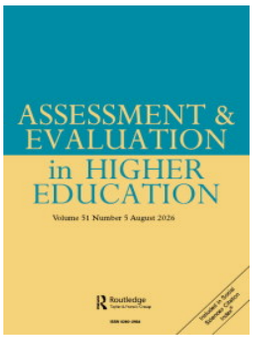
Can GenAI be trained to mimic human markers of extended written assignments in higher education?

KAY, William P <http://orcid.org/0000-0001-6855-7153>, RUTHERFORD, Stephen M <http://orcid.org/0000-0002-5572-8854>, SMITH, David P <http://orcid.org/0000-0001-5177-8574>, SHORE, Andrew M <http://orcid.org/0000-0001-7115-2050> and FRANCIS, Nigel J <http://orcid.org/0000-0002-4706-4795>



Routledge
Taylor & Francis Group

Assessment & Evaluation in Higher Education

ISSN: 0260-2938 (Print) 1469-297X (Online) Journal homepage: www.tandfonline.com/journals/caeh20

Can GenAI be trained to mimic human markers of extended written assignments in higher education?

William P. Kay, Stephen M. Rutherford, David P. Smith, Andrew M. Shore & Nigel J. Francis

To cite this article: William P. Kay, Stephen M. Rutherford, David P. Smith, Andrew M. Shore & Nigel J. Francis (11 Aug 2026): Can GenAI be trained to mimic human markers of extended written assignments in higher education?, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2026.2710688

To link to this article: https://doi.org/10.1080/02602938.2026.2710688

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
https://doi.org/10.1080/02602938.2026.2710688

Can GenAI be trained to mimic human markers of extended written assignments in higher education?

William P. Kay[a], Stephen M. Rutherford[a], David P. Smith[b], Andrew M. Shore[a] and Nigel J. Francis[a,c]

[a]School of Biosciences, Cardiff University, Cardiff, UK; [b]School Of Biosciences and Chemistry, Sheffield Hallam University, Sheffield, UK; [c]School of BioSciences, Faculty of Science, The University of Melbourne, Melbourne, Australia

ABSTRACT

The rapid rise of Generative Artificial Intelligence (GenAI) has promoted interest in its use for marking student work in Higher Education. Harnessing the pattern-recognition capabilities of Large Language Models (LLMs) may potentially facilitate objective grading of students' work at scale. This study investigated whether LLMs could mimic human marking of extended written work sufficiently to function as a formative quality benchmarking tool for students. Two versions of a popular GenAI platform, ChatGPT, were used to mark 50 undergraduate bioscience essays against seven assessment criteria under four different prompting conditions. LLM-assigned marks were compared to human marks in both their mean scores and mark variability using distributional modelling. While overall composite marks were often broadly similar between humans and LLMs, substantial discrepances emerged at the individual criterion and essay levels. LLMs generally awarded higher marks than humans, but with significantly reduced variability, resulting in systematic compression of marks towards the centre of the distribution. Lower-scoring essays tended to receive inflated marks, whereas higher-scoring essays received reduced marks relative to human assessment. These findings suggest that current LLMs do not reliably reproduce human judgement in the marking of extended written work, with important implications for assessment practice and GenAI-assisted benchmarking of marks.

KEYWORDS

Artificial intelligence; assessment; marking; large language model

Introduction

Assessment in higher education (HE) is expected to be educationally meaningful and procedurally fair (Bloxham and Boyd 2007; Boud 1995). Essays and other extended written tasks have historically played important roles across many disciplines, where they are used to evaluate students' conceptual understanding, critical

CONTACT Nigel J. Francis nigel.francis@unimelb.edu.au School of BioSciences, Faculty of Science, The University of Melbourne, Melbourne, Victoria 3010. Australia.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/02602938.2026.2710688.

thinking, and use of evidence. The marking/grading of complex, open-ended work is unavoidably interpretive and vulnerable to inconsistency, bias and human error during the marking process (Bloxham et al. 2016; Hasan and Jones 2024; Sadler 1989). Marking extended written work is also time-consuming, thereby limiting formative feedback opportunities which means methods to increase throughput while retaining quality are often sought.

Generative artificial intelligence (GenAI) and large language models (LLMs) are rapidly reshaping assessment practices in HE. Concerns over misuse of LLMs by students to author assignments have caused major reconsiderations of assessment practices (Francis, Jones, and Smith 2024; Pallant et al. 2026; Winstone, Gravett, and Elkington 2026). Recent reviews also chart accelerating interest in AI-assisted systems to support marking, feedback and exam design in HE, whilst highlighting substantial ethical and practical concerns (Xia et al. 2024; Yan et al. 2024; Zhao 2024). There is growing interest in whether LLMs can act reliably as a marker (either autonomously, or as a decision-support tool for human examiners), especially for assessments where workload pressures are acute (Bearman et al. 2024; Li et al. 2024). The extent to which LLMs can mimic human judgement is central to this question.

Studies of AI-based and automated essay scoring predate GenAI and have demonstrated that algorithmic raters can reach moderate to high agreement with human markers in some high-stakes testing contexts; however, equity and validity concerns remain (Bridgeman, Trapani, and Attali 2012). Recent empirical studies report that contemporary LLMs can approximate human exam markers, at least to the level of overall grades/grade bands, across several disciplines. Flodén (2025) reported that ChatGPT produced exam grades in HE that correlated strongly with teacher marks. Recent work by the 'Opraise project' (Talmi, et al. 2026), comparing the marking of 761 psychology essays by four LLMs, also showed reasonable alignment with human marks. However, the authors also observed LLMs assigned 'middling' marks and were oversensitive to linguistic features, thus deeming them unsuitable to mark work reliably, and with close alignment to human marks, at present. Usher and Faraon (2025) also observed reasonable alignment between human and LLM marks, though found larger discrepancies at lower grade levels. This contrasted between-human agreement (student peers in this case), which was greater at the lower grade levels. In an English as a Foreign Language context, rubric-based grading by LLMs has shown promising levels of reliability and validity relative to human markers, though with a tendency towards regression to the mean and compressed mark ranges (Yavuz, Çelik, and Yavaş Çelik 2025). Other work, including experimental and classroom studies, has examined consistency, bias and educational justice implications of LLM-based essay scoring, reporting mixed but cautiously optimistic findings (Li et al. 2024; Rony, Tan, and Arsovski 2025; Seßler et al. 2025). However, as the complexity of student work increases, the ability of LLMs to produce comparable judgements falls away dramatically. A study of a range of LLMs used to critique doctoral theses showed very poor alignment with human assessment (Tensen, Carey, and Grainger 2026). Across this emerging literature, two patterns are particularly interesting. First, the performance of LLMs as markers appears highly sensitive to

prompt design, rubric clarity and availability of calibration examples (Li et al. 2024; Rony, Tan, and Arsovski 2025; Yavuz, Çelik, and Yavaş Çelik 2025). Secondly, there appears to be limited empirical work that systematically manipulates how the outputs of LLMs are refined or calibrated against human markers, by providing additional contextual information, within authentic HE contexts.

A common question within the HE sector currently is whether GenAI could be used to enhance, or even replace, human assessors for marking student work. However, there are underlying associated practical and ethical questions, that extend beyond the accuracy of the marks awarded. There is a high potential (real or perceived) for GenAI marked work to be more consistent, and objective in its judgements, compared to humans; enabling equity across numerous submissions. Initial indications are that for shorter text submissions, GenAI tools can provide reproducible marks which align with human markers (Chen and Wan 2025; Impey et al. 2025; Sreedhar et al. 2025). However, this would be dependent on evidence to show that outcomes of GenAI marking are reproducible and consistent. Another potential benefit could be the use of GenAI-assisted marking as a potential way to manage workload and improve consistency (Alsalem 2024; Bearman et al. 2024; Farazouli et al. 2024; Li et al. 2024). However, this is countered by concern over potential loss of professional judgement, accountability, and risks to academic standards. Students' perspectives of GenAI-assisted marking are similarly ambivalent (Zhang, Li, and Huang 2026). While some studies indicate that, under certain conditions, students may perceive GenAI-based evaluators as fairer than human markers, particularly when evaluation processes are transparent (Chai et al. 2024), others report substantial scepticism about LLM grading, especially around its ability to understand nuance, creativity, disciplinary context, and around its accountability for errors (Rose et al. 2026; Thomas, Yildirim-Erbasli, and Hariharan 2026; Yang 2026). Students also emphasise the need for transparency, human oversight and clear appeals mechanisms if LLMs are used in grading (Usher and Faraon 2025; Yan et al. 2024). There are also considerations around the ethics of delegating marking to GenAI, rather than the human 'expert', where students may expect feedback from tutors who know them, their context and prior learning. Finally, concerns regarding data integrity, particularly whether students' work, personal information, or research data should be entered into LLMs, which may then utilise those inputs to augment the overall centralised training set. Many HE institutions now have enterprise agreements with GenAI platforms to secure student data. However, concerns remain around this security, for example a recent data protection impact assessment by SURF (2026) assigned an 'orange' compliance rating for one commonly-used platform.

