bea-jh at BEA 2025 Shared Task: Evaluating AI-powered Tutors through Pedagogically-Informed Reasoning

Jihyeon Roh
Kakao
166, Pangyoyeok-ro, Bundang-gu, Seongnam-si, Gyeonggi-do, 13529, Korea
noa.h@kakaocorp.com

Jinhyun Bang*
Samsung Research
56 Seongchon-gil, Seocho-gu, Seoul, 06765, Korea
j_h.bang@samsung.com

Abstract

The growing use of large language models (LLMs) for AI-powered tutors in education highlights the need for reliable evaluation of their pedagogical abilities. In this work, we propose a reasoning-based evaluation methodology that leverages pedagogical domain knowledge to assess LLM-generated feedback in mathematical dialogues while providing insights into why a particular evaluation is given. We design structured prompts to invoke pedagogically-informed reasoning from LLMs and compare base model candidates selected for their strengths in reasoning, mathematics, and overall instruction-following. We employ Group Relative Policy Optimization (GRPO), a reinforcement learning method known to improve reasoning performance, to train models to perform evaluation in four pedagogically motivated dimensions, ***Mistake Identification***, ***Mistake Location***, ***Providing Guidance***, and ***Actionability***. Experimental results show that our GRPO-based models consistently outperform the base model and GPT-4.1, and surpass models trained using supervised fine-tuning in three out of four dimensions. Notably, our method achieved top-ranked performance in ***Actionability*** and competitive performance in two other dimensions in the BEA 2025 Shared Task under the team name `bea-jh`, underscoring the value of generating pedagogically grounded rationales for improving the quality of educational feedback evaluation.

1 Introduction

With the rapid development of large language models (LLM) and their text generation performance, research on employing LLM as an evaluation tool, or LLM-as-a-judge (Zheng et al., 2023), is actively being conducted. Specifically, LLMs have been adopted in evaluating overall quality (Gao et al., 2023), safety (Wang et al., 2024b), factual correctness, and fluency (Jain et al., 2023) of machine-generated texts. Furthermore, other works have applied similar methodologies to evaluate and revise texts from students (Bai and Stede, 2023; Awidi, 2024), and introduced artificial intelligence (AI) and LLMs into the field of education.

Although studies have shown that LLM-based feedback can enhance student motivation, evoke positive emotions (Meyer et al., 2024), and provide personalized learning experiences (Liu et al., 2025b), the question of how to evaluate the educational quality of such feedback remains open (Tack and Piech, 2022). Without rigorous evaluation, deploying LLM-based AI systems in education may expose students to biased content, overly simplistic pedagogical approaches (Angwaomaodoko, 2023), or confusing and unhelpful feedback (Denny et al., 2024). However, the educational AI market is rapidly expanding, with an estimated global value of 1.63 billion USD and a projected growth rate of over 30% within the next five years. (Grand View Research, 2025). This calls for the urgent need for LLM-generated student feedback evaluation, starting with defining the evaluation criteria.

Research on automated evaluation of machine-generated texts has provided some valuable guidance on the criteria, or dimensions, of what makes a good text, including consistency, relevance, fluency, and coherence (Jain et al., 2023; Liu et al., 2023; Lee et al., 2023). However, these dimensions are not sufficient when evaluating educational feedback as they fail to capture pedagogical values (Maurya et al., 2025), highlighting the need for domain-specific criteria.

Several studies have proposed pedagogical evaluation dimensions based on learning science principles (Tack and Piech, 2022; Macina et al., 2023; Wang et al., 2024a; Daheim et al., 2024). In this work, we focus on the problem of evaluating the pedagogical abilities of AI-powered tutors and propose an LLM-generated feedback evaluation frame-

*Corresponding author. Email: j_h.bang@samsung.com

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1049–1059

work based on the criteria defined by Maurya et al. (2025), which encompass dimensions proposed by previous approaches. We leverage the reasoning capabilities of LLMs, where the model generates not only answers but also the rationales behind them, for the following reasons. Firstly, reasoning can help improve the resulting performance, as the model can make use of its own reasons when generating the final output (Ke et al., 2025). In addition, reasoning can produce explainability via natural language feedback, which is highly important for AI systems adopted in education (Khosravi et al., 2022). We employ Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to improve LLM's reasoning performance (Guo et al., 2025).

