

Findings of the BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-powered Tutors

Ekaterina Kochmar¹, Kaushal Kumar Maurya¹, Kseniia Petukhova¹,
KV Aditya Srivatsa¹, Anaïs Tack², Justin Vasselli³

¹Mohamed bin Zayed University of Artificial Intelligence (MBZUAI),

²KU Leuven, ³Nara Institute of Science and Technology

Abstract

This shared task has aimed to assess pedagogical abilities of AI tutors powered by large language models (LLMs), focusing on evaluating the quality of tutor responses aimed at student’s mistake remediation within educational dialogues. The task consisted of five tracks designed to automatically evaluate the AI tutor’s performance across key dimensions of *mistake identification*, *precise location of the mistake*, *providing guidance*, and *feedback actionability*, grounded in learning science principles that define good and effective tutor responses, as well as the track focusing on *detection of the tutor identity*. The task attracted over 50 international teams across all tracks. The submitted models were evaluated against gold-standard human annotations, and the results, while promising, show that there is still significant room for improvement in this domain: the best results for the four pedagogical ability assessment tracks range between macro F1 scores of 58.34 (for *providing guidance*) and 71.81 (for *mistake identification*) on three-class problems, with the best F1 score in the *tutor identification* track reaching 96.98 on a 9-class task. In this paper, we overview the main findings of the shared task, discuss the approaches taken by the teams, and analyze their performance. All resources associated with this task are made publicly available to support future research in this critical domain.¹

1 Introduction and Motivation

Conversational agents offer promising opportunities for education as they can fulfill various roles (e.g., intelligent tutors and service-oriented assistants) and pursue different objectives (e.g., improving student skills and increasing instructional efficiency) (Wollny et al., 2021), among which serving as an AI tutor is one of the most prevalent

tasks (Tack et al., 2023). Recent advances in the development of large language models (LLMs) provide our field with promising ways of building AI-based conversational tutors, which can generate human-sounding dialogues on the fly. The key question posed in previous research (Tack and Piech, 2022; Tack et al., 2023), however, still holds: *How can we test whether state-of-the-art generative models are good AI teachers, capable of replying to a student in an educational dialogue?*

Evaluating dialogue systems in general presents a significant challenge. While human evaluation is still considered the most reliable method for assessing dialogue quality, its high cost and lack of reproducibility have led to the adaptation of both reference-based and reference-free automatic metrics, originally used in machine translation and summary evaluation, for dialogue evaluation (Lin, 2004; Popović, 2017; Post, 2018; Gao et al., 2020; Liu et al., 2023). When it comes to Intelligent Tutoring Systems (ITSs), which also function as dialogue systems with the specific role of acting as tutors, these general metrics are insufficient. In the educational context, we need to assess complex pedagogical aspects and abilities of such systems, ensuring that they provide students with sufficient, helpful, and factually correct guidance and do not simply reveal answers when the student makes a mistake, among other aspects. Therefore, developing automatic metrics to evaluate these nuanced aspects is essential for creating effective and helpful tutoring systems.

Due to the lack of a standardized evaluation taxonomy, previous work has used different criteria for evaluation. For example, Tack and Piech (2022) and Tack et al. (2023) evaluated models in terms of whether they *speak like a teacher*, *understand a student*, and *help a student*, while in Macina et al. (2023), responses of models playing roles of tutors were evaluated by human annotators using *coherence*, *correctness*, and *equitable tutoring*. At the

¹https://github.com/kaushal0494/UnifyingAITutorEvaluation/tree/main/BEA_Shared_Task_2025_Datasets

same time, Wang et al. (2024) assessed *usefulness*, *care*, and *human-likeness*, and Daheim et al. (2024) used *targetedness*, *correctness*, and *actionability* of a tutor response as quality evaluation criteria. Such lack of standardization makes it difficult to compare different systems, and, therefore, defining evaluation criteria and developing automatic metrics for them is a crucial task for advancing the field, which we have aimed to address in this task.

2 Task Description and Goals

Following the successful BEA 2023 Shared Task on *Generating AI Teacher Responses in Educational Dialogues* (Tack et al., 2023), we revisit the question of quality assessment of the tutor responses generated with the AI models (specifically, LLMs) in the context of educational dialogues. We believe that (1) the topic is timely and important; (2) LLMs have significantly advanced in the past couple of years, making it important to revisit this topic after the competition run in 2023; and (3) there is a need to establish a pedagogically motivated benchmark for this task. In contrast to the BEA 2023 Shared Task, our focus is not on the *generation of educational dialogues* using state-of-the-art LLMs, but rather on **comprehensive evaluation of AI-tutor responses using a set of pedagogically motivated metrics**.

In this shared task, we have focused on educational dialogues between a student and a tutor in the mathematical domain. Specifically, the conversations are grounded in student mistakes or confusion, where the AI tutor aims to remediate such mistakes or confusion. Each dialogue in the datasets provided in this shared task includes: (i) the context consisting of several prior turns from both the tutor and the student; (ii) the last utterance(s) from the student containing a mistake; and (iii) a set of possible responses to the last student’s utterance(s) from a range of LLM-based tutors and, where available, human tutors, aimed at mistake remediation. The dialogues (parts i-ii) are extracted from two popular datasets of educational dialogues in the mathematical domain – MathDial (Macina et al., 2023) and Bridge (Wang et al., 2024), while the LLM-based tutor responses are generated by the shared task organizers using a set of state-of-the-art LLMs of various sizes and capabilities, including: GPT-4 (Achiam et al., 2023), Gemini (Reid et al., 2024), Sonnet (Anthropic, 2024), Mistral (Jiang et al., 2023), Llama-3.1-8B and Llama-3.1-405B

(Dubey et al., 2024), and Phi-3 (Abdin et al., 2024). To avoid any biases, the tutor responses in the data have been shuffled. In addition to the responses themselves, the dataset contains annotation of their quality along several pedagogically motivated dimensions defined in Maurya et al. (2025). Below, we are reiterating the definitions for these dimensions from Maurya et al. (2025) for completeness:

- *Mistake identification*: Since all dialogues in the dataset contain a mistake made by the student, we expect a good quality response from the tutor to include relevant and clear mistake identification. This aligns with *student understanding* defined in Tack and Piech (2022) and *correctness* in the schemata of Macina et al. (2023) and Daheim et al. (2024).
- *Mistake location*: In addition to notifying the student about the committed error, a good tutor response should also point to its location in the answer and explain what the error is to help the student remediate it. This corresponds to *targetedness* in Daheim et al. (2024).
- *Providing guidance*: Ideally, a good tutor should not reveal the answer immediately and instead should provide the student with relevant and helpful guidance, consisting, for example, of a hint, an explanation, or a supporting question. This aspect is related to *helping a student* in Tack and Piech (2022) and *usefulness* in Wang et al. (2024).
- *Actionability*: Finally, once the guidance is provided to a student, a good tutor response should make it clear to the student what they are supposed to do next. I.e., the tutor’s response should not be vague, unclear, or a conversation stopper. This aspect directly corresponds to *actionability* in Daheim et al. (2024).

Moreover, the proposed evaluation schema aligns with the core pedagogical principles derived from learning sciences. Specifically, the tutor should: (1) *encourage active learning* (Chi and Wylie, 2014; Oakley and Sejnowski, 2021) by not directly revealing the correct answer, (2) *adapt to learners’ goals and needs* (King and South, 2017) through accurate mistake identification and exact location pointing, (3) *manage cognitive load*

Example 1: Spotted a Mistake?	
Conversation topic: Simple Expressions	
Conversation History: Tutor: We have to solve the inner parentheses first. Student: ok Tutor: What is 5 times 6? Student: 50	
Tutor response: Ah, not quite. 5 x 10 is 50. 5 x 6 is something else. Could you give it another try?	
Question: Has the tutor identified the mistake in the above response?	
Answer	Reasoning
✓ (1) Yes	The tutor clearly identified the mistake by explaining how to arrive at 50.
✗ (2) To some extent	
✗ (3) No	

Figure 1: An example on mistake identification from Maurya et al. (2025)

(Mayer, 2002) and enhance *metacognitive skills* (Dehaene, 2020; Cohen et al., 2021) by providing appropriate guidance, and (4) foster *motivation and stimulate curiosity* (Keller, 1987; Patall et al., 2008) by offering clear and actionable steps to the student. Thus, the schema adopted from Maurya et al. (2025) covers all the relevant aspects of a good tutor response proposed in previous work (Tack and Piech, 2022; Macina et al., 2023; Wang et al., 2024; Daheim et al., 2024), while also being supported by the learning science principles. We do not explicitly include such aspects as *speak like a teacher* (Tack and Piech, 2022), as we believe that a tutor that identifies student’s mistakes, points to them accurately, and can explain them to a student in an actionable way *does* speak like a teacher. We also do not explicitly cover *human-likeness* (Wang et al., 2024) as, based on our preliminary analysis, state-of-the-art LLMs are capable of producing overwhelmingly human-like responses.

All aspects are annotated on a 3-point scale, where "No" denotes that the particular aspect of the tutor response is *bad* (e.g., the mistake is not identified at all), "Yes" denotes that it is *good* (e.g., the mistake is identified clearly and correctly), and "To some extent" denotes that the quality of the response, according to the particular aspect, is medium (e.g., there are clarity issues with the mistake identification). Figure 1 provides an example of the annotation for the *mistake identification* aspect.

