

A Preliminary Empirical Study of Large Language Models for Grading Debugging Problems in Programming Education

Qixiang Pang, Linrui Zhang, Belinda Copus

University of Central Missouri
{qpang, lzhang, copus}@ucmo.edu

Shan Du

University of British Columbia
shan.du@ubc.ca

Abstract

Debugging problems are essential for assessing code semantic understanding, yet grading these heterogeneous responses is labor-intensive and prone to inconsistency. This poster presents a preliminary empirical study evaluating five Large Language Models (LLMs)—ChatGPT, Claude, Gemini, Grok, and DeepSeek—as automated grading assistants. Using authentic student submissions from two university Python courses, we compare LLM performance against rubric-based human benchmarks using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Pearson correlation. Results show all models achieve strong correlation ($\gamma > 0.90$), indicating reliable preservation of student rankings. While variance in partially correct solutions persists, the findings suggest LLMs are effective for preliminary scoring and triage, provided human oversight is maintained to mitigate occasional grading deviations.

Introduction and Motivation

Debugging tasks assess higher-order skills like logical reasoning and error diagnosis, which strongly correlate with programming proficiency (Ahmadzadeh, Elliman, and Higgins 2007) (Alqadi 2024). However, grading open-ended student responses is labor-intensive and difficult to automate via traditional unit tests (Keuning, Jeuring, and Heeren 2018) (Alqadi 2024). This manual burden often results in delayed feedback and grading inconsistency in large-scale courses.

Recent Large Language Models (LLMs) offer a potential solution through demonstrated code-reasoning capabilities (Chen et al. 2021) (Chu et al. 2025). Yet empirical evidence of their performance on real-world debugging remains limited, particularly for grading tasks in programming education (Jukiewicz 2025) (Chuang and Chang 2024) (Impey et al. 2025) (Cofini et al. 2025). This preliminary empirical study investigates the feasibility of using multiple LLMs (Anthropic 2024) (DeepSeek-AI 2024) (Anil et al. 2023) (OpenAI 2023) (xAI 2025) as grading assistants by quantitatively evaluating their alignment with human instructor benchmarks.

Copyright © 2026 by the authors. Open access article published under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Methodology and Experimental Design

Dataset & Ethics

- Source: Authentic student submissions from two university Python courses (Foundational & Advanced).
- Scope: Debugging Problems - DBP1 (9 bugs) and DBP2 (6 bugs).
- Ethics: Study conducted with University of Central Missouri Institutional Review Board approval (Protocol No. 2816). All student submissions were anonymized prior to analysis.

Tools

- LLMs: ChatGPT (GPT-5.3), Claude (Sonnet 4.6), Gemini (3 Flash), Grok (4 Auto), and DeepSeek (V3)
- Setup: Standardized zero-shot prompts using free-tier primary chat interfaces to evaluate “out-of-the-box” performance.

Grading Framework & Rubric

A standardized rubric was applied to ensure objective evaluation across both human and AI graders:

- Scoring: 1 pt per bug identified, 1 per bug fixed
- Penalty: -0.5 pt if a ‘fix’ breaks a correct part
- Min score: 0; Max score: $2 \times N$ (the number of bugs)

Prompt Architecture

- Role: You are an expert computer science instructor.
- Task: Grade the student response to a debugging problem based on the provided rubric.
- Output: Provide a numerical score and a brief rationale for the grade.

Validation & Rigor

- Expert baseline: All submissions were graded by the lead instructor as a consistent reference.
- Rubric Alignment: LLMs used the identical fine-grained rubric as instructors to isolate reasoning performance from prompt variance.

Table 1: Agreement analysis between LLM-generated scores and human grading – DBP1

Model	MAE	RMSE	StdError	Correlation
ChatGPT	1.750	2.369	1.611	0.962
Claude	0.987	1.329	0.968	0.987
DeepSeek	1.026	1.259	1.132	0.983
Gemini	0.724	0.916	0.802	0.992
Grok	1.244	1.664	1.367	0.973

Table 2: Agreement analysis for DBP2

Model	MAE	RMSE	StdError	Correlation
ChatGPT	0.96	1.23	1.229	0.940
Claude	1.429	1.626	0.862	0.971
DeepSeek	1.061	1.464	1.265	0.955
Gemini	0.96	1.175	0.907	0.971
Grok	1.141	1.416	1.183	0.958

- **Self-Consistency Protocol:** To reduce variability from LLM non-determinism, each response was graded in triplicate with a fresh chat session; reported scores represent the mean.

Evaluation Metrics

The performance is assessed against human benchmarks:

- **Accuracy:** Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
- **Consistency:** Standard deviation of grading error.
- **Alignment:** Pearson correlation coefficient. A correlation value close to 1 indicates strong agreement between the LLM and the human grader.

