Optimizing Automated Essay Scoring: Balancing Accuracy and Cost at Scale

1st Raghav Manoj Gaur
Data Science Lab
Toronto Metropolitan University
Toronto, Canada
raghav.gaur@torontomu.ca

2nd Mucahit Cevik
Data Science Lab
Toronto Metropolitan University
Toronto, Canada
mcevik@torontomu.ca

3rd Sojin Lee
Co-founder & CEO
Blees Technologies Inc. (Blees AI)
& Olive AI Limited (Olive AI)
Toronto, Canada
ms.sojinlee@gmail.com

***Abstract*—This study explores the effectiveness of Large Language Models (LLMs) for automated essay scoring of student essays in the finance domain. The focus is on generating grades and explanations for six Assessment Indicators (AIs) related to finance and accounting, and providing feedback in places where improvement is needed for each student essay. Our research highlights the capabilities of LLMs and showcases the effectiveness of custom prompt engineering in domain-specific automated essay scoring. We propose an advanced Retrieval-Augmented Generation (RAG) method that can retrieve relevant text from student essays for evaluation in a cost-effective manner. We perform a comparative analysis with several open-source and commercial LLMs and assess their performance and the associated costs for our essay scoring task. Our analysis also involves investigation of prompt engineering techniques and effective prompting structures.**

. *Index Terms*—Large Language Models (LLMs), Automated Essay Grading, In-Context Learning, Generative AI, Natural Language Processing (NLP)

I. Introduction

With the rise of Large Language Models (LLMs) in tasks involving text analysis, detection, and classification, there has been an increase in the application of LLMs in education, both for generating educational content and for evaluating and grading students' work. One of the most common forms of educational content evaluated using LLMs today is student essays. The terms Automated Essay Scoring (AES) and Automated Essay Grading are often used synonymously, both referring to the use of computational methods to evaluate written content [1]. These techniques are used for a wide variety of academic tasks ranging from standardized test evaluation to long-form responses and school-level essays. The scoring criteria differ depending on the context; some emphasize grammatical precision, while others focus on the way core subject concepts are articulated [2].

In this study, we focus on assessing student essays that are domain-specific and are evaluated based on conceptual understanding in finance and accounting. Unlike traditional essay scoring systems, our approach leverages Large Language Models (LLMs) to identify whether a student has addressed six predefined Assessment Indicators (AIs) relevant to key financial topics. To evaluate the essays using LLMs, we employ prompts that classify each assessment indicator as either correctly addressed ('Y') or not ('N'), and then derive a final grade based on this breakdown.

Previous AES research has primarily aimed to assign a single holistic score to an essay, evaluating features such as syntax, cohesion, and narrative flow [3, 4]. These methods are not very effective when domain relevance is crucial, such as in the fields of accounting and finance [2]. Our work diverges by targeting specific conceptual markers, including financial reporting, inventory, and performance metrics. While Helmeczi et al. [5] also explored AES in a finance context, their sentence-level classification approach contrasts with our document-level, concept-based grading method.Similarly research by Garima Malik [23] also built an AES system by generating grades for different Assessment Indicators along with custom prompt engineering for finance related essays but it didn't explore the methods for cost reduction and performance optimization through Retrieval-Augmented Generation (RAG), the use of open-source LLMs, and prompt engineering with feedback generation

Historically, automated grading has been either looked at as a regression or classification problem [6], often implemented using traditional machine learning or transformer-based models [5, 7]. Recent advancements in NLP, particularly with LLMs, enable deeper semantic evaluation across specialized domains. This work demonstrates how LLMs can be adapted for context-sensitive essay evaluation, shifting from surface-level scoring to in-depth conceptual assessment [7].

Our AES approach uses a Retrieval-Augmented Generation (RAG) framework to extract only the most relevant essay content, improving accuracy and reducing token-related computation. We compare cutting-edge open-source models like LLaMA 3.3 70B with commercial LLMs such as GPT-4o and Gemini 2.5 Pro, given the formers suitability for in-house use and data privacy. We also evaluate the quality of LLM-generated feedback against human responses to assess alignment in depth and accuracy.Based on our review of the literature, prior studies have not explored the use of Retrieval-Augmented Generation (RAG) in conjunction with both open-source and commercial LLMs for domain-specific concept-level AES tasks involving multiple assessment indicators along with feedback and explanation generation particularly in the context of finance and accounting education.

II. Literature Review

The understanding of the knowledge gained through academic learning by a student can be easily tested by looking at their capacity of communicating their subject-matter understanding through written responses. As Hyland [14] notes, writing is not just a medium of expression but also a reflection of intellectual engagement within a discipline. Despite its importance, manual grading still remains a very time-intensive task which is prone to subjectivity and inconsistency as different evaluators have a slightly different definition of the correct answer. AES tools have been proposed to mitigate these challenges by streamlining the grading process and improving scoring reliability [1].

Earlier AES systems were applying supervised learning techniques to grade student essays by transforming them into numerical features and modelled them against manually assigned scores [16]. With the rise of deep learning, more sophisticated architectures have emerged. These include CNNs [15], LSTMs [17], and, more recently, pre-trained transformer models [18]. BERT model out of all the architectures has shown superior performance in AES by demonstrating its strong capabilities in mapping text to accurate grade predictions.

Increasingly, researchers are exploring how large, open-source LLMs can be adapted for educational tasks, including grading. Most of the experiments in different studies are done using ASAP dataset, which contains nearly 13,000 essays graded on an 8-point scale and the evaluation is performed using the Quadratic Weighted Kappa (QWK) metric. While ASAP remains a central benchmark, AES research has expanded to essays in multiple languages such as Chinese [7], Japanese [18], and Turkish [6], addressing multilingual contexts and diverse scoring rubrics.

