Evaluating the Evaluators: Metrics for Automated Essay Feedback Generation

Maryam Berijanian[1,*], Christopher G. Shaltry[2], and Dirk Colbry[1]

[1]Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI, USA
[2]College of Human Medicine and College of Osteopathic Medicine, Michigan State University, East Lansing, MI, USA
Email: berijani@msu.edu (M.B.); shaltryc@msu.edu (C.G.S.); colbrydi@msu.edu (D.C.)
[*]Corresponding author

Manuscript received August 14, 2025; revised October 9, 2025; accepted November 26, 2025; published April 10, 2026

***Abstract*—Automated Essay Scoring (AES) systems have improved with advances in Natural Language Processing (NLP), but they often prioritize grade prediction over qualitative feedback, which is crucial for student learning. This study evaluates the capabilities of Large Language Models (LLMs) in generating detailed, context-specific feedback, with the goal of improving student understanding. More importantly, it focuses on validating the efficacy of widely used NLP metrics for assessing feedback quality, analyzing their alignment with human judgments. The findings highlight both the strengths and limitations of these metrics in evaluating qualitative feedback within AES contexts. By comparing LLM performance under few-shot learning and fine-tuning conditions, the study identifies both promising directions and persistent challenges in automated feedback generation. Overall, the results emphasize that while LLMs can enhance feedback generation, current metrics remain inadequate for reliably guiding such improvement, underscoring the need for more robust evaluation frameworks.[1]**

***Keywords*—automated essay scoring, textual feedback generation, meta-metrics, human-in-the-loop**

I. Introduction

Since the creation of the first Automated Essay Scoring System (AES) in 1966, which had limited grammar evaluation capabilities [1], AES has evolved significantly. Modern systems leverage advances in Natural Language Processing (NLP) to assess essay responses in standardized testing and online learning platforms [2]. These systems save instructor time and reduce grading inconsistencies, making them valuable tools in educational settings.

Despite these advancements, many AES models continue to focus primarily on assigning numerical scores, often overlooking the generation of qualitative, context-sensitive feedback that explains the rationale behind these scores. Such feedback is crucial for supporting student learning by identifying strengths and areas for improvement. However, the integration of high-quality qualitative feedback has been hindered by two interrelated challenges: its frequent omission in AES systems and the absence of standardized metrics to assess feedback quality.

In practice, many instructors currently use commercial platforms, such as e-rater® Scoring Engine, IntelliMetric®, Grammarly, and PackBack to automate feedback on student writing. While these tools can quickly flag grammar and spelling errors or surface-level style issues, they are typically evaluated using proprietary benchmarks or internal user-satisfaction surveys. Crucially, they rarely offer the nuanced, argument-level critique or context-aware suggestions that human reviewers provide. This gap underscores the need for research capable of both generating and rigorously evaluating deeper, qualitative feedback.

On one hand, qualitative feedback is often missing because traditional AES systems are designed primarily for score prediction. On the other hand, even when qualitative feedback is generated, its evaluation is problematic. By nature, qualitative feedback is subjective and highly dependent on context, making it difficult to define clear, objective evaluation criteria. The development of standardized metrics is essential because it provides a consistent framework for evaluating and comparing the quality and relevance of feedback across different AES systems. Such standardization would help bridge the gap between automated assessments and human judgment, ensuring that feedback not only explains the scoring process but also genuinely aids in learning.

This study shifts the focus from developing an assessment tool toward validating how effectively existing NLP metrics capture the quality of feedback produced by Large Language Models (LLMs). To achieve this, we incorporate several NLP models, including Transformer-based models such as LLaMA and DeepSeek, which have shown strong capabilities in understanding and generating natural language. These models serve as the basis for exploring different learning methods, such as fine-tuning and few-shot learning. Rather than attempting to train the best possible feedback model, the aim is to evaluate how well commonly used NLP metrics, such as BLEU [3], ROUGE [4], and BERTScore [5], reflect human judgments of feedback quality. By analyzing the correspondence between these metrics and human evaluations, the study reveals their limitations and highlights the need for more specialized metrics that can reliably assess qualitative feedback. Ultimately, this work seeks to enhance the evaluation frameworks that underpin AES research and, in turn, improve the educational value of automated feedback.

A. Objectives and Contributions

The primary objective of this study is to evaluate the effectiveness of existing evaluation metrics in assessing the quality of feedback generated by LLMs within AES systems, rather than to develop or train a new scoring model. Specifically, the study examines how well various metrics align with human judgments of feedback quality. Providing detailed, context-specific essay feedback from LLMs has the potential to support deeper student learning compared to the isolated numerical scores typically produced by traditional AES systems. By systematically assessing both the

[1]The complete codebase is publicly available on GitHub for reproducibility at https://github.com/maryambrj/essay_feedback_meta_metric.git

performance of LLMs and the reliability of evaluation metrics, this study seeks to clarify the strengths and limitations of current approaches and to inform the development of more accurate frameworks for automated feedback evaluation.

A key contribution of this study is the comprehensive evaluation of commonly used NLP metrics such as BLEU [3], ROUGE [4], METEOR [6], GLEU [7], and BERTScore [5]. While these metrics are widely applied in other text generation tasks, their suitability for measuring feedback quality in essay grading remains largely unexamined. By correlating metric scores with human evaluations, the study provides empirical evidence on the reliability, interpretability, and limitations of these metrics in the AES context. This analysis advances understanding of their practical applicability and highlights the need for more specialized, feedback-oriented evaluation standards.

