AITutor-EvalKit: Exploring the Capabilities of AI Tutors

Numaan Naeem*, **Kaushal Kumar Maurya***,
Kseniia Petukhova and **Ekaterina Kochmar**
Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{Numaan.Naeem, kaushal.maurya, kseniia.petukhova, ekaterina.kochmar}@mbzuai.ac.ae

Abstract

We present AITutor-EvalKit, an application that uses language technology to evaluate the pedagogical quality of AI tutors, provides software for demonstration and evaluation, as well as model inspection and data visualization. This tool is aimed at education stakeholders as well as *ACL community at large, as it supports learning and can also be used to collect user feedback and annotation.

1 Introduction

Personalized one-on-one tutoring has long been recognized as a highly effective educational approach (Bloom, 1984). Yet, its widespread adoption is constrained by the limited availability of qualified tutors (Wang et al., 2024b) and the high costs associated with tutor training (Kelly et al., 2020), among other impediments (Yoon et al., 2007; Boyd et al., 2008). An alternative to human tutoring is provided by AI tutoring systems, especially those relying on recent advances in large language models (LLMs), such as Khanmigo (Khan, 2024) and Tutory.[1] Despite the remarkable success of LLMs in various tasks (Minaee et al., 2024), their adoption in education is hindered by lack of a clear understanding of what these models are capable of (Tack et al., 2023; Jurenka et al., 2024) and how pedagogically useful they are (Macina et al., 2023b), which results in lack of trust on the part of key educational stakeholders. With the fast development of LLMs and their easy integration into learning tools, questions about the evaluation of AI-driven tutor performance become increasingly relevant (Kosmyna et al., 2025). The goal of our tool is two-fold: (1) via the open-source and open-access code, we provide a practical, customizable and versatile evaluation tool that can be applied to a variety of educational scenarios; and (2) via informative demonstrations, we aim to raise awareness of the stakeholders (e.g., students and teachers) as well as practitioners interested in the state of the art in AI in Education (i.e., *ACL community at large) of the current capabilities of LLMs-as-tutors. Although our tool is an *early research prototype*, through its extendable nature, we aim to facilitate research in this exciting and emerging area of applied NLP research.

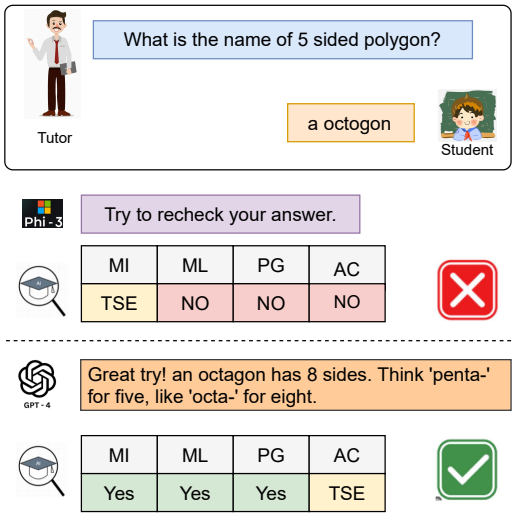

Figure 1: This example shows a sample dialogue and its pedagogical-ability evaluation by the LoMTL model using the AITutor-EvalKit. The evaluation follows the four dimensions from Kochmar et al. (2025): **MI** (Mistake Identification), **ML** (Mistake Location), **PG** (Providing Guidance), and **AC** (Actionability). **TSE:** To some extent.

The current version of our tool focuses on the pedagogical quality evaluation of tutor responses in the context of Student Mistake Remediation (SMR) (Boaler, 2013; Handa et al., 2023) in the mathematical domain, where tutors address errors or misconceptions that hinder students' progress (Wang et al., 2024a,b; Macina et al., 2023a). As the foundation for evaluation, we use the established taxonomy from Maurya et al. (2025), which is grounded in the learning

* Equal contribution.
[1] https://tutory.io

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 3: System Demonstrations, pages 457–479
March 24-29, 2026

sciences principles and which allows us to assess the quality of tutor responses along four key SMR dimensions (Kochmar et al., 2025): (1) *mistake identification*, concerned with whether tutor's response notifies the student of the committed mistake; (2) *mistake location*, focusing on whether the tutor clearly points to the erroneous part in the student's solution; (3) *providing guidance*, evaluating the quality of the pedagogical guidance; and (4) *actionability*, assessing whether tutor's response makes it clear what the student should do next. These dimensions are further outlined in Table 3, with an illustrative example provided in Figure 1.

The structure and implementation of the front- and backend of our tool are described in Section 3. Using `MRBench` dataset (Maurya et al., 2025) and taking inspiration from the BEA 2025 shared task on AI tutor response evaluation (Kochmar et al., 2025), we introduce a novel, efficient, and lightweight multi-task learning model that addresses the four dimensions of pedagogical quality evaluation (§3.1). The outputs of the model, as well as those of an open-source (`Prometeus` (Kim et al., 2024)) and a commercial (`GPT-5`) LLMs used as judges, are displayed using an interactive browser-based UI with helpful visualizations (§3.3). The evaluation results (§4) suggest that while our model achieves competitive results when evaluated against gold standard annotation, its outputs are also perceived by users to be at least as accurate as those of the commercial LLM-as-judge models, and the UI is considered informative and easy to use. We present and publicly release:

- The first of its kind, open-access and open-source model aimed at evaluation of the pedagogical quality of AI tutor responses available at MIT-licensed python repository: `https://github.com/kaushal0494/AITutor-EvalKit`. We believe it to be useful for AI-in-Education practitioners and developers, as it is highly customizable, allowing researchers to apply it to further educational contexts and dialogues, as well as to extend it to other scenarios and domains.
- An interactive UI available at `https://demo-ai-tutor.vercel.app`, which communicates the results and showcases the capabilities of AI tutors in an interpretable way, which we consider to be of interest to education stakeholders and the *ACL community at large. The demo tool can also be run locally with the user's own data and models following the instructions provided in the GitHub repository.
- A short video demonstrating the tool available at `https://www.youtube.com/watch?v=9qgDfrhzOvg`.

2 Related Work

Over the years, the NLP community has seen significant advances in the development of publicly available toolkits for modeling and evaluation, including Hugging Face Transformers (Wolf, 2019), NLTK (Bird, 2006), and Scikit-learn (Pedregosa et al., 2011). These toolkits have enhanced code reusability, enabling researchers to focus on developing more sophisticated models and metrics.

However, in the educational domain, especially in conversational dialogues, there remains a lack of robust tools to push research boundaries. Some progress has been made with frameworks like ConvoKit (Chang et al., 2020), which facilitates the manipulation and analysis of general conversational data, and the social interactions embedded within. More recently, Edu-ConvoKit (Wang and Demszky, 2024) was developed for preprocessing, annotation, and analysis specifically for educational dialogues. While these toolkits contribute significantly to data handling and analysis, they fall short of addressing the evaluation of pedagogical quality in AI-driven educational systems. To the best of our knowledge, there is no toolkit that supports *on-the-fly* assessment of the pedagogical abilities of AI tutors. With `AITutor-EvalKit`, we aim to fill this critical gap by providing an open-source toolkit for the systematic evaluation of AI tutors. The toolkit integrates a popular taxonomy proposed by Maurya et al. (2025) into an automated evaluation framework. It is also designed to be extensible to additional evaluation aspects such as the ones proposed by Macina et al. (2025), who augment pedagogical assessment with measures of student expertise and understanding. By supporting and unifying such evaluation methods, `AITutor-EvalKit` aims to facilitate progress in this underexplored yet important research area.

3 System and Demonstration Description

`AITutor-EvalKit` consists of two major modules: **backend** and **frontend**, with the pipeline illustrated in Figure 2. The backend module includes several models for tutor response evalu-

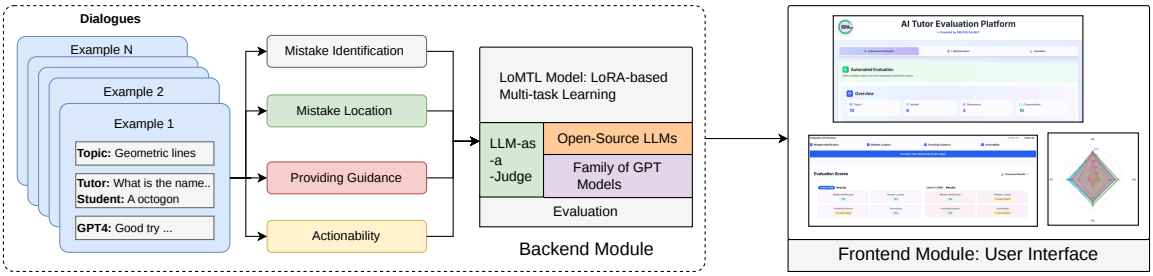

Figure 2: `AITutor-EvalKit` pipeline: The backend module includes several model options to assess the pedagogical soundness of tutors' responses, and the frontend presents evaluation outputs in an interactive user interface.

ation, and frontend seamlessly integrates these evaluation outputs with an interactive, customizable, and flexible UI. Different audience groups can benefit from different functionalities of the toolkit: e.g., researchers and developers can use both modules, train their own automated models, use their own datasets, and launch the demo UI locally, while teachers, policymakers and other educational stakeholders can use the frontend module to understand the capabilities of LLMs-as-tutors and make decisions accordingly.

3.1 Backend Module: Evaluation Models

The backend module has two components: *a specialized automated evaluation model* and *an evaluation pipeline for LLMs-as-judges*. In this section, we provide details on the task, the training data (for the automated evaluation model), the test data, and the automated evaluation model itself, as well as the usage of an open-source LLM and a closed-source GPT model as judges.

3.2 Student Mistake Remediation Task

This task considers mathematical educational dialogues between a student and a tutor, where interactions are driven by student's mistakes or confusions, and the AI tutor aims to remediate them through pedagogically appropriate responses. Formally, let the conversation history be $H = \{(T_1, S_1), (T_2, S_2), \ldots, (T_t, S_t)\}$, where T_i and S_i denote the tutor's and student's i-th utterances, respectively. Let S_k, $k \in [1, \ldots, t]$, denote the most recent student's utterance containing a mistake or confusion; the tutor then produces T_{t+1} to address it. The proposed toolkit evaluates the pedagogical quality of T_{t+1} along eight dimensions defined by Maurya et al. (2025).

