A Generative Transformer-Based Approach to Automated Essay Scoring: Evaluating GPT-2's Performance with Pre- and Post- Data Augmentation

by

Aysegul Gunduz

A thesis submitted in partial fulfilment of the requirements for the degree of

Master of Education

in

Measurement, Evaluation and Data Science

Department of Educational Psychology

University of Alberta

Abstract

Recent advancements in artificial intelligence and language modeling have revolutionized the domain of educational technology, with a special focus on the utility of automated essay scoring (AES) systems. The potential of GPT-based model architectures, including different versions or iterations of the ChatGPT tool, has become an important research topic. My research is designed to investigate the performance of the GPT-2 small model in AES and examine how the back translation technique between English and Turkish can improve its performance on the Hewlett-sponsored ASAP dataset (https://www.kaggle.com/c/asap-aes). The evaluation is based on both Cohen's kappa and Quadratic Weighted Kappa (QWK) for agreement reliability, with additional metrics such as accuracy, precision, sensitivity, and the F-1 score providing further insight into the classification accuracy. Findings indicate a QWK range of 0.60 to 0.80 across most ASAP essay sets, with Essay Set 5 reaching a peak QWK of 0.77. Back Translation techniques showed a significant increase in the model's performance, especially in Essay Set 8, where there was a QWK score increase of 33%. The study highlights the limited capacity of GPT-2 small model and emphasizes the importance of conducting future research with more advanced GPT versions. It also underscores the importance of balanced class distributions to achieve high QWK scores, where the use of balanced essay sets is recommended for future research to enhance AES performance.

Key words: artificial intelligence, language modeling, automated essay scoring (AES), GPT-based models, ChatGPT, GPT-2, data augmentation, back translation, ASAP

Dedication

This thesis is dedicated both to my dearest ones—the Gunduz family, for instilling in me the values of love and gratitude, which are indispensable parts of my life, and to the Ministry of Education of the Republic of Turkiye, for shaping me into the person I am today!

Acknowledgements

Inspired by the saying *'Gratitude is when memory is stored in the heart'*, I will always keep those who have positively influenced my life during my master's program in my heart.

The greatest fortune for a student in academia is having a distinguished supervisor. I would like to express my deepest gratitude to Dr. Mark J. Gierl, who has profoundly shaped my academic journey and career path. During the challenging university acceptance process, he gave me the first opportunity by accepting the university and he was the initial catalyst for my growth. I will forever be grateful for his empathy, his prioritization of my well-being, and the academic freedom and nurturing environment he provided over the past two years.

Learning from experiences is the shortest way to find the truth and sharing information spreads goodness. I am deeply thankful to Dr. Okan Bulut, Dr. Ying Cui, Dr. Ayfer Sayin, and Dr. Maria Cutumisu for the invaluable academic insights and experiences they have provided in my academic journey.

Colleagues with shared objectives can inspire one another and achieve significant milestones. I extend my sincere thanks to all my CRAMER friends, for our mutual support to each other both socially and academically.

Outside of the University, my dear special friends —Nazli Deniz, Gamze Kocak, Yesim Gul, Batu Bora and Mirac Tugcu will always have a special place in my heart. The saying *'Hard times always reveal true friends'* has been validated through our friendship.

The greatest wealth in life is having a loving and supportive family, where challenges are overcome together. My deepest gratitude is reserved for my dearest ones —the Gunduz Family. To my beloved family—Fatma, Halil, Irem, Ahmet and my sweet grandmother, Ayse Gunduz — your unconditional support and love have been the pillars of my life! And lastly, my profound gratitude also extends to the Ministry of Education of the Republic of Turkiye, which has provided me with equal educational opportunities and shaped me into the person I am today!

Table of the Contents

List of Tables

List of Figures

Chapter 1: Introduction

Revolutionary advancements in artificial intelligence, particularly through the development of large language models such as GPT, have profoundly influenced educational technologies, notably in the domain of Automated Essay Scoring (AES). AES systems deploy computer algorithms to evaluate and score written texts, with the goal of emulating the accuracy of trained human raters. This system has increased the potential to assess students' writing skills more objectively, quickly, and consistently (Kumar & Boulanger, 2020; Klebanov & Madnani, 2022).

After the introduction of Chat GPT in 2022, GPT-based models have generated a great deal of interest in the field of language modeling, and the potential of the GPT model for AES has been explored (Gaddipati et al., 2020; Yancey et al., 2023; Mizumoto & Eguchi, 2023). These models have been presented as alternatives to traditional human assessment methods. They serve as viable alternatives because of their improved accuracy and reliability in scoring students' written-response essays. A recent study, conducted by Gunduz and Gierl (2024) using a small set of essays from the Hewlett-sponsored ASAP dataset, revealed that GPT-3.5 and GPT-4, even without bespoke fine-tuning for AES, achieved Quadratic Weighted Kappa (QWK) scores ranging from 0.60 to 0.80, which signify acceptable reliability (Williamson et al., 2012). Nevertheless, the potential of an earlier iteration such as GPT-2 to exceed the performance of its more advanced successors remains unexplored. This study aims to assess the viability, applicability, and efficacy of GPT-2 in the context of AES, specifically using the ASAP dataset. The following section will provide a comprehensive background of the study, clearly define the research problem, and precisely outline the study's objectives and research questions.

Background of the Study

Revolutionary advances in artificial intelligence and large language modeling have led to the development of new models in the field of educational technologies, especially in the field of AES. These systems are being used to score student essays in order to find solutions to the scalability and subjectivity problems faced by traditional human-based assessment methods. The development of AES continues in parallel with advances in natural language processing (NLP) and machine learning (ML). Initially limited to simple statistical methods and linear models, AES has evolved to include deep learning techniques and more sophisticated AI models, particularly transformer-based models. These models have improved the accuracy and reliability of AES systems by allowing researchers to model the deep structure of language and extract rich information from text (Shin & Gierl, 2021; Ramesh & Sanampudi, 2022).

In particular, transformer-based encoders such as BERT and DistilBERT have played a significant role in the development of AES systems by providing a more comprehensive understanding of the text. The decoder-structured Generative Pre-trained Transformer (GPT) series, introduced by OpenAI in 2018, has made significant breakthroughs in the field of language modeling and text generation. These GPT models have demonstrated deeper language understanding and text generation capabilities, opening up innovative possibilities in applications such as AES.

However, the integration of GPT models into AES systems is still an ongoing research challenge. Adapting these models to the AES system introduces several uncertainties, such as interpretability, fairness, and alignment with educational standards. The purpose of my research is to evaluate the potential of GPT-2 in AES. Specifically, I will focus on how accurately and reliably the GPT-2 model scores essays on the Hewlett-sponsored ASAP dataset

(https://www.kaggle.com/c/asap-aes). By evaluating the performance of GPT-2 as an AES system, I aim to contribute to existing work in the field of educational technology and to better understand the potential of transformer-based models in educational assessment.

Research Problem

The purpose of my study is to evaluate the potential of GPT-2 to improve the efficiency and accuracy of AES by utilizing artificial intelligence and a large language modeling approach. Although the success of various language models and transformer models in AES has been evaluated, the effectiveness of GPT models in this area is an under-researched topic.

The use of transformer models such as BERT and DistillBERT as a contextual embedding technique for text representation (Devlin et al., 2019; Sanh et al., 2019) has been evaluated in the AES literature (Firoozi et al., 2022; Klebanov & Madnani, 2022). However, very little work has been done on the essay scoring capabilities of the GPT model, especially after the release of ChatGPT in 2019 (e.g. Mizumoto & Eguchi, 2023). As a result, the extent to which GPT can reproduce the scores given by human raters in the evaluation of AES systems and important dimensions such as scoring consistency that can contribute to the literature have not yet been sufficiently investigated. Furthermore, there is uncertainty as to whether GPT-executed AES outperforms research-based linguistic features in predicting the scores given by human raters.

This research seeks to bridge the gaps identified in existing literature by examining the performance and potential of GPT models within the AES framework. While Gunduz and Gierl (2024) have explored the capabilities of ChatGPT, specifically GPT-3.5 and GPT-4, by directly submitting two essays for assessment without model fine-tuning, there has been a noticeable absence of research on the applicability and efficiency of the antecedent model, GPT-2, in this

domain. The current study is poised to significantly advance the field of educational technology by providing new insights into the development of AES systems. More concisely, the objective is to contribute meaningfully to the body of educational assessment literature, enhancing understanding of the capabilities of a fine-tuned GPT-2 model in AES and addressing existing limitations.

Purpose of the Current Study

Advances in large language modelling and artificial intelligence technologies in the field of AES have provided new opportunities for more efficient and effective educational assessment. In particular, Generative Transformer Models (GPTs) have received a great deal of attention in areas such as language understanding and text generation. However, the effectiveness and applicability of these models in AES is an under-researched topic. Although existing AES methods have the potential to reduce the time and cost requirements of scoring systems that typically rely on human raters, the accuracy and reliability of these systems are critical. There is a knowledge gap regarding the advantages or challenges that models such as GPT-2 offer in terms of accuracy and reliability in AES.

Therefore, my study will fill this gap by evaluating the applicability and performance of GPT-2 in AES on the ASAP dataset. The research will focus on understanding how accurately and consistently the GPT-2 model can produce results under a variety of essay types and scoring rubrics and examine the impact of data augmentation techniques applied through back-translation on its performance. In this context, the research questions are as follows:

1. How accurately and reliably does the GPT-2 model score essays on the ASAP dataset?

2. How do data augmentation techniques applied with back-translation affect the performance of GPT-2 and what are the characteristics of these effects?

In conclusion, this study aims to fill existing gaps in the field of AES, explore the potential of GPT-2 and data augmentation techniques, and provide valuable contributions on how these technologies can be used more effectively in educational evaluation processes.

Chapter 2: Literature Review

Overview of Automated Essay Scoring

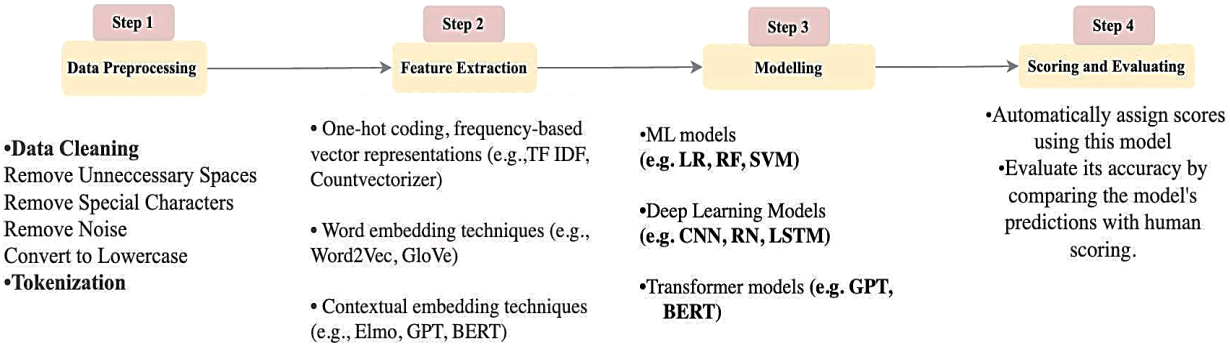
Automated Essay Scoring (AES) is the process of automatically evaluating and scoring student essays using computer software (Shermis, 2014). This technology is often used to assess language skills, composition abilities, and knowledge of specific topics. AES--characterized by its speed, efficiency, objectivity, and big data processing capabilities--is also preferred for efficiently and objectively assessing student essays (Yan, Rupp, & Foltz, 2020). While the manual assessment process involves scoring student essays individually by educational experts, AES significantly speeds up this process by automatically and quickly assessing written essays. Subjective differences and inconsistencies in the manual assessment process are other common challenges, especially among human raters. However, AES minimizes such problems by applying the scoring criteria in a consistent manner, making the evaluation process more objective. Moreover, the big data processing capabilities of AES provide the ability to quickly evaluate the essays of millions of students, making the measurement and evaluation processes more effective and useful. Therefore, AES has gained an important place in education and has become an important tool for improving assessment processes in this field.

As can be seen in Figure 1, the AES process can be described in four steps (see Gierl et al., 2014). 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 (ML) and natural language processing (NLP) have led to advances in text

representation techniques (step 1 and step 2) and modeling algorithms (step 3) (Firoozi et al., 2023), filling the gaps in this field.

