# Evaluating LLM-Based Automated Essay Scoring: Accuracy, Fairness, and Validity

**Yue Huang**
Measurement Incorporated
yueh@udel.edu

**Joshua Wilson**
School of Education
University of Delaware
joshwils@udel.edu

## Abstract

This study evaluates large language models (LLMs) for automated essay scoring (AES) in comparison with a traditional feature-based system (PEG) and human ratings. Drawing on 541 essays from Grades 3–4, we examined three generative LLMs (including GPT-4o, Gemini 2.5 Flash and Claude Sonnet 4) under baseline, context-enhanced, and few-shot chain-of-thought prompting strategies. Results show that carefully designed prompting, particularly context-enhanced few-shot chain-of-thought, substantially improved LLM performance, approaching PEG in human–machine agreement and human–human agreement. Fairness analyses revealed that PEG produced larger disparities for English language learners (ELLs), while LLMs showed smaller but still persistent subgroup bias. Beyond these findings, the study contributes recent evidence on fairness and validity in LLM-based AES and extends research to younger students, a group rarely examined in prior work. Together, these results highlight both the promise and the challenges of integrating LLMs into educational assessment.

## 1 Introduction

*Automated essay scoring* (AES) refers to the use of computational methods and/or AI techniques to evaluate student-generated writing and assign scores in place of or alongside human raters (Uto and Okano, 2020). In the field of educational measurement, automatic scoring has become a cutting-edge approach to evaluating written content without manual grading. This strategy is especially valuable in large-scale assessments and classroom contexts where manual scoring is impractical (Latif and Zhai, 2024; Susanti et al., 2023). Early AES systems were built on *natural language processing* (NLP) and machine learning methods with handcrafted features and large labeled datasets (Uto, 2021). More recently, deep learning models such as recurrent and transformer-based architectures have improved scoring performance by leveraging text embeddings and contextual representations (El-Massry et al., 2025). Still, these approaches face limitations in interpretability, scalability, and fairness.

The rise of *large language models* (LLMs) offers new potential for AES. Pretrained on massive corpora, LLMs capture sophisticated semantic and discourse-level features, eliminating the need for manual feature engineering. With careful prompt design or fine-tuning, these models can be guided to produce not only holistic scores but also rubric-aligned scores. However, critical challenges remain: model outputs can vary depending on prompt engineering, and concerns about fairness, subgroup performance, and transparency persist (Huang et al., 2025).

This paper addresses these gaps by systematically evaluating several LLM-based AES approaches and comparing them with a more traditional feature-based AES system. We examined: (1) accuracy, measured by their alignment with human ratings; (2) fairness, focusing on differences between English language learners (ELLs) and non-ELLs. By analyzing prompting strategies across different LLMs and subgroup outcomes, we contribute empirical evidence to guide both AES research and educational assessment practice.

## 2 Related Work

### 2.1 AES in Educational Assessment

Research on automated essay scoring (AES) has a long history in educational assessment, beginning with the initial version of Project Essay Grade (PEG; Page, 1966) scoring engine. PEG relied on surface-level textual features, such as word counts, sentence length, and syntactic structures, combined with statistical modeling to predict scores. Subsequent systems, such as e-rater, extended this approach by incorporating more linguistically informed features grounded in NLP, including grammar, usage, mechanics, style, and organizational development (Hussein et al., 2019). These early systems demonstrated that linguistic features, when coupled with statistical models, could produce levels of agreement with human raters comparable to inter-rater reliability, establishing the feasibility of automated scoring for constructed responses, short answers, and essays.

With advances in artificial intelligence, AES shifted toward feature-rich regression and classification models trained on large corpora during the mid-2010s. The adoption of deep neural networks, particularly recurrent architectures (such as LSTM) and convolutional neural networks (CNN), enabled models to capture sequential dependencies in student writing (Dong et al., 2017; Taghipour and Ng, 2016). The emergence of transformer-based large language models (LLMs), including BERT and XLNet, further advanced AES by leveraging contextual embeddings that outperformed prior deep learning methods (Rodriguez et al., 2019; Uto, 2021). Building on this foundation, Yang et al. (2020) introduced R$^2$BERT, a BERT-based model that combined regression and ranking objectives, achieving strong performance on the widely used Automated Student Assessment Prize (ASAP) dataset (Hamner et al., 2012). Extending this line of work, Xie et al. (2022) proposed Neural Pairwise Contrastive Regression (NPCR), a contrastive learning approach that modeled score differences across essay pairs and achieved state-of-the-art results. These models reduced reliance on manual feature engineering and improved generalizability across essay tasks.

Despite notable advances, AES still faces important limitations. First, systems struggle to capture the full range of features that characterize high-quality writing, particularly when holistic scoring and rubric-based analytic scoring demand different forms of feature representation (Kumar and Boulanger, 2021). Elements such as rhetorical intent, coherence, and creativity remain especially difficult to model (Huang et al., 2025). Second, much empirical research relies heavily on benchmark datasets such as the ASAP corpus (focusing on essays from students in Grades 7–8), which facilitate standardized comparisons but offer limited insight into writing at earlier developmental stages where challenges are most acute. Finally, fairness continues to be a major concern. Feature-based AES systems frequently inherit biases present in the human ratings used for training, leading to systematically different outcomes for subgroups such as ELLs (Wilson and Huang, 2024). Ensuring equitable scoring across diverse student populations therefore remains a central challenge for AES in educational measurement.

## 2.2 Generative Large Language Models for AES

The recent development of generative LLMs such as GPT-4 and Llama-3 (referred to hereafter as GPT-family models for simplicity) has demonstrated remarkable capabilities in language understanding, reasoning, and text generation. Unlike earlier encoder-based LLMs, GPT-family models adopt decoder-based, autoregressive architectures (Minaee et al., 2025). This design enables them to generate coherent and contextually rich text, capture nuanced semantic relationships, and adapt flexibly to varied writing genres and proficiency levels, which holds particular promise for evaluating essays in ways that attend not only to surface features but also to deeper rhetorical and logical structures.

A growing body of research has examined the performance of generative LLMs for AES, though findings remain mixed. Results vary depending on prompting strategies, fine-tuning methods, and system adaptation (Huang et al., 2025). Proprietary models such as GPT-3.5 and GPT-4 show reasonable performance with few-shot prompting, especially when combined with rubric descriptions, explicit task instructions, and/or chain-of-thought (CoT) reasoning (Mansour et al., 2024; Quah et al., 2024; Wei et al., 2022). However, they often underperform compared to fine-tuned models and raise concerns about transparency and replicability. Designing prompts that ensure reliability remain an open challenge.

