DOI: 10.54254/2755-2721/142/2025.KL22313

Research on AI-Powered Essay Writing Assistant with Contextual Revision System

Yicen Liu

Faculty of Computer and Mathematical Sciences, The Hong Kong Polytechnic University, Hong Kong, China
maths.liu@connect.polyu.hk

Abstract: This report presents an AI-powered essay writing assistant that combines three advanced modules: a fine-tuned DeepSeek-R1-Distill-Llama-8B language model for contextual revisions, retrieval-augmented generation (RAG) for evidence-based suggestions, and a DeBERTaV3-based ordinal regression model for essay scoring. By combining the two models, the system demonstrates high grammatical accuracy in revisions and achieves a satisfactory value of quadratic weighted Kappa in scoring while providing detailed suggestions for further improvement, significantly outperforming baseline approaches. Through systematic integration of these components, the writing assistant offers holistic writing support while maintaining computational efficiency. It can provide real-time feedback, contextual suggestions, and rubric-aligned scoring, significantly enhancing the writing process as well as the user experience. The work may offer a method of using AI techniques in essay writing, demonstrating how context-aware systems can enhance writing quality while maintaining computational practicality. By streamlining the writing and grading processes, this project has the potential to boost productivity and improve learning outcomes, making it a valuable tool in modern educational environments.

Keywords: AI-powered, essay writing assistant, contextual revision system.

1. Introduction

The development of AI-powered writing assistants has gained worldwide attention across various academic fields, addressing many challenges in text composition and quality assurance. Pereira and Barcina demonstrated the value of automated quality control in academic writing through Ikastenbot, a chatbot that detects and revises common errors in technical reports, including plagiarism, poor reference, and grammatical issues [1]. Their work highlights the critical need for tools that can identify writing flaws, particularly for inexperienced students. With sufficient funds and resources, Shi et al. expanded the scope of writing assistance with Effidit, an AI platform offering advanced functions such as context-aware text polishing, paragraph paraphrasing, and retrieval-based sentence completion [2]. While Effidit represents a leap forward in functional diversity, its focus on general writing tasks leaves gaps in domain-specific academic support, particularly in rubric-aligned evaluation and ethical safeguards. Recent work by Han et al. addresses some of these limitations in language learning scenarios through their framework in English as a Foreign Language (EFL) education [3]. By integrating pedagogical principles to evaluate student-LLM interactions, their study establishes metrics for assessing feedback quality and learning outcomes, emphasizing the need for AI tutors to balance technical accuracy with educational efficacy.

These AI assistants have shortages. For instance, they operate in isolation, providing fragmented feedback that fails to address the holistic needs of writers. Moreover, the increasing demand for personalized and context-aware writing assistance has highlighted the limitations of rule-based systems [4]. Recent advancements in natural language processing (NLP), particularly in large language models (LLMs), offer new opportunities to develop intelligent writing assistants [5]. These systems can provide real-time feedback, contextual suggestions, and essay scoring, significantly enhancing the writing process. Synthesizing insights from these approaches, this project aims to design an integrated essay writing assistant that combines fine-tuned language models, Retrieval Augmented Generation (RAG), and ordinal regression-based scoring to address these challenges.

2. Methodology

The structure of the essay writing assistant comprises three core components, as summarized below. The revision and suggestion-providing system is a fine-tuned DeepSeek-R1-Distill-Llama-8B model, which is particularly useful for grammar correction and style improvement. Retrieval-augmented generation is an important module. It is based on topic-specific scholarly papers that contain domain-specific terminology to provide evidence-based suggestions. The scoring module is a DeBERTaV3-based ordinal regression model for rubric-aligned essay scoring. These components enable real-time feedback and iterative improvement of essays while combining cutting-edge AI technology to enhance the learning experience for students and educators [6].

2.1. DeepSeek-R1-Distill-Llama-8B Instruction Tuning

Low-Rank Adaptation (LoRA) is employed to preserve the base model's general capabilities while specializing it for academic writing tasks. To achieve the target module, focusing on attention projection matrices and feed-forward network parameters, which can capture task-specific attention patterns while maintaining linguistic knowledge, is implemented throughout the process.

2.2. Hyperparameter Optimization:

Low-Rank Adaptation:

$$\Delta W = BA \tag{1}$$

Where:

$$B \in \mathbb{R}^{d\times r}\ A \in R^{r\times k}\ (\mathrm{r} = 16) \tag{2}$$

Original weight matrix: $\boldsymbol{W} \in \mathbb{R}^{\boldsymbol{d}\times\boldsymbol{k}}$. Final weight: $W' = W + \alpha\Delta W$ $(\alpha = 16)$.