In the short term, GenAI may be best conceptualised as a complement to human judgement, rather than a replacement. 'Human-in-the-loop' models leverage GenAI's efficiency and consistency, while retaining human responsibility for final decisions (Bearman et al. 2024; Li et al. 2024; Yan et al. 2024). One attractive role for LLMs is the provision of timely formative feedback and indicative marks that help students calibrate the quality of their own work against the academic standards for an assessment, in situations where opportunities for formative feedback from human tutors

are limited due to time constraints and staff workload. The generation of formative grades from LLMs to aid students in benchmarking their performance in preparation for high-stakes assessments could be useful. This could mean rapid return of an indicative mark to students, without negatively impacting educator workload, and avoiding students having to wait for markers to evaluate formative work. Then formal summative decisions could remain with human markers (Bearman et al. 2024; Zhao 2024). However, institutions need robust evidence that LLMs can approximate human judgements with reasonable precision.

This study examined how two versions of a widely-used LLM (ChatGPT4o and GPT5.1 Instant) applied a written rubric to 50 first-year biosciences essays. ChatGPT was chosen because it was the most well-known and popular choice for students at the time the work was undertaken, and therefore the more likely choice for a student to use. ChatGPT4o was the current model at the time, and GPT5.1 Instant was released before the end of the analysis, and so was also included in the analysis.

LLM marks were compared to marks from original academic markers across a range of LLM prompting conditions. The findings indicate that LLMs broadly approximate human marks, but there are overriding trends of raising marks and compressing mark ranges that suggest they cannot, at present, be used as reliable predictors of human marks in extended written assignments.

Methods

Context and assessment task

The study was conducted in a UK university on outputs from an essay-based assignment in a first-year undergraduate biosciences module. Each student submitted an essay (*c.*1500 words), addressing one of 21 bioscience-specific questions, alongside a production log detailing their research/development process. Each essay was marked by a single academic using a standardised rubric with 7 separate criteria relating to aspects of the essay or production log (see Appendix 1 for details of criteria). Each criterion was graded using categorical marking to 2/5/8 points on a 0–100 scale (e.g. 58, 62, 65, etc.), with students typically achieving marks within the 52–75 range. The criterion-level marks were combined, using pre-determined differential weightings (see Appendix 1), for an overall composite mark.

Fifty essays were selected from a pool of 450, based on their composite mark from the original marker. Essays were selected aiming to ensure that there were samples representing low, medium and high marks within each 10-mark band. In practice, equal numbers of essays within each 10-mark range across the full range of marks could not be achieved, due to there being few very low (<50) or very high (>80) marks within the available pool of essays. The actual range of marks for the essays selected is shown in Appendix 2). Essays were submitted in Spring 2024, just over a year after LLMs became widely available to the public. Due to the relative novelty of LLMs at this time, extensive use of LLMs to author the essays is unlikely. We can be reasonably assured, therefore, that the essays were written by the students themselves, although use of GenAI in the writing of the essays is possible.

Ethical considerations

Ethical approval was received from the School Research Ethics Committee (Reference: Francis 24-08-03). Essays were selected with author permission through a robust opt-out process. Author permission included use of their essay for research, and approval to enter the essay into LLM platforms. 422/450 consented to inclusion in the study. All essays had personal identifiers removed prior to use. Academic staff markers of the 50 essays selected for analysis gave consent for use of their marks and feedback.

Marking of essays

Aligned with the human-marking approach, LLMs determined criterion-level marks only. The 7 criterion marks were then combined *post-hoc* using weightings from the original assignment to create a composite mark ('Overall mark' subsequently). Two LLMs were evaluated: ChatGPT4o and GPT5.1 Instant (subsequently referred to as 'GPT4' and 'GPT5′ respectively); accessed *via* the provider's web interface. Four prompt conditions were used (see Figure 1 and Appendix 3a) using standardised prompt templates. Prompts were iteratively developed by the authors, and tested on an essay that was not included in the final dataset (see Appendix 3b for the final prompt texts and LLM settings). 'Prompt 1′ was used as the initial prompt for Conditions A–D. In addition, 'Prompt 2a' and 'Prompt 2b' were then used sequentially for Conditions C and D. If the LLM produced a mark that did not correspond to one of the permitted discrete values within the rubric (i.e. a 2, 5 or 8 point), it was re-prompted, within the same chat, and told to ignore that mark and re-mark the criterion. However, such instances were infrequent, and only marks consistent with the rubric were included in the final analysis.

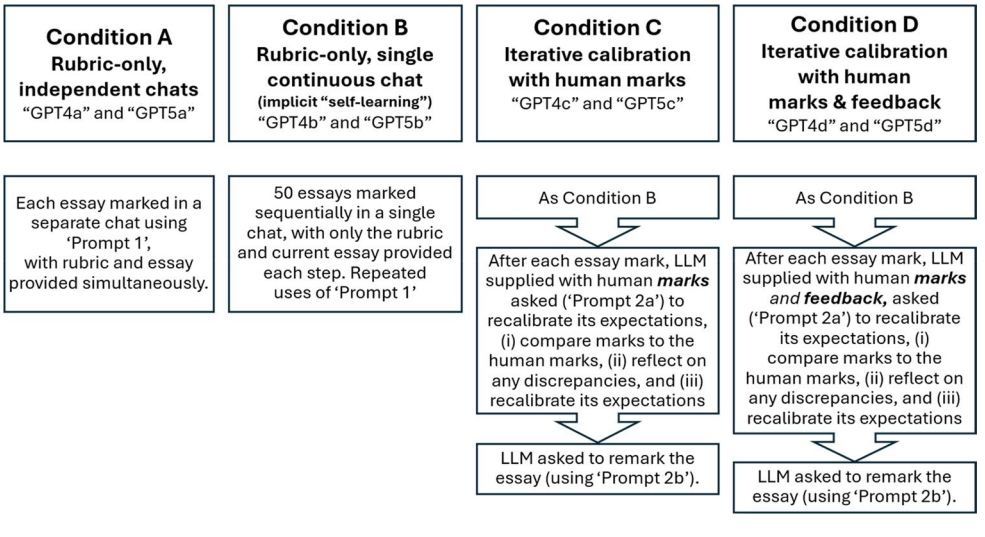


Figure 1. Flowchart showing four 'conditions' (A-D) used for producing criterion marks for essays. Each condition is subsequently referred to by the LLM model followed by a subscript letter (e.g. Condition a for ChatGPT 4.o is presented as 'GPT4a').

For each condition, essays were presented to each LLM in the same fixed order (1–50), held constant across conditions on the assumption that any position-related effects on agreement between human and LLM marks (e.g. increasing alignment as more essays were seen) would be consistent across conditions. Potential ordering effects were not explicitly analysed.

Reliability of LLMs

To assess within-model reliability of LLM outputs under identical prompt conditions, a subset of 15 of the 50 essays was selected, spanning a wide mark range. Each LLM marked these essays five times using Condition A, with each run conducted in a new chat. Reliability was evaluated at the criterion level. For each marker-essay-criterion combination, the standard deviation (SD) of marks across runs was calculated, and these essay-level SDs were summarised using the median SD. Agreement across runs for each marker-criterion combination was assessed using the intraclass correlation coefficient (ICC). All statistical analyses were undertaken using R version 4.5.3 (R Core Team 2026) *via* RStudio version 2026.4.0.526 (Posit Team 2026).

Comparison of human versus LLM marks

Distributional model

To compare how humans and LLMs graded the 50 selected essays, differences were analysed in both the mean mark (μ) and mark variability (σ) using regression models estimating these two features simultaneously (Stasinopoulos and Rigby 2007). Marks were modelled as a function of the marker (human or LLM) and criterion, with essay included as a random effect to account for the non-independence arising from repeated marking of the same essays across markers.

Several possible model structures were considered, differing in how human and LLM marks were allowed to vary, including across marking criteria. The best-supported model was selected using a standard information-theoretic approach, based on the Akaike Information Criterion (AIC; Burnham and Anderson 2004). Model fit was evaluated using residual diagnostics. See Appendix 4 for full details of the modelling approach.