Figure 1

The AES process described in four steps (see Gierl et al., 2014)



Step 1 and step 2 are used to implement text representation techniques to convert the written text into a numerical format and then transform it into suitable inputs for machine learning and deep learning models, respectively. Various techniques used in this step include one-hot coding, frequency-based vector representations (Salton, 1974), word embedding techniques (Mikolov et al., 2013), and contextual embedding techniques (Peters et al., 2018, Radford et al., 2018). While frequency-based techniques have certain shortcomings because they cannot fully reflect semantic relations and semantic similarities, these shortcomings can be addressed using word embedding techniques (e.g., Word2Vec, GloVe). Likewise, word embedding techniques, which struggle to fully understand context at the word level, can be overcome with contextual embedding techniques (e.g., Elmo, GPT, BERT). Contextual embedding techniques, such as BERT and GPT are able to vectorize word positions in the text more efficiently through their positional encoding and embedding structures. These features allow transformer models to perform at a high level in the field of deep learning, which is a topic that should also be explored in terms of vectorization.

For the modeling process in Step 3, prior to transformer models, deep neural network (DNN) and recurrent neural network (RNN) models, which incorporate previous advanced

encoder-decoder structures for sequential data tasks, were used (Alikaniotis et al., 2016; Tay et al., 2018; Shin & Gierl, 2021). Recurrent neural network (RNN) models are specifically designed to operate on sequential data and have the ability to memorize previous states and information. However, RNNs are limited because they cannot effectively manage long-term dependencies (Nugaliyadde et al., 2019). To overcome this challenge, transformer models, especially models such as GPT, which include an attention mechanism, offer structures that can take into account the relations of a word with other words and thus handle longer-range connections. Vaswani et al. (2017) published a paper titled 'Attention Is All You Need' for the NeurIPS conference. They presented a transformer architecture based on the attention mechanism that outperformed RNNs in areas such as learning, language generation, and text classification (Irissappane, 2020). Thus, the GPT model and similar transformer-based models are important advances in terms of effectively solving the long-term dependency challenge of DNN models.

The Mechanism of Transformer-Based Systems

Transformer and Attention Mechanism

In 2017, the paper 'Attention Is All You Need' revolutionized the field of natural language processing (NLP) by introducing the transformer architecture. A transformer architecture is particularly notable for its impressive success in NLP tasks such as machine translation and time series prediction, text summarization, and text classification (Chernyavskiy et al., 2021). The attention mechanism is a distinctive mechanism that sets transformers apart from previous models.

Before the advent of transformers, the models used for sequence-to-sequence tasks such as language translation were usually deep neural network (DNN) and recurrent neural network (RNN) based. For example, for machine translation, the seq2seq model proposed by Sutskever (2014) consisted of an LSTM-based encoder that takes a sequence of tokens and

converts it into a vector and an LSTM-based decoder that converts this vector into a sequence of tokens. However, the sequential nature of seq2seq models does not allow for parallel computation and the generation of longer sequences has brought problems of information loss and vanishing gradients (Hochreiter et al., 2001). In 2016, steps were taken to solve this problem with sequential training by using a parallel structure for encoding input sequences of different lengths through an attention mechanism (Wu et al., 2016).

In 2016, Google Translate started to replace its old statistical machine translation approach with a new neural network-based approach incorporating LSTM and an 'additive' type of attention mechanism. In 2017, Vaswani et al. (2017) revolutionized the original (100M-dimensional) encoder-decoder-transformer model by introducing a faster (parallel or decodable) attention mechanism and transformer model, as described in their paper 'Attention is All You Need'.

For an RNN, the phrase is processed sequentially, one token at a time. Consider the sequence $x = (x_1, x_2, x_3, x_4, x_5)$, corresponding to the words "Love"(x_1), "is"(x_2), "all"(x_3), "you"(x_4), "need"(x_5). The RNN processes each word based on its current input and the previous hidden state, thus sequentially building up the context. The model starts with an initial hidden state h_0 (often zeros), At each timestep t, the hidden state ht is updated based on the current input xt and the previous hidden state h t-1 as follows:

$$h_t = f(W_{hh}.h_{t-1} + W_{xh}.x_t + b_h) \quad (1)$$

where f is activation function, W_{hh} is the weight matrix for the hidden state, W_{xh} is the weight matrix for the input, and bh is the bias term (see Equation 1). This iterative process allows the Recurrent Neural Network (RNN) to maintain a form of memory by capturing and carrying forward the context from one token to the next. However, the standard RNN architecture, as defined by Equation 1, can encounter difficulties with long-range dependencies. The primary issue is the diminishing influence of initial inputs on subsequent

states.For instance, the initial token ‘ x_1’ gradually has less impact on downstream tokens such as ‘x_4’. This diminishing influence is a result of the compounding effects that can lead to vanishing gradients, posing a significant challenge for the model to learn dependencies over longer sequences.

In contrast, a transformer model processes the entire sequence in parallel, allowing each word to directly attend to every other word in the sequence, thereby capturing the contextual relationships more comprehensively. Each token x_i is encoded simultaneously with positional information to maintain the sequence order. For each word, the transformer calculates attention scores reflecting how much focus it should put on other parts of the sentence. The key formula for this calculation is given in Equation 2. As Equation 2 demonstrates, the attention mechanism determines the level of focus or attention, ‘love’ should give to other words such as ‘is’, ‘all’, ‘you’, ‘need’ simultaneously, rather than limiting attention to only the previous context,

$$Attention(Q, K, V) = softmax\left(\frac{Q.K^T}{\sqrt{d_k}}\right)V \quad (2)$$

Where $Q(query)$, $K(key)$, $V(value)$ are matrices representing the input sequence, and d_k is the dimension of the key vectors. The output is a set of vectors where each vector is a weighted sum of the entire sequence's representations, reflecting both local and global context. This allows the token ‘x_2’ to be directly influenced by ‘x_1’ as much as it is by ‘x_3’ or ‘x_4’, effectively capturing long-range dependencies. Hence, in encoding or decoding the representation of an input sequence, the attention mechanism allows transformers to learn the context of the input by parallelizing all the surrounding inputs within training examples (Vaswani et al., 2017).

In conclusion, while RNNs process the phrase ‘Love is all you need’ sequentially, potentially losing context from earlier parts of the sequence due to limitations like vanishing gradients. Transformers process each word in the context of every other word in the sequence

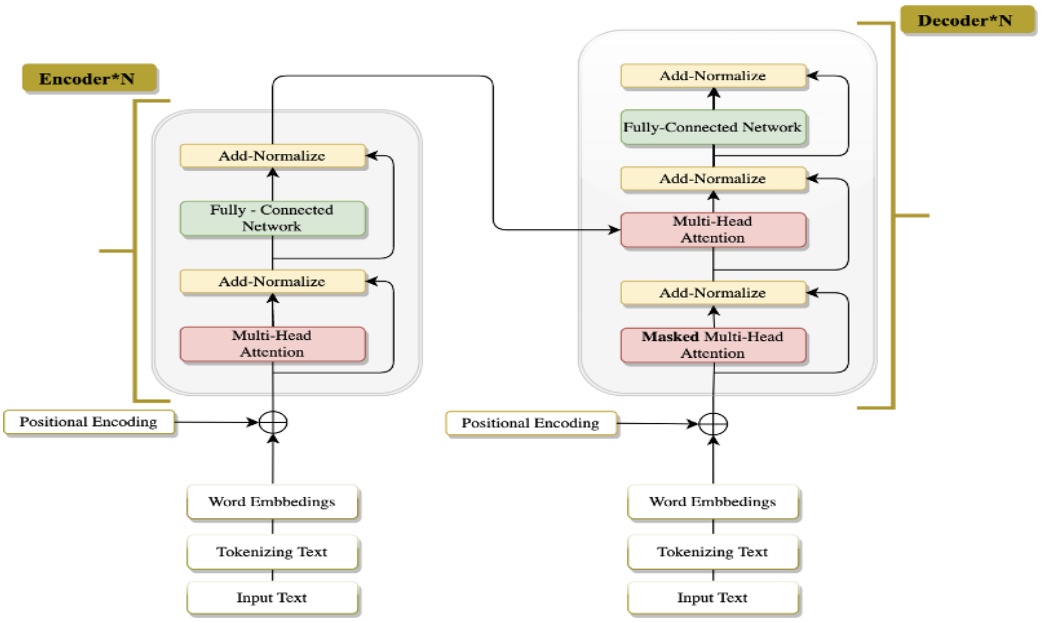
simultaneously. This parallel processing enables a more holistic understanding of the sequence, better capturing the nuanced meanings and relationships between words across the entire phrase. Such capabilities make transformers particularly powerful in handling complex linguistic constructs and long-range dependencies in natural language understanding and generation tasks.

Transformer Architecture

A summary of the transformer architecture is presented in Figure 2. The left half is designated as the encoder, while the right half constitutes the decoder stack structure. Both the encoder and decoder consist of a block that is repeated Nx times (where Nx is specifically set to 6). Each block encompasses layers including a multi-head attention mechanism and a location-based fully connected feed-forward network. Surrounding each layer is a residual connection coupled with a normalization layer.

Figure 2

Transformer Architecture (Vaswani et al., 2017)



Beyond the encoder layers, the decoder incorporates a masked self-attention layer at its base. This masked self-attention mechanism maintains the autoregressive structure of the

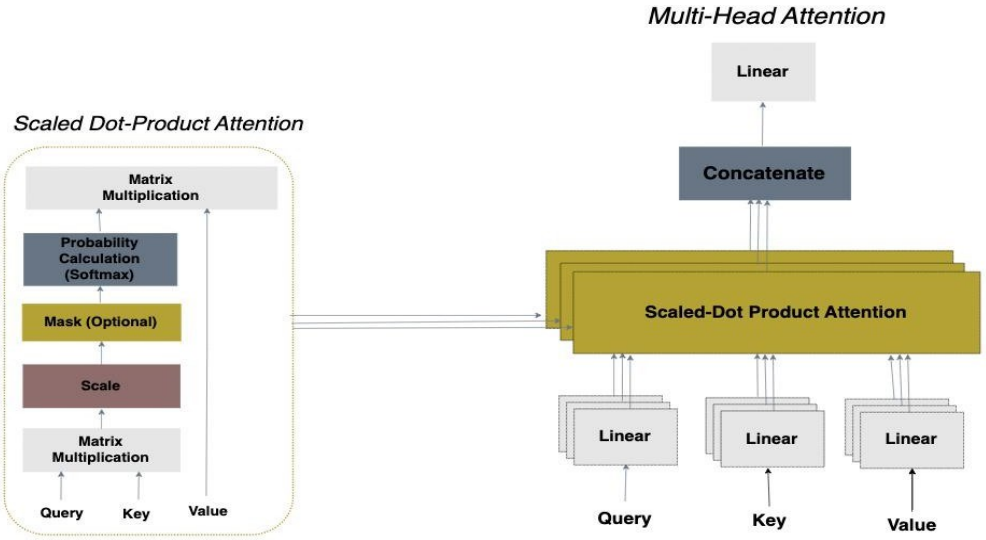
decoder, enabling the language model to predict the next word by solely considering the current information for each prediction, masking future positions. Consequently, this ensures that the prediction for the i-th position relies exclusively on the known outputs at positions below i.

Scaled dot-product attention is a key component in the self-attention mechanism of the transformer model (see Figure 3). This attention mechanism enables the model to focus on different parts of the input sequence with varying levels of importance. Given a query matrix Q, a key matrix K, and a value matrix V, the attention scores are calculated by taking the dot product of the query and key matrices, divided by the square root of the dimension of the key vectors.

The scaled dot-product attention mechanism is efficiently summarized by Equation 2. In this equation, $\sqrt{dk}$ represents the dimensionality of the key vectors. By applying the softmax function, the attention scores are normalized, allowing the model to compute a weighted sum of the values to produce the final output. This mechanism enables the model to assign varying attention weights across different segments of the input sequence. It adeptly captures dependencies and relationships between words, thereby enhancing the model's capacity for understanding context in a context-aware manner.

Figure 3

Scaled Dot-Product Attention and Multi-Head Attention Mechanism (Vaswani et al., 2017)



In the preceding self-attention mechanism, a single attention head is employed which includes processing query (Q), key (K), and value (V) matrices to determine the relationships between one word and the other words. Conversely, the multi-head attention mechanism operates by concurrently processing the same Q, K, and V matrices under different attention heads, producing parallel sets of information outputs. Each attention head focuses on distinct features or relationships, representing various facets of a word by attending to diverse aspects. The output of these multiple attention computations is then combined in a process defined by Equation 3 (see Equation 3), where each head is an individual attention computation with its own set of projected Q, K, and V matrices (Vaswani et al., 2017):

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_n).W_0 \quad (3)$$

where $head_i = Attention(Q.W_{Qi}, K.W_{Ki}, V.W) \; V_i.$

Multi-Head Attention in Equation 3 improves upon the traditional attention mechanism by utilizing multiple attention heads to capture diverse aspects of the input data.The process begins with the linear transformation of the query (Q), key (K), and value (V) inputs using learned weight matrices W_Q, W_K, and W_V for each head independently.

