Evaluating Open-Source LLMs for Automated Essay Scoring: The Critical Role of Prompt Design

Mahmoud Abujadallah * , Motaz Saad , Shadi Abudalfa

Posted Date: 19 November 2025

doi: 10.20944/preprints202511.1429.v1

Keywords: automated essay scoring (AES); large language models (LLMs); prompt engineering

 doi:10.20944/preprints202511.1429.v1

Article

Evaluating Open-Source LLMs for Automated Essay Scoring: The Critical Role of Prompt Design

Mahmoud Abujadallah [1,*], Motaz Saad [2] and Shadi Abudalfa [3]

[1] Department of Information Technology, University College of Applied Sciences, Gaza, Palestine
[2] Department of Data Science, Faculty of Information Technology, Islamic University of Gaza, Gaza Strip, Palestine
[3] SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI), King Fahd University of Petroleum & Minerals, Academic Belt Road, Dhahran, 31261, Kingdom of Saudi Arabia
* Correspondence: m.abujadallah@gmail.co

Abstract

This paper evaluates the Automated Essay Scoring (AES) performance of five open-source Large Language Models (LLMs)—LLaMA 3.2 3B, DeepSeek-R1 7B, Mistral 8×7B, Qwen2 7B, and Qwen2.5 7B—on the PERSUADE 2.0 dataset. We assess each model under three distinct prompting strategies: (1) rubric-aligned prompting, which embeds detailed, human-readable definitions of each scoring dimension; (2) instruction-based prompting, which names the criteria and assigns a grading role without elaboration; and (3) a minimal instruction-based variant, which omits role priming and provides only a concise directive. All prompts constrain the output to a single numerical score (1–6) to ensure comparability. Performance is measured using standard AES metrics, including Exact Match, F1 Score, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Pearson and Spearman correlation coefficients, and Cohen's k. Results demonstrate that prompt design critically influences scoring accuracy and alignment with human judgments—with rubric-aligned prompting consistently outperforming instruction-based alternatives. Among the models, DeepSeek-R1 7B and Mistral 8×7B achieve the strongest overall results: DeepSeek-R1 attains the highest F1 Score (0.93), while Mistral 8×7B leads in correlation with human scores (Pearson = 0.863, Spearman = 0.831). Human comparison experiments further confirm that rubric-aligned prompting yields the closest alignment with expert graders. These findings underscore the potential of lightweight, open-source LLMs for reliable and equitable educational assessment, while highlighting explicit rubric integration—not model scale—as the key driver of human-aligned AES performance.

Keywords: automated essay scoring (AES); large language models (LLMs); prompt engineering

1. Introduction

Automated Essay Scoring (AES) has long held promise as a scalable solution for formative and summative assessment in educational settings. By providing timely, consistent, and cost-effective feedback, AES systems can alleviate instructor workloads and expand access to high-quality writing evaluation.

Large Language Models (LLMs) have transformed artificial intelligence by enabling advanced language understanding, reasoning, and generation across domains such as healthcare, finance, and education [1].

Despite significant progress in areas such as contextual reasoning and ethical alignment, many current evaluation strategies remain narrow in scope. They often rely on limited benchmarks or assess individual models in isolation, failing to capture broader issues such as algorithmic bias, and performance scalability across diverse linguistic and cultural contexts [2].

AES represents a particularly compelling use case for LLMs in education, with the potential to address persistent challenges in manual essay grading namely inconsistency, subjectivity, and inefficiency. Traditional grading methods often suffer from grader fatigue, implicit biases, and delayed

 doi:10.20944/preprints202511.1429.v1

feedback, all of which can hinder both the reliability of assessment and student learning outcomes [3]. LLM-powered AES systems offer scalable, fast, and consistent alternatives that, when aligned with robust scoring rubrics, can approach or even exceed the performance of human raters [4].

Initially, this study intended to leverage both the Automated Student Assessment Prize (ASAP) 2.0 and PERSUADE 2.0 datasets. While ASAP offers a single holistic score that bundles various writing dimensions together [5], PERSUADE 2.0 separates the evaluation into six distinct criteria Position, Claim, Evidence, Counterclaim, Rebuttal, and a Holistic score enabling finer-grained assessment and interpretability for argumentative writing tasks. Given the focus of this study on prompt customization strategies as detailed in the Experiments chapter, PERSUADE 2.0 provides a more suitable framework for evaluating how well Large Language Models respond to targeted prompts, especially those aligned with argumentative structure. [6]

Crucially, emerging evidence suggests that how we prompt LLMs may matter as much as—or more than—the models themselves. Techniques such as Chain-of-Thought (CoT) reasoning or explicit rubric integration have shown promise in structured scoring tasks, but their impact across diverse open-source architectures has not been systematically evaluated in the context of essay assessment.

To address these gaps, we present a comprehensive evaluation of five representative open-source LLMs—LLaMA 3.2 3B, DeepSeek-R1 7B, Mistral 8×7B, Qwen2 7B, and Qwen2.5 7B—for AES on the PERSUADE 2.0 dataset.