Calibration model

To examine whether differences between LLM and human marks depended on essay quality, a conditional calibration model was fitted, focusing on LLM-assigned marks. Specifically, LLM marks were modelled as a function of marker, criterion, and corresponding human-assigned mark (taken as a proxy for underlying essay quality). This complemented the distributional model by providing a mechanistic, essay-by-essay assessment of how LLM marks tracked human marks; allowing investigation of whether agreement between LLMs and humans changed systematically across the range of essay quality. See Appendix 5 for further details.

Results

Data summary

The dataset comprised 3600 observations. Each essay contributed 72 marks: 7 assessment criteria and the 'overall' mark, each evaluated by 9 'markers' (the human marker, plus ChatGPT4o ('GPT4') and ChatGPT5.1 Instant ('GPT5')), using 4 Conditions (A-D respectively).

For every criterion, mean mark ± SE, and SD ± SE, were derived for each marker from 50 essay-level observations. Marks are presented in a range of 0–100.

LLMs produced consistent marks under repeated prompting

GPT4 and GPT5 showed only modest variability in marks over repeated runs; the median standard deviation ranged from 1.7–4.8 for GPT4 and 1.9–5.9 for GPT5, depending on the criterion (Table 1). This indicated that, for a given essay and criterion, repeated runs of the same LLM typically varied by approximately ± 5. Intraclass correlation coefficients (ICCs) indicated consistently excellent agreement across runs for GPT4 (between 0.84–0.97), and moderate to excellent agreement for GPT-5 (0.50–0.91), with some variation across criteria (Table 1).

The LLMs were therefore considered sufficiently reliable to support using a single output per essay per Condition to compare human versus LLMs.

Overall essay marks differed modestly between LLMs and humans

Model comparison supported a distributional model with normally distributed errors; both the mean mark (μ) and its variability (σ) were associated with marker and criterion, with an essay-level random effect. The model explained 72.3% of the variation in the data. Full details of model selection and performance are provided in Appendices 4 and 6.

There were numerous modest, statistically significant, differences in the mean overall essay mark awarded by LLMs compared to the mark given by humans, with LLMs awarding a mean mark up to 7.3 ± 0.9 higher or 3.6 ± 1.1 lower (Figure 2a).

Table 1. Test-retest reliability of LLM-generated marks across criteria median SD, and ICC values for each LLM across criteria, based on five runs per essay under identical prompt conditions.

	GPT4		GPT5	
Criterion	Median SD	ICC	Median SD	ICC
Overall	1.47	0.97	2.04	0.82
Content	1.64	0.96	1.64	0.91
Organisation	3.13	0.89	5.36	0.61
Presentation	4.98	0.86	5.08	0.81
Academic	3.08	0.92	4.42	0.59
Strategy	3.50	0.84	4.38	0.75
Credibility	3.08	0.89	4.55	0.50
Justification	2.12	0.93	3.51	0.80

Smaller median SD and higher ICC values indicate greater reliability.

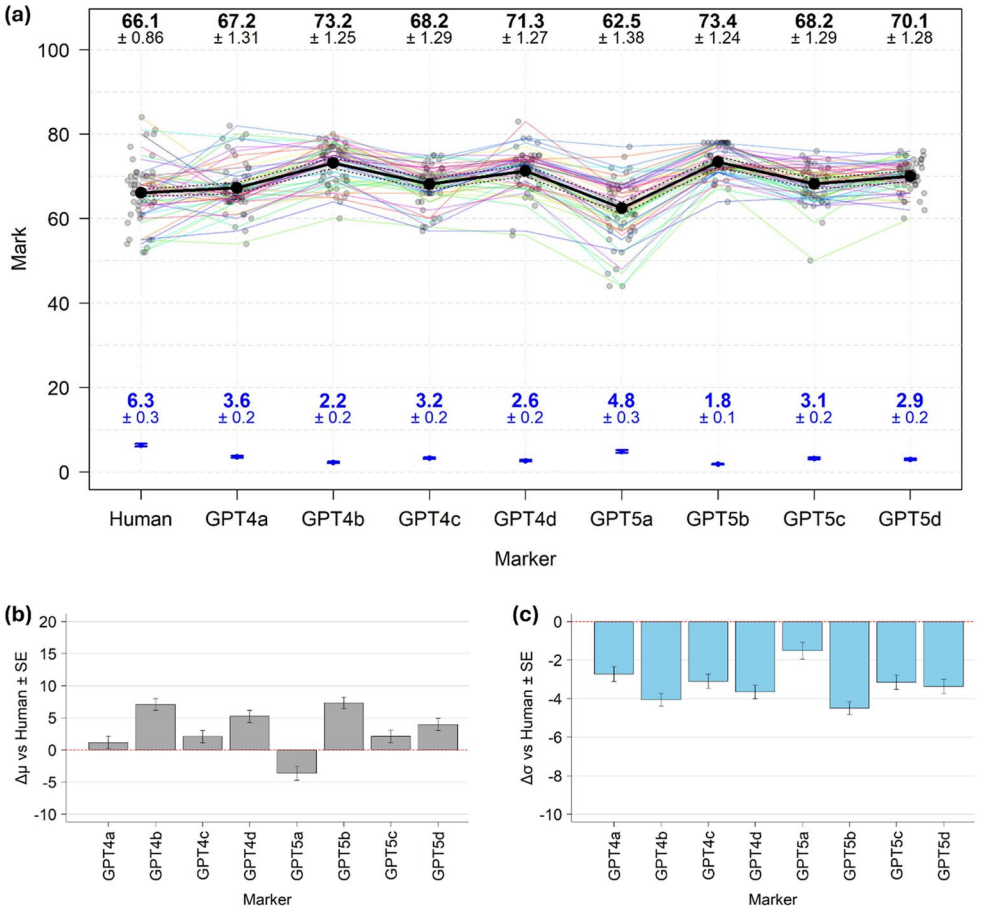


Figure 2. Overall marks awarded: (a) the overall essay mark awarded by humans and LLMs. Small grey dots connected with individually coloured lines show individual essay marks ($n = 50$). The large black dots and bold black line show the mean (μ) mark awarded (± SE in shaded region); these values are also shown in black text. The small blue dots with error bars show the point estimates for the mark variability (σ) ± SE; this value is also included in blue text. (b,c) The difference in μ and σ, respectively, between LLMs and humans (error bars show ± SE).

All but one of the LLMs awarded a higher mean overall mark than humans (Figure 2b).

The mean overall mark given by humans was 66.1 ± 0.9. Condition B saw the highest mark increase compared to humans, with GPT4b's mean overall mark 7.1 ± 0.9 higher (t = 7.78, $p < 0.001$; i.e. increasing from 66.1 to 73.2) and the GPT5b mark 7.3 ± 0.9 higher (t = 8.20, $p < 0.001$). GPT4d provided a 5.2 ± 0.9 higher mean mark (t = 5.61, $p < 0.001$). Smaller increases in marks were observed for GPT4c, GPT5c and GPT5d (increases between 2 – 4; all $p < 0.05$). In contrast, the GPT5a mark was 3.6 ± 1.1 less (t = −3.35, $p < 0.001$), while the GPT4a mark was not significantly different (1.2 ± 1 higher, t = 1.18, $p = 0.239$).

All LLMs also exhibited reduced variability of the overall mark (i.e. a narrowing of the mark range) compared to humans (Figure 2c and Appendix 7). The predicted

SD of overall marks was 1.5 – 4.5 lower for LLMs than humans (all $p < 0.001$). GPT4b and GPT5b saw the largest compression of overall mark variability compared to humans, at −4 and −4.5 respectively.

In summary, LLMs tended to be slightly more generous than humans on average, but also less discriminating in the spread of marks they awarded, especially GPT4b and GPT5b. This means they may produce marks that look broadly plausible, but may not capture the full range of student performance in the same way as human markers.

Criterion-level marks varied more widely between LLMs and humans

While differences in the mean and variability of overall marks between LLMs and humans were relatively small, differences for individual assessment criteria were comparatively large (Figures 3 and 4). Statistically significant marker × criterion interactions indicated criterion-specific differences in mean marks awarded between LLMs and humans. Here, we report the most substantive results (Figure 3), with a summary of the differences in mean marks for all criteria and conditions shown in Figure 4. Graphs for criteria not shown in Figure 3 are provided in Appendices 7 and 8.

The largest positive difference in mean mark between LLMs and humans was 16.1 ± 1.2 (GPT4b marking the 'Content' criterion; Figure 3a, Figure 4 second row), and the largest negative mean mark difference was −8 ± 1.5 (GPT4a marking the 'Presentation' criterion; Figure 3c, Figure 4 fourth row). The 'Justification' criterion saw the smallest differences on average between humans and LLM mean marks, from −7.4 ± 1.4 to 3.6 ± 1.4 (Figure 3b. Figure 4 final row). Differences in mean mark between humans and LLMs for other individual assessment criteria were typically more modest, and across all LLM conditions, the mean mark given by LLMs was generally higher than humans (Figure 4, Appendices 7 and 8). GPT4b and GPT5b, and GPT4a and GPT5a, were exceptional in that they awarded disproportionately higher and lower marks, respectively, compared to other LLM conditions (Figure 4, Appendices 7 and 8).