Open-source models such as Llama-3 introduce new opportunities. Research by Ormerod and Kwako (2024) demonstrated that smaller open-source models, when fine-tuned, can achieve performance comparable to traditional best-performing models while running on modest hardware. This approach enhances transparency and allows researchers to integrate explainable AI methods, addressing some of the limitations of closed GPT-family models. Yet, fine-tuning requires technical expertise, and performance still lags behind state-of-the-art models on benchmark datasets.

As with earlier LLMs, the adoption of GPT-family models raises broader concerns about fairness and validity (Huang et al., 2025). Few studies have examined subgroup differences, and those that exist focus mainly on multilingual learners (e.g., Tate et al., 2024). Ethical concerns are also mounting, particularly around data privacy, consent, and intellectual property.

Overall, research on LLM-based AES is still emerging. Current evidence suggests that while LLMs can approximate human scoring with careful prompt design or fine-tuning, their performance remains inconsistent across contexts, and fairness outcomes are underexplored. The next phase of research must therefore integrate technical advances with principles of educational measurement to ensure that LLM-based automated scoring is both effective and equitable. The present study contributes to this effort by examining LLM scoring across student subgroups, specifically ELLs versus non-ELLs.

## 3 Research Questions

This study evaluates three large language models (LLMs) alongside a traditional feature-based AES system (PEG) to examine their alignment with human ratings and their fairness for ELLs compared to non-ELLs, under three different prompt engineering strategies.

RQ1: How do prompt engineering strategies affect human–machine agreement across LLMs, PEG, and human raters?

RQ2: Do LLMs exhibit performance differences or subgroup bias between ELL and non-ELL students?

## 4 Methods

### 4.1 Sample

This study draws on data from an evaluation of an automated writing evaluation system in Grades 3–5 in a U.S. school district in school year 2017–2018. The district implemented the system in conjunction with a Common Core–aligned English language arts curriculum to support writing instruction for all students. A subsample of 541 de-identified essays from third and fourth graders ($N$ = 233 and 308, respectively) written between April 1 to May 31, 2018, was analyzed; each grade responded to a separate grade-level informative essay task. For the writing tasks, Grade 3 students read two short texts about national parks—one emphasizing their value for recreation, wildlife protection, and science, and the other highlighting challenges such as pollution and overcrowding—and were asked to write an informative essay explaining what national parks are and why they matter. Similarly, Grade 4 students read texts introducing invertebrates and describing the features, habitats, and life cycle of crabs, and were asked to write an informative essay about the key characteristics of crabs and how they live. Essays were scored by six approaches (see details below). Ten percent of the essays ($N$ = 57) were randomly double scored by a second human rater. ELLs comprised 32% of third graders and 46% of fourth graders.

### 4.2 Measures

Six scoring approaches were evaluated: (1) human rater 1, (2) human rater 2 (10% of the sample), (3) PEG, (4) LLMs with baseline CoT prompting, (5) LLMs with context-enhanced CoT prompting, and (6) LLMs with context-enhanced + few-shot CoT prompting. Three LLMs were considered: GPT-4o, Gemini 2.5 Flash and Claude Sonnet 4.

Human raters were professional scorers employed by the company operating the automated writing evaluation system. They received extensive training and were continuously monitored through rater management systems designed to ensure scoring accuracy and consistency. Human raters applied a six-trait rubric assessing development of ideas, organization, style, sentence fluency, word choice, and conventions. Each trait was scored on a 1–5 scale, and a holistic score was obtained by summing the six traits (range = 6–30).

The most recent PEG scoring engine has advanced substantially beyond its earlier, simpler versions. Current PEG scores are produced using a proprietary model that integrates more than 800 linguistic features with deep learning algorithms, trained on a large corpus of historical student essays from the same grade band and curriculum-aligned tasks.

Figure 1 presents the flowchart for the three prompting strategies. In the baseline CoT condition, prompts included the scoring task instructions, essay task description, rubric details, and a CoT component guiding the model to reason step by step about how to apply the rubric. The system was then asked to generate a score and provide the scoring output as specified. For the context-enhanced CoT strategy, one additional component was introduced: the model was assigned the role of an experienced essay rater familiar with the writing proficiency levels of third- and fourth-grade stu-

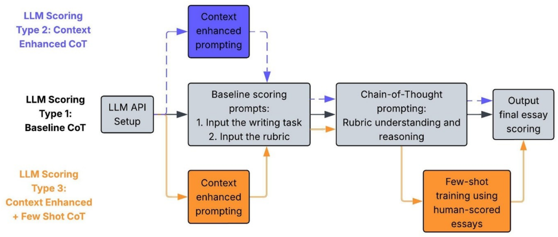

LLM Scoring Type 2: Context Enhanced CoT
Context enhanced prompting
LLM Scoring Type 1: Baseline CoT
LLM API Setup
Baseline scoring prompts: 1. Input the writing task 2. Input the rubric
Chain-of-Thought prompting: Rubric understanding and reasoning
Output final essay scoring
LLM Scoring Type 3: Context Enhanced + Few Shot CoT
Context enhanced prompting
Few-shot training using human-scored essays

Figure 1: Flow Chart for the Design of Prompt Engineering Strategies Using LLMs

dents. For the context-enhanced + few-shot CoT strategy, another component was added: the model was provided with five sample essays along with their corresponding trait-level and holistic scores to guide its scoring decisions. All prompts were executed iteratively for each individual essay through API interactions with the models using Python.

### 4.3 Data Analysis

For RQ1, we evaluated human–machine agreement by calculating *Quadratic Weighted Kappa* (QWK) and *exact/adjacent agreement rates* between each scoring method (**S**: PEG and all LLMs across prompting strategies, as well as human rater 2) and human scores (**H**, from rater 1). QWK, a widely used reliability index, ranges from 0 to 1, with higher values indicating stronger alignment between two sets of ratings. Exact agreement reflects the proportion of cases where the AES score matches the human score exactly, while adjacent agreement reflects cases where the AES score is within ±1 point of the human score. Together, these measures provide complementary perspectives on model accuracy relative to human raters.

For RQ2, QWKs were calculated separately for ELL and non-ELL students to assess subgroup-specific performance. To further evaluate fairness, we applied Litman et al.'s (2021) metrics:

- *Overall Score Accuracy* (OSA): Measures whether AES scores are equally accurate across groups by regressing squared error $(S - H)^2$ on student group. A significant positive coefficient indicates systematic differences in accuracy between groups.

