Stability of LLM-Based Essay Scoring Under Input Reordering and Few-Shot Prompting

Arnav Palkhiwala*, Maxx Ferrian*, Mohit Mohanraj*

*University of Washington

Keywords: Maximum 5 keywords placed before the abstract.

Abstract

[Placeholder]

1 Introduction

With the growing capabilities of large language models (LLMs) to process and evaluate large volumes of data, educators and reviewers are increasingly turning to these tools to reduce grading workload and generate scores efficiently. In academic settings, including classroom assessment, college admissions, and research conference review, LLMs offer substantial gains in speed and scalability. While automated grading significantly improves efficiency, important questions arise regarding the quality and reliability of these evaluations, particularly whether models produce consistent outputs across runs. The stakes of these determinations are high, as they directly influence academic outcomes, admissions decisions, and conference acceptances, which are contexts where fairness and consistency are critical. Ensuring that LLM-based grading systems are stable and robust is therefore essential before broader deployment in high-impact decision-making environments.

Previous literature demonstrates that LLMs are sensitive to input ordering in structured tasks. In multiple-choice settings, simply reordering answer options has been shown to produce substantial performance gaps, ranging from **13–75**% [1]. This finding highlights the extent to which LLMs exhibit positional bias when selecting among a list of options. Related work further shows that reshuffling inputs across tasks such as paraphrasing and relevance judgment can also degrade performance, and while few-shot prompting may reduce this sensitivity, it does not eliminate the effect entirely [2]. Studies comparing zero-shot and few-shot prompting in specialized domains, such as medical clinics and biotech, indicate that model architecture and prompting strategy both play a significant role in performance quality [3]. However, most prior work focuses on classification, reasoning, or option-selection tasks. Compared to these settings, essay grading represents a more holistic and complex evaluation problem. Essays require longer contextual understanding, more nuanced reasoning, and the assignment of a continuous score rather than selection from fixed alternatives. As a result of the entire evaluative and scoring responsibility resting with the LLM, determining whether order-related biases persist in such complex judgment tasks is crucial.

In this paper, we extend prior work by investigating whether the order in which essays are presented within a grading batch affects the score assigned to each essay. Specifically, we examine scenarios in which multiple essays are evaluated simultaneously and test whether reordering the same set of essays leads to changes in assigned grades. In addition to measuring order effects, we analyze how prompting strategy influences model stability by comparing zero-shot prompting with few-shot prompting. Our examples for this include essays representing exceptional, average, and poor quality writing. We define consistency as the stability of assigned scores when identical essays are presented in different input sequences, with or without supporting demonstrations. While previous research has documented order sensitivity in classification and reasoning tasks, its implications for holistic evaluation tasks such as automated essay grading remain underexplored. We provide a systematic analysis of input-order sensitivity in LLM-based automated grading under both zero-shot and few-shot prompting conditions.

2 Experimental Design

To evaluate the consistency and potential biases of LLM-based grading under different prompting configurations, we designed a controlled experimental framework that systematically varies both document order and prompting strategy. The primary objective of this study is to assess whether grading outcomes differ when documents are presented in ascending versus descending order and when the model is prompted using zero-shot versus few-shot techniques. All experiments were conducted using the **Gemini-2.5-Flash-Lite [TO BE UPDATED]** model, accessed via the Gemini API under a fixed inference configuration.

Documents were sampled from our database without replacement in batches of ten at a time. Sampling without replacement ensured that each document was evaluated only once within a given experimental cycle, thereby preventing duplication and preserving independence across observations. This batch-based sampling procedure was repeated iteratively until all documents in the dataset had been evaluated.

The dataset was constructed to span a broad range of writing quality while maintaining diversity in content and complexity. Poor-quality essays consisted of K–6 narrative writing samples. Average-quality essays included 10th–12th grade and undergraduate-level essays drawn from English courses and

college application writing. Exceptional-quality essays were drawn from introductions and conclusions of research papers in science and technology domains sourced from arXiv. Our dataset consists of **100** examples for each essay quality type, giving us **300** essays that are being scored in total. This stratified sampling approach ensured clear variation in writing proficiency across the evaluation set.