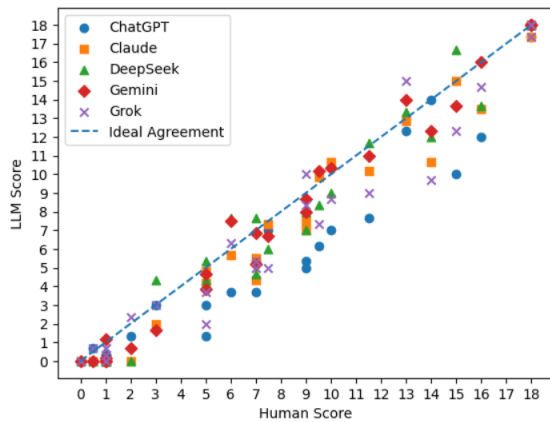


Figure 1: Unified Score Correlation Scatter Plot (DBP1)

Results Analysis and Discussion

Functional Capability Analysis

All evaluated LLMs were able to analyze student submissions and generate numerical scores and brief explanations.

Grading Accuracy and Agreement Analysis

The quantitative results are presented in Tables 1 and 2 and Figure 1. The key findings are summarized below:

- **Strong Alignment:** Although the dataset is small (~ 60 students), the evaluated models exhibit strong correlations ($\gamma > 0.90$), indicating reliable preservation of student rankings.
- **Partial Credit Variance:** Most scores cluster closely along the ideal-agreement diagonal, with the highest accuracy observed for extreme cases (i.e., near-perfect or clearly failing submissions). Deviations occur more frequently for non-canonical or partially correct solutions, suggesting that LLM outputs are not fully interchangeable.
- **Stricter Interpretation:** The models tended to slightly under-score submissions, often penalizing minor syntactic or typo errors that human instructors—who prioritize student intent—might overlook (Figure 1).
- **Need for Human Oversight:** Occasional large deviations emphasize that human review remains essential to ensure overall assessment integrity.
- **Comparative Performance:** No single tool consistently outperformed the others across all aspects.
- **Role as Grading Assistants:** While not yet suitable for fully autonomous grading, LLMs can effectively support preliminary scoring, reduce instructor workload, and help triage outlier submissions for targeted human review.

Conclusion and Future Work

This preliminary empirical study demonstrates that LLMs can serve as effective grading assistants for preliminary scoring in programming education. While all models achieved strong correlation with human benchmarks and preserved student rankings, occasional large deviations in non-canonical or partially correct solutions highlight the need for continued human oversight.

Future Work

- **Scale & Diversity:** Expand the dataset to larger cohorts and additional languages (e.g., C++, Java).
- **Grader Rigor:** Conduct multi-expert studies to assess inter-rater reliability and reduce grading subjectivity.
- **Technical Determinism:** Move to API-based testing to minimize probabilistic variance.

Acknowledgement

Funding provided by a UCM Alumni Foundation Opportunity Grant Program.

References

- Ahmadzadeh, M.; Elliman, D.; and Higgins, C. 2007. The impact of improving debugging skill on programming ability. *Innovation in Teaching and Learning in Information and Computer Sciences* 6(4):72–87.
- Alqadi, B. S. 2024. Enhancing novice programmers' debugging skills through systematic education: A comparative study. *IEEE Access* 12:181192–181204.
- Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.
- Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. D. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Chu, Z.; Wang, S.; Xie, J.; Zhu, T.; Yan, Y.; Ye, J.; Zhong, A.; Hu, X.; Liang, J.; Yu, P. S.; et al. 2025. Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733* 2.
- Chuang, Y.-T., and Chang, H.-Y. 2024. Analyzing novice and competent programmers' problem-solving behaviors using an automated evaluation system. *Science of Computer Programming* 237:103138.
- Cofini, V.; Jobe, T.; Letteri, I.; and Vittorini, P. 2025. Preliminary evaluation of an llm-based system for grading and providing feedback on short-text answers in data science exercises. In *International Conference in Methodologies and intelligent Systems for Technology Enhanced Learning*, 95–106. Springer.
- DeepSeek-AI. 2024. Deepseek-v3 technical report.
- Impey, C.; Wenger, M.; Garuda, N.; Golchin, S.; and Stamer, S. 2025. Using large language models for automated grading of student writing about science. *International Journal of Artificial Intelligence in Education* 35(4):1825–1859.
- Jukiewicz, M. 2025. A systematic comparison of large language models for automated assignment assessment in programming education: Exploring the importance of architecture and vendor. *arXiv preprint arXiv:2509.26483*.
- Keuning, H.; Jeurig, J.; and Heeren, B. 2018. A systematic literature review of automated feedback generation for programming exercises. *ACM Transactions on Computing Education (TOCE)* 19(1):1–43.
- OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- xAI. 2025. Grok 3. <https://docs.x.ai/docs/models/grok-3>. Official model documentation.