All LLMs exhibited significantly reduced variability of individual criterion marks relative to humans, with reductions in predicted SD of between 1.5 ± 0.4 to 6.8 ± 0.7 (Figures 3 and 5, Appendices 7 and 9). LLMs thus retained as little as 28–76% of the variability observed in human marks. GPT5b and GPT4b exhibited the greatest reductions in mark variability relative to humans (ΔSD = −5.8 and −5.3, respectively, when averaged across criteria). GPT4c, GPT4d, GPT5c, and GPT5d demonstrated smaller reductions, between −4.0 to −4.7. GPT5a and GPT4a saw the smallest reductions compared to humans, of −3.5 and −2.9, respectively. In contrast to criterion-specific effects observed for mean marks, there was no marker × criterion interaction on mark variability, indicating that LLMs exhibited a general tendency across criteria to compress mark variability relative to humans (Figure 5, which shows only negative observations). However, because marks for individual assessment criteria differed in their human-assigned variability, this uniform compression resulted

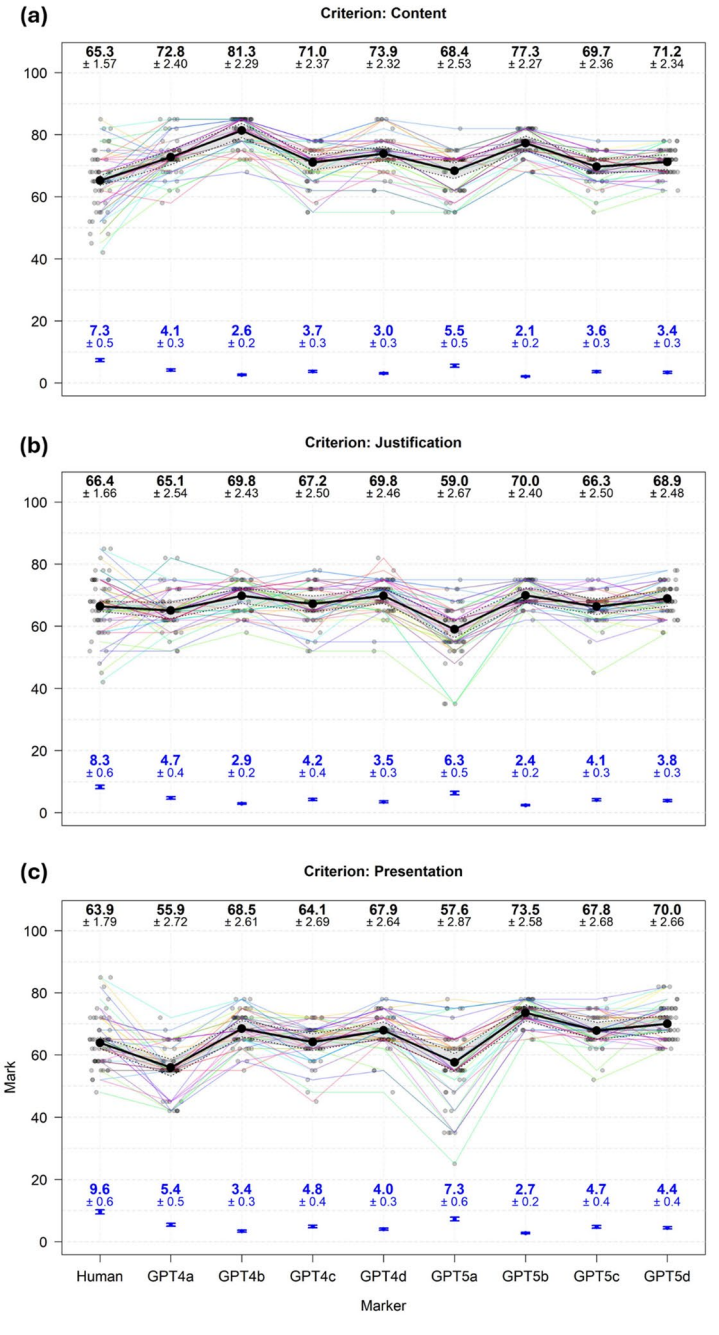


Figure 3. The mark awarded by humans and LLMs for the 'Content', 'Justification' and 'Presentation' criteria. Small grey dots connected with individually coloured lines show individual essay marks ($n = 50$). The large black dots and bold black line show the mean (μ) mark awarded (± SE in shaded region); these values are also shown in black text. The small blue dots with error bars show the point estimates for the mark variability (σ) ± SE; this value is also included in blue text. Panels show contrasting activities: (a) 'Content' shows the most extreme example of LLMs awarding higher marks; (b) 'justification' shows the most modest differences, compared to the human markers; (c) 'presentation' shows the most extreme example of lower marks.

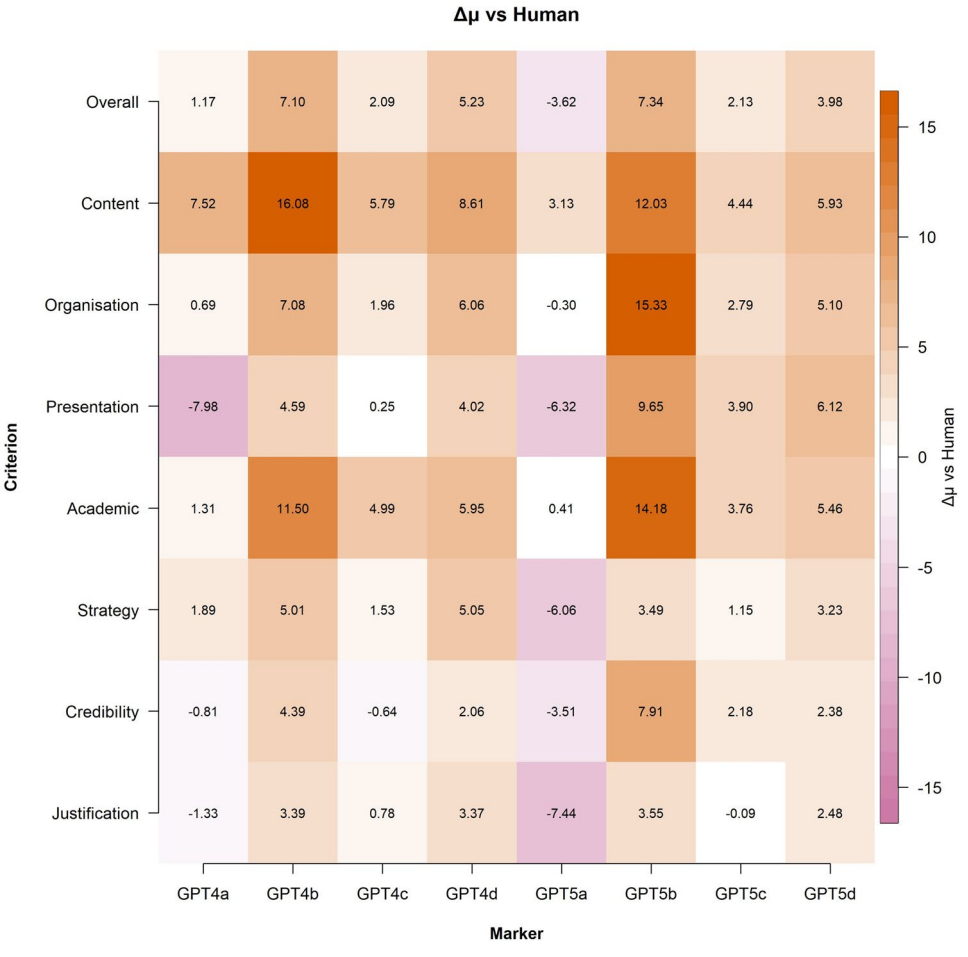


Figure 4. Differences in mean marks. A summary of differences in mean marks (Δμ) between each LLM condition and human markers for individual assessment criteria and the overall essay mark. Positive values (orange) indicate higher mean marks awarded by LLMs relative to humans, whereas negative values (mauve) indicate lower mean marks. Colour intensity reflects the magnitude of the difference.

in larger absolute reductions in variability for criteria with higher baseline variability. For example, the 'Presentation' criterion, which exhibited relatively high human-assigned variability, showed a correspondingly large absolute reduction under LLM marking, consistent with scale-dependent effects rather than differential compression (Figure 5).

