K-NLPers at BEA 2025 Shared Task: Evaluating the Quality of AI Tutor Responses with GPT-4.1

Geon Park[*1], **Jiwoo Song**[*2], **Gihyeon Choi**[*2], **Juoh Sun**[*2] **and Harksoo Kim**[1,2],

[1]Department of Computer Science and Engineering, Konkuk University
[2]Department of Artificial Intelligence, Konkuk University

Correspondence: nlpdrkim@konkuk.ac.kr

Abstract

This paper presents automatic evaluation systems for assessing the pedagogical capabilities of LLM-based AI tutors. Drawing from a shared task, our systems specifically target four key dimensions of tutor responses: Mistake Identification, Mistake Location, Providing Guidance, and Actionability. These dimensions capture the educational quality of responses from multiple perspectives, including the ability to detect student mistakes, accurately identify error locations, provide effective instructional guidance, and offer actionable feedback. We propose GPT-4.1-based automatic evaluation systems, leveraging their strong capabilities in comprehending diverse linguistic expressions and complex conversational contexts to address the detailed evaluation criteria across these dimensions. Our systems were quantitatively evaluated based on the official criteria of each track. In the Mistake Location track, our evaluation systems achieved an Exact macro F1 score of 58.80% (ranked in the top 3), and in the Providing Guidance track, they achieved 56.06% (ranked in the top 5). While the systems showed mid-range performance in the remaining tracks, the overall results demonstrate that our proposed automatic evaluation systems can effectively assess the quality of tutor responses, highlighting their potential for evaluating AI tutor effectiveness.

1 Introduction

Recent advancements in Large Language Models (LLMs) have significantly enhanced performance across various tasks in natural language processing and artificial intelligence (Kim et al., 2025, 2024; Das et al., 2025). These developments have spurred interest in applying LLMs within educational settings, aiming to leverage their capabilities for personalized learning, intelligent tutoring, and educational assessment (Macina et al., 2023; Chevalier et al., 2024; Wang et al., 2024b; Gan et al., 2023). However, despite these promising developments, how well LLMs can provide educational feedback and guidance in authentic tutoring scenarios remains underexplored. To address this gap, there is a growing need for systematic evaluation methods that can rigorously assess the pedagogical quality of LLM-generated tutor responses. To address this need, we participated in the BEA 2025 Shared Task on Pedagogical Ability Assessment of AI Tutors (Kochmar et al., 2025), which aims to systematically evaluate the educational quality of AI tutor responses across multiple dimensions. Our system was submitted to four subtasks (Tracks 1–4), each corresponding to the pedagogical evaluation dimensions defined in the shared task: Mistake Identification, Mistake Location, Providing Guidance, and Actionability.

For this study, we employed prompting techniques using GPT-4.1[1] model to complete each evaluation task. GPT-4.1 is well-known for its superior ability in instruction-following, handling complex contexts, and performing multi-step reasoning (OpenAI, 2025). These capabilities, combined with our prompting strategies, enabled effective evaluation of tutoring performance. This paper details the prompting strategies and methodologies utilized in the evaluation tracks in which we participated. Additionally, we analyze and discuss the performance of our proposed systems, aiming to provide practical insights for future development of LLM-based tutoring systems.

2 Methodology

This section introduces the three prompting-based evaluation systems we developed for the BEA 2025 Shared Task. Each system is designed to align with the pedagogical goals of different evaluation tracks, while sharing a common objective of simulating human-like reasoning in tutoring scenarios.

[*] These authors contributed equally to this work.

[1] gpt-4.1-2025-04-14

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1145–1163
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

Section 2.1 presents the Chain-of-Thought-based Pedagogical Evaluation system with Reasoning Layers. This approach models the step-by-step reasoning process of a human tutor, who first analyzes the student's thinking, constructs a correct solution, and then evaluates the tutor's response in context. This system was applied to **Track 1**, which focuses on evaluating whether the tutor successfully identifies student mistakes, and **Track 4**, which assesses the actionability of the feedback provided.

Section 2.2 describes the Multi-Perspective Reflective Evaluation, developed for **Track 2**. Inspired by the reflective feedback behavior of human tutors, this method simulates internal deliberation among distinct reasoning perspectives to assess whether the tutor accurately identifies the location of a student's mistake.

Section 2.3 details the Rubric-Based Evaluation Method, which targets **Track 3**. This approach decomposes the Providing Guidance criterion into multiple rubric-based sub-questions. It extracts structured features from LLM-generated probability distributions and enhances scoring consistency by using a downstream classifier trained to align model judgments with human evaluation patterns.

These methodologies establish a comprehensive framework for evaluating tutor responses, enhancing interpretability, pedagogical alignment, and educational validity in open-ended dialogue settings.

2.1 Chain-of-Thought-based Pedagogical Evaluation with Reasoning layers

This system is designed to automatically evaluate the pedagogical appropriateness of a tutor's final utterance in a math lesson dialogue. Instead of using a single prompt or a simple classification-based approach, we adopted a step-by-step processing structure that emulates how human tutors interpret student solutions and determine appropriate feedback. The design of this structure is based on two key observations: First, large language models (LLMs) show improved performance on complex problems when explicitly guided through intermediate reasoning steps, a technique known as chain-of-thought prompting (Wei et al., 2022). Second, LLMs tend to exhibit conformity bias—favoring only a single "standard" solution path and struggling to respond appropriately to diverse or alternative reasoning strategies (Li et al., 2024).

To address these issues, we designed the flow so that the model first analyzes the student's reasoning process, then generates a correct solution path based on that reasoning, and finally evaluates the tutor's utterance in light of the student's thinking. All stages are implemented using the GPT-4.1 model, which was selected for its strong performance in instruction following, conversational context retention, and multi-step reasoning. Our proposed system is composed of the following four stages:

1. **Problem Extraction**: Extract the math problem from the dialogue. The extracted problem serves as the foundation for all subsequent reasoning, and functions as a critical preprocessing step for maintaining contextual coherence and semantic consistency.

