Prompt Sensitivity in LLM-Based Essay Scoring is Model-Specific

Giulio Corsi
University of Cambridge
Cambridge, UK
gc540@cam.ac.uk

Alexandru Marcoci
University of Cambridge
Cambridge, UK
am3159@cam.ac.uk

Maryam Abo-Tabik
Univ. of Central Lancashire
Preston, UK
mabo-tabik@lancashire.ac.uk

Roni Tibon
University of Nottingham
Nottingham, UK
Roni.Tibon@nottingham.ac.uk

Yael Benn
Manchester Metropolitan Univ.
Manchester, UK
Y.Benn@mmu.ac.uk

Deborah Talmi
University of Cambridge
Cambridge, UK
dt492@cam.ac.uk

ABSTRACT

The emergence of Large Language Models (LLMs) has renewed interest in automated essay scoring, yet a barrier to adoption persists: LLM-generated scores are sensitive to prompt formulation, and it is unclear which aspects of a prompt drive this effect or whether findings from one model or institution transfer to another. For educators and institutions considering deployment, this uncertainty is consequential, as without understanding the structure of prompt sensitivity, there is no principled basis for deciding whether to trust an LLM-generated score. This study evaluates three frontier LLMs (GPT-5.4, Claude Opus 4.6, Gemini 3 Flash) across a $3 \times 3 \times 3$ factorial design varying criteria specificity, calibration intervention, and scoring strategy on a calibration sample of 153 undergraduate essays from three UK universities with distinct marking conventions. A main-effects ANOVA on the 27 per-model conditions reveals that prompt sensitivity is substantial but strikingly model-specific: criteria specificity accounts for the largest share of condition-level variance for Gemini (57%, $p < .001$), scoring strategy for GPT (35%, $p < .01$), and scoring strategy and calibration together for Claude (32% and 24% respectively), with no single dimension dominating across models. The optimal configuration also differs across institutions, even within the same model. At each institution, the gap between best and worst configurations spans 4–5 RMSE points on the calibration set, with the worst performing worse than a naïve baseline. For the models tested, these findings indicate that no universal prompt recipe exists, and that responsible deployment requires empirical prompt calibration as a prerequisite rather than an optional refinement – a process more similar to hyperparameter tuning than to simple instruction design.

Giulio Corsi, Alexandru Marcoci, Maryam Abo-Tabik, Georgiana Thorpe Apreutesei, Lyba Razzaq, Roni Tibon, Yael Benn, and Deborah Talmi. Prompt Sensitivity in LLM-Based Essay Scoring is Model-Specific. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 760–764. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21039772

Keywords

automated essay scoring, large language models, prompt engineering, hyperparameter tuning

1. INTRODUCTION

The emergence of LLMs has renewed interest in automated essay scoring, with recent work demonstrating that individual models can approximate human scores on standardised tasks [5, 3]. However, a practical barrier to deployment persists: LLM-generated scores are sensitive to prompt formulation, with minor variations in rubric framing or role instructions capable of shifting outputs by a full grade band [1]. For institutions considering the use of LLM-based scoring, this sensitivity is consequential yet poorly understood, and it is unclear whether the effect follows patterns that can inform principled prompt design or whether it is contingent on the specific model and context in use.

Despite this known limitation of LLM-based grading, prompt design for essay scoring is typically treated as a fixed authoring choice, selected on the basis of practitioner intuition. Where studies do evaluate prompt effects, they tend to do so by sequentially adding prompt components within a single model and selecting whichever configuration maximises agreement – an approach that can identify that prompts matter, but not which dimensions of a prompt matter, nor whether those dimensions generalise across models [6, 4]. This leaves two open questions. The first is about decomposition: given that prompt formulation affects LLM-generated scores, which types of prompt intervention are actually responsible for the effect, and how much of the variance does each explain? The second is about generalisability: do the answers to the first question hold across models, or is prompt sensitivity model-contingent in ways that make universal recommendations impossible? The distinction is consequential: if one prompt dimension consistently dominates, practitioners can follow a simple prompting framework; if the dominant dimension varies by model, then empirical calibration is not a convenience but a necessity.

