# Enhancing Automated Essay Scoring through Fine-tuning and Prompt Engineering

Author

April 6, 2025

**Abstract**

This paper presents an improved approach to automated essay scoring using the DREsS rubric through a combination of model fine-tuning and prompt engineering techniques. We fine-tuned the Microsoft Phi-2 model on a subset of the DREsS_New dataset and enhanced the system's performance through optimized prompting strategies. Our results demonstrate significant improvements in scoring accuracy compared to the baseline model, with the enhanced system achieving scores closer to expert evaluations. The paper details our methodology, presents experimental results, and discusses future directions for research in this domain.

## 1 Introduction

Automated essay scoring (AES) systems have become increasingly important in educational settings, providing efficient, consistent evaluation of student writing while reducing the time and resources required for manual grading. Despite advances in natural language processing techniques, accurately assessing the nuanced aspects of essay quality remains challenging.

This research addresses these challenges by leveraging the DREsS dataset [Yoo et al., 2024] and applying fine-tuning and prompt engineering techniques to the Microsoft Phi-2 language model to enhance scoring accuracy. Our approach focuses on three key evaluation criteria: content, organization, and language usage, as defined by the DREsS rubric.

# 2 Method

## 2.1 Dataset

For this study, we utilized the DREsS_New dataset from the Dataset for Rubric-based Essay Scoring on EFL Writing [Yoo et al., 2024]. DREsS is a comprehensive collection designed specifically for automated essay scoring research and comprises three sub-datasets: DREsS_New, DREsS_Std, and DREsS_CASE. For our experiments, we focused exclusively on DREsS_New, which contains 1.7K essays authored by English as a Foreign Language (EFL) undergraduate students and scored by English education experts.

The DREsS evaluation framework assesses essays on a scale of 1 to 5 (with 0.5 increments) across three dimensions:

1. **Content**: Evaluates whether paragraphs are well-developed and relevant to the argument, with strong supporting reasons and examples.
2. **Organization**: Assesses how effectively the argument is structured and developed, focusing on readability and coherence.
3. **Language**: Examines the sophistication of vocabulary, adherence to grammar and usage rules, and correctness of spelling and punctuation.

Due to computational constraints when running on a Mac M3 Pro using CPU processing, we limited our dataset to the first 500 rows of DREsS_New.tsv for fine-tuning.

## 2.2 Fine-tuning Process

We selected the Microsoft Phi-2 model for fine-tuning due to its relatively small size (2.7B parameters) while maintaining strong performance on language tasks. This model provides a good balance between computational efficiency and effectiveness for specialized tasks like essay evaluation.

Our fine-tuning process consisted of the following steps:

1. **Dataset preparation**: We augmented the selected essays with score justifications generated using the GPT-4o-mini model. These justifications included specific examples from the essays highlighting strengths or weaknesses for each scoring dimension.
2. **Dataset formatting**: The data was formatted as instruction-response pairs in JSONL format. Each pair included the essay prompt, student essay, and expert-assigned scores along with generated justifications.

3. **Low-Rank Adaptation (LoRA)**: We utilized Parameter-Efficient Fine-Tuning (PEFT) techniques, specifically LoRA, to efficiently fine-tune the model while minimizing computational requirements and avoiding catastrophic forgetting.

The hyperparameters used for fine-tuning were:

```
LORA_R = 16
LORA_ALPHA = 32
LORA_DROPOUT = 0.05
LEARNING_RATE = 2e-4
NUM_TRAIN_EPOCHS = 3
PER_DEVICE_TRAIN_BATCH_SIZE = 1
GRADIENT_ACCUMULATION_STEPS = 4
MAX_GRADIENT_NORM = 0.3
```

We targeted the query, key, value projection matrices and dense layers in the model's attention mechanism for adaptation.

## 2.3 Prompt Engineering

After fine-tuning, we enhanced our system through targeted prompt engineering. Our baseline prompt was:

```
"Instruction: You are an expert essay evaluator.
   Evaluate the following essay based on the DREsS
   rubric."
```

Based on recent research on prompt optimization, we implemented two key techniques:

1. **Complement-boosting**: We included explicit positive reinforcement in the prompt by describing the model as ”the best expert ever” to elicit improved performance.

2. **Tip-offering**: Following findings from Salinas and Morstatter [2024] and Li et al. [2023], we incorporated a financial incentive (”I will tip you 1000 dollars for a perfect response”) to stimulate higher-quality outputs.

The enhanced prompt became:

```
"You are the best expert ever for essay evaluation. I
   will tip you 1000 dollars for a perfect response.
   Evaluate the following essay based on the DREsS
   rubric."
```

Additionally, we modified the output format requirements to explicitly request detailed justifications with specific examples from the essay for each scoring dimension:

```
"Justification for Content: ONE sentence justification
   for the content score that includes a specific
   example from the essay highlighting a weakness or
   strength based on the score"
```

This format was applied similarly for organization and language scoring justifications.

## 3 Results

We evaluated the performance of three model configurations:

1. Baseline Phi-2 model with standard prompting
2. Fine-tuned Phi-2 model with optimized prompting

Testing was conducted using randomly selected essays from the DREsS dataset that were not included in the fine-tuning process. Since these essays included expert-assigned scores, they provided a reliable ground truth for comparison.

