Enhancing Essay Scoring with GPT-2 Using Back Translation Techniques

Aysegul Gunduz
University of Alberta
gunduz@ualberta.ca

Mark J. Gierl
University of Alberta
mark.gierl@ualberta.ca

Okan Bulut
University of Alberta
okan.bulut@ualberta.ca

Abstract

Advancements in artificial intelligence and transformer-based language models have significantly influenced educational assessment, particularly in the development of Automated Essay Scoring (AES) systems. This study examines the effectiveness of the GPT-2 small model in evaluating student essays from the Automated Student Assessment Prize (ASAP) dataset[1].It also explores the effect of a back-translation data augmentation technique(translating essays into Turkish and then back into English) On model performance. Evaluation metrics include Cohen's kappa and Quadratic Weighted Kappa (QWK). The model achieved QWK scores ranging from 0.60 to 0.80 across essay sets, with a peak of 0.77 on Essay Set 5. Notably, back translation led to substantial improvements, particularly in Essay Set 8, where QWK increased by 33%. These findings highlight the potential of data augmentation to mitigate class imbalance and improve scoring robustness. However, the limited semantic depth of the GPT-2 small model points to the need for more advanced, rubric-aware architectures. The study underscores the importance of balanced data distributions in enhancing the validity and fairness of AES systems.

Keywords: *artificial intelligence, language modeling, automated essay scoring (AES), GPT-based models, GPT-2*

1 Introduction

Recent advances in large language models (LLMs), particularly those developed under the Generative Pretrained Transformer (GPT) architecture, have significantly influenced Automated Essay Scoring (AES). Early AES systems relied on surface-level linguistic features and traditional machine learning algorithms (Kumar and Boulanger, 2020; Klebanov and Madnani, 2022), while more recent approaches have incorporated transformer-based models capable of capturing deeper semantic and syntactic patterns (Taghipour and Ng, 2016). Among these, encoder-only architectures such as BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2019) have been widely applied to AES tasks, achieving strong performance and serving as reliable baselines (Firoozi et al., 2023; Wang et al., 2022).

In contrast, decoder-based generative models, particularly GPT variants, have received comparatively limited attention in AES despite their proven success in other natural language processing applications. Recent studies have demonstrated that advanced generative models such as GPT-3.5 and GPT-4 can achieve near human-level performance in essay scoring benchmarks (Mizumoto and Eguchi, 2023; Xiao et al., 2025; Yamashita, 2025). However, these models are proprietary and resource-intensive, limiting their accessibility for educational researchers and practitioners. Smaller, open-source alternatives like GPT-2 remain underexplored in AES, even though evidence from related classification tasks indicates that fine-tuned GPT-2 can rival or surpass BERT-based hybrids in performance (Bouchiha et al., 2025). This observation supports the need for systematic evaluation of GPT-2 in AES, both as a practical and a methodological contribution.

Another persistent challenge in AES is the limited size and imbalance of available datasets, which can compromise model generalization and fairness (Jong et al., 2022; Guo et al., 2024). Data augmentation techniques such as back-translation have been proposed as a potential solution, offering greater linguistic diversity and reducing the effects of class imbalance (Lun et al., 2020). Yet, it remains unclear whether performance improvements attributed to back-translation derive from genuine linguistic variation or simply from the increased size of training data. Addressing this ambiguity requires careful experimental design that controls

[1] https://www.kaggle.com/c/asap-aes

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 406–416

for training size and duplication.

This study makes two key contributions. First, it provides a controlled evaluation of GPT-2 for AES on the ASAP dataset, positioning it against established encoder-based baselines and recent parameter-efficient fine-tuning approaches such as LoRA (Liu et al., 2024). Second, it investigates the effect of back-translation as a data augmentation strategy under controlled conditions, clarifying whether observed improvements stem from data diversity rather than dataset expansion alone.

The study is guided by two research questions:

1. How reliably can a fine-tuned GPT-2 model score essays from the ASAP dataset compared to established baselines?
2. To what extent does back-translation improve GPT-2's AES performance beyond the effect of increasing training set size?

2 Related Work

2.1 Overview of Automated Essay Scoring

Automated Essay Scoring (AES) refers to the use of computational methods to evaluate and score student essays (Shermis, 2014). While manual scoring is often time-consuming and prone to rater inconsistency, AES offers efficiency, objectivity, and scalability, making it an increasingly valuable tool in educational contexts (Yan et al., 2020).