This study addresses that question directly. Three prompt dimensions – targeting what evaluative structure the model

receives, how it is distributionally anchored, and how it is asked to reason – are independently varied in a $3 \times 3 \times 3$ factorial design across three LLMs, yielding 81 model-prompt conditions evaluated on a calibration set of 153 undergraduate essays from three UK institutions. The factorial structure permits decomposition of not only which prompt components matter, but whether they matter in the same way across models.

The contribution of this work, and what attendees will see at the poster, is twofold. First, a per-model variance decomposition that makes the dimensions of prompt sensitivity visible rather than aggregating them into a single "prompts matter" headline. Second, an empirical demonstration that the optimal prompt configuration differs not only across models but also across institutions with distinct marking conventions, with the practical implication that institutions cannot import a prompt validated elsewhere and expect it to behave consistently on their own essays. The poster will display the full grid of 81 conditions alongside the per-institution best/worst gaps, allowing attendees to inspect the model-specific patterns directly and to consider what an empirical calibration workflow would look like in their own context.

2. METHODS

2.1 Dataset

The data is drawn from a corpus of 764 long-form undergraduate psychology essays from three UK universities, spanning 50 modules. All were online submissions of coursework and exam scripts marked by human assessors using the standard UK percentage scale with discrete tick marks (e.g., 62, 65, 68, 72; $M = 63.5$, $SD = 10.1$). A stratified 80/20 split produced a calibration set ($n = 153$, approximately 50 per institution) and a held-out test set ($n = 611$). All analyses in this paper use the calibration set only. The decision to reserve the test set is deliberate rather than incidental: the research question addressed here concerns the structure of prompt sensitivity, which requires evaluating the same essays under all 81 model–prompt conditions and decomposing the resulting variance. The held-out essays are reserved for a separate, downstream question – whether configurations selected through calibration generalise to unseen submissions – which will be addressed in subsequent work and would be premature to evaluate before the present decomposition is established.

2.2 Factorial Design

The three LLMs used – GPT-5.4 (OpenAI), Claude Opus 4.6 (Anthropic), and Gemini 3 Flash (Google) – were selected from different providers to maximise diversity of biases and failure modes. Three prompt dimensions were varied in a fully crossed $3 \times 3 \times 3$ design (Figure 1). The dimensions were chosen to span three theoretically distinct sources of prompt influence: *what* evaluative structure the model receives (criteria specificity), *how* it is distributionally anchored (calibration intervention), and *how* it is asked to arrive at a judgement (scoring strategy). Together, these cover the input-side, distributional, and process-level determinants of LLM scoring behaviour.

(A) Criteria specificity addresses how much evaluative structure the model receives. Prior work has shown that rubric provision can either improve consistency or introduce dimension level artefacts [1], but the relationship between rubric granularity and accuracy has not been isolated. Three levels were tested: no rubric (A1), summary grade boundaries (A2; e.g., "70+: demonstrates critical analysis and original insight"), and the full analytic rubric for a given assignment with per-dimension performance descriptors (A3).

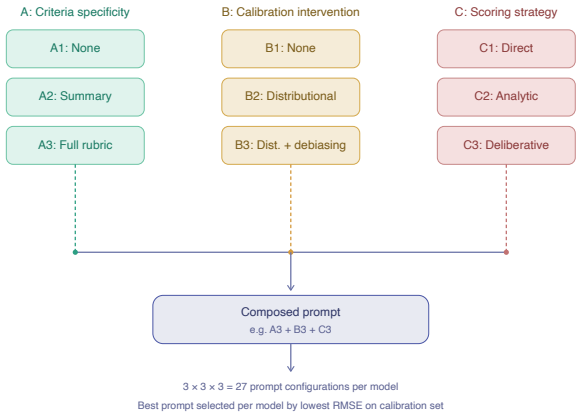

Figure 1: Factorial prompt composition. Three dimensions are crossed in a $3 \times 3 \times 3$ design, yielding 27 prompt configurations per model.