2. **Student Reasoning Process Reconstruction**: Based on the student's response and the flow of the conversation, reconstruct the reasoning path that the student followed to solve the problem. Even in the absence of explicit explanations, infer a plausible line of reasoning. This mirrors how a human tutor might infer a student's thought process in real instructional settings to provide targeted feedback.

3. **Correct Reasoning Process Generation**: Using the reconstructed student reasoning as a foundation, generate a correct solution path. If the student's approach is partially valid, it is preserved and only the errors are corrected. If the approach is fundamentally flawed, a new solution is generated. This stage serves both as a reference point for comparison and as a mechanism to mitigate the conformity bias described earlier.

4. **Tutor Response Evaluation**: Finally, the tutor's final utterance is evaluated using the following four criteria:

    - Mistake Identification
    - Mistake Location
    - Providing Guidance
    - Actionability

These criteria, while based on the definitions provided by the task organizers, are redefined in our approach to focus on how utterances actually function within the student's learning process, moving beyond simple sentence-level evaluation. **Mistake Identification** is judged not merely on whether a mistake was mentioned, but on whether this recognition was perceptible and significant to the student. **Mistake Location** is assessed not by whether the error's position is explicitly stated, but by whether the student can reasonably infer where the mistake occurred based on the tutor's response. **Providing Guidance** is assessed not by the mere provision of a correct answer, but by whether it was a method that stimulated and broadened student thinking. **Actionability** uses as its criterion whether the student can actually understand and follow the guidance, rather than the mere presence or absence of a suggested action.

These redefinitions allow the LLM to evaluate the tutor not as a mere provider of correct answers (tutor-as-answerer), but as a facilitator of reasoning and learning (tutor-as-guide). To ensure consistent scoring across levels, especially for nuanced categories like "To some extent", concrete judgment criteria were clearly designed. The specific prompts corresponding to each criterion, along with detailed evaluation guidelines, are provided in detail in the Appendix A.

2.2 Multi-Perspective Reflective Evaluation Method

To accurately determine whether a tutor's response correctly identifies the location of a mistake in a student's solution, our system proposes a multi-perspective reasoning process inspired by how human tutors approach student feedback. Rather than relying on static classification, this system simulates a dynamic reasoning process, decomposing the evaluation into distinct functional perspectives such as recalling relevant context, analyzing logic, assessing clarity, and monitoring emotional tone.

2.2.1 Human-Like Multi-Step Reasoning

When human tutors assess a student's response, they typically do not evaluate it in a single step. Instead, they engage in a layered cognitive process: understanding the problem, reconstructing the student's reasoning, identifying discrepancies, and delivering feedback that balances correctness and pedagogical clarity. One recent attempt to emulate this human-like multi-step reasoning within a single LLM is **S**olo **P**erformance **P**rompting (SPP), which activates diverse personas to facilitate self-collaboration and mimic human reasoning (Wang et al., 2024c). Building on this idea, our system mirrors such behavior by simulating a group of internal "reasoning participants", each representing a specific evaluative function. These participants collaborate iteratively to reach a decision regarding the quality of the tutor's feedback.

2.2.2 Reasoning Process

Given a conversation history, including the original question, the student's response, and the tutor's follow-up, the system performs the following steps:

- **Perspective Initialization**: Depending on the complexity of the student's reasoning and the characteristics of the tutor's feedback, a set of internal perspectives is dynamically activated. These perspectives represent distinct reasoning roles (e.g., logical analysis, memory retrieval, contextual interpretation).
- **Independent Assessment**: Each perspective independently analyzes whether the tutor's response points to the specific step where the mistake occurred. The analysis includes not only factual correctness but also the interpretability and relevance of the feedback.
- **Collaborative Deliberation**: After the initial assessments, the perspectives engage in a multi-turn collaborative discussion. They provide critical feedback on one another's reasoning, refine interpretations, and critique or support conclusions.
- **Final Decision**: Based on this internal collaboration, the system synthesizes a final judgment: "Yes", "To some extent", or "No", depending on how clearly and precisely the tutor's response identifies the location of the student's mistake.

2.2.3 Prompting Strategy

We implement the above reasoning process through a carefully designed prompt that guides the language model to simulate human-like evaluation. Rather than instructing the model to directly respond to a tutor's utterance, the prompt breaks the evaluation into distinct reasoning roles. It encourages the model to adopt multiple perspectives. The prompt explicitly instructs the model to initiate internal reflection by assigning roles such as logical analysis, memory recall, and clarity evaluation. It then simulates a collaborative discussion where these roles critique and refine one another's views before converging on a final judgment. This structured interaction is carried out entirely within a single language model, enabling it to reason through the task in a self-contained yet multi-faceted manner. By prompting the model to consider both explicit and implicit forms of feedback, as well as emotional tone and pedagogical clarity, this design elicits more interpretable and human-aligned judgments. It ensures the model reflects on why a tutor response is effective or not, rather than simply what label to assign.

2.3 Rubric Based Evaluation Method

In this section, we aim to evaluate whether tutor LLMs provide correct and relevant guidance within the context of tutoring dialogue. We apply a method that predicts high-dimensional judgments through item-specific probability distributions, such as those used in LLM-Rubric (Hashemi et al., 2024), to assess the educational validity of tutor LLM responses. Specifically, for the prediction of Providing Guidance, we designed five detailed questions (Q_{rubric}) and a single comprehensive question ($Q_{overall}$) utilizing statistical information. For each item, we constructed prompts such that the LLM outputs the probabilities of "Yes", "To some extent", "No". However, the evaluation labels generated by the LLM may not completely align with the labels of human evaluators. Therefore, we use the item-wise probability distributions as input features for a subsequent classifier, aiming to calibrate the LLM's judgments to be more consistent with human evaluation.

2.3.1 Feature Extraction via Structured Prompting

The feature extraction step based on structured prompting consists of two components: rubric-based evaluation criteria and statistical information. The prompts used for each task are presented in Appendix B, and the responses were generated using the GPT-4.1 model.

