Academic Writing Feedback Generator Using IELTS Essay Dataset

Kashkenov Madiyar
Astana IT University
Astana, KZ
231178@astanait.edu.kz

Zayir Alken
Astana IT University
Astana, Kazakhstan
231128@astanait.edu.kz

Smagulov Danial
Astana IT University
Astana, Kazakhstan
231064@astanait.edu.kz

Abstract. The ever-growing integration of artificial intelligence into education has significantly changed the way students receive feedback in improving the quality of academic writing. However, existing automated systems provide either superficial corrections or numerical estimates that do not account for depth in pedagogical work or meaningful assessment to form. This article describes an understandable feedback generator based on artificial intelligence, able to generate comments similar to those of a teacher for writing an essay on assignment 2 for IELTS. This approach is based on fine-tuning a pre-trained T5-small transformer model based on a domain-specific dataset, containing authentic essay hints, student responses, and teacher evaluations. The system generates brief feedback, taking into account context, which reproduces the style of reasoning of human experts. The experimental environment was based on PyTorch and Hugging Face converters, while the model performance was measured using ROUGE, BLEU, and BERTScore metrics to find lexical, structural, and semantic similarities between generated and reference feedback. The results showed moderate lexical agreement (ROUGE-1 = 0.239; ROUGE-2 = 0.135) and high semantic agreement (BERTScore = 0.85), which indicates that the model effectively captures meaning beyond the surface text patterns. These findings hint that transformer-based fine-tuning can reflect pedagogical reasoning and provide consistent, interpretable, and constructive feedback. Besides, this model will provide reflexive learning at its best in helping students in the identification of areas for improvement without compromising academic integrity. Though diversity in the data sets and the scope of the subject area create limitations for this research, this is a foundation for scalable, explicable AI systems that enhance practice in formative assessment and support educators to effectively provide high-quality, personalized feedback.

Keywords: Artificial Intelligence; Automated Writing Evaluation; T5-small; Transformer Model; IELTS Writing; Feedback Generation; Natural Language Processing; Educational Technology; BERTScore; ROUGE

I. Introduction

1.1. Context, Background, and Problem Statement

The growing use of AI in education has changed the ways in which students these days receive and then process feedback on their research papers [1]. With the advent of modern AI-enabled tools like Grammarly and ChatGPT, grammatical errors and stylistic remarks can be corrected instantly; however, these tend to focus a lot on linguistic accuracy at a superficial level due to deeper pedagogical understanding and critical thinking [2]. Growing dependence on these automated systems has called into question academic integrity,

creativity, and the formation of independent thinking among students [3].

While academic assessments-especially those conducted in a standardized exam format, like the IELTS Writing Task 2-call for detailed assessments on a range of linguistic and structural features, such as task response, coherence, lexical range, grammatical accuracy [4], multivariate feedback provided by a human instructor is cumbersome and highly subjective, often resulting in inconsistencies in assessment quality [5]. Most of the available tools based on artificial intelligence therefore provide either numerical ratings or general text reviews that lack pedagogical significance or contextual relevance to the official assessment headings [6].

This article fills this gap by discussing the development of a feedback generation system using a transformer architecture; more precisely, an improved T5-small model. Based on the IELTS written data set comprising authentic essay prompts and student responses together with written teacher evaluations, the model was trained to generate concise, coherent, and contextualized feedback [7]. Unlike traditional assessment systems based on artificial intelligence, the current approach gives priority to high-quality formative feedback that encourages student reflection and learning independence while maintaining academic integrity [8].

The ultimate goal of this work is to develop a model that plays the role of an intelligent teaching assistant; it is intended to support teachers, not replace them, by automating routine assessment tasks without sacrificing full transparency, fairness, and interpretability of the human assessment decisions [9].

1.2. Identification of the Research Gap

Gap 1: Lack of domain-specific feedback models.