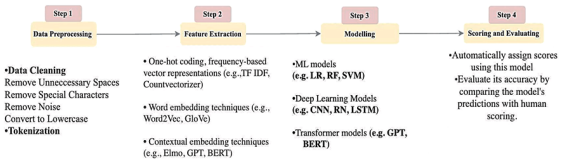

Figure 1: The AES Process described in Four Steps (Gierl et al., 2014).

As can be seen in Figure 1, the AES process consists of four steps: text preprocessing, feature extraction, model training, and performance evaluation (Gierl et al., 2014). To detail, it involves the preprocessing (Step 1) and conversion of essays written in a training environment into numerical vectors using text representation techniques (Step 2), combining these vectors with machine learning algorithms or deep learning networks to create a scoring model (Step 3), and automatically assigning scores using this model and evaluating the scoring model to see if it can predict human scoring (Step 4). Advances in machine learning and natural language processing have significantly improved the first three stages, particularly through enhanced text representation and modeling techniques (**?**).

Early feature extraction methods employed frequency-based techniques, such as term frequency (TF) and TF-IDF (Salton et al., 1975), but these approaches were unable to capture semantic meaning. Later, word embedding models like Word2Vec and GloVe (Mikolov et al., 2013) improved semantic representation but were still context-independent. Contextual embedding models such as ELMo, BERT, and GPT addressed this limitation by incorporating surrounding context into each word's representation (Peters et al., 2018; Radford et al., 2018; Liu et al., 2020). These advances improved the quality of input features used in scoring models.

Earlier AES studies employed deep neural networks (DNNs) and recurrent neural networks (RNNs) to model sequential patterns in text (Alikaniotis et al., 2016; Tay et al., 2018). RNNs often struggle with capturing long-range dependencies and information across time steps; however, they are designed to suit the sequence-to-sequence design effectively (Nugaliyadde et al., 2019). This limitation has motivated the use of transformer-based architectures, which replace recurrence with attention mechanisms. The self-attention mechanism introduced by Vaswani et al. (2017) enables the model to learn dependencies across all positions in a sequence simultaneously, resulting in a richer representation of global structure and semantic relationships. As a result, transformer models such as GPT have become increasingly prevalent in recent AES research.

2.2 Transformer-Based Architectures in AES

In 2017, the paper 'Attention Is All You Need' revolutionized the field of natural language processing (NLP) by introducing the transformer architecture (see Figure 2) (Vaswani et al., 2017). This model leveraged self-attention mechanisms to capture long-range semantic and syntactic dependencies in text. In AES tasks, encoder-only transformers such as BERT (Devlin et al., 2019) and RoBERTa have also demonstrated state-of-the-art performance in both predictive and analytic scoring (Firoozi et al., 2023; Klebanov and Madnani, 2022).

These models provide robust contextual embeddings and strong baselines for AES. These models are typically fine-tuned on prompt-specific essay datasets, where only the top classification layer is

updated while the encoder layers provide contextualized embeddings. This parameter-efficient strategy has proven effective in score prediction, especially under constrained computational resources. However, their reliance on bidirectional masked language modeling may limit their utility in generative tasks and document-level coherence modeling. Although rubric-integrated encoder architectures have improved interpretability and alignment with human scoring rubrics (Liu et al., 2020), their generalization across unseen prompts and diverse discourse structures remains limited.

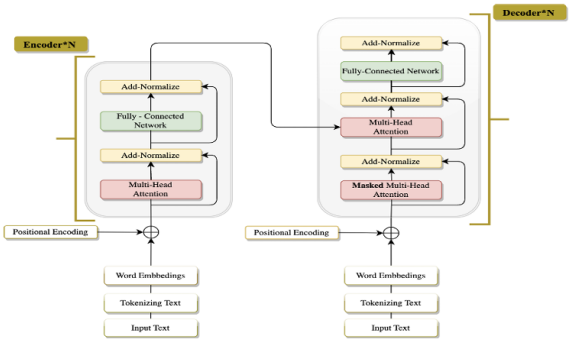

Figure 2: Transformer Architecture (Vaswani et al., 2017).

These limitations have motivated research on decoder-based models, which are inherently more suited to sequence-level generation and whole-document representation. Importantly, GPT-2 has not only been effective in generative tasks but has also demonstrated competitive performance in classification settings. For instance, GPT-2 has matched or even surpassed BERT-based hybrids in hierarchical text classification (Bouchiha et al., 2025), performed strongly in text classification and natural language inference benchmarks (Montesinos, 2020), and shown competitive results against BERT in low-resource classification tasks (Wang et al., 2024). Such findings indicate that GPT-2 is a viable model for AES, where both classification accuracy and generative capabilities are critical.