We benchmark their performance under three prompting strategies: (1) basic instruction-based prompting, (2) Chain-of-Thought (CoT) prompting, and (3) rubric-aligned prompting that explicitly embeds scoring criteria. Using a suite of standard AES metrics—including F1 Score, Mean Absolute Error (MAE), Pearson and Spearman correlations, and Cohen's k—we assess both scoring accuracy and alignment with human judgments. In addition, we conduct human comparison studies to evaluate which prompting approach yields outputs most consistent with expert graders.

The rest of this paper is organized as follows: Section 2 review related works, Section 3 describe our approach, Section 4 presents experimental results. Finally, Section 5 draw conclusions.

2. Related Works

Automated Essay Scoring (AES) has undergone a fundamental transformation over the past two decades—from rule-based systems relying on surface-level features to modern Large Language Models (LLMs) capable of nuanced semantic and rhetorical analysis. This section situates our work within this trajectory, reviewing (i) traditional and neural AES approaches, (ii) the rise of LLMs in educational assessment, and (iii) the emerging role of prompt design. We conclude by identifying key gaps that motivate our study.

2.1. Traditional AES Approaches

Early AES systems such as Project Essay Grade (PEG) [7] and e-rater [8] relied on handcrafted linguistic features—including grammar counts, lexical diversity, and syntactic complexity—combined with regression or decision tree models. While interpretable and effective for constrained tasks, these systems struggled to capture higher-order writing traits such as argument coherence, logical development, or rhetorical sophistication. Later machine learning methods, including SVMs, Random Forests, and Gradient Boosting [9,10], improved generalization but remained bottlenecked by manual feature engineering, limiting their scalability and cross-prompt adaptability. As a result, these systems often failed to align with holistic or analytic rubrics, particularly for argumentative writing [11,12].

2.2. Neural and Deep Learning Models

Neural architectures such as LSTMs [13] and CNNs reduced reliance on handcrafted features by learning representations directly from raw text. These models captured sequential and local discourse patterns more effectively but still struggled with long-range dependencies and holistic reasoning. Transformer-based architectures [14] addressed these limitations through self-attention mechanisms, enabling superior contextual modeling and paving the way for LLM-based AES.

 doi:10.20944/preprints202511.1429.v1

2.3. LLMs in Automated Essay Scoring

Large Language Models—including GPT, BERT, LLaMA, Mistral, DeepSeek, and Qwen—leverage large-scale pretraining to capture rich linguistic and semantic patterns [15–17]. In AES, they support both holistic and analytic scoring, as well as feedback generation and rubric alignment [18,19]. Recent studies show that GPT-4 can achieve human-level agreement in specific contexts [20,21]. However, reliance on closed, proprietary models raises concerns about transparency, reproducibility, and equitable access—especially in public education. Consequently, open-source LLMs offer a compelling alternative, yet research on their AES capabilities remains sparse, particularly in systematic, multi-model comparisons under standardized conditions.

2.4. Prompting Strategies in AES

Prompt engineering has emerged as a critical factor in LLM-based AES performance [19]. Instruction-based prompting explicitly directs models to apply scoring criteria; rubric-aligned prompting embeds detailed definitions of each dimension (e.g., Lead, Counterclaim); and Chain-of-Thought (CoT) prompting elicits step-by-step reasoning before scoring. Recent work demonstrates that structured, rubric-integrated prompts significantly improve scoring accuracy and interpretability [22]. However, most evaluations lack systematic comparison across multiple open models and diverse prompt formulations.

Notably, [23] found significant misalignment between LLMs and human graders in analytic rubric-based scoring, attributing this gap to LLMs' limited logical reasoning. They emphasize the need to better mirror human evaluators' judgment processes. Similarly, [22] proposed a hybrid evaluation framework combining automatic metrics and human assessment, showing that LLMs can approach expert-level performance when carefully prompted—but also highlighting risks of overfitting and poor generalization across prompts.

2.5. Research Gap

Despite these advances, three key gaps remain: (i) limited benchmarking of modern open-source LLMs for AES across consistent datasets and metrics; (ii) insufficient understanding of how prompt design choices—from minimal instructions to full rubric integration—affect scoring reliability and human alignment; and (iii) a lack of reproducible, rubric-grounded evaluation frameworks that combine quantitative metrics with qualitative human comparison.

This study directly addresses these gaps through a comparative evaluation of five open-weight LLMs on the PERSUADE 2.0 dataset under three distinct prompting strategies—rubric-aligned, instruction-based, and minimal instruction—using standard AES metrics and human-graded validation.

3. Our Approach for Automated Essay Scoring

The primary aim of this work is to assess the effectiveness of LLMs in replicating human scoring, particularly within the context of argumentative writing. Rather than using fine-tuning techniques.

Five state-of-the-art open-source LLMs **LLaMA 3.2 3B, DeepSeek-R1 7B, Mistral 8x7B, Qwen2 7B, and Qwen2.5 7B** were selected for their advanced natural language Each model was asked to give an overall essay score from 1 to 6, using grading rules similar to those used by human graders.