In summary, at the level of individual marking criteria, LLMs differed from human markers considerably more than they did for overall essay marks. Hence, LLMs may appear reasonably similar to humans when looking at overall essay marks, but they behave less consistently at the criterion level, tending to flatten differences between essays.

LLM and human marks diverged systematically across the mark distribution While mean differences in marks between humans and LLMs were often modest at the

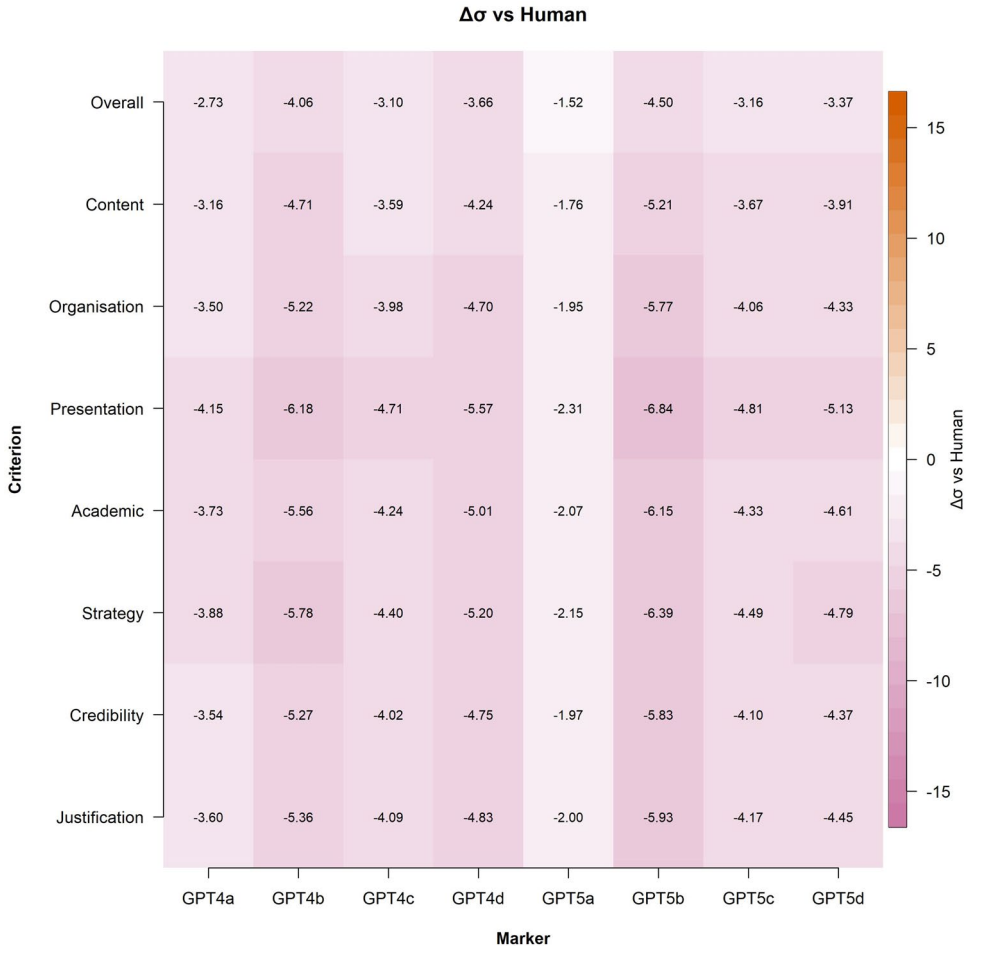


Figure 5. Difference in mark variability. A summary of differences in mark variability (Δσ) between each LLM condition and human markers for individual assessment criteria and the overall essay mark. Negative values (mauve) indicate lower variability. Colour intensity reflects the magnitude of the difference. Note: There are no positive values in the data.

cohort level, discrepancies at the individual essay level were frequently large. Again, this pattern was more pronounced for individual criteria than the overall essay mark. Across all essays, differences between LLM and human overall marks ranged from −24 to +21 (Figure 6a), while at the individual criterion level the largest range, observed for the 'Organisation' criterion, spanned −37 to +40 (Figure 6b); Appendices 10 and 11).

Mark discrepancies were not random, instead varying systematically in association with the human-assigned mark. The calibration model revealed a small, but statistically significant, positive association between LLM-assigned marks and human-assigned marks. For every 1-point increase in the human-assigned mark, the predicted LLM mark increased by 0.038 ± 0.008 ($t = 3.6$, $p < 0.001$); whereas perfect agreement between LLM and human marks would produce a

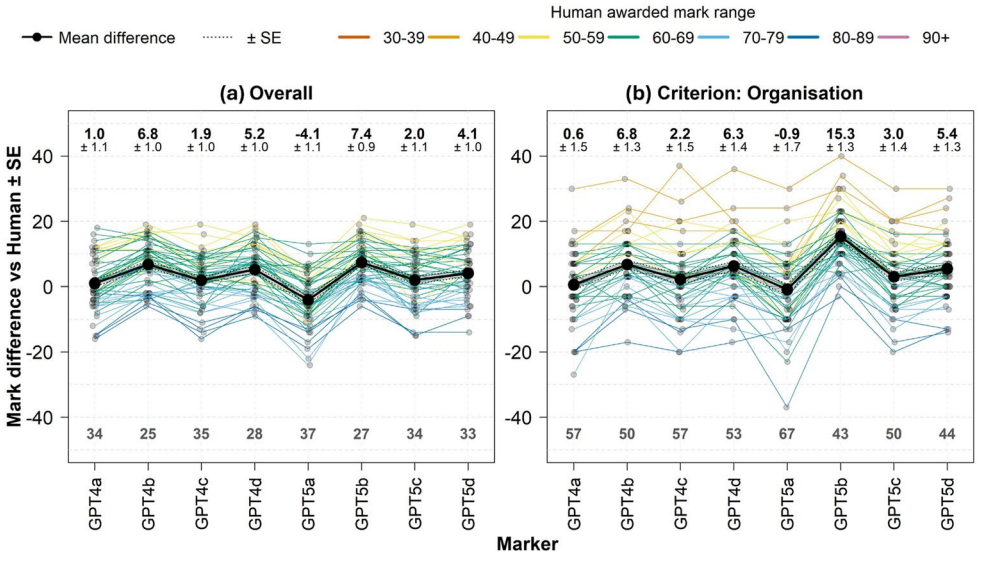


Figure 6. Mark differences between human and LLM marks. The difference in the mark awarded by humans and LLMs for (a) the overall essay mark, and (b) the 'Organisation' criterion. Small grey dots connected by lines show individual LLM-human discrepancies ($n = 50$). The large black dots and bold black line show the mean difference (± SE in shaded region); these values are also shown in black text at the top. Individually-coloured lines represent individual essays, with their colour representing the mark given by the human. See also appendix 10.

slope of 1. LLMs appeared to track underlying essay quality, as defined by humans, but did not separate weak and strong essays clearly enough. Consequently, essays receiving lower marks from humans tended to receive relatively inflated marks from LLMs, whereas essays receiving higher human marks tended to receive relatively reduced marks (Figure 6; Appendices 10 and 11). This phenomenon accounts for the compression of marks towards the centre of the mark distribution (Figure 5).

While the association was small in magnitude, across the full range of human-assigned marks it produced substantial divergence between LLM and human marks, with the largest discrepancies occurring at the extremes of the distribution and the smallest around the centre. For example, essays awarded an overall mark of 50 by humans were predicted to receive LLM marks approximately 16 points higher, whereas essays awarded a human mark of 80 overall were predicted to receive an LLM-assigned mark approximately 13 points lower (Appendices 10 and 11). The 'Academic' and 'Credibility' criteria, which contained the lowest (35) and highest (95) human-assigned marks, respectively, were associated with LLM marks that differed by approximately +30 and −30 marks (Appendices 10 and 11). Across all criteria and LLMs, the maximum observed range of LLM-human discrepancies was 67 marks (Figure 7). LLMs may therefore broadly recognise better and weaker essays, but they compress the mark range, inflating weaker work and under-rewarding stronger work.