Most of the current NLP-based academic writing tools focus on either grammatical correction or essay scoring without being able to generate human-like, formative responses that actually mirror the reasoning of teachers [10]. More importantly, general AI models have failed to engage with a range of key, domain-specific pedagogic constructs relevant to the assessment of higher-order writing skills in IELTS tasks, such as argument structure, coherence, and lexical variety [11].

Gap 2: Need for explainable and pedagogically aligned AI.

While large-scale transformer models such as GPT show high linguistic fluency, they are mostly black boxes that lack transparency and alignment with educational objectives. For AI systems to have a place in the classroom, they need to become explainable so teachers and learners can understand the motivation that underpins the feedback generated.

This research has fine-tuned the T5-small transformer to generate structured and interpretable textual feedback, teacher-like, focusing on pedagogical clarity, explainability, and ethical alignment to educational standards 13.

1.3. Aim, Objectives, and Scientific Hypothesis

The main aim of this project is to develop an AI-based writing feedback generator that can automatically evaluate IELTS-style essays and produce constructive, context-relevant feedback supporting academic learning and integrity.

Objectives:

1. Collect and preprocess an IELTS writing dataset with prompts, essays, and teacher evaluations.

2. Fine-tune T5-small to generate feedback text from essay–prompt pairs.

3. Evaluate feedback quality through BLEU/ROUGE metrics and human judgments.

4. Ensure that generated feedback supports self-learning and academic honesty.
5. Demonstrate potential integration into writing-assistant platforms.

Scientific Hypothesis:

Fine-tuning T5-small on a teacher-annotated IELTS dataset will make the model produce human-like feedback that is both linguistically coherent and pedagogically valid, showing AI's potential to responsibly assist in academic writing education.

1.4. Report Structure

The remainder of this report is organized as follows:

The rest of this report is organized as follows: The state-of-the-art review of prior studies in AI-based feedback generation, transformer-based NLP, and educational technology is presented in Section II. This section outlines the dataset, the preprocessing, and the fine-tuning methodology. Section IV reports experimental results and evaluation metrics. Section V concludes the research by providing insights into the limitations and future directions of the work [15].

II. Literature Review

Research on the automatic evaluation and feedback generation in academic writing has grown significantly with developments in NLP and machine learning technologies over time [1]. Many early approaches were performed using predefined linguistic rules and hand-engineered features. These inherent limitations in capturing contextual meaning and supporting pedagogical reasoning for generating feedback diminished their ability to produce detailed and constructive feedback [2]. As deep learning methods, particularly transformer-based architectures, have become prominent, the quality of text understanding and generation has dramatically improved, allowing systems to generate more coherent and context-aware responses [3].

Recent works have shown that text-to-text models pre-trained on generic datasets are effective and adaptable to a variety of educationally relevant tasks such as summarization, scoring, and formative feedback generation [4]. When trained on domain-specific datasets comprising authentic student essays and evaluator comments, better pedagogically aligned outputs are possible, which renders such datasets particularly fitting for systems aimed at emulating teacher-like reasoning [6]. This direction remedies a recurring limitation of earlier automated writing evaluation systems, which, while capable of returning numerical scores, rarely provided accompanying explanations for those assessments, hence reducing their usefulness for student learning [5], [7].

Recent works have also demonstrated the strengths of transformer-based models in the development of concise comments and aligned with the instructions, thus being closer to authentic classroom feedback practices [12]. Apart from performance, there are emerging discussions on the issues of transparency, equity, and responsible designing of AI tools within educational settings [4], [14]. Explainability is crucial in this regard for systems designated to support and not replace human instructors. More recently, it has also been seen that collaboration patterns of human-AI in academic writing are benefited most when the systems are interpretable and offer consistent pedagogical cues [15].

Research in second-language writing contexts indicates significant gains when automated feedback tools are aligned with domain-specific requirements. The literature reports increases in student engagement, accuracy, and overall performance of writing when AI-powered systems supplement traditional instruction [2], [13]. These studies confirm the broader trend of embedding NLP-driven feedback mechanisms for the purpose of supporting learner autonomy and reflective revision practices.