- *Overall Score Difference* (OSD): Assesses whether AES scores are consistently higher or lower than human scores for different groups using the absolute difference $|S - H|$ as the outcome. Significant differences suggest systematic over- or under-prediction for a subgroup.

- *Conditional Score Difference* (CSD): Extends OSD by controlling for student proficiency (approximated by human scores). Two models are compared—with and without student group. A significant increase in $R^2$ indicates that group membership affects AES accuracy beyond proficiency, signaling potential subgroup bias.

Based on RQ1 findings (see Results section), only LLM scores generated with context-enhanced + few-shot CoT prompting—the highest-performing strategy overall—were subjected to fairness analyses (OSA, OSD, CSD).

## 5 Results

### 5.1 RQ1–Comparisons across Scoring Approaches

Figure 2 shows QWK values for total scores. Human–human agreement was the highest benchmark (QWK = .91), with PEG next in line (QWK = .76). The LLMs, while trailing PEG, demonstrated a clear upward trend across prompting strategies: GPT-4o improved from .46 under baseline CoT to .72 with context-enhanced + few-shot prompting, Gemini 2.5 Flash rose from .43 to .60, and Claude Sonnet 4 from .30 to .69. These results indicate that structured prompts, especially those combining context and few-shot examples, substantially strengthen the alignment of LLM-generated scores with human ratings.

Trait-level analyses (Figure 3) reveal similar patterns. PEG maintained strong agreement across all traits (QWK = .61–.74), consistently falling between human–human agreement (.77–.86) and LLM performance. Among the LLMs, GPT-4o

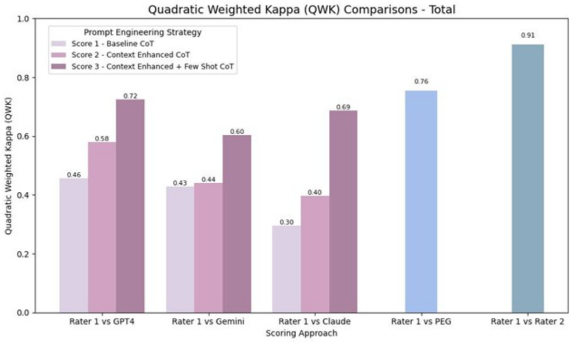


Figure 2: Comparisons of Quadratic Weighted Kappa (QWK) between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs – Total Score