In more recent studies, researchers have discovered that using LLMs to first extract only the relevant text from essays and then prompting them provides better results [2]. With the exponential increase in the number of LLMs being released there is also a rise in a number of highly effective open-source LLMs which can provide great results depending on the task [9]. LLMs are not only being proven to be highly effective in grading student essays but they also provide feedback to the students which can help them improve and learn [8].

Recent advancements have also shifted from scalar scoring to pairwise comparisons, where LLMs rank essay quality and models like RankNet convert these preferences into continuous scores, as seen in the LCES framework [20]. Other studies propose hybrid pipelines that combine ranking and scoring stages. For example, the Rank-Then-Score (RTS) framework fine-tunes LLMs by first producing ranked outputs and then assigning scores, yielding strong results across English and Chinese datasets [21]. Additionally, TRATES introduces a trait-specific, rubric-based AES framework that leverages LLM-generated features to assess specific dimensions of student writing, achieving state-of-the-art performance in trait-level evaluation [22].

Emerging research have also explored few-shot learning techniques to address the high data requirements of AES. Approaches like PET and SetFit have shown promise in improving performance when labeled data is scarce, with PET offering strong results despite its computational cost [5]. Additionally, domain-specific efforts, such as in the finance sector, have applied custom prompting strategies and in-context learning to score multiple assessment indicators, demonstrating the adaptability and generalizability of LLMs for specialized AES tasks [23].One particular study by Yoshida [24] explores whether detailed rubrics are essential for automated essay scoring using large language models. Through experiments on the TOEFL11 dataset, the study finds that simplified rubrics can achieve comparable scoring accuracy to full rubrics while reducing token usage.Furthermore, a new research leverages multimodal large language models to assess lexical-, sentence-, and discourse-level traits, addressing key limitations of traditional AES. By utilizing trait-specific scoring and multimodal context understanding, it enables more precise and context-rich evaluations without manual feature engineering [25]. Do et al. [26] further advance trait-level AES by introducing RaDME, a framework that prompts LLMs to generate both scores and explanatory rationales. This self-explainable approach enhances transparency and interpretability in multi-trait evaluation. It reflects a growing focus on aligning automated scoring with human reasoning.

In this study, we developed a concept-driven AES framework specifically tailored for finance and accounting case studies.Our approach evaluates student essays based on six domain-specific Assessment Indicators (AIs), each representing a critical financial concept. These include key topics like financial reporting, inventory assessment, and performance metrics. Instead of assigning an overall grade, we determine whether each AI is correctly addressed (Y) or not (N), enabling a granular and concept-sensitive evaluation of the student's understanding. To further enhance the assessment process, we adopt a Retrieval-Augmented Generation (RAG) strategy that isolates only the most relevant sections of text, improving both evaluation accuracy and computational efficiency by limiting the number of tokens sent to the model. We also explore whether advanced open-source models like LLaMA 3.3 70B can achieve grading performance comparable to commercial models such as GPT-4o and Gemini 2.5 Pro, with the added advantages of in-house deployment and stronger data privacy. Finally, our framework investigates the ability of LLMs to generate explanatory feedback and compares it against human-generated responses to assess alignment in quality and depth.

III. Methodology

This section provides the methodology details, including the essay dataset representation, LLMs considered, RAG mechanism, and evaluation metrics.

A. Essay Dataset Representation

We use a proprietary dataset of finance-related essays written by students. Each essay is identified with a unique name and stored as a dictionary, where:

- Key: Essay name (e.g., "essay_001")
- Value: Essay text (string)

Each essay is manually labeled with a binary mark: Y (meets criteria) or N (does not meet criteria). We do not include the essay text in this paper due to confidentiality, but synthetic examples are provided for illustration. The synthetic essay examples provided below reflect the type of responses found in the actual data. Each essay represents a different student's perspective on a common topic related to banking regulations.

```
{
   "essay_001": "Basel III was introduced
       after the 2008 financial crisis to
       prevent future banking collapses. It
       increased the required capital
       reserves for banks and introduced
       liquidity standards like the LCR and
       NSFR. These changes aimed to reduce
       systemic risk while maintaining
       lending capacity.",

   "essay_002": "The Basel III regulations
       helped fix banks after the financial
       crisis. Banks had to keep more money
       in reserve, which made them safer.
       They also had to follow new rules
       about lending. This helped make the
       system more stable.",

   "essay_003": "Basel III is a financial
       rule for banks. It says banks need to
       save money. This is good so they dont
       run out. Banks didnt like this at
       first but now they do it."

}
```

Ground Truth Labels: The table below shows synthetic ground truth labels corresponding to each essay across three evaluation criteria.

TABLE I: Ground Truth Labels for Synthetic Essays

Essay ID	Criteria 1	Criteria 2	Criteria 3
essay_001	Y	Y	N
essay_002	Y	N	N
essay_003	N	N	N

B. Retrieval-Augmented Generation (RAG)

RAG is a technique where a model's input is enriched by fetching relevant context passages before generating an output. This is especially helpful when the input is lengthy and only a subset is needed for a particular task (e.g., answering a question or scoring a criterion). We designed a two-stage RAG method, consisting of Lexical RAG and LLM RAG, that is both effective and cost efficient, as illustrated in Figure 1.