(B) Calibration intervention addresses the well-documented tendency of LLMs to compress scores towards the centre of the scale [2], which can in principle be mitigated by providing distributional anchoring. Three levels were tested: no calibration (B1), distributional base-rate information (B2; e.g., "approximately 25% of essays in this cohort receive a First"), and distributional information with an explicit debiasing instruction (B3; e.g., "...be aware that LLMs tend to compress scores towards the mean").

(C) Scoring strategy addresses how the model is asked to arrive at its judgement, drawing on the distinction between analytic and holistic scoring approaches in educational assessment. Three levels were tested: holistic-direct scoring with no structured reasoning (C1), analytic-decomposed scoring in which the model evaluates the essay separately on substantive quality, use of evidence, and clarity of expression before synthesising a single overall score (C2), and holistic-deliberative scoring in which the model first identifies the appropriate grade band, weighs evidence for and against that classification, and then assigns a specific mark within it (C3).

Each model scored the calibration set ($n = 153$) under all 27 conditions. The best-performing prompt per model was selected by lowest RMSE, independently for each institution. Both prompt selection and the variance decomposition reported below use this same calibration partition. This is appropriate to the research question – which prompt dimensions drive scoring variance – but means that the per-institution best/worst comparisons reflect in-sample performance and are subject to optimistic selection bias, as discussed in Section 3.1.

Table 1: Best and worst model–prompt conditions per institution on the calibration set ($n = 153$). Naïve baseline = predicting the cohort mean. Best/worst are selected as min/max RMSE across 27 conditions; the observed gap is therefore an in-sample upper bound on the true best–worst difference (see text).

Institution	Condition	Model	RMSE	BA (%)	Δ RMSE	Δ BA
Inst. A	Best (A3_B3_C1)	Gemini	5.84	74.1		
	Worst (A1_B1_C1)	Claude	10.94	48.1	+5.09	−25.9
	Naïve baseline	–	6.95			
Inst. B	Best (A3_B1_C3)	GPT	9.74	42.4		
	Worst (A3_B2_C1)	Gemini	13.89	32.2	+4.16	−10.2
	Naïve baseline	–	11.44			
Inst. C	Best (A1_B3_C1)	Gemini	6.40	50.0		
	Worst (A1_B1_C2)	Claude	10.19	32.4	+3.79	−17.6
	Naïve baseline	–	7.02			

3. RESULTS

3.1 The Magnitude of Prompt Effects

Before decomposing which prompt dimensions drive scoring variance, it is important to establish the practical magnitude of the effects at stake. Table 1 presents the best and worst model–prompt conditions for each institution on the calibration set. The in-sample gap between optimal and worst-case conditions spans 3.8–5.1 RMSE points and up to 26 percentage points of band accuracy. At Institution A, the best condition (Gemini, A3_B3_C1) achieves an RMSE of 5.84 with 74.1% band accuracy; the worst (Claude, A1_B1_C1) yields 10.94 and 48.1%.

An important caveat applies: because the "best" and "worst" configurations are selected as the extremes across 27 conditions, the observed gap is subject to selection optimism and will tend to overstate the true best–worst difference on new data. The gap is too large to be explained entirely by sampling variability – the best conditions consistently fall below the naïve baseline while the worst fall above it – but the precise magnitude should be interpreted as an upper bound. Future work will evaluate whether the gap holds on the reserved test set ($n = 611$).

