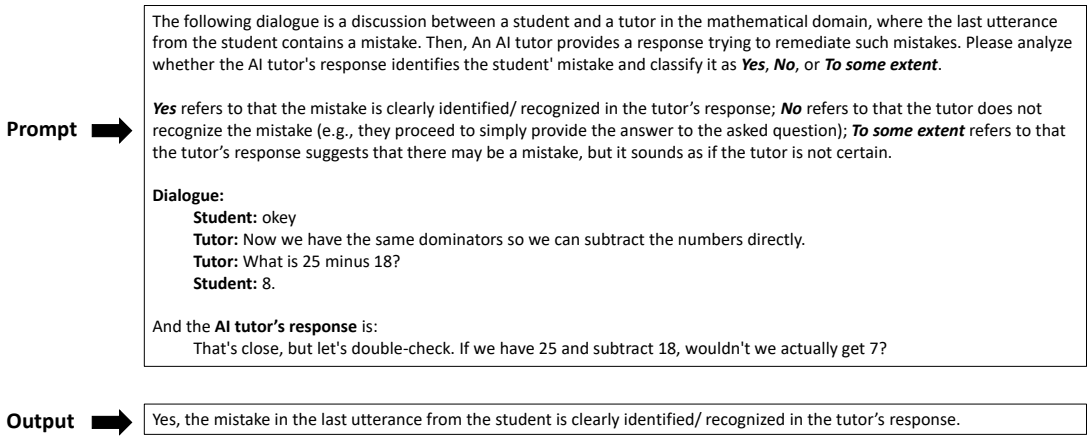

Figure 2: Example of a prompt for the task of mistake identification.

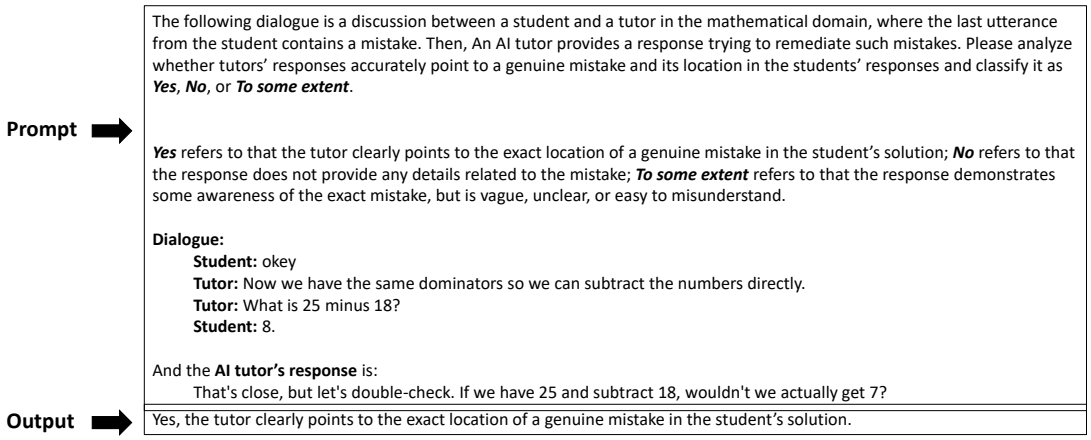

Figure 3: Example of a prompt for the task of mistake location.

"*The following dialogue is a discussion between a student and a tutor in the mathematical domain, where the student's last utterance contains a mistake. A tutor then provides a response aimed at remediating the error. Please analyze the response and classify the origin of the tutor as one of the following: Novice, Expert, Mistral, Phi, Sonnet, Llama318B, GPT-4, Gemini, or Llama31405B.*"

When employing LoRA for PEFT, we set the r dimension to 16 and the $LoRAalphas$ to 16. This configuration was chosen based on our available GPU resources and the adapter's representation capability. The training parameters were set as follows: the number of epochs was configured to 4, the batch size to 8, the learning rate to 2e-4, and the optimizer used was AdamW. The fine-tuning process was executed on an NVIDIA L20 GPU.

3 Results

Table 2 presents the performance outcomes of the parameter-efficiently fine-tuned Llama-3.2-3B across five distinct tracks.Among the first four racks, the organizers of BEA 2025 have utilized exact macro F1, exact accuracy, lenient macro F1, and lenient accuracy as evaluation metrics. The exact macro F1 and exact accuracy involve assessing predictions using three classes: "Yes," "To some extent," and "No." Conversely, lenient macro F1 and lenient accuracy consolidate "Yes" and "To some extent" into a single class, thus evaluating predictions within a two-class framework ("Yes + To some extent" vs. "No"). For the track of tutor identification, the evaluation metrics were exact macro F1 and exact accuracy scores. Specifically, in the task of mistake identification, our team achieved an exact macro F1 score of 0.6983 and an exact accuracy of 0.8675. In the task of actionability, we attained an exact macro F1 score of 0.6843 and an exact accuracy of 0.7285. Conversely, the tasks of mistake location and providing guidance proved to be more challenging. In mistake location, our team scored a macro F1 of 0.5450 and an exact accuracy of 0.7104, whereas in providing guidance, we obtained a macro F1 score of 0.5416 and an exact accuracy of 0.6464. In the task of tutor identification, our team achieved an exact macro F1 score of

Prompt

The following dialogue is a discussion between a student and a tutor in the mathematical domain, where the last utterance from the student contains a mistake. Then, An AI tutor provides a response trying to remediate such mistakes. Please analyze whether tutors' responses offer correct and relevant guidance, such as an explanation, elaboration, hint, examples, and so on and classify it as ***Yes***, ***No***, or ***To some extent***.