1. **Lexical RAG:** As the name suggests we would be using keywords to extract the relevant text from our essay. We start off by first dividing our essay into 4 parts which can be called as chunks. We tested different chunk sizes like 3, 4, 5, 6 and found that 4 chunks hit the perfect balance of accuracy of not missing the relevant text and removing as much non-relevant text as possible, leading to potential cost benefits. We take a set of keywords which convey the idea of what we are assessing on and which should definitely occur in the students answer. It is important to note that we should not select keywords that are very generic as that might extract more chunks than needed for assessment.Then we select the chunks which contain these keywords so in our case mostly it was chunk 3 or 3 and 4 or 2 and 3. Once these chunks are selected we pass these chunks forward to the LLM RAG.
2. **LLM RAG:** In this step we utilize an LLM (llama 3.3 70B for cost effectiveness). We collect the reference material which is taught by teachers and it can be considered as the right answer and provide it to the LLM along with the relevant chunks using the following prompt structure:

   Extract and output only the block of text from the student essay which is similar to the Reference material.
   Reference material:
   . . .
   Student Essay:
   . . .
   Block of Text:

   This helps improve precision and narrows the models focus, enhancing grading accuracy and minimizing irrelevant content.

Token Efficiency and Cost Impact: To demonstrate the impact of our hybrid retrieval approach on token usage and cost, we create a hypothetical cost analysis. Let's assume the following for cost calculation purposes:

- Number of tokens per essay = 2000
- Number of prompts = 9
- Number of student essays = 90
- Number of tokens in reference material = 500
- Number of tokens after Lexical RAG = 1500

Without RAG, we have $2000 \times 9 \times 90 = 1,620,000$ tokens, and with Lexical + LLM RAG, we have $(2000 \times 90) + (500 \times 9 \times 90) = 585{,}000$ tokens. Hence, token savings is $\frac{1{,}620{,}000-585{,}000}{1{,}620{,}000} \approx 63.8\%$, which is a significant reduction in token usage and cost.

C. Large Language Models

LLMs are rapidly revolutionizing the space of education, business and technology by providing machines the power to understand and generate human-like language at scale. From customer service automation to academic grading and content creation, LLMs are at the core of AI-driven innovation. Their ability to process text data and provide coherent outputs which are context aware makes them a very powerful tool in the context of automating complex cognitive tasks. In what follows, we briefly discuss the open-source and commercial LLMs considered in our analysis.

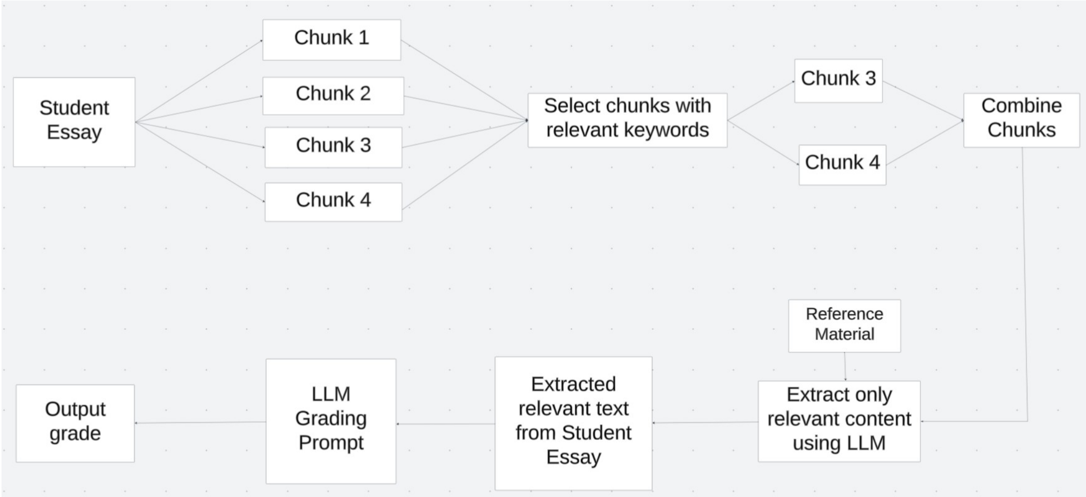
Student Essay
Chunk 1
Chunk 2
Chunk 3
Chunk 4
Select chunks with relevant keywords
Chunk 3
Chunk 4
Combine Chunks
Reference Material
Extract only relevant content using LLM
Extracted relevant text from Student Essay
LLM Grading Prompt
Output grade

Fig. 1: Flow chart explaining the overall RAG process

1) Open-Source LLMs: Open-source LLMs act as a very powerful alternative to commercial LLM solutions like gpt-4o, gemini-2.5 pro, claude etc by providing greater control, transparency, and cost-effectiveness. Organizations can deploy these models on custom infrastructure (e.g., Groq, vLLM) to achieve low-latency, scalable performance while maintaining data privacy, an essential feature for educational and healthcare applications. We consider following open-source LLMs:

- **LLaMA 3.3 70B-Instruct:** LLama 3.3 70B [13] is a very powerful LLM which supports a context window length of size 8k tokens and has been trained over 15T tokens. Llama 3.3 70B offers near-commercial performance at a fraction of the operation cost when it is deployed efficiently.This model has been fine-tuned specifically for improved reasoning and multilingual capabilities.
- **Qwen 2.5 72B:** Qwen 2.5 72B [12] is known for its strong multilingual support and alignment with coding tasks. Qwen 2.5 supports up to 128K context tokens, making it especially useful in document-heavy scenarios. It is trained on a blend of curated open-domain and code datasets, showing strong performance across a wide range of tasks.