The original weight matrix W is part of the parameters of a neural network layer. The idea is to adapt the weight matrix W using a low-rank update. This is done by decomposing the update ΔW into two smaller matrices B and A. The rank r is a hyperparameter that determines the size of the matrices B and A. The final weight matrix W′ is computed by adding a scaled version of the low-rank update to the original weight matrix and α is a scaling factor (another hyperparameter) that controls the magnitude of the update.

Memory efficiency is also important for training [5]. The 4-bit quantization reduces VRAM consumption by approximately 60% while maintaining high full-precision performance through QLoRA compensation. Some other critical implementation details during the finetuning will be analyzed. As for the foundation part, Masked language modeling on the corpus is completed, enabling the model to learn contextual representations of words and capture nuances and meanings that depend on surrounding words. The main part is Supervised fine-tuning with the chain of thought prompting, which can reason more like humans, providing both accurate answers and the logical pathways that lead to those answers. Some technical methods during the training process will be discussed. Firstly, Gradient Checkpointing manages memory usage in deep learning and enables larger batch sizes through memory recomputation. Secondly, Flash Attention 2 improves both speed and memory efficiency, reducing attention computation. Finally, the Warmup Strategy is applied. At the start of training, the learning rate is set to a very low value, and over the first 15% of the total training steps, the learning rate is gradually increased linearly until it reaches the target learning rate. prevents early divergence and enhances the stability and performance of neural network training.

2.3. Ordinal Regression Framework

The system employs ordinal regression to model the inherent ordered nature of essay scores (1-5), addressing limitations of standard multi-class classification. The scoring model extends standard classification through cumulative probability modeling:

Mathematical Formulation:

$$\mathit{For\ score\ classes}, \boldsymbol{c} \in \{\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}, \mathbf{6}\}:$$

$$P(y \le c \mid x) = \sigma(\theta c - f(x)) \tag{3}$$

Loss Function:

$$\mathcal{L} = -\frac{1}{5N}\sum_{i=1}^{N}\sum_{c=1}^{5}[yi \le c]log\left(P(\,y \le c \mid xi\,)\right) + [yi > c]log\left(1 - P(\,y \le c \mid xi\,)\right) \tag{4}$$

Where: $f(x)$: DebertaV3 embedding output, σ: Sigmoid activation, θc: Learnable class thresholds, N: Batch size.

Here, the probability P represents the probability that the true label y for input x is less than or equal to class c and is modeled using a sigmoid function. The sigmoid activation function maps any real-valued number into the range (0, 1). The DeBERTaV3 model provides an embedding of the input. Additionally, the learnable thresholds for each class help determine the boundaries between the ordered classes. As for the loss function, it is designed to handle the ordinal nature of the problem by considering the cumulative probabilities. The indicator function [yi≤c] is 1 if the true label yi is less than or equal to c. Similarly, [yi>c] is 1 if the true label yi is greater than c. This loss function penalizes the model based on how well it predicts the cumulative probabilities for each class threshold.

In this approach to ordinal regression, the process begins with a label transformation where scores are converted into cumulative binary labels. For instance, a score of 3 is transformed into the binary vector [1,1,0,0,0], which indicates the thresholds that have been surpassed. The structure of the model is built upon the DeBERTaV3-extra-small, which consists of 91 million parameters, providing a robust foundation for feature extraction. The model is trained using a binary cross-entropy loss function applied across these thresholds, ensuring that the predictions align with the ordinal nature of the dataset. To evaluate the model's performance, a specific implementation of the Weighted Kappa metric is employed. This metric penalizes large discrepancies between predicted and true scores, with quadratic weighting that aligns with educational assessment standards, ensuring the accuracy of the predictions.

Table 1: Key Hyperparameters

Parameter	Value	Rationale
Sequence Length	512	Captures full essay context
Batch Size	32	Balances VRAM usage and gradient stability

Table 1: (continued).

Parameter	Value	Rationale
Learning Rate	5×10^{-6}	Prevents catastrophic forgetting
Mixed Precision	FP16	Around 37% faster training

As shown in Table 1, the selection for the parameter of the model implementation includes several key strategies to optimize performance and efficiency. The learning rate is managed using cosine learning rate annealing, which peaks at 6×10^{-6} over the three epochs. This approach helps in gradually adjusting the learning rate to improve convergence. Additionally, early stopping is employed, with model checkpointing based on validation Kappa, ensuring the model retains the best weights during training. To further enhance the training process, TensorFlow dataset optimization techniques such as caching and prefetching are applied, resulting in an improvement in throughput. Dynamic padding is used to accommodate sequences up to 512 tokens, which helps in efficiently managing varying input lengths. Throughout the training dynamics, careful attention is paid to convergence behavior, ensuring that the model learns effectively and achieves optimal performance on the task at hand.