Another contribution of this work is the comparison of multiple LLMs, including T5 [8], BART [9], LLaMA [10], and DeepSeek [11], under different learning settings such as few-shot learning and fine-tuning. This comparative analysis identifies the relative strengths and weaknesses of each model in generating coherent, context-aware, and pedagogically meaningful feedback, offering guidance for future improvements in automated writing assessment.

II. Related Work

AES systems have evolved through both commercial and academic implementations. For instance, well-known commercial systems such as Educational Testing Service's (ETS) e-rater® and Vantage Labs' IntelliMetric® have seen widespread adoption in standardized testing. In parallel, several lab-based systems and research prototypes, such as the Intelligent Essay Assessor (IEA) [12] and systems developed as part of the Automated Student Assessment Prize (ASAP-AES) challenge [13–15], continue to push the boundaries of research in this area.

Most AES systems fundamentally function as scoring mechanisms where the primary objective is to predict the final grade of an essay [16, 17]. To do so, these systems consider factors such as essay length, lexical diversity, prompt relevance, readability, syntactic structure, argumentation quality, semantic coherence, and discourse features. Early and traditional implementations relied on established NLP techniques, including bag-of-words (BOW) models and Long Short-Term Memory (LSTM) networks [18], which capture basic text features through statistical representations of word occurrences. Moreover, transformer-based architectures such as BERT [19], XLNet [20], and similar models have also been employed in AES [17] to capture deeper contextual and semantic nuances. Such approaches provided a foundation that, while effective to a degree, often depends on hand-engineered features or simpler learning models and thus may miss the complex linguistic patterns present in high-quality writing.

Building on these traditional methods, state-of-the-art AES models incorporate carefully selected subsets of these features as training objectives to generate a holistic score [16]. Training objectives refer to the specific tasks or attributes that a model is optimized for during the learning process. By focusing on crucial writing qualities such as syntactic structure, semantic depth, and discourse organization, these objectives allow a model to capture complex dimensions of writing. This ensures that the resulting score reflects a comprehensive evaluation rather than merely a superficial count of easily measurable characteristics.

More recent research has explored the integration of Large Language Models (LLMs) into AES systems. LLMs, such as DeepSeek and LLaMA, represent a significant shift from traditional NLP approaches. Unlike earlier models that only partially capture context through fixed feature sets, LLMs leverage transformer architectures and extensive pretraining to generate context-specific, coherent, and explanatory feedback. Their ability to mimic human evaluators by providing personalized explanations for grading decisions has been demonstrated in both experimental systems and emerging real-world applications [21–23]. This shift highlights a move from systems that merely output numerical scores to those that also offer rich qualitative feedback—a capability increasingly recognized as essential for educational impact.

However, the shift toward generating qualitative feedback presents new challenges. Unlike numerical scores, textual feedback is inherently unstructured, variable and more difficult to evaluate. The lack of standardized metrics for assessing its quality complicates the automated evaluation process, raising concerns about reliability and consistency across models and contexts.

This study aims to address these challenges by validating the effectiveness of existing NLP metrics in evaluating feedback quality rather than training a new scoring model. Specifically, it investigates: (1) the capability of LLMs to generate pedagogically valuable feedback, and (2) the alignment of widely used evaluation metrics with human judgments. By analyzing how well current metrics capture the nuances of feedback quality, this study seeks to establish a clearer foundation for developing reliable, standardized feedback systems that enhance the educational utility of AES systems.

III. Methodology

This section outlines the models, dataset, and experimental methodologies used in this study.

A. Models

This study focuses on fine-tuning and evaluating several LLMs: T5, BART, LLaMA, and DeepSeek. DeepSeek is included for its few-shot learning capabilities, which enable performance assessment with minimal retraining. These models were selected because they are freely available and require comparatively lower computational resources than larger commercial alternatives.

It is important to note that our choice of models is opportunistic rather than central to the research focus. The primary goal is to evaluate the effectiveness of various evaluation metrics, not to benchmark the LLMs themselves.

Section IV-C provides detailed explanations of the training process for each model.

B. Dataset

The primary dataset for this study is PeerRead, a large-scale peer-review corpus introduced by [24], which contains over 14,000 anonymized reviews of academic

manuscripts from top-tier venues including ACL, CONLL and ICLR. PeerRead includes full review texts, reviewer ratings across multiple dimensions (e.g., significance, clarity, originality), and editorial decisions, providing researchers with rich, human-authored qualitative feedback that extends beyond numerical scores. This makes PeerRead particularly well-suited for evaluating context-specific, textual feedback generation within AES systems. It serves as the main resource for feedback generation, contributing to a comprehensive evaluation framework.