2) Commercial LLMs: These are typically more refined and optimized through proprietary training pipelines, often incorporating reinforcement learning from human feedback (RLHF), extensive alignment tuning, and real-world deployment experience. Some advantages of commercial LLMs are having higher overall performance and reliability across a wide range of tasks, providing access to proprietary data and tooling (e.g., integrations, plugins, APIs), and scalability, thanks to being deployed on powerful cloud infrastructure optimized for inference at global scale. We consider following commercial LLMs:

- **GPT-4o (OpenAI):** GPT-4o [10] was launched by OpenAI in May 2024, represents the latest advancement in the GPT-4 series. It is a model that has capabilities of processing input of different modalities such as vision, text, and audio inputs, with faster response times and significantly reduced costs compared to GPT-4-turbo. GPT-4o provides supports 128K token context, improved reasoning ability, and native multilingual fluency. It offers a solution that is balanced for enterprises needing top-tier model performance with wide-ranging capabilities.
- **Gemini 2.5 Pro (Google):** Gemini 2.5 pro [11] was released by Google DeepMind in early 2024, is a powerful multimodal model that excels in complex reasoning, image interpretation, and long-context handling. With a context window of 1 million tokens in some configurations, Gemini is designed for enterprise-scale applications. Its integration into Google's ecosystem and support for advanced tools make it a strong choice for developers looking for extensibility and deep tool integration.

3) Prompting Strategy for Grading: To make sure that our grading is consistent and interpretable, we have adopted a structured prompting framework across LLMs for our comparative study. The essay of every student is evaluated on multiple rubric-based assessment criteria using a clearly defined prompt structure. This setup helps the model to not only provide a binary grade (Mark: Y/N) but also it helps the model to generate meaningful explanations and, when necessary, improvement-oriented feedback.

- **Context:** Role instruction to establish that the LLM is acting as a grader.
- **Assessment Criterion:**A specific grading question aligned with the rubric.
- **Few-shot Examples:** One example each for Mark: Y and Mark: N, including detailed explanations.
- **Feedback Reference:** Guidance on what a good answer should containonly invoked if the model marks N.
- **Student Essay:** The full essay text to be evaluated.
- **Final Instruction:** A directive to provide the mark (Y/N)

and explanation.

Prompt Template:

> You are a grader tasked with evaluating student essays based on specific questions about [topic, e.g., revenue recognition]. Only evaluate based on the content of the essay and *always* provide explanation for the evaluation.
>
> Assessment Criterion: `[Insert Criterion]`
>
> Examples:
> Example 1: Essay: `...` Mark: Y
> Example 2: Essay: `...` Mark: N
>
> If the answer is Mark N, also provide feedback on how to get Mark Y for the student essay under a feedback section using the following as reference: `[Insert Feedback Guidelines]`
>
> Student Essay: `[Insert Essay]`
>
> Mark:

The above template will consecutively fill the actual data of the student essay and send it to the LLM using the python-based pipeline. The same prompt structure was maintained across all the models to make sure we do a fair and standardized evaluation, allowing for reliable performance comparison and feedback quality assessment.

D. Strategies for Improving Prompts

A high-quality LLM system should not only assign accurate grades but also generate explanations and improvement feedback that reflect true rubric alignment. In this section, we explore how LLM-generated feedback and explanations can be used to both interpret and improve prompt design.

1) Feedback-Guided Prompt Engineering: To improve prompt effectiveness and ensure alignment with grading intent, we adopted a feedback-driven prompt engineering approach. By carefully analyzing the explanations and feedback generated by LLMs, we could assess whether the model was correctly interpreting the assessment criterion and applying it accurately to student essays. In cases where the model's rationale or feedback appeared misaligned with rubric expectations, this served as a signal that the prompt needed adjustment. For example, vague or overly general prompts often resulted in surface-level evaluations or incorrect grading logic. Through iterative prompt refinement, we clarified language, tightened rubric questions, and adjusted examples to better guide the models attention. Below, we provide an illustrative example on minor language change leading to correct grading.

Prompt Version A (Original Vague Wording)

Student Response:

"Assets go down in value because of market changes or wear and tear, and this is called depreciation."

Prompt Guidance:

- Mark Y: If the student mentions depreciation as a reduction in asset value over time.
- Mark N: If the student misunderstands depreciation or confuses it with market price changes.

LLM Output:

> *Explanation:* The student mentioned depreciation as a reduction in asset value due to market changes or wear and tear.
> *Mark: Y → Incorrect prediction*

We note that the model incorrectly accepts this as the prompt did not clearly distinguish between depreciation and market-driven revaluation.

Prompt Version B (More Precise Wording)

Prompt Guidance:

- Mark Y: If the student describes depreciation as the allocation of an assets cost over its useful life solely due to wear and usage.
- Mark N: If the student equates depreciation with fluctuating market prices or fails to mention cost allocation.

LLM Output:

> *Explanation:* "The student mistakenly associated depreciation not only with cost allocation over an asset's useful life, but also with changes in market value."
> *Mark: N → Correct prediction*

With just a minor rephrasing to emphasize cost allocation, the model correctly identifies the error in the students explanation. The example above shows how making *minor changes to the wording of prompts*, such as adding a clarifying clause or reframing a sentence, can significantly affect the model's interpretation. Further, it highlights that LLMs are highly sensitive to linguistic cues, and the clarity of a prompt plays a crucial role in aligning model output with expected grading behavior.

2) Evaluating Explanation Quality Using BERTScore: In our use case, LLMs are being used to evaluate student essays by assigning grades based on rubric-aligned criteria. While we validate grading accuracy by comparing LLM-generated grades with human-assigned grades, it is equally important to also check the quality of the accompanying explanations. High-quality explanations are very useful for interpretability, trust, and reliability in automated educational assessments. To assess whether LLMs can generate explanations comparable in quality to those written by human graders, we evaluate the semantic similarity between LLM-generated and human-generated explanations using BERTScore. BERTScore is a well-established metric that utilizes the contextual embeddings from pre trained language models (such as BERT) to compare a candidate explanation with a reference explanation and unlike traditional n-gram based metrics it captures deeper semantic meaning through token-level matching and contextualized embeddings.

BERTScore computes three key measures:

- Precision: How much of the LLM's explanation is semantically present in the human reference.
- Recall: How much of the human explanation is captured by the LLM's version.
- F1 Score: The harmonic mean of precision and recall, reflecting overall similarity.