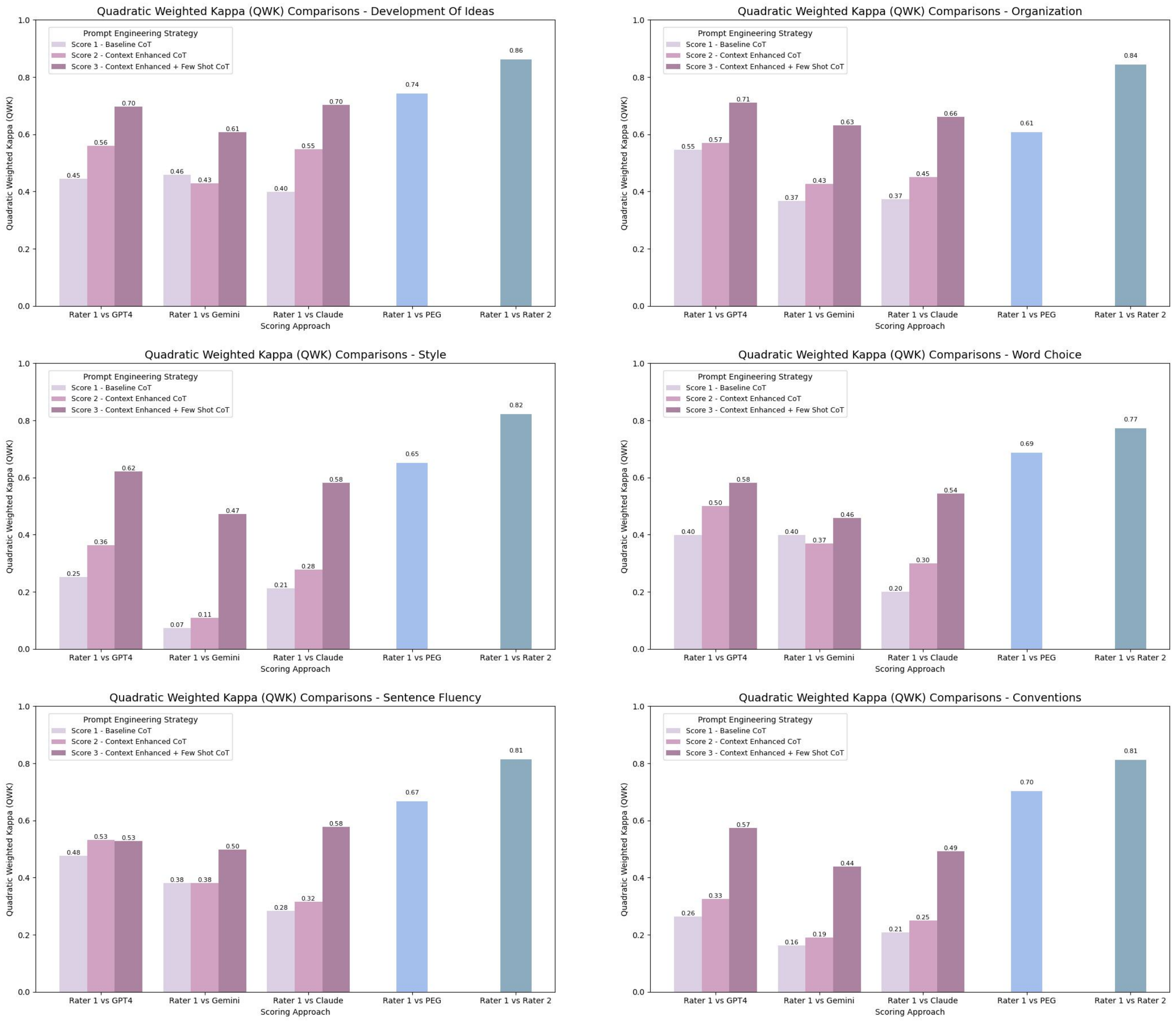


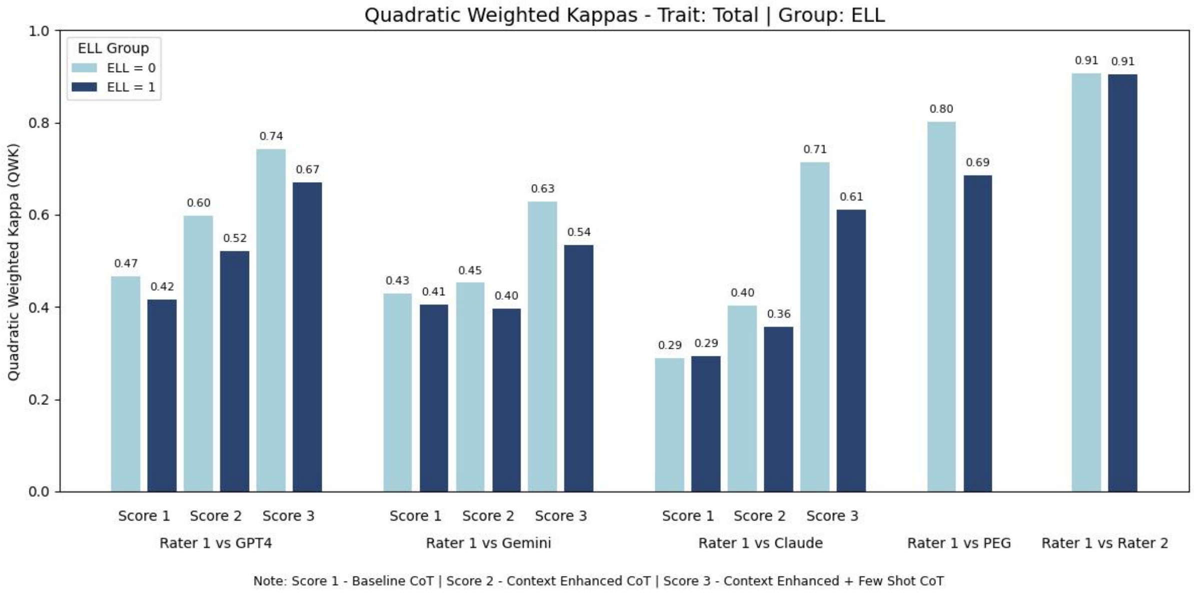
Figure 3: Comparisons of Quadratic Weighted Kappa (QWK) between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs – by Trait

again showed the highest alignment, particularly for development of ideas (.70) and organization (.71), while Claude Sonnet 4 performed competitively for development of ideas (.70). Gemini 2.5 Flash generally lagged behind, though its agreement improved under structured prompting.

Patterns in exact and adjacent agreement (see Appendix A) further support these findings. Exact agreement was highest for human–human (.28 for total score) and PEG (.21), with LLMs showing smaller but improving proportions as prompting strategies became more structured (e.g., GPT-4o rising from .14 to .20). Adjacent agreement was consistently stronger for total scores and trait scores. For example, human–human reached .63 in total score, PEG achieved .44, and LLMs again improved with prompting, with GPT-4o and Claude Sonnet 4 approaching PEG's level for traits including development of ideas, organization, style and word choice. Overall, these results suggest that while PEG remains the most reliable automated scorer, LLMs (particularly GPT-4o) can achieve meaningful gains through contextually enriched, few-shot prompting, with the largest improvements seen on traits tied to style and conventions.

### 5.2 RQ2–Fairness across ELL Group

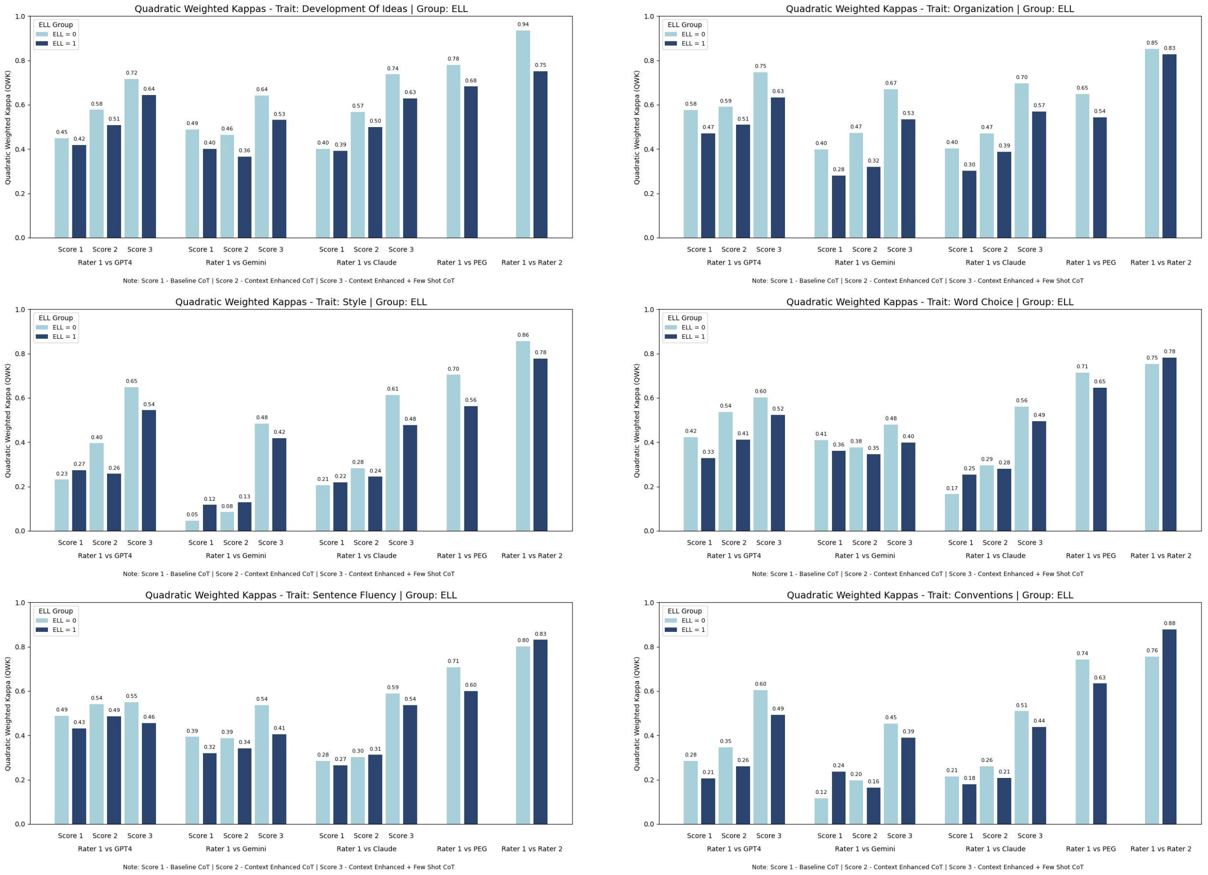
Figures 4 and 5 show QWK comparisons by ELL status. Across nearly all models and traits, agreement between AES scores and human ratings was higher for non-ELLs than for ELLs, indicating modest subgroup disparities. For total scores, GPT–human agreement reached .74 for non-ELLs versus .67 for ELLs under context-enhanced + few-shot prompting, Claude–human agreement achieved .71 versus .61, and Gemini–human agreement .63 versus .54, while PEG-human agreement



