Diverse LLM Approaches in Essay Scoring: A Comparative Exploration of Many-Shot Prompting, LLM Jury Panels, and Model Fine-Tuning

Stanford CS224N {Custom} Project

Alexa Sparks
School of Education
Stanford University
absparks@stanford.edu

Rizwaan Malik
School of Education
Stanford University
rizmalik@stanford.edu

Matías Hoyl
School of Education
Stanford University
mhoyl@stanford.edu

Abstract

This paper explores diverse approaches to automatic essay scoring (AES) utilizing large language models (LLMs). We investigate three novel methods: many-shot prompting using Gemini 1.5 Flash, a jury of models, and model fine-tuning with Llama 3. Our experiments run on the Learning Agency Lab - Automated Essay Scoring 2.0 dataset, consisting of 24,000 student-written argumentative essays, to evaluate the effectiveness of these techniques in improving scoring accuracy. The many-shot prompting approach with Gemini 1.5 Flash achieved a quadratic weighted kappa (QWK) of 0.82, outperforming the baseline. The jury of models strategy, which combines scores from Mixtral-8x7b-Instruct-v0.1, Gemini 1.5 Flash, and Llama-3-7b-chat, resulted in a QWK of 0.76, and also outperforms the GPT-4 baseline performance. Conversely, fine-tuning Llama 3 using the Low-Rank Adaptation (LoRA) method yielded a QWK of only 0.31, highlighting the challenges and limitations of this approach. Our findings suggest that leveraging extended context and multiple examples in prompts can significantly enhance the accuracy and reliability of AES systems, while ensemble methods offer a cost-effective alternative to single large models.

1 Key Information to include

- Mentor: Zhoujie Ding

2 Introduction

Automatic essay scoring (AES) has gained attention in recent years, particularly with the rise of large-scale educational assessments and the potential for efficient grading solutions. The task of AES is inherently challenging due to the nature of essays, which are long, unstructured, and vary widely in style and content. Even with a detailed rubric, there is always a degree of subjectivity in scoring, making it difficult to achieve consistent and reliable results. Essays often require nuanced understanding and contextual analysis, which are difficult for traditional models to capture.

Despite these challenges, the potential benefits of developing effective AES systems are substantial. Teachers spend a significant amount of time reviewing essays and providing feedback. An effective AES system could offer preliminary scores and feedback, allowing teachers to focus more on instructional strategies and personalized support for students. This can lead to more efficient classroom management and better educational outcomes.

With recent advances in large language models (LLMs), there is potential to leverage these models for the AES scoring task. We believe there is substantial room for innovation in three key areas. Firstly, leveraging LLMs with large context windows, such as Gemini 1.5 Flash, and utilizing many-shot

prompting can significantly improve performance by providing the model with more contextual examples during inference [1]. Secondly, recent evidence suggests that a jury of smaller models can be more effective and cost-efficient than a single large model, reducing bias and enhancing scoring reliability [2]. Lastly, new open models, such as Llama 3, offer promising alternatives to BERT for fine-tuning, potentially leading to better baseline models for AES tasks.

In this work, we utilize a dataset of 24,000 student-written argumentative essays provided by The Learning Agency Lab for a 2024 automated essay scoring competition on Kaggle. Our experiments compare the performance of each approach and find the following results:

- **Many-shot prompting with Gemini 1.5 Flash:** This approach achieved a quadratic weighted kappa (QWK) of 0.82, outperforming the baseline QWK of 0.59 obtained with the same model but without examples in the prompt.
- **Panel of juries:** Implementing a jury of diverse models resulted in a QWK of 0.76, compared to a baseline of 0.67 using GPT-4 with zero-shot prompting.
- **Fine-tuned Llama 3 model:** This model achieved a QWK of 0.31, highlighting the potential but also the current limitations of this approach in comparison to more sophisticated prompting strategies.

Our findings indicate that the long context window approach with multiple examples in the prompt performs the best, demonstrating the highest QWK score among the tested methods. This suggests that leveraging extended context and multiple examples can significantly enhance the accuracy and reliability of automated essay scoring systems.