3.1. Dataset

This study employs the *PERSUADE 2.0 corpus*, a large-scale dataset collaboratively developed by The Learning Agency and Vanderbilt University [6]. Released under the CC BY-NC-SA 4.0 license, PERSUADE 2.0 substantially extends the earlier PERSUADE 1.0 version by including both holistic essay scores and fine-grained discourse annotations for more than 25,000 argumentative essays written by students in grades 6–12 across the United States. The dataset spans 15 writing prompts covering both independent and source-based argumentative tasks.

 doi:10.20944/preprints202511.1429.v1

Each essay is annotated with a holistic score on a 1–6 scale and discourse-level labels derived from a rubric grounded in Toulmin's argumentative framework [24] and refined using prior work by Nussbaum et al. [25] and Stapleton and Wu [26]. Annotated elements include:

- *Lead* (introductory segment),
- *Position* (explicit stance),
- *Claim* (supporting argument),
- *Counterclaim* (opposing viewpoint),
- *Rebuttal* (response to counterclaim),
- *Evidence* (facts/examples supporting claims),
- *Concluding Summary* (restatement of key arguments)

Annotation followed a double-blind rating process with full adjudication, where two expert raters independently evaluated each essay, and disagreements were resolved by a third rater to ensure high reliability.

Additionally, the dataset provides comprehensive demographic metadata, including grade level, gender, race/ethnicity, English language learner (ELL) status, disability status, socioeconomic indicators, and prompt/task identifiers, enabling detailed analyses of fairness and bias in automated essay scoring (AES) systems.

For this study, the main file utilized was:

- `persuade_2.0_human_scores_demo_id_github.csv`, which contains full essay texts, holistic scores, demographic profiles, writing task types, source text information, and prompt metadata.

Overall, *PERSUADE 2.0* offers a rich, multi-layered benchmark for evaluating Large Language Models (LLMs) in AES, supporting investigations into both scoring accuracy and equity across diverse demographic subgroups [6].

Table 1. Sample Snippets from the PERSUADE 2.0 Corpus.

Essay ID	Prompt Text	Essay Excerpt	Discourse Elements	Score
423A1CA112E2	Do people use cell phones daily?	Modern humans today are always on their phone...	Lead, Claim	3
BC75783F96E3	Mandatory recycling in cities?	Mandatory recycling programs are essential...	Lead, Claim	4
74C8BC7417DE	Limit teen social media use?	Social media should be limited ... it causes mental health problems.	Claim, Evidence	2
A8445CABFECE	Banning homework—beneficial?	Banning homework would harm students' ability to develop...	Counterclaim, Evidence	3
6B4F7A0165B9	Teach coding in elementary schools?	Introducing coding classes ... will prepare students for future careers.	Claim, Evidence	4

Table 1 presents a representative snippet of the dataset, illustrating the structure and content of the essays along with their corresponding scores. This excerpt is provided to give readers a clear understanding of the dataset's format, features, and the type of textual input used in the experiments.

 doi:10.20944/preprints202511.1429.v1

3.2. Model Selection

We evaluate five open-source LLMs spanning diverse architectures and design objectives; their key characteristics are summarized in Table 2. Models were selected to represent diverse architectures and parameter scales while remaining lightweight enough for resource-constrained educational deployment. Their open-source nature ensures transparency and reproducibility in AES research.

Table 2. Overview of evaluated open-source LLMs. Context window sizes reflect standard configurations reported by model providers.

Provider	Model Name	Size	Description	Reference
Meta	LLaMA 3.2 3B	3B	Instruction-tuned, efficient generative transformer.	[27]
DeepSeek	DeepSeek-R1 7B	7B	Optimized for classification and reasoning tasks.	[28]
Mistral AI	Mistral 8×7B	7B	High-accuracy model for nuanced text understanding.	[29]
Alibaba	Qwen2 7B	7B	Streamlined for low-latency NLP applications.	[30]
Alibaba	Qwen2.5 7B	7B	Enhanced for numerical accuracy and rubric alignment.	[31]

3.3. Prompt Engineering

To systematically investigate the impact of prompt design on automated essay scoring, we implement three distinct prompting strategies that vary in the level of guidance and structural scaffolding provided to the model. These range from minimal directives that merely name scoring dimensions to fully explicit rubric-embedded instructions. Table 3 presents representative examples of each prompt type used in our experiments, along with their corresponding classifications. Notably, all prompts enforce a constrained output format—requiring only a numerical score—to ensure consistency and comparability across conditions, while differing in how (or whether) they articulate the underlying assessment criteria or reasoning process. This design allows us to isolate the effect of prompt engineering on scoring accuracy and alignment with human judgments.

Table 3. Prompt strategies evaluated in this study. All prompts constrain output to a single numerical score; only the rubric-aligned variant embeds detailed scoring criteria.