3 Shared Task Structure

This shared task consisted of two major phases:

- **Development phase:** In the development phase, we released annotated tutor responses for 300 dialogues extracted from the MathDial and Bridge datasets (approximately 75% examples from MathDial and 25% examples from Bridge). For each di-

alogue, responses from 7 LLM-based tutors (see Section 2 for more details) as well as expert (for both datasets) and novice (provided in the Bridge dataset only) tutor responses were released together with the annotations for 4 pedagogical aspects following the scheme and guidelines proposed in Maurya et al. (2025). This sums up to a total of 2,476 tutor responses. During the development phase, participating teams could build their systems aiming to predict the quality values for any or all of the pedagogical aspects.

- **Test phase:** In the test phase of the competition, we released an additional set of 191 dialogues extracted from the MathDial and Bridge datasets, following the distribution in the development set, together with the tutor responses (1,547 in total), but the annotations were not provided for this data. The participating teams were asked to run their systems and submit their predictions, which were then evaluated using the shared task official metrics (see Section 6).²

In addition, the task included the *5th track* on the tutor identity identification, aimed at automated detection of which model or human tutor an anonymous response in the test data originated from. This sub-task was inspired by our observations that various AI tutors have very specific tutoring and linguistic styles (Maurya et al., 2025).

The task used *open-data and model strategy*: as there were no explicit training phase, the teams were allowed to use any external data in addition to the released annotated dialogues during the development phase, as well as build traditional machine learning as well as large language model-based solutions.

The test phase of the task was hosted on the CodaBench platform, with a separate track for (1) *Mistake Identification*,³ (2) *Mistake Location*,⁴ (3) *Providing Guidance*,⁵ (4) *Actionability*,⁶ and (5) *Tutor Identification*.⁷ Each team was allowed up to 5 submissions in each track.

²Development and test sets are available at https://github.com/kaushal0494/UnifyingAITutorEvaluation/tree/main/BEA_Shared_Task_2025_Datasets

³<https://www.codabench.org/competitions/7195/>

⁴<https://www.codabench.org/competitions/7200/>

⁵<https://www.codabench.org/competitions/7202/>

⁶<https://www.codabench.org/competitions/7203/>

⁷<https://www.codabench.org/competitions/7206/>

4 Data Description

As described in Section 2, we used the data from two publicly available datasets – MathDial (Macina et al., 2023) and Bridge (Wang et al., 2024). Both datasets allow for adaptation, modification and (re)sharing without any restrictions: MathDial⁸ is distributed under the Creative Commons Attribution-ShareAlike 4.0 International License, while the Bridge⁹ dataset is licensed under the MIT license.

Annotations of the pedagogical aspects were provided by the organizing team following the scheme and guidelines established by Maurya et al. (2025). Of the 300 development set dialogues, responses in 200 were doubly-annotated by four annotators, reaching an average Fleiss’ kappa of 0.65, which indicates substantial agreement and shows reliability of this task (see the description of the annotation experiment in Maurya et al. (2025)). An additional set of tutor responses for further development and test set dialogues were annotated by the six shared task organizers using the same scheme and approach. A subset of 83 tutor responses in 10 dialogues were annotated by all six co-organizers, showing substantial agreement on the use of the scheme with Fleiss’ kappa of 0.64. After this initial annotation round, co-organizers discussed and resolved disagreements before proceeding to annotate the rest of the data.

5 Teams

Over 50 teams participated in the shared task, with 11 teams submitting to all five tracks. The task attracted participation from all over the world, with teams from Asia (e.g., Bangladesh, China, India, Indonesia, Philippines, and South Korea), Australia, Europe (e.g., France, Germany, and Romania), the MENA region (e.g., Egypt, Lebanon, and the UAE), the North (e.g., USA and Canada) as well as South America (e.g., Chile and Uruguay) taking part in it. The submissions were distributed as reported in Table 1.¹⁰ The next section briefly summarizes the main trends in the approaches adopted by the teams, while more details can be found in the individual system reports submitted by 26 teams as well as in Section 6.

⁸<https://github.com/eth-nlped/mathdial>

⁹<https://github.com/rosewang2008/bridge>

¹⁰The official leaderboards can be found in Appendix A.

Track	# Submissions	# Teams
Track 1	153	44
Track 2	86	32
Track 3	105	36
Track 4	87	30
Track 5	54	20

Table 1: Number of submissions and participating teams in each track

5.1 Main Trends

Based on the overall analysis of the approaches taken by the participating teams, we have identified the following major trends:

- A few teams used *LLMs*, both commercial (GPT-4o (Hurst et al., 2024), Gemini (Reid et al., 2024), Claude (Anthropic, 2024)) and open-source (Mistral (Jiang et al., 2023), LLaMa (Dubey et al., 2024), Qwen (Bai et al., 2023)) extensively. Examples include teams BJTU (Fan et al., 2025), BLCU-ICALL (An et al., 2025), NeuralNexus (Naeem et al., 2025), Henry (Pit, 2025), and LexiLogic (Bhattacharyya et al., 2025), among others.
- *LoRA-based fine-tuning* (Hu et al., 2022) has also been popular among the participants, including teams TutorMind (Dekmak et al., 2025), Archaeology (Roşu et al., 2025), Wonderland_EDU@HKU (Wang et al., 2025), Averroes (Yasser et al., 2025), and MSA (Hikal et al., 2025).
- *Data augmentation and imbalance handling* were used, including methods like synthetic data generation by TutorMind (Dekmak et al., 2025) and Henry (Pit, 2025), random downsampling by BJTU (Fan et al., 2025), oversampling by Thapar Titans (Dadwal et al., 2025) and NLIP (Saha et al., 2025), and class-weighted loss by Jinan Smart Education (Chen, 2025) and SYSUpporter (Chen et al., 2025).
- *Ensemble methods* were also applied: this included majority voting by Jinan Smart Education (Chen, 2025), stacking by NLIP (Saha et al., 2025), and disagreement-aware inference by MSA (Hikal et al., 2025).
- Finally, *hybrid and multi-stage architectures* were used, including integration of simpler

models for initial prediction followed by escalation to more powerful LLM judges as in the approach by Emergent Wisdom (Jain and Rengarajan, 2025), or use of architectures that combine embeddings and classification models as in the dual-encoder setup used by Jinan Smart Education (Chen, 2025).

6 Evaluation, Results, and Summary of Approaches

Tracks 1-4 used **macro F1 as the main metric**, with accuracy being the secondary metric. These were used in two settings:

- *Exact evaluation*: predictions submitted by the teams were evaluated for the exact prediction of the three classes (“Yes”, “To some extent”, and “No”)
- *Lenient evaluation*: since for these dimensions tutor responses annotated as “Yes” and “To some extent” share a certain amount of qualitative value, we considered “Yes” and “To some extent” as a single class, and evaluated predictions under the 2-class setting (“Yes + To some extent” vs. “No”)

Track 5 on *Tutor Identification* used **macro F1 as its main metric**, and accuracy of the tutor identity prediction as its secondary metric, in an exact multi-class scenario without the *lenient* setting.

This section overviews and discusses the results achieved by the teams in each track. For the full leaderboards, see Appendix A.

6.1 Track 1: Mistake Identification

Table 2 presents the results of a majority-class baseline prediction model for the development (Dev maj.) and test (Test maj.) sets. Since the data is heavily imbalanced, with “Yes” being the dominant class, we find such a baseline informative, as it shows what level of performance is achievable by a very simple system that always predicts the majority class. We report exact (strict) macro F1 (Ex. F1) and accuracy (Ex. Acc), as well as lenient F1 (Len. F1) and accuracy (Len. Acc).

77 participants registered in this track, and 44 teams submitted 153 system predictions in total. Table 2 reports the best results achieved by the teams (Best test) on all four metrics: exact F1 of 0.7181, exact accuracy of 0.8798, lenient F1 of 0.9185, and lenient accuracy of 0.9541. The winning team in this track, according to the main

shared task metric (exact F1), as well as according to the secondary metric of lenient F1, is BJTU (Fan et al., 2025).¹¹ The winners according to exact accuracy are TutorMind (Dekmak et al., 2025) and MSA (Hikal et al., 2025), with TutorMind scoring first in terms of lenient accuracy as well.

Category	Ex. F1	Ex. Acc	Len. F1	Len. Acc
Dev maj.	0.2922	0.7803	0.4596	0.8506
Test maj.	0.2827	0.7363	0.4522	0.8255
Best test	0.7181 ⁽¹⁾	0.8798 ^(12,13)	0.9185 ⁽¹⁹⁾	0.9541 ⁽³⁵⁾

Table 2: Results for *Track 1: Mistake Identification*

In this track, the 1st-place BJTU team used zero-shot prompting combined with dialogue-shuffling, random downsampling, and task-oriented prompt refinement (Fan et al., 2025). The 2nd-place TutorMind team fine-tuned GPT-4o-mini and Mistral-7B with LoRA and augmented their training data synthetically, significantly improving model performance (Dekmak et al., 2025). Averroes, ranked 3rd, benchmarked multiple instruction-tuned models, demonstrating that compact, carefully tuned models could outperform larger ones (Yasser et al., 2025). The 4th-place MSA team used Mathstral-7B with LoRA and introduced disagreement-aware ensemble strategy (Hikal et al., 2025). Finally, the 5th-place BD team combined MPNet fine-tuning (Song et al., 2020) with cross-validation and ensemble voting (Rohan et al., 2025).

6.2 Track 2: Mistake Location

Table 3 presents the results of a majority-class baseline prediction model for the development and test sets, as well as the best results achieved by the participating teams on the test set.