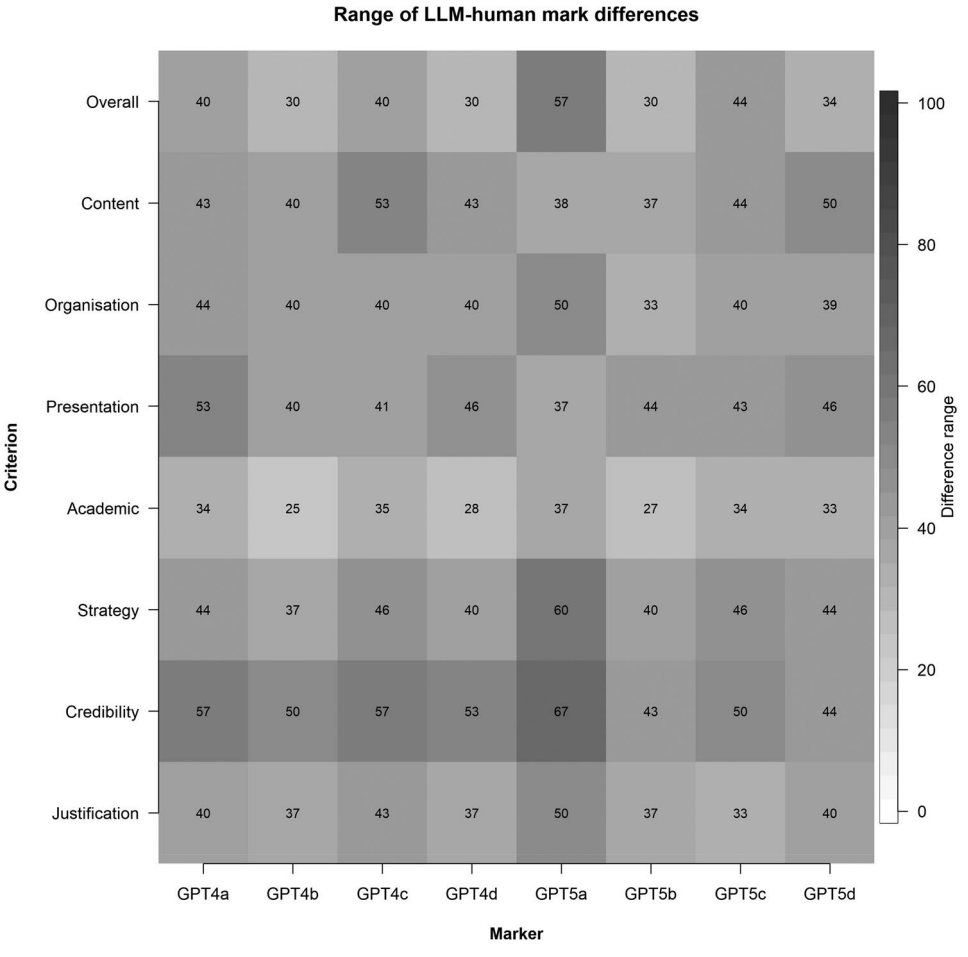


Figure 7. Mark differences between human and LLM marks. Observed range of LLM-human mark differences across individual assessment criteria and the overall essay mark. Values represent the difference between the maximum and minimum LLM-human discrepancies within each marker-criterion combination. Colour intensity reflects the magnitude of the range. Large ranges typically reflect a combination of over- and under-marking rather than one-sided bias (see also Figure 6, Appendices 10 and 11).

Discussion

The findings of this study indicate that the LLMs tested varied considerably in marks awarded, and were inadequate predictors of the human mark awarded to essays. Despite being relatively consistent at producing similar marks when using the same prompt on the same essay, when used to mark a range of essays of differing standards, there was poor alignment between the LLM-provided marks and the marks assigned by the original human marker. Although the 'overall' mark for the essays were relatively well aligned on average with human marks, differences for individual criteria were considerable, reflecting the aggregating effect of a composite score.

Differences in agreement between individual criteria suggest that some aspects of marking are more sensitive to stochastic variation in LLM outputs than others. The largest difference in the mark awarded by any LLM compared to a human-awarded mark was 16.1 at the cohort-level, and 40 at the individual essay level. In all but one case, LLMs typically returned higher average marks than humans across criteria (Figure 4). GPT5a was the exception, returning mean marks that were typically lower compared to the human mark. Interestingly, the overall essay mark (derived from a weighted combination of the 7 individual criteria) was approximately ± 5 compared to the human mark at the cohort-level, which might suggest a tolerable performance of the LLMs as predictors of human marking. However, when evaluating performance at the level of individual criteria and on an essay-by-essay basis, discrepancies between the LLM- and human-assigned marks were substantial. So, despite the LLMs awarding, on average, reasonably similar overall essay marks, they would potentially reach that overall mark from a much more varied combination of marks for each criterion in comparison to human markers, and differences for individual essays could be much larger.

There were some criteria on which LLMs were less reliable than others in awarding average marks that mapped to human marks. For example, the Content and Academic criteria were substantially different, whereas for Strategy, Credibility, or Justification, differences were considerably smaller. These discrepancies suggest that LLMs are better able to mark some criteria than others. However, there was no consistent feature between the more-reliable criteria that could explain this pattern of behaviour. It might be expected that content, structure and use of citations would be easier for pattern-recognition or content-mapping features of an LLM to identify. However, alignment seemed to be better for the more-reflective elements of the assessment, which was surprising.

The most variable mark differences were awarded by the LLMs when supplied with only the essay and the rubric (GPT4a and GPT5a; Figure 4). Notably, under these conditions, the LLM demonstrated a relatively greater tendency to award marks that were lower than those of humans (see first and fifth columns of Figure 4). In both cases where only the rubric was provided but where essays were given sequentially within the same chat (GPT4b and GPT5b), the LLMs consistently exhibited the largest compression effect and awarded relatively higher marks on average than all other Conditions across all criteria except Strategy (Figures 4 and 5). There was no clear explanation for this; we speculate that the LLM was recalibrating itself internally with repeat iterations of the marking process, somehow leading to an upwards drift. However, the 'black box' algorithms that underpin the functioning of commercial LLMs make it impossible to determine why this phenomenon was observed (Winstone, Gravett, and Elkington 2026). An investigation of the trend in the sequential marks given for these two conditions would be useful to determine if there is an ordering effect present, or if these conditions simply provide higher marks.

In Conditions C and D, where additional contextual information was provided in the form of human marks alone or marks accompanied by feedback, respectively, discrepancies between mean LLM- and human-assigned marks were typically similar (within ± 5 marks) to those observed in Conditions A and B (Figure 3). A small

number of model-criterion combinations showed a notable improvement in alignment. However, despite these isolated improvements, mean marks awarded under Conditions C and D remained significantly higher typically than human marks, exceeding them by up to 8.61 marks depending on the model and criterion. These findings suggest that while the inclusion of contextual information may improve LLM performance relative to more rudimentary prompting approaches, any gains are generally modest and inconsistent. Consequently, alignment with human marking remains unsatisfactory and therefore the potential of these LLMs to act as a precise formative benchmarking tool remains highly questionable.

One of the critiques of using a 0–100 range for marking subjective work, such as extended written assignments, is that human markers typically do not use the full range of marks available (Webb, Paul, and Chessey 2020; Yorke, Bridges, and Woolf 2000), with marks below 30 and above 70 being relatively rare. Within the sample of essays analysed in this study (which it should be noted was chosen to represent a wide mark range), mark variability was compressed by all LLMs and across all assessment criteria, and compression was particularly substantial at the extremes of the mark distribution. This has clear implications for the potential of LLMs to serve as formative predictors of quality, as they may fail to identify excellent or very poor work reliably (Talmi, et al. 2026). It was clear that the main cause of this mark compression was the raising of lower marks and the reduction of higher marks by the LLMs (Figure 6).

Despite differences between LLMs in the overall extent to which they compressed the mark range relative to humans, there was no marker × criterion interaction on mark variability. This means that, while each LLM differed in their baseline level of compression, each model applied that compression consistently across assessment criteria. However, because human-assigned mark variability differed between criteria, the same proportional reduction in variability could still produce different absolute reductions across criteria and LLMs, with larger reductions occurring for criteria that were more variable under human marking. Thus, even in the absence of criterion-specific compression effects, variability reductions may appear uneven simply because the baseline levels of human variability differ across criteria.

The findings presented here align with Talmi, et al. (2026), who also observed compression at both the upper and lower mark ranges for psychology essays. Other studies report contrasting findings; Yavuz, Çelik, and Yavaş Çelik (2025) observed that lower marks showed closer alignment to human-derived marks than average or good marks. Conversely, Usher and Faraon (2025) observed that LLMs tended to raise lower-level marks, but that humans and LLMs became more aligned as the quality of the work increased. However, it should be noted that the work being graded in Usher & Faraon's study was a student team-authored questionnaire, not an extended essay, and therefore likely to be more structured and objectively aligned to marking criteria and thus potentially easier for an LLM to mark.