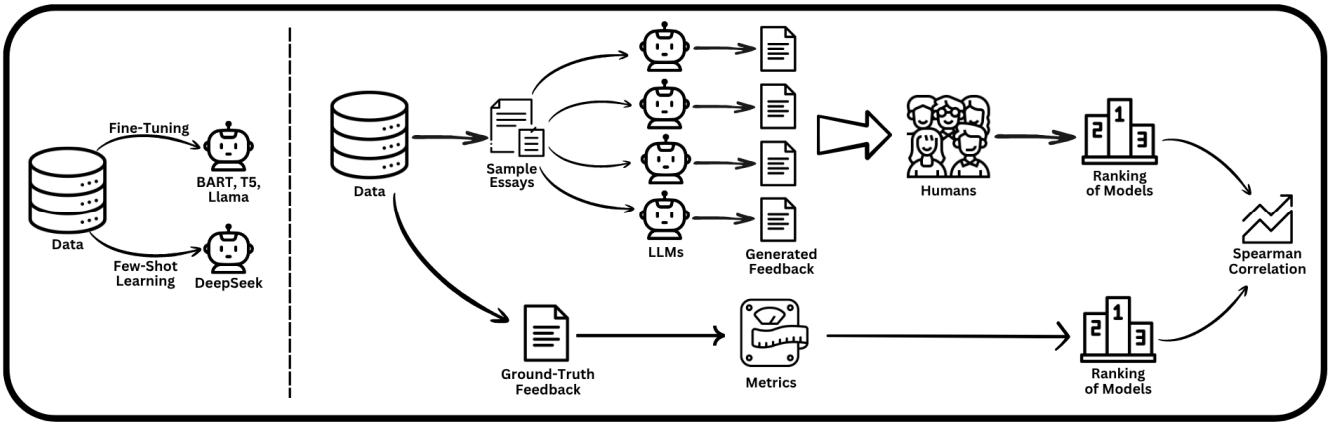

Fig. 1. Overall workflow of the study. *Left:* model preparation stage where BART, T5, and LLaMA are fine-tuned and DeepSeek is applied under a few-shot learning setting using the PeerRead dataset. *Right:* evaluation stage where each trained or prompted model generates qualitative feedback for selected essays, which is then assessed using both human evaluation and automated metrics. The results of these evaluations are compared via Spearman correlation to validate the efficacy of the metrics.

Fig. 1 provides an overview of the experimental workflow. As shown in the left panel, the PeerRead dataset forms the foundation of the model preparation process, supplying input–output pairs used for fine-tuning BART, T5, and LLaMA, and for few-shot prompting with DeepSeek.

For this study, the PeerRead dataset was analyzed to extract meaningful insights for pre-processing and model training, focusing on the ICLR 2017, CONLL 2016, and ACL 2017 sub-datasets, which include the necessary peer reviews for this work. The analysis involved identifying the distribution of paper and review lengths, as well as the proportion of accepted versus rejected papers.

Histograms were generated to visualize the distribution of both paper and review lengths. Word frequency analysis was also conducted to identify common vocabulary and domain-specific terminology. These insights were used to refine the tokenization process and prepare the dataset for model training. This helped in understanding the dataset's structure and complexity, and guided pre-processing strategies.

During the analysis, a class imbalance was observed—of the 586 papers, 172 were accepted and 414 were rejected. While this imbalance could affect certain tasks, such as classification, it wasn't a primary concern for this study, since the focus is on feedback generation rather than acceptance prediction. Consequently, the dataset was not artificially balanced by removing rejected papers, as this would have resulted in significant data loss.

The data pre-processing steps converted the raw dataset into a format suitable for fine-tuning the models. The dataset was processed to extract training input-output pairs consisting of the paper text (including title, abstract, and sections) and the associated reviews and meta-reviews. The following number of training pairs were extracted:

- ICLR 2017: 4,496 pairs
- CONLL 2016: 33 pairs
- ACL 2017: 248 pairs

In total, 4,777 training pairs were extracted from 586 papers. The dataset, already split into training, validation, and test sets, resulted in the following distribution:

- Training Set: 4,777 pairs
- Validation Set: 521 pairs
- Test Set: 500 pairs

All models used the same train/validation/test splits and basic preprocessing (e.g., whitespace normalization), but differed in tokenization and data formatting:

BART & T5: Both used their native tokenizers on input (paper) and output (review) texts, with dynamic padding to the longest sequence in each batch.

LLaMA: Employed its SentencePiece-based Byte-Pair Encoding (BPE) tokenizer and an instruction-tuning format of:

```
[INST] <paper> [\INST] <review>
```

DeepSeek: Applied a custom preprocessing pipeline that merged parsed essay content, metadata, and filtered review comments into unified text entries.

Finally, each model's tokenized data was saved in its required format, ensuring consistency across splits and compatibility with both few-shot and fine-tuning workflows.

C. Training

All models followed a common training pipeline that involved loading preprocessed, tokenized data with dynamic padding and consistent batching. However, they differ in their fine-tuning or prompting strategies. Importantly, the goal of these experiments was not to train a production-ready assessment system, but rather to generate controlled feedback outputs from different models so that the validity of evaluation metrics could be tested under comparable conditions.

As illustrated in Fig. 1–left, this stage encompasses the fine-tuning of BART, T5, and LLaMA models and the few-shot evaluation of DeepSeek. The resulting trained or prompted models serve as the basis for the evaluation process depicted in the right panel.

BART & T5: Both models were fine-tuned for three

epochs using AdamW optimization with linear warm-up, gradient clipping, and early stopping based on validation loss. The best checkpoints and corresponding tokenizers were saved to ensure reproducibility.