Yes refers to that the tutor provides guidance that is correct and relevant to the student's mistake; ***No*** refers to that the tutor's response does not include any guidance, or the guidance provided is irrelevant to the question or factually incorrect; ***To some extent*** refers to that guidance is provided but it is fully or partially incorrect, incomplete, or somewhat misleading.

Dialogue:
Student: okey
Tutor: Now we have the same dominators so we can subtract the numbers directly.
Tutor: What is 25 minus 18?
Student: 8.

And the **AI tutor's response** is:
That's close, but let's double-check. If we have 25 and subtract 18, wouldn't we actually get 7?

Output

Yes, the tutor provides guidance that is correct and relevant to the student's mistake.

Figure 4: Example of a prompt for the task of providing guidance.

Prompt

The following dialogue is a discussion between a student and a tutor in the mathematical domain, where the last utterance from the student contains a mistake. Then, An AI tutor provides a response trying to remediate such mistakes. Please analyze whether tutors' feedback is actionable, i.e., it makes it clear what the student should do next, and classify it as ***Yes***, ***No***, or ***To some extent***.

Yes refers to that the response provides clear suggestions on what the student should do next; ***No*** refers to that the response does not suggest any action on the part of the student (e.g., it simply reveals the final answer); ***To some extent*** refers to that the response indicates that something needs to be done, but it is not clear what exactly that is.

Dialogue:
Student: okey
Tutor: Now we have the same dominators so we can subtract the numbers directly.
Tutor: What is 25 minus 18?
Student: 8.

And the **AI tutor's response** is:
That's close, but let's double-check. If we have 25 and subtract 18, wouldn't we actually get 7?

Output

Yes, the response provides clear suggestions on what the student should do next.

Figure 5: Example of a prompt for the task of actionability.

Table 2: The performance of fine-tuned Llama-3.2-3B in each task in the test set.

	Exact macro F1	Exact accuracy	Lenient macro F1	Lenient accuracy	Ranking
Mistake Identification	0.6983	0.8675	0.9109	0.9496	7/44
Mistake Location	0.5450	0.7104	0.7649	0.8003	9/31
Providing Guidance	0.5416	0.6464	0.7456	0.7886	7/35
Actionability	0.6843	0.7285	0.8613	0.8888	6/29
Tutor Identification	0.8795	0.8778	N.A.	N.A.	7/20

0.8795 and an exact accuracy of 0.8778.

According to the exact macro F1 score, the BEA 2025 organizers ranked all participating teams. Our team achieved a top 10 ranking in each track, demonstrating the robustness of the PEFT technique employed in this report. The ranking further illustrates that the fine-tuned Llama-3.2-3B achieved superior performance in the tasks of mistake identification, providing guidance, and actionability, ranking 7th out of 44 teams, 7th out of 35 teams, and 6th out of 29 teams, respectively. In contrast, its performance in the task of mistake location was comparatively lower, ranking 9th out of 31 teams.

4 Discussion

Researchers have increasingly employed AI to automatically analyze educational dialogues (Wang et al., 2024b), aiming to provide timely feedback and enhance student learning. Existing studies predominantly utilize supervised machine learning techniques to develop models for educational dialogue analysis, which often suffer from limited generalizability. Typically, researchers and engineers must collect and annotate data and train AI models to analyze specific dimensions within dialogues. This process needs to be repeated for each new dimension under investigation or in a new educational context (Wang et al., 2023). The advent of LLMs offers a potential solution to these challenges, given their general and versatile capabilities in understanding human language and executing various natural language processing tasks. Consequently, researchers have begun exploring the use of LLMs to analyze diverse aspects of educational dialogues through prompt engineering techniques (e.g, Wang and Demszky, 2023; Wang et al., 2023). Additionally, the ability of LLMs to respond to user queries positions them as potential AI tutors to facilitate student learning. Thus, assessing whether LLMs can effectively serve as teachers is a critical question in educational practice, with numerous studies examining their impact in various contexts (Wang and Fan, 2025). Furthermore, Wang and Chen (2025) suggests that the relationship between AI model performance and educational outcomes may not always be positively correlated; less accurate AI models can sometimes achieve similar educational impacts to their more accurate counterparts if designed into learning activities appropriately. It is therefore essential to evaluate the pedagogical capabilities of LLMs across different dimensions, enabling educators to determine which versions of LLMs should be adopted for specific types of analysis and activities for teachers and students. In response to this need, the BEA 2025 conference organized a shared task to comprehensively assess the pedagogical potential of AI-powered tutors.

