Automatic Grading of Student Work Using Simulated Rubric-Based Data and GenAI Models

Yiyao Yang[1], Yasemin Gulbahar[2]

{yy3555[1], yg2918[2]}@tc.columbia.edu

Teachers College, Columbia University, New York, NY, USA

Abstract

Grading assessment in data science faces challenges related to scalability, consistency, and fairness. Synthetic dataset and GenAI enable us to simulate realistic code samples and automatically evaluate using rubric-driven systems. The research proposes an automatic grading system for generated Python code samples and explores GenAI grading reliability through human-AI comparison.

Keywords

Automated Assessment, Generative AI, Rubric-Based Grading, Inter-Rater Reliability, Human-AI Assessment Comparison, Hybrid Assessment Systems

1 Introduction

Digital technologies have significantly influenced educational assessment, leading to a growing interest in the automatic grading of student work. Traditionally, evaluating student submissions, especially coding and open-ended responses, has been labor-intensive and often subjective for educators. Manual grading tends to be inconsistent, biased, and a considerable time investment, particularly in large-enrollment courses. Automatic grading systems, primarily designed to address these challenges, aim to enhance efficiency, consistency, and objectivity in the assessment process, streamlining educational workflows and providing timely student feedback. Automatic grading is subject to the nature of text, code, and evaluating program codes might use different approaches for automated grading, where human cognition and human experience shed light on the process, and we will also assess the reliability of GenAI coding grading through a series of comparisons between human and GenAI evaluations.

Automated Grading of Student Work

The development of automatic grading systems is deeply rooted in advancements in artificial intelligence (AI), particularly in the fields of Natural Language Processing (NLP) and Machine Learning (ML) (V. Nikhil et al., 2025; Kumar et al., 2024; Alqahtani et al., 2023). While early systems relied on rule-based approaches and statistical models to analyze surface-level features such as grammar, spelling, and word count, more recently, deep learning techniques, including models like BERT and RoBERTa, have enabled more sophisticated semantic analysis, allowing systems to better understand the content and coherence of written responses (Ayaan & Ng, 2025; Faseeh et al., 2024; Bayer et al., 2022). The emergence of Large Language Models (LLMs) further promises to revolutionize this domain by offering enhanced capabilities in processing complex sentences, establishing relationships between text elements, and even discerning the intent behind writing.

The benefits of automatic grading systems are substantial and widely discussed in the literature. Foremost among these is the significant reduction in grading time and workload for instructors, freeing instructors to focus on more personalized student interactions and curriculum development (Messer et al., 2025). Automated systems also provide unparalleled consistency and objectivity, applying uniform criteria across all submissions and minimizing human biases that can inadvertently affect grades (Vetrivel et al., 2025). These advantages, which incl • ude scalability, rapid feedback, and reduced bias, have been extensively synthesized in recent systematic reviews (Marcelo Guerra Hahn et al., 2021), underscoring their central role in modern online and large-scale learning environments.

Crucially, these systems offer immediate feedback to students, a feature essential for reinforcing learning and enabling prompt self-correction. Previous research suggests that these practices lead to better long-term retention and skill development (Demszky et al., 2023; Wilson et al., 2021). Furthermore, the scalability of automatic grading makes it a crucial tool for large classes and online learning environments (Lin et al., 2024; Messer et al., 2023).

Despite these advantages, previous studies also highlight several challenges and limitations associated with automatic grading (Padó et al., 2023). A primary concern is the inherent difficulty in replicating the nuanced judgment of human graders, especially concerning subjective aspects of writing such as creativity, originality, critical thinking, and subtle rhetorical devices like humor or irony. Critics argue that over-reliance on algorithms might inadvertently incentivize students to adopt formulaic writing styles that appeal to the machine's evaluative criteria rather than foster genuine intellectual development. Moreover, while automated systems excel at quantitative feedback, they often struggle to provide the rich, qualitative, and constructive suggestions that human graders can offer, which are essential for deep learning and improvement (Figueras et al., 2025; Bato & Pomperada, 2025; Fagbohun et al., 2024).