LLaMA: The 1B-parameter LLaMA-3.2 model, with a 128K-token context window was instruction-tuned to generate essay feedback. Essays exceeding 10K tokens were excluded for efficiency. Training employed DeepSpeed-ZeRO3 [25] across eight A6000 GPUs (batch size 1 with four-step gradient accumulation), AdamW (learning rate = 1e-5, weight decay = 0.01), a brief warm-up schedule, and FlashAttention [26] to optimize memory usage and speed. Training was limited to four epochs, sufficient to elicit stable outputs for metric validation rather than to maximize model performance.

DeepSeek (Few-Shot): Instead of fine-tuning, DeepSeek was evaluated using a prompt-based few-shot approach. Three randomly selected essay-feedback examples, along with role instructions, were prepended to each input prompt. These were sent via API calls at maximum token capacity, requiring no parameter updates or retraining. Sample prompts used in this approach are included in the supplementary materials (Section A) for reference.

D. Experiments and Evaluation Framework

Fig. 1–right outlines the evaluation process, where feedback generated by each model is compared against human and metric-based assessments.

The experimental design aimed to validate the efficacy of feedback evaluation metrics by comparing automated metric outputs against human judgments of feedback quality. The experiments were not intended to train or optimize AES systems for deployment, but rather to provide a controlled environment in which multiple LLMs could produce comparable feedback samples for analysis. We combined human assessments with automated metrics following methodologies from [27] to evaluate how well each model's feedback aligns with human judgment.

Human Evaluation: We selected 10 essays from the PeerRead dataset and generated four feedback outputs per essay, one from each model (40 samples in total). Five human evaluators participated in the study, and each essay was independently reviewed by three different evaluators. Evaluators ranked the four feedback outputs based on accuracy, coherence, relevance, and educational value. The majority ranking among evaluators was used to produce a gold-standard ranking for each essay. This approach ensured that the final rankings reflected a consensus of human judgment rather than individual bias.

Metric-Based Evaluation: In parallel, automated metrics (summarized in Table S1) were used to evaluate the same feedback samples. For each of the 10 essays, every metric computed a similarity score between each model's feedback and the corresponding ground-truth review, yielding one ranking of the four models per essay (10 rankings per metric).

Correlation Analysis: To determine how closely the automated metrics aligned with human judgments, we calculated Spearman's rank correlation coefficients [28]. For each essay, the metric-based rankings were compared with the gold standard human rankings. Coefficients close to 1 indicate strong agreement with human judgment, while values near 0 or –1 indicate weak or inverse alignment.

IV. Results

This section presents the study's findings from model training, feedback generation, and evaluation.

A. Training

First, the distributions of the paper and review lengths were examined to guide tokenization, padding, and maximum-length settings. Fig. 2 shows that papers cluster around 6,000 words, with the vast majority under 10,000, whereas reviews are much shorter, typically peaking between 50 and 200 words. These trends had two practical implications for the setup: (1) dynamic padding was applied to the inputs, and only exceptionally long essays were excluded during the instruction-tuning of long-context models (Section IV-C); and (2) conservative maximum generation lengths and decoding constraints were applied to the outputs to ensure concise reviews.

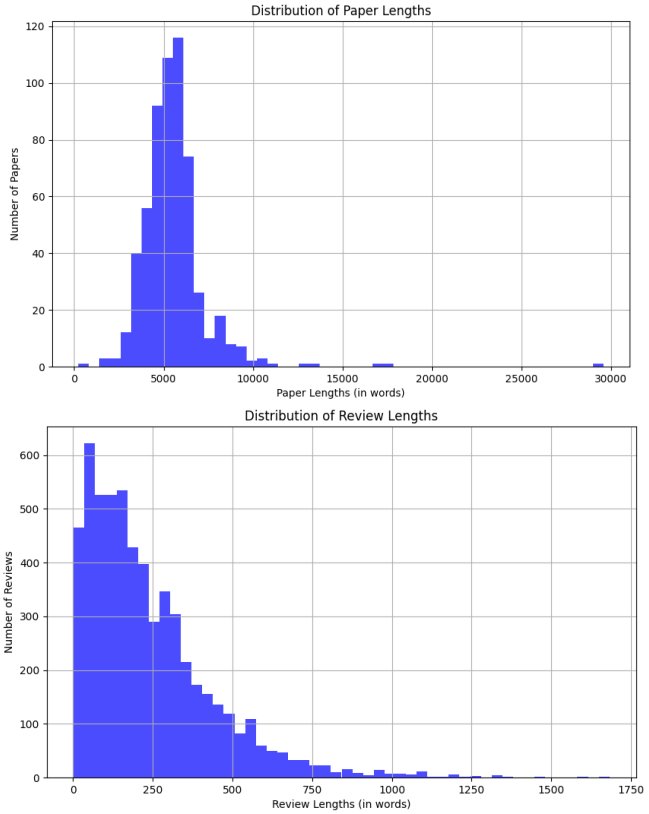

Fig. 2. Length distributions of source papers and reviews. Top: histogram of paper word counts across ICLR 2017, CONLL 2016, and ACL 2017. Bottom: histogram of review word counts. Papers are long with a long tail (most <10,000 words), while reviews are short (peak ~50–200 words). These distributions motivated our use of dynamic padding and selective filtering of extremely long inputs (Section 4.3) and informed conservative maximum generation lengths for review outputs.