Quadratic Weighted Kappas - Trait: Total | Group: ELL

Note: Score 1 - Baseline CoT | Score 2 - Context Enhanced CoT | Score 3 - Context Enhanced + Few Shot CoT

Figure 4: Comparisons of Quadratic Weighted Kappa (QWK) between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs by ELL Status – Total Score



Quadratic Weighted Kappas - Trait: Development Of Ideas | Group: ELL

Note: Score 1 - Baseline CoT | Score 2 - Context Enhanced CoT | Score 3 - Context Enhanced + Few Shot CoT

Quadratic Weighted Kappas - Trait: Organization | Group: ELL

Note: Score 1 - Baseline CoT | Score 2 - Context Enhanced CoT | Score 3 - Context Enhanced + Few Shot CoT

Quadratic Weighted Kappas - Trait: Style | Group: ELL

Note: Score 1 - Baseline CoT | Score 2 - Context Enhanced CoT | Score 3 - Context Enhanced + Few Shot CoT

Quadratic Weighted Kappas - Trait: Word Choice | Group: ELL

Note: Score 1 - Baseline CoT | Score 2 - Context Enhanced CoT | Score 3 - Context Enhanced + Few Shot CoT

Quadratic Weighted Kappas - Trait: Sentence Fluency | Group: ELL

Note: Score 1 - Baseline CoT | Score 2 - Context Enhanced CoT | Score 3 - Context Enhanced + Few Shot CoT

Quadratic Weighted Kappas - Trait: Conventions | Group: ELL

Note: Score 1 - Baseline CoT | Score 2 - Context Enhanced CoT | Score 3 - Context Enhanced + Few Shot CoT

Figure 5: Comparisons of Quadratic Weighted Kappa (QWK) between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs by ELL Status – by Trait

| Metric | Human Rater 2 | PEG | GPT-4o (Context+Few-Shot CoT) | Gemini (Context+Few-Shot CoT) | Claude (Context+Few-Shot CoT) |
|---|---|---|---|---|---|
| *Total Score* | | | | | |
| OSA – ELL | 1.047 | **6.293**$^*$ | -0.077 | 0.658 | 1.502 |
| OSD – ELL | -0.014 | **0.693**$^*$ | -0.211 | -0.240 | -0.354 |
| CSD – $\Delta R^2$ | 0.001 | 0.006 | **0.010**$^*$ | **0.009**$^*$ | **0.016**$^*$ |
| *Trait 1 – Development of Ideas* | | | | | |
| OSA – ELL | **0.260**$^*$ | 0.146 | -0.008 | 0.072 | 0.080 |
| OSD – ELL | **0.262**$^*$ | **0.133**$^*$ | -0.098 | -0.073 | -0.039 |
| CSD – $\Delta R^2$ | 0.064 | 0.004 | **0.012**$^*$ | 0.006 | 0.006 |
| *Trait 2 – Organization* | | | | | |
| OSA – ELL | -0.003 | **0.198**$^*$ | 0.061 | 0.043 | 0.040 |
| OSD – ELL | -0.146 | 0.105 | 0.004 | -0.079 | **-0.120**$^*$ |
| CSD – $\Delta R^2$ | 0.033 | 0.001 | 0.002 | **0.008**$^*$ | **0.018**$^*$ |
| *Trait 3 – Style* | | | | | |
| OSA – ELL | 0.128 | **0.214**$^*$ | -0.026 | -0.115 | -0.031 |
| OSD – ELL | -0.043 | **0.124**$^*$ | -0.001 | -0.019 | -0.113 |
| CSD – $\Delta R^2$ | 0.003 | 0.002 | **0.006**$^*$ | **0.008**$^*$ | **0.026**$^*$ |
| *Trait 4 – Word Choice* | | | | | |
| OSA – ELL | 0.040 | 0.137 | 0.028 | 0.056 | 0.080 |
| OSD – ELL | 0.073 | **0.124**$^*$ | -0.005 | -0.027 | -0.024 |
| CSD – $\Delta R^2$ | 0.000 | 0.002 | **0.007**$^*$ | **0.008**$^*$ | **0.009**$^*$ |
| *Trait 5 – Sentence Fluency* | | | | | |
| OSA – ELL | -0.077 | **0.200**$^*$ | -0.004 | 0.097 | 0.014 |
| OSD – ELL | -0.072 | **0.117**$^*$ | -0.054 | -0.034 | -0.023 |
| CSD – $\Delta R^2$ | 0.008 | 0.001 | **0.016**$^*$ | **0.009**$^*$ | **0.009**$^*$ |
| *Trait 6 – Conventions* | | | | | |
| OSA – ELL | -0.208 | **0.165**$^*$ | 0.030 | -0.054 | -0.013 |
| OSD – ELL | -0.088 | 0.090 | -0.057 | -0.007 | -0.036 |
| CSD – $\Delta R^2$ | 0.019 | 0.000 | **0.012**$^*$ | **0.006**$^*$ | **0.009**$^*$ |

Table 1: Fairness Evaluation Results by ELL Status. Values are coefficients for OSA and OSD (ELL effect) and $\Delta R^2$ for CSD. Significant values are in **bold** and marked with $^*$ ($p < .05$).

also favored non-ELLs (.80 vs. .69). Gaps between ELL and non-ELL for human–human agreement also varied somewhat across traits, with smaller subgroup differences for organization, word choice, and sentence fluency. Notably, prompting did not eliminate subgroup gaps, and both LLMs and PEG continued to score ELLs less consistently than non-ELLs. These comparisons with human–human agreements should be interpreted cautiously, however, given the limited size of the double-scored sample (10%) by a second human rater.

Table 1 presents results from the fairness evaluation. Based on RQ1 findings, only LLM scores generated with context-enhanced + few-shot CoT prompting—the most accurate overall—were examined further. Specifically, PEG showed significant ELL-based differences in both OSA and OSD for most traits, but these differences were not significant regarding CSD. In contrast, the three LLMs with context-enhanced + few-shot prompting displayed few significant results for OSA and OSD, implying more balanced performance across groups at the overall level. Yet, CSD revealed persistent disparities: GPT-4o showed effects for total score, development of ideas, style, word choice, sentence fluency, and conventions; Gemini 2.5 Flash for total score, organization, style, word choice, and conventions; and Claude Sonnet 4 for nearly all traits except development of ideas. These findings suggest that while LLMs reduced overt subgroup bias relative to PEG, subtler inequities remained once proficiency was considered.