2.3 Decoder-Only Transformers: The GPT Family

Decoder-only models, particularly the GPT series introduced by OpenAI, are trained with autoregressive objectives and unidirectional attention, which makes them inherently generative (Radford et al., 2018, 2019). GPT-2 expanded this architecture to 1.5 billion parameters and demonstrated strong transfer performance. Subsequent models such as GPT-3, GPT-3.5, and GPT-4 further scaled capacity and achieved near human-level accuracy in AES benchmarks under zero-shot and few-shot prompting conditions (Mizumoto and Eguchi, 2023; Yancey et al., 2023; Gunduz and Gierl, 2024). GPT-4 also introduced multimodal input processing, although its architecture and training data remain undisclosed

While these larger models have shown impressive results, their proprietary nature restricts reproducibility and accessibility. In contrast, GPT-2 remains fully open-source and scalable across different sizes, making it a practical option for academic and educational research. Importantly, GPT-2 has proven effective beyond generative applications. Prior studies have shown that fine-tuned GPT-2 can outperform BERT-based hybrids in hierarchical text classification (Bouchiha et al., 2025), perform strongly in text classification and natural language inference benchmarks (Montesinos, 2020), and achieve competitive results against BERT in low-resource classification tasks (Wang et al., 2024).

These findings indicate that GPT-2 is not only cost-efficient and accessible but also capable of delivering robust performance in classification-oriented tasks. Nevertheless, systematic evaluations of GPT-2 on AES benchmark datasets such as ASAP remain limited. This study addresses this gap by providing a controlled and reproducible assessment of GPT-2 for essay scoring, with particular attention to the role of data augmentation.

2.4 Recent Applications of GPT-Based AES

In recent years, GPT models have been increasingly applied to educational assessment tasks, including both short-answer and essay scoring. One line of research has focused on data augmentation to mitigate class imbalance.Fang et al. (2023) employed GPT-4 to generate synthetic responses for minority scoring classes, which improved the performance of a DistilBERT scoring model. Similarly, Gaddipati et al. (2020) compared transfer learning models such as ELMo, BERT, GPT, and GPT-2 for short-answer grading, showing that while ELMo provided strong baselines, transformer-based models offered greater scalability for downstream use. Several studies have investigated larger GPT models for direct scoring. Mizumoto and Eguchi (2023) evaluated GPT-3 on TOEFL essays and reported that combining linguistic features with

Model	Architecture	Parameters	Training Data	Release Date
GPT-1	12H Decoder	117M	BookCorpus	2018
GPT-2	12–48H Decoder	1.5B	WebText	2019
GPT-3	Modified GPT-2	175B	CC + WebText	2020
GPT-3.5	Undisclosed	175B	–	Mar 2022
GPT-4	Undisclosed	∼1.7T	–	Mar 2023

Table 1: Overview of OpenAI's GPT-n series.

Model	Params	Layers	Hidden	Input
Small	117M	12	768	768
Medium	345M	24	1024	1024
Large	774M	36	1280	1280
XL	1558M	48	1600	1600

Table 2: GPT-2 Model Configurations across Four Sizes.

model outputs improved agreement with human raters.Yancey et al. (2023) assessed GPT-3.5 and GPT-4 on essays from English language learners, finding that GPT-4 achieved performance comparable to state-of-the-art Automated Writing Evaluation (AWE) systems, though alignment varied by learners' first language. Henkel et al. (2023) used GPT-4 for scoring short-answer reading comprehension tasks in low- and middle-income countries, demonstrating its potential in resource-limited educational contexts.Obata et al. (2023) tested ChatGPT for essay scoring in English and Japanese and showed that validity improved when combined with linguistic features.Xiao et al. (2025) further argued that GPT-3.5 and GPT-4 are most effective when augmenting human raters in hybrid scoring systems.

Despite these advances, most work has concentrated on proprietary models such as GPT-3.5 and GPT-4, limiting reproducibility and transparency. Benchmark studies on open-source models remain scarce. Gunduz and Gierl (2024) compared GPT-3.5 and GPT-4 under different prompting conditions on the ASAP dataset, but no systematic evaluation of GPT-2 has yet been conducted. Considering GPT-2's accessibility, scalability, and demonstrated competitiveness in classification tasks (Bouchiha et al., 2025; Wang et al., 2024), further investigation is warranted. This study addresses this gap by fine-tuning GPT-2 on the ASAP dataset and evaluating the effects of back-translation as a data augmentation strategy, offering a reproducible and transparent alternative to proprietary systems.