3.2.1 Dataset

As discussed in Section 1, we used `MRBench` dataset by Maurya et al. (2025), which has 491 dialogues (300 in the development set and 191 in the test set), each paired with seven LLM-generated tutor responses and one or two human tutor responses. Each response is annotated by human annotators for each of the four dimensions using the categories "Yes," "To some extent," and "No." We also created a randomly selected *demonstration set* consisting of 10 dialogues, which is a subset of the test set. More details on the dimensions, annotation, and data statistics can be found in Appendix A.

3.2.2 Specialized Automated Model: `LoMTL`

This component implements a ternary classifier to evaluate tutor responses across the four considered dimensions. The models developed by participants in the BEA shared task provide a good starting point, as summarized by Kochmar et al. (2025). However, upon closer inspection, we found that several teams relied on closed-source models, fully fine-tuned and LoRA-tuned models, and some even used multiple models or ensembles for each dimension, which makes these approaches difficult to scale across all dimensions.

Considering this, we propose a novel LoRA-based multi-task model (called `LoMTL`) that fine-tunes a single small LLM in a LoRA setting across all four dimensions, each treated as a task in a multi-task learning setting. This modeling is naturally suited to the four tasks, all of which are ternary, closely related classification problems, allowing them to benefit from shared learning during training. Additionally, we use sampling and balanced batching methods to improve performance. We observe that a small `google/gemma-2-2b-it` (Team et al., 2024) model using `LoMTL` achieves performance compet-

itive with the top-performing teams in the shared task while remaining highly efficient and scalable. Appendix B presents more details on model development, prompts, configuration, and comparative results with the BEA shared task's top-performing teams.

3.2.3 LLM-as-a-Judge Evaluation

This component provides functionality for evaluation using LLMs-as-judges. Following Maurya et al. (2025), we have selected `Prometheus2` (Kim et al., 2024) as the primary open-source LLM. However, the implementation of this module is flexible enough to support the family of `Llama` and other open-source causal LLMs for evaluation. Additionally, we have included an option to evaluate using the closed-source OpenAI `GPT-5` model, although the implementation supports any model from the GPT family. Users must provide their OpenAI API key to run evaluations. We selected the above two models considering their open- and closed-source nature, as well as their human-like performance on public benchmarks (Kim et al., 2024). However, the codebase is flexible enough to require only minimal adaptation to support other LLMs.

3.2.4 Flexible Design

Each of the three evaluation setups (automated evaluation, evaluation with open-source LLMs, and evaluation with closed-source GPT models) provides high degree of flexibility in the choice of the base LLM, prompting strategy, and hyperparameter tuning. For example, each evaluation setup has its own prompting file and configuration, allowing users to customize and use the components of the tool as per their needs. Evaluation can be run using the following short commands:

```
# Training LoMTL model
cd src/autoeval && ./lora_finetune_runner.sh

# Evaluation with automated (i.e., LoMTL) model
cd src/autoeval && ./lora_evaluation_runner.sh

# Evaluation with open-source LLM
cd src/llmeval &&
python run_open_llm_as_judge_evaluation.py

# Evaluation with GPT5
cd src/llmeval && ./gpt5_eval_runner.sh
```

3.3 Frontend Module: Demo App

We present an interactive prototype built on top of `MRBench`, designed to evaluate the pedagogical abilities of AI tutors in educational dialogues. Users can explore the demo in two modes: (1) a **Static Mode**, where educators can access a deployed version containing 10 conversations from the `MRBench` test set evaluated by our fine-tuned model and share feedback on the helpfulness of tutor responses; and (2) a **Customized Mode**, where developers can run the interface locally to analyze their own datasets using either our `LoMTL` model or their own one, supported by the provided code (the exact step-by-step details are supplied in the official GitHub README file[2]). The UI consists of three key modules **Automated Evaluation**, **LLM Evaluation**, and **Visualizer**, that together enable exploration, analysis, and visualization of the key aspects of educational dialogues.

- **Automated Evaluation** provides automated evaluation results using our `LoMTL` model on 4 evaluation dimensions.
- **LLM Evaluation** leverages LLMs-as-judges for human-aligned evaluations.
- **Visualizer** enables rich visual analytics for interpreting evaluation scores across 4 pedagogical dimensions on the `MRBench` development set.

The UI supports the selection of evaluation methods and provides instant visualization of performance metrics through plots, bar charts, and spider graphs.

3.3.1 Automated Evaluation UI

The Automated Evaluation module presents results generated by our `LoMTL` model, assessing AI tutors across four pedagogical dimensions. As shown in Figure 6, users are first presented with an overview panel summarizing key statistics such as the number of topics, models, dimensions, and conversations. Users can evaluate a single tutor or compare two tutors side by side.

Single Tutor Evaluation: In this mode, users select a problem topic to view the complete student-tutor dialogue in the "Context" block (Figure 7). A drop-down menu allows selection of a tutor model, and the corresponding response is displayed in the "Tutor Response" block. Users can rate the usefulness of the response (*Helpful*, *To Some Extent*, or *Not Helpful*) and optionally view the ground truth solution for better context. Upon clicking "Get Auto-Evaluation Results," the

[2] https://github.com/kaushal0494/AITutor-EvalKit/blob/main/README.md

system generates performance evaluations across the chosen dimensions, with results downloadable in a PNG, JPG, or JSON format. The "Best Performance by Dimension" panel highlights the top tutor(s) for each dimension, helping users quickly identify their pedagogical strengths.

Tutor Comparison Mode: This mode allows users to directly compare two tutors by enabling the "Tutor Comparison Mode" option. The selected tutors' responses are displayed side by side (Figure 8), and users can provide quick feedback indicating which tutor performed better or mark both as equally good or bad. After choosing the evaluation dimensions and clicking "Compare Tutor Responses," the interface presents a detailed two-column comparison of scores across all selected dimensions. To facilitate interpretation, the "Comparison Visualization" panel (Figure 9) provides four interactive views: the **Summary** view highlights the leading tutor for each evaluation dimension and identifies the overall winner; the **Spider Chart** offers a radar-style visualization comparing performance patterns across dimensions; the **Bar Chart** displays side-by-side scores for each dimension; and the **Differences** view illustrates the magnitude of score gaps between tutors. All visualizations can be exported in PNG or JPG formats for reporting or analysis. Finally, the "Best Performance by Dimension" panel summarizes the comparative strengths of the tutor pair, providing a concise overview of pedagogical differences. This mode supports structured, interpretable, and visually grounded benchmarking of AI tutor performance.

3.3.2 LLM Evaluation UI

The LLM Evaluation module enables advanced pedagogical assessment of AI tutor responses using LLMs as judges. It extends the functionality of the automated evaluation pipeline by leveraging LLMs to judge and compare tutor responses across four pedagogical dimensions. The UI supports three evaluation modes: *single tutor evaluation*, *tutor model comparison*, and *LLM judge comparison*. As shown in Figure 10, the overview panel summarizes available topics, conversations, tutor models, evaluation dimensions, and judge LLMs, providing a quick snapshot of the evaluation setup. Currently, two LLM judges are supported: `GPT-5` and `Prometheus-7B-v2.0` (Kim et al., 2024).

Single Tutor Evaluation: This mode allows users to analyze how a selected tutor performs on a specific problem using an LLM as a judge. After selecting the problem, tutor, and LLM judge, users can generate dimension-wise evaluation results that reflect the LLM's assessment of the tutor's pedagogical performance. A "Best Performance by Dimension" panel highlights the top-performing tutor(s) for each dimension.

Tutor Comparison Mode: This mode enables side-by-side comparison of two tutor models on the same problem, judged by a selected LLM. Tutor responses are displayed together (Figure 11), and upon comparison, the system presents dimension-wise evaluations and visualizations such as Summary, Spider Chart, Bar Chart, and Differences to clearly show relative strengths across pedagogical aspects.

Judge Comparison Mode: This mode compares how different LLM judges evaluate the same tutor response. The tutor's response is displayed once, and evaluations from both judges are presented in parallel with corresponding visualizations. This feature helps assess consistency between LLM judges and identify possible biases in their evaluations.

Together, these modes enable fine-grained, interpretable analysis of tutor behavior, offering insights into both model performance and evaluation reliability across different judging LLMs.

3.3.3 Visualizer UI

The Visualizer module provides a high-level overview of the `MRBench` development set, using gold-standard annotations across four evaluation dimensions through intuitive visual analytics. Users are first presented with a "Dataset Overview" panel summarizing key statistics, including the number of conversations, tutor models, and evaluation dimensions. This module includes three main visualization panels: *Tutor Performance Summary*, *Visualization Controls*, and *Dataset Visualization*.

The **Tutor Performance Summary** panel presents average scores for each tutor model across all four dimensions, where "*Yes*," "*To some extent*," and "*No*" correspond to 1.0, 0.5, and 0.0, respectively (Figure 12). The **Visualization Controls** panel allows users to select specific tutors and dimensions to generate detailed visualizations (Figure 13). The **Dataset Visualization** panel then displays the results through spider and bar charts, where spider plots (the *most informative* visualizations from our perspective) summarize tutors'

strengths and weaknesses across dimensions (Figure 14), while bar charts show detailed score distributions for selected dimensions (Figure 15), along with averages for each response label.

This visual analytics module enables users to explore and interpret pedagogical quality effectively, supporting comparative analysis and data-driven insights. The same functionality is also available in the customized mode, allowing developers to visualize and analyze their own datasets locally in a similar way.

4 Evaluation

To assess the toolkit and evaluation models, we first measure models' performance using the metrics from the BEA 2025 shared task (Kochmar et al., 2025) – accuracy and macro-F1, and then conduct a human evaluation study, in which participants assess both the prediction quality of the LoMTL evaluation model and the usability of our demo tool and UI.