Prompt #	Type	Prompt Example (condensed)	Key Features
1	Rubric-aligned	"As an expert essay grader, assess the student's response using the following criteria: **Lead**: Does the essay introduce the topic effectively? **Position**: Is the stance clear? ...Provide a score from 1 to 6. Output only the numerical score."	Explicit definitions of each scoring dimension
2	Instruction-based	"You are an experienced English teacher tasked with grading a student's argumentative essay. Assign a total score from 1 to 6 based on the following criteria: Lead, Position, Counterclaim, Rebuttal, Evidence, and Concluding Summary. Please provide only the final numerical score."	Role assignment + named criteria (no definitions)
3	Instruction-based (minimal)	"Grade this student's argumentative essay by evaluating lead, position, Counterclaim, Rebuttal, Evidence, and Concluding Summary. Assign a final grade from 1 to 6, reflecting how well the student constructs and supports their argument. Provide only the numerical score."	No role; criteria named only; most concise

 doi:10.20944/preprints202511.1429.v1

3.4. Evaluation Metrics

To rigorously evaluate the performance of each Large Language Model (LLM) in predicting essay scores, a comprehensive set of evaluation metrics was employed. These metrics were selected to assess not only the accuracy of the models but also their alignment with human grading and their reliability across diverse student responses. Together, they offer a robust and multidimensional framework for assessing LLM effectiveness in Automated Essay Scoring (AES).

3.4.1. Exact Match (EM)

Exact Match (EM) serves as a straightforward baseline metric that quantifies the percentage of model-generated scores that exactly match those assigned by human raters [32].

$$EM = \frac{\text{Number of exact matches}}{\text{Total number of essays}}$$

3.4.2. F1 Score

The F1 Score, calculated as the harmonic mean of precision and recall, offers a balanced evaluation of the model's performance [33].

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.4.3. Mean Absolute Error (MAE)

MAE evaluates the average magnitude of deviation between predicted and actual human scores [34].

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

3.4.4. Root Mean Squared Error (RMSE)

RMSE, like MAE, measures the deviation between predicted and actual scores but places a greater emphasis on larger errors [35].

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$$

3.4.5. Average Absolute Deviation (AAD)

AAD measures the average dispersion between the predicted and human-assigned scores [36].

$$AAD = \frac{1}{N} \sum_{i=1}^{N} \left| (y_i - \bar{y}) - (\hat{y}_i - \bar{\hat{y}}) \right|$$

3.4.6. Pearson and Spearman Correlation

The Pearson correlation assesses the linear relationship between model predictions and human scores, while the Spearman correlation evaluates rank-order consistency [37].

$$r_{pearson} = \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{\hat{y}})^2}}$$

$$\rho_{spearman} = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)}$$

 doi:10.20944/preprints202511.1429.v1

3.4.7. Cohen's Kappa

Cohen's Kappa is a statistical measure of inter-rater agreement that adjusts for chance-level agreement [38].

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed agreement and p_e is the expected agreement by chance.

In summary, this study employed a comprehensive suite of evaluation metrics to assess the performance of Large Language Models (LLMs) in Automated Essay Scoring. Metrics such as Exact Match, F1 Score, MAE, RMSE, and Average Absolute Deviation provided insights into the models' scoring accuracy and error magnitude. Pearson and Spearman correlations assessed the alignment of predictions with human scoring trends, while Cohen's Kappa offered a measure of inter-rater agreement beyond chance. Collectively, these metrics enabled a rigorous, multi-dimensional evaluation of model reliability, consistency, and alignment with human judgment - critical qualities for the effective integration of LLMs in educational assessment systems.

4. Experimental Results

This chapter presents the experimental setup and evaluation results for assessing the performance of various large language models (LLMs) in the task of Automated Essay Scoring (AES). The evaluation focuses on five models described in Section 3.2 using various prompting strategies described below to assess their effectiveness against human grader scores.

4.1. Prompt 1 Results

Table 4 presents the performance of all five open-source LLMs under Prompt 1 (rubric-aligned prompting), which explicitly embeds descriptive scoring criteria for each dimension of argumentative writing. The results reveal substantial variation across models, with DeepSeek-R1 7B and Mistral 8×7B emerging as the strongest performers. DeepSeek-R1 achieves the highest F1 Score (0.93), lowest MAE (0.593), and highest precision and recall (0.93 each), demonstrating balanced classification accuracy across score levels. Meanwhile, LLaMA 3.2 3B attains the strongest correlation with human graders, yielding the highest Pearson (0.863) and Spearman (0.831) coefficients—despite its smaller size—suggesting that rubric alignment enables even lightweight models to capture holistic scoring patterns. In contrast, Qwen2 7B exhibits notably weaker performance across all metrics (e.g., Pearson = 0.371, MAE = 1.000), indicating sensitivity to prompt formulation or domain alignment. Interestingly, Cohen's k remains low overall (max = 0.16 for Mistral), reflecting the well-documented challenge of achieving strong inter-rater agreement in ordinal AES settings—even with explicit rubrics. Nonetheless, the strong correlations and F1 scores underscore that rubric-aligned prompting significantly enhances the reliability of open-source LLMs for automated scoring, particularly when paired with models like DeepSeek-R1 and Mistral.

Table 4. Performance metrics of all five models using Prompt 1 (rubric-aligned prompting).