3 Related Work

AES has evolved significantly over the years, leveraging advancements in artificial intelligence (AI), machine learning (ML), and natural language processing (NLP). The timeline of efforts in AES illustrates the progression from simple statistical methods to sophisticated AI techniques.

One of the earliest AES systems, Project Essay Grade (PEG), was developed by Ellis Page in the 1960s. PEG utilized predefined features, such as essay length and word count, to predict human scores using statistical methods [3]. In the 1990s, the Intelligent Essay Assessor (IEA) by Landauer and Foltz employed Latent Semantic Analysis (LSA) to evaluate the semantic content of essays [4].

Developed by Educational Testing Service (ETS), e-rater is a well-known AES system that combines NLP techniques and machine learning algorithms to evaluate essays based on grammar, usage, mechanics, style, and organization [5]. The system has been continuously updated to incorporate more sophisticated NLP techniques and models.

Recent advancements in deep learning have led to neural network-based AES systems. Taghipour and Ng proposed an approach using convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to capture both local and global features of essays, showing promising results in terms of accuracy and robustness [6].

The introduction of Bidirectional Encoder Representations from Transformers (BERT) by Devlin and colleagues significantly impacted AES. BERT-based models could understand the context of words in a sentence more effectively, leading to more accurate essay scoring [7].

Two recent papers stand out for LLMs and advanced AI workflows to achieve better results in AES. The first paper by Verga and colleagues proposes the use of a panel of diverse models to evaluate LLM outputs, reducing bias and improving accuracy through collective decision-making [2]. The second paper by Agarwal and colleagues explores many-shot in-context learning, utilizing multiple examples to leverage extended context windows in LLMs, demonstrating significant improvements in various NLP tasks [1].

Despite these advancements, the integration of LLMs in AES has not been extensively explored. Our research aims to fill this gap by evaluating and comparing the effectiveness of many-shot prompting, LLM jury panels, and model fine-tuning in automated essay scoring.

4 Approach

To ensure consistency across our experiments, we separated our dataset into a training set (13,845 rows) and a test set (3,462 rows). Additionally, for practical purposes such as speed of testing and inference cost of the API models, we created an abbreviated test dataset of 20 essays. This subset was used to ultimately test the approaches.

4.1 Long Context Window

First, we established a baseline using the Gemini 1.5 Flash model without examples in the prompt (zero-shot). For this, we passed the rubric and the essay to be scored to the model. This process was repeated five times on the abbreviated dataset of 20 essays to ensure consistent results. We computed the QWK for each run and averaged the scores to obtain the final baseline.

Next, we used the same Gemini 1.5 Flash model but with 3,000 examples randomly chosen from the training dataset, ensuring a balanced distribution of scores, included in the prompt. Although Gemini 1.5 Flash has a large context window of 1 million tokens, it maxed out around 3,000 essays, which is why we used only 3,000 examples instead of the entire training dataset. For each essay in the abbreviated dataset, we fed the model the examples along with the rubric and the essay to be scored. This process was also repeated five times, and the QWK scores were averaged to get the final score for this approach.

4.2 Jury of Models

We first established a baseline using zero-shot prompting with GPT-4, conducting five separate trials to ensure the reliability and stability of our results. The QWK scores for these baseline trials and outcomes from our model ensemble are reported in Table 1.

Our ensemble approach involved aggregating scoring results from three diverse models: Mixtral-8x7b-Instruct-v0.1, Gemini 1.5 Flash, and Llama-3-7b-chat. Each model independently scored essays using a holistic scoring rubric provided by The Learning Agency—this rubric was also utilized by human assessors to provide the original essay scores. To maintain consistency with our baseline assessments, we applied zero-shot prompting across all models.

To determine the final ensemble essay scores, we experimented with two methods: majority voting and mean score calculation. The majority voting method selected the most common score provided by the models as the final score, while the mean score method computed the average of the scores. Across all five trials, the mean score calculation method consistently resulted in higher QWK scores. This finding aligns with the research by Verga and colleagues, who demonstrated the higher performance of averaging techniques in ensemble model scenarios when dealing with rank scores [2].

