

A Multiple LLM Approach to Automated Essay Scoring: Comparing Seven Open-Source Models with Human Graders

Edmond YEUNG ¹, Jialing Juley XIAO ¹, Jonathan LEE ², Marco BETTONI ², Rachel WONG ¹

¹ Department of Management and Marketing, The Hong Kong Polytechnic University, Hong Kong, China

² Industrial Centre, The Hong Kong Polytechnic University, Hong Kong, China

Abstract

This study evaluates whether open-source large language models (LLMs) can support automated essay scoring (AES) of reflective writing in higher education. Forty-nine reflective essays from a foundation entrepreneurship and innovation course were scored by lecturers and independently graded by seven open-source LLMs (Llama 3 8B, DeepSeek R1 32B, Ministral 3 8B, Qwen 3 8B, GPT-OSS 20B, Gemma 3 27B, and Phi 3 14B) using the same rubric. Model performance was assessed via exact match, mean absolute error (MAE), Pearson and Spearman correlations, and paired t-tests/Wilcoxon tests against human scores. Ministral produced complete outputs and demonstrated the most balanced alignment with human grading, showing no significant difference in mean or median scores. Other models exhibited larger deviations, conservative bias, or missing outputs, raising scalability concerns. Results suggest that LLMs may serve as decision-support tools but do not yet replace human judgement for reflective assessment.

Keywords: Automated Essay Scoring (AES), reflective essay scoring, open-source LLM, higher education assessment, AI grading

1. Introduction

Recent advances in artificial intelligence are reshaping education by automating routine tasks, thereby allowing educators to focus on mentoring and curriculum design (Awidi, 2024). Essay grading is one of the most time-intensive tasks; in large classes, it may invite fatigue and inconsistent ratings (Awidi, 2024). In our undergraduate management courses, summative assessment relies on reflective essays. With cohorts numbering in the thousands, this creates a substantial marking load. If AI could assist scoring at scale while preserving validity and fairness, it could alleviate stress and release time for formative feedback.

Automated Essay Scoring (AES) has long promised rapid, consistent scoring; however, most evidence stems from K–12 settings or standardized English tests (e.g., TOEFL/IELTS), rather than discipline-specific university writing. Where higher education is concerned, research often emphasises surface language accuracy rather than higher-order thinking and reflective integration (Uyar & Büyükahıska, 2025; Ouyang et al., 2022). Large Language Models (LLMs) such as ChatGPT have renewed interest in AES because they can interpret context-rich text and produce human-like rationales. Yet, whether they can validly and reliably evaluate reflective essays—requiring meaning-making, disciplinary alignment, and critical self-assessment—remains uncertain (Awidi, 2024; Yeung et al., 2025).

Automated Essay Scoring (AES) and ChatGPT

ChatGPT is arguably the most widely discussed LLM since its introduction, and its application to AES has attracted considerable attention. Numerous studies have explored ChatGPT’s ability to score essays and compared its outputs with human raters. Findings, however, are mixed. For instance, Awidi (2024) examined ChatGPT’s performance in grading short reflective essays in a first-year engineering course at an Australian university. Significant discrepancies persisted between AI-generated and human scores, leading Awidi (2024) to recommend human moderation and to position AI as a supplementary tool rather than a replacement. Similarly, Uyar and Büyükahıska (2025), using IELTS Writing Task 2 descriptors in an EFL context, reported substantial differences between ChatGPT and human ratings across genres, with correlations ranging from weak to moderate—insufficient for practical deployment. While some studies echo these limitations (e.g., Liu et al., 2023; Yoon et al., 2023), others suggest potential improvements under certain conditions (e.g., Tate et al., 2023), creating a contradictory evidence base that underscores the need for new approaches to AES.

A Multiple LLMs Approach

Recent research has shifted from evaluating a single proprietary model to benchmarking multiple LLMs against human graders to assess AES feasibility across diverse contexts. For example, Abujadallah et al. (2025) analysed a large dataset of grade 6–12 argumentative essays to evaluate five open-source LLMs. Their findings indicate that rubric-aligned prompting significantly enhances performance, with models achieving correlations above 0.83 and F1 scores near 0.93. In contrast, Gaggioli et al. (2025), studying psychology essays written in Italian in higher education, found that LLMs struggle to assess tasks requiring “disciplinary insight and contextual sensitivity” (p. 1). These contexts differ substantially from ours: a higher education setting where English is the primary

language of instruction and reflective essays serve as evidence of learning. This scenario offers high transferability to similar global contexts. However, no existing research compares multiple LLMs' performance in evaluating reflective essays within higher education.