2.5 Data Augmentation and Back-Translation in AES

Data augmentation is widely used in NLP to improve generalization and mitigate label imbalance through techniques such as synonym replacement, paraphrasing, and translation-based methods (Wei and Zou, 2019). In AES, augmentation helps balance score distributions and enrich training data diversity (Lun et al., 2020; Jong et al., 2022).

Back-translation, which generates paraphrases by translating text into a target language and back, has been shown to increase linguistic variety and robustness in low-resource tasks (Sennrich et al., 2016; Edunov et al., 2018). In AES, augmentation methods have been applied to enrich training data (e.g., (Firoozi, 2023; Guo et al., 2024)), yet the specific impact of back-translation on score distributions, particularly under imbalanced data conditions, remains underexplored. This study addresses this gap by applying back-translation to the ASAP dataset under controlled conditions, clarifying its contribution beyond simple dataset expansion.

3 Method

3.1 Dataset

This study utilizes the Automated Student Assessment Prize (ASAP) dataset, developed under the sponsorship of the Hewlett Foundation in 2012, to encourage scalable and reliable approaches to AES (Shermis, 2014). The dataset comprises eight distinct essay sets written by students in Grades 7 through 10, encompassing various genres, including narrative, persuasive, and expository writing. Each essay set varies in terms of grade level, rubric

type, essay length, and scoring range (see Table 3). Essays were scored by two or three expert raters using holistic, trait-based, or composite rubrics. The score ranges and aggregation methods for domain scores differ across sets. Table 3 summarizes the specific score ranges for Rater 1, Rater 2, and the derived domain score used for model training. Among the essay sets, Set 4 stands out for its relatively balanced score distribution across all score categories. As shown in Table 3, both individual rater scores and the domain score span the full range from 0 to 3, with sufficient representation in each category. This balanced distribution is particularly beneficial for training reliable AES models, as it reduces the risk of class imbalance and supports more effective learning dynamics.

3.2 Data Preprocessing

Text Preprocessing. To prepare the essays for model input, standard text preprocessing steps were applied. All texts were lowercased and lemmatized using the NLTK library (Bird et al., 2009). The cleaned essays were then tokenized using the GPT-2 tokenizer from the Hugging Face Transformers library (Wolf et al., 2019). Since transformer models require fixed-length input, padding and truncation were used to standardize sequence lengths.

Score Preprocessing. Each essay was scored by two or three raters, and domain scores were computed according to the scoring rules in Table 3. However, some sets (Essay Sets 1, 7, and 8) had wide or unbalanced score ranges. To improve model performance, these scores were rescaled into fewer ordinal categories. For instance, Set 1 domain scores (2–12) were converted to a 1–6 ordinal scale. Similarly, Set 7 scores (0–24) were mapped to a 0–3 scale, and Set 8 scores (0–60) were compressed into six ordinal categories based on trait aggregation logic (see Table 3).

3.3 Model Development

GPT-2 Architecture. The model developed in this study builds upon the GPT-2 architecture (see Figure 3), a decoder-only transformer pretrained on over 8 million web pages (Radford et al., 2019). GPT-2 generates contextualized word embeddings using masked self-attention and is optimized for predicting the next token. Among its four variants, the smallest version—GPT-2 Small (124M parameters, 12 decoder layers, 768 hidden units)—was selected due to computational efficiency. All training was conducted using Google Colab Pro (Tesla V100 GPU, 32GB RAM). Each decoder block in

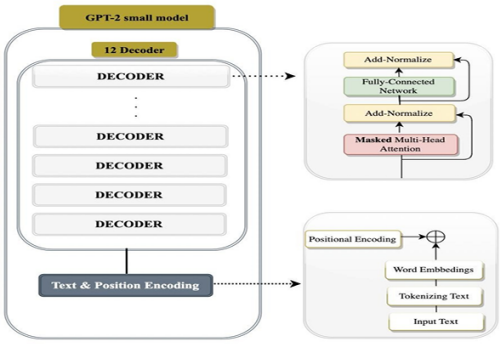

Figure 3: GPT-2 Small Architecture.

GPT-2 includes masked multi-head self-attention, a feedforward network, residual connections, and layer normalization (Vaswani et al., 2017). The input sequence is processed from left to right, making the model suitable for both generative and classification tasks.