4.3 Finetuned LLM

For the baseline, we used two out-of-the-box models: Llama-3-7b-chat and Llama-3-7b-Instruct. To enhance their performance for the essay scoring task, we employed the Low-Rank Adaptation (LoRA) technique, which modifies a subset of the model's parameters. Specifically, we adjusted 10% of the parameters, aiming to make the models more adept at handling the nuances of essay scoring.

LoRA is a method designed to efficiently fine-tune large language models by focusing on low-rank updates. Instead of updating the entire model, LoRA introduces trainable low-rank matrices into each layer of the transformer architecture. This approach significantly reduces the number of parameters that need to be trained, making the fine-tuning process more resource-efficient while still effectively adapting the model to new tasks. LoRA maintains the pre-trained weights of the original model, which helps in preserving the knowledge acquired during the initial large-scale training phase.

Our fine-tuning process was conducted in two stages. Initially, we fine-tuned both Llama-3-7b-chat and Llama-3-7b-Instruct models on a smaller subset, consisting of 1% of the available training data. This preliminary step was crucial for evaluating the feasibility and impact of our fine-tuning strategy without incurring high computational costs. After confirming the viability of our approach, we proceeded to fine-tune the models on the full training dataset, aiming to fully leverage the available data to enhance model performance.

Data Preparation for Fine-Tuning

- **Data Cleaning:** We removed any incomplete or corrupted entries to ensure the quality of the training data.
-
- **Tokenization:** The essays were tokenized into sequences suitable for input into the LLMs, ensuring consistency with the models' pre-trained tokenization schemes.
- **Formatting:** Each essay was paired with its corresponding score from the rubric. The data was then formatted into a structured input-output format compatible with the fine-tuning process.
  - **Input Format:** Each input was structured as follows:

    ```
    system
    {{ system_prompt }}
    user
    {{ user_msg_1 }}
    assistant
    {{ model_answer_1 }}
    ```

    where $system\_prompt$ is the main prompt, $user\_msg\_1$ is the essay, and $model\_answer\_1$ is the score.

This data preparation aimed to ensure that the fine-tuning process provided the models with high-quality, structured input to learn from.

5 Experiments

5.1 Data

Our experiments utilize the Learning Agency Lab - Automated Essay Scoring 2.0 dataset, which comprises 24,000 student-written argumentative essays. Each essay is scored on a scale of 1 to 6, and the dataset includes a rubric with guiding instructions for evaluating each essay. This dataset is publicly available on Kaggle.

5.2 Evaluation method

To measure the effectiveness of our automated scoring methods, we employ the quadratic weighted kappa (QWK) metric. This metric is particularly suitable for our the AES task because it accounts for the degree of disagreement between scores, offering a more nuanced comparison of ordinal data. By considering the distance between scores, QWK ensures that minor discrepancies are weighted less than major ones, making it ideal for evaluating the accuracy and reliability of our automated scoring systems compared to human evaluations.

5.3 Results

Long Context Window: For the long context window approach, the baseline QWK using Gemini 1.5 Flash without examples in the prompt (zero-shot) was 0.59. When we included 3,000 examples in the prompt, the QWK improved significantly to 0.82. This demonstrates the effectiveness of utilizing a large context window with many-shot prompting to enhance the scoring accuracy.

Jury of Models: In the jury of models approach, the baseline QWK using GPT-4 in a zero-shot configuration was 0.67. By employing an ensemble of smaller models (Mixtral-8x7B-Instruct-v0.1, Gemini 1.5 Flash, and Llama-3-7b-chat), the QWK improved to 0.76. This result aligns with our expectations based on the literature, showing that the jury mechanism can provide more reliable and cost-effective scoring compared to a single large model [2].

Finetuned LLM: For the finetuned LLM approach, the baseline QWK using the out-of-the-box Llama-3-7b-chat model was 0.31. Our first attempt at fine-tuning this model using LoRA yielded a very low QWK score of 0.17. After improving the prompt engineering and refining our approach, the

second attempt resulted in a QWK of 0.31. Despite these efforts, the fine-tuned model's performance remained significantly lower than the out-of-the-box Llama-3-7b-chat model, which scored 0.72.