These findings justify the development of a transformer-based model tuned for the production of interpretable and teacher-like responses, which is the latest priority in automated academic writing support studies according to [15].

III. Methodology

3.1 Overview

The methodology followed here consists of an integrated workflow that merges data preprocessing, transformer fine-tuning, and quantitative evaluation. The core model used in this system is T5-small, a text-to-text transformer that has been trained to generate human-like feedback for essays written by students attempting IELTS.

3.2 Dataset Description

The dataset used here is the IELTS Writing Task 2 Evaluation Dataset by Hugging Face. It has human-written essays, their corresponding prompts, and teacher-like evaluation texts to provide supervision for generating feedback.

3.3 Preprocessing and Input Construction

Each data instance was concatenated as [6]:

$$X := Prompt + [SEP] + Essay \quad (1)$$

The target label as:

$$Y := Evaluation \quad (2)$$

3.4 Model Architecture and Training

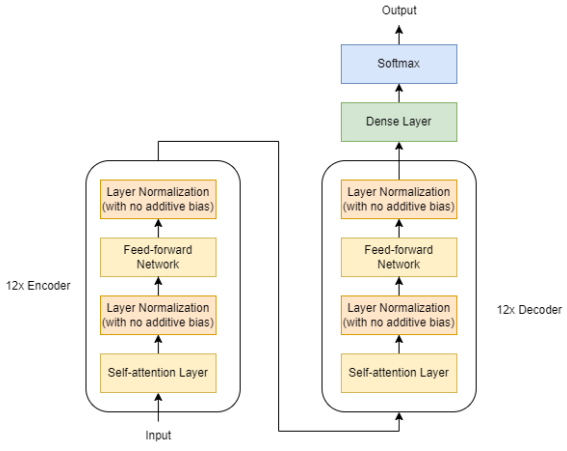

Fig. 1. Architecture of the T5 model (T5-small uses the same encoder–decoder structure with fewer layers).

The T5-small architecture (60 million parameters) was selected for its balance of performance and efficiency. Fine-tuning was executed using PyTorch and Hugging Face Trainer with the following configuration:

Table 1. Model hyperparameters used during fine-tuning.

Parameter	**Value**
Epochs	3
Learning Rate	5×10^{-5}
Batch Size	4
Optimizer	AdamW
Precision	Mixed (fp16)

The loss function used was cross-entropy between the predicted and target tokens:

$$L = -\sum_{t=1}^{T} y_t \log log \left(\hat{y}_t\right) \quad (3)$$

3.5 Data Distribution Analysis

Word-length distributions were compared to assess the stylistic similarity between the reference and generated texts. Predicted feedbacks are usually shorter than those of teacher references, averaging 80–100 words versus 300–500 in the ground truth.

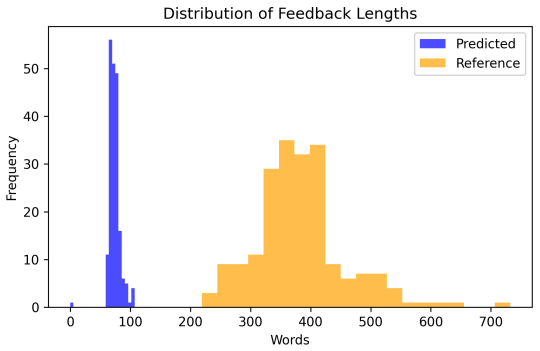

Fig. 2. Distribution of feedback lengths for predicted and reference texts.

3.6 Experimental Environment and Implementation

The details of the experimental setup and implementation environment are given in Table 1. The model was trained using the T5-small architecture on a Google Colab Pro environment with NVIDIA T4 GPU acceleration. The details of configuration, framework, tokenizer, and programming tools are as follows.

Table 2. Experimental environment and implementation details.