Our contribution can be summarized as follows. Firstly, we introduce a state-of-the-art training methodology for producing explainable evaluations on LLM-generated feedback. Secondly, we provide system prompts, engineered based on pedagogical studies, that were used to train LLMs for evaluation. Our team, `bea-jh`, participated in four tracks of the BEA 2025 Shared Task (Kochmar et al., 2025). According to the official leaderboard of the shared task[1], we ranked 1st in Track ***Actionability***, 6th in Tracks ***Mistake Location*** and ***Providing Guidance***, and 13th in Track ***Mistake Identification*** on the shared task's main metric, strict macro-F1.

In the following section, related work, composed of previous approaches on machine-generated text evaluation, GRPO, and reward modeling, is introduced. In Section 3, we detail our system prompts, base model candidates, model selection rationale, and rewards mechanisms, where the effectiveness of the models resulting from the proposed approach is shown in Section 4. Finally, we conclude the paper in Section 5 together with future work.

2 Related Work

2.1 Machine-Generated Text Evaluation

Evaluating machine-generated text has been a central focus in natural language processing (NLP), with common approaches relying on dimensions such as fluency, coherence, consistency, and relevance (Liu et al., 2023; Kryściński et al., 2019). Frameworks such as UniEval (Zhong et al., 2022) provide evaluators for various natural language generation tasks—such as summarization and dialogue generation—by focusing on these core dimensions. However, these general-purpose metrics often fall short when applied to domain-specific texts, thus highlighting the need for more specialized evaluation frameworks.

Mathematical reasoning tasks require evaluation methods that assess not only the correctness of the final answer but also the stepwise logic and clarity of explanation. Benchmarks such as MATH[2] (Hendrycks et al., 2021), U-MATH[3] (Chernyshev et al., 2024), and GSM8K[4] (Cobbe et al., 2021) have emphasized the need for fine-grained evaluation of intermediate reasoning steps. Recent surveys (Lee and Hockenmaier, 2025) and methods such as ReasonEval (Xia et al., 2025) further underscore the importance of systematic evaluation of intermediary reasoning steps in mathematical problem solving.

In educational settings, machine-generated feedback should align with pedagogical principles, making its evaluation distinct from that of general text generation. Dimensions such as actionability, providing guidance, mistake identification, and mistake location are critical in determining the educational effectiveness of AI-generated feedback (Maurya et al., 2025). Other studies also emphasize additional aspects such as the tone (Han et al., 2024) and human-likeness (Wang et al., 2024a) of educational feedback.

2.2 Group Relative Policy Optimization

Reinforcement learning (RL) has played a central role in aligning large language models (LLMs) with human preferences. A widely adopted framework is Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017), which fine-tunes LLMs to produce outputs that are better aligned with human judgments (Ouyang et al., 2022). The standard RLHF pipeline consists of three stages: (1) training a reward model using human preference data, (2) generating outputs from the base model and scoring them with the reward model, and (3) fine-tuning the policy model via reinforcement learning, often using Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017). Despite its effectiveness, RLHF suffers from several well-known limitations. These include instability during training (Henderson et al., 2018), over-optimization of the reward model (reward hacking) (Casper et al., 2024), and sensitivity to biases in

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

[2] https://github.com/hendrycks/math/

[3] https://huggingface.co/datasets/toloka/u-math

[4] https://huggingface.co/datasets/openai/gsm8k

the human-labeled preference data (Barnhart et al., 2025).

To address these limitations, Group Relative Policy Optimization (GRPO) (Shao et al., 2024) has been proposed as an alternative reinforcement learning approach. Unlike traditional methods that rely on trained reward models, GRPO can leverage rule-based reward signals to guide optimization if correctness can be validated in an objective and deterministic fashion (Guo et al., 2025). GRPO has shown particular promise in reasoning-intensive tasks, such as mathematical problem solving (Shao et al., 2024).

As GRPO promotes the generation of coherent and interpretable reasoning chains, models can refer to their own rationales when generating the final output, thereby guiding themselves towards more reliable responses (Ke et al., 2025; Wei et al., 2022). Moreover, since the method does not explicitly train the reasoning traces, it enables models to produce novel rationales that can lead to improved performance (Guo et al., 2025). Such reasoning capabilities can also enhance the transparency of model decisions, offering better interpretability (Jie et al., 2024).