In total, 56 participants registered in this track, and 32 teams submitted 86 system predictions. Table 3 reports the best results achieved by the teams (Best test) on all four metrics: exact F1 of 0.5983, exact accuracy of 0.7679, lenient F1 of 0.8404, and lenient accuracy of 0.8630. The winning team in this track according to the main shared task metric (as well as exact and lenient accuracy) is BLCU-ICALL (An et al., 2025). The winner according to lenient F1 is K-NLPers (Park et al., 2025).

In this track, the 1st-place BLCU-ICALL

¹¹The notation in brackets indicates the place according to the main (exact F1-based) ranking of the submission showing the best result for each individual metric.

Category	Ex. F1	Ex. Acc	Len. F1	Len. Acc
Dev maj.	0.2560	0.6232	0.4159	0.7120
Test maj.	0.2450	0.5811	0.3974	0.6593
Best test	0.5983 ⁽¹⁾	0.7679 ⁽¹⁾	0.8404 ⁽⁵⁾	0.8630 ⁽¹⁾

Table 3: Results for *Track 2: Mistake Location*

team used a combination of in-context learning (ICL) with advanced prompting using the Gemini-2.5-pro model, supervised fine-tuning on large models like Qwen2.5-32B, and reinforcement learning from human feedback (RLHF) (An et al., 2025). The 3rd-place K-NLPers used GPT-4.1 combined with a specialized Multi-Perspective Reflective Evaluation approach, modeling internal deliberation among distinct reasoning perspectives (Park et al., 2025). The 5th-place team SG used Gemma-3-27B-IT in a two-step approach, where the model was first prompted to produce bulleted steps on the correct solution to the problem discussed in the dialogue, and then the tutor response was rated according to the specific rubrics. Finally, BJTU (2nd) and MSA (4th) used the same approaches as those described for the Mistake Identification track.

6.3 Track 3: Providing Guidance

As before, Table 4 presents the results of a majority-class baseline prediction model for the development and test sets and the best results achieved by the participating teams on the test set.

62 participants registered in this track, and 36 teams among them submitted 105 system predictions in total. Table 4 reports the best results achieved by the teams (Best test) on all four metrics: exact F1 of 0.5833, exact accuracy of 0.7052, lenient F1 of 0.7860, and exact accuracy of 0.8222. The winning team in this track according to the main shared task metric is MSA (Hikal et al., 2025). The winners according to other metrics are: SG, which scored first in terms of exact accuracy and lenient F1, and BLCU-ICALL (An et al., 2025), who scored first on lenient accuracy.

Category	Ex. F1	Ex. Acc	Len. F1	Len. Acc
Dev maj.	0.2416	0.5683	0.4355	0.7714
Test maj.	0.2313	0.5314	0.3995	0.6652
Best test	0.5834 ⁽¹⁾	0.7052 ⁽²⁾	0.7860 ⁽²⁾	0.8222 ⁽⁶⁾

Table 4: Results for *Track 3: Providing Guidance*

We note that this dimension is the only one where the distribution of annotations for the major-

ity class (“Yes” and “To some extent” combined) is substantially different from that in the test set. We attribute this to the inherent difficulty in judging the quality and appropriateness of pedagogical guidance provided by tutors in various contexts.

In this track, the top-ranked MSA (1st place), SG (2nd place), and BJTU (4th place) teams applied previously described generalizable training and prompt-based augmentation approaches. BLCU-ICALL (3rd place) specifically leveraged advanced ICL strategies, using models like Gemini-2.5-pro to excel in more open-ended instructional tasks. Meanwhile, K-NLPers (5th place) implemented a structured, rubric-based evaluation approach that decomposes guidance criteria into sub-questions, subsequently training a downstream Random Forest classifier to enhance scoring consistency.

6.4 Track 4: Actionability

In Table 5, we present the results of a majority-class baseline prediction model for the development and test sets and the best results achieved by the participating teams on the test set.

In total, 51 participants registered in this track, and 30 teams among them submitted 87 system predictions. Table 5 reports the best results achieved by the teams (Best test) on all four metrics: exact F1 of 0.7085, exact accuracy of 0.7557, lenient F1 of 0.8659, and lenient accuracy of 0.8940. The winning team according to the main shared task metric as well as exact accuracy is bea-jh (Roh and Bang, 2025). The winners according to other metrics are: MSA (Hikal et al., 2025) with the best score for lenient F1, and BJTU (Fan et al., 2025) scoring the highest in terms of lenient accuracy.

Category	Ex. F1	Ex. Acc	Len. F1	Len. Acc
Dev maj.	0.2307	0.5291	0.4041	0.6781
Test maj.	0.2198	0.4919	0.4095	0.6936
Best test	0.7085 ⁽¹⁾	0.7557 ⁽²⁾	0.8659 ⁽⁴⁾	0.8940 ⁽³⁾

Table 5: Results for *Track 4: Actionability*

In the Actionability track, the top-ranked bea-jh team implemented Group Relative Policy Optimization (GRPO) using GLM-4-9B (GLM et al., 2024), enhancing their predictions with explicit rationales in structured tags (Roh and Bang, 2025). BJTU (2nd) and MSA (3rd) continued using their prompting and fine-tuning frameworks. LexiLogic (4th place) experimented with multiple transformer-based models, achieving their best performance

with the Phi model (Bhattacharyya et al., 2025). The Phaedrus team (5th place) created an ensemble of seven LLMs, fine-tuned with LoRA on preference data, and integrated task-specific components such as generalized mean pooling and multi-sample dropout.

6.5 Track 5: Tutor Identification

Finally, in Table 6, we present the results of a majority-class baseline prediction model for the development and test sets, as well as the best results achieved by the participating teams on the test set.

50 participants registered in this track, and 20 teams submitted 54 system predictions in total. Table 6 reports the best results achieved by the teams (Best test) on the two metrics: exact F1 of 0.9698, and exact accuracy of 0.9664. The winning team according to both metrics is Phaedrus (Tiwari and Rastogi, 2025).

Category	Ex. F1	Ex. Acc
Dev maj.	0.0240	0.1212
Test maj.	0.0244	0.1235
Best test	0.9698 ⁽¹⁾	0.9664 ⁽¹⁾

Table 6: Results for *Track 5: Tutor Identification*

In the Tutor Identification track, the Phaedrus team (1st place) used an ensemble of seven LLMs with cross-response context augmentation, constraint satisfaction post-processing, and a specialized greedy label assignment. SYSupporter (2nd place) augmented training data with synthetic noise and used class-weighted loss, applying the Hungarian algorithm for unique label assignment at inference (Chen et al., 2025). Two Outliers (3rd place) developed DiReC, a two-stage model separating content and style features via supervised contrastive learning, followed by predictions with a CatBoost classifier and Hungarian algorithm (Tjitrahardja and Hanif, 2025). JInan_Smart Education (4th place) used a dual-encoder setup based on DeBERTa-v3, fusing dialogue and tutor-response representations before ensemble voting (Chen, 2025). Lastly, BLCU-ICALL (5th place) integrated supervised fine-tuning with large-scale models (Qwen2.5-32B) to specifically enhance performance on tutor authorship identification.

6.6 Best Teams across Tracks

Teams BJTU (Fan et al., 2025), MSA (Hikal et al., 2025), and BLCU-ICALL (An et al., 2025) emerged as the top-performing teams among those that participated in at least four out of five tracks, each achieving an average ranking within the top five. Notably, BJTU achieved the highest performance with an average rank of 2 participating in four tracks (including *mistake identification*, *mistake location*, *providing guidance*, and *actionability*), while MSA achieved an average rank of 4 across all five tracks. These teams employed cutting-edge techniques – such as diverse prompting, supervised fine-tuning, and RLHF – alongside traditional methods like data augmentation and output ensembling using state-of-the-art LLMs. The success of these strategies offers methodological insights and practical ideas for future research aimed at evaluating tutor responses.

6.7 Most Generalizable Approaches across Tracks

Teams MSA (Hikal et al., 2025), Wonderland_EDU@HKU (Wang et al., 2025), and TBA (Gombert et al., 2025) are the top-performing ones with the most generalizable approaches, having participated in at least four tracks and achieving average rankings within the top 10. The MSA model is an instruction-tuned variant (using LoRA) of Mathstral-7B-v0.1 (Mistral AI Team, 2024). To improve prediction reliability, they introduced a disagreement-aware ensemble inference strategy that enhances the coverage of minority labels. Wonderland_EDU@HKU proposed a LoRA-based instruction-tuned model using LLaMA-3.2-3B, where appropriate label-specific descriptions were added to improve performance. Finally, TBA fine-tuned FLAN-T5-x1 models on each evaluation dimension separately, then merged them using the DARE-TIES algorithm (Yu et al., 2024) to exploit task interdependencies. This merged model was further fine-tuned per task to produce the final submissions. These models show great promise for the development of generalized approaches for these challenging tasks and similar future benchmarks. For more details, please refer to Appendix B.

7 Analysis and Discussion

In this section, we conduct a detailed analysis of the results and data across all five tracks, highlighting

notable trends, challenging cases, and differences across dialogue contexts and LLMs.

Most Difficult and Easiest Cases Four tutor responses across the tracks proved particularly challenging, with **none** of the teams correctly classifying these cases according to the gold-standard annotations. Specifically, three of these difficult cases were originally annotated as “To some extent” – one each in the dimensions of *mistake identification*, *mistake location*, and *actionability*. Interestingly, there was also one challenging case originally annotated as “Yes” in the *Actionability* track, which was universally misclassified. In the tutor identification track, two cases involving responses from Llama-3.1-8B and Llama-3.1-405B were especially difficult, as none of the teams successfully identified these tutors. See Tables 30 and 31 in Appendix C for illustrative examples of the most challenging cases in the *Mistake Identification* and *Tutor Identification* tracks, respectively.