The underlying similarity is measured using cosine similarity between token embeddings, where Cosine Similarity = $\frac{A\cdot B}{\|A\|\times\|B\|}$ with A and B representing the embedding vectors of the LLM and human explanations, respectively.

During the evaluation, we restrict the comparison between the human and LLM explanations to only cases where the grades match. For each such instance, we compute the BERTScore between the LLM explanation and the corresponding human explanation. This process helps quantify how well the LLM's explanation resembles human reasoning not just in conclusion (i.e., the grade) but also in the underlying justification. This approach provides a scalable method to evaluate explanation quality across large datasets and helps highlight areas where LLMs may still lack the depth, nuance, or context sensitivity of human evaluators.

E. Experimental Setup

Evaluation Metrics: To assess the alignment between LLM-generated and human-assigned grades, we employed two evaluation metrics: Macro F1 Score and Quadratic Weighted Kappa (QWK). The Macro F1 Score provides a balanced view of performance across all classes, regardless of class frequency, making it particularly useful in the presence of class imbalance as in our case. QWK, on the other hand, measures the agreement between ordinal labels while penalizing larger discrepancies more heavily. Unlike simple accuracy, which treats all errors equally, QWK accounts for the ordered nature of the grading scale, recognizing that misclassifying a grade by two levels (e.g., from "Excellent" to "Poor") is more severe than a one-level difference. This makes QWK particularly suitable for evaluating tasks where the labels have a meaningful ranking, as it better reflects the real-world impact of grading inconsistencies. Together, these metrics offer complementary insights into both classification accuracy and the severity of grading disagreements.We conducted automated essay scoring on a dataset of 100 student essays, evaluating each response across six distinct assessment indicators (AI-1 to AI-6). Each indicator is designed to evaluate the students understanding of a specific aspect of the subject matter. These criteria encompass a range of cognitive and analytical skills, such as analyzing the impact of specific events, examining the effect of an event on particular metrics, proposing improvements to given scenarios, and discussing key domain-relevant topics. This multi-dimensional evaluation framework allows for a more nuanced and comprehensive assessment of student performance, extending beyond overall essay quality. The final grade for each essay is computed as an aggregation of the scores assigned across all six assessment indicators.

Experiments: We designed three experiments to evaluate the effectiveness, efficiency, and alignment of LLMs in our AES task. The first experiment tested our two-stage RAG framework using GPT-4o as the grading model and LLaMA 3.3 70B for semantic retrieval, comparing grading performance and cost between RAG and non-RAG settings. The second experiment compared commercial (GPT-4o, Gemini 2.5 Pro) and open-source (LLaMA 3.3 70B) models on the same set of 100 student essays, assessing both grading quality and cost-effectiveness. Finally, the third experiment evaluated the alignment between LLM-generated and human-written explanations using BERTScore, to determine how closely model rationales reflect human reasoning when the grade assigned was consistent. Each experiment was designed around a shared grading task based on six rubric-aligned criteria for essays from finance domain.

IV. Results

In this section, we provide our findings based on our detailed numerical study.

A. Effect of RAG Application on Performance and Cost

To evaluate the impact of Retrieval-Augmented Generation (RAG) on grading performance and computational efficiency, we conducted a controlled experiment using the GPT-4o model. This experiment compared the models performance under two conditions: a baseline setting in which the full essay was directly input to the model, and a RAG-enhanced setting where only retrieved, semantically relevant chunks were provided. The RAG pipeline involved both lexical and LLM-based retrieval stages, with LLaMA 3.3 70B used in the second stage to refine relevance via semantic filtering.

The evaluation was conducted on 100 student essays written on revenue recognition, each scored across six rubric-aligned criteria. Identical prompts were used across both settings to ensure comparability, with only the input content differing. Performance was measured using Macro F1 and Quadratic Weighted Kappa (QWK), two metrics widely used in automated grading for assessing classification accuracy and ordinal agreement, respectively.

As shown in Table II, the RAG-enhanced setup achieved comparable or slightly improved performance across both metrics, particularly for criteria with higher variability (e.g., MPI-6). These results suggest that focusing the model's input on contextually relevant content does not degrade scoring accuracy, and may help improve alignment in more complex rubric dimensions.

TABLE II: Model Performance With and Without RAG

RAG	Metric	AI-1	AI-2	AI-3	AI-4	AI-5	AI-6	Average
No	F1 Score	1.000	0.642	0.686	0.757	0.871	0.689	0.774
	QWK	1.000	0.662	0.552	0.665	0.748	0.643	0.712
Yes	F1 Score	1.000	0.644	0.684	0.750	0.874	0.766	0.786
	QWK	1.000	0.664	0.554	0.665	0.752	0.694	0.721

Beyond performance, a key motivation for RAG was reducing inference cost through lower token consumption. Using GPT-4o's published pricing ($5 per 1M input tokens, $15 per 1M output tokens), we estimated a blended average of $10 per 1M tokens. The total cost for grading 100 essays was $18.35 in the baseline condition and $10.20 with RAGa 44% cost reduction.

Our token usage analysis reveal the following:

- Without RAG: Approx. 2,039 tokens per evaluation

- With RAG: Approx. 1,133 tokens per evaluation

These numbers reflect a reduction of approximately 906 tokens per evaluation. Although our theoretical estimate had projected a 63.8% reduction, the observed 44% reduction is directionally consistent, with deviations attributed to essay length variability and differing generation lengths across prompts.

Overall, the RAG framework demonstrates promising benefits in reducing computational costs while preserving grading quality, making it a viable strategy for scalable, cost-sensitive educational assessment applications.