4.1 Intrinsic Evaluation: Quantitative Analysis

For intrinsic evaluation, we compare our LoMTL model's predictions on the test set from Kochmar et al. (2025) with the gold human annotations and also run GPT-5 and Prometheus2 on the same data. The average results across evaluation dimensions, presented in Table 1, show that Prometheus2 performs substantially worse than both our model and GPT-5. Our model achieves the highest averaged accuracy and macro-F1 on the full test set, and it also performs competitively on the demonstration subset. While GPT-5 achieves a slightly higher averaged macro-F1 on the demonstration subset, the overall trend indicates that our model provides more reliable evaluations than Prometheus2 and performs on par with, or better than, GPT-5. A close inspection of the confusion matrix between human annotations and the LoMTL model shows strong overall agreement, with the majority of instances concentrated on the diagonal, particularly for clear *Yes* (3,106) and *No* (1,057) cases. However, the model exhibits systematic confusion on borderline instances and a mild tendency to overpredict *Yes*, especially when humans assign *To some extent* or *No* labels. Extended results and a related discussion, including average precision and recall scores, are provided in Table 7 in Appendix C.

Model	Full Test Set		Demonstration Set	
	Accuracy	Macro-F1	Accuracy	Macro-F1
LoMTL (ours)	**0.72**	**0.60**	**0.68**	0.55
Prometheus2	0.47	0.41	0.41	0.34
GPT-5	0.66	0.58	0.66	**0.59**

Table 1: Accuracy and macro-F1 scores (averaged across Mistake Identification (MI), Mistake Location (ML), Providing Guidance (PG), and Actionability (AC)) for LoMTL, Prometheus2, and GPT-5 on the full test set from Kochmar et al. (2025) and on the demonstration set. Best results are shown in **bold**.

Performance across dimensions: The per-dimension results, presented in Table 2, show that Prometheus2 performs poorly, while our model notably outperforms GPT-5 on the Mistake Identification and Actionability dimensions. However, it underperforms GPT-5 by 3 and 6 percentage points in terms of macro-F1 for Mistake Location and Providing Guidance, respectively. A similar trend is observed in the ten dialogues used for demonstration.

4.2 Extrinsic Evaluation: Human Study

Participants were given access to the demo website, detailed guidelines, and an evaluation form. The form instructed them to assess three components: the Automated Evaluation tab, the LLM Evaluation tab, and the Visualizer tab, as detailed in Section 3.3. For the first two tabs, participants were asked to explore at least five different dialogues, review at least two tutor responses per dialogue, and use the feedback field to rate each response as *helpful*, *helpful to some extent*, or *not helpful*. They also examined the model's evaluation for each response across at least one assessment dimension. Additionally, for each dialogue, participants compared at least one pair of tutor responses in comparison mode and indicated which response they considered a better one. After using Automated Evaluation tab, they reported how frequently they agreed with the model's judgments on a 1-5 scale, both in single-response and comparison modes. In the LLM Evaluation tab, participants evaluated how often they agreed with GPT-5 and Prometheus2, again in both modes. They were also asked which evaluation model they perceived as more accurate: GPT-5 or the model from the first tab (which corresponds to our LoMTL model), and Prometheus2 or the model from the first tab. Finally, in the Visualizer tab, participants explored visualizations for at least two tutors and

Model	Full Test Set								Demonstration Set							
	Accuracy				Macro-F1				Accuracy				Macro-F1			
	MI	ML	PG	AC	MI	ML	PG	AC	MI	ML	PG	AC	MI	ML	PG	AC
LoMTL (ours)	**0.86**	0.67	0.63	**0.70**	**0.66**	0.55	0.54	**0.65**	**0.76**	**0.69**	0.60	**0.68**	**0.57**	0.50	0.47	**0.65**
Prometheus	0.58	0.53	0.31	0.46	0.48	0.42	0.32	0.43	0.48	0.41	0.42	0.32	0.34	0.30	0.38	0.30
GPT-5	0.67	**0.68**	**0.70**	0.58	0.53	**0.58**	**0.61**	0.55	0.67	0.64	**0.67**	0.66	0.53	**0.56**	**0.59**	0.64

Table 2: Accuracy and macro-F1 scores of our model, Prometheus, and GPT-5 across Mistake Identification (MI), Mistake Location (ML), Providing Guidance (PG), and Actionability (AC) on the full test set from Kochmar et al. (2025) and on the demonstration set. Best results are shown in **bold**.

rated how informative they found them on a 1-5 scale. For every tab, participants also rated the ease of use on a 1-5 scale. The full questionnaire is provided in Appendix C.

A total of 14 participants took part in the study. Their educational background included individuals pursuing a Master's degree, those holding a Master's degree, and those holding a PhD. Eleven participants had teaching experience, and eight had prior experience using an AI tutor.

Most participants perceived the LoMTL evaluation model as more accurate in its judgments than both GPT-5 and Prometheus2. As shown in Figure 3, the majority of participants agreed with the LoMTL model's assessments more than half of the time in both single-response and comparison modes. A similar trend was observed for GPT-5: most participants agreed with its judgments more than half of the time, although a larger proportion reported agreement only about half of the time. In contrast, participants tended to agree with Prometheus2 less than half of the time in the single-response mode, but more than half of the time in the comparison mode. Overall, most participants rated the ease of use of all tabs as *very easy* and found the visualizations *very informative*.

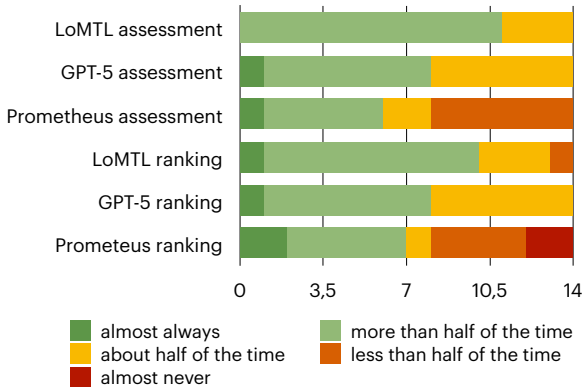

Figure 3: Participants' responses indicating how often they agreed with the models' judgments in single-response and comparison modes.

In total, we collected 95 annotations for single-tutor responses and 115 for pairwise comparisons. The analysis of these annotations is provided in Appendix C.2. We do not present it here because evaluating tutor performance is not the focus of this work. Instead, our goal is to demonstrate how our interactive UI can be used for annotation purposes by collecting data that can later be used to train evaluation models or align language models.

5 Conclusions and Future Work

This paper introduces the first open-access, open-source model for pedagogical quality evaluation of AI tutor responses, released under an MIT license. It is grounded on pedagogical principles and presents multiple evaluation choices including our proposed light-weight multi-tasking LoMTL model. The toolkit is highly customizable, allowing researchers and practitioners to extend it to diverse educational contexts and dialogues across domains, and it can be easily set up and run locally. In addition, we provide an interactive web-based UI that offers interpretable evaluations of the pedagogical capabilities of state-of-the-art LLMs acting as AI tutors for education stakeholders, policymakers, and non-technical audience.

Future work will extend the toolkit by (1) integrating new user dialogues and LLMs in the frontend module to generate real-time responses and evaluations; (2) expanding the evaluation dimensions; (3) enabling new data upload option in UI; and (4) broadening coverage to additional subjects, grade levels, and languages.

Limitations

We acknowledge that our work has several limitations. Below, we summarize the major ones among them.

Domain and grade level: In this work, we focus on the mathematical tasks at the middle-school

level. This decision is motivated by the availability of the data at this level and in this domain, but we plan to extend our work to other domains and levels as we elaborate in Section 5. Moreover, since users can run our tool on their own data, new domains and levels can already be integrated on the user's side.

Language: Similarly, our current work focuses on English only. In the future, we hope to extend it to other languages, as we specify in Section 5, while users of our tool can also apply it to data in other languages with their own evaluation models and LLMs-as-judges on their side.

Educational scenario: Building on previous work in this domain (Maurya et al., 2025; Kochmar et al., 2025), we only address student mistake remediation as an educational scenario. While this is one of the most salient and challenging scenarios in educational dialogues, we recognize this as one of the limitations of our work and plan to address it in the future.

Context length: Similarly, following up on the previous work, our current evaluation approach is limited to a single turn in the dialogue. We acknowledge this as a limitation, and believe that future work should extend single-turn approaches to multiple-turn or full-dialogue ones.

Taxonomy: We have built our prototype tool around a well-established evaluation taxonomy of Maurya et al. (2025). While using a single taxonomy is a limitation, our code is open-source and can be extended to incorporate other data, dialogues, taxonomies, and evaluation models on the user's side.

Models: Finally, via our demo tool, we only showcase a few LLMs-as-tutors and deploy only two LLMs-as-judges. Since our code is open-source and extendable, more tutor models and LLMs-as-judges can be integrated on the user's side.

Ethical Considerations

As this work is exploratory, we do not anticipate any significant ethical risks associated with it. Moreover, one of our main goals in this research is to raise awareness of the education stakeholders as well as practitioners interested in the state of the art in AI in Education and the current capabilities of LLMs-as-tutors. Through transparent evaluation of such models, we aim to improve their interpretability, which we hope will help avoid potential future risks associated with a wider adoption of these models in education.

This work uses the MRBench dataset (Maurya et al., 2025), which integrates the MathDial (Macina et al., 2023a) and Bridge (Wang et al., 2024a) datasets. In these datasets, the identities of the tutors are not revealed, while the student profiles are either synthetically created or anonymized. As a result, we do not anticipate any direct ethical risks associated with the datasets used.

Acknowledgments

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

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 Altenschmidt, 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 (BEA 2025)*, pages 1084–1097, Vienna, Austria. Association for Computational Linguistics.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. In `https://api.semanticscholar.org/CorpusID:268232499`.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Benjamin Samuel Bloom. 1984. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, 13:16 – 4.

Jo Boaler. 2013. Ability and mathematics: The mindset revolution that is reshaping education. Forum.

Donald J. Boyd, Pam L. Grossman, Hamilton Lankford, Susanna Loeb, and James Humphrey Wyckoff. 2008. Teacher Preparation and Student Achievement. *Educational Evaluation and Policy Analysis*, 31:416–440.

Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Z Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. *arXiv preprint arXiv:2005.04246*.

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 (BEA 2025)*, pages 1073–1077, Vienna, Austria. Association for Computational Linguistics.