In the *Mistake Identification* track, three cases were correctly classified by **all** participating teams. These were annotated as “Yes” in the gold standard and featured explicit mistake identification phrasing such as *It seems like there’s a small mistake in your solution*. Similarly, in the *Actionability* track, there were two universally correctly classified “No” cases. One involved the tutor response *That was a very good try!*, which lacked any guidance. The other case involved a response that did not identify the student’s mistake but instead simply praised the student’s solution.

The Most Difficult Dialogue The conversation shown in Table 7 was the most challenging for teams across all pedagogical dimensions, with the majority of team predictions being incorrect for all responses. This difficulty likely stems from the subtly ambiguous problem statement, which led to a plausible but incorrect student interpretation that many tutors failed to explicitly correct. Tutor responses varied considerably: some correctly identified the student’s error, others implicitly reinforced the misunderstanding, and most lacked clear guidance or actionable feedback.

Difficulty Evaluation across LLMs Our analysis revealed substantial variability in evaluation difficulty across different tutor models, as measured by the rate at which team predictions misaligned with the gold-standard annotations. Responses from models like Llama-3.1-8B (42.35%

misalignment) and Gemini (40.57%) proved particularly challenging for the teams to classify accurately. Even Expert responses exhibited a high misalignment rate (37.14%), highlighting the inherent complexity and nuance of expert pedagogical dialogue. In contrast, models such as GPT-4 and Phi-3 showed much lower misalignment rates (20.45% and 17.72%, respectively), suggesting more consistent and predictable styles.

Difficulty Evaluation across Subsets and Tracks

Table 8 shows aggregate performance across tracks and subsets. The scores reported are average exact label match scores across all submissions and test examples. *Mistake Identification* and *Tutor Identification* show the highest scores, suggesting that these are somewhat easier tasks. In contrast, *Providing Guidance* has the lowest scores, likely due to its open-ended nature requiring explanations, examples, or strategies. *Mistake Location* and *Actionability* fall in between, indicating moderate difficulty. While MathDial generally shows higher scores than Bridge, across most tracks except for *Tutor Identification*, the performance gap is small, highlighting model adaptability to diverse dialogue formats and underscoring future scalability potential.

Tutor Identification Analysis Table 9 shows tutor identification accuracy for individual tutors across all submissions. Phi-3 is the most easily identifiable, likely due to its generally incoherent responses. Closely following is Expert, the human tutor, where indicators like spelling errors may have contributed to easier identification. Llama-3.1-405B, Mistral, Gemini, and GPT-4 yield similar scores, indicating moderate confusion among these models. In contrast, Llama-3.1-8B and Novice receive the lowest scores, suggesting that their responses closely resemble those of other tutors, making tutor identification more difficult.

8 Conclusions

In this paper, we have overviewed the BEA 2025 shared task on pedagogical ability assessment of AI-powered tutors in the context of student mistake remediation in the mathematical domain. Specifically, this shared task has focused on four key pedagogical dimensions of *mistake identification*, *precise location of the mistake*, *providing guidance*, and *feedback actionability*, grounded in learning science principles. In addition, a separate track has

Dialogue Context					
Tutor: Hi, could you please provide a step-by-step solution for the question below? The question is: Pat is having a picnic with her family. She has 42 cookies. She also has 63 pieces of candy and 21 brownies. There are 7 people in her family. If each person gets the same number of each dessert, how much will each person get?					
Student: The total number of desserts is $42 + 63 + 21 = 126$. If each person gets the same number of each dessert, then each person will get a total of $126/7 = 18$ desserts. Since there are three types of dessert, each person will get $18/3 = 6$ of each dessert. Therefore, each person will get 6 cookies, 6 pieces of candy, and 6 brownies.					
Tutor: I think you were confused by this question. Please talk me through your solution.					
Tutor: You correctly added the total number of desserts as 126 and divided it amongst 7 people to give 18 desserts each. What would happen if you calculated each dessert separately?					
Student: If I calculated each dessert separately I would get $42/7 = 6$ cookies for each person, $63/7 = 9$ pieces of candy for each person, and $21/7 = 3$ brownies for each person. However, this would mean that each person would not get the same number of each dessert.					
Tutor	Tutor Response	Mistake Id.	Mistake Loc.	Guidance	Actionable
Llama-3.1-405B	Actually, that's not a problem – the question states that each person gets the same number of each dessert, but it doesn't say that the number of each dessert has to be the same, so your second solution, where each person gets 6 cookies, 9 pieces of candy, and 3 brownies, is actually the correct one!	No	No	No	No
Llama-3.1-8B	That's correct – when you divide each type of dessert separately, you get a different amount for each, which means they won't be equal among the 7 people.	No	No	No	No
Expert	Ok, the question is confusing you. They just have to have a share each. The answer should be 18.	To some extent	Yes	No	No
GPT-4	That's correct! The question was asking how many of each type of dessert each person would get, not that everyone had to get the same number of each dessert. So, each person will get 6 cookies, 9 pieces of candy, and 3 brownies. Good job!	No	No	No	No
Sonnet	Great job on calculating each dessert separately – that's actually the correct approach here, as the question asks for each person to get the same number of each type of dessert, not the same total number of desserts.	No	No	No	No
Gemini	You're right, that would mean everyone doesn't get the same amount of each type of dessert! Let's think about how we can make sure everyone gets an equal share of cookies, candy, and brownies.	To some extent	No	No	No
Mistral	It seems there was a misunderstanding in the question; it's asking for each person to get the same number of each type of dessert, not the same total number of desserts.	Yes	No	No	No
Phi-3	To ensure each student receives an equal number of cookies, divide the total number of cookies by the number of students.	To some extent	No	No	No

Table 7: The dialogue from the test set that proved to be the most challenging for participating teams

Tracks	Bridge	MathDial
Mistake Identification	80.36	81.17
Mistake Location	63.97	67.48
Providing Guidance	56.16	59.29
Actionability	64.84	65.70
Tutor Identification	78.54	76.43

Table 8: Aggregate submission performance across tracks and subsets. The reported scores are average exact match scores across submissions and test examples.

addressed *detection of the tutor identity* based on the inherent linguistic and stylistic properties of tutor responses. Over 50 international teams have

Tutor	Accuracy (in %)	Tutor	Accuracy (in %)
Llama-3.1-8B	61.4	GPT-4	70.9
Novice	66.5	Sonnet	74.5
Llama-3.1-405B	68.8	Expert	79.1
Mistral	69.1	Phi-3	79.5
Gemini	69.4	-	-

Table 9: Tutor identification accuracy for each tutor across all submissions

participated in this shared task across all tracks, and in this paper, we have discussed the approaches adopted and the results achieved, highlighting the general trends in this challenging domain as well as the most promising avenues for research.

Limitations

We hope that the findings of this shared task will help the community advance research in pedagogically oriented AI-powered tutoring systems. However, we recognize that this task has been subject to several limitations, including:

Specific pedagogical dimensions and educational scenarios: In this task, we have specifically focused on the mistake remediation scenario in educational dialogues. As a result, only particular pedagogical properties of the responses (such as the ability of a tutor to indicate that there is a mistake in the student’s solution and point to it, providing pedagogically useful, actionable guidance) were considered. We acknowledge that, in broader educational scenarios, additional properties of tutor responses may be considered important, and we hope that future work will take this into account.

Limited contextual window: Another important limitation of the scheme used in this shared task is that, at the moment, we are considering pedagogical values of tutor responses in terms of addressing a specific student’s mistake or confusion exemplified in a limited number of previous student turns. Future work should consider extending tutor response evaluation to the extent of the full dialogue.

Domain limitations: This shared task has focused on the mathematical domain only. We acknowledge that applications to other subject domains may present researchers with different challenges.

Language limitations: Similarly, we acknowledge that this shared task has focused on dialogues in English only.

Limited number of LLMs-as-tutors: Finally, despite the fact that this shared task has considered a set of diverse LLMs-as-tutors, this set is necessarily limited.

Ethics Statement

Although we do not foresee any ethical risks or implications related to this shared task, we acknowledge that this task relies on the outputs from LLMs, and there are certain risks associated with such outputs in general: these models may generate outputs that, although plausible, may be factually incorrect, nonsensical, or even offensive. For instance, hallucinations can misguide students and propagate biases, which is especially dangerous in educational

settings. Nevertheless, we strongly believe that this shared task will help shed light on the current LLM capabilities in the context of educational dialogues, and the insights gained from this task may help mitigate issues related to the use of LLMs in the educational domain in the future.

Acknowledgments

We are grateful to Google for supporting this research through the Google Academic Research Award (GARA) 2024.