In this task, our team applied parameter-efficient fine-tuning to Llama-3.2-3B to automatically evaluate the quality of LLM-generated feedback on student-teacher dialogues, focusing on mistake identification, mistake location, guidance provision, and guidance actionability. The final leaderboards revealed that the fine-tuned Llama-3.2-3B achieved notable performance, particularly in the areas of mistake identification, mistake location, and guidance actionability. Our team ranked within the top ten across all tracks, underscoring the robustness and considerable potential of the PEFT method in educational dialogue analysis.

5 Limitation

While our application of a PEFT technique to fine-tune a widely recognized LLM yielded notable performance in this shared task, several limitations warrant acknowledgment. First, we exclusively evaluated the Llama-3.2-3B model. The generalizability of our findings to larger or alternative models, such as Mistral or Gemma, remains uncertain, and comparative analyses could reveal performance variations across LLMs. Second, the investigation focused solely on a single PEFT method. A broader exploration of alternative PEFT strategies—as well as full fine-tuning approaches—could strengthen the robustness of the proposed method and provide more comprehensive empirical validation. Third, the experiments relied on a uniform prompt design. As previous research, such as Tran et al. (2024), has demonstrated, the design of prompts significantly influences LLM performance. Incorporating diverse prompting techniques (e.g., chain-of-thought, role-based instructions) could mitigate bias and improve the reliability of experimental outcomes. To address these gaps, future work should prioritize (1) benchmarking across multiple LLM architectures, (2) systematically evaluating diverse fine-tuning paradigms, and (3) integrating advanced prompt engineering strategies, to rigorously assess the potential of LLMs as pedagogical tools.

Acknowledgments

This work was supported by Hong Kong Research Grants Council, University Grants Committee (Grant No.: 17605221).

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328.

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.

Christine Howe and Manzoorul Abedin. 2013. Classroom dialogue: A systematic review across four decades of research. *Cambridge journal of education*, 43(3):325–356.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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.

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.

Neil Mercer and Karen Littleton. 2007. *Dialogue and the development of children's thinking: A sociocultural approach*. Routledge.

Baptiste Moreau-Pernet, Yu Tian, Sandra Sawaya, Peter Foltz, Jie Cao, Brent Milne, and Thomas Christie. 2024. Classifying tutor discursive moves at scale in mathematics classrooms with large language models. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 361–365.

Tung Phung, Victor-Alexandru Pădurean, José Cambronero, Sumit Gulwani, Tobias Kohn, Rupak Majumdar, Adish Singla, and Gustavo Soares. 2023. Generative ai for programming education: benchmarking chatgpt, gpt-4, and human tutors. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 2*, pages 41–42.

Dapeng Shan, Deliang Wang, Chenwei Zhang, Ben Kao, and Carol KK Chan. 2023. Annotating educational dialog act with data augmentation in online one-on-one tutoring. In *International Conference on Artificial Intelligence in Education*, pages 472–477. Springer.

Yang Tao and Gaowei Chen. 2023. Coding schemes and analytic indicators for dialogic teaching: A systematic review of the literature. *Learning, Culture and Social Interaction*, 39:100702.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsay Clare Matsumura. 2024. Analyzing large language models for classroom discussion assessment. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 500–510, Atlanta, Georgia, USA. International Educational Data Mining Society.

Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

Deliang Wang, Cunling Bian, and Gaowei Chen. 2024a. Using explainable ai to unravel classroom dialogue analysis: Effects of explanations on teachers' trust, technology acceptance and cognitive load. *British Journal of Educational Technology*.

Deliang Wang and Gaowei Chen. 2025. Evaluating the use of bert and llama to analyse classroom dialogue for teachers' learning of dialogic pedagogy. *British Journal of Educational Technology*.

Deliang Wang, Dapeng Shan, Ran Ju, Ben Kao, Chenwei Zhang, and Gaowei Chen. 2025a. Investigating dialogic interaction in k12 online one-on-one mathematics tutoring using ai and sequence mining techniques. *Education and Information Technologies*, page 9215–9240.

Deliang Wang, Dapeng Shan, Yaqian Zheng, Kai Guo, Gaowei Chen, and Yu Lu. 2023. Can chatgpt detect student talk moves in classroom discourse? a preliminary comparison with bert. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 515–519.

Deliang Wang, Yang Tao, and Gaowei Chen. 2024b. Artificial intelligence in classroom discourse: A systematic review of the past decade. *International Journal of Educational Research*, 123:102275.

Deliang Wang, Chao Yang, and Gaowei Chen. 2025b. Using lora to fine-tune large language models for analyzing collaborative argumentation in classrooms. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale (L@S '25)*, Palermo, Italy. ACM.

Deliang Wang, Yaqian Zheng, Jinjiang Li, and Gaowei Chen. 2025c. Parameter-efficiently fine-tuning large language models for classroom dialogue analysis. *IEEE Transactions on Learning Technologies*.

Jin Wang and Wenxiang Fan. 2025. The effect of chatgpt on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. *Humanities and Social Sciences Communications*, 12(1):1–21.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024c. 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.

Rose E Wang and Dorottya Demszky. 2023. Is chatgpt a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. *arXiv preprint arXiv:2306.03090*.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*.