

Evaluating Machine Learning, Transformer, and LLM Approaches for Large-Scale Automated Essay Scoring

Cyrus Sai-Cheong Chan¹, Charles Anifowose¹,
Su Zhang², Winnie Wing-Yee Tse², Nizam Radwan², Hyunah Kim²

¹ Vretta

² Education Quality and Accountability Office

Abstract

This study compared machine learning, transformer fine-tuning, prompt-based LLM, and novel ensemble approaches for automated essay scoring in a Canadian large-scale provincial assessment. Fine-tuned transformers achieved the highest reliability, followed by machine learning models. Meanwhile, prompt-based LLMs provided greater explainability, and ensemble architectures highlighted opportunities for balancing reliability and explainability.

1 Introduction

1.1 Automated Essay Scoring

Automated essay scoring (AES) has evolved across multiple technical generations, encompassing feature-engineered machine learning (ML) models, fine-tuned transformer architectures, and prompt-based large language models (LLMs). Each approach presents distinct trade-offs regarding scoring reliability, system transparency, and operational feasibility (Atkinson & Palma, 2025; Huang & Wilson, 2025; Li & Ng, 2024; Ramesh & Sanampudi, 2022). Although prompt-based LLM scoring has attracted growing research attention (Chu et al., 2025; Hou et al., 2025; Lee et al, 2024), operational implementations in large-scale K-12 assessments have mainly relied on traditional ML-based engines (Zhang et al., 2026).

1.2 Aim

This study evaluates the operational feasibility of diverse AES approaches within a high-stakes, large-scale Canadian provincial assessment. Specifically, it compares traditional ML,

transformer fine-tuning, prompt-based LLMs, and novel ensemble architectures across four key dimensions: scoring reliability, subgroup consistency, explainability, and operational feasibility.

2 Sample

This study utilized student responses and scoring data across three long-writing items from two assessment administrations in 2023-2024. Student responses for long-writing items were evaluated on two scoring scales: Topic Development (TD; 6-point scale) and Conventions (CV; 4-point scale). Generic, scale-specific rubrics were established and applied across writing prompts.

A stratified sampling strategy based on administration window, gender, and English Language Learner (ELL) status ensured the representativeness of the sample. Two operational datasets and an expert-audit dataset were constructed to develop and evaluate the AES approaches.

2.1 Operational Dataset A

Operational Dataset A (Operational A) was created for ML model training, transformer fine-tuning, and ensemble architecture evaluation. Comprising 15,000 student responses, it was split into training, validation, and testing sets using a 70/15/15 ratio. The training set was used to fit the ML models and fine-tune the transformers, but was not required for the ensemble approach as it involved no dedicated training phase. The validation set served to compare candidate models within each approach and select the best-performing configurations. The testing set was reserved for final independent model evaluations.

To ensure direct comparability across the four AES approaches on a unified benchmark, all best-performing models, including the prompt-based LLM, were evaluated on the held-out Operational A testing set.

2.2 Operational Dataset B

Operational Dataset B (Operational B) was dedicated exclusively to LLM prompt optimization. Comprising 12,000 student responses, it was independently sampled from the same overarching pool of operational responses as Operational A. Consequently, the two datasets shared a partial overlap of 3,191 samples. Operational B was partitioned into training, validation, and testing sets using a 20/40/40 split. The training set facilitated subject matter expert (SME) feedback collection and iterative prompt optimization, the validation set was used to select the best performing prompt templates, and the testing set provided an independent evaluation of the selected prompt template.

2.3 Expert-Audit Dataset

An expert-audit dataset was established as an authoritative benchmark to ensure model alignment with expert scoring standards. The dataset comprised 404 responses for TD and 387 responses for CV across the three long-writing items, covering all responses used in operational scorer qualifying tests and validity checks. Scores were assigned and validated by expert scorers, reflecting rigorous operational standards. Within this dataset, 21 responses overlapped across the TD and CV scales. Relative to the broader operational datasets, this audit set shared 29 overlapping responses with Operational A and 28 with Operational B.

3 Methodology

Four distinct AES approaches were implemented and evaluated. Within each approach, various model configurations were developed and benchmarked to select optimal candidates.

3.1 Machine Learning Approach

Score prediction was framed as a regression task, with clipping and rounding applied during post-processing to align continuous predictions with valid score points. Feature engineering generated 22 linguistic features for TD and 24 for CV. Three

scale-specific ML models were trained on identical feature sets: Ridge Regression (RR), Support Vector Regression (SVR), and XGBoost Regression (XGB).