During training, validation losses were monitored to identify the best configurations, defined as those with the lowest validation loss; an example loss curve is shown in the supplementary materials (Section C).

B. Generated Outputs

To qualitatively assess model behavior, Fig. 3 presents representative excerpts from the four models, while complete outputs are included in the supplementary materials (Section D).

The snippets in Fig. 3 reveal distinct failure modes across models. BART exhibits repetitive phrasing and semantic drift, often hallucinating unrelated text. T5 produces incoherent, circular statements lacking substantive critique. LLaMA maintains coherence but shows shallow reasoning and limited analytical depth. DeepSeek is generally factual and structured, yet occasionally overgeneralizes.

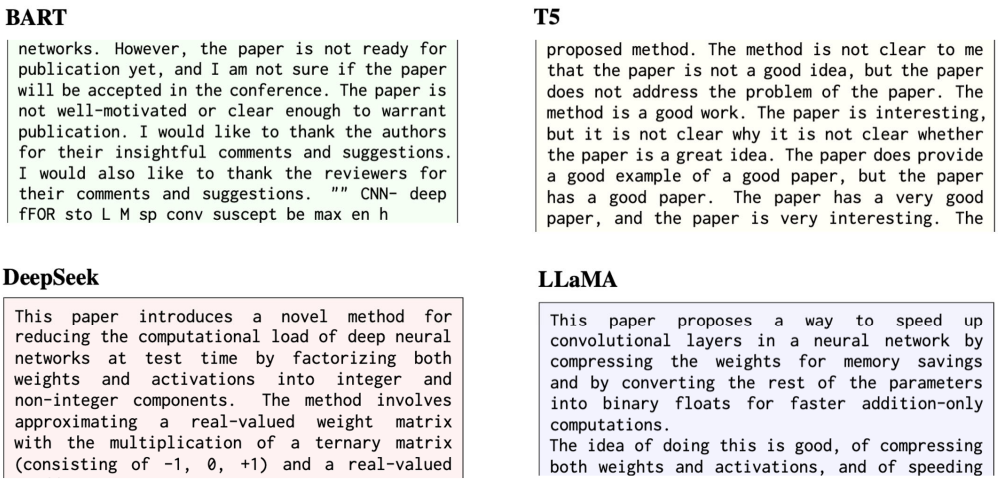

Fig. 3. Representative excerpts from model-generated feedback highlighting common qualitative issues. *BART:* repetitive phrasing and irrelevant acknowledgments. *T5:* circular and vacuous repetition showing loss of semantic control. *LLaMA:* coherent but analytically shallow reasoning. *DeepSeek*: structured and factual feedback, though mildly overgeneralized in scope.

These observations indicate five recurring error types: (1) repetition and verbosity, (2) circular phrasing, (3) hallucinated or fabricated content, (4) shallow reasoning, and (5) contextual overgeneralization. Even models with higher lexical similarity to human feedback often falter in factual grounding and pedagogical relevance, dimensions overlooked by surface-level metrics. Overall, the analysis underscores the need for evaluation metrics that capture coherence, factuality, and educational usefulness beyond lexical overlap.

C. Human Ranking Results

The human ranking results for the 10 essay samples are provided in the supplementary materials (Section E). Rankings for each essay were assigned by three independent evaluators, with the models numbered as follows: 1 = BART, 2 = LLaMA, 3 = T5, and 4 = DeepSeek. In the results table, a ranking such as 4-2-1-3 indicates that model 4 (DeepSeek) was rated the best, followed by LLaMA, BART, and T5. The final gold-standard ranking for each essay was determined based on the majority vote among the evaluators to ensure consensus and reduce individual bias.

Overall, the results show that DeepSeek (model 4) was ranked as the best model in most cases, indicating that it produced feedback perceived as more coherent, relevant, and educationally valuable by human evaluators. However, in two instances, LLaMA (model 2) received the top ranking, suggesting that it can occasionally generate feedback of comparable or superior quality. These findings highlight DeepSeek's overall consistency and reliability, while acknowledging LLaMA's potential strength in certain contexts.

Importantly, this analysis was not designed to determine a "best model" for deployment, but to create a human-judged benchmark against which the reliability of metrics could be assessed. Thus, the human rankings serve as the gold standard for evaluating how well different metrics align with human perceptions of feedback quality.

D. Metric-Based Ranking Results

Each metric generated a ranking of the models based on their similarity to the ground-truth human feedback (i.e., the original peer reviews in the dataset) for each essay. The full metric-based rankings for all 10 essays are presented in the supplementary materials (Section F).

These rankings provide an automated perspective on model performance, with each metric capturing distinct dimensions of text quality such as lexical overlap, semantic similarity, and structural alignment.