The comparison with the naïve baseline – predicting the cohort mean for every essay – is what makes these differences consequential rather than merely descriptive. The best-calibrated conditions fall below this baseline, indicating that they extract genuine evaluative signal from the essay text. The worst conditions fall clearly above it, meaning that a poorly chosen prompt actively destroys signal that a trivial heuristic would preserve. The difference between a good and bad configuration is therefore not one of marginal refinement but of whether the system functions at all – a point that the poster will make visible by displaying the full distribution of the 27 per-model conditions against the baseline, so that the magnitude of the best–worst spread can be inspected rather than inferred.

Table 2: Condition-level ANOVA: proportion of variance (η^2) in per-condition RMSE explained by each prompt dimension, estimated separately for each model ($n = 27$ conditions per model, 20 residual df). Bootstrap 95% CIs (10,000 resamples) in brackets. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Dimension	Claude	GPT	Gemini
A: Criteria	ns	ns	57%***
			[29, 78]
B: Calibration	24%*	ns	ns
	[3, 48]		
C: Strategy	32%**	35%**	ns
	[6, 56]	[8, 60]	

3.2 Prompt Sensitivity is Model-Specific

The central question of this work is whether prompting effects follow general patterns – in which case a universal best-practice recommendation would be possible – or whether they are model-contingent. A condition-level ANOVA on the 27 per-model RMSE values (Table 2) decomposes variance by prompt dimension under a main-effects model. An important caveat applies throughout: with only 27 observations per model and 20 residual degrees of freedom, the η^2 values are point estimates with wide confidence intervals, not precise effect sizes; we report bootstrap 95% CIs (10,000 resamples) to make this uncertainty explicit. With that in mind, the pattern across models is striking. For Gemini, criteria specificity accounts for the largest share of condition-level variance (57%, 95% CI [29, 78], $p < .001$). For GPT, scoring strategy accounts for 35% (95% CI [8, 60], $p < .01$). For Claude, scoring strategy and calibration together account for over half (32%, 95% CI [6, 56], $p < .01$; and 24%, 95% CI [3, 48], $p < .05$). No single dimension is consistently dominant, and dimensions that are significant for one model are non-significant for another.

Two-way interactions (A×B, A×C, B×C) were tested in a full model including all main effects and two-way terms (residual $df = 8$). No individual interaction reached significance at $\alpha = .05$ for any of the three models, and the

combined interaction terms did not improve model fit over the main-effects specification. Given the limited residual degrees of freedom, these null results should not be taken as evidence that interactions are absent, but they support treating the main-effects decomposition as a reasonable approximation.

This pattern is corroborated by a repeated-measures analysis at the essay level ($n = 153$, each scored under all 27 conditions). Gemini shows the largest total prompt sensitivity (partial $\eta^2 = 2.0\%$, driven by criteria at 1.5%, $p < .001$), Claude shows moderate sensitivity (1.2%, spread across calibration and strategy), and GPT shows the least (0.4%). The within-essay effects are naturally small in relative terms – essay quality dominates individual score variance, as it should. But the practical consequence of prompt sensitivity operates at the system level, not the essay level: small per-essay biases that are systematic across a cohort aggregate into the 4–5 point RMSE differences reported in Table 1. A prompt configuration that introduces a consistent directional bias, for instance, may shift each essay by only 2–3 marks, but across a cohort this is enough to meaningfully alter the RMSE. The model-specificity of the pattern is consistent across both levels of analysis, confirming that the condition-level decomposition reflects a real structural difference rather than an artefact of low power.

With that caveat, the practical implication is clear: scoring strategy appears as the largest single dimension for both Claude and GPT, but the two models differ in which other dimensions matter (calibration for Claude, none for GPT), and Gemini is driven by an entirely different dimension (criteria specificity). A practitioner who reads that "scoring strategy is the most important prompt dimension" and applies this as a general rule will miss the dimension that matters most for Gemini and misconfigure calibration for Claude. For the three models tested here, universal prompt guidelines do not merely oversimplify – they actively mislead.