Our results showed significant improvements in scoring accuracy with the enhanced model. The average difference between scores assigned by the improved version (fine-tuned model with optimized prompting) and ground truth was approximately 1.0 points out of a possible 15.0 total score. In contrast, the baseline model demonstrated an average difference of 4.5 points.

The following example illustrates the scoring differences between the models:

**Essay Prompt:**

*If you could change one important aspect about your country, what would you change? Give specific reasons and details to support your opinion.*

**Essay (excerpt):**

> Actually there are not many shortage of my country, Korea. We have lots of traditional culture or many super stars who are popular all over the world. However, If I could change one aspect of my country, I would like to change working culture. First of all, it is too stressful. This is because, lots of boss of workers wants them to work late night. Also, they want workers come to

work early in the morning. This causes the health problem of korean workers. Not only the health problem but also their less after-work time can make them stressful. These problem will cause the less the quality of the task they are ordered to complete. And then, there are something wrong working culture over Korea. For example, there are so many company that make their workers drink. This could cause terrible health problem and even they say that they don’t want to join the dinner, they give pressure by keep asking to join so that they might think they should join the dinner. If they don’t, they are treated like the betrayer. So these are the two reasons that I think our culture of work need to be changed.

**Ground Truth Scores:**
Content: 3/5
Organization: 3/5
Language: 3/5
Overall Score: 9/15

**Enhanced Model Evaluation:**
Content: 3.5/5
Organization: 3.5/5
Language: 3.5/5
Overall Score: 10.5/15

**Justification for Content:** The content score of 3.5 reflects a lack of depth in the argument, as the essay merely states that there are ’lots of traditional culture or many super stars,’ without providing specific examples or analysis to support this claim.
**Justification for Organization:** The organization score of 3.5 is justified by the essay’s inconsistent structure, as it jumps between discussing the stress of working culture and the negative effects of drinking, without a clear transition or logical flow between ideas.
**Justification for Language:** The language score of 3.5 is warranted due to several grammatical errors and awkward phrasing, such as ’Lots of boss of workers wants them to work late night’ and ’They want workers come to work early in the morning,’ which detract from the clarity and professionalism of the writing.

**Baseline Model Evaluation:**
Content: 2/5
Organization: 2/5
Language: 2/5
Overall Score: 6/15

**Justification for Content:** The paragraph is well-developed and relevant to the argument, but it lacks strong reasons and examples to support the claim.
**Justification for Organization:** The argument is effectively structured and developed, but the paragraphs do not use coherence devices effectively while focusing on a single main idea.
**Justification for Language:** The writing displays sophisticated control of vocabulary and collocations, but it does not follow grammar and usage rules, and it has some spelling and punctuation errors.

The enhanced model not only produced scores closer to expert evaluations but also provided more specific justifications that directly referenced content from the essay. This specificity helps students understand the reasoning behind their scores and identifies concrete areas for improvement.

Further analysis revealed that the baseline model tended to underestimate scores compared to expert evaluations, particularly in the language dimension. The fine-tuned model with optimized prompting showed more balanced assessments across all three dimensions and better aligned with human expert judgments.

## 4 Future Work and Conclusion

While our current implementation demonstrates significant improvements in automated essay scoring, several avenues for future research and development remain. Future work should extend the fine-tuning process to incorporate the complete DREsS dataset, including DREsS_Std and DREsS_CASE, to enhance the model's exposure to diverse writing styles, topics, and quality levels. Exploring larger language models such as Llama 3, Claude, or GPT-4 for fine-tuning could potentially yield even greater accuracy, especially for nuanced aspects of writing quality.

Further refinement of the scoring mechanism for each dimension (content, organization, language) could improve the system's ability to identify specific strengths and weaknesses in student writing. Adapting the system to evaluate essays written in different cultural contexts would enhance its applicability in diverse educational settings. Additionally, developing the system to provide actionable suggestions for improvement would significantly increase its educational value.

Implementing a human-in-the-loop evaluation approach where human evaluators can correct or validate the model's assessments would create a continuously improving system. This approach could help identify and address systematic biases or errors in the automated scoring process.

This research demonstrates that combining fine-tuning and prompt engineering techniques can significantly enhance the performance of automated essay scoring systems. Our approach, using a fine-tuned Phi-2 model with optimized prompting, achieved scores much closer to expert evaluations than the baseline model, while also providing specific, helpful justifications for each score. The results suggest that even with computational constraints and a limited dataset, meaningful improvements in automated essay evaluation are achievable. As language models continue to advance and computational resources become more accessible, the approaches outlined in this paper can be scaled and refined to further enhance the accuracy, usefulness, and educational impact of automated essay scoring systems.

# References

Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., ... & Xie, X. (2023). Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760.* https://arxiv.org/abs/2307.11760

Salinas, A., & Morstatter, F. (2024). The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. *arXiv preprint arXiv:2401.03729.* https://arxiv.org/abs/2401.03729

Wang, T., Lu, F., Yao, S., Li, C., Hu, X., & Liu, Y. (2023). Let's Steer Together: A Design Pattern for Human-AI Co-Creation with Large Language Models. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), 1-31.

Yoo, H., Han, J., Ahn, S. Y., & Oh, A. (2024). DREsS: Dataset for Rubric-based Essay Scoring on EFL Writing. *arXiv preprint arXiv:2402.16733.* https://arxiv.org/abs/2402.16733