B. Comparison of Open-Source and Commercial LLMs

To investigate the trade-offs between model performance and inference cost, we conducted a comparative evaluation of three LLMs for our AES task, namely, GPT-4o, Gemini 2.5 Pro and LLaMA 3.3 70B. All three models were evaluated on an identical set of 100 student essays. To ensure fairness in evaluation, the prompt format, scoring template, and rubric guidance were standardized across all models.

The results of this evaluation are summarized in Table III, which reports F1 and QWK scores for each model across the six individual rubric criteria. As shown, GPT-4o achieved consistently strong performance, with F1 scores ranging from 0.644 to 1.000 and QWK values ranging from 0.552 to 1.000. Gemini 2.5 Pro produced comparable results, although slightly lower on more difficult tasks such as MPI-3 and MPI-4. LLaMA 3.3 70B performed marginally below its commercial counterparts across most rubric items but remained within a reasonable range, demonstrating that it can approximate commercial performance on high-level tasks such as rubric-aligned grading.

We visualize the average F1 and QWK scores for each model in Figure 2. The plot highlights that GPT-4o leads in both metrics, followed closely by Gemini 2.5 Pro. LLaMA 3.3 70B trails slightly behind but remains competitive. While these differences are measurable, they are not dramatic, particularly when considering the substantial cost differences discussed below.

In terms of cost-effectiveness, Figure 3 compares the API pricing-based inference cost per million tokens for each model. GPT-4o, with a blended input-output rate of approximately $10.20 per million tokens, was the most expensive. Gemini 2.5 Pro followed at $6.85. In contrast, LLaMA 3.3 70Bserved through Groq's high-speed inference infrastructurecost only $1.37 per million tokens, representing a reduction of over 85% compared to GPT-4o. This cost differential becomes highly significant when grading large volumes of student responses, as would be typical in institutional or statewide deployments.

Taken together, these results suggest that while commercial models currently offer marginally higher grading performance, open-source alternatives like LLaMA 3.3 70B can deliver strong performance at a fraction of the cost. For large-scale, cost-sensitive applications such as formative assessment or curriculum-wide grading pipelines, this performance-cost trade-off makes open-source models a viable and economically attractive alternative.

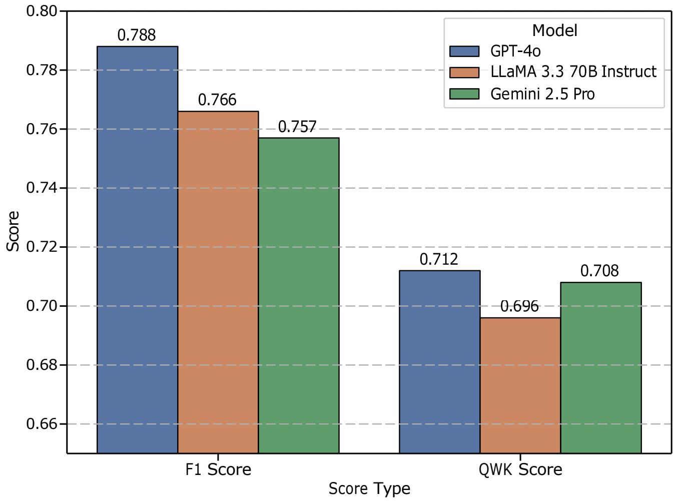

Fig. 2: Performance comparison of GPT-4o, Gemini 2.5 Pro, and LLaMA 3.3 70B.

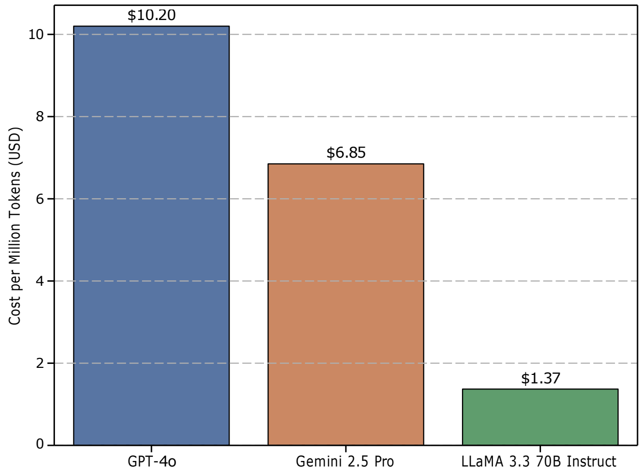

Fig. 3: Cost comparison of GPT-4o, Gemini 2.5 Pro, and LLaMA 3.3 70B.

C. Evaluation of Model-Generated Feedback Quality

While previous sections evaluated the accuracy of grading decisions, another critical dimension of automated assessment lies in the quality of feedback generated by language models. To assess this, we examined whether the explanations produced by `LLaMA 3.3 70B` are semantically aligned with those written by expert human graders. Our goal was not merely to test linguistic fluency or surface similarity, but to assess whether the models rationale exhibits comparable depth of understanding, specificity, and domain-relevance in the context of educational feedback.

To this end, we conducted an experiment using BERTScore, a semantic similarity metric that compares contextual embeddings of candidate and reference texts. Specifically, we selected 100 student essays where both the model and a human evaluator had independently assigned the same binary grade (Y or N) to each rubric criterion. This ensured that the comparison

TABLE III: Comparison of Open Source vs Commercial LLMs using F1 and QWK

LLM	Metric	AI-1	AI-2	AI-3	AI-4	AI-5	AI-6	Average
GPT-4o	F1 Score	1.0000±0.000	0.644±0.091	0.684±0.082	0.752±0.097	0.878±0.073	0.767±0.089	0.788±0.120
	QWK	1.0000±0.000	0.662±0.078	0.552±0.091	0.665±0.074	0.748±0.066	0.643±0.087	0.712±0.141
LLaMA 3.3 70B	F1 Score	1.000±0.000	0.613±0.095	0.639±0.104	0.720±0.088	0.884±0.069	0.746±0.106	0.767±0.136
	QWK	1.000±0.000	0.584±0.085	0.544±0.193	0.620±0.179	0.734±0.063	0.696±0.072	0.696±0.150
Gemini 2.5 Pro	F1 Score	1.000±0.000	0.642±0.087	0.620±0.112	0.679±0.079	0.848±0.076	0.751±0.083	0.757±0.133
	QWK	1.000±0.000	0.594±0.089	0.538±0.198	0.629±0.281	0.737±0.165	0.750±0.070	0.708±0.151

was focused solely on the quality of the explanation, rather than any disagreement in judgment. For each of these matched cases, we collected the explanation generated by the model as well as the corresponding explanation written by a human evaluator.