Kunal Handa, Margarett Clapper, Jessica Boyle, Rose Wang, Diyi Yang, David Yeager, and Dorottya Demszky. 2023. "Mistakes Help Us Grow": Facilitating and Evaluating Growth Mindset Supportive Language in Classrooms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8877–8897, Singapore. Association for Computational Linguistics.

Baraa Hikal, Mohamed Basem, Islam 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 (BEA 2025)*, pages 1194–1202, Vienna, Austria. Association for Computational Linguistics.

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*.

Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, and 1 others. 2024. Towards responsible development of generative AI for education: An evaluation-driven approach. *arXiv preprint arXiv:2407.12687*.

Sean Kelly, Robert Bringe, Esteban Aucejo, and Jane Cooley Fruehwirth. 2020. Using global observation protocols to inform research on teaching effectiveness and school improvement: Strengths and emerging limitations. *Education Policy Analysis Archives*, 28:62–62.

Sal Khan. 2024. Khanmigo.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Ekaterina Kochmar, Kaushal Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack, and Justin Vasselli. 2025. Findings of the BEA 2025 shared task on pedagogical ability assessment of AI-powered tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1011–1033, Vienna, Austria. Association for Computational Linguistics.

Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023a. MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. MathTutorBench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 204–221, Suzhou, China. Association for Computational Linguistics.

Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023b. Opportunities and Challenges in Neural Dialog Tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2357–2372, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuren Mao, Yuhang Ge, Yijiang Fan, Wenyi Xu, Yu Mi, Zhonghao Hu, and Yunjun Gao. 2025. A survey on lora of large language models. *Frontiers of Computer Science*, 19(7):197605.

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.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Mathieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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 (BEA 2025)*, pages 1049–1059, Vienna, Austria. Association for Computational Linguistics.

Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. The BEA 2023 shared task on generating AI teacher responses in educational dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 785–795, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Rose Wang and Dorottya Demszky. 2024. EduConvoKit: An Open-Source Library for Education Conversation Data. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 61–69, Mexico City, Mexico. Association for Computational Linguistics.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024a. Bridging the Novice-Expert Gap via Models of Decision-Making: A Case Study on Remediating Math Mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199, Mexico City, Mexico. Association for Computational Linguistics.

Rose E Wang, Ana T Ribeiro, Carly D Robinson, Susanna Loeb, and Dorottya Demszky. 2024b. Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise. EdWorkingPaper No. 24-1054. *Annenberg Institute for School Reform at Brown University*.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Kwang Suk Yoon, Teresa Duncan, Silvia Wen-Yu Lee, Beth Scarloss, and Kathy L Shapley. 2007. Reviewing the evidence on how teacher professional development affects student achievement. issues & answers. rel 2007-no. 033. ERIC.

A Pedagogical Dimensions and Data

A.1 Evaluation Dimensions

Maurya et al. (2025) proposed an evaluation taxonomy with eight dimensions to assess the pedagogical soundness of T_{t+1} tutor response in the context of SMR. These dimensions are grounded in learning science research and prior work on tutor evaluation, defining the assessment via eight concrete criteria. Furthermore, the authors validated these dimensions as necessary and sufficient through a human pilot study. They also released the associated `MRBench` dataset with human annotations (see §A.2). Following this, Kochmar et al. (2025) focused on four key dimensions from this taxonomy for the BEA shared task, where participating teams were challenged to develop a ternary classification model for each of the four dimensions. These classifiers aimed to assess the pedagogical quality of the T_{t+1} response *on-the-fly*, enabling evaluation scalability with new data and tutors. Details on each of the four dimensions, along with their definitions, annotated labels, and desiderata, are provided in Table 3. Building on this, our `AITutor-EvalKit` toolkit focuses on the four key dimensions used in the BEA shared task.

A.2 Dataset Details and Statistics

We used the `MRBench` dataset to develop the toolkit and the automated evaluation model (i.e., `LoMTL`). The initial version of `MRBench`, released by Maurya et al. (2025), contains 191 dialogues. This was extended by Kochmar et al. (2025) which results in 300 dialogues in the development set and 191 dialogues in the test set. In this work, we use the extended version of the `MRBench` dataset.

The dataset is built on top of two public datasets – `MathDial` (Macina et al., 2023a) and `Bridge` (Wang et al., 2024a) – which provide partial conversational histories from secondary and primary school-level mathematics, respectively, along with human tutor responses as T_{t+1}. `MathDial` includes only one expert tutor response, whereas Bridge includes two responses from expert and novice human tutors. Additionally, each dialogue includes seven T_{t+1} responses generated by seven *state-of-the-art* LLMs-as-tutors, 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 Phi3 (Abdin et al., 2024).

All T_{t+1} responses (whether from human tutors or LLMs) are annotated by human annotators across four selected dimensions with labels "*Yes*," "*To some extent*," and "*No*," as detailed in Table 3. The proposed `LoMTL` model is trained on the development set and evaluated on the test set; all results presented in this work are reported on the test set. We further split the development set into a 9:1 ratio for training and validation when developing the `LoMTL` model. All model checkpoints were selected based on validation performance. Finally, we randomly selected a subset of 10 dialogues from the test set as a demonstration set, which is used in the demo app. These details are summarized in Table 4.

B `LoMTL` Evaluation Model

B.1 Motivation

Building `LoMTL` has a two-fold motivation: (1) The current human annotation-based pedagogical ability assessments presented by Maurya et al. (2025) are static in nature. They are not scalable to new LLMs and tutoring systems, which are being developed very frequently nowadays. We need a reliable *automated* evaluation model that can provide pedagogical assessment *on-the-fly* for new tutors or responses and help track the progress of AI tutor abilities. (2) The BEA shared task (Kochmar et al., 2025) is a good starting point for developing an automated evaluation model. More than 50 international teams participated in the challenge and proposed several novel modeling approaches, including diverse prompting strategies, full instruction tuning, LoRA-based finetuning, supervised finetuning, data augmentation, label balancing, ensembling and so on. However, most teams that participated in all four tracks did not develop a unified approach (with the exception of the MSA team (Hikal et al., 2025)) and instead used models with a large number of parameters. This hinders the adaptability of these approaches, as their deployment becomes challenging and costly.

These limitations motivated us to develop `LoMTL`, a lightweight model with only 2 billion parameters, created by training the `google/gemma-2-2b-it` model using LoRA in a multi-task learning setting. It achieves comparable performance while being significantly more efficient (see Table 5 for comparison). For instance, the top-performing BJTU team (Fan et al., 2025) achieved a macro-F1 score of 0.645 using 288 billion parameters (across all

Dimension	Definition	Labels	Desiderata
Mistake identification	Has the tutor identified a mistake in a students response?	(1) Yes (2) To some extent (3) No	Yes
Mistake location	Does the tutor's response accurately point to a genuine mistake and its location?	(1) Yes (2) To some extent (3) No	Yes
Providing guidance	Does the tutor offer correct and relevant guidance, such as an explanation, elaboration, hint, examples, and so on?	(1) Yes (2) To some extent (3) No	Yes
Actionability	Is it clear from the tutor's feedback what the student should do next?	(1) Yes (2) To some extent (3) No	Yes

Table 3: An overview of the evaluation taxonomy, associated definitions, annotation labels, and desired labels from Maurya et al. (2025).

Parameters	Value/Details
Number of dialogues (Dev / Test / Total)	300 / 191 / 491
Number of tutor responses (Dev / Test / Total)	2,476 / 1,547 / 4,023
Number of tutors (Total)	9
Number of human tutors	2 (1 expert, 1 novice)
Number of LLM tutors	7
LLM tutor models	GPT-4, Gemini, Sonnet, Mistral, Llama-3.1-8B, Llama-3.1-405B, Phi-3
Source datasets	`MathDial`, `Bridge`
`MathDial`	Expert human tutor only
`Bridge`	Expert and Novice human tutors
Demonstration set size	10 dialogues (from test set)

Table 4: Key details of the extended `MRBench` dataset.

four dimensions). In contrast, the `LoMTL` model achieved 0.60 with only 2 billion parameters, approximately 0.7% of the BJTU model's parameter count.

B.2 Training and Inference

In this section, we provide details on the training and evaluation of the `LoMTL` model. Inspired by the success of LoRA-based modeling from the BEA shared task (Kochmar et al., 2025) and by the community (Mao et al., 2025), we adapted a LoRA-based fine-tuning approach. Since we have a small amount of training data (approximately 2500 examples for each dimension) and the different tasks are somewhat related, the natural modeling choice is multi-task learning, where each evaluation dimension is formulated as a task. This LoRA-based multi-task fine-tuning approach (called `LoMTL`) resulted in a compact single model and enabled flexibility in model deployment. Further, we experimented with a small `google/gemma-2-2b-it` model, which resulted in fast inference. We observed two major issues during training: *task imbalance* and *label imbalance*. To mitigate task imbalance, we implemented a balance batching where each batch has an uniform number of examples from each task. For label imbalance, we explored several approaches such as focal loss, label sampling, loss weighting, and sampling methods. We obtained the best performance with oversampling where we randomly sample underrepresented examples in the training dataset. Model training and inference were done with a single 48GB A6000 GPU. The best checkpoints were obtained using validation data (10% of the development dataset).

B.3 Prompts and Configurations

Prompt structure and training/evaluation configurations for the `LoMTL` model are shown in Figure 4 and Table 6, respectively.