2.3 Prompt Engineering

Prompt engineering is the practice of strategically designing task-specific instructions as inputs to steer generative AI models towards producing desired outputs (Sahoo et al., 2024). Effective prompts typically incorporate clear instructions (Lo, 2023), contextual information (Yi et al., 2022), and relevant reference examples (Schick and Schütze, 2022). Incorporating domain-specific knowledge into prompts enhances LLM's ability to generate outputs that are not only accurate but also contextually appropriate, particularly in specialized fields (Marvin et al., 2023; Liu et al., 2025a), including education (Cain, 2024; Chen et al., 2024).

3 Methodology

3.1 Problem Definition

This work aims to evaluate feedback provided by AI-powered tutors, specifically LLMs, within the context of educational dialogues in mathematics. Traditional metrics used in dialogue systems are often inadequate for capturing pedagogical intent (Maurya et al., 2025), such as recognizing and locating students' misconceptions, guiding learning, and offering actionable feedback. To address this limitation, the 2025 BEA (Workshop on Innovative Use of NLP for Building Educational Applications) Shared Task[5] (Kochmar et al., 2025) proposes a benchmark for assessing tutor responses using a set of pedagogically motivated evaluation dimensions.

The evaluation focuses on four key abilities:

- ***Mistake Identification***: whether the tutor correctly identifies a student's mistake.
- ***Mistake Location***: whether the tutor correctly points out where in the student's response the mistake occurs.
- ***Providing Guidance***: whether the tutor offers helpful educational support such as hints or explanations.
- ***Actionability***: whether the tutor's feedback clearly indicates what the student should do next.

The development dataset consists of 300 multi-turn dialogues excerpted from two mathematics-focused datasets, MathDial (Macina et al., 2023) and Bridge (Wang et al., 2024a), where a mistake made by a student is included in every dialogue. Tutor responses from human and LLM sources are annotated across the four dimensions and categorized into three labels: "Yes", "To some extent", and "No". Accuracy and macro-F1 scores are used as core evaluation metrics under both strict and lenient evaluation settings, where the lenient setting merges "Yes" and "To some extent" as a single label.

3.2 Prompt Engineering

3.2.1 Prompt Design Principles

To effectively evaluate the pedagogical abilities of AI-powered tutors, we carefully designed the system prompts to encourage models to generate reasoning traces before producing final answers. Each prompt explicitly instructs the model to indicate its rationales by wrapping them between the following tag-like sequences: `<think>` and `</think>`, inspired by Deepseek-r1 (Guo et al., 2025). The prompt also instructs the model to wrap the final answer between `<answer>` and `</answer>` in a similar fashion. This structure facilitates the generation of coherent reasoning chains, and allows the final answer to be easily parsed and evaluated.

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

Moreover, each prompt includes an example illustrating the expected format of both rationale and answer. LLMs tend to respond better to the desired output format when shown examples following the specific format requirements (OpenAI, 2024). The following is an example excerpted from the prompt used for ***Mistake Identification***:

```
<think>The tutor response offers a follow-up
     question that directly targets the
    student's misunderstanding and
    encourages deeper thinking. The
    question is relevant and accurate,
    helping the student make progress.</
    think>
<answer>Yes</answer>
```

In addition to the format considerations, we emphasize the importance of incorporating domain-specific knowledge into the prompts (Marvin et al., 2023; Liu et al., 2025a; Cain, 2024; Chen et al., 2024). We embedded the details of the evaluation dimensions and corresponding labels into our system prompts. By doing so, we aim to focus the model's rationales on pedagogical assessment, rather than general linguistic assessment.

In the following sections, we describe in detail how the prompts were designed for each evaluation dimension.