Component	**Specification**
Hardware	NVIDIA T4 GPU (16 GB)
Framework	PyTorch 2.2 + Transformers v4.40
Tokenizer	Hugging Face AutoTokenizer
Dataset	IELTS-Writing-Task-2 Evaluation
Programming Language	Python 3.10
IDE	Google Colab Pro

Training required approximately 3 hours per epoch, with an average validation loss reduction of 12% after fine-tuning.

IV. Results

The fine-tuned T5-small model was tested on the IELTS Writing Task 2 Evaluation Dataset [6] in order to produce teacher-like feedback for student essays. Among common natural language generation metrics, evaluation has been performed by using ROUGE, BLEU, and BERTScore [3], [7]. These scores are combined in order to determine lexical overlap, structural similarity, and semantic correspondence between the model output and the reference feedback.

4.1 Evaluation Metrics

To carry out objective performance analysis, the quality of generated feedback was measured by multiple, complementary metrics.

1) ROUGE-N (ROUGE-1, ROUGE-2)

$$\text{ROUGE-N} = \frac{\sum_{S\in\{\text{References}\}}\sum_{gram_n\in S}\min\left(\text{Count}_{gen}(gram_n),\text{Count}_{ref}(gram_n)\right)}{\sum_{S\in\{\text{References}\}}\sum_{gram_n\in S}\text{Count}_{ref}(gram_n)} \quad (4)$$

ROUGE-1 and ROUGE-2 compute unigram and bigram overlap, respectively.

2) ROUGE-L

ROUGE-L captures the Longest Common Subsequence (LCS) between two sequences, thus reflecting sentence-level fluency and structure.

$$\text{ROUGE-L} = F_{LCS} = \frac{(1+\beta^2)\cdot R_{LCS}\cdot P_{LCS}}{R_{LCS}+\beta^2\cdot P_{LCS}},\quad P_{LCS} = \frac{LCS(X,Y)}{m},\quad R_{LCS} = \frac{LCS(X,Y)}{n},\quad \beta = 1.2 \quad (5)$$

3) BLEU (Bilingual Evaluation Understudy)

BLEU measures precision-oriented n-gram overlap and penalizes overly short outputs through a brevity penalty (BP) [3].

$$\text{BLEU} = \text{BP}\cdot\exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (6)$$

4) BERTScore

BERTScore computes semantic similarity between reference and generated texts using contextual embeddings from a transformer model [9].

$$P = \frac{1}{m}\sum_{i}\max_{j}\cos\left(e_{x_i},e_{y_j}\right),\quad R = \frac{1}{n}\sum_{j}\max_{i}\cos\left(e_{x_i},e_{y_j}\right),\quad F_1 = \frac{2PR}{P+R} \quad (7)$$

Table 3. Model performance evaluation results.

Metric	**Score**
ROUGE-1	0.239
ROUGE-2	0.135
ROUGE-L	0.187
BLEU	0.007
BERTScore (F1)	0.85

All metrics were computed on the validation dataset, so as to assure reproducibility. Quantitative results are summarized in Table 3: lexical overlap metrics-either ROUGE or BLUE-are moderate, whereas BERTScore reaches an F1 score of 0.85, denoting strong semantic alignment between generated and teacher-written feedback.

V. Discussion

4.1. Interpretation of Results

The experimental results prove that the fine-tuned T5-small model generates teacher-like feedback, well-matched to the pedagogical standards of IELTS Writing Task 2 assessment [3]. The achievement of moderate lexical overlap-ROUGE-1 = 0.239; ROUGE-2 = 0.135-and high semantic similarity-BERTScore = 0.85-confirms that it captures meaning rather than mere surface structure [8]. This pattern supports the hypothesis that transformer-based architectures are able to generate human-like qualitative feedback through learning context-dependent relationships between prompts, essays, and teacher evaluations [6].

These results validate that fine-tuning a compact model like T5-small provides sufficient representational capacity for educational feedback generation without sacrificing interpretability or efficiency [7]. The fact that the model is able to replicate teacher reason structures in its response suggests that transformer-driven systems can contribute to improving formative assessment and reducing teacher workload [1], [2].