C Toolkit Evaluation Details and Results

C.1 Extended Evaluation Results

In addition to the observations on the accuracy and macro-F1 scores in Section 4.1, a closer inspection of precision and recall (from Table 7) further supports our findings. On the full test set, `LoMTL` achieves the highest macro-precision (0.63) while maintaining competitive recall (0.59), indicating more accurate and consistent positive predictions compared to both baselines. `GPT-5` attains slightly higher recall (0.60) but with lower

Team/Model	Macro-F1	# LLMs	# Parameters	Parameter Size vs. 2B
BJTU (Fan et al., 2025)	0.646	4 (`Qwen/Qwen2.5-72B`)	$4 \times 72\text{B} = 288\text{B}$	$144\times$ larger
MSA (Hikal et al., 2025)	0.643	5×4 (`mistralai/Mistral-7B-v0.1`)	$20 \times 7\text{B} = 140\text{B}$	$70\times$ larger
BLCU-ICALL (An et al., 2025)	0.632	4 (`Qwen/Qwen2.5-7B`)	$4 \times 7\text{B} = 28\text{B}$	$14\times$ larger
bea-jh (Roh and Bang, 2025)	0.625	4 (`zai-org/glm-4-9b`)	$4 \times 9\text{B} = 36\text{B}$	$18\times$ larger
`Prometheus2` (Kim et al., 2024)	0.410	4 (`prometheus-7b-v2.0`)	$4 \times 7\text{B} = 28\text{B}$	$14\times$ larger
`GPT-5` (Achiam et al., 2023)	0.581	-	-	-
`LoMTL` (ours)	0.601	1 (`google/gemma-2-2b-it`)	2B	-

Table 5: Comparison of macro-F1 scores and model parameters between our automated evaluation model (i.e., `LoMTL`) and the top-performing teams in the BEA shared task and LLM-as-a-judge models across four dimensions (aka. tasks). Note that, (1) for a fair comparison, we include only the teams that participated in all four tracks of the shared task, and (2) since GPT-5 is a closed-source model, its parameter details are not publicly available.

SYSTEM PROMPT

You are an expert evaluator of AI tutors.
For the given ### Task, ### Task Definition, ### Label Definition, ### Conversation History and ### Tutor Response, assess the pedagogical appropriateness of the Tutor Response.
Output exactly one label without additional text: Yes, "No", or "To some extent".

TASK DEFINITIONS

Mistake Identification	Has the tutor identified or recognized a mistake in a student's response?
Mistake Location	Does the tutor's response accurately pinpoint the location of a genuine mistake?
Providing Guidance	Does the tutor offer correct and relevant guidance, such as an explanation, elaboration, hint, examples, and so on?
Actionability	Is it clear from the tutor's feedback what the student should do next?

LABEL DEFINITIONS

Mistake Identification	**Yes:**	The tutor correctly identified the mistake in the student's response.
	To some extent:	The tutor partially recognized the mistake but did not fully capture it.
	No:	The tutor failed to identify any mistake.
Mistake Location	**Yes:**	The tutor accurately points to the exact mistake and its location.
	To some extent:	The tutor points to a mistake but imprecisely or partially.
	No:	The tutor fails to indicate the mistake or its location.
Providing Guidance	**Yes:**	The tutor provides correct and relevant guidance, hints, examples, or explanation.
	To some extent:	The guidance is partially correct or not fully helpful.
	No:	The tutor fails to provide relevant guidance.
Actionability	**Yes:**	It is clear what the student should do next.
	To some extent:	The next steps are somewhat unclear or incomplete.
	No:	The feedback does not indicate any actionable steps.

FINAL PROMPT STRUCTURE

```
content = (
        f"{cls.SYSTEM_PROMPT_WITH_LABEL}\n\n"
        f"### Task: {task}\n"
        f"### Task Definition: {task_def}\n\n"
        f"### Label Definition: \n{label_def_str}\n\n"
        f"### Conversation History: {conversation.strip()}\n\n"
        f"### Tutor Response: {response.strip()}\n\n"
        f"Now provide the classification label:"
    )
```

Figure 4: Overview of the prompt components, their associated definitions and details, and the final prompt structure used in `LoMTL`.

precision (0.58), suggesting a more recall-oriented behavior. On the demonstration subset, `GPT-5` achieves the highest precision (0.60) and recall (0.61), whereas `LoMTL` remains competitive (0.58 precision, 0.56 recall) and substantially outperforms `Prometheus2`. Overall, these results show that LoMTL maintains a balanced precisionrecall trade-off, reinforcing its robustness across evaluation dimensions.

C.2 Human Annotation Analysis

We collected 95 annotations for single-tutor responses and 115 for pairwise comparisons. Since annotators were free to choose any dialogues and models, the number of annotations per tutor is not uniform. In pairwise comparisons, `Gemini` was selected most frequently – 49 out of 115 comparisons included `Gemini` as one of the tutors. Other tutors appeared in 17-36 comparisons, except for Novice, which was chosen only five times. Over-

Category	Parameter	Value
Common Settings		
Model	MODEL_NAME	`google/gemma-2-2b-it`
Task Dimensions	DIMENSIONS	Mistake_Identification, Mistake_Location, Providing_Guidance, Actionability
Input Length	MAX_LENGTH	1024
Prompt	include_label_definitions	Enabled
Training-Only Settings		
Batching	BATCH_SIZE	4
Batching	GRAD_ACCUM	1
Training Schedule	EPOCHS	3
Training Schedule	LEARNING_RATE	1e-4
Training Schedule	WEIGHT_DECAY	0.1
Logging	LOGGING_STEPS	50
Saving	SAVE_STEPS	300
Evaluation Cycle	EVAL_STEPS	300
Oversampling	OVERSAMPLE_METHOD	“random”
Metric for best model	METRIC_FOR_BEST	“eval_loss"
LoRA	LORA_R	8
LoRA	LORA_ALPHA	16
LoRA	LORA_DROPOUT	0.1
Early Stopping	EARLY_PATIENCE	5
Early Stopping	EARLY_THRESHOLD	0.0
Evaluation-Only Settings		
Generation	TEMPERATURE	1.0

Table 6: Summary of training and evaluation configurations along with their corresponding parameter names.

all, Expert responses were preferred most often, winning in 60% of the cases in which the Expert's response appeared. `Sonnet` and `Llama-3.1-405B` were also selected more than half of the time. Novice's responses never won.

In the single-tutor mode, 37 annotations marked responses as *helpful*, 31 as *not helpful*, and 27 as *to some extent helpful*. The most helpful responses came from Expert, `Sonnet`, `Gemini`, and `Mistral`, each of which was rated *helpful* in at least half of the corresponding cases. The least helpful responses were from `Phi3` and `Llama-3.1-8B`. `Mistral` and `Llama-3.1-405B` were the only tutors without any *not helpful* annotations, as they had the highest proportion of *to some extent helpful* ratings.

Model	Full Test Set				Demonstration Set			
	Accuracy	Macro-F1	Precision	Recall	Accuracy	Macro-F1	Precision	Recall
LoMTL (ours)	**0.72**	**0.60**	**0.63**	0.59	**0.68**	0.55	0.58	0.56
Prometheus2	0.47	0.41	0.44	0.45	0.41	0.34	0.38	0.36
GPT-5	0.66	0.58	0.58	**0.60**	0.66	**0.59**	**0.60**	**0.61**

Table 7: Accuracy, macro-F1, macro-precision and macro-recall scores (averaged across Mistake Identification (MI), Mistake Location (ML), Providing Guidance (PG), and Actionability (AC)) for our LoMTL model, Prometheus2, and GPT-5 on the full test set from Kochmar et al. (2025) and on the demonstration set. Best results are shown in **bold**.

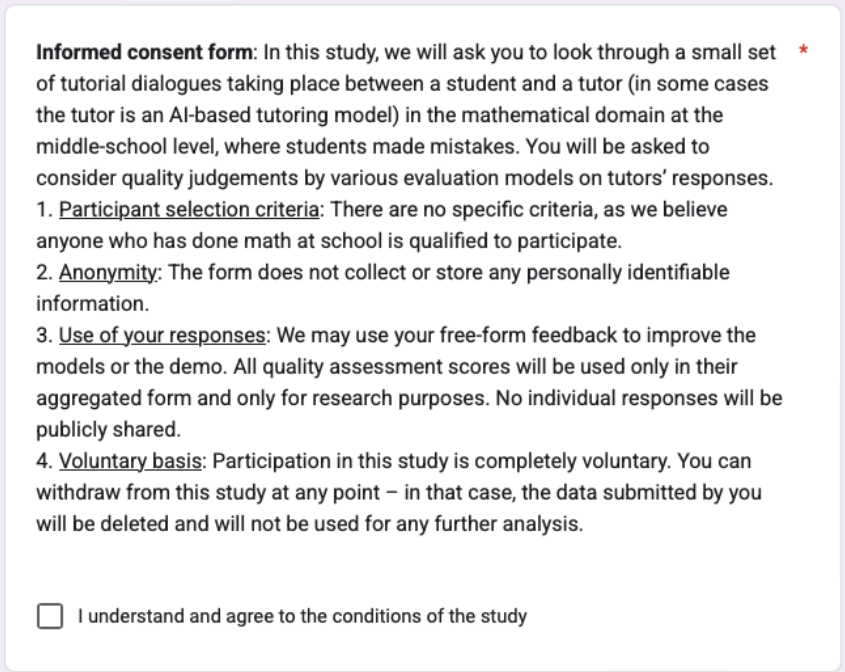

Figure 5: Informed consent form that participants were required to accept before proceeding with their feedback and annotations.

C.3 Full Questionnaire

Below are the informed consent forms that participants were required to read and accept, as well as the full questionnaire.

Background

Item	Question / Response Options
What is your highest qualification?	Response options: *Bachelor's degree; Master's degree; PhD degree; Other.*
Do you have teaching experience (e.g., lecturing, supervising or mentoring students, TA-ing, or similar)?	Response options: *Yes; No.*
Have you ever used an AI tutor before?	Response options: *Yes; No.*

Automated Evaluation

Item	Question / Response Options
Instructions	This tab gives you an opportunity to select among 10 short dialogues on relatively simple (no higher than middle-school level) math problems, check students' misconceptions, and explore various human and AI tutor responses. The quality of these responses is evaluated using a fine-tuned evaluation model.
Steps	1) Explore at least 5 different dialogues (the correct answer is also available to help you spot the student's mistake). 2) For each dialogue, check at least 2 different tutor responses. 3) Rate each response as *Helpful*, *Not Helpful*, or *To some extent*. 4) Select at least one quality dimension to view the model's assessment of that response. 5) For each dialogue, use the comparison mode at least once.
How many dialogues have you checked?	Free-form response.
How often did you agree with this model's assessment for a single tutor response?	Scale: 1 = almost never, 2 = less than half of the time, 3 = about half of the time, 4 = more than half of the time, 5 = almost always.
Response options: *1, 2, 3, 4, 5.*	
How often did you agree with this model's ranking of the tutors in the comparison mode?	Same scale as above.
Response options: *1, 2, 3, 4, 5.*	
On a scale from 1 to 5, how easy was it to use the "Automated Evaluation" tab?	Scale: 1 = not easy at all, 5 = very easy.
Response options: *1, 2, 3, 4, 5.*	
Please feel free to give us any further feedback on this tab.	Free-form response.