Feature Extraction from Rubric-Based Evaluation Criteria The prompt for feature extraction based on rubric-defined evaluation items consists of role specification, presentation of dialogue context, definition of label criteria and output format, and a list of evaluation questions.

- **Role specification**: By assigning the expert role of "expert evaluator analyzing a tutor's response in a learning dialogue," the model is encouraged to think critically from the perspective of an evaluator rather than as a simple generator.

- **Presentation of dialogue context**: The dialogue context is presented sequentially and consists of the entire conversation between the tutor and student, the student's last utterance, and the tutor's response to that utterance. This allows the LLM to conduct evaluations based on a sufficient understanding of the context.

- **Definition of label criteria and output format**: For each item, the judgment consists of three options: Yes, To some extent, and No. The definitions of these labels are based on criteria defined by the annotator. For each item, the model outputs the probability value for each of the three labels in decimal form, based on the rationale for its judgment. The sum of all probability values is designed to be 1.0, and these values are used as input features for the subsequent classifier.

- **List of questions**: The prompt includes five questions designed to capture various aspects of the Providing Guidance criterion. Each question is constructed to evaluate specific elements of detailed feedback, as follows.

  Q1 Did the tutor attempt to provide any explanation, hint, or example?

  Q2 Was the guidance factually correct and appropriate given the student's error?

  Q3 Did the tutor's response directly address the student's specific mistake?

  Q4 Did the guidance help the student figure out what to do next, without directly giving the final answer?

  Q5 Was the tutor's response clear and unlikely to confuse the student?

This prompt design enables the LLM to consistently perform structured evaluations. The extracted features serve as inputs for subsequent classifiers, thereby enhancing the precision and reliability of the automated evaluation framework.

Feature Extraction Using Statistical Information

Antecedent	Consequent	Support	Confidence	Lift
ML = Yes	MI = Yes	0.6163	0.9889	1.2674
MI = Yes	ML = Yes	0.6163	0.7898	1.2674
PG = Yes	MI = Yes	0.5399	0.9502	1.2178

Table 1: Results of Association Rule Analysis among Mistake Identification, Mistake Location, and Providing Guidance

In the feature extraction step utilizing statistical information, features were constructed based on association rules among items analyzed from the development dataset. To this end, the Apriori algorithm (Agrawal and Srikant, 1994) was applied using the label information of the three items: Mistake Identification, Mistake Location, and Providing Guidance. Based on the calculated support and confidence values, the three most reliable association rules were extracted. The main association rules derived from this analysis are shown in Table 1. The top three association rules all exhibit high confidence values, suggesting that the relationships between items possess meaningful associations beyond mere coincidence. Among these, the relationship between "Mistake Location = Yes" and "Mistake Identification = Yes" demonstrates particularly high confidence, confirming a strong association between the two items.

To reflect these statistically significant associations, the following elements were added within the same prompt structure used in previous tasks:

- **Insertion of prior prediction results**: Predictions from previous Tracks (Mistake Identification and Mistake Location) were included in the prompt, allowing the LLM to perform evaluations based on this prior knowledge.

- **Provision of statistical associations**: Confidence values derived from association analysis were explicitly presented in the prompt to numerically illustrate conditional relationships among the three items. This allows the model to reference the likelihood of specific judgments influencing others during response evaluation.

- **Presentation of a single comprehensive question**: A question designed to elicit an overall assessment of "Providing Guidance" was included. The model was prompted to holistically assess whether the response attempted meaningful guidance, based on the given content, prior Track predictions, and statistical information.

Through this prompt, the model can extract comprehensive judgment features for Providing Guidance by simultaneously considering both existing prediction results and quantitative association information.

2.3.2 Improving Consistency in LLM Evaluation Using Classifiers

To calibrate the evaluation results of LLM responses with human assessors' judgments, we adopted an approach utilizing a subsequent classifier and conducted experiments to select an optimal classification model. The features extracted in Section 2.3.1 consist of probability values for the three categories—“Yes,” “To some extent,” and “No”—for each of the six sub-questions(Q_{rubric} and $Q_{overall}$). Each sub-question is represented as a three-dimensional vector (i.e., three probabilities), and concatenating these yields an 18-dimensional real-valued vector (6 questions × 3 classes = 18), which serves as the input feature for the response quality classifier. We compared the performance of three classification models — Random Forest (Breiman, 2001), Logistic Regression (Cox, 1958), and XGBoost (Chen and Guestrin, 2016) — using 5-fold cross-validation on the development dataset.

Classifier	Gold		Pred		w/o $Q_{overall}$	
	F1	Acc	F1	Acc	F1	Acc
Logistic Regression	0.61	0.71	0.49	0.66	0.49	0.66
Random Forest	0.63	0.70	0.55	0.61	0.54	0.60
XGBoost	0.61	0.69	0.52	0.66	0.50	0.64

Table 2: Comparison of Classifier Performance According to the Use of $Q_{overall}$. **Gold** denotes that ground-truth labels for $Q_{overall}$ were supplied, whereas **Pred** uses the values predicted by the methods in Sections 2.1 and 2.2.

Table 2 compares classifier performance across three experimental conditions. The first condition inputs gold labels for Mistake Identification and Mistake Location into $Q_{overall}$. The second condition uses predicted values from previous Tracks for $Q_{overall}$, while the third entirely excludes $Q_{overall}$ from input features to analyze its performance impact. Under the gold-label configuration for $Q_{overall}$, Random Forest achieved the highest Macro F1 score of 0.63. This indicates strong alignment between $Q_{overall}$ and the final Providing Guidance label, representing the upper performance bound of the proposed methodology. In the simulated test environment using predicted values for $Q_{overall}$, Random Forest maintained superior performance with a Macro F1-score of 0.55, though lower than the gold-label scenario. This performance gap underscores the influence of prediction uncertainty in $Q_{overall}$ and highlights its critical role in overall accuracy. Experiments excluding $Q_{overall}$ resulted in performance degradation across all models, demonstrating that $Q_{overall}$ facilitates comprehensive judgment rather than isolated item assessment, thereby making substantial contributions to classifier efficacy.

