Can Language Models Grade Algebra Worked Solutions? Evaluating LLM-Based Autograders Against Human Grading

Shreya Bhandari
School of Education
University of California, Berkeley
shreya.bhandari@berkeley.edu

Zach Pardos
School of Education
University of California, Berkeley
pardos@berkeley.edu

ABSTRACT

In this study, we investigate the feasibility of using LLMs as autograders by evaluating multiple models—OpenAI's GPT-4o, GPT-4.1-mini, and GPT-4.1-nano—on a manually graded open-source dataset of college algebra worked solution responses. For each model, we assess performance both with and without a self-consistency grading approach. We compare LLM-generated correctness labels to human annotations across 18,000 responses. Results show that GPT-4.1-mini achieves the highest accuracy (94.47% with self-consistency), followed by GPT-4.1-nano (93.07%) and GPT-4o (91.93%). These findings suggest that self-consistency can slightly improve grading reliability, and that even compact models like GPT-4.1-mini and GPT-4.1-nano can approach human-level agreement in algebra autograding.

Keywords

Generative AI, Large Language Models (LLMs), Automated Grading

1. INTRODUCTION

Grading constitutes a significant portion of teachers' overall workload. For example, the TALIS 2013 report highlights that teachers spend approximately 6 hours per week marking and correcting students' work, while the Teacher Workload Survey 2016 reports that secondary teachers spend around 8 hours on grading [6]. Prior to large-scale advancements in large language models, auto-scorers were developed using pattern matching to compare student responses against predefined answers or by parsing responses to detect specific phrases to infer correctness [17, 18]. However, these methods face several limitations, such as the inability to recognize equivalent variations of a response and the challenge of identifying the exact position where the correct answer appears within the student's response. Since large language models (LLMs) can generate and process text in a manner more akin to natural language than past mechanical methods, have proved effective in various aspects of educational content production (e.g., hints, items, skill-tagging), and can broadly support mathematics instruction in higher education, the rise of generative AI and LLMs has sparked interest in their potential to automate grading processes [15, 11, 10, 13].

Researchers have conducted studies on the efficacy of using LLMs as autograders, yielding mixed results, with most research focusing on applications in open-ended non-STEM subjects (e.g., humanities) [9, 2, 4] and open-ended STEM subjects (e.g., computer science and engineering) [12, 14, 1, 3]. Historically, it has been assumed that STEM subjects with closed-ended responses do not require sophisticated autograding because answers can be checked verbatim. Popular mastery-based learning platforms like Cognitive Tutor [8], ASSISTments [5], and OATutor [16] have leveraged this approach, relying on exact answer matching to assess student responses. However, when students show their work, grading becomes more complex. Research has shown that requiring students to show their work is an effective learning strategy [7], but existing tutoring systems often do not support such responses because they are not trivially scorable. Evaluating shown work requires parsing the student's solution and determining whether it contains the correct answer. This raises the question: now that LLMs exist, can they be leveraged to grade responses where students demonstrate their reasoning? In this work, we focus on the use case of closed-ended autograding for responses that include shown work.

Despite LLMs demonstrating average performance in solving algebraic problems [11], we argue that grading algebraic responses should be a significantly easier task than actually solving the problems, as it primarily involves an equivalency check rather than problem-solving. To evaluate the feasibility of LLM-based autograding, we compare the grading outputs of OpenAI's GPT-4o against a manually graded open-source dataset. If the LLM-generated correctness labels closely align with human annotations, this would support the viability of using LLMs as scalable autograders, reducing the need for manual grading while maintaining reliability. In this work, we extend previous research in two ways: (1) by evaluating three OpenAI models—GPT-4o, GPT-4.1-mini, and GPT-4.1-nano—on a shared algebra dataset and (2) by testing the effect of an error mitigation technique called self-consistency [19], where each response is graded across multiple completions (10 iterations) and the modal

Shreya Bhandari, and Zach Pardos. Can Language Models Grade Algebra Worked Solutions? Evaluating LLM-Based Autograders Against Human Grading. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 554–558. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870250

result is returned.