3.2 Transformer Fine-Tuning Approach

The transformer fine-tuning approach represented a middle ground between feature-engineered ML engines and prompt-driven LLMs. Three pretrained transformers, BERT, DistilBERT, and RoBERTa, were fine-tuned for each scale using uniform hyperparameter configurations: 4 epochs, a learning rate of $1e-5$, weight decay of 0.01, and FP16 mixed precision.

3.3 Prompt-Based LLM Approach

The prompt-based approach utilized Llama 3 70B, an open-weight foundation model, configured with zero-shot chain-of-thought prompting. Prompt context included scoring rubrics, scorer training guides, additional scoring instructions, item prompts, and student responses. The model generated both numerical scores and qualitative scoring rationales.

Prompt optimization for the TD scale involved four iterative rounds on non-overlapping subsets of the Operational B training set. SME feedback on discrepant cases was systematically incorporated to refine the additional scoring instructions, yielding four distinct prompt templates.

3.4 Ensemble Approach

The ensemble approach integrated traditional ML with prompt-based LLMs through three mechanisms: linguistic feature injection, reference score injection, and combined injection.

For linguistic feature injection, rubric-relevant linguistic features were aggregated into qualitative performance indicators and injected into the top-performing LLM prompt template as objective evidence to assist model reasoning.

For reference score injection, predictions from top-performing ML models were provided within the prompt to inform, but not dictate, LLM evaluations. The LLM retained autonomy to override reference scores if its internal reasoning diverged from ML predictions.

For combined injection, both qualitative performance indicators and ML-predicted reference scores were integrated into the augmented LLM prompt template.

Dataset	Item	Metric	Human	ML	Transformer	LLM	Ensemble
Operational A	Item 1	QWK	0.76	0.83	0.84	0.67	0.76
		SMD	-0.13	-0.08	-0.03	-0.33	-0.14
		EA	57.82	64.67	65.87	44.27	52.40
	Item 2	QWK	0.74	0.78	0.80	0.64	0.71
		SMD	0.02	-0.03	-0.01	0.15	0.16
		EA	59.74	66.40	66.40	49.33	53.87
	Item 3	QWK	0.75	0.85	0.87	0.69	0.76
		SMD	-0.12	0.01	0.04	-0.14	-0.04
		EA	55.04	66.53	67.47	46.27	50.80
Operational B	Item 1	QWK	0.76	-	-	0.70	-
		SMD	-0.13	-	-	-0.25	-
		EA	57.82	-	-	49.00	-
	Item 2	QWK	0.74	-	-	0.67	-
		SMD	0.02	-	-	0.12	-
		EA	59.74	-	-	47.69	-
	Item 3	QWK	0.75	-	-	0.75	-
		SMD	-0.12	-	-	-0.10	-
		EA	55.04	-	-	47.94	-
Expert-audit	Item 1	QWK	0.97	0.97	0.98	0.88	0.92
		SMD	-0.02	-0.06	-0.02	-0.04	-0.13
		EA	82.61	83.96	85.85	50.94	59.43
	Item 2	QWK	0.95	0.94	0.97	0.83	0.90
		SMD	-0.03	-0.13	-0.06	0.14	-0.16
		EA	81.03	76.19	86.39	44.90	55.78
	Item 3	QWK	0.96	0.96	0.98	0.89	0.92
		SMD	-0.05	-0.11	-0.03	0.05	-0.14
		EA	84.23	78.15	86.09	51.66	52.98

Table 1: Model testing results with human-human agreement baseline on Topic Development.

3.5 Evaluation Framework

Human-human agreement baselines were established to benchmark automated model performance. The operational baseline was calculated using all available double-scored operational responses rather than dataset-specific samples, thereby reflecting overall human reliability and providing a common reference benchmark for both Operational Datasets A and B. For the expert-audit dataset, the baseline was calculated by comparing human scores against expert scores, using bootstrapping with 10,000 iterations to derive stable mean agreement statistics across multi-scored responses.

Automated model performance was evaluated relative to the human-human agreement baselines using industry-standard operational metrics with acceptance thresholds based on cross-jurisdictional practices (Zhang et al., 2026):

- Quadratic Weighted Kappa (QWK) ≥ 0.70

- Absolute value of Standardized Mean Difference (SMD) ≤ 0.15
- Exact Agreement (EA) rate $\geq 65\%$
- QWK degradation from the human-human baseline ≤ 0.10
- EA degradation from the human-human baseline ≤ 5.25 percentage points (pp)

Subgroup performance was monitored across gender and ELL status to evaluate scoring equity. Subgroup analyses were limited to Operational Datasets A and B due to sample size constraints in the expert-audit dataset.