Metric	LLaMA 3.2 3B	DeepSeek-R1 7B	Mistral 8×7B	Qwen2 7B	Qwen2.5 7B
Exact Match (EM)	0.60	0.70	0.40	0.35	0.55
F1 Score	0.92	**0.93**	0.91	0.92	0.89
MAE ↓	1.000	**0.593**	0.658	1.000	0.538
RMSE ↓	1.140	**0.629**	0.835	1.304	0.734
Pearson r ↑	**0.863**	0.834	0.657	0.371	0.672
Spearman ρ ↑	**0.831**	0.738	0.634	0.392	0.512
Cohen's κ ↑	-0.06	0.06	**0.16**	0.07	0.14
Precision	0.85	**0.93**	0.89	0.85	1.00
Recall	1.00	0.93	**0.94**	1.00	0.80

 doi:10.20944/preprints202511.1429.v1

Table 5 presents a qualitative comparison of model predictions against human-assigned scores for five randomly selected argumentative essays from the PERSUADE 2.0 dataset. All models were prompted using the rubric-aligned strategy (Prompt 1), which embeds explicit scoring criteria and requests a single numerical score. Each row shows the human holistic score (on a 1–6 scale) alongside the predictions generated by the five evaluated LLMs. This snapshot illustrates both alignment and divergence in model behavior: for instance, DeepSeek-R1 7B matches the human score in three of the five cases and never deviates by more than 2 points, while Mistral 8×7B shows greater variability (e.g., predicting a score of 3 for an essay rated 5 by humans). Such examples highlight how prompt design and model choice jointly influence scoring fidelity at the instance level.

Table 5. Human scores versus model predictions for five sample essays using rubric-aligned prompting (Prompt 1).

Essay Snippet	Human	LLaMA	DeepSeek	Mistral	Qwen2	Qwen2.5
New software has been created...	5	6	5	3	5	6
When you need advice?...	4	4	3	4	5	4
Cars—no driver...	1	2	1	1	2	1
Some may say they need cars...	3	3	5	4	3	4
Does the Electoral College work?...	3	3	3	4	4	3

4.2. Prompt 2 Results

Table 6 reports model performance under Prompt 2 (instruction-based prompting), which provides a clear grading directive and lists the six scoring dimensions but omits their detailed definitions. Compared to the rubric-aligned condition (Prompt 1), all models exhibit a noticeable drop in performance, confirming that mere mention of criteria—without explicit elaboration—is insufficient to guide accurate scoring. Nevertheless, consistent trends emerge: Qwen2.5 7B achieves the highest correlation with human judgments (Pearson = 0.526, Spearman = 0.542) and the best inter-rater agreement (Cohen's k = 0.38), while DeepSeek-R1 7B leads in classification-oriented metrics such as F1 Score (0.55) and Exact Match (0.57). The generally low k values (0.31–0.38) indicate only fair agreement with human graders—a known limitation in AES when models lack fine-grained guidance. Notably, even the best-performing models under Prompt 2 fall short of the correlations observed with Prompt 1 (e.g., Pearson > 0.83), reinforcing the central finding that embedding explicit rubric language is far more effective than naming criteria alone. This gap underscores the sensitivity of open-source LLMs to prompt specificity, especially in complex, multi-dimensional assessment tasks.

Table 6. Performance metrics of all five models using Prompt 2 (instruction-based prompting).

Metric	LLaMA 3.2 3B	DeepSeek-R1 7B	Mistral 8×7B	Qwen2 7B	Qwen2.5 7B
Exact Match (EM)	0.52	0.57	0.45	0.47	0.39
F1 Score	0.48	0.55	0.52	0.53	0.56
MAE ↓	0.752	0.689	0.701	0.691	0.668
RMSE ↓	0.904	0.802	0.835	0.824	0.793
Pearson r ↑	0.442	0.503	0.487	0.491	**0.526**
Spearman ρ ↑	0.431	0.536	0.464	0.502	**0.542**
Cohen's κ ↑	0.31	0.37	0.34	0.35	**0.38**
Precision	0.46	0.53	0.50	0.51	0.54
Recall	0.44	0.50	0.48	0.49	0.51

Table 7 illustrates model behavior under Prompt 2 (instruction-based prompting) through a side-by-side comparison of human scores and predictions on five representative essays from PERSUADE 2.0. Unlike the rubric-aligned condition (Prompt 1), which embeds explicit scoring definitions, Prompt 2 only names the assessment criteria—resulting in less consistent and often less accurate predictions. For instance, on the high-scoring essay ("New software has been created...", human score = 5), LLaMA

 doi:10.20944/preprints202511.1429.v1

3.2 and Mistral 8×7B under-predict (3 and 2, respectively), while Qwen2.5 over-predicts (6). Similarly, for the lowest-scoring essay ("Cars—no driver..", human score = 1), Qwen2.5 assigns a notably inflated score of 3, suggesting a failure to recognize poorly developed arguments without explicit rubric guidance. DeepSeek-R1 7B demonstrates relatively stable performance, matching the human score in two cases and deviating by at most 1 point in others—consistent with its strong quantitative results in Table 6. Overall, this qualitative snapshot reinforces a key trend: without detailed rubric descriptors, even capable open-source LLMs struggle to replicate human grading logic, particularly at score extremes. This contrasts sharply with Prompt 1, where the same models showed markedly better alignment—highlighting that prompt specificity, not just model capacity, governs AES reliability.