This outcome was surprising, as we initially hypothesized that fine-tuning with the entire training dataset would yield better performance due to the model's adaptation to the specific characteristics of the essay scoring task. The poor performance of the fine-tuned model might be attributed to several factors. The parameters chosen for fine-tuning, including the learning rate, batch size, and number of epochs, might not have been optimal. Incorrect parameter settings can lead to underfitting or overfitting, both of which can negatively impact model performance. Additionally, the fine-tuning process may not have sufficiently adapted the model to the nuances of the essay scoring task. Essays are inherently complex and require a deep understanding of structure, coherence, and content, and the fine-tuning process might have failed to capture these subtleties.

Another potential issue could be overfitting to the training data. Given the relatively small size of the fine-tuning dataset, the model might have overfitted to the specific examples seen during training, leading to poor generalization on the test set. This overfitting could have resulted in the fine-tuned model performing worse than the out-of-the-box model. Furthermore, the quality and distribution of the training data used for fine-tuning could have affected the model's performance. This is explored more in the Analysis section.

Method	**Baseline QWK**	**Improved QWK**
Long Context Window	0.59	0.82
Jury of Models	0.67	0.76
Finetuned LLM	0.17	0.31

Table 1: Comparison of baseline and improved QWK scores for different methods.

The results of the jury of models were consistent with our expectations. Based on the research of Vergara and colleagues, we anticipated the jury of models to outperform the state of the art GPT-4 model. Our results confirmed this, demonstrating the efficacy of the jury mechanism in enhancing scoring reliability.

We were particularly impressed with the results of the long context window approach. We did not expect it to be the best performing method, yet it achieved a QWK of 0.82. This performance can be attributed to the high quality of the Gemini 1.5 Flash model, which is known for its efficiency and accuracy, as well as its ability to leverage and generalize from the examples provided in the prompt (in-context learning).

6 Analysis

We conducted a detailed evaluation of the behavior of each model to understand when and why they perform well or poorly on the AES task.

Mixtral-8x7b-Instruct-v0.1 This model showed strength in scoring essays that were within the 600-700 word range and demonstrated readability at or above a 10th-grade level according to the Flesch-Kincaid Readability test [8]. It seemed to prefer structurally clear and linguistically precise texts. Notably, essays that received accurate scores from Mixtral typically conveyed a positive sentiment, suggesting that the model may be sensitive to the emotional tone of the writing.

Llama-3-7b-chat Llama showed a broad range of capability, accurately scoring essays with diverse structural characteristics. It managed essays with varying sentence lengths (13 to 45 words per sentence) and complexity (about 10% of words in each essay exceeded seven letters). These characteristics indicate Llama's ability to handle essays with different levels of lexical and structural complexity.

Gemini 1.5 Flash Gemini was particularly adept at scoring essays with high lexical diversity, scoring essays accurately where the type-token ratio exceeded 0.4. It could handle a wide variety of sentence lengths, from as short as 7 to as long as 124 words. However, its performance on essays

with lower Flesch-Kincaid readability scores suggests some difficulties with complex texts that could impact its comprehension and scoring accuracy.

Overall, each model displayed distinct strengths and weaknesses, indicating that a single model might not be universally effective across all types of essay content.

6.1 Fine-Tuning

When manually inspecting the fine-tuned data, we observed that the fine-tuned model was assigning scores of 3 to almost all examples in the test set. This consistent assignment led to the poor QWK score observed. To understand this behavior, we analyzed the distribution of essays by score in the training set, as shown in Table 2.

Score	**1**	**2**	**3**	**4**	**5**	**6**
Percentage	7.23	27.29	36.29	22.69	5.60	0.90

Table 2: Distribution of essays by score in the training set.

From the distribution, it is clear that there is an over representation of essays scored as 3. This imbalance likely caused the model to overfit to the predominant score in the training data, leading it to frequently predict a score of 3 for essays in the test set. This overfitting issue highlights the importance of balanced training data in ensuring robust model performance.

