Explanation Of Revisions

In the previous submission, reviewers and the area chair provided many valuable comments, which can be summarized into the following eight points. We have thoroughly considered and improved upon these comments. Below, we will describe the changes we have made in response to these points.

> Limited Generalizability Due to the Elementary-Level Scope and Bilingual (English–Korean) Constraints of the Bi-GSM8K Dataset

In the Limitations section of the paper, under the subsection titled "Limitations of the Dataset: Elementary-Level Scope, Simulated Student Solutions, and Language Selection Considerations", we explicitly state the rationale and research validity behind restricting the dataset to elementary-level mathematics problems and bilingual (English–Korean) content. These constraints were necessitated by the objectives and experimental design of the present study. We acknowledge, however, that such limitations may affect the generalizability of the dataset. Accordingly, we emphasize the need for future work to extend the dataset to encompass a broader range of languages and educational levels.

> Ambiguity in the Definition of "Student Solutions" and Absence of Genuine Student-Generated Data

We acknowledge that the initial version of the manuscript lacked a clear definition of the term student solution. In Section 3, we have clarified that student solutions do not refer to answers written by actual students, but rather to responses generated by data construction experts who simulated typical student error patterns. Furthermore, we explicitly recognize the limitations of this approach—namely, its inability to fully capture the cognitive processes of real students—in the Limitations section, under the subsection "Limitations of the Dataset: Elementary-Level Scope, Simulated Student Solutions, and Language Selection Considerations"

> Lack of Detailed Description Regarding Prompt Design, Example Selection Criteria, and Handling of Ambiguous or Complex Errors in the SED Module

We have provided a detailed explanation of the prompt design and the criteria used for selecting few-shot examples in the SED module in Appendix G. Additionally, the methodology for identifying and handling the initial point of complex or ambiguous errors is thoroughly described in Section 3.

> Precise Definition of the GTA Module

While this paper proposes the GTA module as a novel evaluation metric for measuring similarity between teacher and student solutions, the initial version primarily focused on the alignment outcomes between the two solutions. To address this, Section 4.1 has been expanded to provide a detailed description of the alignment-based similarity score computation method. Furthermore, Section 5.2 quantitatively validates the proposed evaluation metric by analyzing the Pearson and Spearman correlation coefficients between the similarity scores generated by the model and those assigned by human annotators.

> Insufficient Explanation of Input Data Types (Vectors, Text, etc.) and the Rationale for Using Different Similarity Metrics (Cosine Similarity, Pearson Correlation, SemScore, BERTScore) in the GTA Module for Decoder- and Encoder-based LLMs

The input formats and the rationale behind applying different similarity metrics in the GTA module are comprehensively detailed in Appendix C.

Lack of Human Expert Validation for Evaluation Metrics (GTA and SED)

For the GTA module, Section 5.2 experimentally validates the validity of the evaluation metrics by comparing alignment similarity and error detection results between human annotators and the model. Appendix E provides the human evaluation guidelines, and Appendix F presents an analysis of inter-annotator agreement for alignment consistency, quantitatively assessing reliability and consistency. Regarding the SED module, Section 5.3 clarifies that no separate human evaluation is required, as the ground truth for initial error points is provided by mathematics experts and data annotation specialists.

Dependence on a Single Commercial Model (GPT-4o)

We extended the experiments by incorporating additional commercial models, Claude-Sonnet-4 and Gemini-2.5-Flash, into Tables 1 and 2 of Section 5. The evaluation results for these models are reported alongside those of GPT-4o, enabling comparative analysis across commercial LLMs. The corresponding findings are explicitly presented in Section 5.

Omission of Baseline Evaluation for LLM-Only Error Detection Without Access to Teacher Solutions

We have included baseline experimental results in Appendix D that evaluate the performance of the LLM in detecting the first error without access to the teacher solution. This addition enables a clearer comparison of the relative effectiveness of the alignment-based approach proposed in this study.

Lack of Comparative Experiments with Smaller Commercial Models (e.g., GPT-o1-mini) to Support the Claim that “Some Open-Source Models Perform Error Analysis Faster than Large-Scale Commercial Models”

The claim regarding error analysis speed pertains specifically to the GTA module, not the SED module. To prevent potential misunderstanding, we have revised the wording in both the abstract and Section 6 (Conclusion) to clarify this distinction. Furthermore, we have added experimental results for GPT-o1-mini in Table 1 of Section 5, empirically demonstrating that the proposed BERTScore + LM + NW algorithm computes similarity scores more efficiently than smaller commercial models.

Lack of Performance Visualization

To address the lack of performance visualization, we have added supplementary materials in Appendix J. Specifically, we provide radar charts depicting the Pearson and Spearman correlation coefficients of the GTA module, enabling intuitive comparison across both languages and models. For the SED module, accuracy scores are similarly visualized using radar charts, facilitating cross-linguistic and cross-model comparisons. Additionally, we present vertical bar charts illustrating model-wise latency for both the GTA and SED modules, offering a clearer view of the computational efficiency associated with each configuration.