Based on these findings, the study implemented a system incorporating all Q_{rubric} and $Q_{overall}$ items as input features, with Random Forest selected as the final classifier for Predicting Providing Guidance labels. This configuration optimizes robustness while maintaining practical applicability in automated feedback evaluation.

3 Evaluation

In this section, we report and discuss the evaluation results obtained from each of the prompting methodologies applied to Tracks 1 through 4. The performance of each proposed method is presented briefly, highlighting their strengths and identifying areas for improvement.

3.1 Evaluation Metrics

Our evaluation followed the same metrics defined by the shared task organizers. Specifically, accuracy and macro F1 scores were utilized as the primary metrics for evaluating performance across Tracks 1 through 4. These metrics were computed under two distinct settings:

- **Exact evaluation (Ex.)**: Predictions were assessed based on the precise classification into three distinct categories ("Yes," "To some extent," and "No").

- **Lenient evaluation (Len.)**: Considering the qualitative similarities between responses annotated as "Yes" and "To some extent," these two classes were combined into a single category ("Yes + To some extent"), resulting in a simplified binary classification ("Yes + To some extent" vs. "No") for performance evaluation.

3.2 Dataset

We conducted our experiments using the dataset provided by the shared task organizers. The dataset consists of 300 dialogues extracted from the MathDial (Macina et al., 2023) and Bridge (Wang et al., 2024a) datasets, and includes a total of 2,476 tutor responses annotated for four pedagogical aspects based on the scheme proposed by Maurya et al. (2025). These annotated responses were used as the development set. An additional 1,547 tutor responses, constructed in the same manner, were used as the test set.

3.3 Chain-of-Thought-based Pedagogical Evaluation System with Reasoning layers

To evaluate the effectiveness of the proposed assessment system in section 2.1, we conducted experiments on two models: GPT-4.1 and GPT-4.1-mini[2]. The experiments were performed on the entire development set. Since the proposed system does not require a separate training phase, all examples in the dataset were directly used for evaluation. To facilitate comparative analysis of the proposed system's performance, we also conducted experiments using an alternative baseline prompt (see Appendix A for details of the baseline prompt), defined by the following conditions:

- The input consists only of the dialogue history and the tutor's final utterance.

- For each evaluation criterion, the original definitions provided by the task organizers were used, rather than the redefined versions proposed in this study.

This setup allows us to directly compare how variations in prompt design and evaluation criteria definitions affect final performance, under identical language model and dataset conditions.

Task	Prompt	Model	Ex. F1	Ex. Acc	Len. F1	Len. Acc
MI	Base	GPT 4.1 mini	0.5566	0.6975	0.8037	0.8958
		GPT 4.1	0.5850	0.7383	0.8107	0.9055
	Ours	GPT 4.1 mini	0.5699	0.7282	0.7965	0.9079
		GPT 4.1	**0.6225**	**0.7993**	**0.8371**	**0.9204**
ML	Base	GPT 4.1 mini	0.5037	0.5856	0.7447	0.7928
		GPT 4.1	0.4642	0.4851	0.7361	0.8029
	Ours	GPT 4.1 mini	0.4885	0.5166	**0.7581**	0.8146
		GPT 4.1	**0.5238**	**0.5969**	0.7564	**0.8154**
PG	Base	GPT 4.1 mini	0.5286	**0.5428**	0.7347	0.8247
		GPT 4.1	0.4905	0.4758	0.7374	0.8320
	Ours	GPT 4.1 mini	0.5117	0.4956	0.7506	**0.8389**
		GPT 4.1	**0.5398**	0.5355	**0.7583**	0.8384
ACT	Base	GPT 4.1 mini	0.4934	0.5141	0.6889	0.7597
		GPT 4.1	0.4487	0.4378	0.6975	0.7815
	Ours	GPT 4.1 mini	0.5045	0.5250	0.7129	0.7851
		GPT 4.1	**0.5210**	**0.5384**	**0.7253**	**0.7948**

Table 3: Performance comparison across tasks, prompts, and models. **Bold** indicates the best performance within each task, and underline indicates the second-best.

Table 3 presents a performance comparison between the proposed evaluation system and the baseline prompt. The proposed approach demonstrates overall superior results across all four evaluation criteria compared to the base prompt. Notably, when using the GPT-4.1 model, improvements in response quality were observed under both exact and lenient evaluation metrics.

For Mistake Identification, which measures the model's ability to recognize student errors, the proposed system proved more effective in producing

[2] gpt-4.1-mini-2025-04-14

clear and convincing judgments.

For Mistake Location, which assesses how well the tutor's response pinpoints where the student made a mistake, the proposed system also showed better performance when using GPT-4.1. Although the performance gains were more limited with the smaller model (GPT-4.1-mini), the proposed system helped generate responses with more consistent error localization patterns.

For Providing Guidance, which evaluates whether the tutor's response offers not only correct answers but also instructional support — such as explanations, hints, or examples — the proposed system was more effective in assessing responses using this criterion, as it successfully identified a variety of instructional strategies, including explanations, hints, and guiding questions. This indicates that the redefined evaluation criteria were more closely aligned with authentic pedagogical practices.

For Actionability, which assesses whether the student can clearly understand what to do next based on the tutor's feedback, the proposed system demonstrated consistently high performance in evaluating responses that effectively prompted concrete next steps. This result likely stems from the prompt structure and evaluation criteria, which were explicitly designed to reflect a student-centered communicative framework.

Taken together, these results demonstrate that even without fine-tuning, the combination of a structured prompt chain, evaluation criteria redefined from the student's perspective, and a reasoning-guided process can enhance both the reliability and pedagogical validity of tutor response evaluation. However, the experimental findings also imply that distinguishing fine-grained judgment boundaries—such as between "Yes", "To some extent", and "No"—remains a challenge. This highlights the limitation of relying solely on prompt-based inference, as the model may still struggle to fully grasp the nuanced intent behind each evaluation category without task-specific training.