Classification Head. To adapt GPT-2 for AES, we added a task-specific classification head on top of the pretrained transformer (see Figure 4). This consisted of a dropout layer (with rates of 0.1, 0.2, and 0.5 tested) followed by a fully connected linear layer that maps the last hidden state of the model to a fixed number of score classes per essay set (e.g., 4 classes in Essay Set 4). The `num_labels` parameter was dynamically set based on the scoring range of each set. All training was performed on Google Colab Pro using the Hugging Face Transformers library (Wolf et al., 2019). The small-scale GPT-2 variant enabled faster iteration while maintaining competitive performance for AES tasks.

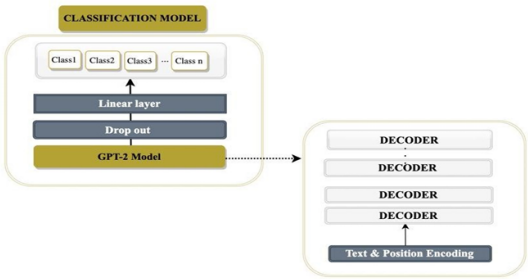

Figure 4: Classification Model Architecture.

3.4 Experimental Setup and Hyperparameter Tuning

All essays were tokenized using the GPT-2 tokenizer from the HuggingFace Transformers library.

Set	Grade	Essay Type	Train Size	Avg. Len.	Rubric Type	Raters	Score Range	Domain Score Explanation
1	8	Persuasive	1783	350	Holistic	2	2–12	Sum of R1 and R2 (2–12)
2a	10	Persuasive	1800	50	Trait	2	1–6	Equals R1's score (1–6)
2b	10	Persuasive	1800	50	Trait	2	1–4	Equals R1's score (1–4)
3	10	Source-Dep.	1726	50	Holistic	2	0–3	Max(R1, R2) (0–3)
4	10	Source-Dep.	1772	50	Holistic	2	0–3	Near max(R1, R2) (0–3)
5	8	Source-Dep.	1805	50	Holistic	2	0–4	Near max(R1, R2) (0–4)
6	10	Source-Dep.	1800	50	Holistic	2	0–4	Near max(R1, R2) (0–4)
7	7	Expository	1569	50	Composite	2	0–12	Sum of R1 and R2 (0–24)
8	10	Expository	723	50	Composite	3	0–30	R1+R2 or R3 used (0–60)

Table 3: Descriptive Statistics and Scoring Guidelines for the Eight ASAP Essay Sets.

Essays exceeding the maximum sequence length of 1,024 tokens (as imposed by the GPT-2 Small architecture) were truncated.

The pre-trained GPT-2 Small model was initialized with default configurations: 12 decoder layers, 768-dimensional hidden states and embeddings, 12 self-attention heads, GELU activation, and dropout probability of 0.1 across embedding, attention, and fully connected layers. Layer normalization used an epsilon value of 1e-5. In total, the model contains approximately 117M parameters. To adapt GPT-2 for essay scoring, a linear classification head with dropout was appended. The number of output classes was defined per essay set.

To adapt GPT-2 for essay scoring, a linear classification head with dropout was appended. The number of output classes was defined per essay set using the `num_labels` parameter. The final hidden state of the first token was passed to the classification layer. Model training was optimized using the AdamW optimizer (Kingma, 2014) with a fixed learning rate of 1e-4 and categorical cross-entropy loss. The loss function for k classes is defined in Equation 1

$$\mathcal{L}(y, \hat{y}) = -\sum_{i=1}^{k} y_i \log(\hat{y}_i) \tag{1}$$

where y is the one-hot true label and $\hat{y}$ is the predicted class distribution.

To ensure robust evaluation, each essay set was randomly partitioned into training (60%), validation (20%), and test (20%) subsets following standard practice.

3.5 Data Augmentation Strategy

The distribution of essays across score levels is utilized by the GPT-2 Small architecture, which features the performance and generalizability of AES models. To address this, we employed data augmentation to enhance the training set, particularly for underrepresented classes.

Text Augmentation. Text data augmentation involves generating additional samples by modifying existing texts, without requiring new data collection. Common methods include synonym replacement (SR), random insertion (RI), random swap (RS), and random deletion (RD), which introduce lexical variability while preserving sentence structure (Firoozi, 2023).

Back-Translation. Among these methods, back-translation has emerged as a particularly effective strategy for producing fluent and semantically consistent variations. This technique translates a sentence into an intermediate language and back to the original, creating paraphrased versions that enrich the training data. In our study, the source language was English, and the target language was Turkish. We translated English essays into Turkish and then back into English using the Google Translate API. Turkish was intentionally chosen as the pivot language due to its agglutinative morphology and syntactic divergence from English, contributing to greater linguistic variety in the augmented texts.