2.4. Retrieval-Augmented Generation (RAG)

The RAG system leverages a knowledge base of academic texts, including arXiv papers and domain-specific PDFs. These texts were preprocessed into 500-token chunks and indexed using FAISS for efficient retrieval. For a given query, the system computes cosine similarity between the query embedding and chunk embeddings to retrieve the most relevant passages. The embeddings were generated using the text-embedding-3-small model from Open AI. After these steps, the retrieved passages are used as context for the fine-tuned language model, enabling evidence-based suggestions [7]. The generation process is guided by a structured prompt template:

2.5. Dataset

All datasets used for training in the project are open sources, accessible from arXiv, Hugging Face, and Kaggle, respectively [8-10]. Scholarly papers from the arXiv database can be downloaded in portable document format by selecting specific keywords and utilizing Python libraries to transform them into combined texts [8]. LongWriter-6k, a dataset that contains 6,000 Supervised Fine-Tuning data with ultra-long output ranging from 2k-32k words in length (both English and Chinese) can support training LLMs to extend their maximum output window size to more than 10,000 words while increasing the ability of generating texts with logical thinking [9]. The dataset on Kaggle includes high-quality, realistic classroom writing samples, from various economic and geographic backgrounds, which may reduce the risk of algorithmic bias [10]. In addition, the dataset for evaluation includes 100 essays from the International English Language Testing System (IELTS) and the Chinese National College Entrance Examination.

3. Experimental Results

Table 2: Correction Rate

Evaluation Metric	Value
Grammar Error Correction	95%
Style Improvement Rate	80%
Useful Suggestion Precision	89%

Table 3: Confusion Matrix Analysis:

True\Pred	1	2	3	4	5	6
1	69%	28%	3%	0%	0%	0%
2	11%	73%	16%	0%	0%	0%
3	2%	19%	64%	15%	0%	0%
4	0%	3%	17%	60%	15%	5%
5	0%	0%	7%	12%	78%	3%
6	0%	0%	0%	10%	40%	50%

The fine-tuned DeepSeek-R1-Distill-Llama-8B model is tested on 100 essays from examinations. Table 2 indicates advanced pattern recognition for syntactic and morphological errors and a high rate of writing style improvement. Dominant error types corrected include subject-verb agreement, tense inconsistency, and specific word misuse. Moreover, the correction rate in short essays could achieve nearly 100% while in longer essays the fine-tuned model cannot discover all the mistakes, suggesting there is still space for improvement. The style improvement rate reflects strong contextual awareness, particularly in voice conversion, redundancy reduction, and academic tone enhancement. Meanwhile, the useful suggestion precision shows a relatively high user adoption rate for AI-proposed edits, with structural reorganization and citation formatting based on the contexts.

As for Table 3, it shows the scoring model achieves satisfactory ordinal consistency while exhibiting class-boundary challenges. About 35% of essays are misclassified and this may be due to overlapping rubric criteria. Focusing on domain-adaptive scoring thresholds and enhancing essay analysis can reduce these errors [11].

4. Discussion

The development of AI-powered essay writing assistants marks a significant transformation in academic support tools, leveraging advanced natural language processing and domain-specific knowledge retrieval [12]. These systems address critical challenges in academic writing by offering real-time grammatical corrections, contextual suggestions, and rubric-aligned evaluations. For example, many AI assistants utilize retrieval-augmented generation (RAG) systems to enhance suggestion quality by grounding feedback in a vast database of academic papers, ensuring that recommendations adhere to academic conventions [13]. Similarly, the ordinal regression approach used in scoring models, as the project implemented in DeBERTaV3, effectively captures the hierarchical nature of rubric-based assessments, achieving a relatively high quadratic weighted Kappa, which surpasses traditional multiclass methods.

Essay writing is a crucial method for assessing student learning and performance, but manually revising them can be time-consuming for educators [14]. The systematic fine-tuning approach illustrates how modern parameter-efficient methods can create high-performance writing assistants. The finetuned DeepSeek-R1 model achieves human-aligned suggestion quality through progressive domain adaptation, while the DeBERTaV3 scoring system demonstrates superior rubric consistency through its ordinal architecture. This dual optimization strategy provides a template for developing specialized Natural Language Processing (NLP) systems that balance accuracy with computational efficiency. Some functions can be added in future versions. The integration of ethical safeguards, such as plagiarism detection and AI-generated content flagging, as seen in OUTFOX, addresses growing concerns about academic integrity while maintaining utility [15]. Thus, these functions may be taken into consideration as improvement.

Emerging studies highlight the importance of workflow efficiency. For instance, some AI tools reduce the obstacle of users through tailored prompts and utilize autocomplete features to accelerate drafting [2-5]. These advancements mirror findings in human-AI collaboration research, where context-aware systems and retrieval-augmented generation, just as what has been done in this project, improve comprehension by linking suggestions to source materials. Nevertheless, the risk of over-reliance on these tools necessitates a balanced design, as emphasized by the hybrid model, which combines AI analysis with human-like argumentation feedback [4].