Research Question:

- RQ: How closely do LLM-based autograders (GPT-4o, GPT-4.1-mini, GPT-4.1-nano, with and without self-consistency) agree with human-labeled correctness of algebra worked solution responses?

2. METHODS

To evaluate the feasibility of using LLMs as autograders, we compare AI-graded and human-graded correctness labels on a large dataset of college algebra worked solution responses. Our methodology has two key components: (1) producing a ground truth dataset to utilize, where correctness labels are evaluated by hand, and (2) implementing a grading pipeline using three different LLMs, each evaluated using standard prompting and self-consistency prompting.

2.1 Dataset Generation

To produce the dataset of worked solution responses to the questions—originally introduced by Liu et al. [11] and available online[1]—a variety of LLMs were utilized to simulate the human learner's answer to 20 college algebra questions. Given LLMs' natural verbosity, prompting LLMs to answer these questions resulted in not only the attempted answer but also the steps taken to arrive at the answer (i.e., a worked solution). The dataset consisted of 21 columns: the first column, Generating Model, specified the model or source that generated the responses, while the remaining 20 columns (Q1 to Q20) indicated the correctness of answers to the 20 questions. The dataset included responses from different generating models: 150 each per question from GPT-4, GPT-3.5, Llama 3, Llama 2, Gemini, and Cohere. The temperature setting for the LLMs was non-zero, allowing for variability in the responses. A sample response to be auto-graded is provided below:

Example input to autograder:

> Given that $m = 4$ and the point $(2, 5)$, the equation of the line in slope-intercept form can be found using the point-slope formula:
>
> $y - y_1 = m(x - x_1)$
>
> $y - 5 = 4(x - 2)$
>
> $y - 5 = 4x - 8$
>
> $y = 4x - 3$
>
> Therefore, the equation of the line passing through the point $(2, 5)$ with slope 4 is $y = 4x - 3$.

The LLM-generated worked solution responses were all hand-graded. The 150 responses per question across six models led to a total dataset size of 18,000 hand-graded responses ($150 \times 6 \times 20$).

[1] https://figshare.com/articles/dataset/Leveraging_LLM-Respondents_for_Item_Evaluation_a_Psychometric_Analysis/27263496?file=49883421

2.2 LLM-Based Autograder

We introduce an LLM-based autograder that evaluates correctness by comparing worked solution responses against a predefined answer key. The autograder is designed to assess equivalence in answers while maintaining strict accuracy. We define equivalence as algebraic equivalence; thus, any response that simplifies algebraically to the provided answer is considered correct. We use three OpenAI models: GPT-4o, GPT-4.1-mini, and GPT-4.1-nano. Each model is evaluated using two prompting strategies:

1. **Standard (Single) Prompting:** A deterministic grading prompt at default temperature.
2. **Self-Consistency Prompting:** The same prompt run 10 times at default temperature; the modal output ("True" or "False") is selected as the final grade.

Leveraging OpenAI's structured schema response format to ensure structured JSON-based outputs, we send a structured prompt to the model, instructing it to compare a given response against the corresponding correct answer. Specifically, we use the following system and user prompts:

> System Prompt: You are an AI grader that evaluates student responses for correctness. Equivalent answers should be considered correct.
>
> User Prompt: Check if the student's response 'response' matches the correct answer 'answer'. Output strictly 'True' or 'False'.

Each model independently grades all 18,000 responses using both prompting strategies. For the standard approach, the model returns a single binary decision ("True" or "False") per response. For the self-consistency approach, we generate 10 completions at the model's default temperature, allowing variability in responses across the 10 completions, and select the majority vote as the final grade.

To evaluate the accuracy of the autograder, we calculate the percentage overlap between the correctness of the 18,000 autograded responses and the 18,000 manually graded responses, providing a quantitative measure of agreement between the AI-based and human grading approaches.

