

Comparing Generative AI and Human Grading of Undergraduate Students' Personal Development Plans

Man Kin LEUNG¹, Sui Leung Walter FUNG², Weihang LIAO³, Yagiz TORCUK⁴

¹ Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China

² Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

³ Department of Data Science & Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong, China

⁴ Department of Mathematics, Technische Universität Dresden, Dresden, Germany

Abstract

The use of Generative AI (GenAI) in assignment grading has brought convenience to teachers but remains controversial due to concerns regarding its accuracy and consistency in providing nuanced grades and feedback. This study compares the scores of Personal Development Plans from 50 students in a Freshman Seminar course, with each plan graded by three human graders and three AI graders. Through quantitative and qualitative analyses, we examine the grading patterns of different GenAI models and evaluate the feasibility of AI-assisted grading.

Keywords: GenAI, assessment, automated grading, reflective writing

1. Introduction

Generative Artificial Intelligence (GenAI), based on Large Language Models (LLMs), shows potential in various aspects of education, including automated assessment (Chan & Hu, 2023). With comprehensive prompts, GenAI can evaluate students' written assessments efficiently, providing detailed scores and feedback. However, studies report inconsistencies and inaccuracies in GenAI-assisted grading compared to human graders (Jauhiainen & Garagorry Guerra, 2024; Usher, 2025). Considering these potential benefits and limitations, the authors have developed an AI grading system to assist in the grading of written assessments. The results from human graders and AI graders based on different LLM models are evaluated and compared in terms of consistency and similarity.

The objectives of this study are addressed through the following research questions:

- **RQ1:** To what extent do grade distributions by AI graders align with human graders?
- **RQ2:** What are the differences in grading styles between AI graders and human graders?

2. Methodology

This study is based on a Freshman Seminar course offered by the Faculty of Computer and Mathematical Sciences at The Hong Kong Polytechnic University. Students were required to submit Personal Development Plans reflecting their motivation for study, inspirations from the seminars, and career goals. Grading criteria were defined on five aspects with a total of 40 marks:

- **C1:** Self-assessment and reflection (10)
- **C2:** Clarity and specificity (10)
- **C3:** Addressing societal needs (5)
- **C4:** Actionable plan (10)
- **C5:** Structure and writing skills (5)

An AI grading system had been developed prior to its offering in Fall 2025. To provide a standard for the system, these criteria were further elaborated into rubrics describing the quality of work required to attain different points in each component. The system was also prompted to provide feedback under each component for each student, followed by a conclusive comment.

A convenience sampling of 50 submissions was selected by chronological order of submission, balancing the ratio of students from three departments. Each submission was graded independently by three teachers from different departments (the first three authors). It was also graded by three AI graders: **M1** (GPT-4.1-mini), **M2** (Grok-4-Fast), and **M3** (Claude-Haiku-4.5), deployed on Poe. Apart from the system prompt, the course introduction document and rubrics of the assignment were uploaded as a knowledge base for the models.

Studies in the literature have verified that the consistency of grading using the same model is high upon multiple trials (Jauhainen & Garagorry Guerra, 2024). However, inconsistency might be significant across different AI models (Wetzler et al., 2025). In this study, the distance between score distributions of the graders was analysed using Kullback-Leibler divergence (Belov & Armstrong, 2011). By comparing the results to those of human grading, the submissions with the greatest difference between the medians of human graders and AI graders were selected as cases for qualitative analysis to provide in-depth

explanation for the difference, based on the detailed feedback provided by the AI grading system.

3. Results and Discussion

We first present the quantitative analysis of the overall score distribution (3.1), followed by a qualitative analysis of a few cases with high inconsistency (3.3).

3.1 Descriptive results of the overall grading

The descriptive statistics for the scores assigned by three human graders (H1 to H3) and three AI graders (M1 to M3) are presented in Table 1. In general, human graders assign higher mean scores (30.22 to 31.76) compared to the AI graders (26.80 to 29.78). Notably, human graders exhibit lower score variability than AI graders, as indicated by the standard deviations. The AI graders, particularly M2 and M3, utilise a broader scoring range, assigning lower minimum scores (14) than human graders (21). Additionally, the descriptive figure suggests greater consistency among human graders. Their medians (30.5, 32, 31) and 75% quartiles (34, 34, 33.75) are closely clustered. In contrast, AI graders exhibit larger differences in their central tendencies and spreads. This suggests that AI graders have a stricter grading tendency and less consensus among themselves compared to human graders.