Next, scaled dot-product attention is computed separately for each of the h heads. The resulting attention outputs from all heads are then concatenated. Finally, this concatenated result undergoes a linear transformation using a weight matrix W_0 to produce the final output of the multi-head attention mechanism. This multi-head approach allows the model to attend to various parts of the input data simultaneously, thereby enhancing performance and understanding.

Generative Pre-Trained Transformer (GPT) from Transformers

The Emergence and Development of GPT. A generative pre-trained transformer (GPT) is an artificial neural network introduced by OpenAI in 2018 and belongs to the class of large language models (Yenduri et al., 2023). This transformer serves as a pioneering approach in the field of generative AI, pre-trained on large amounts of unlabeled text data and used in tasks such as language understanding, sentence completion, translation, and text generation (Radford et al., 2019).

GPT models consist of a decoder-only structure, which gives them the ability to generate text based on input context. While the transformer architecture introduced in 2017 included encoder and decoder blocks (Vaswani et al., 2017), other large language models that emerged in 2018, notably BERT, focused primarily on the encoder structure and not on context understanding, which limited its ability to generate text. Around the same time, OpenAI, in its paper "Generative Pre-Learning Language Comprehension Improvement", introduced GPT-1, which consists of a decoder structure, laying the foundation for generative pre-trained models capable of generating novel and human-like content (Radford et al., 2019).

The GPT series has evolved significantly over time, with each new version having an increasing number of parameters and being trained on larger training sets. GPT-1, the first model introduced in 2018, has 117 million parameters and is based on the 12-head

transformer decoder structure. Trained on the BookCorpus dataset using 4.5 GB of text data, GPT-1 stood out for its language understanding and text generation capabilities. GPT-2, the successor to GPT-1, was introduced in 2019, and while retaining its basic architecture, GPT-2 has come to the fore with a significant increase in the number of parameters and the size of the training dataset. With 1.5 billion parameters, GPT-2 was trained on 40 GB of text data on the WebText dataset.

GPT-2 has been successfully used in many fields with its language modeling capabilities. In particular, Sahin (2021) showed that GPT-2 can perform well in classification tasks, which highlights the potential of GPT-2 in text classification. This implies that the GPT-2 model can be effectively used for score classification tasks in the field of AES. Although GPT-2 has been shown to give positive results in classification problems, the fact that it has not been tested in the field of AES motivated me to conduct my study based on GPT-2. Moreover, the fact that the GPT-2 tokenizer and GPT-2 model functions are readily available in the transformer library provided for free by Hugging Face (Wolf et al., 2019), whereas the tokenizers of the GPT-3 and GPT-3.5 models are provided for a fee by OpenAI, is another reason for choosing the GPT-2 model for this study. Moreover, although Gunduz and Gierl (2024) measured the performance of ChatGPT-3.5 and ChatGPT-4 in AES using the same ASAP dataset, this thesis aims to explore whether the GPT-2 model can perform comparably to GPT-3.5 and GPT-4 when appropriately fine-tuned. The primary objective of this thesis is to evaluate the performance of the GPT-2 small model in AES using the ASAP dataset while the GPT-2 model series include four different architectures with 12, 24, 36, and 48-headed Transformer decoders.

GPT-3, the successor to GPT-2, is an artificial neural network introduced in 2020 and contains 175 billion parameters. This model extends the features of the previous GPT series to build a large language model. The training data consisted of 499 billion tokens from

CommonCrawl, WebText, English Wikipedia and two book collections. In 2020, Microsoft announced the exclusive licensing of GPT-3 for Microsoft's products and services, following a billion-dollar investment in OpenAI. This means that only Microsoft is authorized to provide exclusive access to the GPT-3 base model. After GPT-3, the GPT-3.5 version was developed and used to implement the successful chat bot ChatGPT. The latest version, GPT-4, released on March 14, 2023, is notable for having image input capability in addition to the language model. This feature suggests that the model has been trained on a larger dataset than previous versions. However, although OpenAI does not disclose technical details and statistics about GPT-4, George Hotz claims that GPT-4 has 1.76 trillion parameters. Table 1 shows detailed information about the GPT series.

Table 1

OpenAI's "GPT-n" Series

Model Name	Architecture	Parameter Count	Training Data	Release Date
GPT-1	12-headed Transformer decoder	117M	BookCorpus: 4.5GB text	2018
GPT-2 *S-M-L-XL*	12-24-36-48-headed Transformer decoder+ GPT-1, but with modified normalization	1.5B	WebText: 40GB text	2019
GPT-3	6-headed Transformer decoder+ GPT-2, but with modification to allow larger scaling	175B	499B tokens (CommonCrawl, WebText, English Wikipedia, Books1, Books2)	2020
GPT-3.5	Undisclosed	175B	Undisclosed	March 15, 2022
GPT-4	Undisclosed	Estimated 1.7T	Undisclosed	March 14, 2023

GPT-2 Architecture. Generative Pre-trained Transformer 2 (GPT-2) is the second major language model in OpenAI's series of basic GPT models and is a GPT version with around 1.5 billion parameters (Radford et al., 2019). The Small GPT-2 model consists of 117

million parameters, 12 layers, an input size of 768, and 768 hidden units. The model has four different configurations depending on the number of parameters it contains, as shown in Table 2.

Table 2

Hyperparameters for Four Different Sizes of GPT-2 Model

Model Name	Parameters	Layers	Input Size	Hidden Units
Small GPT-2	117M	12	768	768
Medium GPT-2	345M	24	1024	1024
Large GPT-2	774M	36	1280	1280
XL GPT-2	1558M	48	1600	1600

Note. M represents a million.

GPT-2 uses a word embedding layer to convert each word into a vector representation. In addition, positional encodings are used to add the position information of the symbols in the sentence and these encodings are integrated into the input embedding layer of each symbol. The GPT-2 small model provides each word representation with a 768-dimensional embedding and a 768-dimensional positional encoding to obtain the appropriate input to the decoder blocks.

Before the attention mechanism, layer normalization is applied to speed up the training process of the model and provide a more stable progression. Layer normalization (LN) standardizes the outputs with the results obtained from the intermediate layers, called multi-headed attention sublayers or feed-forward network sublayers. This approach allows the model to be trained more quickly and helps the optimization process. Next, the word vectors enter the decoder blocks, which consist of attention and feed-forward neural networks.

The attention mechanism uses query, key, and value vectors to contribute to understanding the context of each word in relation to other words. The attention scores within

the attention mechanism are normalized through the softmax layer, which determines which words should be given more attention than others. However, this attention mechanism, unlike mechanisms in other models, includes a masked self-attention mechanism that prevents interaction on tokens to the right of a position. That is, such a self-attention mechanism limits information from tokens to the right of the computed location. After each attention block, the model uses a feed forward network structure to process language features in a more complex and meaningful way. The outputs of the attention blocks are passed through two matrices of size (n x 4n) and (4n x n), respectively, resulting in an n-dimensional vector. Each neural network layer is specifically designed to process the language features learned by the model in more depth.

To summarize, each "block" is a combination of these operations and the GPT-2 small model consists of 12 blocks. In GPT-2, a single "decoder block" consists of a layer normalization followed by a multi-head attention block, followed by another layer normalization and a feedforward layer. The feed-forward layer and the multi-head attention block have jump connections.

Recent Applications of GPT-Based AES

Recently, GPT has become an important language model for evaluating performance in automated scoring. In particular, score generation with data augmentation techniques for minority scoring groups has shown potential for use in short answer and essay-style written assessments. GPT has been used to generate augmented responses for minority scoring classes (e.g. Fang et al., 2023), to score reading comprehension, short answer tests (e.g. Gaddipati, et al., 2020; Yancey et al., 2023; Henkel et al., 2023), and long answer tests such as essays (e.g. Mizumoto & Eguchi, 2023).

In a study on Automatic Short Answer Grading (ASAG), Gaddipati et al. (2020) evaluated the effectiveness of pre-trained transfer learning models. In this study, the

performance of transfer learning models such as ELMo, GPT, BERT and GPT-2 on ASAG using pre-trained embeddings was investigated. The results showed that ELMo outperformed the other models. Yancey et al. (2023) evaluated the usability of large language models (LLMs), specifically GPT-3.5 and GPT-4, for the automated scoring of short essay responses written by learners of English as a second language (ESL). The results show that GPT-4 can perform almost identically to modern Automated Writing Evaluation (AWE) methods when calibration samples are provided. However, it is emphasized that agreement with human raters may vary depending on the test-taker's first language. Similar work is also investigating the use of generative large language models (LLMs) such as GPT-4 to assess short answer reading comprehension questions in low and middle-income countries (LMICs) (Henkel et al., 2023).

There are also studies evaluating the potential of GPT models for AES. Mizumoto and Eguchi (2023) performed automated scoring of 12,100 essays using the ETS Corpus of Non-Native Written English (TOEFL11) dataset and compared the scores obtained using the GPT-3 text-davinci-003 model with human evaluations. The findings showed that GPT has a certain level of accuracy and reliability in AES and can provide a valuable supplement to human evaluations. Furthermore, it was found that the use of language features such as lexical diversity, lexical sophistication, syntactic complexity, and cohesion in combination with GPT improves AES performance.

They used GPT-4 to generate augmented responses that specifically student-written responses for the minority of the scoring classes (Fang et al., 2023). The augmented data was trained by the DistilBERT model for automated scoring. The results showed that GPT-4 augmented data significantly improved the performance of the scoring model. The use of GPT-4 can improve the performance of automated scoring models and has the potential to be applicable in various educational contexts.

It is clear that GPT has significant potential for use in educational assessment areas such as data augmentation, short answer, and essay assessment. However, in the field of AES, the performance of GPT-based models, particularly on the Automated Student Assessment Prize (ASAP) dataset supported by Hewlett in 2012, has only been examined in the poster study by Gunduz and Gierl (2014). This study has comparatively analyzed the automated scoring performance of different versions of the GPT models (GPT-3.5 and GPT-4) on the essay sets within the ASAP dataset and evaluated the models' adaptation to three different prompt scenarios (zero-shot, one-shot, few-shot). ASAP is a significant dataset for AES. Although the performance of a number of models has been evaluated in the literature, the contribution of GPT-based models to AES has still not been explored in detail. This paper aims to make a significant contribution to the literature on training evaluation by understanding the potential of GPT-based models, namely the GPT-2 model, in AES and addressing the limitations in this area.

Chapter Summary

In this chapter, I provided an overview of AES, an important tool in educational assessment that leverages computer software to evaluate and score essays. The AES process consists of four steps, involving preprocessing, conversion of essays to numerical vectors, creating a scoring model using machine learning or deep learning, and evaluating the model's performance to determine the model's predictive power relative to human scoring. The chapter underscores advancements in machine learning and NLP, particularly the transition from traditional text representation techniques to more advanced methods like contextual embeddings, exemplified by BERT and GPT models, which offer more effective vectorization of words in texts.

The limitations of recurrent neural networks (RNNs), such as the ineffective handling of long-term dependencies, are discussed and contrasted with transformer models which

utilize attention mechanisms to maintain complex relationships within the text. The attention mechanism is recognized for its success in various NLP tasks and is detailed as a revolutionary approach that has surpassed previous models like RNNs in learning, language generation, and text classification.

Chapter 3: Method

In this chapter, I provide a detailed explanation of the dataset used in the present study, data preprocessing, the model development and architecture, and the evaluation metrics used for model validation.

Dataset for Automated Essay Scoring

The dataset used in this study was collected and released in 2012 with sponsorship of the Hewlett Foundation for the Automated Student Assessment Prize (ASAP) competition. The objective of the ASAP competition was to encourage data scientists and machine learning experts to create cost-effective, rapid, and efficient solutions for the automated assessment of student essays. The focus of the competition was to analyze whether AES systems could produce results similar to those of trained human raters (Shermis, 2014). The summary of the descriptive characteristics in the dataset is provided in Table 3.

This dataset encompasses a collection of eight different essay sets, each of which was written by students at different age groups ranging from 7th to 10th grade (Grade Level). These essays cover a range of essay types, including persuasive, narrative, and source-dependent topics (Essay Type). Notably, five of these composition sets were authored by 10th-grade students, two by 8th-grade students, and one by 7th-grade students.

The training dataset comprises a total of approximately 13,000 student essays. Each of these eight essay sets varies in size, containing essays between 700 and 1800 (Training Set Size). The average length of these essays falls within the range of 150 to 650 words (Average Essay Length), reflecting the varying complexity and content of the essays.

Essays were scored by two or three human raters using different types of rubrics with varying ranges. The raters conducted their evaluations in accordance with the scoring rubrics

categorized into three main headings: holistic, specific, and composite (Rubric Type). Only essay 8 was assessed by three human raters, while the remaining seven essays were evaluated by two human raters each, and the classification score ranges differed from each other (raters' range). Lastly, the score ranges provided by the raters vary according to the type of graded rating key and the domain score, which is a combination of rater 1 and rater 2 (Domain Score Range).