To ensure transparency and reproducibility, we provide all source code along with all question text [2].

3. RESULTS AND DISCUSSION

We compute the agreement of the LLM with the human correctness labels. Table 1 summarizes the agreement scores across three OpenAI models—GPT-4o, GPT-4.1-mini, and GPT-4.1-nano—each evaluated with and without self-consistency.

GPT-4.1-mini achieves the highest agreement at 94.47% when using self-consistency, followed closely by GPT-4.1-nano (93.07%) and GPT-4o (91.93%). Without self-consistency, performance drops slightly for all models. Across all models, self-consistency yields marginal accuracy improvements. No-

[2] https://github.com/CAHLR/Autograder

Table 1: Model-wise Agreement with Human Grading Labels (Across 18,000 Responses)

Model and Grading Method	Overall Accuracy (Agreement)
GPT-4.1-mini (Self-Consistency)	94.47%
GPT-4.1-mini (Standard)	94.42%
GPT-4.1-nano (Self-Consistency)	93.07%
GPT-4.1-nano (Standard)	92.26%
GPT-4o (Self-Consistency)	91.93%
GPT-4o (Standard)	90.64%

tably, GPT-4.1-mini outperforms both GPT-4o and GPT-4.1-nano, even in its standard mode, suggesting that these models can offer both efficiency and high grading reliability. Although the gains from self-consistency are modest (less than 1%), they appear consistent across all models.

Interestingly, previous studies have reported GPT-4o achieving approximately 92% accuracy in algebra problem-solving tasks [20], making our observed grading accuracy (91.93%) notably close to these prior findings. It is somewhat surprising that the accuracy is not closer to 100% given that the ground truth correct answer was known by the autograder. Table 2 provides a detailed breakdown of agreement percentages for each of the 20 questions. While most questions exhibit high overlap (>90%), a few questions, such as Q6, Q13, and Q14, show lower agreement. Questions with notably lower agreement may involve more complex algebraic steps, notation inconsistencies, or multiple equivalently correct forms, making grading decisions more subjective. Table 3 compares LLM-assigned correctness percentages to human-graded values across all 20 questions.

There are several limitations and areas for future research. While our results show promising agreement across models, future work should examine misclassified responses—especially for questions with lower agreement scores such as Q6 and Q16—to identify specific error patterns. Although we explored self-consistency as a strategy to improve reliability, further investigations could assess whether alternative prompting methods (e.g., prompt engineering or fine-tuning) could improve the autograder's reliability. Another major limitation is that the evaluated responses were generated by LLMs themselves, not by real students. Thus, further studies using authentic student-generated worked solutions should validate these findings. Third, our subject scope is fairly limited since we focus exclusively on college algebra. Future work should explore other mathematical topics or entirely different disciplines (e.g., physics, chemistry) to assess the generalizability of LLM-based autograding. Additionally, due to variability in LLM-generated responses, particularly in structure and notation, there were formatting inconsistencies. This may have an impact on grading consistency. Finally, while this study compares three LLMs from OpenAI, broader evaluations that include models from other families (e.g., Claude, Gemini, or Llama) could yield further insight into which models are best suited for educational autograding tasks.

4. CONCLUSION

In this study, we leverage an existing dataset with manually graded worked solution responses to 20 college algebra questions as the ground truth for evaluating LLM-based autograding. We develop a grading pipeline using three OpenAI models—GPT-4o, GPT-4.1-mini, and GPT-4.1-nano—and compare their output correctness labels to human-provided labels. Each model is evaluated under two conditions: standard prompting and self-consistency. Results show that GPT-4.1-mini with self-consistency achieves the highest agreement with human grading (94.47%), followed by GPT-4.1-nano (93.07%) and GPT-4o (91.93%). Accuracy modestly improves with self-consistency across all models. While this accuracy approaches that which would be usable in real-world scenarios, it is notable that these accuracy rates closely align with GPT-4o's previously reported performance on algebra tasks, especially given the autograder is told the ground truth correct answer. Further methods to mitigate grading inaccuracies should be explored, such as prompt iteration, model selection, and post-processing techniques. Since perfect agreement is not achieved, educators and researchers should exercise caution when relying solely on LLM-based autograder outputs until a method for perfect accuracy has been introduced.