Table 1: Descriptive Statistics for all Graders (out of 40 marks)

Grader	Mean	Std	Min	25%	50%	75%	Max
H1	30.22	4.82	21	26.25	30.5	34	38
H2	31.76	3.16	24	31	32	34	38
H3	30.62	3.87	21	28	31	33.75	38
M1	29.78	4.25	18	27	30	33	37
M2	27.34	5.31	14	24	28	30.75	37
M3	26.80	5.83	14	23	26	31.75	38

3.2 Consistency and comparison among graders

To further investigate the consistency among graders, we analyse the divergence between score distributions and the confidence intervals for the population mean. The symmetric Kullback-Leibler divergence matrix quantifies the dissimilarity between score distributions of each pair of graders, presented in Table 2. A larger number indicates more dissimilarity, while 0 indicates an identical distribution. The divergence values among human graders are relatively low (all smaller than 1), confirming their scoring alignment. Conversely, higher divergence is observed between certain AI graders (e.g., M1-M3: 1.39) and between some hu-

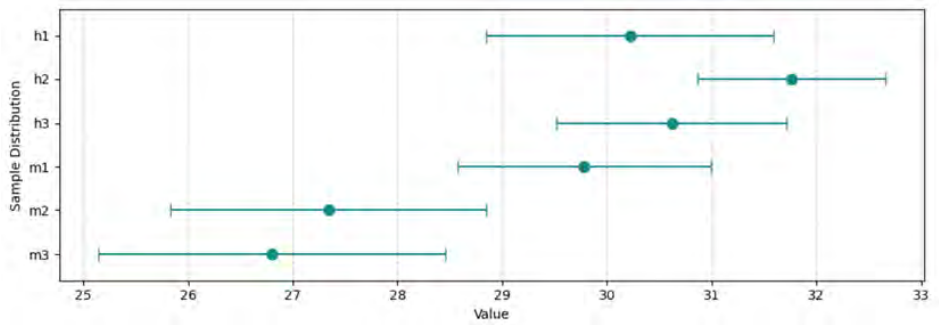
man-AI pairs (e.g., H2-M3: 1.52), indicating significant differences in their grading patterns. Interestingly, the distribution of M1 is closer to H1-H3 rather than M2 and M3, suggesting a similar grading profile with human graders.

Table 2: Symmetric KL Divergence Matrix of Results by all Graders

	H1	H2	H3	M1	M2	M3
H1	0	0.87	0.15	0.69	0.13	0.23
H2		0	0.29	0.21	1.22	1.52
H3			0	0.24	0.24	0.42
M1				0	0.95	1.39
M2					0	0.10
M3						0

The 95% confidence intervals for the estimated population mean are shown in Figure 1. The intervals for human graders show substantial overlap, implying they are likely estimating a similar mean score for the student population. The intervals for the AI models are more dispersed. Notably, the CIs for M2 and M3 are wider, reflecting their higher score variance, and they do not overlap with the CI for H2. This suggests a statistically significant difference between the mean scores assigned by these AIs and the most generous human grader.

(Figure 1: 95% Confidence Intervals for the Population Mean)



3.3 Cases with significant differences between human and AI grading

For each submission, the score given by each AI grader is compared to the median of scores by human graders. This difference is regarded as significant if it exceeds 4 out of 40 marks. The frequency of such cases is shown in Table 3. Three cases with the greatest difference are selected for case study. M2 was instructed to regrade and provide detailed explanation for the scores of each component. The case study is based on M2’s feedback, reviewed by a human.