Despite these limitations, we applied the proposed system to the official evaluations of Track 1 and Track 4 without any additional training, in order to see whether it would perform reliably in a real evaluation setting. As a result, the system maintained stable performance on the test set, achieving Exact macro F1 scores of 0.6669 and 0.5664 for Track 1 and Track 4, respectively, thereby demonstrating that the performance observed on the development set was consistently replicated in the official evaluation.

3.4 Multi-Perspective Reflective Evaluation System

We submitted our system, developed under the team name K-NLPers, to the Mistake Location track of the BEA Shared Task. It was built upon our proposed multi-perspective reasoning framework and evaluated using the GPT-4.1 model.

Team	Ex. F1	Ex. Acc	Len. F1	Len. Acc
BLCU-ICALL	**0.5983**	**0.7679**	0.8386	**0.8630**
BJTU	0.5940	0.7330	0.7848	0.8261
K-NLPers	0.5880	0.7641	**0.8404**	0.8610
MSA	0.5743	0.6975	0.7848	0.8209
SG	0.5692	0.7602	0.8118	0.8416

Table 4: Evaluation Results on the Mistake Location Track under Multi-Perspective Reflective Evaluation. **Bold** indicates the best performance and underline indicates the second-best.

As shown in Table 4, our system achieved competitive results, ranking 3rd overall among participating teams. In particular, it showed strong performance in Exact Accuracy (0.7641), Lenient macro F1 (0.8404), and Lenient Accuracy (0.8610), with scores closely comparable to those of the top two teams. These results suggest that our system produces predictions with consistent structure and high lexical accuracy, demonstrating that the proposed approach can effectively compete with state-of-the-art systems. However, the Exact macro F1 score (0.5880) was slightly lower than that of the top-ranked teams, primarily due to difficulty in distinguishing responses labeled as "To Some Extent". Despite this, the results confirm that our system is robust and generalizable, yielding strong overall performance across evaluation metrics in a competitive setting.

3.5 Rubric Based Evaluation System

This section evaluated the Providing Guidance dimension (Track 3) using the rubric-based system proposed in Section 2.3.

Team	Ex. F1	Ex. Acc	Len. F1	Len. Acc
BLCU-ICALL	0.5741	0.6716	0.7487	0.8061
BJTU	0.5725	0.6490	0.7445	0.8100
K-NLPers	0.5606	0.6270	0.7446	0.8003
bea-jh	0.5451	0.6387	0.7253	0.7977

Table 5: Performance and ranking of our models in predicting "Providing Guidance" on the test set.

Table 5 presents the leaderboard results on the test set for the final system proposed in Section 2.3. Our system achieved an Exact macro F1 score of 0.5606 and a Lenient macro F1 score of 0.7446 on the test set. Despite employing a straightforward approach that relies solely on prompt-based probability distribution outputs and a post-processing classifier, the system demonstrates the capability to secure a satisfactory level of precision and consistency in real-world settings. Notably, attaining an Exact macro F1 score of 0.5606—a stringent evaluation criterion—indicates that the structured multi-dimensional features derived from the rubric-based items Q_{rubric} and $Q_{overall}$ effectively capture the educational validity of tutor responses.

These findings suggest that the prompt-based multi-dimensional judgment methodology not only generates responses but also effectively aligns the evaluation and classification of responses with human raters. Furthermore, the methodology maintains a certain degree of generalization performance even on inputs that were not seen during training, thereby illustrating that evaluations leveraging large language models can function as assessments with genuine educational validity.

4 Conclusion

This study proposed a set of prompting-based automatic evaluation methods to assess the pedagogical quality of AI tutor responses across four key dimensions: Mistake Identification, Mistake Location, Providing Guidance, and Actionability. Leveraging the capabilities of GPT-4.1, the methods were designed to emulate human-like reasoning through chain-of-thought prompting, multi-perspective reflection, and rubric-based probability estimation, aligning large language model outputs with authentic educational feedback standards.

Our approaches demonstrated competitive performance in the BEA 2025 Shared Task across multiple evaluation tracks. The Multi-Perspective Reflective Evaluation showed strong performance in Mistake Location, while the Rubric-Based Evaluation validated the effectiveness of structured feature extraction and post-classification for nuanced feedback analysis in Providing Guidance. These findings confirm that prompt engineering—when guided by educational theory and structured evaluation logic—can significantly improve the interpretability and reliability of LLM-based tutor assessments. Although fine-grained distinctions between evaluation categories remain challenging, the results underscore the feasibility of using large language models for scalable, pedagogically sound evaluation in open-ended educational dialogues.

Future work may explore integrating these methods into real-time tutoring systems, applying task-specific fine-tuning to improve classification sensitivity, and extending the framework to multi-modal or domain-specific educational contexts. Ultimately, this line of research contributes to developing AI systems that are not only linguistically fluent but also aligned with human learning objectives.

Limitations

As the proposed methods relies on the model's internal reasoning to perform evaluations, it may yield interpret evaluation criteria differently depending on the model. This is especially true for intermediate categories such as "To some extent", where subjective interpretation can lead to ambiguity, indicating a limitation in ensuring the reliability of automatic assessment.

Acknowledgments

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2023-00216011, Development of artificial complex intelligence for conceptually understanding and inferring like human). This work also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00553041, Enhancement of Rational and Emotional Intelligence of Large Language Models for Implementing Dependable Conversational Agents).

References

Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, page 487–499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Leo Breiman. 2001. Random Forests. *Machine learning*, 45:5–32.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd acm sigkdd international conference on* *knowledge discovery and data mining*, pages 785–794.

Alexis Chevalier, Jiayi Geng, Alexander Wettig, Howard Chen, Sebastian Mizera, Toni Annala, Max Jameson Aragon, Arturo Rodríguez Fanlo, Simon Frieder, Simon Machado, Akshara Prabhakar, Ellie Thieu, Jiachen T. Wang, Zirui Wang, Xindi Wu, Mengzhou Xia, Wenhan Xia, Jiatong Yu, Jun-Jie Zhu, and 3 others. 2024. Language Models as Science Tutors. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

D. R. Cox. 1958. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. 2025. Security and Privacy Challenges of Large Language Models: A Survey. *ACM Computing Surveys*, 57(6):1–39.

Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large Language Models in Education: Vision and Opportunities. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4776–4785, Los Alamitos, CA, USA. IEEE Computer Society.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.

Hongjin Kim, Jeonghyun Kang, and Harksoo Kim. 2025. Can Large Language Models Differentiate Harmful from Argumentative Essays? Steps Toward Ethical Essay Scoring. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8121–8147, Abu Dhabi, UAE. Association for Computational Linguistics.

Hongjin Kim, Jai-Eun Kim, and Harksoo Kim. 2024. Exploring Nested Named Entity Recognition with Large Language Models: Methods, Challenges, and Insights. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8670, Miami, Florida, USA. Association for Computational Linguistics.

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 *In Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.

Hang Li, Tianlong Xu, Kaiqi Yang, Yucheng Chu, Yanling Chen, Yichi Song, Qingsong Wen, and Hui Liu. 2024. Ask-Before-Detection: Identifying and Mitigating Conformity Bias in LLM-Powered Error Detector for Math Word Problem Solutions. *arXiv preprint arXiv:2412.16838*.

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, Singapore. Association for Computational Linguistics.

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.

OpenAI. 2025. Introducing GPT-4.1 in the API. `https://openai.com/index/gpt-4-1/`. Accessed: 2025-05-20.

Rose E. 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*. Association for Computational Linguistics.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024b. Large Language Models for Education: A Survey and Outlook. *arXiv preprint arXiv:2403.18105*.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024c. Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. 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 257–279, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

A Chain-of-Thought-based Pedagogical Evaluation with Reasoning layers Prompts

This section presents the detailed prompt used in the Chain-of-Thought-based Pedagogical Evaluation with Reasoning Layers methodology described in Section 2.1. The prompt serves as a core component of the proposed step-by-step structure, which emulates how human tutors interpret students' reasoning and determine appropriate feedback. It guides the model to first reconstruct the student's reasoning, then generate a correct solution path, and finally evaluate the pedagogical appropriateness of the tutor's response in light of the student's thinking process.

Promblem Extraction

# Identity

You are an expert in analyzing conversations and extracting specific information precisely from textual inputs.
Your task is to read through a dialogue transcript carefully and extract a math problem as-is, without modifying any part of it. The conversation always contains exactly one math problem.

# Instructions

* Read the conversation carefully and identify the one math problem embedded within.
* Copy the entire text of that math problem verbatim, exactly as it appears in the dialogue.
* Do not add any explanation, paraphrasing, or interpretation.
* Output only the extracted math problem and nothing else.

Student Reasoning Process Reconstruction

# Identity

You are an expert in analyzing educational conversations to reconstruct the reasoning processes behind students' mathematical answers.
Your task is to read a conversation between a tutor and a student, along with the math problem discussed. Then, from the student's point of view, explain the reasoning process the student might have used to arrive at their final answer.
Your goal is to reconstruct the student's reasoning path — whether correct or incorrect — as faithfully and coherently as possible based on the conversation and the problem given.

# Instructions

1. Do not modify, correct, or reinterpret the student's final answer — even if the answer is incorrect.

2. Base your reasoning entirely on what was stated in the conversation.

3. If no reasoning was explicitly given by the student, infer a likely and plausible thought process they could have followed to reach their answer.

4. Ensure that your explanation is logically consistent with the content of the conversation. Avoid introducing contradictions.

5. Write your reasoning as a clear, step-by-step explanation, emulating how the student may have thought through the problem.

Correct Reasoning Process Generation

# Identity

You are a logical reasoning assistant specialized in mathematical thinking and student misconception analysis.

You are given:

1. A transcript of a conversation between a student and a tutor
2. The math problem discussed in the conversation
3. The student's reasoning process, which has been reconstructed based on the conversation, not written directly by the student.
   - Parts explicitly mentioned in the dialogue can be trusted as the student's actual reasoning.
   - Parts not mentioned directly have been inferred based on the dialogue and should be treated as plausible, but not definitive.

Your task is to carefully analyze the student's reasoning process and perform the following instructions:

# Instructions

Step 1: Identify Reasoning Errors Review the student's reasoning. Clearly point out any logical, mathematical, or conceptual errors. If there are no errors, state that explicitly.

Step 2: Reconstruct the Correct Reasoning (Based on Student's Thought Process) If the student made partial progress or had a valid approach but made an error along the way, retain and respect their original reasoning path. Correct the specific mistakes and continue the reasoning from where they deviated. If the student's reasoning is fundamentally flawed from the beginning or completely irrelevant to the problem, it is acceptable to construct a new, correct reasoning path.

Step 3: Solve the Problem Using the corrected reasoning (rooted in the student's approach if applicable), solve the math problem and provide the correct final answer.

Step 4: Output Format Provide your response in the following structure:
- Student Reasoning Error(s): [List and explain]
- Corrected Reasoning (Respecting Student's Logic): [Step-by-step, rooted in their original path]
- Final Answer: [Answer]

Tutor Response Evaluation

# Identity

You are a senior math tutor and tutor coach with expertise in evaluating instructional quality. You will receive the following inputs:

1. A dialogue between a student and a novice tutor.
2. The math problem discussed in the dialogue.
3. A senior tutor's analysis of the student's likely reasoning and a revised correct solution.
4. The final utterance made by the novice tutor.

# Instructions

Evaluation Criteria

- Evaluate the novice tutor's final utterance, using the following four criteria:
- broader dialogue context, including the student's previous responses and the progression of the conversation.
- In other words, evaluate how well the tutor's final utterance functions as a response within the instructional flow and in light of the student's reasoning process.