4 Results

Among the trained ML models, XGB achieved the highest scoring performance on TD, while SVR performed best on CV. DistilBERT emerged as the top-performing fine-tuned transformer model across both scales. For prompt-based LLMs, four rounds of prompt optimization were completed on

Dataset	Item	Metric	Human	ML	Transformer	LLM	Ensemble
Operational A	Item 1	QWK	0.61	0.70	0.75	-	-
		SMD	-0.01	0.05	0.14	-	-
		EA	54.38	55.33	61.07	-	-
	Item 2	QWK	0.62	0.63	0.67	-	-
		SMD	0.03	0.13	0.18	-	-
		EA	56.21	54.00	55.60	-	-
	Item 3	QWK	0.60	0.67	0.70	-	-
		SMD	-0.15	0.10	0.16	-	-
		EA	52.05	55.07	56.67	-	-
Operational B	Item 1	QWK	0.61	-	-	0.54	-
		SMD	-0.01	-	-	0.24	-
		EA	54.38	-	-	51.50	-
	Item 2	QWK	0.62	-	-	0.40	-
		SMD	0.03	-	-	0.23	-
		EA	56.21	-	-	46.00	-
	Item 3	QWK	0.60	-	-	0.48	-
		SMD	-0.15	-	-	0.22	-
		EA	52.05	-	-	49.50	-
Expert-audit	Item 1	QWK	0.90	0.88	0.92	0.65	0.72
		SMD	-0.05	0.03	0.03	-0.03	0.02
		EA	77.79	72.26	79.56	45.26	48.91
	Item 2	QWK	0.89	0.85	0.87	0.70	0.79
		SMD	-0.05	0.05	0.12	0.29	0.19
		EA	77.86	70.77	71.54	54.62	58.46
	Item 3	QWK	0.94	0.89	0.89	0.78	0.85
		SMD	0.00	0.13	0.18	0.11	0.17
		EA	85.41	74.17	76.67	61.67	71.67

Table 2: Model testing results with human-human agreement baseline on Conventions.

TD with the final prompt template version producing the strongest performance, whereas the CV prompt remained unoptimized. Among ensemble strategies, reference score injection achieved the highest scoring agreement on TD, while combined injection performed best on CV.

Table 1 and Table 2 present testing results for the top-performing models across all AES approaches on the TD and CV scales, respectively, alongside the human-human baseline. The top-performing ML models demonstrated strong, stable predictive performance. Fine-tuned transformers achieved even slightly higher reliability while eliminating manual feature engineering. Prompt-based LLMs generated human-readable scoring rationales, provided an auditable prompt optimization process, and incorporated complex scoring standards via natural-language instructions. The ensemble approach outperformed standalone prompt-based LLMs while achieving slightly

lower reliability than pure ML models, successfully preserving rationale generation capabilities.

Table 3 and Table 4 summarize subgroup performance on the testing sets. ML and transformer models exhibited consistent performance across Female, Male, and ELL subgroups, though performance dropped slightly on the non-ELL subgroup. Prompt-based LLM and ensemble approaches showed lower consistency across Female and Non-ELL subgroups, exhibiting a pronounced drop on the latter.

Prompt-based LLM and ensemble results on CV are omitted from Operational A in Table 2 and Table 4 because no prompt optimization was conducted on CV, leaving no optimized template to evaluate or integrate into the ensemble architecture for that dataset. The LLM results reported for CV on Operational B represent the performance of the unoptimized initial prompt template. Ensemble injection methods for CV were evaluated and compared exclusively on the expert-audit dataset.

Dataset	Subgroup	Metric	ML	Transformer	LLM	Ensemble
Operational A	Female	QWK	0.81	0.82	0.61	0.70
		SMD	-0.07	-0.01	-0.11	-0.01
		EA	67.50	66.96	47.87	52.75
	Male	QWK	0.83	0.85	0.71	0.77
		SMD	0.00	0.01	-0.12	-0.01
		EA	64.23	66.19	45.37	51.96
	ELL	QWK	0.85	0.87	0.72	0.79
		SMD	-0.01	-0.01	-0.13	-0.04
		EA	65.10	66.34	44.58	52.04
	Non-ELL	QWK	0.76	0.78	0.54	0.63
		SMD	-0.07	0.02	-0.09	0.03
		EA	66.64	66.81	48.67	52.67
Operational B	Female	QWK	-	-	0.68	-
		SMD	-	-	-0.09	-
		EA	-	-	49.24	-
	Male	QWK	-	-	0.72	-
		SMD	-	-	-0.08	-
		EA	-	-	48.16	-
	ELL	QWK	-	-	0.73	-
		SMD	-	-	-0.11	-
		EA	-	-	47.43	-
	Non-ELL	QWK	-	-	0.65	-
		SMD	-	-	-0.05	-
		EA	-	-	49.92	-

Table 3: Model subgroup performance on Topic Development.