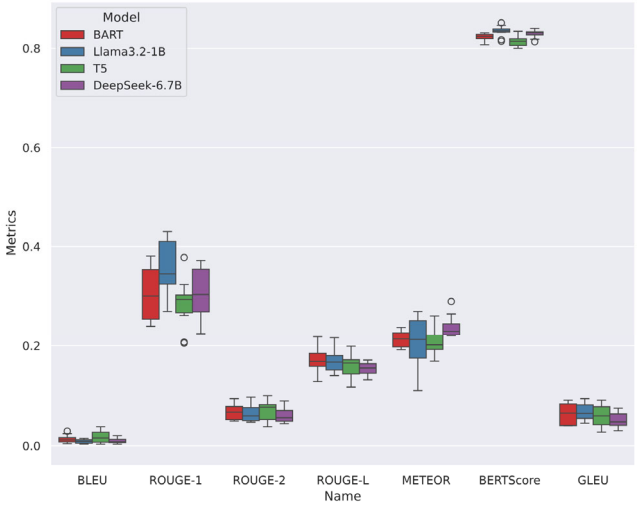

Fig. 4. Average metric performance across models. Each bar shows the mean similarity score (0–1 scale) of model-generated feedback compared to reference reviews across 10 test essays. Higher values indicate greater textual or semantic similarity. Among all metrics, BERTScore achieves the highest overall values, reflecting its stronger sensitivity to contextual semantics, while N-gram–based metrics such as BLEU and ROUGE yield lower scores, suggesting limited surface-level overlap between generated and human feedback.

Fig. 4 presents the average of metric scores across the 10 essays. Metric values range from 0 to 1, where 0 indicates no similarity with the reference feedback and 1 indicates perfect equivalence. The relatively low average scores across most

metrics suggest that the generated feedback differs substantially from human-written feedback in surface form.

BERTScore stands out with an average of 0.825, significantly higher than the others, suggesting that it better captures meaning-level similarity despite limited lexical overlap. This result supports prior findings that embedding-based metrics align more closely with human perception of quality than purely token-based ones.

In addition to the figure, Table 1 reports the average scores for each metric across the four models (BART, T5, LLaMA, and DeepSeek). This comparison provides further insight into each model's feedback characteristics. For example, LLaMA achieves the highest ROUGE-1 score, reflecting stronger lexical overlap, while DeepSeek demonstrates competitive performance across multiple metrics.

Table 1. Average scores for different metrics across different models

Metric	BART	T5	Llama3.2-1B	DeepSeek-6.7B
ROUGE-1	0.305±0.054	0.285±0.049	0.358±0.052	0.304±0.052
ROUGE-2	0.067±0.015	0.069±0.020	0.064±0.016	0.061±0.014
ROUGE-L	0.170±0.024	0.159±0.025	0.169±0.022	0.153±0.012
METEOR	0.213±0.016	0.207±0.027	0.207±0.051	0.239±0.022
BERTScore	0.823±0.007	0.814±0.007	0.834±0.011	0.829±0.008
GLEU	0.064±0.022	0.060±0.021	0.068±0.016	0.051±0.014

However, since the goal of this study is to assess the validity of these metrics rather than the models themselves, these differences primarily serve to test how sensitively and consistently each metric reflects the variations identified by human evaluators.

E. Spearman Correlation

To assess the alignment between metric-based rankings and human gold-standard rankings, the Spearman rank correlation coefficient was calculated for each metric and essay. The detailed rankings used for this analysis are presented in Tables S2 and S3. The resulting correlation values for each metric, along with their averages across the 10 essays, are summarized in Table 2.

Table 2. Spearman correlation values between different metrics and human rankings

Essay	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore	GLEU
1	0.20	0.40	0.40	0.20	0.80	0.80	0.40
2	0.40	−0.80	−0.20	−0.80	0.40	0.40	−1.00
3	−0.20	0.00	−0.40	0.00	0.60	0.80	0.00
4	0.40	−0.20	0.20	−0.20	1.00	0.60	0.40
5	−0.60	−0.20	0.00	−1.00	0.00	0.40	−0.80
6	−0.40	0.80	−1.00	−0.80	−0.80	1.00	0.20
7	−0.40	−0.40	−0.80	−0.80	0.20	−0.40	−1.00
8	−1.00	0.40	−0.80	−0.80	0.40	0.80	−0.20
9	−0.60	0.40	−0.60	−0.60	0.40	0.40	0.00
10	1.00	0.80	0.80	0.80	0.80	1.00	1.00
Avg	−0.12±0.57	0.12±0.49	−0.24±0.56	−0.40±0.55	0.38±0.49	**0.58**±0.39	−0.10±0.63

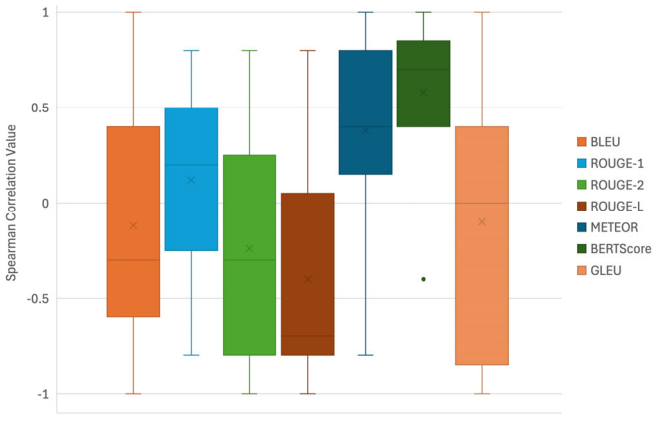

Fig. 5. Alignment of automated metrics with human judgment. Box plots of Spearman correlation coefficients between metric-based and human-derived model rankings across 10 essays. Each box shows the spread and average of correlations for a given metric. BERTScore exhibits the highest and most consistent correlation with human rankings (average ≈ 0.58), indicating its superior ability to capture contextual and semantic similarity. By contrast, surface-level metrics show lower averages and greater variance, revealing weaker alignment with human preferences.