Addressing this gap, we pose the following research questions:

1. Which open-source LLMs demonstrate the strongest performance in evaluating reflective essays?
2. How does their accuracy compare to human raters' grades?

2. Methodology

The primary aim of this research is to assess the effectiveness of LLMs in scoring reflective essays within a higher education context. We evaluated seven open-source LLMs available to us. We selected the latest versions: **Llama 3 8B**, **DeepSeek R1 32B**, **Ministral 3 8B**, **Qwen 3 8B**, **GPT-OSS 20B**, **Gemma 3 27B**, and **Phi 3 14B**. Models were selected for diverse representation, while their open-source nature ensures transparency and reproducibility in AES research.

Sample: The Reflective Essays

The dataset comprises 49 reflective essays written by students and previously graded by human raters. The essays were submitted as the summative assessment for a foundation-level university course in Hong Kong introducing concepts related to entrepreneurship and innovation. Students were instructed to reflect on (1) what they learned from the course and (2) how that learning connects to their academic pathway and/or career plans.

The grading rubric contains three components: (i) reflection, (ii) language, and (iii) citation. Each LLM was prompted to assign a single overall score using the same rubric applied by human graders. Scores range from 1 to 12, corresponding to letter grades from F to A+. Human grading was completed by two course lecturers, who also designed the rubric and taught the course.

Data Analysis

All analyses were conducted using SPSS. Each model's score distribution is summarised using descriptive statistics such as mean, standard deviation, minimum, maximum, median, and quartiles. To evaluate scoring accuracy relative to human grades, we calculated:

1. **Exact Match:** The proportion of essays where the model grade equaled the human grade.

2. **Mean Absolute Error (MAE):** To quantify the model’s average grading deviation relative to human scores; here, *deviation* refers to the differences between model and human grades (ignoring whether the model over- or under-estimated the score).
3. **Pearson’s correlation coefficient (r) and Spearman’s rank correlation (ρ):** To examine linear association and rank-order consistency.
4. **Paired sample t-test (p) and Wilcoxon signed-rank test (p):** To test whether the differences between the means and medians were statistically significant.

3. Results and Discussion

Table 1 summarises the descriptive statistics of grades assigned by the human raters and the seven LLMs. Three models—DeepSeek, GPT-OSS, and Phi—produced output errors and therefore graded fewer than the full 49 essays: DeepSeek failed to score nine essays (18.37%), with smaller shortfalls for GPT-OSS and Phi. Although these error rates appear modest in a small sample, they could become a substantial reliability issue when grading thousands of essays.

In terms of score distributions, **Ministral** appears most aligned with human grading: it successfully graded all essays, and its mean and SD closely match the human raters’, with identical quartiles, minimum, and maximum values. Phi shows a comparable mean but a notably larger SD and wider range, indicating greater variability than the human raters. LLaMa displays a compressed distribution with a smaller SD and narrower range, suggesting limited differentiation among essays relative to humans. Qwen, GPT-OSS, and Gemma show lower average scores than humans, consistent with a tendency toward more conservative grading. Finally, DeepSeek exhibits the greatest dispersion—more than double the human SD—alongside missing outputs, indicating both stability and reliability concerns for large-scale deployment.

Table 1. Descriptive statistics of all seven models

	Human	LLaMa	DeepSeek	Ministral	Qwen	GPT	Gemma	Phi
n	49	49	40	49	49	46	49	48
Mean	9.5102	9.0612	9.175	9.6939	8.9796	8.3478	8.5714	9.3333
SD	0.7671	0.3168	1.5834	0.7693	1.2160	1.8405	0.7906	1.4192
Min	8	8	5	8	5	4	7	5
Q1	10	9	9	10	9	9	9	9
Median	10	9	10	10	10	9	9	10
Q3	12	10	12	12	12	12	10	12
Max	12	10	12	12	12	12	10	12

Table 2 provides converging evidence that Ministral is the most promising model overall in this dataset. Ministral achieves the lowest Mean Absolute Error (MAE = 0.6327), indicating the smallest average deviation from human grades. Although both correlation coefficients indicate only moderate correlation, Ministral shows the strongest linear alignment with human scores (Pearson’s $r = 0.4114$) and a comparatively high rank-order association (Spearman’s $\rho = 0.3140$). Importantly, both the paired-samples t-test ($p = 0.130$) and Wilcoxon signed-rank test ($p = 0.127$) are non-significant for Ministral, implying no statistically significant difference from human graders in terms of mean or median scores.