References

- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, 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 Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Jiyuan An, Xiang Fu, Bo Liu, Xuquan Zong, Cunliang Kong, Shuliang Liu, Shuo Wang, Zhenghao Liu, Liner Yang, Hanghang Fan, and Erhong Yang. 2025. BLCU-ICALL at BEA 2025 Shared Task: Multi-Strategy Evaluation of AI Tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.
- Anthropic. 2024. *The Claude 3 Model Family: Opus, Sonnet, Haiku*.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Souvik Bhattacharyya, Billodal Roy, Niranjan M Kumar, and Pranav Gupta. 2025. LexiLogic at BEA 2025 Shared Task: Fine-tuning Transformer Language Models for the Pedagogical Skill Evaluation of LLM-based tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.
- Sofía Correa Busquets, Valentina Córdova Véliz, and Jorge Baier. 2025. IALab UC at BEA 2025 Shared Task: LLM-Powered Expert Pedagogical Feature Extraction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.
- Lei Chen. 2025. Jinan Smart Education at BEA 2025 Shared Task: Dual Encoder Architecture for Tutor Identification via Semantic Understanding of Pedagogical Conversations. In *Proceedings of the 20th*

- Workshop on Innovative Use of NLP for Building Educational Applications.*
- Longfeng Chen, Zeyu Huang, Zheng Xiao, Yawen Zeng, and Jin Xu. 2025. SYSUpporter Team at BEA 2025 Shared Task: Class Compensation and Assignment Optimization for LLM-generated Tutor Identification. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications.*
- Micheline TH Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4):219–243.
- Richard K Cohen, Deanne Kildare Opatosky, James Savage, Susan Olsen Stevens, and Edward P Darrah. 2021. *The Metacognitive Student: How to Teach Academic, Social, and Emotional Intelligence in Every Content Area.* ERIC.
- Harsh Dadwal, Sparsh Rastogi, and Jatin Bedi. 2025. Thapar Titan/s : Fine-Tuning Pretrained Language Models with Contextual Augmentation for Mistake Identification in Tutor–Student Dialogues. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications.*
- Nico Daheim, Jakob Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. Stepwise Verification and Remediation of Student Reasoning Errors with Large Language Model Tutors. *arXiv preprint arXiv:2407.09136.*
- Stanislas Dehaene. 2020. *How we learn: The new science of education and the brain.* Penguin UK.
- Fatima Dekmak, Christian Khairallah, and Wissam Antoun. 2025. TutorMind at BEA 2025 Shared Task: Leveraging Fine-Tuned LLMs and Data Augmentation for Mistake Identification. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications.*
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The LLaMA 3 herd of models. *arXiv preprint arXiv:2407.21783.*
- Yuming Fan, Chuangchuang Tan, and Wenyu Song. 2025. BJTU at BEA 2025 Shared Task: Task-Aware Prompt Tuning and Data Augmentation for Evaluating AI Math Tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications.*
- Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. *arXiv preprint arXiv:2009.06978.*
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chat-GLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv preprint arXiv:2406.12793.*
- Sebastian Gombert, Fabian Zehner, and Hendrik Drachler. 2025. TBA at BEA 2025 Shared Task: Transfer-Learning from DARE-TIES Merged Models for the Pedagogical Ability Assessment of LLM-Powered Math Tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications.*
- Santiago Góngora, Ignacio Sastre, Santiago Robaina, Ignacio Remersaro, Luis Chiruzzo, and Aiala Rosá. 2025. RETUYT-INCO at BEA 2025 Shared Task: How Far Can Lightweight Models Go in AI-powered Tutor Evaluation? In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications.*
- Baraa Hikal, Mohmaed Basem, Islam Abdulhakeem Oshallah, and Ali Hamdi. 2025. MSA at BEA 2025 Shared Task: Disagreement-Aware Instruction Tuning for Multi-Dimensional Evaluation of LLMs as Math Tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications.*
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. LoRA: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276.*
- Raunak Jain and Srinivasan Rengarajan. 2025. Emergent Wisdom at BEA 2025 Shared Task: From Lexical Understanding to Reflective Reasoning for Pedagogical Ability Assessment. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications.*
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825.*
- John M Keller. 1987. Development and use of the ARCS model of instructional design. *Journal of instructional development*, 10(3):2–10.
- John King and Joseph South. 2017. Reimagining the role of technology in higher education: A supplement to the national education technology plan. *US Department of Education, Office of Educational Technology*, pages 1–70.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruo Chen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- Zhihao Lyu. 2025. CU at BEA 2025 Shared Task: A BERT-Based Cross-Attention Approach for Evaluating Pedagogical Responses in Dialogue. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.
- 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*.
- Maria Monica Manlises, Mark Edward Miranda Gonzales, and Lanz Yong Lim. 2025. DLSU at BEA 2025 Shared Task. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.
- 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.
- Richard E Mayer. 2002. Multimedia learning. In *Psychology of learning and motivation*, volume 41, pages 85–139. Elsevier.
- Mistral AI Team. 2024. Mathstral 7B v0.1: A Math Reasoning and Scientific Discovery Model. <https://mistral.ai/news/mathstral>. Accessed: 2025-06-09.
- Numaan Naeem, Sarfraz Ahmad, Momina Ahsan, and Hasan Iqbal. 2025. NeuralNexus at BEA 2025 Shared Task: Retrieval-Augmented Prompting for Mistake Identification in AI Tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.
- Barbara Oakley and Terrence J Sejnowski. 2021. *Uncommon sense teaching: Practical insights in brain science to help students learn*. Penguin.
- Geon Park, Jiwoo Song, Gihyeon Choi, Juoh Sun, and Harksoo Kim. 2025. K-NLPers at BEA 2025 Shared Task: Evaluating the Quality of AI Tutor Responses with GPT-4.1. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.
- Erika A Patall, Harris Cooper, and Jorgianne Civey Robinson. 2008. The effects of choice on intrinsic motivation and related outcomes: a meta-analysis of research findings. *Psychological bulletin*, 134(2):270.
- Henry Pit. 2025. Henry at BEA 2025 Shared Task: Improving AI Tutor’s Guidance Evaluation Through Context-Aware Distillation. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.
- Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. *arXiv preprint arXiv:1804.08771*.
- Md. Abdur Rahman, Md Al Amin, Sabik Aftahee, Muhammad Junayed, and Md Ashiqu Rahman. 2025. SmolLab_SEU at BEA 2025 Shared Task: A Transformer-Based Framework for Multi-Track Pedagogical Evaluation of AI-Powered Tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Jihyeon Roh and Jinhyun Bang. 2025. bea-jh at BEA 2025 Shared Task: Evaluating AI-powered Tutors through Pedagogically-Informed Reasoning. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.
- Shadman Rohan, Ishita Sur Apan, Muhtasim Ibteda Shochcho, Md Fahim, Mohammad Ashfaq Ur Rahman, AKM Mahbubur Rahman, and Amin Ahsan Ali. 2025. BD at BEA 2025 Shared Task: MPNet Ensembles for Pedagogical Mistake Identification and Localization in AI Tutor Responses. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.
- Ana Maria Roşu, Iani Gabriel Ispas, and Sergiu Nisioi. 2025. Archaeology at BEA 2025 Shared Task: Are Simple Baselines Good Enough? In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.
- Trishita Saha, Shrenik Ganguli, and Maunendra Sankar Desarkar. 2025. NLIP at BEA 2025 Shared Task: Evaluation of Pedagogical Ability of AI Tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.
- Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. The BEA 2023 shared task on generating AI teacher responses in educational dialogues. *arXiv preprint arXiv:2306.06941*.

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

Rajneesh Tiwari and Pranshu Rastogi. 2025. Phaedrus at BEA 2025 Shared Task: Assessment of Mathematical Tutoring Dialogues through Tutor Identity Classification and Actionability Evaluation. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.

Eduardus Tjitrahardja and Ikhlusal Akmal Hanif. 2025. Two Outliers at BEA Shared Task 2025 Task 5: Tutor Identity Classification using DiReC, a Two-Stage Disentangled Contrastive Representation. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.

Deliang Wang, Chao Yang, and Gaowei Chen. 2025. Wonderland_EDU@HKU at BEA 2025 Shared Task: Fine-tuning Large Language Models to Evaluate the Pedagogical Ability of AI-powered Tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.

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.

Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachslar. 2021. Are we there yet? - A systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, 4:654924.

Mazen Yasser, Mariam Saeed, Hossam Elkordi, and Ayman Khalafallah. 2025. Averroes at BEA 2025 Shared Task: Verifying Mistake Identification in Tutor, Student Dialogue. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. JMLR.org.

A Leaderboards

For more details, please also check the full official leaderboards at <https://sig-edu.org/sharedtask/2025#results>.

A.1 Track 1: Mistake Identification

The main leaderboard is presented by Table 10. The top 5 results for each secondary metric in this track are shown in Tables 11 to 13.

A.2 Track 2: Mistake Location

The main leaderboard is presented by Table 14. The top 5 results for each secondary metric in this track can be found in Tables 15 to 17.

A.3 Track 3: Providing Guidance

Table 18 presents the main leaderboard, while the top 5 results for each secondary metric in this track can be found in Tables 19 to 21.

A.4 Track 4: Actionability

Table 22 presents the main leaderboard, while Tables 23 to 25 report on the top 5 results for each secondary metric in this track.

A.5 Track 5: Tutor Identification

Tables 26 and 27 present the main leaderboard and the top 5 results for the secondary metric, respectively.

B Analysis of the Approaches

Table 28 provides a comprehensive overview of the modeling approaches and LLMs employed by participating teams. Table 29 and Figure 2 further summarize the methodologies and models adopted by the top-performing teams in each track. Notably, Table 29 highlights the highest-scoring teams across tracks. It also details instances where a single modeling approach demonstrated robust performance across multiple tracks, underscoring the potential generalizability of certain approaches.

C Examples of Particularly Challenging Cases

Tables 30 and 31 provide illustrative examples of the most challenging cases in the *Mistake Identification* and *Tutor Identification* tracks, respectively.