Fig. 5 illustrates the distribution of Spearman correlation coefficients for each evaluation metric, showing how closely automated metric rankings align with human gold-standard judgments. This visualization highlights both the variability and reliability of each metric across the 10 evaluated essays.

The results show that BERTScore achieved the highest average correlation with human rankings, with a Spearman coefficient of approximately 58%. This indicates that BERTScore, by leveraging contextual embeddings, is better suited to capture semantic and contextual nuances than surface-level metrics such as BLEU and ROUGE.

Other metrics demonstrated notably weaker correlations, suggesting that they struggle to reflect human assessments of feedback quality. In particular, N-gram–based metrics showed greater variance and lower average alignment, emphasizing their limited ability to evaluate qualitative or meaning-driven aspects of text.

The findings reinforce that metrics relying solely on lexical similarity are insufficient for assessing nuanced, pedagogically meaningful feedback. Consequently, these results underscore the need for new evaluation frameworks that integrate semantic understanding and educational

relevance, providing a more accurate reflection of human judgment in automated feedback systems.

V. Discussion

This study examined the potential of LLMs to generate qualitative feedback within AES systems, shifting the focus from traditional grade prediction to feedback evaluation. While the findings underscore the promise of LLMs in enhancing educational applications, they also reveal fundamental limitations that must be addressed for these systems to become reliable and pedagogically meaningful.

The experimental results demonstrate notable variability in LLM performance. DeepSeek consistently produced detailed and coherent feedback, showing strong capability in handling the nuanced task of qualitative feedback generation. LLaMA also performed well in several cases, particularly when instruction-tuned for the task. In contrast, T5 and BART, despite their general text generation strengths, often produced repetitive or semantically weak feedback, reflecting limitations in capturing the contextual and instructional depth required for effective educational commentary. These observations highlight that while some LLMs show potential, reliable feedback generation remains an open challenge that requires careful task-specific adaptation.

Beyond model performance, the findings reveal a more critical issue concerning evaluation metrics. Traditional metrics such as BLEU and ROUGE showed poor alignment with human evaluations, as they emphasize surface-level lexical overlap rather than semantic or pedagogical quality. Even embedding-based metrics like BERTScore, although showing higher correlations, only moderately reflected human judgments. This persistent misalignment indicates that current metrics are insufficient for assessing feedback quality and offer limited guidance for improving model outputs. Consequently, the study reaffirms that evaluating LLM-generated feedback requires more sophisticated frameworks that account for semantic coherence, contextual relevance, and educational effectiveness rather than textual similarity alone.

A further challenge identified in this research is the scarcity of suitable datasets. The PeerRead corpus served as the primary dataset due to its availability of human-authored reviews, but locating alternative resources for qualitative feedback evaluation proved exceptionally difficult. The absence of diverse, domain-specific, and well-annotated datasets restricts both the generalizability of experiments and the development of reliable evaluation frameworks. Without such datasets, progress in generating and validating high-quality feedback remains constrained.

Despite these limitations, the study provides valuable insights into the current capabilities of LLMs and the deficiencies of existing evaluation practices in AES research. The results suggest several key directions for future work. First, addressing the dataset gap by creating new corpora specifically designed for qualitative feedback generation should be prioritized. Second, developing metrics that align more closely with human judgments—by integrating dimensions such as semantic fidelity, pedagogical usefulness, and human preference modeling—will be essential to advance the field. Ultimately, the findings highlight that while LLMs hold significant promise for enhancing AES systems, progress depends on the development and validation of more robust evaluation metrics capable of accurately reflecting the educational value of generated feedback.

VI. Conclusion

This study examined the use of LLMs for generating qualitative, context-aware feedback in AES and evaluated the reliability of common automatic metrics against human judgments. While models such as DeepSeek and instruction-tuned LLaMA showed promise for producing coherent feedback, the broader findings reveal a fundamental mismatch between current automatic metrics and the qualities that humans value in feedback. This misalignment limits the utility of metric-based evaluation for guiding model improvement.

A central limitation identified in this research is the inadequacy of existing evaluation metrics for assessing qualitative feedback. Although widely used in NLP tasks, metrics such as BLEU, ROUGE, and similar N-gram–based measures primarily capture surface-level properties like lexical overlap and word order. Consequently, they fail to account for deeper semantic, contextual, and pedagogical aspects that define meaningful feedback. The correlation analysis with human rankings confirms that these traditional metrics often diverge from human judgment, underscoring their limited validity in educational contexts.

Even the best-performing metric in this study, BERTScore, achieved only moderate agreement with human evaluations. This finding further underscores the need for novel evaluation metrics that prioritize semantic understanding, discourse coherence, and instructional usefulness over syntactic similarity. Future metrics should integrate semantic similarity measures like embedding-based representations with pedagogical relevance and coherence-scoring components, potentially enhanced by human preference modeling to approximate how instructors assess feedback quality. Until such metrics are developed, AES research risks over-reliance on evaluation criteria that do not accurately reflect educational value.