Table 7. Human scores versus model predictions for five sample essays using instruction-based prompting (Prompt 2).

Essay Snippet	Human	LLaMA	DeepSeek	Mistral	Qwen2	Qwen2.5
New software has been created...	5	3	5	2	3	6
When you need advice?...	4	4	3	4	5	3
Cars—no driver...	1	2	1	2	1	3
Some may say they need cars...	3	3	2	3	2	3
Does the Electoral College work?...	3	2	3	4	3	2

4.3. Prompt 3 Results

Table 8 presents model performance under Prompt 3 (minimal instruction-based prompting), which provides only a concise directive to grade the essay based on key argumentative dimensions—without assigning a role (e.g., "teacher") or defining the criteria. Surprisingly, this stripped-down approach yields modest but consistent improvements over Prompt 2 (the role-assigned instruction variant) across several metrics, particularly in inter-rater agreement and correlation. DeepSeek-R1 7B again emerges as the strongest performer, achieving the highest F1 Score (0.63), Cohen's k (0.49), and Precision (0.59), indicating robust classification capability even with minimal guidance. Qwen2.5 7B leads in correlation metrics (Pearson = 0.582, Spearman = 0.577) and achieves the lowest MAE (0.598), suggesting it is especially adept at capturing the ordinal structure of human scores when not constrained by potentially misleading role framing. Notably, all models show higher k values under Prompt 3 (0.43–0.49) than under Prompt 2 (0.31–0.38), implying that the addition of a "teacher" persona in Prompt 2 may have introduced unintended biases or inconsistencies—particularly for models less attuned to role-based reasoning. Although performance under Prompt 3 still lags significantly behind the rubric-aligned condition (Prompt 1), these results reveal a nuanced insight: for some open-source LLMs, simplicity in prompting can outperform overly prescriptive or role-heavy instructions, especially when the role is not grounded in explicit task definitions. This underscores that effective prompt engineering for AES is not merely about adding more information, but about providing the right kind of guidance.

Table 8. Performance metrics of all five models using Prompt 3 (minimal instruction-based prompting).

Metric	LLaMA 3.2 3B	DeepSeek-R1 7B	Mistral 8×7B	Qwen2 7B	Qwen2.5 7B
Exact Match (EM)	0.35	0.43	0.40	0.41	0.42
F1 Score	0.54	**0.63**	0.60	0.61	0.62
MAE ↓	0.681	0.603	0.628	0.612	**0.598**
RMSE ↓	0.852	0.726	0.755	0.743	**0.714**
Pearson r ↑	0.508	0.578	0.536	0.371	**0.582**
Spearman ρ ↑	0.483	0.562	0.536	0.542	**0.577**
Cohen's κ ↑	0.43	**0.49**	0.47	0.46	0.48
Precision	0.54	**0.59**	0.57	0.56	0.58
Recall	0.52	**0.56**	0.54	0.54	0.56

 doi:10.20944/preprints202511.1429.v1

Table 9. Human scores versus model predictions for five sample essays using minimal instruction-based prompting (Prompt 3).

Essay Snippet	Human	LLaMA	DeepSeek	Mistral	Qwen2	Qwen2.5
New software has been created...	5	4	5	4	5	4
When you need advice?...	4	2	3	4	5	4
Cars—no driver...	1	1	2	2	3	1
Some may say they need cars...	3	3	4	2	3	2
Does the Electoral College work?...	3	2	3	3	2	2

Table 9 shows model predictions versus human scores for five representative essays under Prompt 3 (minimal instruction-based prompting)—a bare-bones directive that names scoring dimensions but provides no role assignment or rubric definitions. Despite its simplicity, this prompt yields more calibrated predictions than Prompt 2 (the role-enhanced variant), particularly for mid- to high-scoring essays. For instance, on the top-rated essay ("New software has been created...", human = 5), three models (DeepSeek-R1, Mistral, Qwen2) predict 5 or 4, showing reasonable fidelity—unlike Prompt 2, where the same essay triggered more extreme errors (e.g., Mistral scoring it 2). Similarly, DeepSeek-R1 7B matches the human score in two cases and stays within ±1 point in all others, aligning with its strong quantitative performance in Table 8. However, challenges persist at the score extremes: Qwen2 7B over-predicts the lowest-quality essay ("Cars—no driver...", human = 1) as a 3, while LLaMA 3.2 under-predicts the solid mid-range essay ("When you need advice?...", human = 4) as a 2. Notably, no model consistently outperforms others across all items, suggesting that minimal prompting reduces—but does not eliminate—model idiosyncrasies. Crucially, even in this simplified setting, performance remains substantially below Prompt 1, reinforcing that explicit rubric integration provides irreplaceable scaffolding for reliable AES. These observations support a nuanced conclusion: while stripping away unnecessary role framing (as in Prompt 3) can improve stability over more verbose but under-specified instructions (Prompt 2), detailed criteria remain essential for human-aligned scoring.