Each sampled document was graded by the LLM under four distinct experimental conditions. Specifically, every document received a score under ascending order with zero-shot prompting, descending order with zero-shot prompting, ascending order with few-shot prompting, and descending order with few-shot prompting. In all conditions, the model assigned a numerical grade on a scale from **1 to 10**, where 1 represented the lowest performance and 10 represented the highest performance according to the evaluation criteria defined in the prompt. The grading criteria focuses on writing quality over content as we did not want specific topics to be confounded with higher scores. Instead, the primary goal is simply looking at writing flow scores and if they remain the same under different conditions.

LLM Grading Prompt

```
You are a college-level writing instructor
grading a student reflection essay.

Grade holistically using these dimensions:

• 1) Flow (clarity/readability across
  sentences)
• 2) Transitions (connections between
  ideas/paragraphs)
• 3) Content Quality & Focus (relevance,
  specificity, alignment to the week's
  concepts)
• 4) Spelling & Grammar (correctness,
  professionalism)
• 5) Knowledge & Depth (understanding +
  thoughtful engagement)
• 6) Structure (organization, paragraphing,
  logical progression)

Score rules:

• - Provide ONE overall score from 1 to 10
  (integers only).
• - 1 = fundamentally flawed; minimal
  understanding; very poor writing.
• - 5 = mixed/adequate; some understanding;
  clear weaknesses.
• - 10 = exemplary; polished, insightful,
  well-structured; no meaningful
  weaknesses.
• - Avoid score inflation:  9--10 only if
  nearly flawless.
```

Few-shot only (inserted before essays):

```
Calibration anchors (use these to calibrate
your scoring scale):

[ANCHOR A --- 10/10 EXAMPLE]

{anchor_good}

[ANCHOR B --- 5/10 EXAMPLE]

{anchor_mid}

[ANCHOR C --- 1/10 EXAMPLE]

{anchor_bad}

Now grade EACH essay below independently.

IMPORTANT RULES:

• - Return a JSON ARRAY only.
• - The array must have EXACTLY
  {len(essays)} items.
• - Items must be in the SAME ORDER as the
  essays appear below.
• - Use the essay ID exactly as provided.
• - pred_score must be an integer 1--10.

Essays (in order):

{essays_block}

Output STRICT JSON ONLY (no markdown, no
extra text).

Each object must match:

{"id":"<string>","pred_score":<integer
1-10>,"justification":"<2-4 sentences>"}
```

The ascending and descending order conditions refer to the sequence in which documents were presented to the model within a batch, allowing us to test for potential order effects in grading behavior. The zero-shot condition provided only task instructions, while the few-shot condition included example graded documents intended to guide the model's evaluation. The few-shot prompt included one exemplar essay from each quality category (poor, average, and exceptional) to provide representative calibration across the writing spectrum.

For each document, we recorded its unique identifier along with the four grades assigned under the different experimental conditions. The LLM outputs for each batch is stored in JSON format file containing **document ID, score, and a short justification**. All results were then stored in a structured dataset indexed by document ID and condition, enabling direct within-document comparisons across prompting strategies and ordering configurations. Because each document was evaluated under all four conditions, the design supports paired statistical analyses that control for document-level variability.

The final dataset consists of four numerical scores per document, corresponding to the four prompting conditions. These

data serve as the basis for subsequent statistical analysis, including paired comparisons and variance analyses, to determine whether prompting strategy, document order, or their interaction significantly influences grading outcomes.

References

[1] Pouya Pezeshkpour and Estevam Hruschka. Large Language Models Sensitivity to the Order of Options in Multiple-Choice Questions. 2023. https://arxiv.org/abs/2308.11483

[2] Bryan Guan, Tanya Roosta, Peyman Passban, and Mehdi Rezagholizadeh. The Order Effect: Investigating Prompt Sensitivity to Input Order in LLMs. 2025. https://arxiv.org/abs/2502.04134

[3] Yanis Labrak, Mickael Rouvier, and Richard Dufour. A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks. 2023. https://arxiv.org/abs/2307.12114

Acknowledgements