3.2.2 Mistake Identification

Mistake Identification aims to evaluate whether a tutor has correctly captured the correctness of a student's solution. As the task of identifying the correctness of a mathematical solution is objective (Macina et al., 2025), we prompted the model to identify the student's mistake by itself before comparing its result with the given feedback. We also included the label descriptions provided by the shared task (Kochmar et al., 2025) to guide the model on where to draw the line between labels. Here is the corresponding segment excerpted from the prompt used for evaluating ***Mistake Identification***:

```
Step 1. Identify the student's mistake in <
    CONVERSATION_HISTORY>
Step 2. Assess whether <LAST_TUTOR_RESPONSE>
     **recognizes and identifies the
    student's mistake**. Use the criteria
    below:

### Evaluation Criteria:
- Yes: In <LAST_TUTOR_RESPONSE>, the mistake
     is clearly identified/recognized in
    the tutor's response.
- To some extent: <LAST_TUTOR_RESPONSE>
    suggests that there may be a mistake,
    but it sounds as if the tutor is not
    certain.
- No: In <LAST_TUTOR_RESPONSE>, the tutor
    does not recognize the mistake (e.g.,
    they proceed to simply provide the
    answer to the asked question).
```

3.2.3 Mistake Location

Mistake Location aims to evaluate whether a tutor accurately identifies where errors occur in a student's response. In designing the prompt, we incorporated the definition of this dimension, along with explanations on how locating mistakes correctly can support a student's learning process (Maurya et al., 2025). The following paragraphs are drawn from the prompt employed in the evaluation of ***Mistake Location***:

```
Your goal is to assess whether <
    LAST_TUTOR_RESPONSE> is **locating
    student's mistake**-that is, whether it
     not only notifies the student of the
    committed error, but also points to its
     location in the answer and outline
    what the error is to help student
    remediate it in their next response.

Use the following definitions:

- Yes: In <LAST_TUTOR_RESPONSE>, the tutor
    clearly points to the exact location of
     a genuine mistake in the student's
    solution.
- To some extent: <LAST_TUTOR_RESPONSE>
    demonstrates some awareness of the
    exact mistake, but is vague, unclear,
    or easy to misunderstand.
- No: <LAST_TUTOR_RESPONSE> does not provide
     any details related to the mistake.
```

3.2.4 Providing Guidance

Providing Guidance evaluates a tutor's ability to offer helpful guidance to students. Similar to ***Mistake Location***, we adopted the dimension descriptions from Maurya et al. (2025) as shown below:

```
Your goal is to assess whether <
    LAST_TUTOR_RESPONSE> is **providing
    guidance**-that is, whether it provides
     the student with relevant and helpful
    guidance, such as a hint, an
    explanation, a supporting question, and
     the like.

Use the following definitions:

- Yes: <LAST_TUTOR_RESPONSE> provides
    guidance that is correct and relevant
    to the student's mistake.
- To some extent: Guidance is provided in <
    LAST_TUTOR_RESPONSE> but it is fully or
```

```
     partially incorrect, incomplete, or
    somewhat misleading.
- No: <LAST_TUTOR_RESPONSE> does not include
     any guidance, or the guidance provided
     is irrelevant to the question or
    factually incorrect.
```

3.2.5 Actionability

Actionability aims to evaluate whether the tutor's feedback provides clear guidance on what students should do next, rather than simply giving away the answer. The description of the dimension from Maurya et al. (2025) was also incorporated in the prompt as shown below:

```
 Your goal is to assess whether <
     LAST_TUTOR_RESPONSE> is **actionable**-
     that is, whether it provides clear
     guidance on what the student should do
      next to improve or correct their work.


Use the following definitions:

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

Furthermore, we explicitly guided the model throughout the reasoning process using the following criteria and references. Feedback must be (1) useful and (2) clear, (3) make students want to receive further similar feedback (Broos et al., 2017), and (4) make students feel like they know what to do next (Maurya et al., 2025) for it to be actionable. Accordingly, the prompt is augmented with the following paragraph:

```
In your thinking process, imagine yourself
    being a student.
When you listen to the tutor's response
(1) Do you find this information useful?
(2) Do you find this information clear?
(3) After hearing this information, would
    you like to receive more of this type
    of information?
(4) Do you feel like you know what to do
    next?