4. CONCLUSION

Prompt sensitivity in LLM-based essay scoring appears to be real, substantial, and model-specific. No single prompt dimension dominates across the three models tested: scoring strategy is the primary driver for both Claude and GPT, but Claude is additionally sensitive to calibration, while Gemini is driven by an entirely different dimension – criteria specificity. The gap between optimal and worst-case configurations spans 4–5 RMSE points per institution on the calibration set, with poorly chosen prompts performing worse than a naïve baseline – in effect, making the difference between a scoring system that extracts evaluative signal and one that removes it completely. For the models tested here, the results suggest that no universal prompt recipe exists, and that empirical calibration may be better understood as a prerequisite for considering deployment than as an optional refinement.

This pattern carries implications that extend beyond predictive accuracy. Because the optimal configuration varies across both models and institutions, a prompt validated in one setting and imported into another may yield systematically different scores for substantively similar work, with consequences for grading consistency at the cohort level and for the comparability of marks across institutions. The institution-level variation in optimal configurations also suggests that prompt sensitivity is not a purely technical artefact but interacts with local marking conventions, which raises fairness considerations when LLM-generated scores are used to compare students who have been assessed under different rubrics or grade distributions. These implications are tentative given the calibration-only scope of the present analysis, but they indicate that the choice of prompt may need to be treated as part of the assessment design itself rather than as a downstream implementation detail. Attendees at the poster will be able to inspect the per-institution differences directly and to consider what an empirical calibration workflow might look like in their own assessment context.

Several limitations constrain these conclusions. All reported comparisons are in-sample: the "best" configuration per model-institution pair is selected by lowest RMSE on the same calibration set used for evaluation, introducing optimistic selection bias. The magnitude of the best–worst gap should therefore be read as an upper bound, and the findings as preliminary evidence rather than deployment-ready conclusions; held-out validation on the reserved test set ($n = 611$) is needed to establish a de-biased estimate. The calibration set comprises approximately 50 essays per institution, yielding η^2 estimates with wide confidence intervals (Table 2); the qualitative pattern of model-specificity is more robust than any individual percentage. Interaction effects between prompt dimensions were tested and found to be non-significant, but with only 8 residual degrees of freedom the test has limited power, and the per-dimension attributions should be treated as approximate. The findings are limited to three models from three providers evaluated at a single point in time, and to undergraduate psychology essays marked under UK conventions; whether the model-specificity pattern extends to other disciplines, marking scales, or model generations warrants further investigation. These limitations notwithstanding, the core practical implication appears clear: prompt configuration for LLM-based scoring may be better treated as an empirical calibration problem than as an authoring decision.

5. ADDITIONAL AUTHORS

Additional authors: Georgiana Thorpe Apreutesei (University of Cambridge, email: `gt485@cam.ac.uk`) and Lyba Razzaq (Manchester Metropolitan Univ., email: `lyba.razzaq@stu.mmu.ac.uk`).

6. REFERENCES

[1] M. Abujadallah, M. Saad, and S. Abudalfa. Evaluating open-source llms for automated essay scoring: The critical role of prompt design. 2025.

[2] S. Lee, Y. Cai, D. Meng, Z. Wang, and Y. Wu. Unleashing large language models' proficiency in zero-shot essay scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198, 2024.

[3] W. Li and H. Liu. Applying large language models for automated essay scoring for non-native japanese.

Humanities and Social Sciences Communications, 11(1):1–15, 2024.

[4] P. Y. Liew and I. K. Tan. On automated essay grading using large language models. In *Proceedings of the 2024 8th international conference on computer science and artificial intelligence*, pages 204–211, 2024.

[5] A. Mizumoto and M. Eguchi. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050, 2023.

[6] W. Xia, S. Mao, and C. Zheng. Empirical study of large language models as automated essay scoring tools in english composition_taking toefl independent writing task for example. *arXiv preprint arXiv:2401.03401*, 2024.