4.4. The Primacy of Prompt Design in AES

Across all evaluated models, prompt design exerts a decisive influence on Automated Essay Scoring (AES) performance—often outweighing differences in model architecture or size. Rubric-aligned prompting (Prompt 1), which embeds explicit, human-readable definitions of each scoring dimension, consistently yields the strongest alignment with human judgments, achieving Pearson correlations above 0.83 and F1 scores near 0.93 for top models. In contrast, both instruction-based variants—whether enhanced with a grading role (Prompt 2) or stripped to a minimal directive (Prompt 3)—produce markedly lower agreement, with correlations generally below 0.60 and Cohen's k in the "fair" range (0.31–0.49). Interestingly, the minimal prompt (Prompt 3) often outperforms the role-assigned version (Prompt 2), suggesting that role priming without substantive criteria may introduce noise or misalignment, particularly for models not fine-tuned for pedagogical tasks. These findings collectively demonstrate that the most effective path to reliable, transparent, and equitable AES with open-source LLMs is not model scaling, but prompt precision: providing models with the same descriptive rubric used by human graders enables even modest-sized LLMs to emulate expert scoring behavior with surprising fidelity.

5. Conclusion and Future Work

This work demonstrates that prompt engineering—not model scale or proprietary access—is the critical lever for achieving reliable, human-aligned Automated Essay Scoring (AES) with open-source Large Language Models. Through a systematic evaluation of five contemporary open-weight LLMs—LLaMA 3.2 3B, DeepSeek-R1 7B, Mistral 8×7B, Qwen2 7B, and Qwen2.5 7B—on the PERSUADE 2.0 dataset, we show that embedding explicit rubric descriptors directly into the prompt (rubric-aligned

 doi:10.20944/preprints202511.1429.v1

prompting) dramatically improves scoring accuracy, correlation with human judgments, and inter-rater agreement, consistently outperforming both role-based and minimal instruction strategies. Notably, lightweight models like DeepSeek-R1 7B and Mistral 8×7B achieve Pearson correlations above 0.83 and F1 scores near 0.93 under this approach—performance levels that rival those reported for much larger, closed commercial systems.

Our findings carry significant implications. First, they challenge the assumption that AES requires massive, black-box models, offering a viable pathway toward transparent, auditable, and locally deployable assessment tools—a crucial step toward equitable educational technology. Second, they reveal that not all prompting is equally effective: simply naming scoring criteria or assigning a grading persona is insufficient; explicit, human-readable rubric integration is essential. Surprisingly, we also find that overly verbose or role-heavy instructions can underperform minimalist directives, underscoring that prompt design must prioritize semantic precision over surface-level scaffolding.

Several promising avenues emerge from this work. First, adaptive or dynamic prompting—where the prompt is tailored based on essay characteristics (e.g., length, topic, or predicted difficulty)—could further refine scoring fidelity. Second, enabling models to generate justifications alongside scores (e.g., via constrained Chain-of-Thought) would enhance interpretability and support formative feedback, moving beyond mere assessment to pedagogical utility. Third, evaluating prompt robustness across diverse student populations, grade levels, and writing genres—including non-argumentative tasks—is essential for real-world deployment. Finally, extending this framework to multilingual and low-resource educational contexts could democratize access to high-quality automated feedback, especially where human grading capacity is limited.

These results position prompt engineering as a low-cost, high-impact intervention for educators, developers, and researchers seeking to adopt open-source LLMs in real-world assessment settings. Nevertheless, this study establishes a clear principle: to make LLMs grade like humans, we must first teach them the rules humans use.

References

1. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology* **2025**, *16*, 1–72.
2. Stahl, M.; Biermann, L.; Nehring, A.; Wachsmuth, H. Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. In Proceedings of the Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024); Kochmar, E.; Bexte, M.; Burstein, J.; Horbach, A.; Laarmann-Quante, R.; Tack, A.; Yaneva, V.; Yuan, Z., Eds., Mexico City, Mexico, 2024; pp. 283–298.
3. Zhang, T.; Jiang, Z.; Zhang, H.; Lin, L.; Zhang, S. MathMistake Checker: A Comprehensive Demonstration for Step-by-Step Math Problem Mistake Finding by Prompt-Guided LLMs. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 29730–29732.
4. Pack, A.; Barrett, A.; Escalante, J. Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence* **2024**, *6*, 100234. https://doi.org/10.1016/j.caeai.2024.100234.
5. Bu, J.; Ren, L.; Zheng, S.; Yang, Y.; Wang, J.; Zhang, F.; Wu, W. ASAP: A Chinese review dataset towards aspect category sentiment analysis and rating prediction. *arXiv preprint arXiv:2103.06605* **2021**.
6. Crossley, S.; Tian, Y.; Baffour, P.; Franklin, A.; Benner, M.; Boser, U. A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Assessing Writing* **2024**, *61*, 100865. https://doi.org/https://doi.org/10.1016/j.asw.2024.100865.
7. Semire, D. Automated Essay Scoring. *The Turkish Online Journal of Distance Education* **2006**, *7*.
8. Leidner, D.E. Globalization, culture, and information: Towards global knowledge transparency. *The Journal of Strategic Information Systems* **2010**, *19*, 69–77. https://doi.org/https://doi.org/10.1016/j.jsis.2010.02.006.
9. Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; Mihalcea, R. Automatic Detection of Fake News. In Proceedings of the Proceedings of the 27th International Conference on Computational Linguistics; Bender, E.M.; Derczynski, L.; Isabelle, P., Eds., Santa Fe, New Mexico, USA, 2018; pp. 3391–3401.