This method was selectively applied: in balanced sets (e.g., Set 4), each score class was augmented by 20% following the strategy proposed in Firoozi's Doctoral Thesis (Firoozi, 2023), while in imbalanced sets, score levels with fewer than 50 samples were doubled. This targeted approach aimed to reduce class imbalance, minimize model bias, and improve performance across the entire score spectrum. The augmentation process is illustrated in Figure 5.

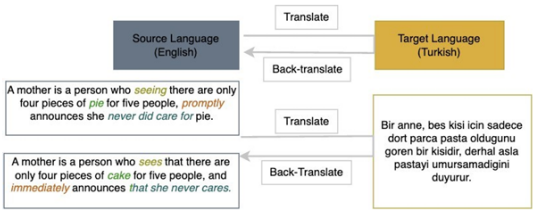

Figure 5: Back-Translation Data Augmentation Pipeline.

3.6 Performance Metrics

To evaluate the effectiveness of the AES model, we employed multiple metrics capturing both agreement with human raters and classification performance.

Cohen's Kappa. Cohen's Kappa (κ) measures inter-rater agreement corrected for chance, and is commonly used to assess the consistency between model predictions and human scores. It is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{2}$$

where P_o is the observed agreement and P_e is the expected agreement by chance. Agreement levels are interpreted based on the guidelines by Landis and Koch (1977).

Quadratic Weighted Kappa (QWK). QWK extends Cohen's Kappa by penalizing disagreements based on the distance between score levels, making it especially suitable for ordinal tasks, such asutilized the GPT-2 Small architecture, featurings:

$$w_{ij} = \frac{(i-j)^2}{(N-1)^2} \tag{3}$$

The QWK score is then defined by:

$$QWK = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}} \tag{4}$$

Accuracy. Accuracy reflects the proportion of essays for which the predicted score exactly matches the human-assigned score. Although it does not account for ordinal distance between misclassified levels, it remains a useful baseline metric for evaluating overall classification correctness.

4 Results

4.1 Hyperparameter Settings

To optimize performance, we fine-tuned key hyperparameters for each essay set, as detailed in Table 4. All models used the GPT-2 Small architecture with 768-dimensional token embeddings, 1024-dimensional positional encodings, and 12 decoder layers.

A classification head consisting of a dropout and a linear layer was appended to map outputs to a variable number of score classes (`num_labels`) per essay set. Dropout rates were adjusted individually; Sets 1 and 2a performed best with 0.5, while 2b, 3, 5, 7, and 8 performed best with 0.1.

A fixed learning rate of 1e-4 was used across sets, optimized via the `AdamW` optimizer. Epochs and batch size varied by set, reflecting differences in convergence behavior and dataset size. For example, Set 6 performed best with 30 epochs, dropout 0.3, and batch size 2.

4.2 RQ1: AES Model Performance

The fine-tuned GPT-2 model demonstrated moderate scoring reliability across the eight essay sets. On average, it achieved a Quadratic Weighted Kappa (QWK) of 0.68, Cohen's Kappa of 0.43, and classification accuracy of 61%. According to the interpretability thresholds proposed by Williamson et al. (2012), the model explained a substantial portion of human scoring variance. These results suggest that even the most minor GPT-2 variant can offer competitive performance in AES tasks under computational constraints.

4.3 RQ2: Effect of Data Augmentation

Back-translation-based data augmentation led to notable performance improvements. The average QWK score increased from 0.68 to 0.74 (+0.06), while Cohen's Kappa rose from 0.43 to 0.48 (+0.05). This gain was most evident in essay sets with initially imbalanced score distributions, confirming the effectiveness of targeted augmentation in enhancing agreement between machine predictions and human raters.

5 Discussion

This study examined the performance of the open-source GPT-2 model for AES on the ASAP dataset, with a focus on fine-tuning and back-translation-based data augmentation. Results show that even the most minor GPT-2 variant, when fine-tuned with optimized hyperparameters, achieved a competitive average QWK score of 0.68 (close to human-level performance at 0.74) and outperformed GPT-3.5 in certain sets.The model performed best in balanced essay sets with sufficient

Parameter	1	2a	2b	3	4	5	6	7	8
Embedding Dim.	768	768	768	768	768	768	768	768	768
Positional Encoding	1024	1024	1024	1024	1024	1024	1024	1024	1024
Decoder Layers	12	12	12	12	12	12	12	12	12
Num Labels	6	6	4	4	4	5	5	4	6
Dropout Rate	0.5	0.1	0.5	0.2	0.2	0.1	0.1	0.3	0.1
Learning Rate	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4
Epochs	20	25	35	20	30	20	30	20	30
Batch Size	2	2	2	4	2	2	2	1	2

Table 4: Final selection of Hyperparameters used for Fine-tuning GPT-2 across all Essay Sets.