Table 3

Descriptive Statistics of the ASAP Dataset

					Evaluation Methods		
Essay Set	**Grade Level**	**Essay Type**	**Training Set Size**	**Average Essay Length**	**Rubric Type**	**Raters' Range**	**Domain Score Range**
1	8	Persuasive	1783	350	Holistic	2 raters 1-6	2-12
2a	10	Persuasive	1800	50	Trait	2 raters 1-6	1-6
2b	10	Persuasive	1800	50	Trait	2 raters 1-4	1-4
3	10	Dependent	1726	50	Holistic	2 raters 0-3	0-3
4	10	Dependent	1772	50	Holistic	2 raters 0-3	0-3
5	8	Dependent	1805	50	Holistic	2 raters 0-4	0-4
6	10	Dependent	1800	50	Holistic	2 raters 0-4	0-4
7	7	Expository	1569	50	Composite	2 raters 0-12	0-24
8	10	Expository	723	50	Composite	3 raters 0-30	0-60

Analysis of Score Ranges for Each Essay Set

Raters employing the holistic grading method focus on the overall communicative aspect of the essay and the central message conveyed by the author (Cooper, 1977; Weigle, 2002). In this type of rubric, a single score representing the overall quality of the essay is assigned (Ebel & Frisbie, 1986; Goulden, 1994; Plakans & Gebril, 2015). The composite scales require raters to assign separate scores for specific aspects such as the organization, language usage, and writing conventions of the essay.

Next, I provide a detailed analysis of the essay sets in this dataset, along with the human raters' score ranges and domain score range rules. Table 4 shows the relationship between rater 1, rater 2, and Domain Score within the ASAP dataset.

Table 4

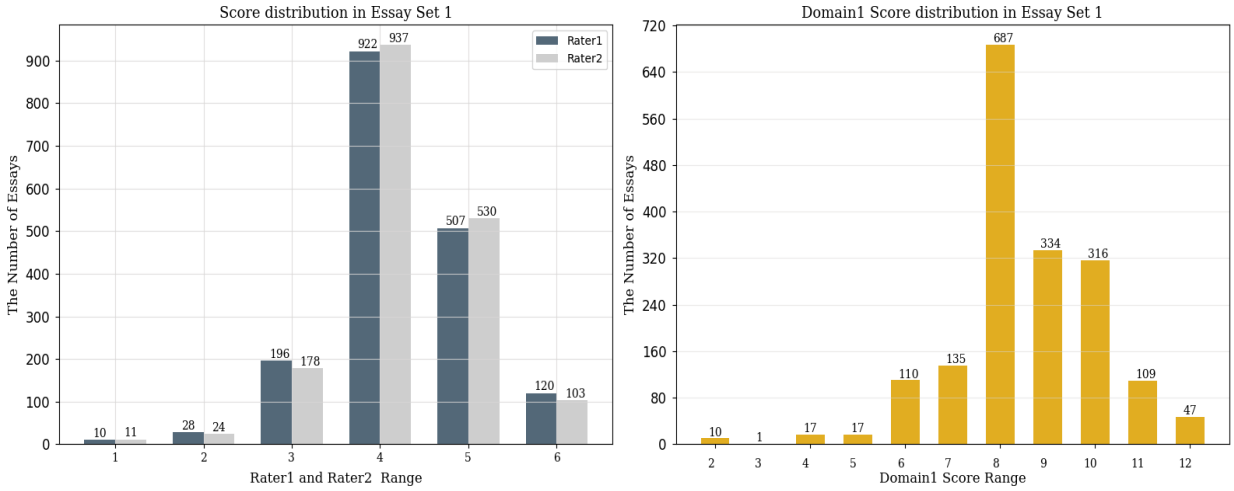
Raters' and Domain Score Ranges for Essay Sets

Essay Set	Rubric Type	Raters' Score Range	Domain Score Range	Additional Information (For Each Row)
1	Holistic	1-6	2-12	Domain Score equals the sum of rater 1's and rater 2's.
2a	Trait	1-6	1-6	Domain Score equals rater 1's.
2b	Trait	1-4	1-4	Domain Score equals rater 1's.
3	Holistic	0-3	0-3	Domain Score equals max (rater 1, rater 2)
4	Holistic	0-3	0-3	Domain Score nearly equals max(rater1, rater 2)
5	Holistic	0-4	0-4	Domain Score nearly equals max (rater1,rater 2)
6	Holistic	0-4	0-4	Domain Score nearly equals max (rater 1, rater 2)
7	Composite	0-12	0-24	Domain Score equals the sum of rater 1's and rater 2's.
8	Composite	0-30 (for rater 1-2) 0-60 (for rater 3)	0-60	Resolved Score equals the sum of rater 1's and rater 2's Resolved Score equals rater 3's.

Analysis of Essay Set 1. This essay set has been evaluated by two raters using a holistic rubric scale. In line with the characteristics of the Rating Scale, both rater 1 and rater 2 provided scores within the range of 1-6. The score for this set, referred to as "Domain1_score," was obtained from the sum of the two raters' scores, resulting in a score that varies from 2 to 12.

Figure 4

Score Distribution of Essay Set 1



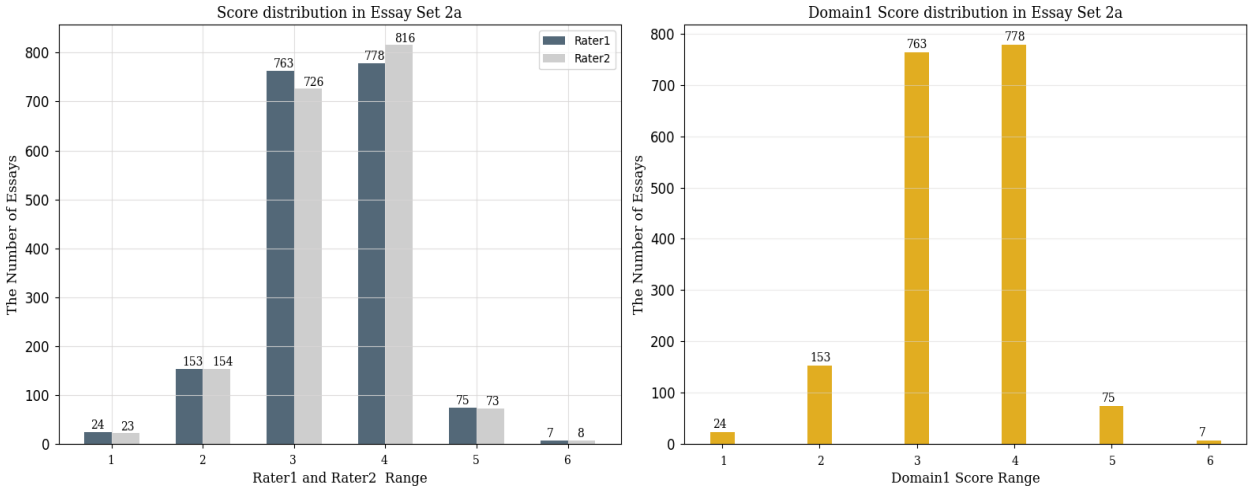
The score distribution of Essay Set 1 is shown in Figure 4. When examining the 'Domain1 Score Range,' it was observed that there was an insufficient sample in the low-score category. While there was a higher number of students receiving scores between 6 and 10 within the sample, there was a lack of an adequate number of samples in the low-score category. This outcome introduced challenges during the model training process. As an example, only one student achieved a score of 3, and just 10 student essays obtained a score of 2.

Analysis of Essay Set 2a. The rubric for "Written Expression" was designed to assess students' written expression skills within a scoring framework ranging from 1 to 6. Rater 1 and rater 2 scored students' essays in the range from 1 to 6 in Essay Set 2a, as shown in Figure 5.

It's worth noting that "Domain_score1" corresponds to rater 1's score for each essay. Upon analyzing the score distributions, it becomes apparent that only seven students achieved the highest score of 6, while 778 students received an average score of 4, indicating an unbalanced distribution of scores.

Figure 5

Score Distribution of Essay Set 2a

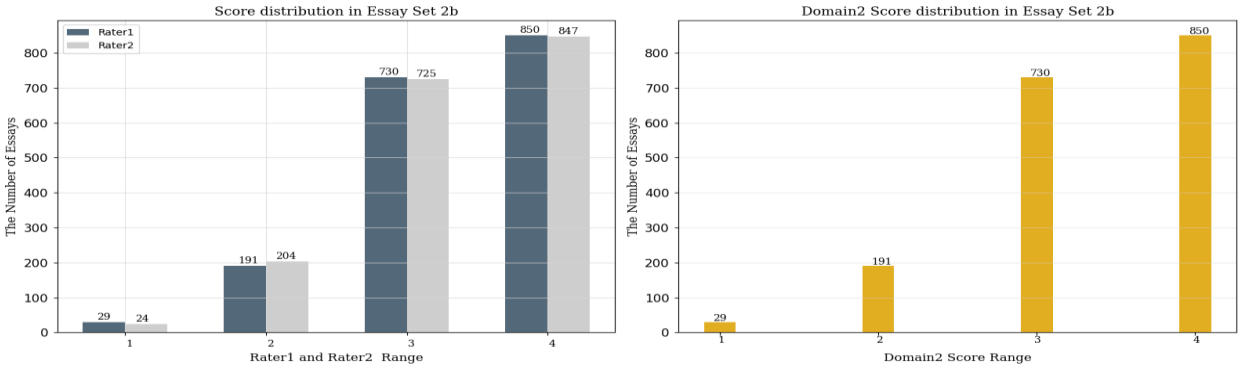


Analysis of Essay Set 2b. The "Language Conventions" rubric has been designed to evaluate a student's language proficiency using a rating scale ranging from 1 to 4. In Essay Set 2b, as shown in Figure 6, in the context of essays assessed by two different raters, "domain2_score" represents the score assigned by rater 1. This rubric evaluated the student's mastery of grammatical structures, spelling, punctuation, and sentence construction.

Furthermore, in Essay Set 2b, there was a notable disparity in the score distribution, with 29 students receiving the lowest score of 1, while 850 students achieved the highest score of 4. This unbalanced distribution could potentially exert a detrimental influence on training the model and evaluating its subsequent classification accuracy.

Figure 6

Score Distribution of Essay Set 2b

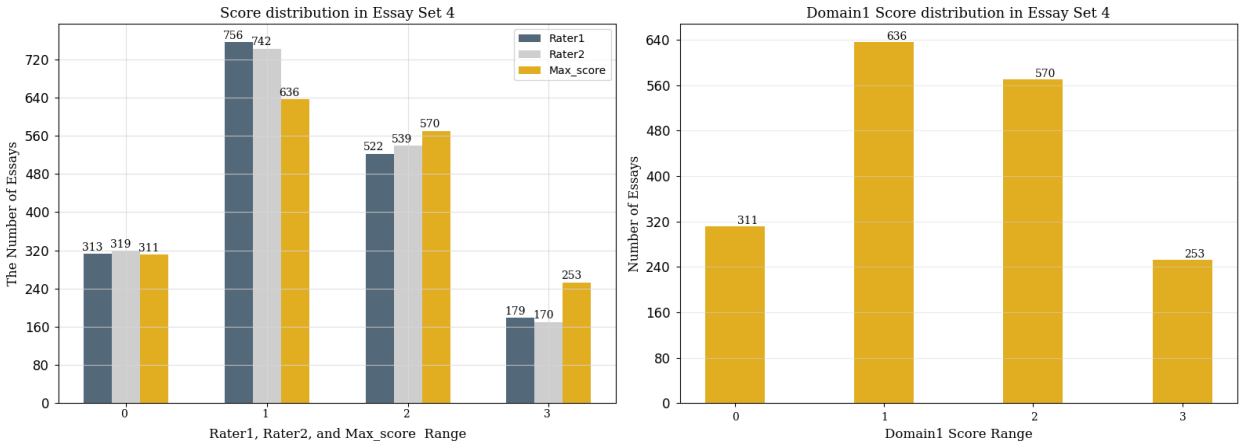


Analysis of Essay Sets 3, 4, 5, and 6. In the Essay Set 3 and 4, human raters scored the essays within the range of 0 to 3. The "Domain1_score" in these sets was determined as the maximum score provided by rater 1 or rater 2 for each essay. In Essay Set 5 and 6, human raters evaluated the essays within the range of 0 to 4. In these essay sets, the "Domain1_score" was also assigned as the maximum score provided by rater 1 or rater 2 for each essay.

As shown in Figure 7, Essay Set 4 has a noteworthy characteristic: It contains a balanced number of samples in all score categories. This balanced distribution will enhance the effectiveness of the model's training process.