Overall, a good feedback should be clear
    about what the student should do next,
    should not be vague, unclear or a
    conversation stopper.
```

3.3 Base Models

In this paper, we employ three open-source LLMs, GLM-4-9B, GLM-Z1-9B (Zeng et al., 2024), and Qwen2.5 14B Instruct (Yang et al., 2024b) as base model candidates. These models were selected as our candidates for their strengths, which will be described in the following subsections. Note that models with more than 14B parameters were excluded for faster iteration of experiments. Brief descriptions and strengths of the models are detailed in the following subsections. Performance of each base model and the selected model are presented in Subsection 4.2.

3.3.1 GLM-4-9B

GLM-4-9B (Zeng et al., 2024) is a powerful language model trained on over 10 trillion multilingual tokens. Its technical report shows that the model outperforms well-known foundation models, including Llama-3-8B (Grattafiori et al., 2024), in various tasks, including mathematical question answering. Specifically, the latest version released on April 14, 2025[6] is used in this work.

3.3.2 GLM-Z1-9B

GLM-Z1-9B (Zeng et al., 2024) is a reasoning model, which was trained on top of GLM-4-9B using reinforcement learning. The model was also further trained on datasets covering mathematics, code, and other logical domains. Specifically, it has demonstrated excellent capabilities in mathematical reasoning (THUDM, 2025). As with GLM-4-9B, we employed the latest version of GLM-Z1-9B released on April 14, 2025[7] as our candidate base model.

3.3.3 Qwen2.5 14B Instruct

Qwen2.5 14B Instruct (Yang et al., 2024b) is a powerful instruction-tuned model trained on 18 trillion tokens. Upon its release, it was reported to outperform other models of similar or even larger sizes (Qwen Team, 2024). Furthermore, compared to previously released Qwen2 (Yang et al., 2024a), Qwen2.5 demonstrated substantial improvements in mathematics and instruction-following capabilities (Yang et al., 2024b).

3.4 Reward Design

The reward or penalty terms used to train base models via GRPO in this work can be categorized

[6] https://huggingface.co/THUDM/GLM-4-9B-0414
[7] https://huggingface.co/THUDM/GLM-Z1-9B-0414

into two groups. The first group consists of penalty (negative reward) terms that encourage the model to generate outputs in the expected format. This group includes the following terms:

- Penalty for not generating a rationale.
- Penalty for not generating an answer.
- Penalty for generating neither a rationale nor an answer.
- Penalty for producing an unexpected answer (other than "Yes", "To some extent", and "No").

The relative value assigned to each penalty term was defined according to its importance. For example, the penalty term for missing a rationale is lower than that of missing an answer since the latter is critical for obtaining the final classification result.

The second group consists of reward and penalty terms that encourage the model to produce correct classification results, assuming the output is in the expected format. A positive reward is assigned when the model correctly predicts the target label. Since the evaluation metrics include those under a lenient setting, we also provide a smaller reward when the model confuses "Yes" and "To some extent," which are considered to be qualitatively similar. In contrast, the model receives a negative reward for any other incorrect predictions.

To help the model recognize the ordinal relationship among the labels, we conducted experiments in which a smaller penalty was applied for confusing "To some extent" with "No" than for confusing "Yes" with "No". However, this setup led the models to converge to a conservative solution, in which most examples were classified as "To some extent."

4 Experiments

4.1 Experiment Settings

We conducted all experiments using `trl` library[8] (von Werra et al., 2020) for training and `vllm` library[9] (Kwon et al., 2023) for serving a reference model for GRPO. We fine-tuned all models for 7 epochs using AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1e-5 and a cosine learning rate scheduler (Loshchilov and Hutter, 2022) with 128 examples in each training step. Since the test dataset is not open to public and submission attempts were limited, reported results are obtained using either the official test set or the evaluation set split from the development set. Details on the dataset used in each experiment are provided in the caption of each table.

[8] https://github.com/huggingface/trl

[9] https://github.com/vllm-project/vllm

4.2 Experiment Results

4.2.1 Base Model Selection

For the selection of the base model, we randomly selected a task to compare the performance of the candidate base models. Table 1 presents the results of the base models on the selected task, ***Mistake Location***. GLM-4-9B was selected as our base model as it outperformed other two candidates. Note that the subpar performance of GLM-4-Z1-9B was primarily due to its failure to follow the required formatting guidelines—such as generating labels outside the set "Yes", "No", "To some extent", or omitting the final decision altogether.