LLM Evaluation

Item	Question / Response Options
Instructions	This tab provides access to the same dialogues and tutor responses as in Tab 1, but this time they are evaluated using LLMs as judges. You can choose between GPT-5 and Prometheus.
Steps	1) Explore at least 5 different dialogues (they may be the same ones as before). 2) For each dialogue, check at least 2 different tutor responses. 3) Use each LLM-as-judge at least twice on different dialogues. 4) Select at least one quality dimension for each response. 5) For each dialogue, use the comparison mode at least once.
How many dialogues have you checked?	Free-form response.
How often did you agree with GPT-5's assessment for a single tutor response?	Scale: 1 = almost never, 2 = less than half of the time, 3 = about half of the time, 4 = more than half of the time, 5 = almost always.

Item	Question / Response Options
Response options: *1, 2, 3, 4, 5.*	
How often did you agree with GPT-5's ranking of the tutors in the comparison mode?	Same scale as above.
Response options: *1, 2, 3, 4, 5.*	
How often did you agree with Prometheus' assessment for a single tutor response?	Same scale as above.
Response options: *1, 2, 3, 4, 5.*	
How often did you agree with Prometheus' ranking of the tutors in the comparison mode?	Same scale as above.
Response options: *1, 2, 3, 4, 5.*	
Which evaluation model did you perceive to be more accurate in its judgments of tutor responses GPT-5 or the fine-tuned model from Tab 1 ("Automated Evaluation")?	Response options: *The model from Tab 1; GPT-5; Hard to say: they perform similarly.*
Which evaluation model did you perceive to be more accurate in its judgments of tutor responses Prometheus or the fine-tuned model from Tab 1 ("Automated Evaluation")?	Response options: *The model from Tab 1; Prometheus; Hard to say: they perform similarly.*
On a scale from 1 to 5, how easy was it to use the "LLM Evaluation" tab?	Scale: 1 = not easy at all, 5 = very easy.
Response options: *1, 2, 3, 4, 5.*	
Any other feedback is welcome.	Free-form response.

Visualizer

Item	Question / Response Options
Instructions	This tab visualizes statistics from the full dataset of tutorial dialogues and annotated tutor responses. The dataset contains 300 dialogues with responses from 9 tutors (except for the Novice tutor, who has annotations for 76 dialogues). Participants are asked to explore visualizations for at least 2 tutors.
On a scale from 1 to 5, how informative did you find these visualizations?	Scale: 1 = not informative at all, 5 = very informative.
Response options: *1, 2, 3, 4, 5.*	
On a scale from 1 to 5, how easy was it to use the "Visualizer" tab?	Scale: 1 = not easy at all, 5 = very easy.
Response options: *1, 2, 3, 4, 5.*	
We welcome any further feedback.	Free-form response.

D User Interface (UI) Details

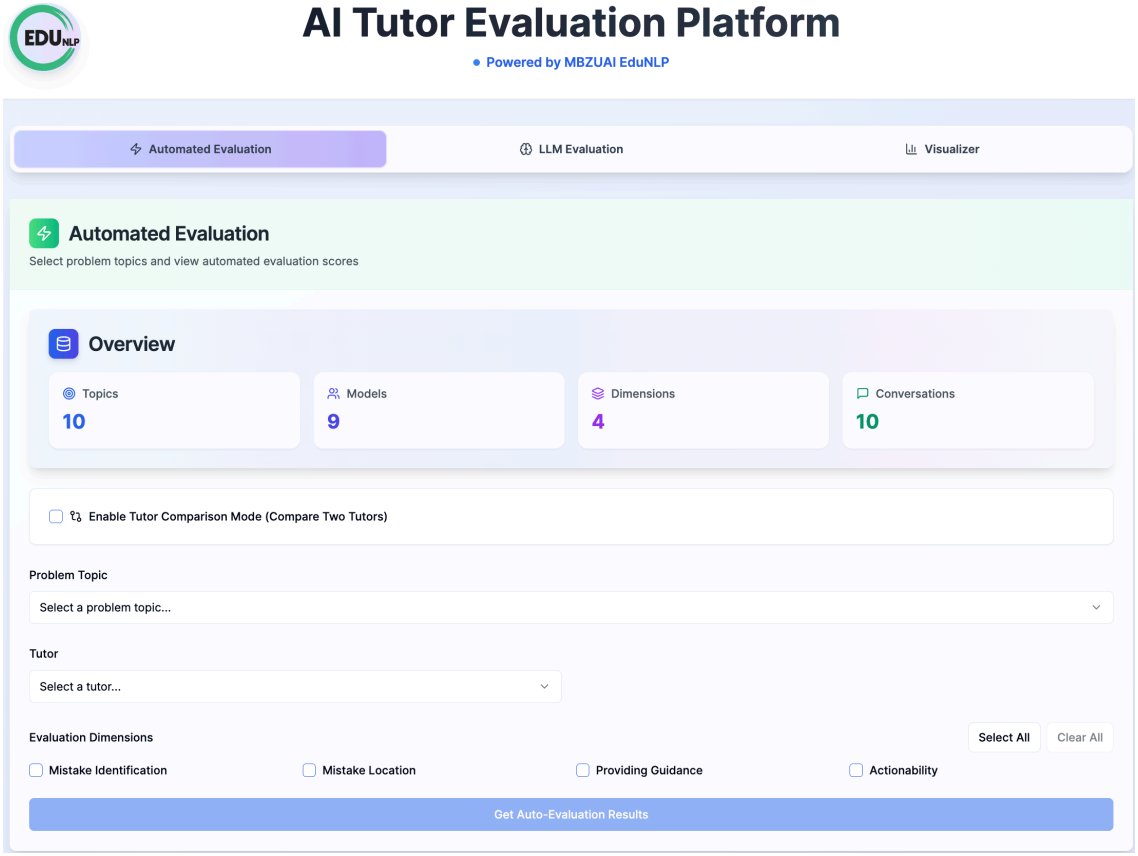

Figure 6: Overview of the UI and the Automated Evaluation Tab.

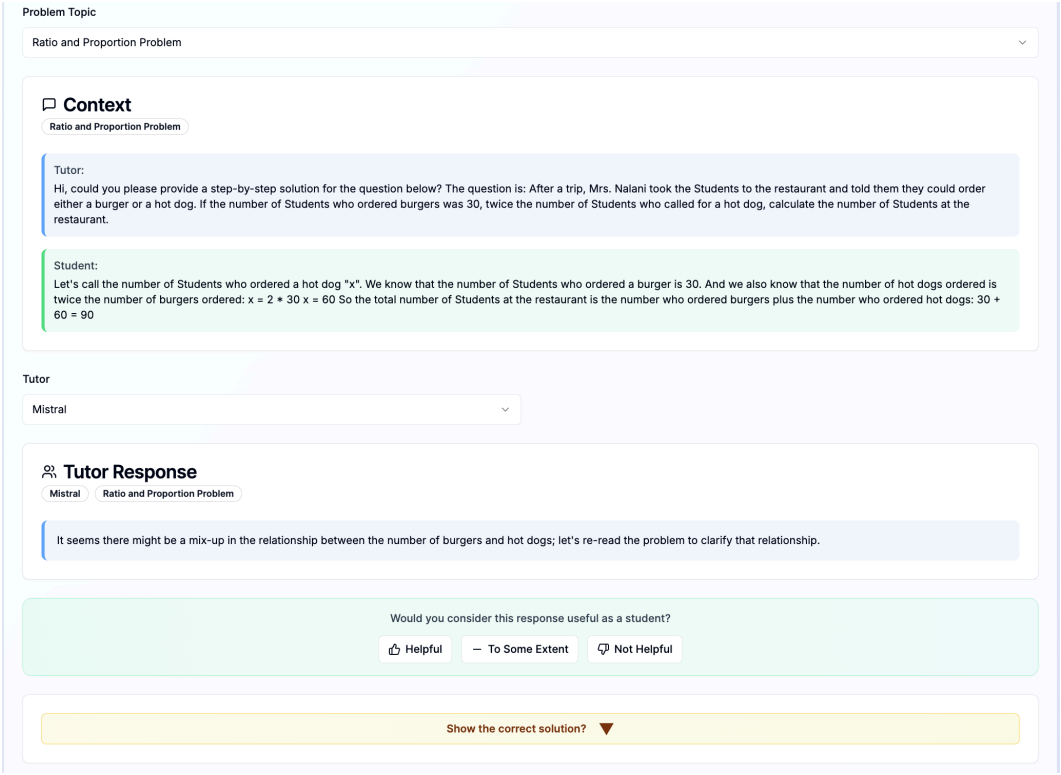

Figure 7: Interface showcasing the selected problem topic with Context, automated Tutor Response, student feedback options, and ground truth verification panel.