Figure 7

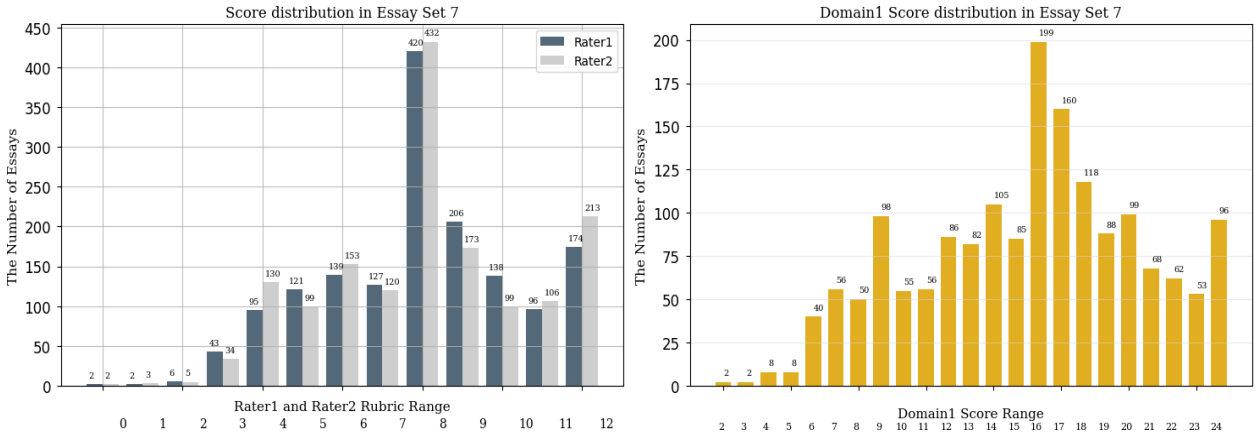
Score Distribution of Essay Set 4



Analysis of Essay Set 7 and 8. For Essay Set 7, essays were scored between 0 and 3 points based on four distinct features (Ideas, Organization, Style, Conventions). Each rater could assign a maximum score of 12 (calculated as 4 features multiplied by a maximum of 3 points each). Consequently, the "Resolved Score" for each essay was computed as the sum of the assessments provided by both rater 1 and rater 2, yielding a score that ranged from 0 to 24. The score distribution of Essay Set 7 is presented in Figure 8.

Figure 8

Score Distribution of Essay Set 7

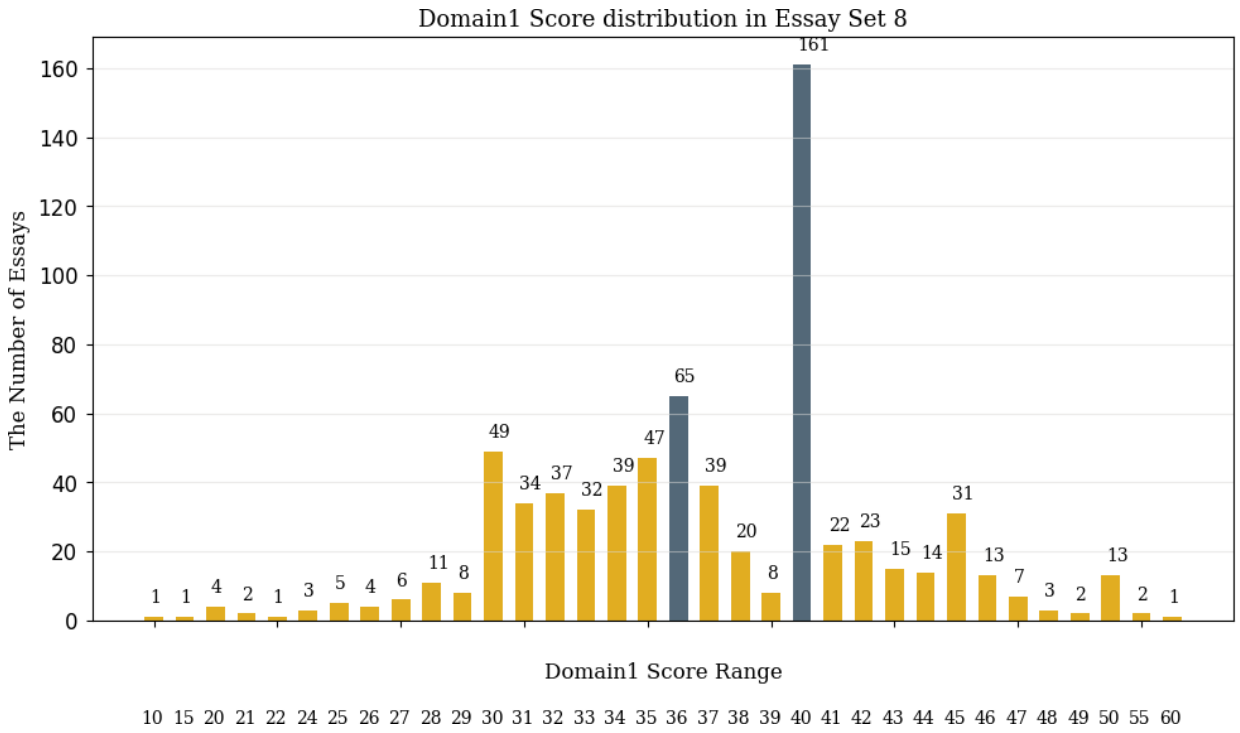


As shown in Figure 9, in Essay Set 8, each essay was scored by two or three human raters across six distinct writing traits (Idea, Organization, Voice, Word Choice, Sentence Fluency, and Conventions), and the scoring range for each trait was set between 0 and 6. Rater 1 and rater 2 assigned scores to the traits Idea (I), Organization (O), Style (S), and Conventions (C) using weighted sums of (1, 1, 1, 2), respectively. Since each trait can receive a maximum score of 6, the maximum total score that each rater could give is 30 (6 multiplied by 5 traits). This indicates that the composite scoring range for each rater spans from 0 to 30. Subsequently, the domain 1 score, which falls within the range of 0-60, is obtained by summing the scores given by rater 1 and rater 2. On the other hand, rater 3 employs weighted sums of (2, 2, 2, 4) for the same traits

(I, O, S, and C) to determine scores, resulting in a scoring range of 0 to 60. The only scores with more than 50 essay examples are 36 and 40, and these have been shown in blue.

Figure 9

Score Distribution of Essay Set 8



Different rubric types and wide score ranges can potentially complicate the model's learning process, which could, in turn, adversely affect the model's performance. Therefore, domain scores have been standardized by narrowing their intervals. Notably, this adjustment has been implemented in specific essay sets, namely Essay Sets 1, 7, and 8. The details of the scaling process are described in the Data Preprocessing section.

Data Preprocessing

Text-Based Preprocessing

Prior to converting the essay responses into word embedding vectors, I applied essential preprocessing procedures to reduce any noise in the model's learning process and predictions. Specifically, I converted all words to lowercase and performed lemmatization using the Python NLTK library (Bird et al., 2009). Following these steps, the processed responses underwent tokenization.

Tokenization is the process of dividing text into smaller units called tokens, which can be words, phrases, subwords, or characters. (Vaswani et al., 2017). I utilized the GPT2 Tokenizer function from the Hugging Face "Transformers" library for tokenizing the preprocessed responses. Each essay within the dataset may possess varying word counts or character length. However, it is important to recognize that deep learning models typically require inputs of fixed dimensions (Hartford, 2016).

Hence, to align the tokenized texts with this standard, we implemented "padding" and "truncation" procedures. Padding and truncation are preprocessing techniques used in transformers to ensure that all input sequences have the same length. Padding refers to the process of adding extra tokens (usually a special token such as [PAD]) to the end of short sequences so that they all have the same length. Truncation, on the other hand, refers to the process of cutting off the end of longer sequences so that they are all the same length (Chollet, 2018).

Score-Based Preprocessing

Each essay has been scored by two or three human raters. The domain score in each essay set was calculated based on rater 1 and rater 2, and these equations are constructed by assigning

specific weights from rater 1, rater 2, and rater 3 (See Table 4). However, domain scores in some essay sets, namely Essay Set 1, Set 7, and Set 8 were rescaled as the scores did not exhibit a normal and balanced distribution.

As shown in Table 4, essay sets other than Essay Sets 1, 7, and 8 contain score levels ranging from 0 to 6, while Essay Sets 1, 7, and 8 encompass a wider range of score classes. Specifically, Essay Set 1 includes 11 different score classes ranging from 2 to 12. Essay Set 7 comprises 23 different score classes ranging from 2 to 24. Essay Set 8 consists of 34 distinct score classes ranging from 10 to 60 (see in Table 5).

Consequently, the scores for Essay Set 1 were adjusted to a range of 1-6, those for Essay Set 7 to a range of 0-3, and those for Essay Set 8 to a range of 1-8. The rationale behind this re-scaling process is that the rubrics for sets other than 7 and 8 are holistic, utilizing only a general trait, whereby a single rater's score determines the Domain Score. In contrast, Essay Sets 7 and 8 employ composite rubrics with each trait being scored separately. Therefore, the score ranges were recalibrated based on the internal scaling of each trait.

Across all three essay sets, for Essay Set 1, each rater's score was scaled between 1 and 6, and the combined Domain_Score from rater 1 and rater 2 was transformed from the original range of 2-12 to a range of 1-6. In Essay Set 7, each rater provided scores between 0 and 3 across four distinct traits: "Ideas," "Organization," "Style," and "Conventions." Although each rater could award up to 12 points, it was observed that there were insufficient examples across the entire range of 0-12 classes. For instance, only one student received scores of 2 and 3, whereas 199 students received a score of 16. Thus, similar to the individual traits within "Ideas", "Organization", "Style", and "Conventions" the scores for Essay Set 7 were standardized to a range of 0-3. In Essay Set 8, each student's essay was evaluated by two or three independent

raters (rater 1 and rater 2) across six different traits: "Ideas" (I), "Organization" (O), "Voice" (V), "Word Choice" (W), "Sentence Fluency" (S), and "Conventions" (C). The scoring range for these traits varied from 0 to 6. Accordingly, the domain scores were also scaled within a range of 0 to 6 points, consistent with the traits. The following scores have been converted to an ordinal scale (See Table 5).

Table 5

Score Levels by Domain Ranges and Ordinal Scales

	Essay Set					
	Set 1		**Set 7**		**Set 8**	
Score Level	Domain Score Range	Ordinal Scale	Domain Score Range	Ordinal Scale	Domain Score Range	Ordinal Scale
Class 1	0-2	1	0-5	0	0-9	1
Class 2	3-4	2	6-11	1	10-19	2
Class 3	5-6	3	12-17	2	20-29	3
Class 4	7-8	4	18-24	3	30-39	4
Class 5	9-10	5	-	-	40-49	5
Class 6	11-12	6	-	-	50-60	6

Model Development

In this part of the study, I present a concise overview of the GPT-2 architecture (Vaswani et al., 2017). Following that, I provide a comprehensive explanation of the model developed by augmenting the GPT-2 architecture with a classifier layer. This classifier has been tailored for a specific classification or prediction task, leveraging the text generation capabilities of GPT-2. Additionally, a comprehensive explanation of the hyperparameter configurations employed in training this model is elaborated upon in this section. All analyses were conducted on Google

Colab ProCloud servers, known for their substantial resource capacities, including 32 GB of RAM and Nvidia Tesla V100-SXM2 GPUs.

GPT-2 Architecture

GPT-2 is a large language model created by OpenAI and released in February 2019 for the single purpose of predicting the next word(s) in a sentence. It is a transformer-based language model pre-trained on an extensive dataset containing 8 million web pages. It was based on the transformer architecture, which is a neural network architecture designed for natural language processing tasks such as language modeling and machine translation.

A functional representation of the Transformer model's encoder and decoder layers and the components of each layer is shown in Figure 2. As shown in Figure 2, both encoder- and decoder-style architectures use the same self-attention layers to encode word tokens. However, the main difference is that encoders are designed to learn embeddings that can be used for various predictive modeling tasks such as classification. In contrast, decoders are designed to generate new texts, for example, answering user queries. For instance, the BERT model comprises solely encoder blocks, processing a text bidirectionally from both the right and left directions, thereby enhancing its ability to grasp contextual information comprehensively from both directions. In contrast, the T5 model incorporates both Encoder and Decoder blocks, enabling it to transform textual input into textual output. Conversely, GPT models consist exclusively of Decoder blocks, processing text unidirectionally from left to right, conventionally employed in tasks such as text generation and text classification.