1. Mistake Identification
    - Did the novice tutor demonstrate awareness of a mistake in the student's reasoning?
        - Yes: The tutor reasonably indicates awareness of the student's mistake or explicitly suggests the possibility of an error, even if somewhat general.
        - To some extent: The tutor vaguely hints at a mistake, but the suggestion is overly ambiguous or uncertain.
        - No: The tutor does not identify or suggest any mistake in the student's reasoning.
2. Mistake Location
    - Did the novice tutor pinpoint where the mistake occurred in the student's process?
        - Yes: The tutor appropriately identifies or indicates the step or area where the student's mistake occurred. Exact pinpointing is not required, as long as the general location or nature of the error is clear.
        - To some extent: The tutor provides only a vague or unclear indication of the error's location, potentially leading to student confusion.
        - No: The tutor does not specify or reference where the student's error occurred.
3. Pedagogical Guidance
    - Did the novice tutor provide helpful explanations, hints, or examples to support student learning?
        - Yes: The tutor provides explanations, hints, or examples that meaningfully support student understanding. Slight inaccuracies or imperfections are acceptable as long as the guidance is helpful from the student's perspective.
        - To some extent: The tutor offers guidance, but it contains significant ambiguities, inaccuracies, or potential misconceptions.
        - No: The tutor provides no useful explanation or hints, or the provided guidance is clearly incorrect or misleading.

4. Actionability
    - Can the student clearly understand what to do next based on the tutor's response?
        - Yes: The tutor reasonably suggests a clear next step or strategy that the student can readily understand and follow. Explicit instructions are not required as long as the suggested action is practical and clear enough.
        - To some extent: The tutor suggests a next step, but the recommendation is unclear, confusing, or insufficiently specific.
        - No: The tutor does not suggest any actionable next step or strategy for the student.

Output Format (MANDATORY)
- Respond in exactly the following format. Do not change the structure, headings, or indentation.

"Mistake Identification: [Yes / To some extent / No]
Explanation: [...]

Mistake Location: [Yes / To some extent / No]
Explanation: [...]

Pedagogical Guidance: [Yes / To some extent / No]
Explanation: [...]

Actionability: [Yes / To some extent / No]
Explanation: [...]"

You must follow this format strictly. Any deviation will be considered incorrect.
Now, evaluate the novice tutor's final utterance.

Basic Prompt

# Identity
You are a senior math tutor and tutor coach with expertise in evaluating instructional quality.
You will receive the following inputs:

1. A dialogue between a student and a novice tutor.
2. The final utterance made by the novice tutor.

# Instructions:

Evaluation Criteria
- Evaluate the novice tutor's final utterance, using the following four criteria.

1. Mistake Identification
   - Detect whether tutors' responses recognize mistakes in students' responses. The following categories are included:
     - Yes: The mistake is clearly identified/recognized in the tutor's response.
     - To some extent: The tutor's response suggests that there may be a mistake, but it sounds as if the tutor is not certain.
     - No: The tutor does not recognize the mistake (e.g., they proceed to simply provide the answer to the asked question).
2. Mistake Location
   - Assess whether tutors' responses accurately point to a genuine mistake and its location in the students' responses. The following categories are included:
     - Yes: The tutor clearly points to the exact location of a genuine mistake in the student's solution.
     - To some extent: The response demonstrates some awareness of the exact mistake, but is vague, unclear, or easy to misunderstand.
     - No: The response does not provide any details related to the mistake.
3. Pedagogical Guidance
   - Evaluate whether tutors' responses offer correct and relevant guidance, such as an explanation, elaboration, hint, examples, and so on. The following categories are included:
     - Yes: The tutor provides guidance that is correct and relevant to the student's mistake.
     - To some extent: Guidance is provided but it is fully or partially incorrect, incomplete, or somewhat misleading.
     - No: The tutor's response does not include any guidance, or the guidance provided is irrelevant to the question or factually incorrect.
4. Actionability
   - Assess whether tutors' feedback is actionable, i.e., it makes it clear what the student should do next. The following categories are included:
     - Yes: The response provides clear suggestions on what the student should do next.
     - To some extent: The response indicates that something needs to be done, but it is not clear what exactly that is.
     - No: The response does not suggest any action on the part of the student (e.g., it simply reveals the final answer).

Output Format (MANDATORY)

- Respond in exactly the following format. Do not change the structure, headings, or indentation.

Mistake Identification: [Yes / To some extent / No]
Explanation: [...]

Mistake Location: [Yes / To some extent / No]
Explanation: [...]

Pedagogical Guidance: [Yes / To some extent / No]
Explanation: [...]

Actionability: [Yes / To some extent / No]
Explanation: [...]

You must follow this format strictly. Any deviation will be considered incorrect.
Now, evaluate the novice tutor’s final utterance.

B Rubric Based Evaluation Method prompt

To evaluate the Providing Guidance dimension, we designed a structured prompt that guides the language model to simulate expert judgment across six sub-criteria. The prompt first provides the full dialogue context, the student's final utterance, and the tutor's response. It also includes prediction results from other evaluation dimensions (Mistake Identification and Mistake Location), as well as statistical correlations observed between them. The model is instructed to assess the tutor's response based on six specific questions, each targeting a pedagogically meaningful aspect of guidance. For each item, the model outputs a brief rationale and assigns probabilities to three labels: Yes, To some extent, and No. The final output consists of both the explanation and a normalized probability distribution that sums to 1.0. The sixth question is designed to produce an overall judgment by incorporating both model predictions and prior statistical informations, providing a holistic measure of guidance quality.

Prompt for Rubric-Based Multidimensional Evaluation

You are an expert evaluator analyzing a tutor's response in a learning dialogue.

Below is a conversation between a student and a tutor.

[**Conversation**]
<Full conversation history, if any>

[**STUDENT_UTTERANCE**]
<Student's latest input>

[**TUTOR_RESPONSE**]
<Tutor's response to be evaluated>

[**Prediction Results**]
- Mistake Identification: <Predicted label>
- Mistake Location: <Predicted label>

Note: Based on statistical analysis of past data, the following association rules are observed:

- If Providing Guidance is "Yes", then Mistake Identification is also "Yes" with confidence 0.950.
- If Mistake Location is "Yes", then Mistake Identification is "Yes" with confidence 0.989.
- If Mistake Identification is "Yes", then Mistake Location is "Yes" with confidence 0.790.