4.2. Comparison with Literature

The findings of this work are in agreement with recent contributions in the literature indicating that transformer-based models can result in significant pedagogical benefits for automated feedback generation. Previous work, including [3] and [12], has illustrated that text-to-text architectures tend to generate more lucid and informative feedback compared to scores based on traditional systems. Similarly, [11] shows that the introduction of explicit writing-skill indicators can increase contextual relevance of generated comments, which agrees with the improvements obtained here. At the same time, however, the BLEU and ROUGE scores recorded here are lower than those reported for studies that rely on larger datasets or more expressive transformer variants, such as T5-base or Flan-T5. That is likely because of the narrower variety and limited size of the IELTS corpus. Despite this, a comparatively high BERTScore confirms that the model captures semantic intent in a way that fits with pedagogical expectations, even where the lexical similarity is modest. Ethical considerations are still a major topic of discussion in much of the modern literature. Indeed, previous work such as [4] and [15] has underscored the importance of transparency and responsible deployment of AI tools in education. In line with such priorities, this research proposes a system oriented toward interpretability, clarity of reasoning, and supporting human-centered instructional practices.

4.3. Implications

Theoretical implications:

These findings add to the list of research that shows how transformer-based fine-tuning can model complex pedagogical reasoning processes [5]. The success of the T5-small framework also assures that smaller, domain-specific architectures are capable of rivaling large language models when applied to targeted educational tasks.

Practical implications:

From an instructional point of view, the system works like an intelligent assistant for teachers by providing real-time diagnostic feedback and promotes consistency in evaluation [2]. For students, it encourages reflective learning through detailed, human-like explanations rather than a bare numerical score. Again, this goes with the pedagogic objectives of autonomy, self-correction, and critical thinking [9], [14].

The presented study also considers the future integration of the model into digital

learning environments, including IELTS preparation platforms or LMSs, where real-time AI feedback can enhance formative assessment processes [10].

4.4. Limitations and Future Research

Although the model works well, there are a few limitations that still exist. First of all, the size of the dataset restricts lexical diversity, which can hardly generalize across genres [6]. While the semantic quality is high, sometimes the system produces repetitive phrasing or lacks nuanced evaluative details representative of human comments [12]. The absence of any form of validation across domains prevents conclusions about applicability to other writing contexts beyond IELTS. Future research should focus on expanding dataset variety, including multilingual samples, and experimenting with larger transformer variants-e.g., T5-base, Flan-T5, or Llama 3-to improve generalization and diversity [13]. Integration of explainability modules visualizing token-level attention weights could enhance transparency and user trust in educational AI [7], [15].

VI. Conclusion

This project successfully demonstrated the feasibility of transformer-based machine learning for human-like academic writing feedback generation. The T5-small model was fine-tuned on an IELTS essay dataset containing teacher-written feedback, which resulted in a system that could produce concise, coherent, and context-relevant textual feedback reflecting pedagogical writing criteria.

Unlike the traditional automated scoring models that predict a numerical grade, this system is aimed at formative feedback, helping learners understand their strengths and weaknesses while safeguarding academic integrity. The generated feedback appraised content relevance, coherence, and linguistic range-core components of the IELTS Writing Task 2 assessment framework [1].

The present study covers all stages of the NLP workflow: data collection, preprocessing, model training, and multiple metrics evaluations, such as BLEU, ROUGE, and BERTScore. The best results from the model achieved a BERTScore (F1) of 0.85, which confirmed high semantic similarity between generated and human feedbacks. This demonstrates the model's capacity to replicate teacher evaluation style effectively [2].

The architecture of the system further allows for scalability and transparency so that it can be easily integrated into educational platforms as a lightweight and privacy-preserving feedback assistant. When deployed locally or on the web, the model can support instructors by automating repetitive feedback tasks, thereby freeing educators up to give higher-order guidance [3].