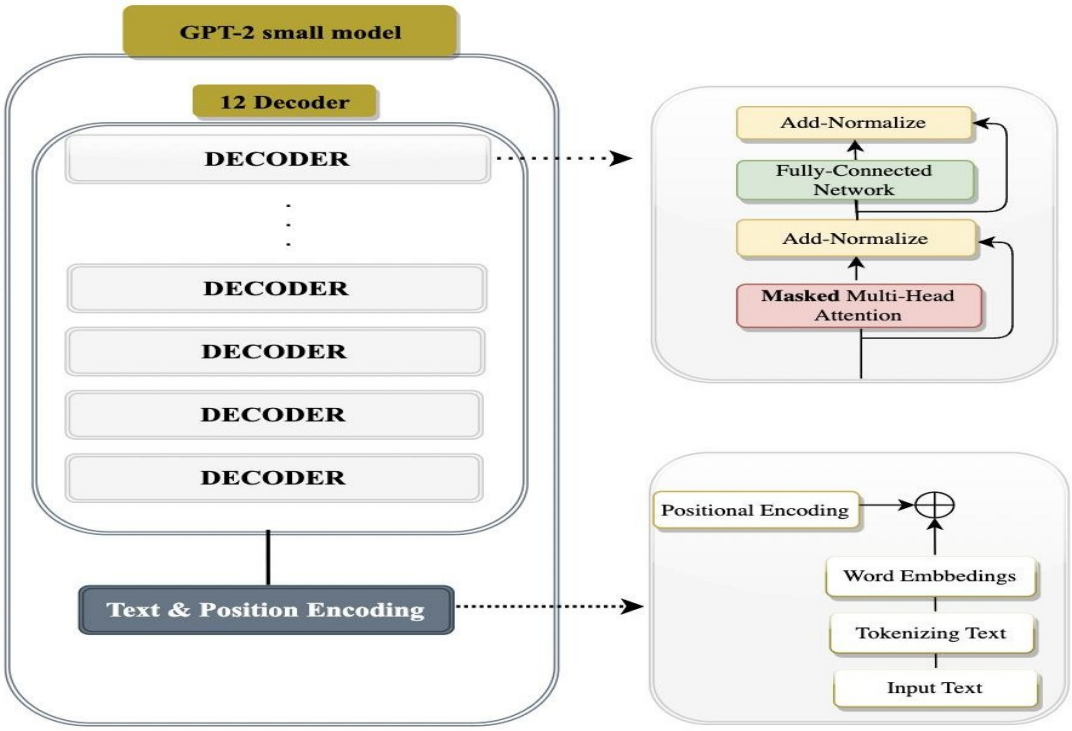
GPT-2 is an enhanced version of GPT with approximately 1.5 billion parameters (Radford et al., 2019). It is capable of performing tasks such as reading comprehension, machine translation, question answering, text generation, and summarization. Ali et al. (2022)

demonstrated the adaptability of GPT-2 for text classification tasks. In this study, we will discuss a model designed by adding a classification layer on top of the GPT-2 model architecture. The GPT-2 model has four different configurations based on the number of parameters it contains (See Table 2). The Small GPT-2 model employs a more straightforward embedding and positional encoding structure when processing input data compared to larger models. For the Small GPT-2 model, each word in the text is represented with 768-dimensional word embedding vectors, and 1024-dimensional positional encoding vectors are utilized to determine the position of each word.

GPT-2's smallest model employs a more straightforward embedding and positional encoding structure when processing input data compared to larger models. For the Small GPT-2, each word in the text is represented with 768-dimensional word embedding vectors, and 1024-dimensional positional encoding vectors are utilized to determine the position of each word. GPT-2 is a Transformer-based model consisting of decoder blocks. In this study, the smallest version of GPT-2 (as illustrated in Figure 10), with 12 decoder blocks, was selected primarily for its computational efficiency. This choice facilitated rapid training and model iteration within the constraints of Google Colab ProCloud's resources. Additionally, the model's reduced size ensures quicker iterative processes and cost-effectiveness, allowing for a more efficient research process while still adequately addressing the complexity of the task.

Figure 10

GPT-2 Small Architecture (Radford et al., 2018)



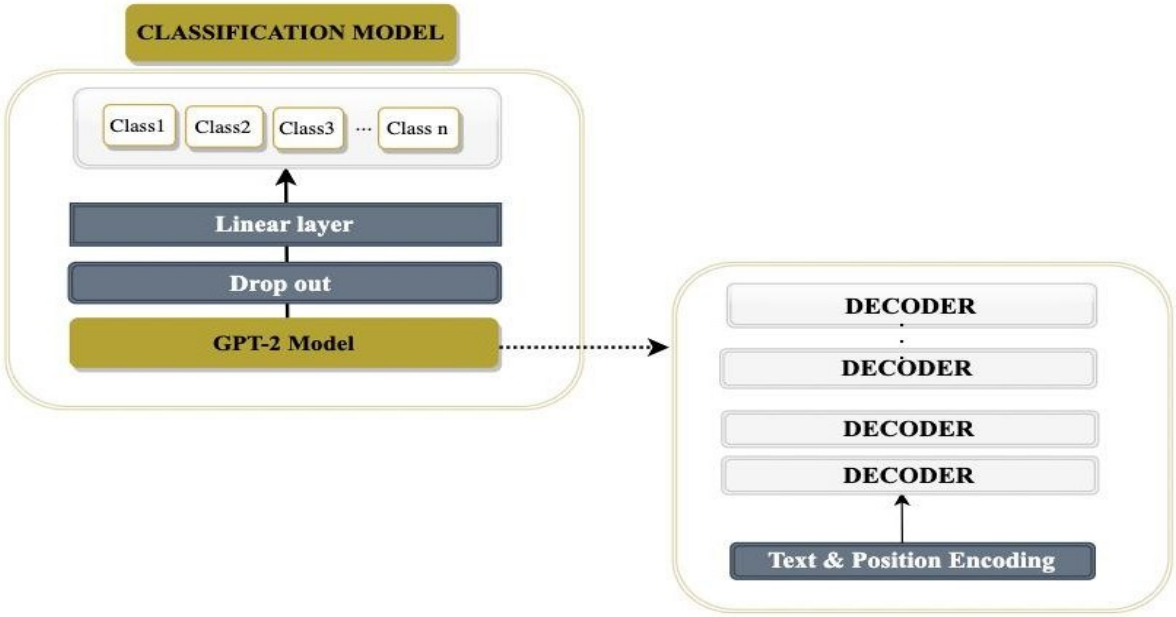
As illustrated in Figure 10, GPT-2 Small consists of 12 decoder blocks. Each block comprises a masked multi-head self-attention layer and Fully-Connected network, consisting of two linear transformations with a ReLU activation in between (Vaswani et al., 2017), with a layer normalization (LayerNorm) layer (Ba et al., 2016) and a residual connection (He et al., 2016). In contrast to the encoder architecture, the decoder employs masked multi-head attention to prevent rightward information flow.

Classification Model Architecture

I expanded the model architecture by incorporating a classification layer subsequent to the transformer decoders, for a more comprehensive understanding of the classification model architecture.

Figure 11

Classification Model Architecture



To construct the GPT-2 classification model, we extended the GPT-2 model by adding layers within a Python class. This Python class, named Classifier Layer, was designed for the specific purpose of implementing a classifier layer within the GPT-2 model. It comprises essential components, including dropout and linear layer (see Figure 11).

Dropout Layer. The Dropout Layer is a critical component that plays a pivotal role in addressing overfitting and enhancing the generalization capability of the model (Srivastava et al., 2014). Specifically, dropout rates of 0.1, 0.2, and 0.5 were employed as hyperparameters to observe their impact on the model's performance.

Linear Layer. This layer was utilized to transform GPT-2 outputs into classification scores. A fully connected layer refers to a neural network in which each neuron applies a linear transformation to the input vector through a weight matrix, calculating scores for each class. This layer takes the 768-dimensional feature vectors generated by the GPT-2 model and produces a

specific number of outputs based on the number of classes in the essay sets. The number of output classes was adjusted based on the number of classes in the essay sets, denoted by the 'num_labels' hyperparameter. For instance, 'num_labels' was set to 4 in Essay Set 4, which had 4 classes with scores of 0, 1, 2, and 3.

Experimental Setup and Hyperparameter Tuning

The textual inputs underwent tokenization utilizing the GPT-2 tokenizer from the Huggingface library. Subsequently, the resultant tokens were supplied as input to the model. Each individual essay within the essay set was subjected to tokenization, and essays that surpassed the maximum sequence length accommodated by the GPT-2 (base) model, set at 1,024 tokens, underwent truncation.

Following that step, I defined a configuration file containing the hyperparameters of the pretrained GPT-2 model. These parameters include a vocabulary size of 50, 257, a hidden layer count of 12, embedding and hidden state dimensions of 768, 12 attention heads for each attention layer, and the use of the GELU activation function instead of ReLU. The dropout probability was set to 0.1 for all fully connected layers, as well as for embedding and attention layers. The epsilon value for layer normalization layers was set to 1e-05. It's worth noting that all of these hyperparameter settings adhered to the default values of GPT-2, which has a total of 117 million hyperparameters.

I employed the same configuration file when fine-tuning a classification model for the GPT-2 model. Token embedding outputs obtained from GPT-2 were fed as input into the classifier layer. To create a classification model tailored to the number of classes in each essay set, I extended the GPT-2 model with a classification layer. The number of classes was

determined by the "num_label" parameter in the linear layer. For instance, if there were 5 classes, then "num_label" was set to 5. A dropout rate of 0.1 was applied.

The cross-entropy loss function, which is used to compute the discrepancy between the model's predictions and the actual class labels, quantifies the distributional difference between two probabilistic distributions. Categorical cross-entropy, which takes into account the number of classes denoted by k, the actual output y, and the predicted output $\hat{y}$, is defined by Equation 4:

$$E(y,\hat{y}) = \sum_{i=1}^{k} y_i . log\ (\hat{y}) \quad (4)$$

I defined an optimizer using AdamW, which applies the Adam algorithm with weight decay (Kingma & Ba, 2014). Weight decay is employed to keep the weights as small as possible, preventing them from diverging and thus mitigating issues related to excessive model fitting. This helps prevent gradient explosions in the network. The calculated loss function, in conjunction with the AdamW optimizer, is used to update the parameters within the model architecture.

In order to accurately assess the overall performance and generalization capability of the model, it is essential to partition the dataset into appropriately balanced training, validation, and test sets. In this study, the dataset has been divided into three subsets as follows: 60% for training, 20% for validation, and 20% for testing. This partitioning was carried out using the widely adopted "train-test split" method, as commonly employed in the literature (Bishop, 2006), which effectively aids in measuring the model's performance. Further details concerning the hyperparameters and architectural aspects of the GPT-2 classification model can be seen in Table 6.

Table 6

GPT-2 Classification Model Hyperparameters

Layer	Parameter Name	Parameter Value
Embedding	Embedding Dimension	768
	Positional Encoding	1024
GPT-2	Decoder	12, 24, 26, 48
Linear Layer	Num_labels	3,4,5,6
Dropout	Dropout Rate	0.1, 0.2, 0.5
Model Compile	Epoch	5,10,15,20
	Learning Rate	1e-2,1e-3, 1e-4, 1e-5
	Batch Size	2,4,8,16,32
	Dropout rate	0.1, 0.3, 0.5

The Effect of Data Augmentation on GPT-2 Model Performance

Data augmentation is the process of artificially modifying data in a training set to increase the size and diversity of the dataset. This method is used in text classification to address imbalanced class distributions in the dataset.

As emphasized by Zhang et al. (2020), when the class distribution of a dataset is not uniformly balanced, the dataset is termed "imbalanced." Imbalance implies that certain classes contain fewer examples compared to others. The data augmentation method assists in mitigating this imbalance in distribution by augmenting the quantity of data pertaining to a specific class Zhu et al. (2018).

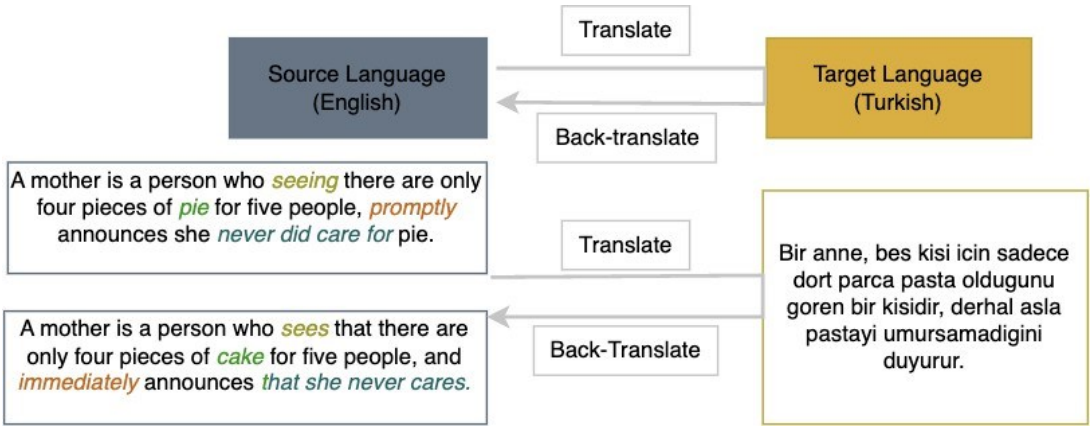
In this research, I utilized the "back-translation" method as one of the data augmentation techniques to enhance the performance of our model. Back-translation involves translating text from the source language into the target language and subsequently translating it back into the source language, effectively serving as an artificial data augmentation approach (Brislin, 1970).