The study also faced constraints related to dataset availability. We relied exclusively on the PeerRead dataset, which, although valuable for its rich feedback annotations, primarily consists of expert academic reviews. Its linguistic style and discourse structure differ substantially from student writing, limiting the generalizability of our findings to broader educational contexts. Future work should therefore test models on more diverse and representative datasets, particularly those reflecting student-level writing and real-world classroom interactions.

Finally, while multiple LLMs were examined, their evaluation served only to create diverse output samples for metric validation, not to optimize or benchmark specific models. As such, model-related findings should be interpreted in light of this primary goal.

Overall, the field would benefit from (1) developing semantically informed evaluation frameworks that align more closely with human judgment and educational priorities and (2) curating richer, feedback-oriented datasets that enable robust assessment of feedback generation quality. Progress in these directions is likely to yield greater

educational impact than refining models under existing, surface-level metrics that inadequately reflect the true instructional value of feedback.

Ethical Considerations

This work raises several ethical considerations related to data usage, model biases, environmental sustainability, and educational impact. All data came from publicly available, open-access sources with no personal or sensitive information, and human evaluations were conducted solely by the authors as part of their academic work, avoiding concerns about external annotator exploitation. However, like all LLMs, the models used may still exhibit biases from their training data, affecting fairness and feedback quality, particularly for diverse linguistic and cultural groups, highlighting a need for future bias mitigation. To reduce environmental impact and energy consumption, the study prioritized metric validation over model optimization to avoid extensive hyperparameter searches, limiting energy-intensive training. Finally, the authors caution that NLP tools in education can be misused, such as automating both essay writing and grading, risking devalued learning and underscoring the need for responsible, socially aware deployment of such systems.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

M.B. developed the research idea, implemented the experiments, analyzed the data, and wrote the first draft of the manuscript. C.S. contributed to developing the original idea and provided feedback. D.C. supervised and guided the research and critically reviewed the manuscript. All authors approved the final version.

Acknowledgment

The authors gratefully acknowledge Soumyadeep Pal, Radhika Shenoy, Abigael Mogusu, and Zhiying Li for their contributions, Dr. Kristen Johnson for guidance in the learning process informing this research, and Dr. John Zubek for his helpful advice and insightful discussions.

References

[1] E. B. Page, "The imminence of... grading essays by computer," *The Phi Delta Kappan*, vol. 47, no. 5, pp. 238–243, 1966.

[2] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: A systematic literature review," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2495–2527, 2022.

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[4] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, pp. 74–81, 2004.

[5] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," presented at International Conference on Learning Representations, 2020.

[6] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. the ACL Workshop on Intrinsic and Extrinsic Evaluation Mseasures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[7] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018.

[8] C. Raffel, N. M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2019.

[9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. R. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *Annual Meeting of the Association for Computational Linguistics*, 2019.

[10] H. Touvron, T. Lavril, G. Izacard *et al*., "LLaMA: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, February 2023.

[11] D. Guo, Q. Zhu *et al*., "Deepseek-coder: When the large language model meets programming—the rise of code intelligence," arXiv preprint arXiv:2401.14196, 2024.

[12] P. W. Foltz, D. Laham, and T. K. Landauer, "The intelligent essay assessor: Applications to educational technology," *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1999.

[13] L. Kong *et al*., "Automated essay scoring via pairwise contrastive regression," in *Proc. the 29th International Conference on Computational Linguistics*, 2022, pp. 2724–2733.

[14] Kaggle, *The Hewlett Foundation: Automated Essay Scoring*, 2012.

[15] S. Mathias and P. Bhattacharyya, "ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores," in *Proc. the Eleventh International Conference on Language Resources and Evaluation (LREC)*, May 2018.

[16] Z. Ke and V. Ng, "Automated essay scoring: A survey of the state of the art," *International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 19, pp. 6300–6308, 2019.

[17] P. U. Rodríguez, A. Jafari, and C. M. Ormerod, "Language models and automated essay scoring," arXiv preprint arXiv:1909.09482, 2019.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *North American Chapter of the Association for Computational Linguistics*, 2019.

[20] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," *Neural Information Processing Systems*, 2019.

[21] A. Pack, A. Barrett, and J. Escalante, "Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability," *Computers and Education: Artificial Intelligence*, vol. 6, 2024

[22] S. X. Xu *et al*., "Human-AI collaborative essay scoring: A dual-process framework with LLMs," arXiv preprint arXiv:2401.06431, 2024.

[23] G. A. Katuka, A. Gain, and Y.-Y. Yu, "Investigating automatic scoring and feedback using large language models," arXiv preprint arXiv:2405.00602, abs/2405.00602, 2024.

[24] D. Kang, W. Ammar, B. Dalvi, M. Zuylen, S. Kohlmeier, E. H. Hovy, and R. Schwartz, "A dataset of peer reviews (PeerRead): Collection, insights and NLP applications," in *Proc. the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 1647–1661.

[25] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He, "Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters," in *Proc. the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 3505–3506.

[26] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré, "Flashattention: Fast and memory-efficient exact attention with io-awareness," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16344–16359, 2022.

[27] M. Berijanian, S. Dork *et al*., "Soft measures for extracting causal collective intelligence," in *Proc. the 1st Workshop on NLP for Science (NLP4Science)*, Miami, FL, USA. Association for Computational Linguistics, 2024, pp. 99–116.

[28] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.