BERTScore was computed using the `roberta-large` model, a widely adopted pre-trained transformer for semantic evaluation tasks. Table IV presents the aggregated results across the 100 matched explanation pairs.

TABLE IV: Average BERTScore precision, recall, and F1 comparing model- and human-generated explanations (N=100).

	Precision	Recall	F1 Score
Average Scores	0.860	0.840	0.850

The average BERTScore F1 score of 0.850 indicates a high degree of semantic alignment between LLaMA 3.3 70Bs generated explanations and those written by human evaluators. Precision and recall scores were similarly strong, suggesting that the model is not only capturing key semantic content from the reference explanations but is also producing sufficiently rich justifications of its own. Importantly, this level of similarity was achieved without explicit fine-tuning of the model on explanation-writing tasks, indicating robust generalization capabilities in instructional contexts.

To contextualize these quantitative findings, Table V illustrates representative examples of both model-generated and human-generated explanations across multiple grading cases. The examples demonstrate the models ability to reference relevant domain concepts (e.g., market liquidity, risk diversification), to offer justification aligned with the prompts criteria, and to do so with a level of clarity and formality consistent with pedagogical expectations.

These examples reinforce the empirical findings: LLaMA 3.3 70B is capable of generating nuanced and contextually appropriate feedback that parallels human explanations not just in content, but in instructional tone and structure. The implications of these findings are significant. In addition to offering accurate grading decisions, open-source models such as LLaMA 3.3 70B can support formative assessment by providing students with constructive feedback that closely mirrors human-generated input. This enhances the practical viability of LLMs in large-scale, feedback-oriented educational settings.

In sum, the BERTScore-based evaluation reveals that the LLMs explanations align well with expert feedback in both semantic content and communicative clarity. When combined with earlier performance and cost findings, this positions open-source LLMs as compelling candidates for integrated use in both summative and formative assessment systems.

V. Conclusion

This work offers a comprehensive exploration of the capabilities and limitations of LLMs in AES. By systematically investigating performance, efficiency, and explanatory quality, we demonstrate that strategically designed LLM-based systems can deliver grading outcomes that are not only accurate and cost-effective but also pedagogically meaningful.

A key contribution of this study is the demonstration that RAG can significantly enhance the efficiency of automated grading. By focusing the model's attention on semantically relevant content, RAG reduces token consumption by over 40% while preserving, and in some cases improving, grading accuracy. This represents a practical advancement for institutions seeking to scale automated assessment without incurring prohibitive computational costs.

Equally important is our comparative analysis of commercial and open-source LLMs. While models like GPT-4o and Gemini 2.5 Pro remain state-of-the-art in terms of performance, we find that open-source models such as `LLaMA 3.3 70B`, when deployed through high-performance inference frameworks, can approximate commercial performance at a fraction of the cost. This cost-performance trade-off is critical for budget-sensitive contexts such as public education systems or large-scale testing environments, where financial constraints and data privacy concerns make proprietary solutions less feasible. In addition, our work also addresses the quality of feedback produced by LLMs, a dimension often overlooked in AES research. Using semantic similarity analysis against human-written explanations, we find that open-source models are capable of generating feedback that aligns closely with expert reasoning in both content and communicative clarity. This suggests that LLMs are not only viable as evaluators but also as instructional support tools capable of enhancing formative assessment.

Together, these findings point toward a future in which open-source LLMs, enhanced with retrieval mechanisms and deployed through efficient infrastructure, can provide scalable, transparent, and pedagogically aligned assessment solutions. They offer the promise of extending high-quality feedback and fair evaluation to broader student populations, especially in under-resourced educational settings.

We note there are certain limitations of our work. The grading tasks addressed here involved clearly defined, instruction-following prompts that emphasized conceptual understanding

TABLE V: Representative feedback explanations generated by LLaMA 3.3 70B and human graders.

Mark	**LLM Explanation**	**Human Explanation**
Y	The essay provides a clear and well-supported argument explaining how interest rate fluctuations influence stock market performance. It accurately links central bank policy to investor sentiment and offers relevant examples.	This response effectively analyzes the relationship between monetary policy and equity markets. The student articulates the impact of interest rate changes on market liquidity and investor behavior.
N	The essay lacks a coherent explanation of how inflation impacts purchasing power and investment decisions. Key concepts are missing or incorrectly applied, and no examples are provided.	The student does not sufficiently address the relationship between inflation and its economic consequences. The explanation is vague, showing limited understanding.
N	The essay fails to explain the concept of risk diversification in investment portfolios. It remains overly generic and does not demonstrate understanding of how diversification reduces unsystematic risk.	The response lacks clarity and misses the core idea of diversification in finance. The student does not provide examples or show comprehension of risk types.

rather than deep logical reasoning or mathematical derivation. The generalizability of these findings to more complex domains remains to be validated. Moreover, prompt engineering and feedback loop refinement remain manual and dependent on domain expertise, posing a barrier to widespread adoption. In this regard, future work should investigate automated prompt optimization strategies, including the use of agentic AI systems capable of iteratively refining grading prompts based on feedback evaluation. Another promising direction lies in dynamically selecting or composing models based on task complexityleveraging the complementary strengths of different LLMs to maximize accuracy, interpretability, and efficiency. Ultimately, the integration of LLMs into educational systems must be guided not only by performance metrics but also by a commitment to transparency, equity, and instructional value.