Ethical considerations and student perception are also critical in the discourse (Farazouli, 2024). Concerns about algorithmic bias, where systems might inadvertently perpetuate or amplify existing societal biases in their training data, are frequently raised. It can lead to unfair or inequitable assessments for specific demographic groups (Baker & Hawn, 2021; Kordzadeh & Ghasemaghaei, 2021). Besides, some studies indicate that students may perceive automated grading as less fair or trustworthy than human evaluation, potentially undermining their confidence in the integrity of the scoring process (Vetrivel et al., 2025; Chai et al., 2024). Addressing these issues requires transparent system design, rigorous validation, and, in many cases, a hybrid approach that integrates human oversight (Kern et al., 2022). Looking ahead, automatic grading is moving towards advanced hybrid systems that combine the strengths of AI with human insight. Researchers focus on how AI can help evaluate more complex skills and adapt to individual learning needs, emphasizing personalized educational paths. A key development area involves integrating these grading tools with existing Learning Management Systems (LMS) to create smooth and effective educational environments. By thoroughly capturing human behavior, these systems can continuously learn from human input, leading to a more efficient, objective, and supportive learning experience for students through detailed and instant feedback, all while carefully managing the associated complexities.

Hence, with the increasing integration of data science and coding instruction into educational curricula, scalable and equitable assessment of student-generated code is becoming prominent. Assignments that involve code development, data analysis, and interpretation pose challenges for large-scale instruction due to the complexity and subjectivity in grading. Although criteria provide a standardized basis for assessment, manual scoring is time-consuming and inconsistent. It emphasizes the requirement for scalable, reliable, and pedagogy-matching grading solutions for education. GenAI and LLMs could develop the perception of code, whereas synthetic student data methods enable the recreation of student submissions in controlled environments, negating student privacy concerns. Moreover, automated grading systems can behave differently depending on the nature of the text and the machine learning approach used to evaluate. Grounded in prior studies, our research proposes an auto-grading framework that integrates GenAI with synthetic data and evaluates the reliability of automated grading in data science by examining differences between human and AI grading outcomes, intending to enhance the efficiency and effectiveness of coding assessment practices.

Hence, our study aims to address the following research questions (RQs):
RQ1: How effectively do rubric-based GenAI grading outputs align with human ratings across all programming code samples regarding total scoring consistency?
RQ2: What are the methodological strengths and practical limitations of GenAI-based rubric grading systems with respect to reliability, scalability, and fairness in programming assessment, and how can educators be guided to integrate such systems effectively into grading practices?

2 Research Methodology

Based on academic and behavioral characteristics, we used GPT-5 to generate 100 synthetic student profiles as the original dataset. These profiles include features such as the number of hours studied per day, lecture attendance rate in percentage, average quiz score, assignment score, final exam score, class participation level, number of hours of internet usage per day, and average number of sleeping hours per day. Each profile was uniquely identified by a distinct student ID.

Based on the synthetic education dataset, we have defined a regression task (supervised task) for further code samples simulation and human-AI grading comparison.

TASK: The Regression Task (Supervised Task)

This task is a supervised regression problem that aims to predict a continuous numerical value, the final exam outcome, based on various student behavioral and academic features. The goal is not just to make accurate predictions, but to build a model that can be easily understood, allowing us to identify which specific student behaviors and indicators have the most significant impact on student final scores. It is different from a classification task, which would predict a discrete category like "pass" or "fail." Instead, the response variable final exam outcome is a numerical variable, such as "92.5". When generating the AI code, you'll need to consider several key details:

Dataset: The input data is in a CSV file named synthetic_education_data.csv. The task includes loading, preprocessing, and analyzing this data.

Response Variable: The column representing the final exam outcome is the response variable we want to predict. You will need to identify this column in the dataset.

Feature Variables: The other columns containing the student behavioral and academic information are the features or independent variables. These will be used to train the model.

Model: Since the goal is interpretability, a good starting point would be models like linear regression, decision trees, or random forests. While more complex models like neural networks might be more accurate, they are often less transparent about predictions.

Evaluation Metrics: The code should use regression-specific metrics to evaluate the model's performance. Common metrics include: Mean Squared Error (MSE): Measures the average of the squared differences between the predicted and actual values. A lower MSE indicates more accurate predictions. Root Mean Squared Error (RMSE): The square root of the MSE, expressed in the same units as the response variable, making it easier to interpret. R-squared (R^2): Indicates how well the model's predictions fit the actual data, ranging in $(-\infty, 1]$, with values closer to 1 indicating that the model explains more of the variability in the outcome. A negative R^2 suggests that the model performs worse than a simple mean predictor.