 doi:10.20944/preprints202511.1429.v1

10. Kahng, M.; Tenney, I.; Pushkarna, M.; Liu, M.X.; Wexler, J.; Reif, E.; Kallarackal, K.; Chang, M.; Terry, M.; Dixon, L. LLM Comparator: Interactive Analysis of Side-by-Side Evaluation of Large Language Models. *IEEE Transactions on Visualization and Computer Graphics* **2025**, *31*, 503–513. https://doi.org/10.1109/TVCG.2024.3456354.
11. Yang, K.; Raković, M.; Li, Y.; Guan, Q.; Gašević, D.; Chen, G. Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability. In Proceedings of the Proceedings of the aaai conference on artificial intelligence, 2024, Vol. 38, pp. 22466–22474.
12. Su, J.; Yan, Y.; Fu, F.; Zhang, H.; Ye, J.; Liu, X.; Huo, J.; Zhou, H.; Hu, X. Essayjudge: A multi-granular benchmark for assessing automated essay scoring capabilities of multimodal large language models. *arXiv preprint arXiv:2502.11916* **2025**.
13. Staudemeyer, R. Understanding LSTM–a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586* **2019**.
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
15. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* **2019**, pp. 4171–4186.
16. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)* **2020**.
17. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* **2023**.
18. Yavuz, F. Utilizing Large Language Models for EFL Essay Grading: An Examination of Reliability and Validity in Rubric-Based Assessments. *British Journal of Educational Technology* **2024**.
19. Fiacco, J. Towards Extracting and Understanding the Implicit Rubrics in Transformer-Based Automated Essay Scoring. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)* **2023**.
20. Lundgren, M. Large Language Models in Student Assessment. *arXiv preprint arXiv:2406.16510* **2024**.
21. Lundgren, M. Large Language Models in Student Assessment. *arXiv preprint arXiv:2406.16510* **2024**.
22. Mansour, W.A.; et al. Can Large Language Models Automatically Score Essays? In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), 2024.
23. Wu, X.; Saraf, P.P.; Lee, G.G.; Latif, E.; Liu, N.; Zhai, X. Unveiling Scoring Processes: Dissecting the Differences between LLMs and Human Graders in Automatic Scoring. *arXiv* **2024**.
24. Toulmin, S.E. *The Uses of Argument*; Cambridge University Press: Cambridge, UK, 1958.
25. Nussbaum, E.M.; Kardash, C.M.; Graham, S. The Effects of Goal Instructions and Text on the Generation of Counterarguments During Writing. *Journal of Educational Psychology* **2005**, *97*, 157–169. https://doi.org/10.1037/0022-0663.97.2.157.
26. Stapleton, P.; Wu, Y.Y. Assessing the Quality of Arguments in Students' Persuasive Writing: A Case Study Analysing the Relationship Between Surface Structure and Argumentative Quality. *Journal of Second Language Writing* **2015**, *30*, 1–11. https://doi.org/10.1016/j.jslw.2015.06.004.
27. Touvron, H.; Martin, L.; Izacard, P.; et al.. LLaMA 3.2: Instruction-Tuned Multilingual Generative Transformer. Meta AI, 2024. Available at: https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct.
28. AI, D. DeepSeek-R1: Open-Source Reasoning Model. DeepSeek AI, 2024. Available at: https://huggingface.co/deepseek-ai/DeepSeek-R1.
29. Jiang, A.; et al.. Mixtral of Experts: High-Accuracy Sparse MoE Model for Text Understanding. *arXiv preprint* **2024**, *abs/2401.04088*.
30. Zhou, Y.; et al.. Qwen2: Optimized for Low-Latency NLP Applications. Alibaba Cloud, 2023. Available at: https://huggingface.co/Qwen/Qwen2-7B.
31. Zhou, Y.; et al.. Qwen2.5: Enhanced for Numerical Accuracy and Rubric Alignment. Alibaba Cloud, 2024. Available at: https://huggingface.co/Qwen/Qwen2.5-7B-Instruct.
32. Zhang, W.; Litman, D. Automated essay scoring: A survey of the state of the art. *International Journal of Artificial Intelligence in Education* **2022**.

 doi:10.20944/preprints202511.1429.v1

33. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* **2011**, *2*, 37–63.
34. Ramesh, D.; Sanampudi, S. An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review* **2022**, *55*, 249–289.
35. Sun, J. A survey of automated essay scoring. *Neurocomputing* **2025**.
36. Attali, Y.; Burstein, J. Validity and reliability of automated essay scoring systems. *Educational Testing Service Research Report* **2016**.
37. Ben-Simon, A.; Bennett, R. Correlation between automated and human essay scoring. *Applied Measurement in Education* **2007**, *20*, 358–381.
38. Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **1960**, *20*, 37–46.