## 6 Conclusions and Implications

This study provides early empirical evidence that large language models (LLMs), specifically generative LLMs such as GPT-family models, when combined with carefully designed prompting strategies, can approach the performance of feature-based AES systems such as PEG. This study compared not only multiple LLMs but also different prompting strategies, offering valuable insights and practical guidance for future research on prompt design. Context-enhanced + few-shot chain-of-thought prompting consistently outperformed baseline approaches, highlighting the central role of prompt engineering in optimizing LLM-based scoring for both accuracy and consistency.

At the same time, fairness analyses revealed that neither PEG nor LLMs fully eliminated subgroup disparities. PEG exhibited larger discrepancies for ELLs in overall accuracy and error magnitude, whereas LLMs appeared more balanced at the surface level. However, conditional score difference analyses showed that subtle, proficiency-adjusted disparities persisted across traits, suggesting that fairness concerns remain in LLMs. Importantly, this study examined both holistic scores and rubric-based analytical scores, contributing evidence on how LLMs perform across different scoring dimensions. Furthermore, it provides some of the most up-to-date findings on subgroup fairness in LLM-based scoring, adding important validity evidence to ongoing debates about their educational use. These findings underscore the importance of evaluating LLMs with multiple fairness metrics and designing safeguards that ensure equitable performance across student populations.

Finally, this study focused on students in Grades 3–5, a population often overlooked in AES research, thereby extending the scope of evidence to younger learners who are at a critical stage in writing development. Future work should extend these findings to additional grade levels, writing genres, and more diverse student populations. There is also a need for clearer evaluation frameworks and design guidelines to ensure prompt quality and subgroup fairness in LLM-based scoring. As LLMs gain traction in educational measurement, this study underscores the need to pair advanced modeling with thoughtful design to support scoring accuracy, fairness, and validity.

## 7 Limitations

Several limitations should be acknowledged. First, only 10% of essays were double-scored, limiting the reliability of human–human benchmarks, particularly for subgroup comparisons. Second, the analysis focused solely on informative writing tasks, leaving other genres such as argumentative or narrative unexamined. Finally, only three prompting strategies were tested, while other approaches, such as extended rubric prompts or fine-tuning, remain unexplored. These constraints suggest caution in interpreting findings and point to directions for future research.

## References

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, Vancouver, Canada. Association for Computational Linguistics.

Ahmed M. ElMassry, Nazar Zaki, Negmeldin AlSheikh, and Mohammed Mediani. 2025. A systematic review of pretrained models in automated essay scoring. *IEEE Access*, pages 1–1. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

Ben Hamner, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. 2012. The hewlett foundation: Automated essay scoring [data competition]. Kaggle. Accessed: 2025-02-15.

Yue Huang, Corey Palermo, Ruitao Liu, and Yong He. 2025. An early review of generative language models in automated writing evaluation: Advancements, challenges, and future directions for automated essay scoring and feedback generation. *Chinese/English Journal of Educational Measurement and Evaluation*, 6(2).

Mohamed Abdellatif Hussein, Hesham Hassan, and Mohammad Nassef. 2019. Automated language essay scoring systems: a literature review. *PeerJ Computer Science*, 5:e208.

Vivekanandan S. Kumar and David Boulanger. 2021. Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education*, 31(3):538–584.

Ehsan Latif and Xiaoming Zhai. 2024. Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100210.

Diane Litman, Haoran Zhang, Richard Correnti, Lindsay Clare Matsumura, and Elaine Wang. 2021. A fairness evaluation of automated methods for scoring text evidence usage in writing. In Ido Roll, Danielle McNamara, Sergey Sosnovsky, Rose Luckin, and Vania Dimitrova, editors, *Artificial Intelligence in Education*, volume 12748, pages 255–267. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Watheq Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? *arXiv preprint*. ArXiv:2403.06149 [cs].

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large Language Models: A Survey. *arXiv preprint*. ArXiv:2402.06196 [cs].

Christopher Ormerod and Alexander Kwako. 2024. Automated text scoring in the age of generative AI for the GPU-poor. *Chinese/English Journal of Educational Measurement and Evaluation*, 5(3).

Ellis B. Page. 1966. The imminence of grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Bernadette Quah, Lei Zheng, Timothy Jie Han Sng, Chee Weng Yong, and Intekhab Islam. 2024. Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Medical Education*, 24(1). Publisher: Springer Science and Business Media LLC.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. 2019. Language models and Automated Essay Scoring. *arXiv preprint*. ArXiv:1909.09482 [cs].

Meilia Nur Indah Susanti, Arief Ramadhan, and Harco Leslie Hendric Spit Warnars. 2023. Automatic essay exam scoring system: a systematic literature review. *Procedia Computer Science*, 216:531–538.

Kaveh Taghipour and Hwee Tou Ng. 2016. A Neural Approach to Automated Essay Scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas. Association for Computational Linguistics.

Tamara P. Tate, Jacob Steiss, Drew Bailey, Steve Graham, Youngsun Moon, Daniel Ritchie, Waverly Tseng, and Mark Warschauer. 2024. Can AI provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence*, 7:100255. Publisher: Elsevier BV.

Masaki Uto. 2021. A review of deep-neural automated essay scoring models. *Behaviormetrika*, 48(2):459–484.

Masaki Uto and Masashi Okano. 2020. Robust neural automated essay scoring using item response theory. In *Artificial Intelligence in Education: Proceedings of the 21st International Conference*, pages 549–561, Cham. Springer International Publishing.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 24824–24837, Red Hook, NY, USA. Curran Associates Inc.

Joshua Wilson and Yue Huang. 2024. Validity of automated essay scores for elementary-age English language learners: Evidence of bias? *Assessing Writing*, 60:100815.

Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2724–2733, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online. Association for Computational Linguistics.