Rank	Team	Ex. F1	Ex. Acc	Len. F1	Len. Acc
1	BJTU	0.7181	0.8623	0.8957	0.9457
2	TutorMind	0.7163	0.8759	0.9108	0.9528
3	Averroes	0.7155	0.8675	0.8997	0.9425
4	MSA	0.7154	0.8759	0.9152	0.9535
5	BD	0.7110	0.8772	0.8966	0.9412
6	Gooby-Snoob Guysz	0.7105	0.8481	0.8901	0.9373
7	Wonderland_EDU@HKU	0.6983	0.8675	0.9109	0.9496
8	Archaeology	0.6976	0.8675	0.8959	0.9405
9	test	0.6948	0.8400	0.8947	0.9451
10	Someone	0.6926	0.8520	0.8964	0.9438
11	TBA	0.6858	0.8740	0.9060	0.9476
12	BLCU-ICALL	0.6822	0.8578	0.8909	0.9418
13	bea-jh	0.6802	0.8708	0.9069	0.9457
14	JiNan_Smart Education	0.6790	0.8688	0.9052	0.9470
15	jeez	0.6735	0.8623	0.8957	0.9418
16	MT-NLP	0.6677	0.8636	0.8885	0.9354
17	K-NLPers	0.6669	0.8113	0.8671	0.9270
18	Thapar Titan/s	0.6647	0.8520	0.8840	0.9328
19	Squirrel Ai Learning	0.6646	0.8539	0.8748	0.9315
20	Smollab_SEU	0.6617	0.8397	0.8782	0.9315
21	bnl	0.6578	0.8494	0.8806	0.9302
22	LexiLogic	0.6549	0.8487	0.8806	0.9302
23	Retuyt-InCo	0.6535	0.8449	0.8395	0.9192
24	777	0.6534	0.8526	0.8731	0.9283
25	CU	0.6514	0.8701	0.8957	0.9425
26	NLP Group 7	0.6499	0.8462	0.8605	0.9276
27	NLIP	0.6438	0.8546	0.8723	0.9257
28	ALA	0.6361	0.8423	0.8493	0.9140
29	mucai	0.6285	0.8067	0.8354	0.8985
30	AGS	0.6251	0.8390	0.8640	0.9211
31	Tutorify	0.6247	0.8261	0.8502	0.9173
32	Promptly Educated	0.6196	0.7104	0.8479	0.9224
33	wyn	0.6184	0.8384	0.8434	0.9095
34	Emergent Wisdom	0.6100	0.8546	0.8799	0.9321
35	Georgia Tech EDU	0.6049	0.8171	0.8386	0.9102
36	SG	0.5896	0.7919	0.8258	0.8875
37	NeuralNexus	0.5840	0.8268	0.8142	0.8972
38	presidency	0.5807	0.7570	0.8070	0.8804
39	NLP_UNH	0.5708	0.8358	0.8358	0.9089
40	letstea	0.5376	0.6593	0.8109	0.8681
41	Patriots	0.5345	0.8028	0.7923	0.8921
42	AUST_NLP	0.4819	0.7085	0.6929	0.7576
43	WhyIamHere	0.4562	0.7931	0.7126	0.8824
44	RAGthoven	0.2949	0.4350	0.4349	0.5365

Table 10: Official leaderboard for *Track 1: Mistake Identification*

Rank	Team	Ex. Acc
1	TutorMind MSA	0.8798
2	BD	0.8772
3	BJTU	0.8765
4	Archaeology	0.8746
5	TBA	0.8740

Table 11: Top 5 results according to exact accuracy for *Track 1: Mistake Identification*

Rank	Team	Len. F1
1	BJTU	0.9185
2	MSA	0.9152
3	TutorMind	0.9143
4	BLCU-ICALL	0.9110
5	Wonderland_EDU@HKU	0.9109

Table 12: Top 5 results according to lenient F1 for *Track 1: Mistake Identification*

Rank	Team	Len. Acc
1	TutorMind	0.9541
2	MSA	0.9535
3	BJTU BLCU-ICALL	0.9515
4	Wonderland_EDU@HKU	0.9496
5	TBA	0.9476

Table 13: Top 5 results according to lenient accuracy for *Track 1: Mistake Identification*

Rank	Team	Ex. F1	Ex. Acc	Len. F1	Len. Acc
1	BLCU-ICALL	0.5983	0.7679	0.8386	0.8630
2	BJTU	0.5940	0.7330	0.7848	0.8261
3	K-NLPers	0.5880	0.7641	0.8404	0.8610
4	MSA	0.5743	0.6975	0.7848	0.8209
5	SG	0.5692	0.7602	0.8118	0.8416
6	bea-jh	0.5658	0.6723	0.7792	0.8197
7	bd	0.5543	0.7143	0.7699	0.8054
8	TBA	0.5490	0.7091	0.7702	0.8035
9	Wonderland_EDU@HKU	0.5450	0.7104	0.7649	0.8003
10	Averroes	0.5366	0.6348	0.7587	0.7822
11	Whyamher	0.5325	0.6910	0.7370	0.7802
12	NLIP	0.5319	0.6878	0.7495	0.7951
	Archaeology	0.5319	0.6568	0.7558	0.8009
13	JiNan_Smart Education	0.5274	0.6968	0.7502	0.7809
14	Squirrel Ai Learning	0.5272	0.6904	0.7306	0.7692
15	Thapar Titans	0.5215	0.6943	0.7374	0.7796
16	jeez	0.5187	0.6833	0.7416	0.7854
17	CU	0.5148	0.6807	0.7358	0.7789
18	777	0.5114	0.6710	0.7195	0.7486
19	Someone	0.5009	0.7208	0.7590	0.8074
20	Retuyt-InCo	0.4959	0.5863	0.7200	0.7608
21	NLP Group 7	0.4936	0.6348	0.6944	0.7524
22	Smollab_SEU	0.4935	0.6057	0.7051	0.7401
23	lexilogic	0.4844	0.6548	0.7138	0.7447
24	mucai	0.4828	0.5495	0.7086	0.7343
25	Emergent Wisdom	0.4773	0.7188	0.7436	0.7893
26	2	0.4749	0.7279	0.7397	0.8003
27	Promptly Educated	0.4717	0.6432	0.6900	0.7337
28	Tutorify	0.4666	0.6626	0.7116	0.7447
29	NLP_UNH	0.4515	0.6994	0.6962	0.7725
30	Patriots	0.4450	0.6328	0.6548	0.7007
31	AUST_NLP	0.3044	0.4163	0.4759	0.4848

Table 14: Official leaderboard for *Track 2: Mistake Location*

Rank	Team	Ex. Acc
1	BLCU-ICALL	0.7679
2	K-NLPers	0.7641
3	SG	0.7602
4	bea-jh	0.7389
5	BJTU	0.7330

Table 15: Top 5 results according to exact accuracy for *Track 2: Mistake Location*

Rank	Team	Len. F1
1	K-NLPers	0.8404
2	BLCU-ICALL	0.8386
3	SG	0.8118
4	BJTU	0.7861
5	bea-jh	0.7851

Table 16: Top 5 results according to lenient F1 for *Track 2: Mistake Location*

Rank	Team	Len. Acc
1	BLCU-ICALL	0.8630
2	K-NLPers	0.8610
3	SG	0.8416
4	BJTU	0.8274
5	bea-jh	0.8268

Table 17: Top 5 results according to lenient accuracy for *Track 2: Mistake Location*

Rank	Team	Ex. F1	Ex. Acc	Len. F1	Len. Acc
1	MSA	0.5834	0.6613	0.7798	0.8190
2	SG	0.5785	0.7052	0.7860	0.8216
3	BLCU-ICALL	0.5741	0.6716	0.7487	0.8061
4	BJTU	0.5725	0.6490	0.7445	0.8100
5	K-NLPers	0.5606	0.6270	0.7446	0.8003
6	bea-jh	0.5451	0.6387	0.7253	0.7977
7	Wonderland_EDU@HKU	0.5416	0.6464	0.7456	0.7886
8	IALab UC	0.5369	0.6244	0.7379	0.7822
9	JiNan_Smart Education	0.5275	0.6432	0.7336	0.7893
10	Henry	0.5265	0.6238	0.7196	0.7744
11	TBA	0.5212	0.6219	0.7299	0.7906
12	MT-NLP	0.5211	0.6141	0.7142	0.7699
13	Archaeology	0.5208	0.5734	0.7171	0.7770
14	Averroes	0.5134	0.6309	0.7095	0.7751
15	Squirrel Ai Learning	0.5087	0.6005	0.7059	0.7763
16	jeez	0.5071	0.5831	0.7234	0.7763
	bd	0.5071	0.5831	0.7234	0.7763
17	Retuyt-InCo	0.5049	0.5947	0.7057	0.7751
18	woaiyuanshen	0.4974	0.5798	0.7034	0.7841
19	SmolLab_SEU	0.4933	0.5695	0.6990	0.7608
20	CU	0.4926	0.5850	0.7031	0.7692
21	Emergent Wisdom	0.4903	0.6102	0.6919	0.7725
22	NLIP	0.4888	0.6025	0.6927	0.7647
23	batikbabu	0.4873	0.6147	0.7001	0.7615
24	Whyiamhere	0.4856	0.6231	0.6880	0.7738
25	isistanNiem	0.4805	0.5844	0.6715	0.7589
26	Thapar Titans	0.4777	0.5624	0.6846	0.7479
27	DLSU	0.4776	0.5669	0.6755	0.7382
28	Tutorify	0.4731	0.5753	0.6709	0.7511
29	777	0.4711	0.6432	0.7075	0.7725
30	Promptly Educated	0.4674	0.6102	0.6785	0.7647
31	lexiLogic	0.4656	0.5869	0.6803	0.7473
32	GGEZ Lab	0.4596	0.5714	0.6652	0.7492
33	Patriots	0.4508	0.5663	0.6422	0.7311
34	NLP_UNH	0.4301	0.6380	0.6895	0.7692
35	AUST_NLP	0.4045	0.5973	0.6094	0.7259