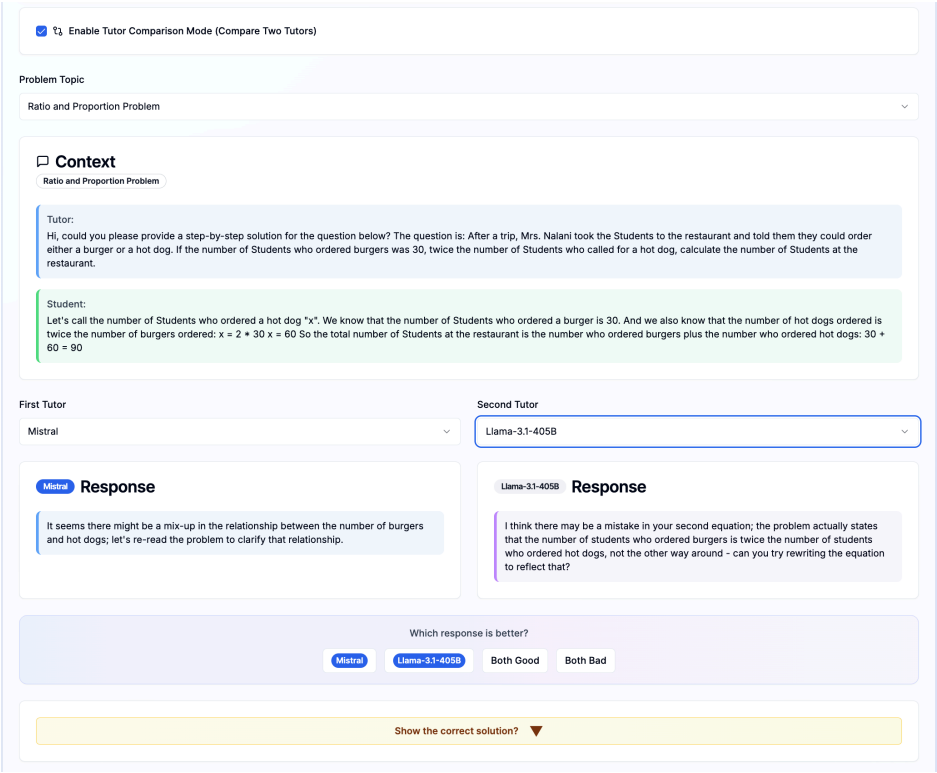

Figure 8: UI displaying the enabled Tutor Comparison Mode, allowing users to compare responses from any two selected tutors for the selected problem topic, along with the Context block, selected tutor responses, feedback options, and ground truth verification option.

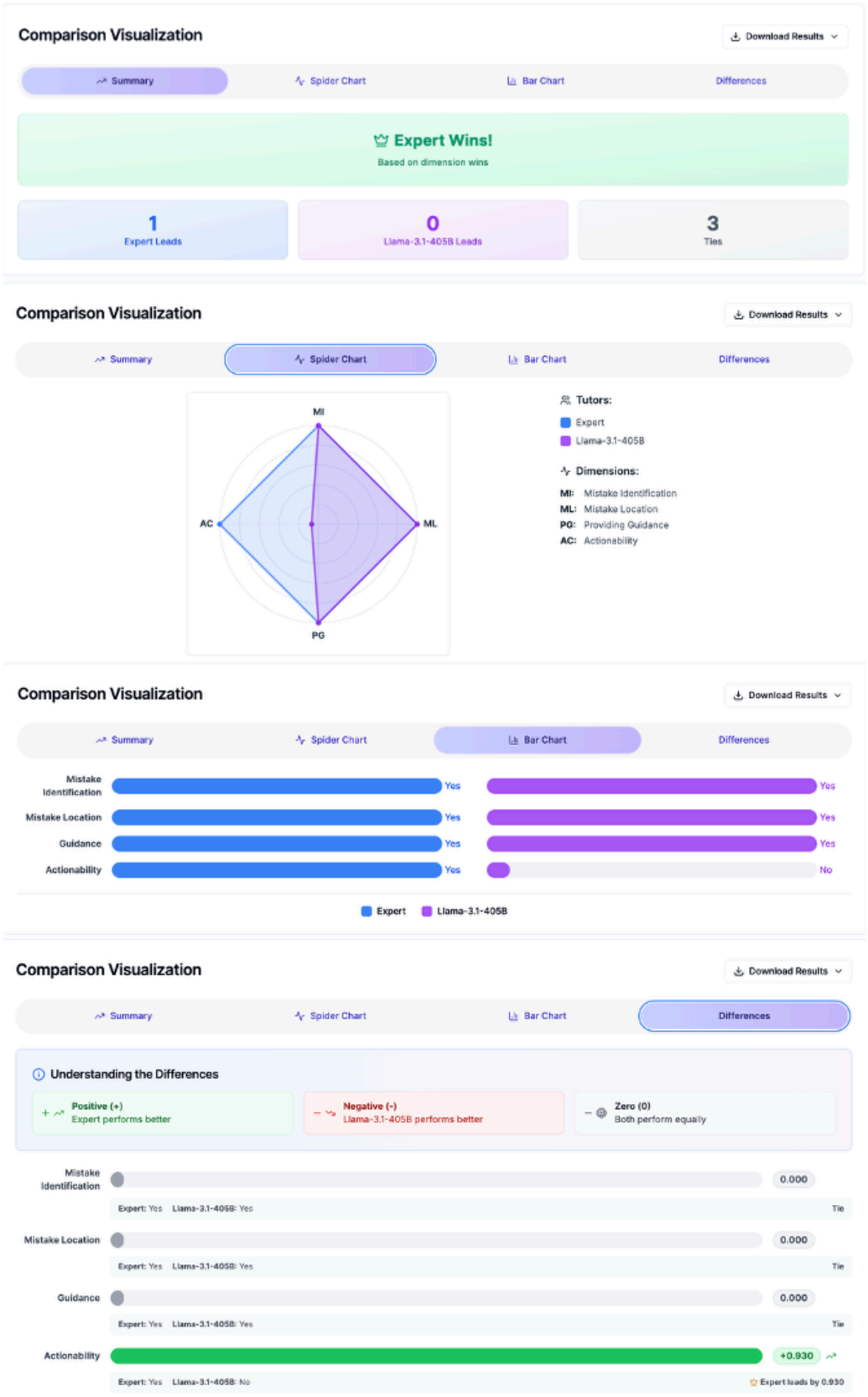

Figure 9: Displayed Tutor Comparison Visualization Panel showcasing Summary metrics, Spider Chart, Bar Chart, and Differences views for the chosen evaluation results.

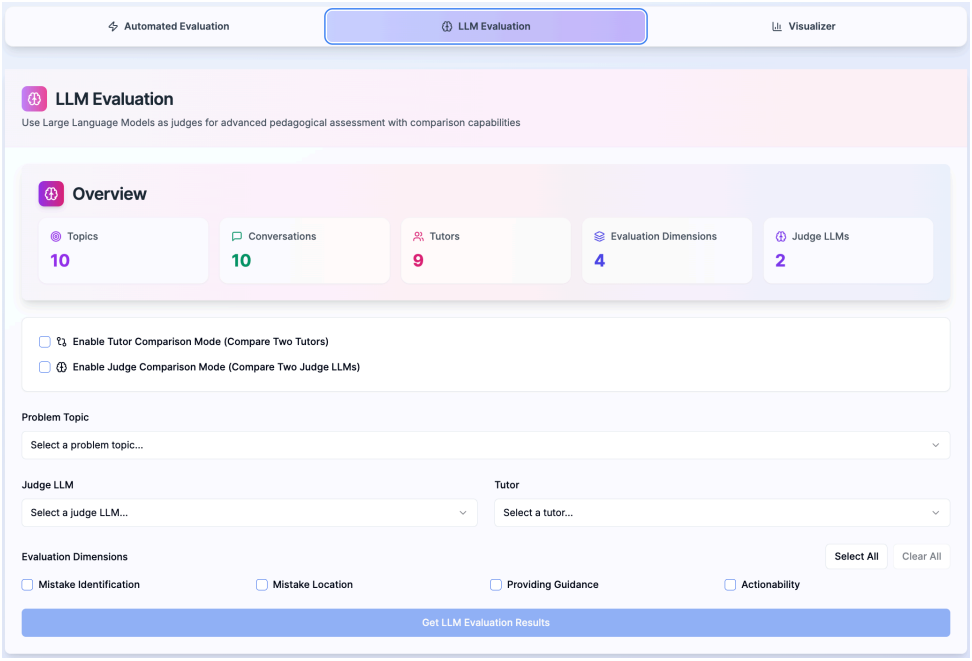

Figure 10: Overview of the LLM Evaluation module showcasing the dashboard panel with statistics on topics, conversations, tutors, and evaluation dimensions. The interface also highlights the provided options for Tutor Comparison Mode and Judge Comparison Mode for advanced pedagogical assessment.

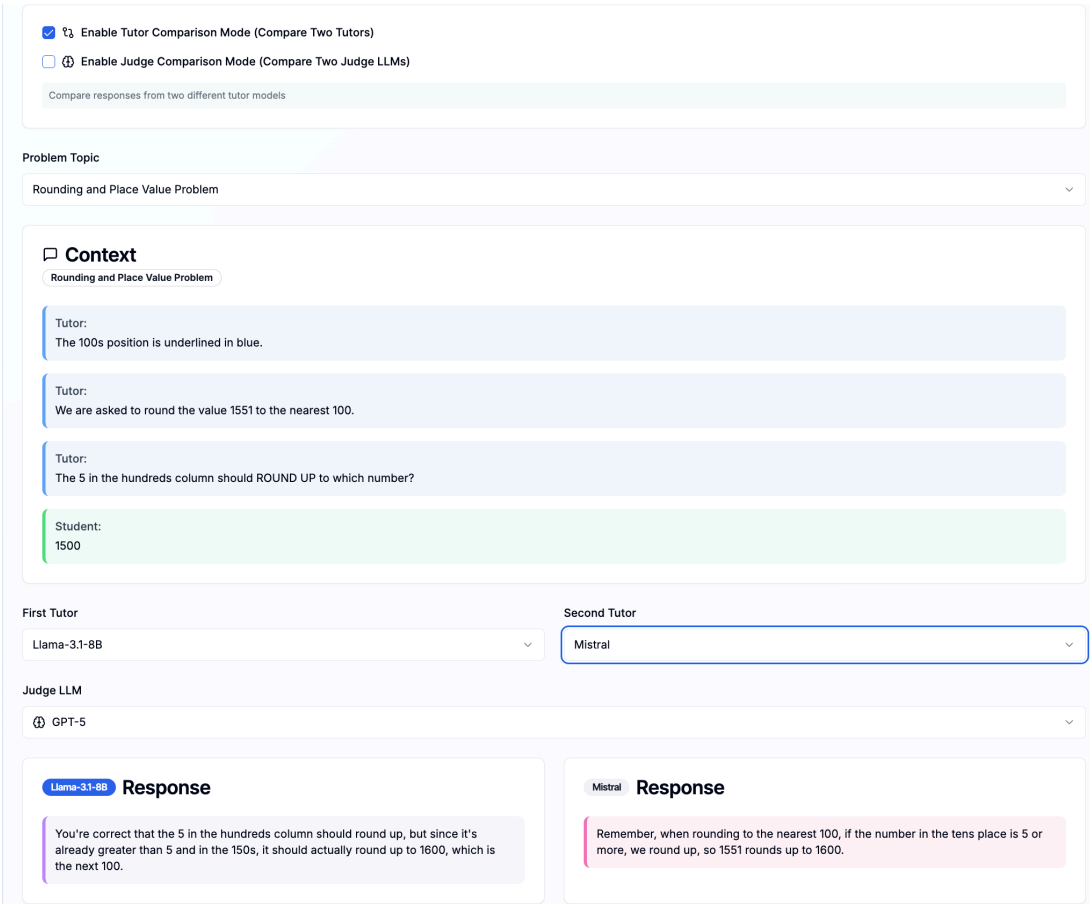

Figure 11: User Interface displaying the enabled Tutor Comparison Mode within the LLM Evaluation module, showing the selected problem, tutor responses from two tutors, and the judge LLM for evaluation.