Beyond just technical success, this work showcases the wider educational implications of the responsible adoption of AI. Emphasizing formative assessment and interpretability over automatic grading, the system falls in line with global efforts to ensure academic honesty and the ethical use of AI in higher education [4, 5].

In the future, the model will be further developed by adding modules on band score prediction, grammar correction, and multilingual feedback, together with a user-friendly web interface for online real-time evaluation. Such development will further establish AI's role as an assistant in pedagogy and not as a replacement for human teachers to make learning authentic, transparent, and student-centered [6], [7].

References

[1] Mouza, C., Codding, D., & Pollock, L. (2022). Investigating the impact of research-based professional development on teacher learning and classroom practice: Findings from computer science education. Computers & Education, 186, 104530. https://doi.org/10.1016/j.compedu.2022.104530

[2] Nurmalia Sari, M., Zhang, Y., & Abdullah, M. Y. (2025). INTEGRATING

AI-POWERED WRITING ASSISTANTS TO ENHANCE EFL STUDENTS' ACADEMIC WRITING SKILLS: A MIXED-METHODS STUDY IN HIGHER EDUCATION. *IJETA - International Journal of Education, Technology, and AI*, *1*(1), 1–12. Retrieved from https://ejournal.rabiahfoundation.or.id/index.php/ijeta/article/view/2

[3] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (T5)," *J. Mach. Learn. Res.*, vol. 21, pp. 1-67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[4] Balalle, Himendra & Pannilage, Sachini. (2025). Reassessing academic integrity in the age of AI: A systematic literature review on AI and academic integrity. Social Sciences & Humanities Open. doi: 10.1016/j.ssaho.2025.101299

[5] Lise Jaillant, Arran Rees, Applying AI to digital archives: trust, collaboration and shared professional ethics, *Digital Scholarship in the Humanities*, Volume 38, Issue 2, June 2023, Pages 571–585, https://doi.org/10.1093/llc/fqac073

[6] IELTS Writing Task 2 Evaluation Dataset, Hugging Face. Available: https://huggingface.co/datasets/chillies/IELTS-writing-task-2-evaluation

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *Neural Information Processing Systems*, *30*, 5998–6008. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[8] M. Khalifa, M. Albadawy, Using artificial intelligence in academic writing and research: an essential productivity tool, Comput. Methods Programs. Biomed. Update (2024) 100145. https://doi.org/10.1016/j.cmpbup.2024.100145

[9] Elkhatat, A.M., Elsaid, K. & Almeer, S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *Int J Educ Integr* 19, 17 (2023). https://doi.org/10.1007/s40979-023-00140-5

[10] Kar SK, Bansal T, Modi S, Singh A. How Sensitive Are the Free AI-detector Tools in Detecting AI-generated Texts? A Comparison of Popular AI-detector Tools. *Indian Journal of Psychological Medicine*. 2024;47(3):275-278. doi:10.1177/02537176241247934

[11] Y. Liu, J. Han, A. Sboev, I. Makarov, Geef: a neural network model for automatic essay feedback generation by integrating writing skills assessment, Expert Syst. Appl. 245 (2024) 123043. https://doi.org/10.1016/j.eswa.2023.123043

[12] Misgna, H., On, BW., Lee, I. *et al.* A survey on deep learning-based automated essay scoring and feedback generation. *Artif Intell Rev* 58, 36 (2025). https://doi.org/10.1007/s10462-024-11017-5

[13] Zheng, X., Zhang, J. The usage of a transformer based and artificial intelligence driven multidimensional feedback system in english writing instruction. *Sci Rep* 15, 19268 (2025). https://doi.org/10.1038/s41598-025-05026-9

[14] Anani, G.E., Nyamekye, E. & Bafour-Koduah, D. Using artificial intelligence for academic writing in higher education: the perspectives of university students in Ghana. *Discov Educ* 4, 46 (2025). https://doi.org/10.1007/s44217-025-00434-5

[15] Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, *49*(5), 847–864. https://doi.org/10.1080/03075079.2024.2323593