We further compared the performance scores of the models trained on top of GLM-4-9B and Qwen2.5 14B Instruct on another randomly sampled task, ***Actionability***, to examine the base model's generalizability. As shown in Table 2, the model fine-tuned from GLM-4-9B outperformed that from Qwen2.5 14B Instruct in three out of four metrics, and demonstrated a comparable level of strict macro-F1 score.

4.2.2 Group Relative Policy Optimization and Reasoning

To examine the effectiveness of the proposed approach, we compared our method with a recently released state-of-the-art proprietary LLM, GPT 4.1 (OpenAI, 2025), released on April 14, 2025. We further compared our approach with conventional supervised fine-tuning (SFT) without rationale outputs. The results are shown in Table 3.

In ***Actionability***, our GRPO-trained model outperforms all other baselines, including GPT-4.1 and conventional SFT-based model, achieving the best scores in all four metrics. In ***Providing Guidance***, our method also achieves the best macro-F1 and accuracy in the lenient setting and the best accuracy in the strict setting, and shows competitive performance in strict macro-F1 as well. A similar trend is observed for ***Mistake Location***, where the proposed method achieves the best strict macro-F1, lenient macro-F1, and accuracy. However, the model trained with conventional SFT performs strongly in ***Mistake Identification***, calling for the need of

Base model	Strict		Lenient	
	Macro-F1	Accuracy	Macro-F1	Accuracy
GLM-4-9B	**0.273**	**0.380**	**0.384**	**0.566**
GLM-Z1-9B	0.095	0.133	0.131	0.162
Qwen2.5 14B Instruct	0.194	0.232	0.338	0.443

Table 1: Initial performance of base model candidates on ***Mistake Location***, obtained on the entire development set. Best score for each metric is marked in **bold**.

Base model	Strict		Lenient	
	Macro-F1	Accuracy	Macro-F1	Accuracy
GLM-4-9B	0.701	**0.756**	**0.861**	**0.888**
Qwen2.5 14B Instruct	**0.709**	0.730	0.853	0.884

Table 2: Performance of different base models on ***Actionability***, obtained on the official test set. Best score for each metric is marked in **bold**.

Methods	***Mistake Identification***	***Mistake Location***
GPT 4.1	0.410 / 0.528 / 0.699 / 0.806	0.342 / 0.355 / 0.639 / 0.673
Base model	0.393 / 0.548 / 0.634 / 0.746	0.390 / 0.468 / 0.582 / 0.641
SFT	**0.715** / **0.899** / **0.900** / **0.952**	0.481 / **0.726** / 0.757 / 0.819
GRPO (ours)	0.564 / 0.867 / 0.805 / 0.919	**0.569** / 0.669 / **0.768** / **0.823**
Methods	***Providing Guidance***	***Actionability***
GPT 4.1	0.532 / 0.613 / 0.704 / 0.790	0.567 / 0.581 / 0.827 / 0.847
Base model	0.409 / 0.516 / 0.583 / 0.738	0.417 / 0.440 / 0.697 / 0.710
SFT	**0.593** / 0.617 / 0.731 / 0.815	0.542 / 0.661 / 0.730 / 0.738
GRPO (ours)	0.571 / **0.649** / **0.764** / **0.859**	**0.664** / **0.758** / **0.854** / **0.875**

Table 3: Performance of different models, obtained on the evaluation set split from the development set. Each cell is composed of strict macro-F1 / accuracy / lenient macro-f1 / accuracy scores. Best score for each metric is marked in **bold**.

further investigation on different characteristics of each dimension. Overall, GRPO-based models consistently outperform the base model and GPT 4.1 across all dimensions, while achieving better performance than SFT-based models in three dimensions, indicating that training a model to produce pedagogically-informed rationales contributes to better evaluation performance.

4.3 Shared Task Leaderboard

The resulting models from experiments, which were submitted under the team name of `bea-jh`, demonstrated strong performance compared to other 2025 BEA Shared Task contestants (Kochmar et al., 2025). Our model ranked 1st in ***Actionability***, and 6th in ***Mistake Location*** and ***Providing Guidance*** in the shared task's official main metric, strict macro-F1 scores. Models trained on top of both GLM-4-9B and Qwen 2.5 14B Instruct achieved better performance than those of other contestants, demonstrating the generalizability of the effectiveness of the proposed prompting and training strategy.

On the other hand, in ***Mistake Identification***, the SFT-based model ranked 13th while the GRPO-based model would have ranked 37th out of 44 contestants. Aforementioned results are summarized in Table 4 along with scores and rankings on the shared task's secondary metrics.

5 Conclusion

In this paper, we proposed a methodology for evaluating the pedagogical abilities of AI-powered tutors in providing helpful feedback across four key dimensions. System prompts were designed to incorporate pedagogical domain knowledge and the base model was selected based on its initial performance to generate rationale-supported evaluation. The selected model was then trained with the system prompts using Group Relative Policy Optimization (GRPO), a state-of-the-art method

