Majumdar, R. et al. (Eds.) (2025). Proceedings of the 1st International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education

# A Rubric-based LLM Automatic Grading of Mathematical Reasoning in Self-explained Answers

**Taisei YAMAUCHI[a*], Brendan FLANAGAN[b, e], Yiling DAI[c], Toshihiro KITA[d] & Hiroaki OGATA[e]**
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Institute for Liberal Arts and Sciences, Kyoto University, Japan*
[c]*Graduate School of Advanced Science and Engineering, Hiroshima University, Japan*
[d]*Research and Education Institute for Semiconductors and Informatics, Kumamoto University, Japan*
[e]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*yamauchi.taisei.28w@st.kyoto-u.ac.jp

**Abstract:** This research addresses challenges in automated math grading by focusing on assessing complex reasoning in high-order math questions through written self-explanations. We developed a rubric-based scoring system using LLMs, incorporating an algorithmic output checker and self-consistency sampling. Twelve self-explanations were scored, with expert grades as the gold standard. Results show the algorithmic checker outperforms the LLM-based method, and self-consistency sampling enhances alignment with expert judgments. Overall, the approach will offer accurate feedback, reducing teacher workload, boosting student engagement, and enhancing scalability, while indicating a need for automated rubric generation.

**Keywords:** Automatic grading, large language models, reasoning, math education

## 1. Introduction

Automated grading and real-time feedback can substantially reduce teacher workload and boost learning outcomes (Celik et al., 2022) yet assessing high order thinking and ensuring broad applicability remain challenging (Langove & Khan, 2024). Existing tools like MathDIP provide step-level feedback on handwritten formulas but focus solely on computation, neglecting reasoning (Pacheco-Venegas et al., 2015). To address this gap, we target students' written self-explanations, which reveal causal and conceptual inferences vital for complex reasoning (Bisra et al., 2018). Although benchmarks such as MathBench (Liu et al., 2024) offer structured evaluation, they may not reflect real-world performance (Porcu & Havlínová, 2024). Recognizing that Nakamoto et al. (2025)'s collaborative feedback system can propagate students' inaccuracies and undermine reliability, we introduce a rubric-based automatic grading framework that combines LLM-based scoring, algorithmic output validation, and self-consistency sampling. Our study investigates: *to what extent can LLMs accurately grade high-order mathematical reasoning in students' self-explanations?*

## 2. Methods

In December 2021, we collected 12 written self-explanations produced as regular homework submissions by Grade 8 students (13–14 years old) using self-explanation box in the LEAF system (Flanagan & Ogata, 2018). We treated these narratives as exercise answers. Three experts independently scored each explanation using a binary rubric with five items: use of mathematical definitions, equation formulation, calculation techniques, mathematical thinking (including higher-order reasoning), and key considerations. Inter-rater agreement, measured by Fleiss's Kappa ($\kappa$ = 0.67), denotes substantial consistency per Landis & Koch's standards (1977). This $\kappa$ value indicates that expert scores were sufficiently consistent.

We developed an automatic scoring system (Figure 1) that evaluates student self-explanations against a provided rubric using self-consistency LLM sampling. Input data include question text, rubrics, a standard answer, and a student self-explanation. From these, we constructed prompts (Figure 2) following OpenAI's best practices and ran them on the "gpt-4o-mini" model. Each LLM output is validated by an output checker—either algorithmic-based or LLM-based—which enforces the required format and, upon failure, triggers regeneration (up to five attempts). The algorithmic checker parses the output for the expected count of grading markers (e.g., total "_o_" and "_x_"), while the LLM-based checker uses another LLM to verify compliance with formatting instructions. To counteract the probabilistic variability of LLM outputs, we applied self-consistency: the same prompt is sampled multiple times, and final scores are determined by majority vote (Wang et al., 2023).

With self-consistency fixed at n = 1, we evaluated format compliance for 36 outputs (3 runs×12 explanations), checking for exactly five "_o_" or "_x_" markers using both algorithmic- and LLM-based checkers and comparing their accuracy. Next, we varied n, generating three runs per setting, and used Fleiss's kappa to identify the optimal n. Finally, we compared modal LLM scores to expert grades via Cohen's kappa.

