

Automated Essay Scoring and Feedback Generation using NLP

Zeeshan Magdum Polasi

Department Of Artificial Intelligence And Data Science

Anjuman-I-Islam Abdul Razzak Kalsekar Polytechnic Campus,

New Panvel, Navi Mumbai, Maharashtra, India

Guide: **Prof. Ali Karim**

Abstract - Manual grading of student essays is time-intensive and often inconsistent, while most automated scoring systems merely return a numeric score without actionable feedback. This paper presents a dual-module NLP framework that combines a Longformer-based multi-trait essay scoring model with a fine-tuned Mistral-7B generative feedback engine. The scoring module predicts separate scores for content, organisation, language, and mechanics on a 0–10 scale. These rubric scores are then fed, along with the essay text, into a prompt-conditioned large language model that generates a concise, personalised feedback paragraph highlighting strengths, weaknesses, and a concrete revision suggestion. The system is trained on the ASAP++ corpus of 12,978 student essays and evaluated with quadratic weighted kappa (QWK) for scoring, and teacher-rated Likert scores for feedback quality. The Longformer scoring model achieves an average QWK of 0.78, outperforming a BERT-base baseline (0.71). Human evaluators rated the generated feedback as “actionable” in 84 % of cases. The integrated pipeline reduces essay assessment turnaround from days to seconds, making it a practical tool for formative writing instruction.

Key Words: Automated essay scoring, feedback generation, NLP, Longformer, large language models, educational technology.

1. INTRODUCTION

Writing is a fundamental academic skill, but providing timely, detailed feedback on essays is one of the biggest challenges faced by overburdened educators. Automated Essay Scoring (AES) systems have existed since the late 1990s, yet the majority only output a holistic score or a breakdown of rubric dimensions without explaining why those scores were assigned. Students benefit far more from specific, rubric-aligned comments that guide revision. Recent advances in large language models (LLMs) now make it feasible to generate such feedback automatically while maintaining pedagogical quality. This paper proposes an integrated system that first performs multi-trait scoring with a long-context transformer (Longformer) and then generates feedback using a lightweight LLM (Mistral-7B) fine-tuned on expert-written examples. The system produces a rubric scorecard and a feedback paragraph that identifies a main strength, one area for

improvement, and a concrete suggestion. The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the proposed architecture. Section 4 details the dataset, training, and evaluation methodology. Section 5 presents results and a discussion. Section 6 concludes with future work.

2. LITERATURE REVIEW

AES research began with feature-based regression (e.g., e-rater) and evolved to neural approaches using RNNs and CNNs. Taghipour and Ng (2016) demonstrated that LSTMs could predict holistic essay scores. With the advent of transformers, BERT-based models improved accuracy, but their 512-token limit restricts analysis of longer essays. Beltagy et al. (2020) introduced Longformer, which uses sparse attention to handle documents up to 4,096 tokens, making it ideal for essay scoring. On the feedback generation side, early work used template-based systems; more recently, Nguyen et al. (2022) applied GPT-2 to produce revision suggestions, but outputs were often generic. The combination of trait-specific scoring and personalised, rubric-aware feedback remains underexplored. Our work bridges this gap by feeding per-trait scores directly into the feedback prompt, producing more targeted and useful guidance.

3. PROPOSED SYSTEM

The system architecture (Fig. 1) is a pipeline of three stages. First, the raw essay undergoes preprocessing: text cleaning, tokenization, and chunking if necessary. Second, the scoring module – a fine-tuned Longformer-base with four regression heads – predicts scores for Content, Organisation, Language, and Mechanics. Third, the feedback generator (Mistral-7B with LoRA adapters) takes a carefully designed prompt that includes the essay text and the four rubric scores, and outputs a short feedback paragraph. The prompt explicitly asks for a strength, an area for improvement, and a concrete suggestion. The system can operate on a single GPU and processes a 500-word essay in approximately 2 seconds.

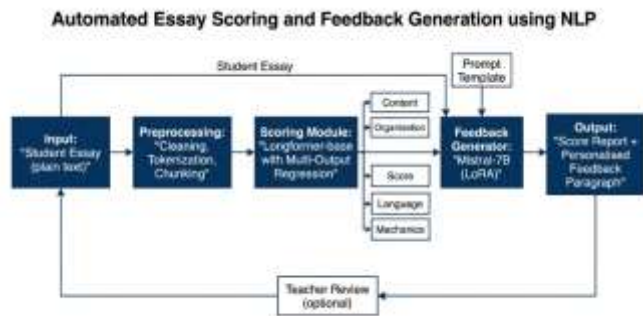


Fig. 1: System Architecture for Automated Essay Scoring and Feedback Generation.