Based on the task descriptions, a total of 25 synthetic Python code samples were generated by GPT-5, differing in syntax, formatting, and comments. The prompt we used was: "Could you generate 25 distinct Python solutions for the Regression Task by simulating 25 different students who have diverse levels of expertise and performance in coding skills, educational data analytics, and data mining methods? "

Evaluation Rubric: Each code sample was assessed using a detailed analytic rubric with 20 evaluation criteria to assess a broad range of coding competencies aligned with learning outcomes in data science education: comments used, number of lines, number of libraries, number of variables, number of visualizations, error-free, clear structure, organized, data cleaning, outlier checking, optimized solution, code complexity, interpretation quality, code readability, predictable variable names, visual readability, code reusability, data accessibility, resource efficiency, and overall quality.

Each code sample received a complete rubric-based score evaluation, and the total score for each code sample was computed by summing the 20 criteria. Based on the 20 criteria, scores ranged from 1 to 5 for each criterion, and with a total score out of 100 for each code sample. GenAI grading was conducted using GPT-5 via OpenAI, guided by the 20 criteria grading scheme with human cognition to enhance reliability and evaluation alignment. To compare

GenAI and human grading results, a human grader evaluated those generated 25 code samples based on the same rubric with 20 evaluation criteria. To better understand the score distribution of GenAI versus Human grading result comparison, we visualized total scores using a multi-line radar and a scatter plot. These visualizations revealed a broad distribution of grades, supporting the diversity in the code samples simulation. Inter-rater agreement between human and AI grading results was analyzed using intraclass correlation coefficients (ICC), Cohen's Kappa, and Cronbach's α to evaluate reliability and consistency between human and AI evaluators. The research establishes a reproducible framework for rubric-based automatic code grading, incorporates realistic grading variability, and evaluates the reliability of AI-based scoring, contributing to the development of hybrid assessment systems that balance efficiency with instructional quality in STEM education.

3 Data Analysis & Results

All data analyses and Python code are stored in a private GitHub repository (*2025-NCME-AIME-Con-Yiyao-Yang*; Yang, 2025), available upon request. The summary statistics of rubric-based GenAI versus human grading across 25 submitted code samples of the regression task (Table 1) indicate a generally consistent trend, but a systematically lower scoring pattern by GenAI. Among all 25 different code samples, the mean score of GenAI grading (81.00) is lower than that of human grading (86.91), with median scores of 80.85 and 86.53, respectively. The score ranges show that both GenAI (74.02 – 91.88) and human (73.47 – 95.79) raters have captured the full spectrum of code quality, although human gradings exhibit higher variability (range = 22.32, IQR = 4.53) compared to GenAI ratings (range = 17.86, IQR = 2.63).

Table 1

Summary Statistics of Rubric-Based GenAI versus Human Grading of 25 Generated Code Samples

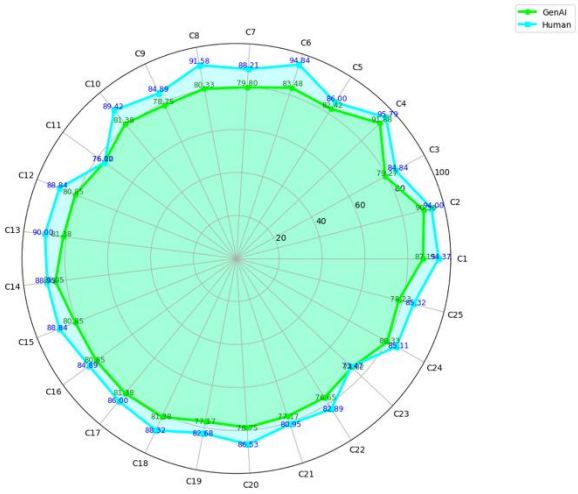
	GenAI	Human
Mean	81.00	86.91
Median	80.85	86.53
Range	17.86	22.32
IQR	2.63	4.53
Max	91.88	95.79
Min	74.02	73.47

Note. All values are reported to two decimal places.

Examining the score distributions, GenAI exhibits a narrower range than human raters, reflecting compressed score variability. Across the 25 code samples, human grading generally assigned higher scores than GenAI grading, with only two tasks (Code Sample # 11 and # 23) receiving similar grades from both evaluators. Visual comparisons, including the multi-line radar plot (Figure 1) and the scatter plot (Figure 2), further confirm that while GenAI grading follows the overall performance trend, it predominantly underestimates scores relative to human evaluation.