5. REFERENCES

[1] U. Alkafaween, I. Albluwi, and P. Denny. Automating autograding: Large language models as test suite generators for introductory programming. *Journal of Computer Assisted Learning*, 41(1):e13100, 2025.

[2] A. Condor and Z. Pardos. Explainable automatic grading with neural additive models. In *International Conference on Artificial Intelligence in Education*, pages 18–31. Springer, 2024.

[3] T. N. B. Duong and C. Y. Meng. Automatic grading of short answers using large language models in software engineering courses. In *2024 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–10. IEEE, 2024.

[4] R. Ferreira Mello, C. Pereira Junior, L. Rodrigues, F. D. Pereira, L. Cabral, N. Costa, G. Ramalho, and D. Gasevic. Automatic short answer grading in the llm era: Does gpt-4 with prompt engineering beat traditional models? In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 93–103, 2025.

[5] N. T. Heffernan and C. L. Heffernan. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24:470–497, 2014.

[6] J. Higton, S. Leonardi, N. Richards, A. Choudhoury, N. Sofroniou, and D. Owen. Teacher workload survey 2016. 2017.

Table 2: Final agreement accuracy (%) between human grading and each model's output across all 20 questions. 'SC' (Self-Consistency) = majority vote from 10 completions; 'Std' (Standard) = single prompt grading without self-consistency.

Question	GPT-4o Std	GPT-4o SC	4.1-mini Std	4.1-mini SC	4.1-nano Std	4.1-nano SC
Q1	99.22	100.00	99.33	99.33	97.33	98.11
Q2	97.89	98.22	99.22	99.11	97.67	98.67
Q3	98.11	99.33	98.89	98.78	99.22	99.33
Q4	91.67	92.89	94.33	94.56	91.00	92.78
Q5	82.89	82.11	95.67	96.56	89.44	93.22
Q6	67.11	67.22	78.78	78.67	93.22	92.89
Q7	97.33	97.67	99.78	99.78	98.78	99.67
Q8	97.11	98.89	98.56	98.67	98.44	98.78
Q9	89.78	90.00	90.00	90.00	89.67	90.11
Q10	98.22	98.78	99.00	99.22	95.33	95.11
Q11	96.56	97.00	97.22	97.22	96.22	96.44
Q12	92.22	93.78	92.33	92.22	93.33	93.67
Q13	74.67	77.56	93.56	93.33	95.22	96.56
Q14	93.67	96.89	99.11	99.11	89.56	89.78
Q15	97.22	97.78	98.78	98.89	98.44	98.56
Q16	87.22	88.22	90.56	89.78	76.33	75.89
Q17	88.11	89.11	91.11	90.78	89.44	89.33
Q18	96.89	98.44	98.89	99.00	97.56	98.22
Q19	86.44	90.11	88.89	89.44	81.44	81.78
Q20	80.44	84.56	84.33	84.89	77.56	82.56
Overall	**90.64**	**91.93**	**94.42**	**94.47**	**92.26**	**93.07**

Table 3: Percent correct per question, as assigned by human raters and each model. 'SC' (Self-Consistency) = majority vote from 10 completions; 'Std' (Standard) = single prompt grading without self-consistency.