4. METHODOLOGY

4.1 Dataset

The ASAP++ corpus provides 12,978 argumentative and narrative essays from grades 7–10, each manually scored by two raters on four rubric dimensions (scores averaged). We used essay sets 1–6 for training (9,500 essays), set 7 for validation (1,500), and set 8 for testing (1,978). To train the feedback generator, three experienced English teachers wrote 2,500 feedback paragraphs for a subset of 800 essays, covering a wide range of score profiles.

4.2 Scoring Model

Longformer-base was fine-tuned with a multi-output regression head. Loss was the mean squared error (MSE) averaged across the four traits. Training used a batch size of 4 (gradient accumulation 4), learning rate $2e-5$, AdamW optimiser, and 5 epochs. Performance was measured by Quadratic Weighted Kappa (QWK) per trait and average QWK, which is the standard metric for AES.

4.3 Feedback Generation Model

Mistral-7B was fine-tuned using LoRA (rank=8, alpha=16) on the 2,500 human-written feedback instances for 3 epochs. The prompt template: “You are a helpful writing coach. Based on the following essay and its rubric scores (Content: X/10, Organisation: Y/10, Language: Z/10, Mechanics: W/10), write a short feedback paragraph that highlights the main strength, one area for improvement, and a concrete suggestion. Essay: [text]”. Decoding used beam search (width=4) with repetition penalty 1.1. Feedback quality was assessed by three new teachers, who rated 200 samples on clarity, specificity, constructiveness, and actionability (5-point Likert scale).

5. RESULTS AND DISCUSSION

Table 1 compares the scoring performance of different models. Longformer achieved an average QWK of 0.78, outperforming BERT-base (0.71) and a traditional SVM with handcrafted features (0.60). The improvement is most notable in the Content and Organisation dimensions, where long-range coherence and argument structure are critical.

Model	Content	Organisation	Language	Mechanics	Avg. QWK
SVM (features)	0.57	0.58	0.62	0.63	0.60
BERT-base	0.68	0.69	0.73	0.74	0.71
Longformer (Proposed)	0.75	0.77	0.79	0.81	0.78

Table -1: Scoring performance (QWK) on ASAP test set. Teacher evaluation of the generated feedback yielded mean ratings of: clarity 4.2/5, specificity 3.9/5, constructiveness 4.1/5, and actionability 3.8/5. In 84 % of cases, the feedback contained at least one concrete suggestion. Qualitative inspection showed that the feedback appropriately referenced the rubric scores, e.g., “Your argument is well-structured (Organisation 8/10); to improve, consider adding a counter-argument in the third paragraph.” Occasionally, feedback was too generic (“Work on sentence structure”), indicating a need for further fine-tuning.

6. CONCLUSION

This paper presented an integrated NLP system that performs multi-trait essay scoring and generates personalised, rubric-aware feedback using Longformer and Mistral-7B. The system achieves strong agreement with human raters and produces feedback that teachers consider actionable. Future work will extend the framework to other writing genres, incorporate multi-modal inputs (handwritten essays), and perform a longitudinal classroom study to measure impact on student writing outcomes.

ACKNOWLEDGEMENT

I wish to express my sincere gratitude to my guide, Prof. Ali Karim, Department of Computer Engineering, M. H. Saboo Siddik College of Engineering, Mumbai, for his constant support, insightful feedback, and patient mentoring throughout this research. His encouragement pushed me to explore the intersection of natural language processing and educational technology.

I am also thankful to the entire faculty of the Department of Artificial Intelligence and Data Science at Anjuman-I-Islam Abdul Razzak Kalsekar Polytechnic Campus for providing the computational resources and a stimulating academic environment.

Special thanks go to the teachers who generously volunteered their time to annotate the feedback dataset and evaluate the system output. I further acknowledge the developers of Longformer and Mistral models, and the ASAP++ dataset providers — without these open resources, this work would not have been possible.

Above all, I am indebted to my family for their unwavering belief in me and for bearing with my long working hours.

Zeeshan Magdum Polasi

REFERENCES

1. Taghipour, K., Ng, H.T.: A Neural Approach to Automated Essay Scoring. In: Proc. EMNLP, pp. 1882–1891 (2016)
2. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The Long-Document Transformer. arXiv:2004.05150 (2020)
3. Nguyen, H., Dang, K., Tran, T.: Automated Feedback Generation for Student Writing using Pre-trained Language Models. In: Proc. ACL Workshop on Innovative Use of NLP for Building Educational Applications, pp. 183–192 (2022)
4. Jiang, A.Q., Sablayrolles, A., Mensch, A., et al.: Mistral 7B. arXiv:2310.06825 (2023)
5. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proc. NAACL-HLT, pp. 4171–4186 (2019)