Figure 1

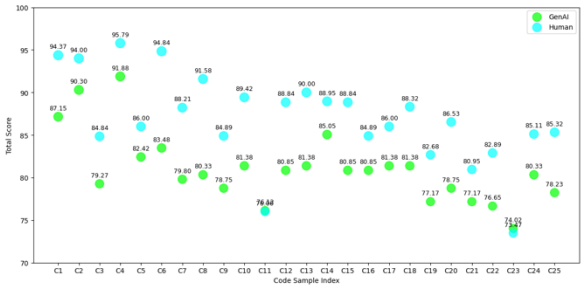
GenAI versus Human Grading: Multi-Line Radar



Note. "C" denotes "Code Sample".

Figure 2

GenAI versus Human Grading: Scatter Plot



Note. "C" denotes "Code Sample".

Overall, GenAI moderately captures relative performance trends but tends to underestimate scores compared to human grading. It provides a useful foundation for consistency checks and supports the development of semi-automated grading workflows. The evaluation of rubric-based GenAI grading systems highlights both methodological strengths and practical constraints. In terms of reliability, the intraclass correlation coefficient (ICC (2,1) ≈ 0.51)

indicates moderate consistency between GenAI and human grading results, suggesting that GenAI reasonably captures relative performance trends, though absolute score alignment remains limited. The low Cohen's Kappa (≈ 0.01) highlights minimal exact agreement on total scores, indicating that categorical consistency between GenAI and human grading outcomes is extremely low. In contrast, the high Cronbach's α (≈ 0.90) demonstrates strong internal consistency across rubric criteria, indicating coherent scoring patterns within the multi-criteria evaluation framework.

Regarding scalability, GenAI efficiently processes large volumes of code samples, producing rapid and reproducible scores without the temporal and cognitive variability of human raters, as a key advantage in large programming courses. For instructional alignment, the moderate total-score reliability suggests that GenAI is best used as a complement, rather than a replacement for human judgment. Educators may use GenAI for first-pass grading, trend identification, and efficient formative feedback, while maintaining human oversight for final scoring decisions. Iterative refinement of rubric prompts can further improve alignment, enabling a collaborative hybrid human-AI grading workflow.

4 Conclusion

Taken together, the findings underscore the necessity of re-evaluating grading practices in programming education. Previous research indicates that human graders often show considerable variability in scoring the same programming assignments, with both inter-rater disagreement and intra-rater inconsistency, suggesting that the notion of a "gold standard" in human grading may be inherently flawed (Messer et al., 2025). A shared rubric alone is insufficient to guarantee consistent evaluation, and additional measures such as assessor training and alternative grading practices are needed to improve reliability. In this context, our research further demonstrates that rubric-based GenAI grading offers a practical complement: While GenAI auto-grading cannot replace human judgment, it can efficiently perform an initial assessment of coding assignments, after which human evaluators can review and adjust the grading results. Such a collaborative human-AI workflow leverages the efficiency of automated scoring while preserving the refined judgment of human graders, providing an effective approach to scalable, semi-automated hybrid assessment of programming tasks. By combining the efficiency of GenAI with the experience and judgement of human evaluators, we can ensure assessment fairness while giving educators the space to guide students meaningfully in data science education, encouraging and inspiring them to grow as passionate programmers and to blossom as inquisitive learners and reflective thinkers, guided by curiosity, courage, and the joy of discovery.

5 References

Alqahtani, T., Badreldin, H. A., Alrashed, M., Alshaya, A. I., Alghamdi, S. S., bin Saleh, K., Alowais, S. A., Alshaya, O. A., Rahman, I., Al Yami, M. S., & Albekairy, A. M. (2023). The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy*,*19*(8). https://doi.org/10.1016/j.sapharm.2023.05.016

Ayaan, A., & Ng, K.-W. (2025). Automated Grading using Natural Language Processing and Semantic Analysis. *MethodsX*, *14*, 103395–103395. https://doi.org/10.1016/j.mex.2025.103395

Baker, R. S., & Hawn, A. (2021). Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*, *32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Bato, B. E., & Pomperada, J. R. (2025). Automated grading system with student performance analytics. *Technium Romanian Journal of Applied Sciences and Technology*, *30*, 58–75. https://doi.org/10.47577/technium.v30i.12871

Bayer, M., Kaufhold, M.-A., & Reuter, C. (2022). A Survey on Data Augmentation for Text Classification. *ACM Computing Surveys*, 3544558. https://doi.org/10.1145/3544558