Model	1	2a	2b	3	4	5	6	7	8	Average
GPT-2	0.75	0.64	0.66	0.71	0.74	0.77	0.73	0.77	0.43	0.68
Human Raters	0.71	0.78	0.72	0.81	0.86	0.74	0.77	0.68	0.63	0.74
Discrepancy	0.04	0.14	0.06	0.10	0.12	0.03	-0.04	0.09	0.18	0.06

Table 5: Comparison of GPT-2 and Human Raters using QWK across all Essay Sets.

Essay Set	1	2a	2b	3	4	5	6	7	8
Before BT	1783	1800	1800	1726	1772	1805	1800	1569	723
After BT	1875	1831	1829	1765	2124	1829	1844	1676	1220
Discrepancy	+92	+31	+29	+39	+352	+24	+44	+107	+497

Table 6: Comparison of Training Dataset Size Before and After Back-Translation for each Essay Set.

Model	Performance	1	2a	2b	3	4	5	6	7	8	Average
GPT-2	Cohen's Kappa	0.52	0.46	0.45	0.41	0.43	0.45	0.40	0.41	0.31	0.43
	QWK	0.75	0.64	0.66	0.71	0.74	0.77	0.73	0.71	0.45	0.68
	Accuracy	0.67	0.61	0.63	0.61	0.60	0.66	0.58	0.60	0.53	0.61
GPT-2 + BT	Cohen's Kappa	0.54	0.50	0.53	0.47	0.48	0.47	0.43	0.49	0.38	0.48
	QWK	0.79	0.65	0.68	0.77	0.81	0.79	0.79	0.74	0.60	0.74
	Accuracy	0.72	0.62	0.74	0.62	0.68	0.64	0.67	0.67	0.59	0.66
Discrepancy	Cohen's Kappa	+0.02	+0.04	+0.08	+0.06	+0.05	+0.03	+0.01	+0.08	+0.11	+0.05
	QWK	+0.04	+0.01	+0.02	+0.06	+0.06	+0.02	+0.06	+0.03	+0.15	+0.06
	Accuracy	+0.05	+0.01	+0.11	+0.01	+0.09	+0.02	+0.06	+0.07	+0.06	+0.05

Table 7: Comparison of GPT-2 Model Performance Before and After Back-Translation (BT) across all Essay Sets.

training data, while lower reliability was observed in sparse or imbalanced sets such as Set 8. Data augmentation proved particularly effective for underrepresented score classes, improving both QWK and Cohen's Kappa scores and reducing class imbalance. These findings affirm the value of tuning smaller, accessible models for educational NLP tasks, highlighting the trade-off between model complexity and interpretability in low-resource contexts.

In conclusion, GPT-2, despite its smaller architecture, offers substantial potential for AES when carefully fine-tuned and supported by data augmentation. Its open-source nature and customizable hyperparameters make it a practical choice for scalable, interpretable assessment systems. Back-translation significantly improved performance in low-resource score categories, demonstrating its value in addressing data sparsity. These results reinforce that high-quality AES systems can be developed without relying solely on larger proprietary models, and suggest future directions in combining linguistic measures and augmentation techniques to enhance model robustness and fairness.

6 Limitations and Future Work

Despite promising results, this study has several limitations. First, it focuses exclusively on classification-based essay scoring and does not incorporate rubric-specific features, which are central to many human scoring protocols. The absence of rubric-aligned modeling limits interpretability and may hinder the effectiveness of feedback-oriented applications. Second, the dataset includes essay sets with imbalanced score distributions and small sample sizes, which may constrain generalizability, particularly in underrepresented categories. Third, experiments were limited to the GPT-2 Small model; while acceptable, tuning significantly improved performance, larger models (e.g., GPT-3, GPT-4) could better capture complex linguistic structures if similarly fine-tuned. Lastly, only one augmentation strategy—back-translation—was explored. Future work should investigate rubric-aware scoring frameworks, incorporate alternative augmentation methods (e.g., synonym substitution, sentence permutation), and evaluate larger-scale models on more balanced datasets to improve the robustness and educational utility of AES systems.