# A Additional Figures

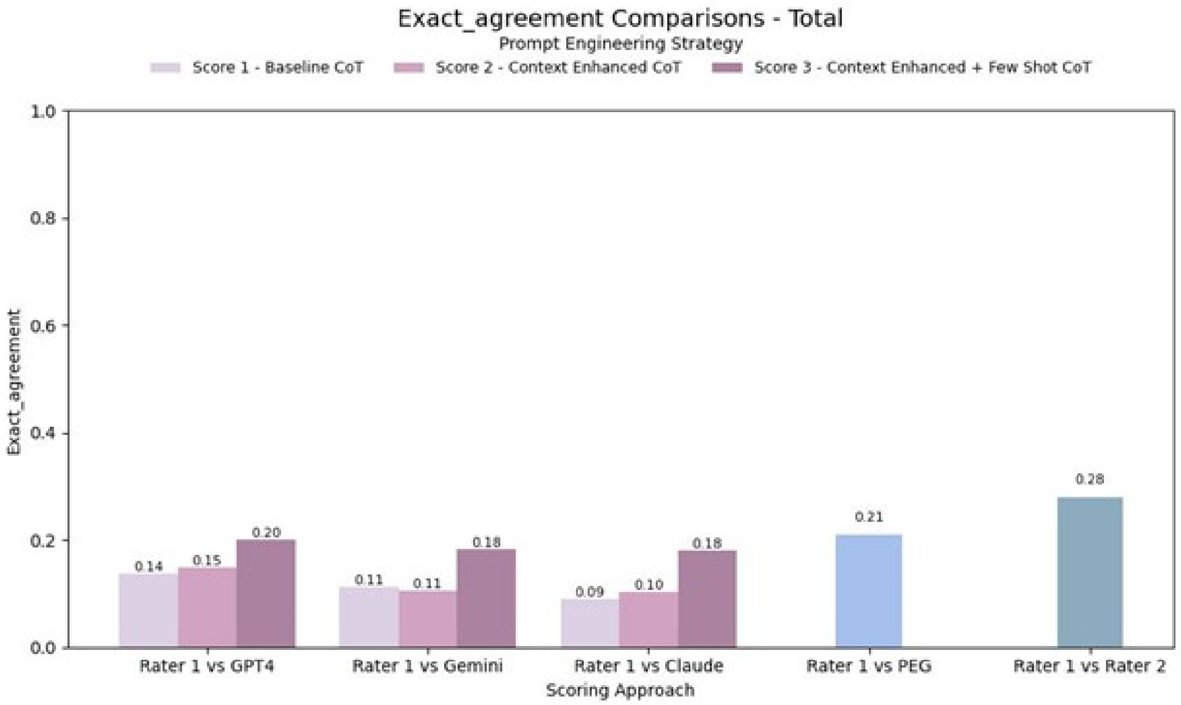


Figure A1: Comparisons of Exact Agreement between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs – Total Score

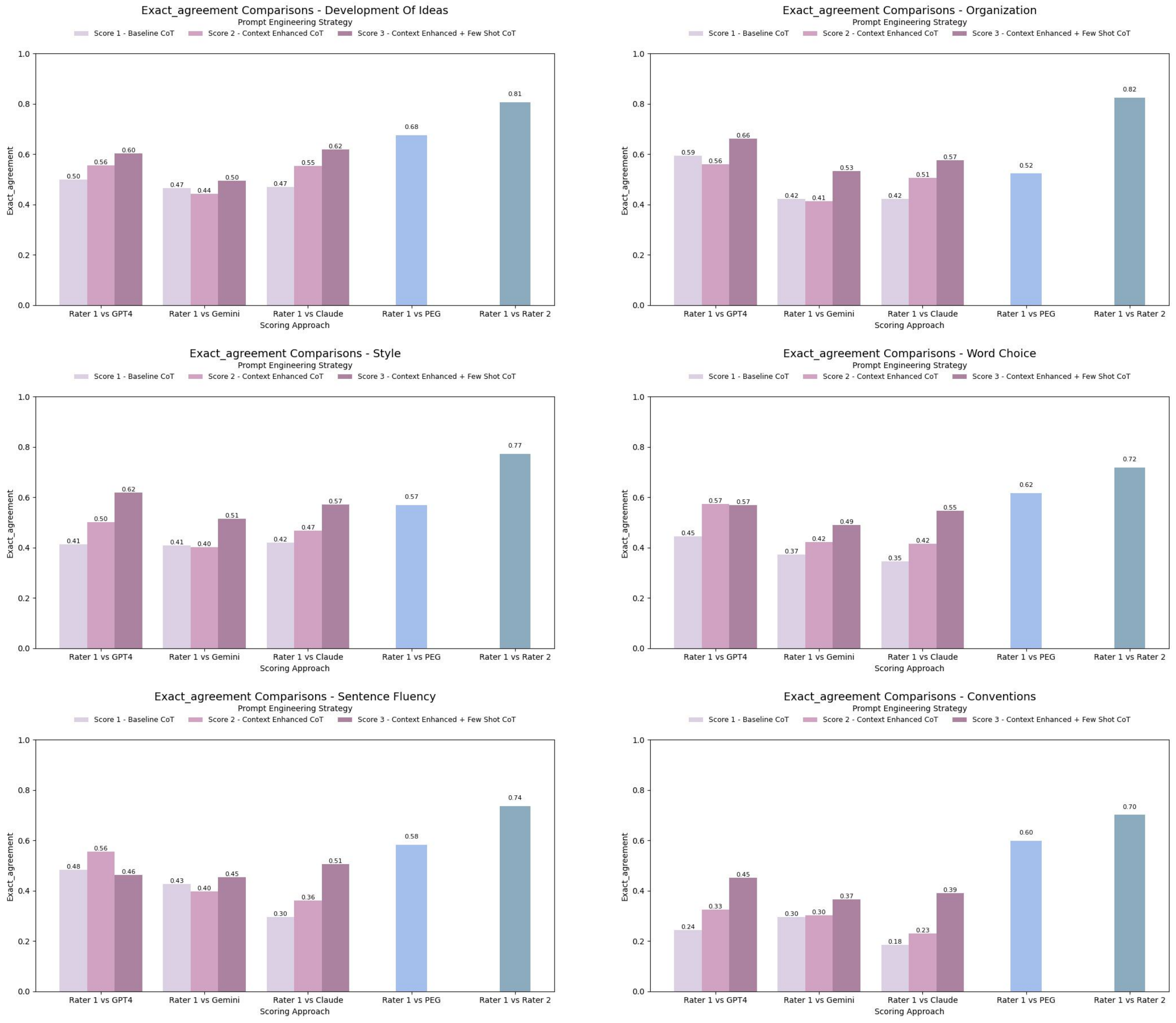


Figure A2: Comparisons of Exact Agreement between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs – by Trait