Future directions should prioritize adaptive personalization, as demonstrated by tools like Effidit that mimic user writing styles and expand multilingual support using some translation frameworks [2]. Additionally, addressing the "cold start" problem through transfer learning is important since specific templates could enhance accessibility for novice researchers [5,14]. By integrating these innovative methods while introducing responsible usage, AI writing assistants can evolve into powerful collaborators rather than mere tools, encouraging higher working efficiency and more educational applications [15].

5. Conclusion

This essay writing assistant represents a significant development in AI-driven writing assistant by incorporating three essential components: a retrieval-augmented generation (RAG) module for context-aware suggestions, a parameter-efficient fine-tuned language model for improving grammar and structure, and an ordinal regression-based scoring system for evaluations. By employing the DeepSeek-R1-Distill-Llama-8B model with Low-Rank Adaptation (LoRA), the system maintains high grammatical accuracy while being computationally efficient, demonstrating that large language models can be effectively tailored to specialized fields without extensive retraining. The DeBERTaV3-based scoring module's quadratic weighted Kappa showcases its superiority over traditional multiclass methods, effectively capturing the writing quality as outlined in educational rubrics. Practical trials revealed significant advantages, such as reducing revision cycles for users and decreasing grading workloads for instructors, highlighting the potential to enhance learning outcomes and teaching efficiency. However, challenges like domain specificity and limited handling of different essay scoring criteria for examinations suggest areas for further improvement. Future versions may consider incorporating multimodal analysis for diagrams and adaptive personalization to accommodate individual writing styles. The essay writing assistant has the potential to provide high-quality writing support, particularly for non-native English speakers and under-resourced institutions. By reducing the time and effort required to improve writing skills and complete essay grading, this writing assistant can enhance productivity and learning outcomes in academic settings.

References

[1] *Pereira J., & Barcina, M.A. (2019). A chatbot assistant for writing good quality technical reports. In Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'19), 59–64.*

[2] *Shi, S., Zhao, E., Tang, D., Wang, Y., Li, P., Bi, W., Jiang, H., Huang, G., Cui, L., Huang, X., Zhou, C., Dai, Y., & Ma, D. (2022). Effidit: Your AI Writing Assistant.*

[3] *Han, J., Yoo, H., Myung, J., Kim, M., Lim, H., Kim, Y., Lee, T.Y., Hong, H., Kim, J., Ahn, S.Y., & Oh. A. (2024). LLM-as-a-tutor in EFL Writing Education: Focusing on Evaluation of Student-LLM Interaction. In Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U), 284–293.*

[4] *Deepika, K., Tilekya, V., Mamatha J., & Subetha, T. (2020). Jollity Chatbot- A contextual AI Assistant. 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 1196-1200.*

[5] *DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.*

[6] *Kostic, M., Witschel, H. F., Hinkelmann, K., & Spahic-Bogdanovic, M. (2024). LLMs in automated essay evaluation: A case study. In Proceedings of the AAAI Symposium Series, 3(1), 143-147.*

[7] Resch, O., & Yankova, A. (2019). Open knowledge interface: a digital assistant to support students in writing academic assignments. In Proceedings of the 1st ACM SIGSOFT International Workshop on Education through Advanced Software Engineering and Artificial Intelligence, 13-16.

[8] Clement, C.B., Bierbaum, M., Keeffe, K.P.O, & Alemi A.A. (2019). On the Use of ArXiv as a Dataset.

[9] Bai, Y.S., Zhang, J.J., Lv, X., Zheng, L.Z., Zhu, S.Q., Hou, L., Dong, Y.X., Tang, J., & Li, J.Z. (2024). LongWriter: Unleashing 10,000+ Word Generation from Long Context LLMs.

[10] Crossley, S., Baffour, P., King, J., Burleigh, L., Reade, W., & Demkin, M. (2024). Learning Agency Lab - Automated Essay Scoring 2.0. https://kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2

[11] Stahl, M., Biermann, L., Nehring, A., & Wachsmuth, H. (2024). Exploring LLM prompting strategies for joint essay scoring and feedback generation. arXiv preprint.

[12] Gruda, D. (2024). Three ways ChatGPT helps me in my academic writing. Nature, 10.

[13] Katsnelson, A. (2022). Poor English skills? New AIs help researchers to write better. Nature 609, 208-209.

[14] Zhao, X. (2023). Leveraging Artificial Intelligence (AI) Technology for English Writing: Introducing Wordtune as a Digital Writing Assistant for EFL Writers. RELC Journal, 54(3), 890-894.

[15] Koike, R., Kaneko, M., & Okazaki, N. (2024). Outfox: LLM-Generated Essay Detection Through In-Context Learning with Adversarially Generated Examples. In Proceedings of the AAAI Conference on Artificial Intelligence, 38(19), 21258-21266.