Table 18: Official leaderboard for *Track 3: Providing Guidance*

Rank	Team	Ex. Acc
1	SG	0.7052
2	BLCU-ICALL	0.7007
3	MSA	0.6729
4	bea-jh	0.6703
5	TBA	0.6652

Table 19: Top 5 results according to exact accuracy for *Track 3: Providing Guidance*

Rank	Team	Len. F1
1	SG	0.7860
2	MSA	0.7798
3	BLCU-ICALL	0.7699
4	K-NLPers	0.7483
5	Wonderland_EDU@HKU	0.7456

Table 20: Top 5 results according to lenient F1 for *Track 3: Providing Guidance*

Rank	Team	Len. Acc
1	BLCU-ICALL	0.8222
2	SG	0.8216
3	MSA	0.8190
4	BJTU	0.8100
5	TBA	0.8035

Table 21: Top 5 results according to lenient accuracy for *Track 3: Providing Guidance*

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
6	Wonderland_EDU@HKU	0.6843	0.7285	0.8613	0.8888
7	Archaeology	0.6776	0.7214	0.8302	0.8565
8	BLCU-ICALL	0.6735	0.7363	0.8596	0.8856
9	TBA	0.6671	0.7324	0.8499	0.8752
10	4	0.6668	0.7033	0.8160	0.8520
	JiNan_Smart Education	0.6668	0.7033	0.8160	0.8520
11	bnl	0.6655	0.6813	0.8172	0.8597
12	woaiyuanshen	0.6651	0.7124	0.8191	0.8533
13	love-peace	0.6562	0.6839	0.8051	0.8352
14	bd	0.6554	0.7182	0.8461	0.8707
15	Thapar Titans	0.6324	0.6774	0.7936	0.8248
16	Smollab_SEU	0.6284	0.6955	0.8223	0.8565
17	Retuyt-InCo	0.6129	0.7033	0.8272	0.8559
18	NLIP	0.6055	0.6897	0.8205	0.8468
19	Squirrel Ai Learning	0.5954	0.6516	0.7639	0.8022
20	Tutorify	0.5681	0.6425	0.7749	0.8190
21	K-NLPers	0.5664	0.5773	0.7346	0.8061
22	Emergent Wisdom	0.5661	0.6645	0.7782	0.8054
23	SG	0.5465	0.6341	0.7545	0.7725
24	SAI	0.5398	0.6277	0.7564	0.8022
25	DLSU	0.5294	0.6089	0.7351	0.7738
26	Patriots	0.4630	0.5727	0.6943	0.7537
27	whyiamhere	0.4306	0.6044	0.7143	0.7938
28	AUST_NLP	0.4196	0.5262	0.6077	0.6833
29	NLP_UNH	0.3798	0.5546	0.6530	0.7524

Table 22: Official leaderboard for *Track 4: Actionability*

Rank	Team	Ex. Acc
1	bea-jh	0.7557
2	MSA	0.7537
3	BJTU BLCU-ICALL	0.7363
4	TBA	0.7324
5	Phaedrus	0.7298

Table 23: Top 5 results according to exact accuracy for *Track 4: Actionability*

Rank	Team	Len. F1
1	MSA	0.8659
2	BJTU	0.8633
3	Wonderland_EDU@HKU	0.8613
4	bea-jh	0.8609
5	BLCU-ICALL	0.8596

Table 24: Top 5 results according to lenient F1 for *Track 4: Actionability*

Rank	Team	Len. Acc
1	BJTU	0.8940
2	MSA	0.8908
3	Wonderland_EDU@HKU	0.8888
4	bea-jh	0.8875
5	BLCU-ICALL	0.8856

Table 25: Top 5 results according to lenient accuracy for *Track 4: Actionability*

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
6	Archaeology	0.8866	0.8882
7	Wonderland_EDU@HKU	0.8795	0.8778
8	MSA	0.8697	0.8649
9	SmolLab_SEU	0.8621	0.8604
10	mucai	0.8602	0.8675
11	Squirrel Ai Learning	0.8432	0.8390
12	Retuyt-InCo	0.8385	0.8475
13	whyiamhere	0.8356	0.8345
14	bnl	0.8247	0.8216
15	Tutorify	0.8212	0.8100
16	LexiLogic	0.8207	0.8145
17	Georgia Tech EDU	0.6468	0.6296
18	DLSU	0.6420	0.6231
19	letstea	0.1749	0.1635
20	zet-epsilon	0.1140	0.1965

Table 26: Official leaderboard for *Track 5: Tutor Identification*

Rank	Team	Ex. Acc
1	Phaedrus	0.9664
2	SYSUpporter	0.9657
3	Two Outliers	0.9412
4	JInan_Smart Education	0.8940
5	BLCU-ICALL	0.8908

Table 27: Top 5 results according to exact accuracy for *Track 5: Tutor Identification*

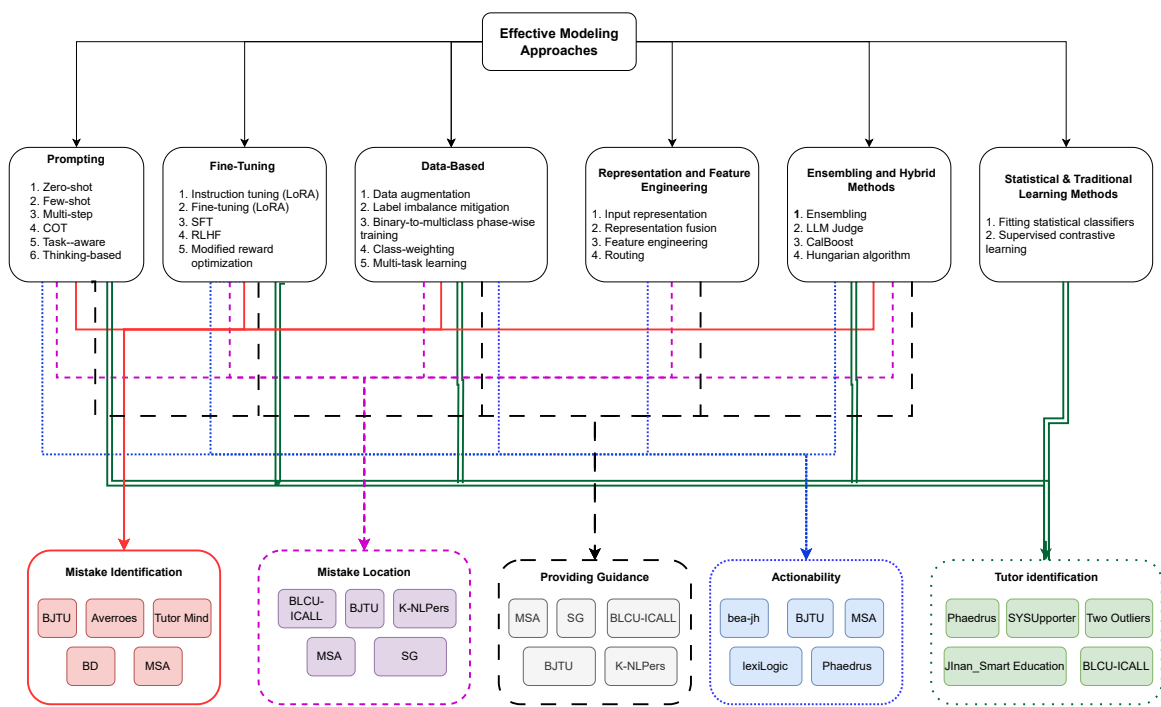


Figure 2: Overview of the effective modeling approaches adopted by top-performing teams for each track