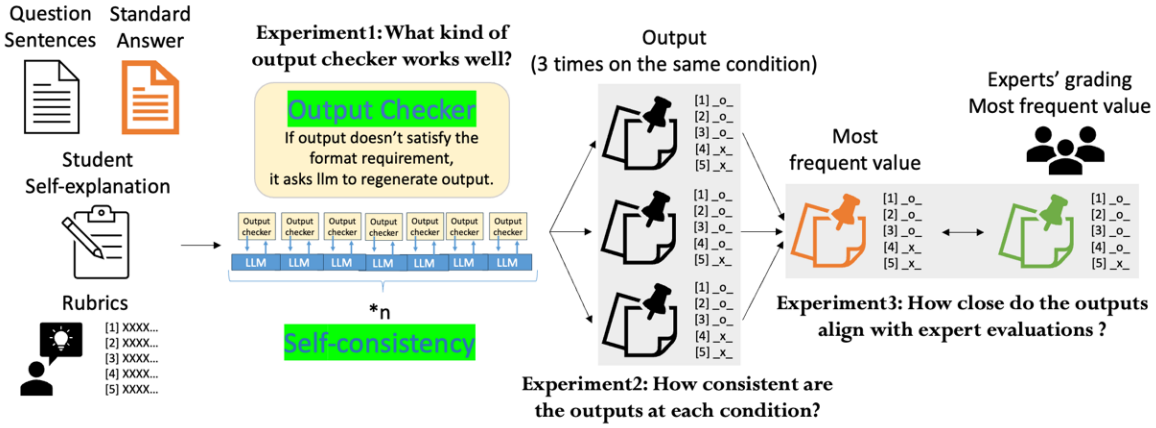


*Figure 1.* An overview of auto-grading method and an experimental flow.

```
The following statement is a math question. {question_sentences}
Rubrics required for these questions are as follows. {rubrics}
A standard answer below satisfies all these rubrics. {standard_answer}
On the other hand, the following statement is the student's answer. It
  may be incomplete. {students_self_explanation}
Determine whether student's answer satisfies each of the rubrics
  required for this question, and output "_o_" if they satisfies,
  otherwise output "_x_".
Be sure to follow the format below.
- For XXXX, enter each rubric title.
- Replace "_*_" with "_o_" or "_x_".
- Do not output anything else.
Format
[1] XXXX _*_
[2] XXXX _*_
...
```

*Figure 2.* A prompt for auto-grading LLM (translated in English).

## 3. Results & Discussion

In this study, we developed a rubric-based automatic grading system using an LLM to assess students' deeper mathematical reasoning—an area not fully addressed in prior work (Nakamoto et al., 2025; Pacheco-Venegas et al., 2015). For output-format validation, our algorithmic-based checker achieved 100.0% accuracy, whereas the LLM-based checker attained 91.7% accuracy, demonstrating the algorithmic method's superior robustness when deterministic checks are feasible. Next, we evaluated the self-consistency of LLM outputs by varying the number of sampled outputs ($n$ = 1, 3, 5, 7, 9, 11). As shown in the left panel of Figure 3, Fleiss's Kappa coefficient increases gradually with $n$ (Mann–Kendall test $p$ = 0.060),

reaching its maximum at $n$ = 11. To benchmark against human graders, we compared the LLM's representative score (the mode of the n outputs) to expert scores using Cohen's Kappa. The right panel of Figure 3 indicates that agreement peaks at $n$ = 7 and then stabilizes around 0.63—close to the experts' inter-rater reliability of 0.67—implying that our LLM-based method can replicate expert grading with similar reproducibility.

Overall, our rubric-driven framework effectively combines algorithmic checks for format with LLM evaluation of reasoning, delivering accurate and timely feedback. While mathematical reasoning provides a natural domain for this approach, future work should explore rubric generation automation to reduce manual effort and extend applicability to other subjects and educational levels.

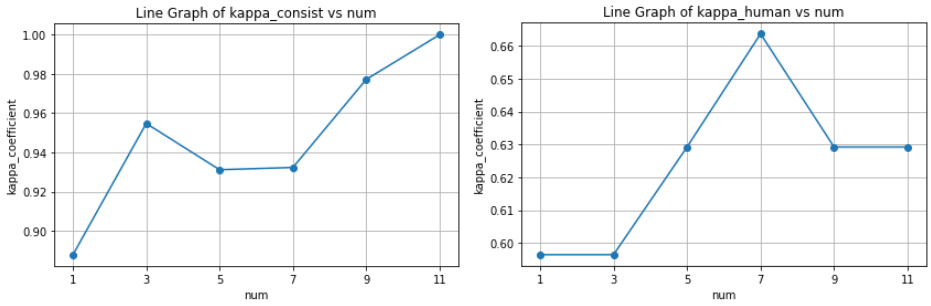


*Figure 3.* Consistency of LLM outputs increases with the number of outputs (left), Degree of agreement between LLM scoring and expert grading (right).

## Acknowledgements

This work was partly supported by JST BOOST JPMJBS2407, JSPS JP24K20902, JP20H01722, JP23K25698 and JP23H01001, JP21K19824, JP23H00505, NEDO JPNP20006, and CSTI SIP Program JPJ012347.

## References

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review, 30*, 703-725.

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends*, *66*(4), 616-630.

Flanagan, B., & Ogata, H. (2018). Learning analytics platform in higher education in Japan. *Knowledge Management & E-Learning: An International Journal, 10*, 4, 469-484.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM computing surveys, 55*(12), 1-38.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Langove, S. A., & Khan, A. (2024). Automated grading and feedback systems: Reducing teacher workload and improving student performance. *Journal of Asian Development Studies*, *13*(4), 202-212.

Liu, H., Zheng, Z., Qiao, Y., Duan, H., Fei, Z., Zhou, F., ... & Chen, K. (2024). MathBench: Evaluating the theory and application proficiency of LLMs with a hierarchical mathematics benchmark. *Findings of the Association for Computational Linguistics ACL 2024*.

Nakamoto, R., Flanagan, B., Dai, Y., Yamauchi, T., Takami, K., & Ogata, H. (2025). Integrating self-explanation and operational data for impasse detection in mathematical learning. *Research and Practice in Technology Enhanced Learning.*

Pacheco-Venegas, N. D., López, G., & Andrade-Aréchiga, M. (2015). Conceptualization, development and implementation of a web-based system for automatic evaluation of mathematical expressions. *Computers & Education, 88*, 15-28.

Porcu, V., & Havlínová, A. (2024). Breaking down the metrics: A comparative analysis of LLM benchmarks. *International Journal of Science and Research Archive, 2024, 13*(02), 777–788.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., ... & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.