5 Discussion

These findings illustrate distinct trade-offs among AES approaches in K-12 large-scale assessment. Fine-tuned transformers achieved the highest overall reliability and subgroup consistency, followed by ML models; both met most operational acceptance thresholds despite offering limited transparency into their scoring decisions. Conversely, the open-weight prompt-based LLM yielded lower scoring agreement and met fewer quantitative thresholds, but generated human-readable rationales that provided useful supporting information for SME review and scoring decisions. These findings align with empirical studies showing that fine-tuned transformers outperform traditional regression-based methods in scoring accuracy (Wang et al., 2022; Yang et al., 2020), whereas prompt-engineered LLMs consistently underperform relative to ML and transformer baselines (Stahl et al., 2024; Yancey et al., 2023).

Despite strong natural language understanding capabilities, prompt engineering alone proved

insufficient to match the predictive performance of traditional ML and fine-tuned transformers, which is consistent with broader literature (Atkinson & Palma, 2025; Li & Ng, 2024; Ramesh & Sanampudi, 2022; Xiao et al., 2024).

However, the ensemble findings demonstrate that these AES approaches can be mutually reinforcing. Integrating ML predictions into prompt-based LLMs via reference score or combined injection offered a potential balance, improving reliability while retaining human-readable rationales that could support SME review.

Subgroup analyses underscored the necessity of evaluating scoring fairness alongside aggregate metrics. All evaluated AES approaches displayed varying degrees of subgroup inconsistency, reflecting a prevalent pattern in recent literature (Huang & Wilson, 2025; Huang et al., 2026; Yang et al., 2024). Notably, certain disparities observed in this study diverged in direction from prior research, even when evaluating identical approaches and demographic groups, such as prompt-based LLM performance across ELL and non-ELL subgroups (Huang & Wilson, 2025). This

Dataset	Subgroup	Metric	ML	Transformer	LLM	Ensemble
Operational A	Female	QWK	0.66	0.69	-	-
		SMD	0.11	0.18	-	-
		EA	57.10	59.86	-	-
	Male	QWK	0.66	0.71	-	-
		SMD	0.08	0.15	-	-
		EA	52.49	55.69	-	-
	ELL	QWK	0.67	0.71	-	-
		SMD	0.13	0.16	-	-
		EA	51.78	54.44	-	-
	Non-ELL	QWK	0.62	0.65	-	-
		SMD	0.06	0.18	-	-
		EA	57.83	61.12	-	-
Operational B	Female	QWK	-	-	0.42	-
		SMD	-	-	0.21	-
		EA	-	-	48.28	-
	Male	QWK	-	-	0.48	-
		SMD	-	-	0.27	-
		EA	-	-	49.85	-
	ELL	QWK	-	-	0.53	-
		SMD	-	-	0.29	-
		EA	-	-	33.33	-
	Non-ELL	QWK	-	-	0.47	-
		SMD	-	-	0.22	-
		EA	-	-	49.74	-

Table 4: Model subgroup performance on Conventions.

divergence indicates that the magnitude and direction of potential bias depend not only on the AES architecture and target subgroups, but also on specific assessment contexts and item prompts. Interpreting these patterns alongside qualitative audits of model reasoning and human scoring behaviors will be essential for identifying underlying causes and promoting equitable scoring.

Beyond predictive performance, operational demands varied considerably across approaches. ML and transformer models require substantial labeled training data and full model retraining for updates. In contrast, prompt-based LLMs facilitate rapid, low-resource iterations via prompt optimization. Ultimately, the transparent rationales offered by LLM and ensemble architectures provide key advantages for system validation, auditing, and stakeholder trust.

A primary limitation of this study is that LLM prompt optimization was conducted exclusively on the TD scale. The CV scale remained unoptimized, leading to incomplete prompt-based LLM and ensemble results that complicate comparisons

against other approaches. Future research should apply consistent optimization protocols across all scoring scales to evaluate domain- and scale-specific generalizability. Additionally, while subgroup analyses identified performance discrepancies across demographics, the root causes remain unexamined. Future studies should integrate model error analyses, human scoring patterns, and student writing characteristics to isolate contributing factors and establish frameworks for mitigating subgroup variance.