Team	Keywords of the Approach	Models / LLMs
BJTU (Fan et al., 2025)	Zero-shot prompting, data augmentation, task-aware prompting	Unspecified
TutorMind (Dekmak et al., 2025)	Instruction tuning (LoRA), data augmentation	GPT-4o-mini, LLaMA-3.1-8B, Mistral-7B
Averroes (Yasser et al., 2025)	Instruction tuning (LoRA)	GTE-ModernBERT-Base, GTE-Qwen2-1.5B, Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen2.5-Math-1.5B
MSA (Hikal et al., 2025)	Instruction tuning (LoRA), ensembling	Mathstral-7B-v0.1
BD (Rohan et al., 2025)	SFT, class weighting, ensembling	MPNet
Wonderland_EDU@HKU (Wang et al., 2025)	Instruction tuning (LoRA)	LLaMA-3.2-3B
Archaeology (Roşu et al., 2025)	SFT, fine-tuning (LoRA), binary-to-multiclass phase-wise training, fitting statistical classifiers	Logistic Regression, LightGBM, String-Kernel-SVM, RoBERTa, DeBERTa, ModernBERT, GritLM, GPT2-XL, Mistral-7B, XGBoost
TBA (Gombert et al., 2025)	SFT, DARE-TIES algorithm	FLAN-T5-XL
BLCU-ICALL (An et al., 2025)	SFT, few-shot prompting, RLHF	GPT-4o, GPT-o3-mini, Gemini-2.5-pro, Grok-3, Deepseek-R1, Claude-3.7, LLaMA-3.1-8B, QwQ-32B, Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B
bea-jh (Roh and Bang, 2025)	GRPO, thinking-based model	GLM-4-9B
JiNan_Smart Education (Chen, 2025)	Input representation, representation fusion, ensembling	DeBERTa-v3
K-NLPers (Park et al., 2025)	Chain-of-thought prompting, multi-strategy ensembling, input representation	GPT-4.1
Thapar Titan/s (Dadwal et al., 2025)	Data augmentation, weighted loss, SFT	BERT, DeBERTa, RoBERTa
SmolLab_SEU (Rahman et al., 2025)	SFT	DeBERTa-V3, EduBERT, RoBERTa-Large, SciBERT
LexiLogic (Bhattacharyya et al., 2025)	SFT, zero-shot prompting, few-shot prompting	Flan-T5, Llama-3.2-3B, Llama-3-8B, ModernBERT, MathBERT, Phi-4-mini-instruct, Qwen2.5-7B-Instruct
Retuyt-InCo (Góngora et al., 2025)	Input representation, SFT, fitting statistical classifiers	Random Forest, SVC, k-NN, Qwen2.5-0.5B-Instruct, XGBoost, DistilBERT, BERT
CU (Lyu, 2025)	SFT, data augmentation, label imbalance mitigation	BERT, GPT-4.1
NLIP (Saha et al., 2025)	SFT, data augmentation, multi-task learning, ensembling	RoBERTa, DeBERTa
ALA*	SFT	BERT
Emergent Wisdom (Jain and Renegarajan, 2025)	Input representation, feature engineering, routing, fitting statistical classifiers, multi-step prompting, LLM judge	XGBoost, T5
SG*	Multi-step prompting	Gemma-3-27B-IT
NeuralNexus (Naeem et al., 2025)	RAG, few-shot prompting, fitting statistical classifiers	k-NN, GPT-4o
IALab UC (Busquets et al., 2025)	Zero-shot prompting, feature engineering, pedagogical theory, fitting statistical classifiers	LearnLM-1.5, Random Forest
Henry (Pit, 2025)	Zero-shot prompting, GRPO, fine-tuning (LoRA), modified reward optimization	GPT-4o, Claude 2.7 Sonnet, Phi-3.5-mini Instruct, MLP
DLSU (Manlises et al., 2025)	Input representation, fitting statistical classifiers	gte-modernbert-base, all-MiniLM-L12-v2, MLP
Phaedrus (Tiwari and Rastogi, 2025)	Zero-shot prompting, instruction tuning (LoRA), ensembling	DeBERTa-v3-large, DeBERTa-v3-base, DeBERTa-v3-small, Longformer-base-4096, BigBird-RoBERTa-large, Qwen-2.5-0.5B, Zephyr-7B-alpha
SYSUpporter (Chen et al., 2025)	Data augmentation, class weighting, ensembling, Hungarian algorithm	Logistic Regression, Random Forest, Extra Trees, XGBoost, DeBERTa
Two Outliers (Tjitrahardja and Hanif, 2025)	Input representation, supervised contrastive learning, ensembling, CalBoost, Hungarian algorithm	DeBERTa
Gooby-Snoob Guysz*	Prompt optimization, failure-driven prompting	OpenAI's O1, GPT-4o

Table 28: Keywords and models associated with the approaches adopted by participating teams across all tracks. SFT = Supervised Fine-Tuning, RAG = Retrieval-Augmented Generation, RLHF = Reinforcement Learning from Human Feedback, MLP = Multilayer Perceptron, GRPO = Guided Reward Prompt Optimization. *Details are obtained via email correspondence. *Statistical classifiers* include traditional models such as Random Forest (RF), XGBoost, etc.

Track/Criteria	Teams	Keywords for Approaches	Model/LLMs
*Mistake Identification	BJTU, Tutor Mind, Averroes, BD, MSA	Zero-shot prompting, data augmentation, task-aware prompting, instruction tuning (LoRA), ensembling, SFT, class weighting	GPT-4o-mini, LLaMA-3.1-8B, Mistral-7B, GTE-ModernBERT-Base, GTE-Qwen2-1.5B, Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen2.5-Math-1.5B, Mathstral-7B-v0.1, MPNet
*Mistake Location	BLCU-ICALL, BJTU, K-NLPers, MSA, SG	SFT, few-shot prompting, RLHF, zero-shot prompting, data augmentation, task-aware prompting, chain-of-thought prompting, multi-step prompting, multi-strategy ensembling, input representation	GPT-4o, GPT-o3-mini, Gemini-2.5-pro, Grok-3, Deepseek-R1, Claude-3.7, LLaMA-3.1-8B, QwQ-32B, Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B, GPT-4.1, Mathstral-7B-v0.1, Gemma-3-27B-IT
*Providing Guidance	MSA, SG, BLCU-ICALL, BJTU, K-NLPers	SFT, few-shot prompting, RLHF, zero-shot prompting, data augmentation, task-aware prompting, chain-of-thought prompting, multi-step prompting, multi-strategy ensembling, input representation	GPT-4o, GPT-o3-mini, Gemini-2.5-pro, Grok-3, Deepseek-R1, Claude-3.7, LLaMA-3.1-8B, QwQ-32B, Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B, GPT-4.1, Mathstral-7B-v0.1, Gemma-3-27B-IT
*Actionability	bea-jh, BJTU, MSA, LexiLogic, Phaedrus	GRPO, thinking-based model, zero-shot prompting, data augmentation, task-aware prompting, instruction tuning (LoRA), ensembling, SFT, few-shot prompting	GLM-4-9B, Mathstral-7B-v0.1, FLan-T5, ModernBERT, MathBERT, Phi-4-mini-instruct, Qwen2.5-7B-Instruct, DeBERTa-v3-large, DeBERTa-v3-base, DeBERTa-v3-small, Longformer-base-4096, BigBird-RoBERTa-large, Qwen-2.5-0.5B, Zephyr-7B-alpha
*Tutor Identification	Phaedrus, SYSupporter, Two Outliers, JInan_Smart Education, BLCU-ICALL	SFT, few-shot prompting, RLHF, zero-shot prompting, instruction tuning (LoRA), ensembling, data augmentation, class weighting, input representation, supervised contrastive learning, CalBoost, Hungarian algorithm, representation fusion	DeBERTa-v3-large, DeBERTa-v3-base, DeBERTa-v3-small, Longformer-base-4096, BigBird-RoBERTa-large, Qwen-2.5-0.5B, Zephyr-7B-alpha, GPT-4o, GPT-o3-mini, Gemini-2.5-pro, Grok-3, Deepseek-R1, Claude-3.7, LLaMA-3.1-8B, QwQ-32B, Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B
**Best Teams Across Tracks	BJTU, MSA, BLCU-ICALL	Zero-shot prompting, data augmentation, task-aware prompting, SFT, few-shot prompting, RLHF, instruction tuning (LoRA), ensembling	Mathstral-7B-v0.1, GPT-4o, GPT-o3-mini, Gemini-2.5-pro, Grok-3, Deepseek-R1, Claude-3.7, LLaMA-3.1-8B, QwQ-32B, Qwen2.5, Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B
***Best Generalizable Models Across Tracks	MSA, Wonderland_EDU@HKU, TBA	Instruction tuning (LoRA), ensembling, SFT, DARE-TIES algorithm	Mathstral-7B-v0.1, LLaMA-3.2-3B, FLAN-T5-XL

Table 29: *Indicates Summary of top-5 team approaches and the models used per evaluation track. **Indicates teams with an average rank within the top 5 and participation in at least four tracks. ***Indicates teams that applied a generalized approach across at least four tracks and achieved an average rank of 10 or higher.

Dialogue Context		
<p>Tutor: Hi, could you please provide a step-by-step solution for the question below? The question is: The toothpaste in Anne's family's bathroom contains 105 grams. Anne's dad uses 3 grams at each brushing, her mom uses 2 grams, Anne and her brother use 1 gram each. Each member of the family brushes their teeth three times a day. How many days will it take for the toothpaste to run out?</p> <p>Student: Each family member uses a total of $3+2+1+1 = 7$ grams of toothpaste per day. For the whole family, they use $7 \times 4 = 28$ grams of toothpaste per day. Therefore, the toothpaste will last for $105/28 = 3.75$ days. Since they can't use a partial day of toothpaste, the toothpaste will last for 3 days. 3</p> <p>Tutor: Hello, Luca! Can you walk me through your solution?</p> <p>Student: Sure! I calculated that each family member uses 7 grams of toothpaste per day. So for the whole family, they use 28 grams of toothpaste per day. Then I divided 105 by 28 to get 3.75 days. Since they can't use a partial day of toothpaste, I concluded that the toothpaste will last for 3 days.</p>		
Tutor	Tutor Response	Mistake Identification
Gemini	That's a great start, Luca! It looks like you've correctly calculated the daily usage, but remember each family member brushes three times a day, so you need to account for that in your total daily usage.	To some extent

Table 30: This case from the test set in the *Mistake Identification* track was the most challenging for participating teams. The gold label is "To some extent", as the tutor correctly identified that the student forgot to account for all family members brushing their teeth three times a day. This captures one part of the student's mistake. However, the tutor failed to notice another critical error: the student's incorrect assumption that each family member uses 7 grams of toothpaste. Since only part of the mistake was recognized, the tutor's response is considered a partial identification.

Tutor	Tutor Response
LLama-3.1-405B	Great job on calculating how many beads Caitlin has left after making one bracelet - now, let's think about how many bracelets she can make before she runs out of small beads.
LLama-3.1-8B	Now, can you multiply 80 by 2?

Table 31: These responses were the most challenging for participating teams in the *Tutor Identification* track – none of the teams correctly identified the tutor identities. The first response was frequently misclassified as coming from Sonnet, while the second was often labeled as Mistral or Expert.