The mechanism underlying the narrowing of LLM-assigned marks was statistically clear, however the reason why LLMs exhibited this behaviour remains uncertain. We anticipate the wording and structure of the rubric may have contributed. Marking criteria within rubrics often have limited variation in descriptors at the upper and lower levels, sometimes differing in only single subjective discriminators (e.g. good,

excellent, outstanding), which may be challenging for LLMs to differentiate. In comparison, a human may have their own innate perspective of those discriminators, as noted by Rutherford et al. (2025). These authors suggest that with more objectively-worded criteria, inter-rater variation would be reduced. We posit that LLMs may be better able to predict extreme marks if provided rubrics consisting of more-detailed criteria, with objectively distinct descriptions for each mark bracket. For example, avoiding differentiators between mark ranges that use progressive descriptors such as 'good', 'excellent', 'outstanding', and 'exceptional' to differentiate between marks. A human may be able to develop an intuitive understanding of these terms, based on marking experience (Rutherford et al. 2025), but such subjective terms are unlikely to be sufficient for an LLM to align to features within written work reliably.

A major, and possibly flawed, assumption when comparing LLM and human marks, is that human marks are reliable, and represent the quality of the work accurately. Previous studies (e.g. Bloxham 2009; Sadler 2009; Tisi et al. 2013) have shown that the marking of subjective work (such as essays) is highly vulnerable to inter-rater variation, which may cause awarded marks to differ substantially between human markers (Rutherford et al. 2025). A more robust approach to assess the agreement between LLMs and humans may be to compare marks awarded by LLMs to a consensus human mark.

Implications for practice

The findings of this study highlight that, at present, GenAI is unable to reliably assign a mark to a subjective piece of written work comparable to that of humans. Even with extensive training of the LLM within a single chat, accuracy is limited. Aside from the ethical issues of submitting student work to GenAI tools without express consent, LLMs cannot, and should not, be relied upon for assigning grades to extended written work by students.

It would be unwise for students to rely on LLMs to benchmark their performance, and certainly, they cannot be relied upon to predict individual students' attainment. With more-subjective assignments, such as essays or written reports, a more reliable formative activity may be to encourage students to focus on engaging more with the marking criteria, and mapping the descriptors of those criteria to their work in order to predict a likely grade (Rust, Price, and O'Donovan 2003), and the use of scaffolded peer-feedback activities (Khan, Payne, and Chahine 2017; Schneider and Preckel 2017). Conversely, the development of more-robust marking criteria, as well as more sophisticated LLMs, may lead to LLMs being better predictors of human marking in subjective written assignments.

Limitations and future work

This study only evaluated two versions of a single GenAI LLM, ChatGPT. This LLM was chosen due to its dominant market share (Cohen et al. 2026). Ideally, the same analyses would have been undertaken on a range of LLMs. In particular,

it would be useful to conduct this analysis on specific LLMs that are being adopted by HE institutions as 'secure' internally-regulated platforms, such as Microsoft's 'Copilot'. However, the two versions of ChatGPT evaluated here performed similarly, and Talmi, et al. (2026) suggest that there is no substantive difference between the four LLMs tested in their study. The versions of ChatGPT used in this study are now less likely to be adopted by users (at the time of submission, the current model is ChatGPT 5.5). However, with the speed of release of updated versions of LLMs, producing data to analyse the current model in real time was not practicable.

The assumption that the human mark is itself a reliable indicator of essay quality is a necessary limitation of this work. Future research should therefore consider comparing LLM-assigned marks against consensus-generated human marks. Also, better-written rubrics, perhaps with less subjective language, more individual criteria (to avoid criteria covering multiple aspects of the essay), and fewer individual grade points within the marking range, may help LLMs discriminate better.

Given growing concerns regarding the impact of GenAI on the validity and authenticity of extended written assessments in HE, the use of the traditional essay as a form of summative assessment may decline (Francis, Jones, and Smith 2024). Consequently, future evaluations of LLM-based marking may be better directed towards more reflective, process-focused, or invigilated forms of assessment.

Conclusions

The findings presented here suggest that the LLMs tested were not suitable to act as GenAI tutors for individual students to provide benchmark grades on their work. There are sufficient levels of variation, grade inflation or reduction, and compression of marks for it to be wise to apply caution to the use of LLMs in the provision of marks on extended written work. Among the range of marks used in this study, the LLMs were typically only well-aligned with human marks for mid-grade essays. In many cases, LLM marks differed substantively from human marks, and the disagreement followed a systematic pattern, namely compression of marks towards the centre of the distribution (i.e. regression to the mean).

Mark compression has major implications for students performing substantially below or above average and would provide poor benchmarking of attainment as a result. In addition, even though there was reasonable agreement on average for the composite mark, the marks for individual criteria and for individual essays varied considerably, emphasising the importance of not relying solely on the composite mark as an indicator of the LLM judgement ability. As LLMs become more sophisticated, their ability to mimic human judgement may enhance. Although with already high levels of inter-rater variation among human markers, this alignment may be hard to achieve.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

William P. Kay http://orcid.org/0000-0001-6855-7153
Stephen M. Rutherford http://orcid.org/0000-0002-5572-8854
David P. Smith http://orcid.org/0000-0001-5177-8574
Andrew M. Shore http://orcid.org/0000-0001-7115-2050
Nigel J. Francis http://orcid.org/0000-0002-4706-4795

References

Alsalem, M. S. 2024. "EFL Teachers' Perceptions of the Use of an AI Grading Tool (CoGrader) in English Writing Assessment at Saudi Universities: An Activity Theory Perspective." *Cogent Education* 11 (1). https://doi.org/10.1080/2331186X.2024.2430865.

Bearman, M., J. Tai, P. Dawson, D. Boud, and R. Ajjawi. 2024. "Developing Evaluative Judgement for a Time of Generative Artificial Intelligence." *Assessment & Evaluation in Higher Education* 49 (6): 893–905. https://doi.org/10.1080/02602938.2024.2335321.

Bloxham, S. 2009. "Marking and Moderation in the UK: False Assumptions and Wasted Resources." *Assessment & Evaluation in Higher Education* 34 (2): 209–220. https://doi.org/10.1080/02602930801955978.

Bloxham, S., and P. Boyd. 2007. *Developing Effective Assessment in Higher Education: A Practical Guide*. Maidenhead, UK: Open University Press.

Bloxham, S., B. den-Outer, J. Hudson, and M. Price. 2016. "Let's Stop the Pretence of Consistent Marking: Exploring the Multiple Limitations of Assessment Criteria." *Assessment & Evaluation in Higher Education* 41 (3): 466–481. https://doi.org/10.1080/02602938.2015.1024607.

Boud, D. 1995. "Assessment and Learning: Contradictory or Complementary?" In *Assessment for Learning in Higher Education*, edited by P. Knight, 35–48. Abingdon: Kogan Page.

Bridgeman, B., C. Trapani, and Y. Attali. 2012. "Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country." *Applied Measurement in Education* 25 (1): 27–40. https://doi.org/10.1080/08957347.2012.635502.

Burnham, K. P., and D. R. Anderson. 2004. "Multimodel Inference." *Sociological Methods & Research* 33 (2): 261–304. https://doi.org/10.1177/0049124104268644.

Chai, F., J. Ma, Y. Wang, J. Zhu, and T. Han. 2024. "Grading by AI Makes Me Feel Fairer? How Different Evaluators Affect College Students' Perception of Fairness." *Frontiers in Psychology* 15: 1221177. https://www.ncbi.nlm.nih.gov/pubmed/38371704. https://doi.org/10.3389/fpsyg.2024.1221177.

Chen, Z., and T. Wan. 2025. "Grading Explanations of Problem-Solving Process and Generating Feedback Using Large Language Models at Human-Level Accuracy." *Physical Review Physics Education Research* 21 (1): 010126. https://doi.org/10.1103/PhysRevPhysEducRes.21.010126.

Cohen, M. C., E. Hage-Youssef, D. McCarthy, and D. D. Sokol. 2026. "Three Strategic Bets on AI's Future." SSRN Electronic Journal. https://doi.org/10.2139/ssrn.6331258.

Farazouli, A., T. Cerratto-Pargman, K. Bolander-Laksov, and C. McGrath. 2024. "Hello GPT! Goodbye Home Examination? An Exploratory Study of AI Chatbots Impact on University Teachers' Assessment Practices." *Assessment & Evaluation in Higher Education* 49 (3): 363–375. https://doi.org/10.1080/02602938.2023.2241676.

Flodén, J. 2025. "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/10.1002/berj.4069.

Francis, N. J., S. Jones, and D. P. Smith. 2024. "Generative AI in Higher Education: Balancing Innovation and Integrity." *British Journal of Biomedical Science* 81: 14048. https://doi.org/10.3389/bjbs.2024.14048.

Hasan, A., and B. Jones. 2024. "Assessing the Assessors: Investigating the Process of Marking Essays." *Frontiers in Oral Health* 5: 1272692. https://doi.org/10.3389/froh.2024.1272692.