Table 3: Submissions with Significant Difference between AI Graders and Median of Human Graders

AI Grader	Significantly above		Significantly below		Average difference	
	<i>n</i>	%	<i>n</i>	%	<i>-m</i>	%
M1	5	10	14	28	-1.24	-3.1
M2	1	2	23	46	-3.68	-9.2
M3	1	2	26	52	-4.22	-10.6

- **Case 1:** (H1: 28; H2: 26; H3: 27; M1: 18; M2: 16; M3: 17)
M2 gives fair scores for C1, C2, and C3 but low scores in C4 and C5. It explains that the essay lacks a clear actionable plan and is below the word limit. Human graders see that the student has put the actionable plan throughout different parts of the essay but not on a specific paragraph. Also, the word count barely reaches the minimum word limit of 500. Without titles and subtitles, it would be under 500 words.
- **Case 2:** (H1: 28; H2: 30; H3: 34; M1: 26; M2: 18; M3: 19)
M2 gives fair to good scores for C1, C3, and C5 but low scores for C2 and C4. It explains that although the student has listed a few goals with a realistic actionable plan, they are too broad and generic, lacking clarity, prioritisation, and measurable parameters.
- **Case 3:** (H1: 29; H2: 34; H3: 25; M1: 19; M2: 14; M3: 14)
M2 gives low scores on all criteria. It points out that the student fails to address own strength and motivation related to major study. The points show vagueness and irrelevance to address disciplinary knowledge and the needs in society and industry. The writing style is also unsatisfactory with various parts not being connected smoothly.

Observations from these three cases demonstrate that AI graders adopt a different grading style from human graders. AI graders tend to adhere strictly to the rubrics provided, with careful evaluation of the whole structure and fulfilment of all components and grading criteria for each submission. In contrast, human graders rely on a general impression when grading after getting an overall understanding of the student’s plan, allowing flexibility in the focus of content and writing style. In addition, AI graders show independence in grading for different submissions. Human graders may unintentionally use previous submissions as reference, hence having bias dependent on the order of grading. The personal experience and knowledge of human graders are also accounted for in the deviation.

4. Implications

Our study shows that an AI grading system developed with precisely defined rubrics can efficiently facilitate grading. However, its consistency compared with human graders varies with different models. Among the three models used, GPT-4.1-mini performs most similarly to human graders. In general, AI graders are stricter in grading and more rigid in adhering to the rubrics. This lack of flexibility results in lower grades compared with human graders. If the context of assessment is highly subjective, such as career development, human experience is essential to accurately assess students' work and provide feedback. Nonetheless, AI provides an objective ground for human graders, especially in large courses with multiple graders. While this study provides an evaluation of AI grading for practitioners in education, its generalisability is limited by the diversity in assessment format and contextual knowledge. The results are more usable for reflective writings in text format of generic knowledge.

Ethical Declaration Statement

This project has been approved by the Institutional Review Board of The Hong Kong Polytechnic University (Reference number: HSEARS20251024002).

Generative AI Policy Statement

The author(s) used POE, GPT-4.1 mini, Grok 4 Fast, and Claude Haiku 4.5 during the preparation of this work to generate AI grading results. After utilising the tool/service, the author(s) thoroughly reviewed and edited the content as necessary and assumed full responsibility for the publication's content.

References

- Awidi, I. T. (2024). Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence*, 6, Article 100226. <https://doi.org/10.1016/j.caeai.2024.100226>
- Belov, D. I., & Armstrong, R. D. (2011). Distributions of the Kullback-Leibler divergence with applications. *British Journal of Mathematical and Statistical Psychology*, 64, 291–309. <https://doi.org/10.1348/000711010X522227>
- Chan, C.K.Y., & Hu, W. (2023). Students' voices on generative AI: perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20, 43. <https://doi.org/10.1186/s41239-023-00411-8>
- Flodén, J. (2024). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, 51(1), 201–224. <https://doi.org/10.1080/000711010X522227>

doi.org/10.1002/berj.4069

- Jauhiainen, J. S., & Garagorry Guerra, A. (2024). Generative AI in education: ChatGPT-4 in evaluating students' written responses. *Innovations in Education and Teaching International*, 62(4), 1377–1394. <https://doi.org/10.1080/14703297.2024.2422337>
- Usher, M. (2025). Generative AI vs. instructor vs. peer assessments: A comparison of grading and feedback in higher education. *Assessment & Evaluation in Higher Education*, 50(6), 912–927. <https://doi.org/10.1080/02602938.2025.2487495>
- Wetzler, E. L., Cassidy, K. S., Jones, M. J., Frazier, C. R., Korbut, N. A., Sims, C. M., Bowen, S. S., & Wood, M. (2024). Grading the graders: comparing generative AI and human assessment in essay evaluation. *Teaching of Psychology*, 52(3), 298–304. <https://doi.org/10.1177/00986283241282696>