Metric		***Mistake Identification***	***Mistake Location***	***Providing Guidance***	***Actionability***
Strict Macro-F1	Score	0.5873 (0.6802*)	0.5658	0.5451	0.7010 (0.7085†)
	Ranking	37 / 44 (13 / 44*)	6 / 32	6 / 36	1 / 30 (1 / 30†)
Strict Accuracy	Score	0.8449 (0.8707*)	0.7389	0.6703	0.7557
	Ranking	28 / 44 (9 / 44*)	4 / 32	4 / 36	1 / 30
Lenient Macro-F1	Score	0.8494 (0.9069*)	0.7851	0.7324	0.8609
	Ranking	32 / 44 (6 / 44*)	5 / 32	10 / 36	4 / 30
Lenient Accuracy	Score	0.9270 (0.9457*)	0.8268	0.8003	0.8875
	Ranking	28 / 44 (11 / 44*)	5 / 32	7 / 36	4 / 30

Table 4: Final scores and rankings on the official test set. Scores are shown to four decimal places, following the official leaderboard format. Along with the scores and rankings for ***Mistake Location*** and ***Actionability*** obtained by evaluating the proposed approach on the official test dataset and reranked based on the official leaderboard, officially recorded scores and rankings under the team name `bea-jh` that are not obtained by the proposed approach are presented inside the parentheses. $*$: score and ranking in the official ***Mistake Identification*** leaderboard, obtained by training GLM-4-9B via supervised fine-tuning. †: macro-F1 score and ranking in the official ***Actionability*** leaderboard, obtained by training Qwen2.5 14B Instruct via our proposed approach.

for optimizing reasoning capabilities in LLMs. As a result, our models demonstrated competitive performance in the BEA 2025 Shared Task, achieving the first place in the ***Actionability*** dimension.

However, the proposed approach exhibits varying performance across different evaluation dimensions. These discrepancies suggest that each dimension may require tailored modeling strategies that reflect its underlying pedagogical definitions. Future work could involve an in-depth pedagogy-based analysis of each dimension to identify how to design a high-quality evaluator. Furthermore, since our approach generates explicit rationales through reasoning, these rationales could potentially be leveraged not only for evaluation, but as a basis for improving the AI tutor's feedback.

Limitations

We believe this study proposes an effective methodology for evaluating the pedagogical abilities of AI-powered tutors. However, the following limitations highlight areas for future investigation.

Thorough investigation of prompt engineering Since system prompts serve as instructions to LLMs, variations in prompt design can lead to different outputs and rationales. While our prompts incorporated pedagogical domain knowledge, further investigation into how each component of a prompt influences the reasoning process could lead to more effective prompt engineering strategies for evaluation tasks.

Analysis of dimension-specific characteristics Although the proposed method achieved strong performance in certain evaluation dimensions, it performed relatively poorly in a particular dimension. This discrepancy may stem from intrinsic differences among dimensions, such as varying levels of subjectivity or difficulty. Analyzing why the method performs better in some dimensions could open the way to the development of evaluation strategies tailored to each dimension.

Analysis of rationale truthfulness Generating rationales provides insights into how models "think", and the rationales generated by an evaluation model may inspire ways to improve the systems under evaluation. However, it remains an open question whether these rationales truly reflect the model's internal decision-making process. Future work could involve further analysis to assess the stability and truthfulness of generated rationales, enabling a more qualitative understanding of reasoning-based evaluation.

References

Ejuchegahi Anthony Angwaomaodoko. 2023. The re-examination of the dangers and implications of artificial intelligence for the future of scholarship and learning. *Traektoriâ Nauki*, 9(10):3021–3028.

Isaiah T Awidi. 2024. Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (ai) tool. *Computers and Education: Artificial Intelligence*, 6:100226.

Xiaoyu Bai and Manfred Stede. 2023. A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. *International Jour-*

nal of Artificial Intelligence in Education, 33(4):992–1030.

Logan Barnhart, Reza Akbarian Bafghi, Stephen Becker, and Maziar Raissi. 2025. Aligning to what? limits to rlhf based alignment. *arXiv preprint arXiv:2503.09025*.

Tom Broos, Laurie Peeters, Katrien Verbert, Carolien Van Soom, Greet Langie, and Tinne De Laet. 2017. Dashboard for actionable feedback on learning skills: Scalability and usefulness. In *Learning and Collaboration Technologies. Technology in Education: 4th International Conference, LCT 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part II 4*, pages 229–241. Springer.

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

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, and 1 others. 2024. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*.

