RESEARCH FORUM
19TH UGRF

2025-AIS411-7

Automated Essay Scoring with LLaMA 3.2: Exploring Fine-Tuning and Prompt Engineering Techniques

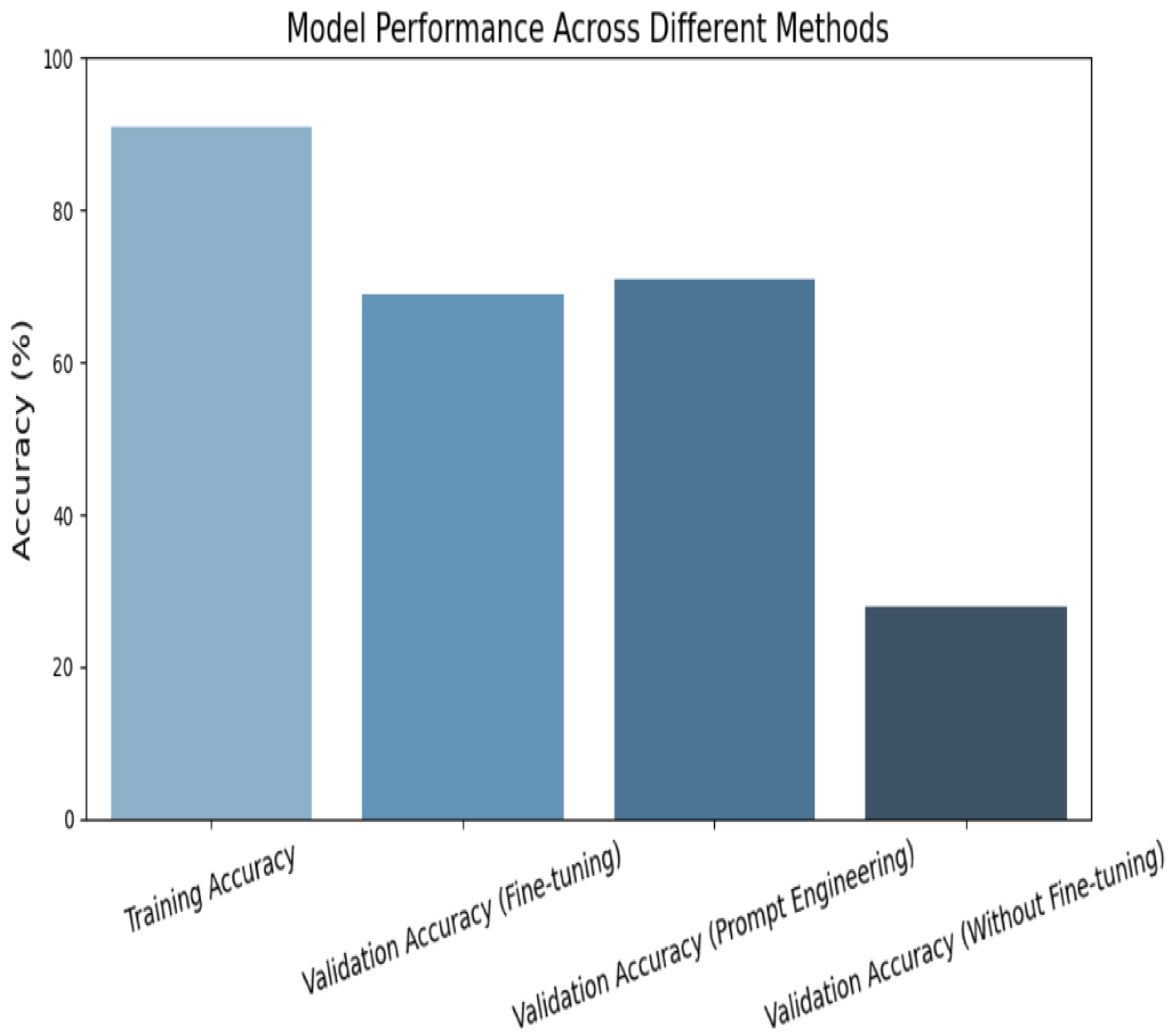
Uomna Hesham- Nada Atef- Farah Fawki- Mai Elgazzar

Abstract

This paper explores Automated Essay Scoring (AES) using LLaMA 3.2 across three configurations: a baseline pre-trained model, a fine-tuned model, and a fine-tuned model with prompt engineering. Results indicate that prompt engineering enhances validation accuracy, emphasizing its role in optimizing AES systems for educational use. Challenges such as fairness, bias, and scalability are discussed alongside potential solutions, with recommendations for further research to address these issues and improve AES effectiveness.

Introduction

Automated Essay Scoring systems evolved from a feature-based approach to a large language model (LLM) approach to provide a more holistic and scalable grading strategy. One example is the LLaMA 3.2 system, which now increases the reliability of AES through tuning and prompt making. The research also establishes how LLMs can improve the consistency and granularity of the generated feedback, making it a platform to deliver equitable and efficient educational assessments. AES hence becomes better placed to consider different writing styles and linguistic complexities, thus making it a fairer and more



Methodology

This study developed an Automated Essay Scoring (AES) system using LLaMA 3.2 to grade essays on a 1-to-6 scale, focusing on accuracy and scalability. Three approaches were implemented:

Baseline Model: A pre-trained LLaMA 3.2 model was used without any modifications to establish a performance baseline.

Fine-Tuned Model: The model was trained with labeled essay datasets to adapt it for the specific scoring task, improving its ability to evaluate essays based on defined rubrics.

Fine-Tuned Model with Prompt Engineering: Tailored prompts were designed to provide clear instructions, optimizing the model's performance in scoring essays more precisely.

Results

The baseline model achieved 28% validation accuracy, highlighting its limitations. Fine-tuning improved validation accuracy to 69%, with training accuracy at 91%, indicating effective learning. Adding prompt engineering further enhanced validation accuracy to 71%, demonstrating its importance in improving scoring precision and reliability.

Conclusion & future work

This study highlights the effectiveness of fine-tuning and prompt engineering in improving Automated Essay Scoring (AES) accuracy using LLaMA 3.2. Future work will address fairness and bias, expand datasets, and introduce multilingual support. Refining prompts and incorporating metrics like creativity and argument strength will enhance the system's fairness and comprehensiveness.

References

[1] W. Xu, R. Mahmud, and W. L. Hoo, ”A Systematic Literature Review: Are Automated Essay Scoring Systems Competent in Real-Life Education Scenarios?” IEEE Access, vol. 12, pp. 77639–77645, 2024.
[2] B. B. Klebanov and N. Madnani, *Automated Essay Scoring*, Morgan & Claypool Publishers, 2024