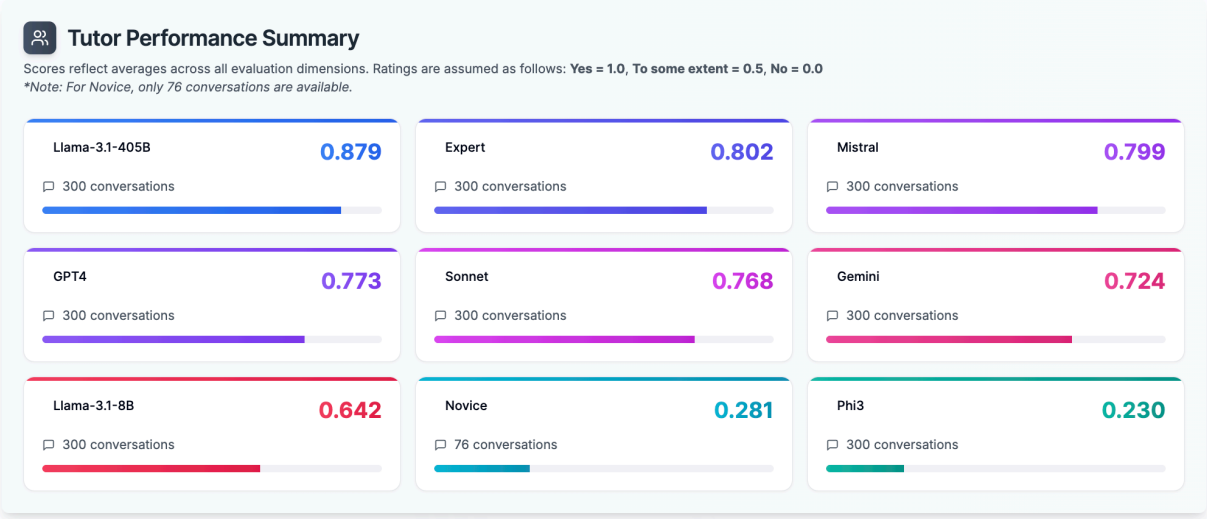

Figure 12: Tutor Performance Summary panel in the Visualizer module, displaying the aggregated evaluation scores for each tutor across all assessment dimensions within the MRBench development dataset.

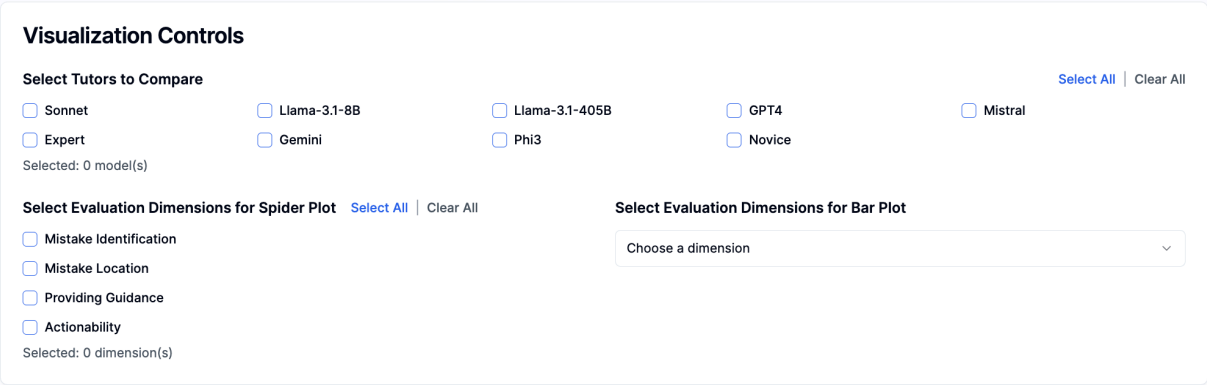

Figure 13: Interface of the Visualization Controls Panel, showing configurable options for selecting tutors and evaluation dimensions to generate comparative spider chart and bar plot visualizations.

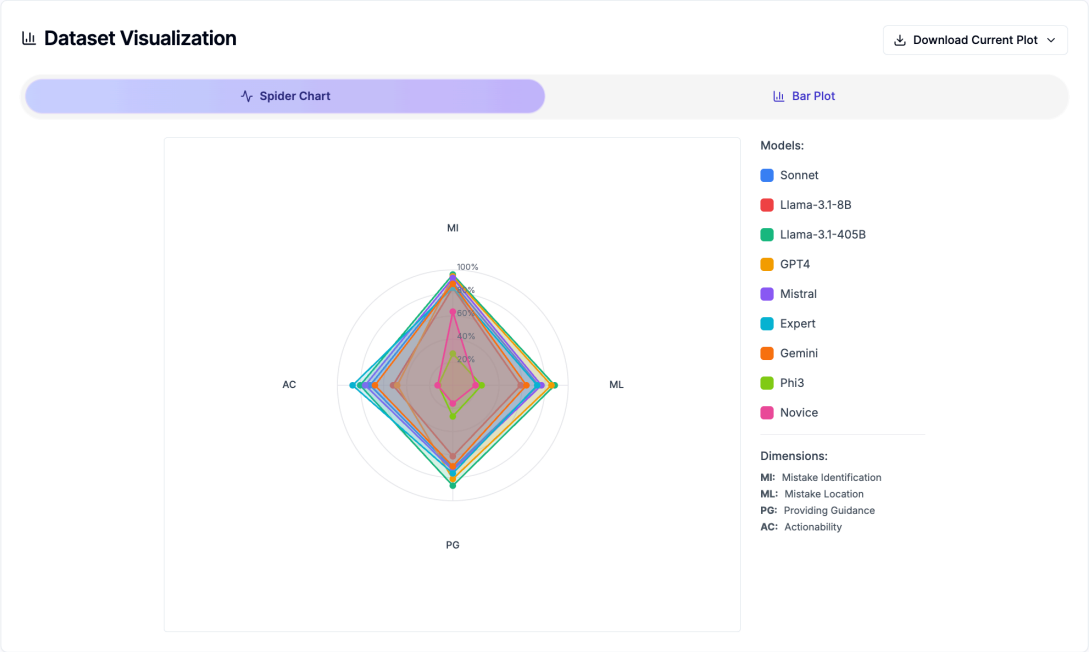

Figure 14: The spider chart representing tutors performance across selected evaluation dimensions, based on configurations chosen in the Visualization Controls Panel.

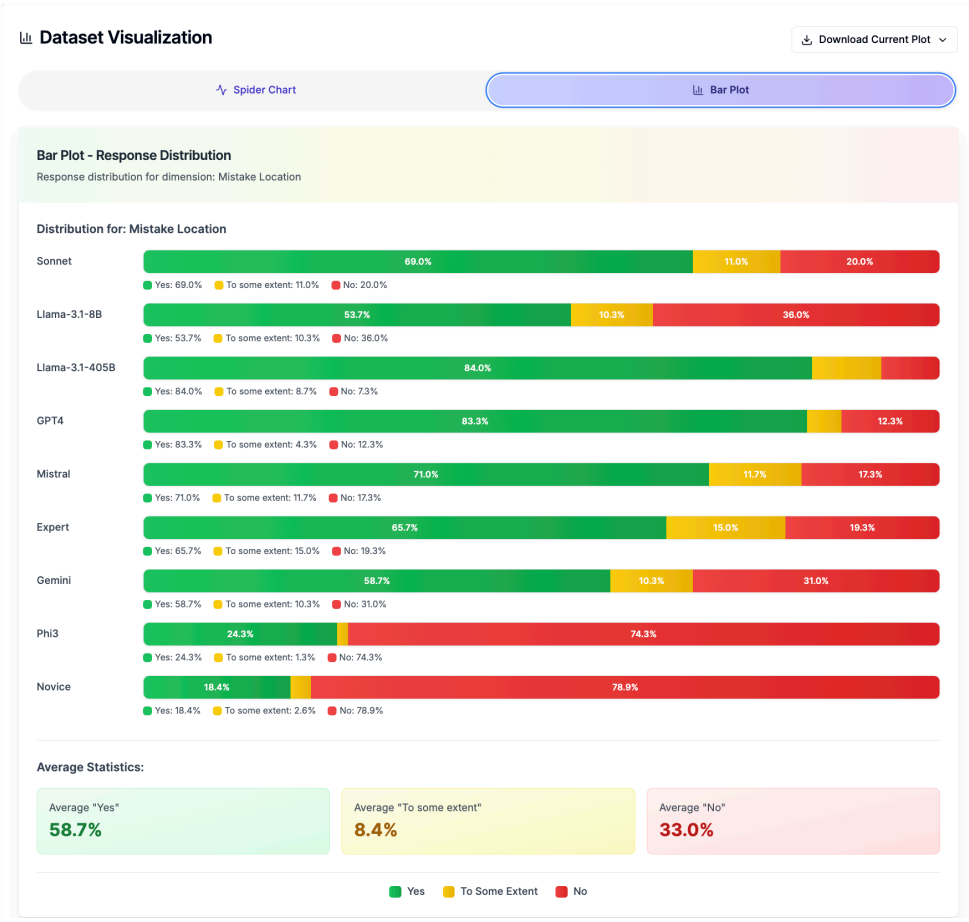

Figure 15: The bar plot representing tutors performance across selected evaluation dimension, based on configurations chosen from the Visualization Controls Panel.