6 Conclusions

This study conducted a comparative evaluation of four AES approaches, including traditional ML, transformer fine-tuning, prompt-based LLMs, and a novel ensemble approach, within a high-stakes, large-scale assessment context.

The results confirm that fine-tuned transformers deliver the highest scoring reliability and subgroup consistency, followed by traditional ML models. Meanwhile, the novel ensemble architecture bridges the gap between ML reliability and LLM explainability, demonstrating that hybrid

architectural designs can effectively leverage the complementary strengths of both approaches.

Finally, operational readiness in AES requires evaluating models beyond raw reliability to ensure fairness, auditability, transparency, and defensible decision-making.

References

- Atkinson, John, & Palma, Diego. (2025). An LLM-based hybrid approach for enhanced automated essay scoring. *Scientific Reports*, 15(1). <https://doi.org/10.1038/s41598-025-87862-3>.
- Chu, SeongYeub, Kim, JongWoo, Wong, Bryan, & Yi, MunYong. (2025). Rationale behind essay scores: Enhancing S-LLM's multi-trait essay scoring with rationale generated by LLMs. *arXiv*. <https://arxiv.org/abs/2410.14202>.
- Hou, Zhaoyi Joey, Ciuba, Alejandro, & Li, Xiang Lorraine. (2025). Improve LLM-based automatic essay scoring with linguistic features. *arXiv*. <https://arxiv.org/abs/2502.09497>.
- Huang, Yue, & Wilson, Joshua. (2025). Evaluating LLM-based automated essay scoring: accuracy, fairness, and validity. *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con)* (pp. 71–83). National Council on Measurement in Education (NCME). <https://aanthology.org/2025.aimecon-wip.9.pdf>.
- Huang, Yue, Palermo, Corey, & Wilson, Joshua. (2026). Accuracy and fairness of generative AI in automated essay scoring: Comparing GPT-4o, feature-based models, and human raters. *Assessing Writing*, 69. <https://doi.org/10.1016/j.asw.2026.101047>.
- Lee, Sanwoo, Cai, Yida, Meng, Desong, Wang, Ziyang, & Wu, Yunfang. (2024). Unleashing large language models' proficiency in zero-shot essay scoring. *arXiv*. <https://arxiv.org/abs/2404.04941>.
- Li, Shengjie, & Ng, Vincent. (2024). Automated essay scoring: a reflection on the state of the art. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 17876–17888). Association for Computational Linguistics. <http://doi.org/10.18653/v1/2024.emnlp-main.991>.
- Ramesh, Dadi, & Sanampudi, Suresh Kumar. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55, 2495–2527. <https://doi.org/10.1007/s10462-021-10068-2>.
- Stahl, Maja, Biermann, Leon, Nehring, Andreas, & Wachsmuth, Henning. (2024). Exploring LLM prompting strategies for joint essay scoring and feedback generation. *arXiv*. <https://arxiv.org/abs/2404.15845>.
- Wang, Yongjie, Wang, Chuang, Li, Ruobing, & Lin, Hui. (2022). On the use of Bert for automated essay scoring: Joint learning of multi-scale essay representation. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3416–3425). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.naacl-main.249>.
- Xiao, Changrong, Ma, Wenxing, Song, Qingping, Xu, Sean Xin, Zhang, Kunpeng, Wang, Yufang, & Fu, Qi. (2024). Human-AI collaborative essay scoring: A dual-process framework with LLMs. *arXiv*. <https://arxiv.org/abs/2401.06431>.
- Yancey, Kevin P., Laflair, Geoffrey, Verardi, Anthony, & Burstein, Jill. (2023). Rating short L2 essays on the CEFR scale with GPT-4. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 576–584). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.bea-1.49>.
- Yang, Kaixun, Raković, Mladen, Li, Yuyang, Guan, Quanlong, Gašević, Dragan, & Chen, Guangliang. (2024). Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20), 22466–22474. <https://doi.org/10.1609/aaai.v38i20.30254>.
- Yang, Ruosong, Cao, Jiannong, Wen, Zhiyuan, Wu, Youzheng, & He, Xiaodong. (2020). Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1560–1569). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.findings-emnlp.141>.
- Zhang, Su, Hove, Jennifer, & Gu, Zhimei. (2026, October 5-7). *A jurisdictional scan of automated scoring practices in K-12 summative assessment programs* [Conference presentation]. Artificial Intelligence in Measurement and Education Conference, Pittsburgh, PA, United States.