Chai, F., Ma, J., Wang, Y., Zhu, J., & Han, T. (2024). Grading by AI makes me feel fairer? How different evaluators affect college students' perception of fairness. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1221177

Dorottya Demszky, Jing, L., Hill, H. C., Jurafsky, D., & Piech, C. (2023). Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a

large-scale online course. *Educational Evaluation and Policy Analysis*, *46*(3), 016237372311692-016237372311692. https://doi.org/10.3102/01623737231169270

Fagbohun, O., Iduwe, N. P., Abdullahi, M., Ifaturoti, A., & Nwanna, O. M. (2024). Beyond Traditional Assessment: Exploring the Impact of Large Language Models on Grading Practices. *Journal of Artificial Intelligence, Machine Learning and Data Science*, *2*(1), 1–8. https://doi.org/10.51219/jaimld/oluwole-fagbohun/19

Farazouli, A. (2024). Automation and Assessment: Exploring Ethical Issues of Automated Grading Systems from a Relational Ethics Approach. *Postdigital Science and Education*, 209–226. https://doi.org/10.1007/978-3-031-58622-4_12

Faseeh, M., Jaleel, A., Iqbal, N., Ghani, A., Abdusalomov, A., Mehmood, A., & Cho, Y.-I. (2024). Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy. *Mathematics*, *12*(21), 3416. https://doi.org/10.3390/math12213416

Figueras, C., Farazouli, A., Cerratto Pargman, T., McGrath, C., & Rossitto, C. (2025). Promises and breakages of automated grading systems: a qualitative study in computer science education. *Education Inquiry*, 1–22. https://doi.org/10.1080/20004508.2025.2464996

Kern, C., Gerdon, F., Bach, R. L., Keusch, F., &. Kreuter, F. (2022). Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. *Patterns*, *3*(10), 100591. https://doi.org/10.1016/j.patter.2022.100591

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: review, synthesis, and Future Research Directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Kumar, M. V., Vichitra Sivaji, B V V Padmavathi, N. Jothi, B. Muthulakshmi, & V R. Yasu Bharathi. (2024). *Natural Language Processing Techniques for Enhancing Automated Essay Scoring Systems*. 1–5. https://doi.org/10.1109/ic3tes62412.2024.10877454

Lin, Z., Yan, H., & Zhao, L. (2024). Exploring an effective automated grading model with reliability detection for large-scale online peer assessment. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.12970

Marcelo Guerra Hahn, Margarita, S., de-la-Fuente-Valentín, L., & Burgos, D. (2021). A Systematic Review of the Effects of Automatic Scoring and Automatic Feedback in Educational Settings. *IEEE Access*, *9*, 108190–108198. https://doi.org/10.1109/access.2021.3100890

Messer, M., Brown, N. C. C., Kölling, M., & Shi, M. (2023). Automated Grading and Feedback Tools for Programming Education: A Systematic Review. *ACM Transactions on Computing Education*, *24*(1), 3636515. https://doi.org/10.1145/3636515

Messer, M., Neil, Kölling, M., & Shi, M. (2025). How Consistent Are Humans When Grading Programming Assignments? *ACM Transactions on Computing Education*. https://doi.org/10.1145/3759256

Nikhil, V., Annamalai, R., & Jayapal, S. (2025). NLP-Driven Approaches to Automated Essay Grading and Feedback [Review of *NLP-Driven Approaches to Automated Essay Grading and Feedback*]. In T. Murugan, K. Periasamy, & A. M. Abirami (Eds.), *Adopting Artificial Intelligence Tools in Higher Education*. Taylor & Francis Group. https://doi.org/10.1201/9781003470304-5

Padó, U., Yunus Eryilmaz, & Kirschner, L. (2023). Short-Answer Grading for German: Addressing the Challenges. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00383-w

Vetrivel, S. C., Arun, V. P., Ambikapathi, R., & Saravanan, T. P. (2025). Automated Grading Systems: Enhancing Efficiency and Consistency in Student Assessments [Review of *Automated Grading Systems: Enhancing Efficiency and Consistency in Student Assessments*]. In T. Murugan, K. Periasamy, & A. M. Abirami (Eds.), *Adopting Artificial Intelligence Tools in Higher Education Student Assessment*. Taylor & Francis Group. https://doi.org/10.1201/9781003470304

Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education*, *168*, 104208. https://doi.org/10.1016/j.compedu.2021.104208