Use the following definitions when choosing a label:

- Yes: The tutor's response fully satisfies the criterion. It is accurate, relevant, and helpful.
- To some extent: The response attempts to satisfy the criterion but is partially incomplete, inaccurate, vague, or not directly useful.
- No: The response does not satisfy the criterion at all, or it is misleading, unrelated, or entirely incorrect.

Please answer the following six questions. For each question, first provide a brief explanation for your judgment. Then, give the probability (in float format) that the response is: Yes, To some extent, or No. Ensure all three values sum to exactly 1.0.

Output format:
`Q1`: <brief explanation>
- Yes: <float>
- To some extent: <float>

- No: <float>

Questions:

Q1. Did the tutor attempt to provide any explanation, hint, or example?

Q2. Was the guidance factually correct and appropriate given the student's error?

Q3. Did the tutor's response directly address the student's specific mistake?

Q4. Did the guidance help the student figure out what to do next, without directly giving the final answer?

Q5. Was the tutor's response clear and unlikely to confuse the student?

Q6. Based on the tutor's response, the model's predictions, and the above statistical information, how likely is it that the tutor attempted to provide meaningful guidance?

C Analysis of Prediction Results on the Development Set

This section presents the classification results on the development set for each of our proposed systems. Although the overall metrics provide a broad overview of performance, they do not sufficiently capture the models' ability to discriminate fine-grained categories—particularly ambiguous ones such as To some extent. Therefore, we provide a detailed analysis of each system's predictions according to the evaluation track.

C.1 Chain-of-Thought-based Evaluation System

As shown in Section 3.3, our proposed Chain-of-Thought-based evaluation system demonstrated effectiveness in generating evaluations that are both consistent and pedagogically valid. However, as previously noted, the model still exhibits limitations in accurately distinguishing between semantically adjacent evaluation categories. To further investigate this issue, we conducted an analysis of how such difficulties manifest in actual prediction outcomes.

Actual / Predict	Yes	To some extent	No	Total
Yes	1,657	237	38	1,932
To some extent	63	68	43	174
No	60	56	254	370

Table 6: Confusion matrix of the CoT-based Evaluation System for the Mistake Identification track.

Table 6 presents the prediction results for the Mistake Identification track in the form of a confusion matrix. In this track, the model accurately classified the majority of "Yes" instances (1,657 out of 1,932), but struggled to distinguish the "To some extent" category. Specifically, only 68 out of 174 instances were correctly identified, while the remaining were misclassified as "Yes" (63 instances) or "No" (43 instances), indicating persistent challenges in delineating fine-grained judgment boundaries.

Actual / Predict	Yes	To some extent	No	Total
Yes	727	547	36	1,310
To some extent	88	245	36	369
No	253	183	361	797

Table 7: Confusion matrix of the CoT-based Evaluation System for the Actionability track.

A similar pattern is observed in the Actionability track, as shown in Table 7. While the model achieved relatively high true positive counts for the "Yes" (727 instances) and "No" (361 instances) categories, "To some extent" cases were frequently misclassified—most notably, among the actual "No" instances, 547 were predicted as "Yes" and 183 as "To some extent".

These results indicate that while the proposed Chain-of-Thought-based evaluation system is effective in producing consistent judgments based on explicit criteria, it still faces limitations in clearly distinguishing semantically adjacent categories. In particular, the frequent misclassification of ambiguous labels such as To some extent highlights the difficulty of inducing fine-grained reasoning solely through prompts without task-specific training. This observation suggests the potential need for improved prompt engineering or subsequent fine-tuning to enhance the model's discriminative precision.

C.2 Multi-Perspective Reflective Evaluation System

This section presents an analysis of the Multi-Perspective Reflective Evaluation System's performance on the Mistake Location track. The goal is to understand how effectively the system distinguishes between clearly defined and semantically adjacent categories within its reflective reasoning framework.

Actual \ Predicted	Yes	To some extent	No	Total
Yes	1,219	140	184	1,543
To some extent	99	33	88	220
No	117	79	517	713

Table 8: Confusion matrix of the Multi-Perspective Reflective Evaluation System for the Mistake Location track.

The analysis of Mistake Location is presented in Table 8. The results show that although the system accurately identifies many instances of "Yes" (1,219 correct predictions), it struggles to distinguish "To some extent" from adjacent categories. Specifically, only 33 out of 220 "To some extent" cases were correctly classified, while the majority were misclassified as either "Yes" (99 instances) or "No" (88 instances). This analysis supports our observations in Sections 3.3 and 3.4 that prompt-based reasoning approaches still face challenges in making fine-grained categorical distinctions.

C.3 Rubric Based Evaluation System

To assess the performance of the Rubric Based Evaluation System in identifying pedagogically meaningful distinctions, we examine its predictions on the Providing Guidance track. This allows us to evaluate the effectiveness of rubric-derived features in capturing subtle differences between response categories.

Actual / Predict	Yes	To some extent	No	Total
Yes	1,034	266	107	1,407
To some extent	250	179	74	503
No	178	83	305	566

Table 9: Confusion matrix of the Rubric Based Evaluation System for the Providing Guidance track.

Table 9 presents the prediction results of our proposed Rubric Based Evaluation System. The system consists of a Random Forest classifier trained using the Q_{rubric} items and the predicted values of $Q_{overall}$ as input features. For the "Yes" class, 1,034 out of 1,407 instances were correctly classified. For the "No" class, 305 out of 566 instances were correctly classified. In contrast, for the "To some extent" class, only 179 out of 503 instances were correctly classified, with 250 instances misclassified as "Yes" and 74 as "No." These results indicate that the classifier struggled to clearly distinguish the "To some extent" class from "Yes" and "No." This suggests that, even when combining information from Q_{rubric} and $Q_{overall}$, additional feature engineering or refinement may be necessary to more precisely delineate the boundaries among the three classes.