Table 2. Performance metrics of all seven models

	LLaMa	DeepSeek	Ministral	Qwen	GPT	Gemma	Phi
Exact Match	0.3265	0.4286	0.2653	0.3265	0.2653	0.3673	0.3061
MAE	0.6939	2.8367	0.6327	0.8980	2.0408	1.0612	1.2245
Pearson r	0.2975	0.2156	0.4114	0.2124	0.3905	0.2307	0.1101
Spearman ρ	0.2260	0.0600	0.3140	0.2440	0.3640	0.0970	0.0670
t-test p	<0.001	0.2890	0.1300	0.006	<0.001	<0.001	0.4030
Wilcoxon p	<0.001	0.3350	0.1270	0.0080	<0.001	<0.001	0.6350

Although several models outperform Ministral on individual indicators, these advantages do not translate into an overall profile consistent with a reliable and valid automated grader. For instance, DeepSeek and Gemma show higher Exact Match rates than Ministral, and GPT-OSS exhibits a higher Spearman’s ρ . However, these apparent strengths are offset by substantive limitations in other metrics and in operational robustness. DeepSeek’s performance is undermined by substantial missing outputs (reduced n), markedly poorer error-based accuracy (high MAE), and agreement indices, raising concerns about scalability and consistency in real-world, high-volume grading. GPT-OSS, despite stronger rank correlation, shows large absolute deviations from human scores (elevated MAE) and a statistically significant difference between its mean and median compared with the human raters’. Therefore, they are less likely to be considered to have sufficient reliability, consistency, and validity for AES purposes.

4. Implications

The findings suggest that **Ministral 3 8B** is the most viable candidate among the evaluated open-source models for supporting automated essay scoring (AES) of reflective writing in a higher education context. Across the seven models, Ministral demonstrated the most balanced performance profile. These

results support the implication that Ministral can approximate human scoring patterns more closely than the other models in this study, and therefore has potential as an AES component for reflective essays when appropriately constrained and monitored.

However, the study also indicates clear boundaries on what the current approach can justify:

1. **Moderate Correlation:** The observed correlations remain moderate rather than strong. While Ministral generally moves in the same direction as human grades, it does not reliably reproduce fine-grained distinctions between adjacent grade bands at the individual-essay level. As a result, Ministral cannot yet replace human grading but may serve as a decision-support tool. Human graders should retain responsibility for final decisions, while the model's value lies in reducing routine workload and enabling instructors to redirect effort toward richer formative feedback.
2. **Need for Replication:** The results underscore the need for replication with a larger dataset before claiming generalisability. With 49 essays from one foundation course, the estimates of accuracy and association are necessarily sample-dependent and may shift with changes in cohort composition, topic prompts, or rater expectations. A larger dataset would allow more stable evaluation of performance and would better reflect the real operational environment where reflective essays vary widely in content and citation practices.

Ethical Declaration Statement

The project has been approved by the Departmental Research Committee of The Hong Kong Polytechnic University (Reference Number: HSEARS20251203008).

Generative AI Policy

The authors used Copilot during the preparation of this work to proofread and generate ideas for the title of the research. After utilising the tool/service, the authors thoroughly reviewed and edited the content as necessary and assumed full responsibility for the publication's content.

References

Abujadallah, M., Saad, M., & Abudalfa, S. (2025). Evaluating Open-Source LLMs for Automated Essay Scoring: The Critical Role of Prompt Design. https://www.preprints.org/frontend/manuscript/90bf5ac43bbcb-975d13a26f61ce24d7e/download_pub

- 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, 100226. <https://doi.org/10.1016/j.caeai.2024.100226>
- Gaggioli, A., Casaburi, G., Ercolani, L., Collovà, F., Torre, P., & Davide, F. (2025). Assessing the Reliability and Validity of Large Language Models for Automated Assessment of Student Essays in Higher Education. <https://arxiv.org/pdf/2508.02442>
- Liu, Y., Moosavi, N. S., & Lin, C. (2024, August). LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 12688-12701). <https://doi.org/10.18653/v1/2024.findings-acl.753>
- Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025, March). Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 462-472). <https://dl.acm.org/doi/pdf/10.1145/3706468.3706527>
- Tate, T. P., Steiss, J., Bailey, D. H., Graham, S., Ritchie, D., Tseng, W., Moon, Y., & Warschauer, M. (2023). Can AI Provide Useful Holistic Essay Scoring? <https://doi.org/10.31219/osf.io/7xpre>
- Yeung, W. E., Xiao, J. L., & Wong, F. R. (2025). Using ChatGPT for Automated Essay Scoring of Reflective Essays in University Management Courses. In *Academy of Management Proceedings 2025*. <https://cdmcd.co/vaA8wJ>
- Yoon, S. Y., Miszoglaid, E., & Pierce, L. R. (2023). Evaluation of ChatGPT Feedback on ELL Writers' Coherence and Cohesion. <https://arxiv.org/pdf/2310.06505>