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] M. Stahl, L. Biermann, A. Nehring, and H. Wachsmuth, “Exploring LLM prompting strategies for joint essay scoring and feedback generation,” *arXiv preprint arXiv:2404.15845*, 2024.

[3] 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.

[4] Z. Ke and V. Ng, “Automated essay scoring: A survey of the state of the art,” in *IJCAI*, vol. 19, 2019, pp. 6300–6308.

[5] R. K. Helmeczi, S. Yildirim, M. Cevik, and S. Lee, “Few shot learning approaches to essay scoring,” in *Canadian AI*, 2023.

[6] T. Firoozi, O. Bulut, and M. Gerl, “Language models in automated essay scoring: Insights for the Turkish language,” *International Journal of Assessment Tools in Education*, vol. 10, no. Special Issue, pp. 149–163, 2023.

[7] C. Xiao, W. Ma, Q. Song, S. X. Xu, K. Zhang, Y. Wang, and Q. Fu, “Human-AI collaborative essay scoring: A dual-process framework with LLMs,” *arXiv preprint arXiv:2401.06431*, 2024. [Online]. Available: https://arxiv.org/abs/2401.06431

[8] M. Kostic, H. F. Witschel, K. Hinkelmann, and M. Spahic-Bogdanovic, “LLMs in automated essay evaluation: A case study,” in *Proceedings of the AAAI Symposium Series*, vol. 3, no. 1, 2024, pp. 143–147.

[9] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen *et al.*, “A comprehensive capability analysis of GPT-3 and GPT-3.5 series models,” *arXiv preprint arXiv:2303.10420*, 2023.

[10] OpenAI Platform, “GPT-4o-2024-05-13,” 2024. [Online]. Available: https://platform.openai.com/docs/models/gpt-4o

[11] Google DeepMind, “Gemini 2.5 Pro,” 2025. [Online]. Available: https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro

[12] Qwen Team, “Qwen2.5 Technical Report,” 2024. [Online]. Available: https://arxiv.org/abs/2412.15115

[13] Meta AI, “Llama 3.3 70B Model Card,” 2024. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[14] K. Hyland, “Writing in the university: Education, knowledge and reputation,” *Language Teaching*, vol. 46, no. 1, pp. 53–70, 2013.

[15] Y. Salim, V. Stevanus, E. Barlian, A. C. Sari, and D. Suhartono, “Automated English digital essay grader using machine learning,” in *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*, pp. 1–6, IEEE, 2019.

[16] F. Dong and Y. Zhang, “Automatic features for essay scoring–An empirical study,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1072–1077.

[17] K. Taghipour and H. T. Ng, “A neural approach to automated essay scoring,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1882–1891.

[18] J. Lun, J. Zhu, Y. Tang, and M. Yang, “Multiple data augmentation strategies for improving performance on

automatic short answer scoring," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 9, 2020, pp. 13389–13396.

[19] N. A. Sheikh and F. Al-Makhadmeh, "LLM-based assessment system for short answer grading: Enhancing automated evaluation with transformers," in *2024 International Conference on Smart Applications and Data Analytics (SADA)*, IEEE. [Online].

[20] T. Shibata and Y. Miyamura, "LCES: Zero-shot Automated Essay Scoring via Pairwise Comparisons Using Large Language Models," *arXiv preprint arXiv:2505.08498*, 2025. [Online]. Available: https://arxiv.org/abs/2505.08498

[21] Y. Cai, K. Liang, S. Lee, Q. Wang, and Y. Wu, "Rank-Then-Score: Enhancing Large Language Models for Automated Essay Scoring," *arXiv preprint arXiv:2504.05736*, 2025. [Online]. Available: https://arxiv.org/abs/2504.05736

[22] S. Eltanbouly, S. Albatarni, and T. Elsayed, "TRATES: Trait-Specific Rubric-Assisted Cross-Prompt Essay Scoring," *arXiv preprint arXiv:2505.14577*, 2025. [Online]. Available: https://arxiv.org/abs/2505.14577

[23] G. Malik, M. Cevik, and S. Lee, "Exploring Large Language Models for Automated Essay Grading in Finance Domain," in *Proceedings of the 34th Annual International Conference on Collaborative Advances in Software and Computing (CASCON 2024)*, Toronto, Canada, Nov. 2024, pp. 1–10, IEEE. doi:10.1109/CASCON.2024.10838105. [Online]. Available: https://www.computer.org/csdl/proceedings-article/cascon/2024/10838105/23zYzrJGvYs

[24] L. Yoshida, "Do We Need a Detailed Rubric for Automated Essay Scoring using Large Language Models?," *arXiv preprint arXiv:2505.01035*, 2025. [Online]. Available: https://arxiv.org/abs/2505.01035

[25] J. Su, Y. Yan, F. Fu, H. Zhang, J. Ye, X. Liu, J. Huo, H. Zhou, and X. Hu, "EssayJudge: A Multi-Granular Benchmark for Assessing Automated Essay Scoring Systems," *arXiv preprint arXiv:2502.11916*, 2025. [Online]. Available: https://arxiv.org/pdf/2502.11916

[26] H. Do, S. Ryu, and G. G. Lee, "Teach-to-Reason with Scoring: Self-Explainable Rationale-Driven Multi-Trait Essay Scoring," *arXiv preprint arXiv:2502.20748*, 2025. [Online]. Available: https://arxiv.org/abs/2502.20748