Impey, C., M. Wenger, N. Garuda, S. Golchin, and S. Stamer. 2025. "Using Large Language Models for Automated Grading of Student Writing about Science." *International Journal of Artificial Intelligence in Education* 35 (4): 1825–1859. https://doi.org/10.1007/s40593-024-00453-7.

Khan, R., M. W. C. Payne, and S. Chahine. 2017. "Peer Assessment in the Objective Structured Clinical Examination: A Scoping Review." *Medical Teacher* 39 (7): 745–756. https://www.ncbi.nlm.nih.gov/pubmed/28399690. https://doi.org/10.1080/0142159X.2017.1309375.

Li, J., N. K. Jangamreddy, R. Hisamoto, R. Bhansali, A. Dyda, L. Zaphir, and M. Glencross. 2024. "AI-Assisted Marking: Functionality and Limitations of ChatGPT in Written Assessment Evaluation." *Australasian Journal of Educational Technology* 40 (4): 56–72. https://doi.org/10.14742/ajet.9463.

Pallant, J. L., J. Blijlevens, A. Campbell, and R. Jopp. 2026. "Mastering Knowledge: The Impact of Generative AI on Student Learning Outcomes." *Studies in Higher Education* 51 (4): 714–735. https://doi.org/10.1080/03075079.2025.2487570.

Posit Team. 2026. *RStudio: Integrated Development Environment for R*. Boston, MA: Posit Software, PBC. http://www.posit.co/

Rony, S. B. Q., X. F. Tan, and S. Arsovski. 2025. "Educational Justice. Reliability and Consistency of Large Language Models for Automated Essay Scoring and Its Implications." *Journal of Applied Learning and Teaching* 8 (1): 67–77. https://doi.org/10.37074/jalt.2025.8.1.21.

Rose, S. E., L. Taylor, G. Pheiffer, Z. Fortune, and N. Wilde. 2026. "Lacking the 'Personal Touch': students' Perceptions of Generative Artificial Intelligence in Assessment Feedback." *Assessment & Evaluation in Higher Education*: 1–17. https://doi.org/10.1080/02602938.2026.2658633.

Rust, C., M. Price, and B. O'Donovan. 2003. "Improving Students' Learning by Developing Their Understanding of Assessment Criteria and Processes." *Assessment & Evaluation in Higher Education* 28 (2): 147–164. https://doi.org/10.1080/02602930301671.

Rutherford, S., C. Pritchard, W. Kay, L. Nelson, H. Shaw, and N. Francis. 2025. "Marking the Markers: Evaluating the Potential of Professional Development through Collaborative Marking Circles." *Emerging Topics in Life Sciences* 9 (5): ETLS20253033. https://doi.org/10.1042/ETLS20253033.

Sadler, D. R. 1989. "Formative Assessment and the Design of Instructional Systems." *Instructional Science* 18 (2): 119–144. https://doi.org/10.1007/BF00117714.

Sadler, D. R. 2009. "Indeterminacy in the Use of Preset Criteria for Assessment and Grading." *Assessment & Evaluation in Higher Education* 34 (2): 159–179. https://doi.org/10.1080/02602930801956059.

Schneider, M., and F. Preckel. 2017. "Variables Associated with Achievement in Higher Education: A Systematic Review of Meta-Analyses." *Psychological Bulletin* 143 (6): 565–600. https://doi.org/10.1037/bul0000098.

Seßler, M., F. Tuma, M. Möller, T. Zöller, and S. Schütt. 2025. "Can AI Grade Your Essays? Large Language Models and Teacher Ratings in Higher Education." In *Proceedings of the 15th International Conference on Learning Analytics & Knowledge (LAK25)*, Dublin. https://doi.org/10.1145/3706468.3706527.

Sreedhar, Radhika, Linda Chang, Ananya Gangopadhyaya, Peggy Woziwodzki Shiels, Julie Loza, Euna Chi, Elizabeth Gabel, et al. 2025. "Comparing Scoring Consistency of Large Language Models with Faculty for Formative Assessments in Medical Education." *Journal of General Internal Medicine* 40 (1): 127–134. https://doi.org/10.1007/s11606-024-09050-9.

Stasinopoulos, D. M., and R. A. Rigby. 2007. "Generalized Additive Models for Location Scale and Shape (GAMLSS). *R*." *Journal of Statistical Software* 23 (7): 1–46. https://doi.org/10.18637/jss.v023.i07.

SURF. 2026. "DPIA Microsoft 365 Copilot for Education: Data Protection Impact Assessment on the Processing of Personal Data with Microsoft 365 Copilot for Education." https://vendorcompliance.surf.nl/wp-content/uploads/2026/05/20260527-SURF-2nd-Update-DPIA-Microsoft-365-Copilot-public-version.pdf

Talmi, Deborah, Giulio Corsi, Alexandru Marcoci, Yael Benn, Mohammad Abo-Tabik, Roni Tibon, Laila Razzaq, Sarah Mattey, et al. 2026. "AI in University Assessment: Evaluating the Opportunities and Risks of Automated Marking." https://www.emotional-cognition.psychol.cam.ac.uk/opraise

R Core Team. 2026. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Tensen, D., M. D. Carey, and P. Grainger. 2026. "Comparing ChatGPT-5 and Other GenAI Tools in Doctoral Confirmation: Variability, Persona Effects, and Alignment with Human Feedback." *Assessment & Evaluation in Higher Education*: 1–17. https://doi.org/10.1080/02602938.2026.2619902.

Thomas, M. L., S. N. Yildirim-Erbasli, and S. Hariharan. 2026. "Exploring Undergraduate Students' Perceptions of AI vs. Human Scoring and Feedback." *The Internet and Higher Education* 68: 101052. https://doi.org/10.1016/j.iheduc.2025.101052.

Tisi, J., G. Whitehouse, S. Maughan, and N. Burdett. 2013. "A Review of Literature on Marking Reliability Research (Report for Ofqual)." Slough. https://assets.publishing.service.gov.uk/media/5a81cc0e40f0b62302699360/0613_JoTisi_et_al-nfer-a-review-of-literature-on-marking-reliability.pdf

Usher, M., and M. Faraon. 2025. "Who Grades Best? Comparing ChatGPT, Peer, and Instructor Evaluations across Varying Levels of Student Project Quality." *Assessment & Evaluation in Higher Education*: 1–20. https://doi.org/10.1080/02602938.2025.2588682.

Webb, D. J., C. A. Paul, and M. K. Chessey. 2020. "Relative Impacts of Different Grade Scales on Student Success in Introductory Physics." *Physical Review Physics Education Research* 16 (2): 020114. https://doi.org/10.1103/PhysRevPhysEducRes.16.020114.

Winstone, N. E., K. Gravett, and S. Elkington. 2026. "Black Box Assessment: Rethinking Integrity and Learning for a Time of Generative AI." *Assessment & Evaluation in Higher Education:* 1–18. https://doi.org/10.1080/02602938.2026.2661365.

Xia, Q., X. Weng, F. Ouyang, T. J. Lin, and T. K. F. Chiu. 2024. "A Scoping Review on How Generative Artificial Intelligence Transforms Assessment in Higher Education." *International Journal of Educational Technology in Higher Education* 21 (1). https://doi.org/10.1186/s41239-024-00468-z.

Yan, Lixiang, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, et al. 2024. "Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review." *British Journal of Educational Technology* 55 (1): 90–112. https://doi.org/10.1111/bjet.13370.

Yang, B. 2026. "If GAI Has Finished Her Job, Why Would I Need Her?': A Mixed-Methods Study on Students' Perceptions of GAI-Using Teachers." *Studies in Higher Education*: 1–15. https://doi.org/10.1080/03075079.2026.2648696.

Yavuz, F., Ö. Çelik, and G. Yavaş Çelik. 2025. "Utilizing Large Language Models for EFL Essay Grading: An Examination of Reliability and Validity in Rubric-Based Assessments." *British Journal of Educational Technology* 56 (1): 150–166. https://doi.org/10.1111/bjet.13494.

Yorke, M., P. Bridges, and H. Woolf. 2000. "Mark Distributions and Marking Practices in UK Higher Education." *Active Learning in Higher Education* 1 (1): 7–27. https://doi.org/10.1177/1469787400001001002.

Zhang, K., S. Li, and Y. Huang. 2026. "Generative AI-Driven Feedback in Higher Education: A Scoping Review." *Assessment & Evaluation in Higher Education*: 1–18. https://doi.org/10.1080/02602938.2026.2617157.

Zhao, C. 2024. "AI-Assisted Assessment in Higher Education: A Systematic Review." *Journal of Educational Technology and Innovation* 6 (4). https://doi.org/10.61414/jeti.v6i4.209.