In this study, the source language was English, and the target language was set to Turkish. This choice was based on the strategic decision to take into account the agglutinative structure of Turkish and its fundamental deconstructive differences from English. This feature of Turkish increases grammatical variety and depth during translation because in Turkish meaning is expressed through affixes, whereas in English, it is mostly expressed through word order and inflections. This fundamental difference between these two languages provides grammatical and semantic enrichment in translations so that natural language processing models are trained with more diverse and complex language structures, which expands the model's ability to understand and process the nuances of the language. For these two reasons, I chose Turkish as the intermediary language, as this diversity will improve the paraphrase and language processing capabilities of the model, contributing to more effective natural language processing performance.

My research primarily focused on examples from score classes that exhibited imbalanced distributions and had limited sample sizes. To address this, we translated essays from English to Turkish and then back from Turkish to English, thereby creating artificial essays. The text translation process was carried out using the Google Translate API, with Turkish specified as the target language. Furthermore, the googletrans library enables translation into 107 additional target languages. Figure 12 below illustrates the translation mechanism of the "Back-Translation" method.

Figure 12

The Translation Mechanism of the Back-Translation Method



I categorized nine different essay sets into two groups: those with balanced distributions and those with imbalanced distributions. Essay Set 4 exhibited a comparatively balanced distribution compared to the other sets. Therefore, as in the dissertation by Firoozi (2023), a 20% data augmentation was applied to each score group in Essay Set 4. For the other sets, since the score classes showed imbalanced distributions, the number of instances for classes with fewer than 50 examples was doubled. This strategy was employed in order to enhance model performance in the datasets.

Performance Metrics

In this study, model performance was evaluated with two main measures: Cohen's Kappa and Quadratic Weighted Kappa scores were used to assess overall consistency. Alongside evaluating consistency, the overall accuracy performance of the model was measured. These scores were utilized to gauge the consistency and accuracy of the model in comparison to human judgment within the AES system. Additionally, metrics such as precision, sensitivity, and F1-Score were employed to evaluate the model's performance at each score level.

Cohen's Kappa is a statistical metric used to measure the consistency of classifications made by an evaluation system or model with human assessments. This metric assesses how much the observed agreement (P_o) differs from the expected agreement (P_e) based on chance. Cohen's Kappa is calculated as the observed agreement divided by the maximum chance-adjusted agreement rate (see Equation 5). The formula is as follows:

$$Cohen's\ Kappa\ (\kappa) = \frac{Po-Pe}{1-Pe}. \quad (5)$$

Cohen's Kappa serves as an indicator of the extent to which the AES system reflects agreement in classifications. Landis and Koch (1977) proposed the following ranges of values for interpreting Cohen's Kappa values:

$\kappa<0$ represents less compliance than chance,

$0.01\leq\kappa\leq0.20$ indicates mild agreement,

$0.21\leq\kappa\leq0.40$ represents fair agreement,

$0.41\leq\kappa\leq0.60$ indicates moderate agreement,

$0.61\leq\kappa\leq0.80$ represents significant agreement, and

$0.81\leq\kappa\leq0.99$ indicates almost perfect agreement.

Therefore, traditionally, a Cohen's Kappa value greater than 0.80 is considered to indicate perfect agreement, while a value greater than 0.60 is considered to indicate good agreement. These values provide a scale to interpret the level of agreement between the AES system and human evaluation based on Cohen's Kappa measurements.

Quadratic Weighted Kappa (QWK) is a derivative of Cohen's Kappa and is a metric that includes weighted calculations shown in Equation 6 to account for misclassifications based on how close the scores are to the correct score level:

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2} \quad (6)$$

Subsequently, a Oi, matrix is established to represent the number of essays receiving an observed human rating (i) and machine rating (j). Following this, the observed count matrix (O) and the expected count matrix (E) are computed. Quadratic Weighted Kappa (QWK) serves as a measure of agreement between the observed and expected matrices and is calculated using Equation 7

$$QWK = \frac{\sum_{i,j} (W_{i,j}.O_{i,j})}{\sum_{i,j} (W_{i,j}.E_{i,j})}. \quad (7)$$

Quadratic Weighted Kappa (QWK) is a measure ranging from random agreement (0) to perfect agreement (1) between raters. As highlighted by Williamson et al. (2012), QWK values between 0.60 and 0.80 indicate an acceptable lower bound of reliability, particularly in high-stakes testing situations involving human raters.

Both Cohen's Kappa and QWK are commonly used together as measures of consistency in AES, with QWK being the most widely used performance evaluation measure. These metrics offer insights into the agreement between the AES system and human evaluation, considering chance agreement and the proximity of scores. The most commonly used performance measurement for AES is QWK (Williamson et al., 2012).

The primary distinction between Cohen's Kappa and Quadratic Weighted Kappa (QWK) lies in the weights assigned to score categories in classification problems. Cohen's Kappa assigns equal weight to all class categories, assuming that each class has the same importance. A confusion matrix is utilized to evaluate the performance of a model at different scoring levels (refer to Table 7). This matrix determines the model's sensitivity, precision, and recall rates while also calculating the F-score index.

Table 7

Confusion Matrix Definition

		Predicted Scores
Actual Scores	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

As illustrated in Equation 8, precision is calculated by dividing the True Positives by the total positive samples. A high sensitivity suggests a minimal occurrence of false positives, although it may not account for instances of false negatives

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

Recall is the sum of a model's true positive predictions divided by the sum of true positives and false negatives (see Equation 9). This metric is employed to evaluate how effectively the model recognizes positive examples and assesses its tendency to miss such instances.

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

The F1-Score aims to mitigate errors by achieving an equilibrium between precision and sensitivity, as shown in Equation 10 (Visa et al., 2011). Its purpose is to strike a judicious balance between minimizing both false positives and false negatives.

$$F_1\ Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (10)$$

Chapter Summary

In this chapter, I described the methodology for AES, detailing the dataset, data preprocessing, model development, and evaluation metrics. I described the use of the ASAP dataset and the preprocessing steps, including text and score-based methods. The chapter introduces the development of a GPT-2-based model, enhanced with a classification layer for AES and hyperparameter tuning. I also described the effect of data augmentation techniques like back-translation. Performance metrics, including Cohen's Kappa and Quadratic Weighted Kappa, were employed to assess the model's accuracy and consistency compared to human raters. I also presented the metrics of accuracy, precision, sensitivity, and the F-1 score providing further insight into the classification accuracy.

Chapter 4: Results

This chapter provides a detailed analysis of the classifier results utilizing the 'small' version of the GPT-2 model. I begin by describing the hyperparameter settings where my model exhibits optimal performance. Next, I use these parameter settings to evaluate the model's performance using the Quadratic Weighted Kappa (QWK) score as the performance measure. Following this, I provide a comprehensive examination of the impact of data augmentation techniques on the model's performance.

Hyperparameter Settings

In this section, I conducted experiments involving various regularization parameters to enhance the performance of the model. Table 8 provides a detailed explanation of the model and the hyperparameters that yield the best performance**.**

The classification layer added to the GPT-2 Small model was employed to categorize the model's outputs into distinct classes that encompassed the dropout and linear layer components. Critical hyperparameters, such as the learning rate, number of epochs, and batch size, were fine-tuned throughout the model's training process.

Table 8

Final Selection of Hyperparameters of the Best GPT-2 Models

Layer	Parameter Name	Essay Set								
		1	**2a**	**2b**	**3**	**4**	**5**	**6**	**7**	**8**
Embedding	Dimension	768	768	768	768	768	768	768	768	768
	Positional Encoding	1024	1024	1024	1024	1024	1024	1024	1024	1024
GPT-2 Model	Decoders	12	12	12	12	12	12	12	12	12
Linear Layer	The number of Layers	6	6	4	4	4	5	5	4	6
Dropout Layer	Dropout Rate	0.5	0.5	0.1	0.1	0.2	0.1	0.3	0.1	0.1
Model Compile	Learning Rate	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4
	Epoch	20	25	35	20	20	30	20	30	30
	Batch Size	2	2	2	4	2	2	2	1	2

In the GPT-2 Small model, tokenized text is represented by embeddings of size 768, each of which carries the meaning of individual tokens (Embedding Dimension). Additionally, these embeddings are augmented with positional encodings of size 1024. This process enhances text processing efficiency by capturing each word's content and positional information. These vectors are subsequently processed within the 12 decoder layers of the GPT-2 Small model (Decoder).

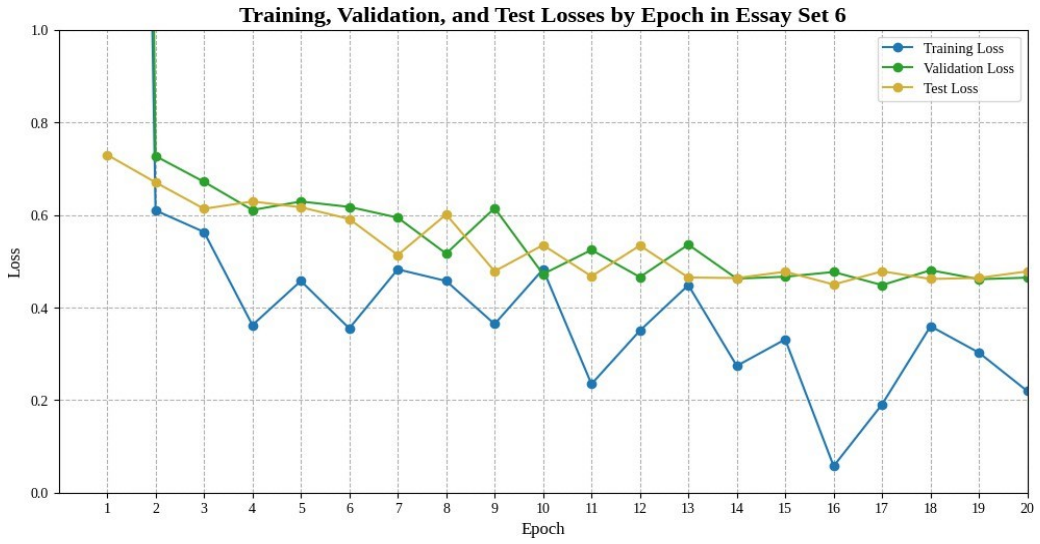
The output from the 12th decoder layer was further subjected to processing through the dropout and linear layer, which played an integral role in the classification task. The 'num_label' parameter within the linear layer dictated the number of categorized classes the model produces

as output. The number of classes varied depending on the essay set class count (linear layer). To mitigate overfitting, distinct dropout rates were applied for each essay set. Specifically, for Essay Sets 1 and 2a, a dropout rate of 0.5 demonstrated the best performance, while for Essay Sets 2b, 3, 5, 7, and 8, the dropout rate of 0.1 produced the best performance.

The learning rate was utilized as a hyperparameter in the ADAMW optimization algorithm. According to the results of the hyperparameter optimization, a learning rate of 1e-4 showed the most effective convergence in the loss function across all essay sets. Epoch and batch size are two interrelated hyperparameters. An epoch signifies the number of times the model processes the entire training dataset from beginning to end, constituting a single iteration. Different essay sets exhibited their best performance at varying epoch values. For instance, in the case of Essay Set 6, the most favourable outcome was obtained with a dropout rate of 0.3, a learning rate of 1e-4, and an epoch number of 20. As the epoch count increased, the variations in training and test loss are illustrated in Figure 13.

Figure 13

Evolution of Loss Values by Epoch in Essay Set 6



During the training process, various batch sizes were investigated. In Essay Set 3, a batch size of 4 was used, while in Essay Set 7, it was set to 1. For the remaining essays, the most favourable results were achieved when utilizing a batch size of 2 data samples in each iteration. The hyperparameter known as "batch size" specifies the number of data samples processed in each iteration. For instance, if the batch size is configured as 2 for a dataset containing a total of 100 data points, then the model processes two data samples in every iteration. Consequently, a full cycle, which is referred to as one epoch, is completed over the entire dataset, totalling 50 iterations. For instance, Table 9 illustrates the impact of changes in batch size on the training time and model performance for Essay Set 4 while keeping the number of epochs at 20 and the dropout rate at 0.2 constant.