The model's tendency to assign a score of 3 to most test set examples resulted in a poor QWK score, as the metric penalizes uniform predictions that do not reflect the true distribution of scores. The poor score indicates that the model failed to accurately capture the variability and nuances present in the essays.

This analysis underscores the need for careful consideration of the training data's distribution when fine-tuning models. Ensuring a more balanced distribution of scores in the training dataset might help in mitigating such overfitting issues and improving the model's ability to generalize better to unseen data. Additionally, further exploration of alternative fine-tuning techniques and parameter optimization could help in addressing these challenges and enhancing the model's performance in AES tasks.

7 Conclusion

Our study of diverse LLM approaches to AES yielded significant insights into the capabilities and limitations of many-shot prompting, jury of models, and fine-tuning evaluations. Among the three methods tested, the many-shot prompting approach using Gemini 1.5 Flash emerged as the most effective. This approach demonstrated a clear advantage in enhancing scoring accuracy, surpassing the baseline model performance. The success of this method highlights the importance of context and examples in improving the reliability of LLM outputs.

The implementation of the jury model evaluation, which combined scores from Gemini 1.5 Flash, Llama-3-7b-chat, and Mixtral-8x7b-Instruct-v0.1, also showed improvement over the baseline GPT-4 accuracy. This ensemble method demonstrated that a collective model approach could offer a robust alternative to relying on a single large, and often more expensive, model. The jury of models aligns with recent research and suggests that diversity in model perspectives can yield consistent and reliable outcomes.

Fine-tuning Llama 3, however, did not lead to an improvement in performance. Despite efforts to refine the model through LoRA and improved prompt engineering, the QWK remained at 0.31, significantly lower than the out-of-the-box performance. This result underscores the challenges associated with fine-tuning large models for specific tasks and indicates a potential area for future research. It highlights the need to explore more targeted approaches for AES fine-tuning, optimize fine-tuning parameters, and ensure balanced training data to prevent overfitting. Future work should also consider the integration of more sophisticated adaptation techniques to better align model capabilities with the complex requirements of essay scoring.

8 Ethics Statement

The AES systems explored in this paper introduce potential risks to ethical educational assessment. One significant risk involves the potential reinforcement of biases present in the training data. Large language models and techniques like many-shot prompting and model fine-tuning may inadvertently amplify these biases, affecting fairness and accuracy in scoring at scale. We recommend future AES researchers to include essays from students with diverse backgrounds to assess the affect of these biases on different demographic groups.

The use of LLMs raises questions around transparency and interpretability. These models, particularly when fine-tuned or used in jury systems, can become "black boxes," where decision-making processes are not easily understood. Ensuring transparency in these systems is essential not only for trust and acceptance but also for identifying and correcting errors in scoring.

Lastly, the focus on optimizing for specific metrics such as the QWK could lead to teaching to the test scenarios, where the educational emphasis shifts from critical thinking and creative writing to mastering the idiosyncrasies of scoring algorithms. This could undermine educational objectives, promoting a narrow set of skills over a broader educational development.

Researchers and education assessment practitioners are urged to approach the integration of these AES technologies with caution, implementing rigorous testing phases, continuous monitoring for unintended consequences, and maintaining an ongoing dialogue with educational experts to ensure that these tools serve to support and enhance educational outcomes, rather than detract from them.

References

[1] Rishabh Agarwal et al. *Many-Shot In-Context Learning*. 2024. arXiv: 2404.11018 [cs.LG].

[2] Pat Verga et al. *Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models*. 2024. arXiv: 2404.18796 [cs.CL].

[3] Ellis B. Page. "The Imminence of Grading Essays by Computer". In: *Phi Delta Kappan* 47.5 (1966), pp. 238–243.

[4] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. "An Introduction to Latent Semantic Analysis". In: *Discourse Processes* 25.2-3 (1998), pp. 259–284.

[5] Yigal Attali and Jill Burstein. "Automated Essay Scoring with e-rater® V.2". In: *Annual Meeting of the American Educational Research Association*. 2006.

[6] Kaveh Taghipour and Hwee Tou 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.

[7] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[8] J. P. Kincaid et al. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. https://doi.org/10.21236/ADA006655. 1975.