Eason Chen, Danyang Wang, Luyi Xu, Chen Cao, Xiao Fang, and Jionghao Lin. 2024. A systematic review on prompt engineering in large language models for k-12 stem education. *arXiv preprint arXiv:2410.11123*.

Konstantin Chernyshev, Vitaliy Polshkov, Ekaterina Artemova, Alex Myasnikov, Vlad Stepanov, Alexei Miasnikov, and Sergei Tilga. 2024. U-math: A university-level benchmark for evaluating mathematical skills in llms. *arXiv preprint arXiv:2412.03205*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. Stepwise verification and remediation of student reasoning errors with large language model tutors. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8386–8411.

Paul Denny, Stephen MacNeil, Jaromir Savelka, Leo Porter, and Andrew Luxton-Reilly. 2024. Desirable characteristics for ai teaching assistants in programming education. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, pages 408–414. ACM.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Grand View Research. 2025. AI Tutors Market Size, Share & Trends | Industry Report 2030 — grandviewresearch.com. https://www.grandviewresearch.com/industry-analysis/ai-tutors-market-report. [Accessed 19-04-2025].

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and 1 others. 2024. Llm-as-a-tutor in efl writing education: Focusing on evaluation of student-llm interaction. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 284–293.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8487–8495.

Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. 2024. How interpretable are reasoning explanations from prompting large language models? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2148–2164.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, and 1 others. 2025. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*.

Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi-Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. 2022. Explainable artificial intelligence in education. *Computers and education: artificial intelligence*, 3:100074.

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

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jinu Lee and Julia Hockenmaier. 2025. Evaluating step-by-step reasoning traces: A survey. *arXiv preprint arXiv:2502.12289*.

SeungJun Lee, Taemin Lee, Jeongwoo Lee, Yoonna Jang, and Heuiseok Lim. 2023. Kullm: Learning to construct korean instruction-following large language models. In *Annual Conference on Human and Language Technology*, pages 196–202. Human and Language Technology.

Hongxuan Liu, Haoyu Yin, Zhiyao Luo, and Xiaonan Wang. 2025a. Integrating chemistry knowledge in large language models via prompt engineering. *Synthetic and Systems Biotechnology*, 10(1):23–38.

Jiayi Liu, Bo Jiang, and Yu'ang Wei. 2025b. Llms as promising personalized teaching assistants: How do they ease teaching work? *ECNU Review of Education*, page 20965311241305138.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Leo S Lo. 2023. The clear path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, 49(4):102720.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2022. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621.

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. *arXiv preprint arXiv:2502.18940*.

Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2023. Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*, pages 387–402. Springer.

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.

Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199.

OpenAI. 2024. Best practices for prompt engineering with the openai api.

OpenAI. 2025. Introducing gpt-4.1 in the api | openai.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Qwen Team. 2024. Qwen2.5: A party of foundation models!

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *CoRR*.

Timo Schick and Hinrich Schütze. 2022. True few-shot learning with prompts—a real-world perspective. *Transactions of the Association for Computational Linguistics*, 10:716–731.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Anaïs Tack and Chris Piech. 2022. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues. *International Educational Data Mining Society*.

THUDM. 2025. Thudm/glm-4: Glm-4 series: Open multilingual multimodal chat lms: .

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

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.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2024b. Do-not-answer: Evaluating safeguards in llms. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2025. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27723–27730.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Jingjie Yi, Deqing Yang, Siyu Yuan, Kaiyan Cao, Zhiyao Zhang, and Yanghua Xiao. 2022. Contextual information and commonsense based prompt for emotion recognition in conversation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 707–723. Springer.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *CoRR*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038.