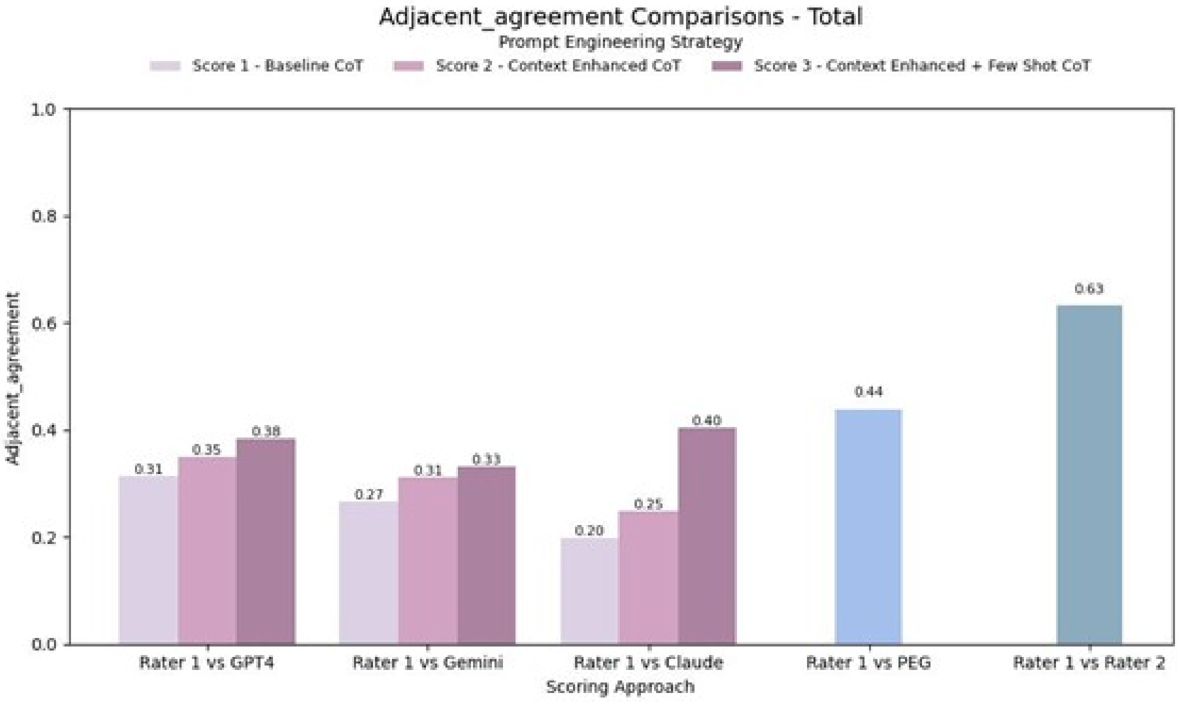


Figure A3: Comparisons of Adjacent Agreement between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs – Total Score

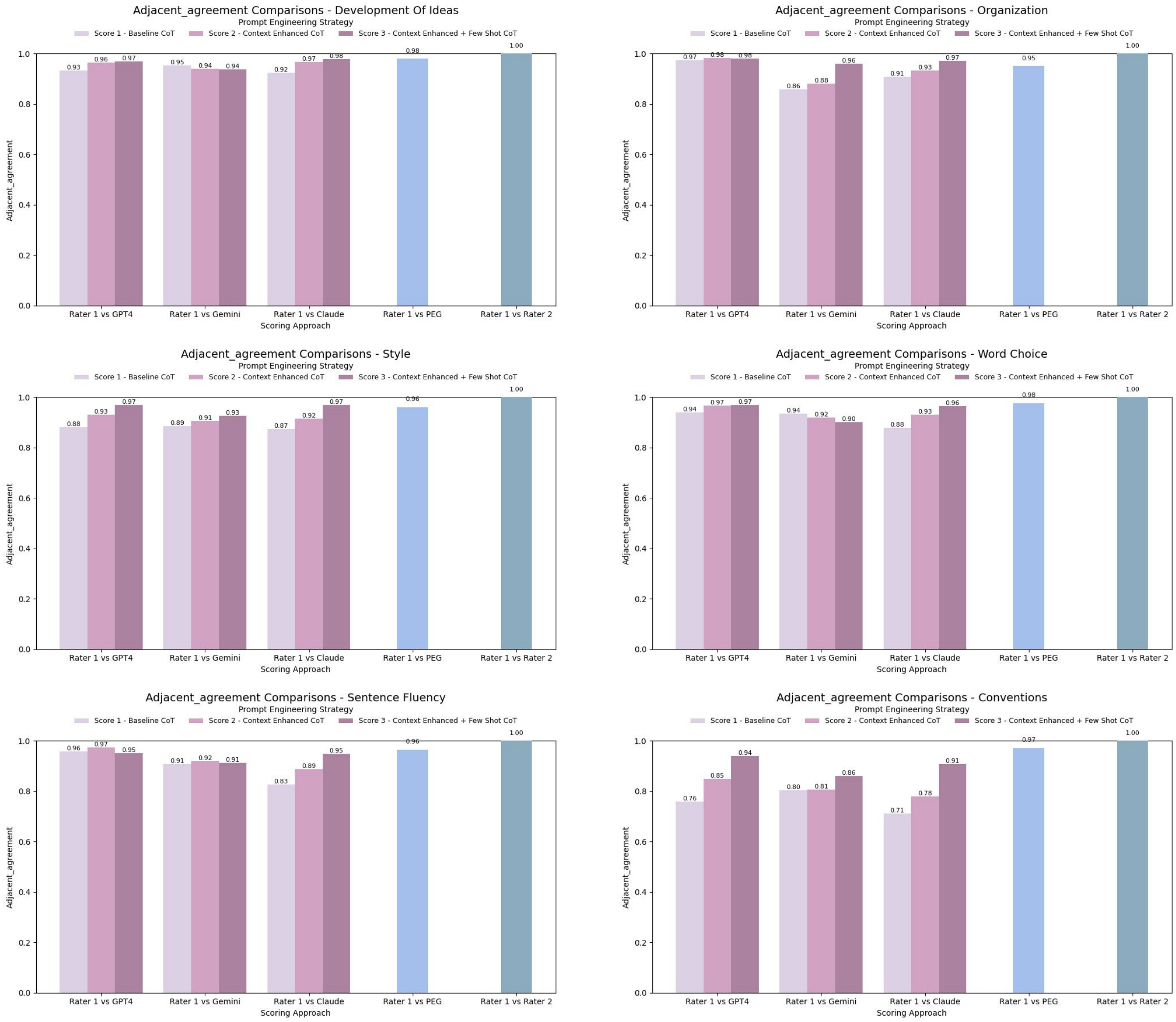


Figure A4: Comparisons of Adjacent Agreement between Human Scores and Machine Scores across Prompt Engineering Strategies and LLMs – by Trait