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. *arXiv preprint arXiv:1606.04289*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Djelloul Bouchiha, Abdelghani Bouziane, Noureddine Doumi, Benamar Hamzaoui, and Sofiane Boukli-Hacene. 2025. Hierarchical text classification: Fine-tuned gpt-2 vs bert-bilstm. *Applied Computer Systems*, 30(1):40–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Luyang Fang, Gyeong-Geon Lee, and Xiaoming Zhai. 2023. Using gpt-4 to augment unbalanced data for automatic scoring. *arXiv preprint arXiv:2310.18365*.

Tahereh Firoozi. 2023. Using automated procedures to score written essays in persian: An application of the multilingual bert system.

Tahereh Firoozi, Okan Bulut, and Mark Gierl. 2023. Language models in automated essay scoring: Insights for the turkish language. *International Journal of Assessment Tools in Education*, 10(Special Issue):149–163.

Sasi Kiran Gaddipati, Deebul Nair, and Paul G Plöger. 2020. Comparative evaluation of pretrained transfer learning models on automatic short answer grading. *arXiv preprint arXiv:2009.01303*.

Mark J Gierl, Syed Latifi, Hollis Lai, André-Philippe Boulais, and André De Champlain. 2014. Automated essay scoring and the future of educational assessment in medical education. *Medical education*, 48(10):950–962.

Aysegul Gunduz and M Gierl. 2024. Automated essay scoring with chatgpt 3.5 and 4.0. Presentation at the UBlberta Graduate Student Research in Education Conference.

Weiqin Guo, Yong Yang, and Ge Ren. 2024. Research of automatic scoring of essays based on data augmentation. In *Proceedings of the 4th Asia-Pacific Artificial Intelligence and Big Data Forum*, pages 635–641.

Owen Henkel, Libby Hills, Bill Roberts, and Joshua McGrane. 2023. Can llms grade short-answer reading comprehension questions: An empirical study with a novel dataset. *arXiv preprint arXiv:2310.18373*.

You-Jin Jong, Yong-Jin Kim, and Ok-Chol Ri. 2022. Improving performance of automated essay scoring by using back-translation essays and adjusted

scores. *Mathematical Problems in Engineering*, 2022(1):6906587.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Beata Beigman Klebanov and Nitin Madnani. 2022. *Automated essay scoring*. Springer Nature.

Vivekanandan Kumar and David Boulanger. 2020. Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in education*, volume 5, page 572367. Frontiers Media SA.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Qi Liu, Matt J Kusner, and Phil Blunsom. 2020. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.

Zequan Liu, Jiawen Lyn, Wei Zhu, Xing Tian, and Yvette Graham. 2024. ALoRA: Allocating low-rank adaptation for fine-tuning large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 622–641, Mexico City, Mexico. Association for Computational Linguistics.

Jiaqi Lun, Jia Zhu, Yong Tang, and Min Yang. 2020. Multiple data augmentation strategies for improving performance on automatic short answer scoring. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13389–13396.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Dimas Munoz Montesinos. 2020. Modern methods for text generation. *arXiv preprint arXiv:2009.04968*.

Anupiya Nugaliyadde, Upeka Somaratne, and Kok Wai Wong. 2019. Predicting electricity consumption using deep recurrent neural networks. *arXiv preprint arXiv:1909.08182*.

Ayaka Obata, Takumi Tagawa, and Yuichi Ono. 2023. Assessment of chatgpt's validity in scoring essays by foreign language learners of japanese and english. In *2023 15th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)*, pages 105–110. IEEE.

Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Gerard Salton, Chung-Shu Yang, and Clement T Yu. 1975. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33–44.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (NeurIPS 2019)*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–96.

Mark D Shermis. 2014. State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. *Assessing Writing*, 20:53–76.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018. Recurrently controlled recurrent networks. *Advances in neural information processing systems*, 31.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yongjie Wang, Chuan Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. *arXiv preprint arXiv:2205.03835*.

Yu Wang, Wen Qu, and Xin Ye. 2024. Selecting between bert and gpt for text classification in political science research. *arXiv preprint*, arXiv:2411.05050.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

David M Williamson, Xiaoming Xi, and F Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2025. Human-ai collaborative essay scoring: A dual-process framework with llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 293–305.

Taichi Yamashita. 2025. Exploring potential biases in gpt-4o's ratings of english language learners' essays. *Language Testing*, 42(3):344–358.

Duanli Yan, André A Rupp, and Peter W Foltz. 2020. *Handbook of automated scoring: Theory into practice*. CRC Press.

Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 576–584.