Can Artificial Intelligence Replicate Human Grading?

Valentina Scorza[0009−0001−9321−8804], Giacomo Cassano[0000−0002−1784−0626], and Nicoletta Di Blas[0000−0002−7987−2930]

Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
valentina.scorza@polimi.it, giacomo.cassano@polimi.it,
nicoletta.diblas@polimi.it
https://www.polimi.it

1 Introduction

Can artificial intelligence take on one of the most human aspects of teaching: grading open-ended student responses? This kind of assignment is essential in education, fostering critical thinking and revealing students' reasoning and conceptual understanding [1]. However, the evaluation of open-ended responses is time-consuming and difficult to scale [2]. With the advent of large language models (LLMs), the educational landscape has been significantly transformed, enabling new possibilities for automating complex tasks [3]. While recent studies have explored the use of LLMs for assessing open-ended responses [4–7], they rarely address adaptation to the specific and often variable evaluation criteria used by individual instructors. This work investigates whether LLMs, specifically GPT-4o and GPT-4o Mini, can replicate a teacher's grading process not only by producing plausible scores, but also by aligning with instructor-specific rubrics. Teachers often evaluate responses either explicitly, using structured criteria [8], or implicitly, based on experience. An effective AI grader must be able to emulate both. To explore this, we implemented two evaluation strategies: a *shot-based approach*, in which models infer grading criteria from scored examples (implicit rubric learning), and a *rubric-based approach*, which relies on structured rules (explicit rubric learning). We compare these strategies to assess the models' ability to reproduce human judgment and support scalable, consistent, and fair assessment in educational contexts.

2 Implementation and Method

A system based on LLMs and using LangChain (a framework that streamlines the integration of LLMs with external data sources) was developed to evaluate open-ended student responses, adopting two distinct approaches. In the **shot-based approach**, the model uses a few scored examples to infer grading criteria and assess new responses. It evaluates the model's ability to generalize and apply human-like judgment, also providing feedback to justify the score. The **rubric-based approach** uses teacher-provided rubrics and scoring formulas to grade

structured responses. It ensures consistent, objective grading and includes feedback explaining the score. Two professors from Politecnico di Milano contributed datasets for this study. The first provided two sets used to evaluate the shot-based approach, each containing six open-ended questions from the Technologies for Information Systems course on communication, answered by 37 and 43 students respectively, with scores and optional feedback but no formal rubric. The second contributed a single-question dataset from the Leadership and Innovation course, including 276 responses. This set included a detailed evaluation rubric and scoring formula, allowing for rubric-based assessment.

3 Results

The findings of the study reveal that both GPT-4o and GPT-4o Mini show strong potential in evaluating student responses, particularly within the *mid-quality range*. In the **shot-based evaluation**, performance improved with more example answers, achieving better alignment with teacher scores for mid-range responses. Accuracy decreased at score extremes, where subtle distinctions and nuanced reasoning are harder to assess. This approach was especially effective for complex questions requiring deep reasoning, such as identifying and explaining key concepts in lengthy texts. In the **rubric-based evaluation**, complementary strengths were exhibited by the models: **GPT-4o Mini** excelled at detecting relevant content and assigning partial scores, leading to better overall alignment with teacher grading; **GPT-4o** was more consistent in computing final grades from partial scores. Human grading variability, due to holistic or interpretive judgments beyond strict rubric criteria, reduced model alignment. Introducing a tolerance margin improved correspondence, particularly for GPT-4o Mini, highlighting its robustness to flexible grading.

4 Conclusions

This study explored whether **GPT-4o** and **GPT-4o Mini** can replicate teacher grading by aligning with either implicit or explicit rubrics. Results show strong agreement with human scores, especially for mid-range responses, capturing both implicit and explicit grading criteria. LLMs demonstrate strong consistency, aligning well with teacher evaluations across a wide range of responses. Nevertheless, they struggle at the extremes, failing to fully recognize partial understanding in weaker answers and creative reasoning in outstanding ones, due to their strict adherence to rubrics. However, in scenarios where human grading is inconsistent or deviates from the rubric, LLMs have the potential to enhance assessment quality by improving fairness and alignment with established evaluation criteria. Rather than replacing human evaluators, these models offer valuable support as intelligent assistants, providing scalable and consistent evaluation. Future work should focus on improving LLMs' ability to balance rubric fidelity with the flexible, nuanced judgment characteristic of human grading, advancing their role as complementary tools in educational assessment.

References

1. Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*, 5–31. https://doi.org/10.1007/s11092-008-9068-5
2. Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student-constructed responses. *Behavior Research Methods*, *44*, 608–621. https://doi.org/10.3758/s13428-012-0211-3
3. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274
4. Chang, L.-H., & Ginter, F. (2024, March). Automatic Short Answer Grading for Finnish with ChatGPT. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(21), 23173–23181. https://ojs.aaai.org/index.php/AAAI/article/view/30363
5. Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100213. https://doi.org/10.1016/j.caeai.2024.100213
6. Impey, C., Wenger, M., Garuda, N., Golchin, S., & Stamer, S. (2025). Using Large Language Models for Automated Grading of Student Writing about Science. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00453-7
7. Henkel, O., Hills, L., Roberts, B., Zhang, Y., Wang, Z., Singh, A., Smith, N., & Chuang, J. (2025). Can LLMs grade open response reading comprehension questions? An empirical study using the ROARs dataset. *International Journal of Artificial Intelligence in Education*, *35*, 651–676. https://doi.org/10.1007/s40593-024-00431-z
8. Wolf, K., & Stevens, E. (2007). The role of rubrics in advancing and assessing student learning. *The Journal of Effective Teaching*, *7*, 3–14.