Question	Human-Graded %	GPT-4o Std	GPT-4o SC	4.1-mini Std	4.1-mini SC	4.1-nano Std	4.1-nano SC
Q1	57.67	56.89	57.67	57.00	57.00	55.00	55.78
Q2	51.33	52.78	52.22	51.67	51.78	50.33	51.11
Q3	91.56	90.56	91.78	91.11	91.00	91.67	91.78
Q4	43.67	43.56	43.00	43.78	44.00	40.89	42.44
Q5	44.78	43.00	37.78	42.89	43.56	38.00	41.78
Q6	16.78	44.33	45.11	33.11	33.00	20.22	20.33
Q7	6.00	8.22	7.89	6.00	6.00	7.22	6.33
Q8	85.11	83.56	85.56	85.00	85.11	84.89	85.22
Q9	89.56	98.89	99.11	99.11	99.11	98.56	99.22
Q10	72.89	71.56	72.11	72.11	72.33	68.89	68.44
Q11	69.67	71.56	72.00	71.33	71.11	73.00	72.78
Q12	19.78	18.89	17.33	19.22	18.89	16.67	17.22
Q13	13.22	36.33	33.44	18.78	19.44	15.78	14.22
Q14	14.78	20.22	17.00	14.56	14.33	24.78	24.56
Q15	10.67	9.22	10.44	10.33	10.22	10.67	10.11
Q16	30.89	38.78	36.67	30.56	30.22	15.22	14.56
Q17	65.89	75.11	76.11	73.67	74.00	75.11	75.89
Q18	16.22	15.78	16.44	16.89	16.78	18.22	17.56
Q19	37.56	38.67	41.67	37.33	37.22	34.33	36.22
Q20	41.44	39.44	43.56	33.78	33.89	38.78	41.56
Overall	**43.97**	**47.87**	**47.84**	**45.41**	**45.45**	**43.91**	**44.36**

[7] M. Kelly. A study of written communication: Showing your steps. 2007.

[8] K. R. Koedinger, J. R. Anderson, W. H. Hadley, and M. A. Mark. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8:30–43, 1997.

[9] M. Kostic, H. F. Witschel, K. Hinkelmann, and M. Spahic-Bogdanovic. Llms in automated essay evaluation: A case study. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 143–147, 2024.

[10] Y. Kwak and Z. A. Pardos. Bridging large language model disparities: Skill tagging of multilingual educational content. *British Journal of Educational Technology*, 55(5):2039–2057, 2024.

[11] Y. Liu, S. Bhandari, and Z. A. Pardos. Leveraging llm respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, 56(3):1028–1052, 2025.

[12] K. Manikani, R. Chapaneri, D. Shetty, and D. Shah. Sql autograder: Web-based llm-powered autograder for assessment of sql queries. *International Journal of Artificial Intelligence in Education*, pages 1–31, 2025.

[13] N. Matzakos, S. Doukakis, and M. Moundridou. Learning mathematics with large language models: A comparative study with computer algebra systems and other tools. *International Journal of Emerging Technologies in Learning (iJET)*, 18(20):51–71, 2023.

[14] J. C. Paiva, J. P. Leal, and Á. Figueira. Automated assessment in computer science education: A state-of-the-art review. *ACM Transactions on Computing Education (TOCE)*, 22(3):1–40, 2022.

[15] Z. A. Pardos and S. Bhandari. Chatgpt-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *Plos one*, 19(5):e0304013, 2024.

[16] Z. A. Pardos, M. Tang, I. Anastasopoulos, S. K. Sheel, and E. Zhang. Oatutor: An open-source adaptive tutoring system and curated content library for learning sciences research. In *Proceedings of the 2023 chi conference on human factors in computing systems*, pages 1–17, 2023.

[17] J. Sukkarieh, S. Pulman, and N. Raikes. Automarking: using computational linguistics to score short ,free-text responses. 2003.

[18] J. Z. Sukkarieh and J. Blackmore. c-rater: Automatic content scoring for short constructed responses. In H. C. Lane and H. W. Guesgen, editors, *Proceedings of the Twenty-Second International Conference of the Florida Artificial Intelligence Research Society*, pages 290–295, Sanibel Island, Florida, USA, 2009. AAAI Press.

[19] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

[20] X. Wei. Evaluating chatgpt-4 and chatgpt-4o: Performance insights from naep mathematics problem solving. *Frontiers in Education*, 9:1452570, Sept. 2024.