Table 9

Effect of Batch Size on QWK and Training Time

Dropout	Leaning Rate	Epoch	Batch Size	QWK	Training Time
0.2	1e-4	20	2	0.74	16:38 min
0.2	1e-4	20	4	0.68	16:24 min
0.2	1e-4	20	8	0.69	15:10 min

Following the findings presented in Table 9, it is evident that when holding the dropout rate and learning rate constant, the configuration employing a batch size of 2 yielded the highest Quadratic Weighted Kappa (QWK) score. Moreover, an increase in batch size is associated with a reduction in training time. This outcome occurs because, with a larger batch size, the model processes more data in each iteration, leading to a shorter training duration.

Result of the AES Model

Williamson, Xi, and Breyer (2012) described two pivotal criteria for assessing Quadratic Weighted Kappa (QWK) scores. The first criterion posits that a QWK score of 0.70 or higher indicates a robust level of agreement, explaining at least half of the variability inherent in human scores. The second criterion states that the absolute discrepancy between human-human and human-machine agreement should not exceed 0.10. Table 10 presents a comprehensive overview of the results derived from QWK scores, and these findings undergo a meticulous assessment concerning their adherence to the aforementioned criteria.

Table 10

Model Performance Comparison Using QWK Score

Model	Essay Set									
	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	0.71	0.45	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	0.03	0.18	0.06

Considering the first criterion, the QWK score of the GPT-2-based model meets or exceeds the threshold of 0.70 in six of the nine essay sets. This result indicates that the model is in compliance with this specific criterion. The model achieved a QWK score of 0.70 only in sets 2a, 2b, and 8, whereas it attained its highest QWK score of 0.77 in Essay Set 5.

When considering the second criterion, the difference between the human raters and the model was less than 10% in six out of nine test sets, indicating that the model meets this criterion. However, for Essay Set 8, the difference between the human raters and the model was quite large and well above the criterion at 0.18.

Consequently, the GPT-2 small-based model fulfilled these two criteria in six of the nine essay sets. Essay Set 2a and Essay Set 8 produced the weakest results with QWK producing scores of 64% and 45%, respectively, while the absolute differences between the model and the human raters in these essay sets were 14% and 18%, respectively. As a result, these two criteria were not satisfied for these specific essay sets.

According to the scale proposed by Landis and Koch (1977), Cohen's Kappa values of the GPT-2 model reflect the level of agreement between the model's AES system and human raters. According to this scale, Cohen's Kappa values between 0.21 and 0.40 represent "fair" agreement, Cohen's Kappa values between 0.41 and 0.60 represent "moderate" agreement, values between 0.61 and 0.80 represent "substantial" agreement and values between 0.81 and 1.00 represent "almost perfect" agreement.

As illustrated in Table 11, the model's performance was compared with that of human raters using metrics such as Quadratic Weighted Kappa (QWK), Cohen's Kappa, and Accuracy. The model's highest Cohen's Kappa score of 0.52 (Essay Set 1) indicates a "moderate" fit, while 0.31 (Essay Set 8) falls into the "fair" fit category. The model's average Cohen's Kappa value for all sets is 0.43. These results suggest that the model's average agreement with human raters represent "moderate" agreement. QWK values provide a measure of agreement corrected for the severity of misclassifications and, therefore offer a more nuanced assessment than Cohen's Kappa. The GPT-2 model achieved the highest QWK value of 0.77 for Essay Set 5, indicating that the model best takes into account the severity of ratings in this set. On the other hand, the QWK value of 0.45 for Essay Set 8 indicates that the model's performance on this set is significantly lower, suggesting that the weight of the classification errors contributes to this poor fit. The model's average QWK value for all sets is 0.68.

Accuracy was used as another performance metric for predicting essay scores across essay sets using the GPT-2 model. The highest accuracy value belongs to Essay Set 1 and 5 with 0.67, while the lowest accuracy value belongs to Essay Set 8 with 0.52. The model's average Accuracy value for all sets is 0.61.

Table 11

Model Comparison using QWK, Cohen's Kappa, and Accuracy

Model	Performance Metrics	Essay Set 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	0.67	0.58	0.62	0.52	**0.61**

The GPT-2 results given in Table 12 are compared with GPT-3.5 and GPT-4 using two of the same datasets, and the results obtained are presented in Table 12 (Gunduz & Gierl, 2024). In the study by Gunduz and Gierl (2024), both GPT-4 and GPT-3.5 have shown significant performance improvement in a single-prompt scenario. This finding underscores the critical importance of correctly tuning learning scenarios in maximizing model performance.

While the highest QWK value achieved by GPT-4 for Essay Set 4 was 0.75, this value was 0.71 for Essay Set 5. Considering the performance of the same essay sets with the fine-tuned GPT-2 model, the fine-tuned GPT-2 model reached QWK values of 0.74 for Essay Set 4 and 0.77 for Essay Set 5. These findings indicate that fine-tuning could enhance the performance of GPT-2, making it competitive with GPT-4, and in some cases, it can even surpass GPT-4.

Table 12

Model Performance Comparison in the study (Gunduz &Gierl, 2024)

Models	Performance Metrics	Essay Set 4			Essay Set 5		
GPT-2	QWK		**0.74**			**0.77**	
	Cohen's Kappa		0.43			0.45	
		Zero-shot	**One-shot**	**Few-shot**	**Zero-shot**	**One-shot**	**Few-shot**
GPT-3.5	QWK	0.12	0.55	**0.73**	0.35	0.64	0.68
	Cohen's Kappa	0.05	0.29	0.44	0.12	0.31	0.32
GPT-4	QWK	0.56	0.75	**0.75**	0.42	**0.71**	0.68
	Cohen's Kappa	0.31	0.49	0.48	0.20	0.35	0.33

Precision measures the proportion of items correctly classified by the model out of the total predicted items. In Essay Set 2b, the model's precision score for Class 3 was the highest (0.78), indicating that the majority of the model's predictions for this class were correct. However, Class 1 had a relatively low precision score (0.33), indicating that many of the model's predictions for this class are incorrect. This could mean that the model shows weaknesses in recognizing essays with a score corresponding to Class 1, and it can be inferred that it needs to be improved to reduce false positives.

Sensitivity measures how many items that actually belongs to that class are correctly classified by the model. In Essay Set 3, the sensitivity score for Class 4 was high (0.79), indicating that the model correctly captures this proportion of true instances of this class. However, the sensitivity for Class 1 in the same set is zero (0.00), indicating that the model failed to classify any correct instances of this class, which is considered a serious deficiency.

The F1 score is the harmonic mean of precision and sensitivity. A high F1 score indicates that the model is both accurate and responsive. For Essay Set 4, Class 4 is characterized by a

high F1 score (0.69), indicating that the model both correctly predicts the items in this class and accurately captures the actual situations. On the other hand, the F1 score for Class 1 is low (0.47), indicating that the model exhibits both low precision and low sensitivity in this class.

Table 13

Average Accuracy Scores of the GPT-2 Model at the Essay Sets with 4 Classes

	Essay Set								
	2b			3			4		
Performance Level	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Class 1	0.33	0.50	0.40	0.00	0.00	0.00	0.65	0.45	0.53
Class 2	0.71	0.60	0.65	0.67	0.68	0.67	0.57	0.53	0.55
Class 3	0.78	0.32	0.45	0.56	0.79	0.66	0.59	0.83	0.69
Class 4	0.59	0.94	0.72	0.73	0.25	0.37	0.70	0.36	0.47
Weighted Average	0.68	0.63	0.59	0.63	0.61	0.58	0.61	0.60	0.59

The data presented in Table 14 details the classification performance of the GPT-2 model on Essay Sets 2a, 5 and 6 with 5-6 classes and includes precision, recall and F1 scores for each class.

Table 14

Accuracy Scores of the GPT-2 Model at the Essay Sets with 6 classes

	Essay Set								
	2a			5			6		
Performance Level	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Class 1	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.20	0.33
Class 2	0.65	0.52	0.58	0.68	0.63	0.65	0.43	0.30	0.35
Class 3	0.70	0.41	0.52	0.66	0.83	0.74	0.54	0.56	0.55
Class 4	0.57	0.91	0.70	0.64	0.60	0.62	0.72	0.59	0.65
Class 5	0.00	0.00	0.00	0.74	0.32	0.50	0.44	0.79	0.56
Class 6	0.00	0.00	0.00	-	-	-	-	-	-
Weighted Average	0.61	0.61	0.58	0.66	0.66	0.65	0.62	0.58	0.58

In terms of precision, the score of 1.00 for Class 1 in Essay Set 6 indicates that the model perfectly correctly classified all the essays belonging to this category, with no false positive predictions. As for the sensitivity metric, the score of 0.91 for Class 4 in Essay Set 2a indicates that the model correctly detected the vast majority of real essays belonging to this class. The F-1 score is an indication of how evenly the model classifies the essays. The F1 score of 0.74 for Class 3 in Essay Set 5 indicates that the model has a good balance between precision and accuracy and performs strongly in this class. When it comes to Table 15 at Essay Sets 1,7, and 8, in terms of precision, the score of 1.00 for Class 2 in Essay Set 8 indicates that the model perfectly correctly classified all the essays belonging to this category, with no false positive predictions.

Table 15

Accuracy Scores of the GPT-2 Model at Essay Sets 1, 7, and 8

	Essay Set								
	1			**7**			**8**		
Performance Level	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Class 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Class 2	0.00	0.00	0.00	0.62	0.59	0.60	1.00	0.08	0.15
Class 3	0.40	0.33	0.36	0.64	0.66	0.65	0.59	0.86	0.70
Class 4	0.72	0.93	0.81	0.72	0.71	0.72	0.60	0.43	0.50
Class 5	0.36	0.18	0.24	-	-	-	0.00	0.00	0.00
Class 6	0.00	0.00	0.00	-	-	-	0.00	0.00	0.00
Weighted Average	0.57	0.67	0.60	0.66	0.67	0.66	0.60	0.59	0.55

In conclusion, it is clear that the GPT-2 model performs remarkably well in certain classes but requires significant improvements in other classes. These findings suggest that the model needs to be improved, especially in terms of reducing false positives and increasing the accuracy of class recognition.

Impact of Data Augmentation on Model Performance

Impact of Augmentation on the Essay Sets

Each set of essays contains different score ranges. While Essay Set 4 has a well-distributed range of scores, the other essay sets have an uneven distribution of samples across scores. This imbalance can affect the model's performance in certain score ranges, which is why data augmentation was employed.

However, the data augmentation process primarily targeted score groups with fewer than 50 samples. For these small sample size groups, a "back-translation" technique was used. Back-translation is one of the text augmentation techniques that involves the process of translating text from the source language to the target language and then back to the source language (Edunov et al., 2018). In the back-translation process, Turkish was chosen as the intermediate language. Once the language choice was determined, using this technique, the essays were first translated into Turkish and then translated back into English, with Turkish being chosen as the intermediary language. It is important to note that this data augmentation did not involve any changes to the original essay scores. The scores of the essays remained unchanged.

In Table 16, the number of samples included in each score class for Essay Sets 2 through 6 is presented. Classes with fewer than 50 examples in each Essay set have been doubled to increase the sample size. However, as the four distinct score classes in Essay Set 4 contained fewer than 50 examples each, a uniform augmentation method of 20% was applied for all score classes. This method of increasing each score class by 20% was also applied in the study (Firoozi & Gierl, 2024). The classification of Domain Score, encompassing up to 6 classes, presents a comparison of sample sizes before and after data augmentation, as demonstrated in Table 16.

Table 16

Number of Essays in Score Classes with 4, 5, and 6 Before and After Back-Translation

Model	Essay Set					
	2a	**2b**	**3**	**4**	**5**	**6**
Domain Score Range	1-6	1-4	0-3	0-3	0-4	0-4
Rater's Score Range	1-6	1-4	0-3	0-3	0-4	0-4
Class 1	24 48	29 58	39 78	311	24 48	44 88
Class 2	153	191	607	636	302	167
Class 3	763	730	657	570	649	405
Class 4	778	850	423	253	572	817
Class 5	75	-	-	-	258	367
Class 6	714	-	-	-	-	-
Discrepancy	+24	+29	+39	*20% +352	+24	+44

In Essay Sets 1, 7, and 8, there are score classes that exceed 6 classes. Essay Set 1 includes a total of 11 different score classes ranging from 2 to 12. Essay Set 7 comprises 23 different scores ranging from 2 to 24 (see Figure 14). Essay Set 8 consists of 34 distinct score classes ranging from 10 to 60. Across all three essay sets, the back translation method was utilized to augment the number of essay samples in score classes that had fewer than 50 examples.

In Essay Set 1, students scoring (2, 3, 4, 5, 12) had the following counts of essays: (10, 1, 17, 17, 47), respectively. Since each of these score classes had fewer than 50 essays, the number of essays was doubled, resulting in an additional +92 artificially generated essays.

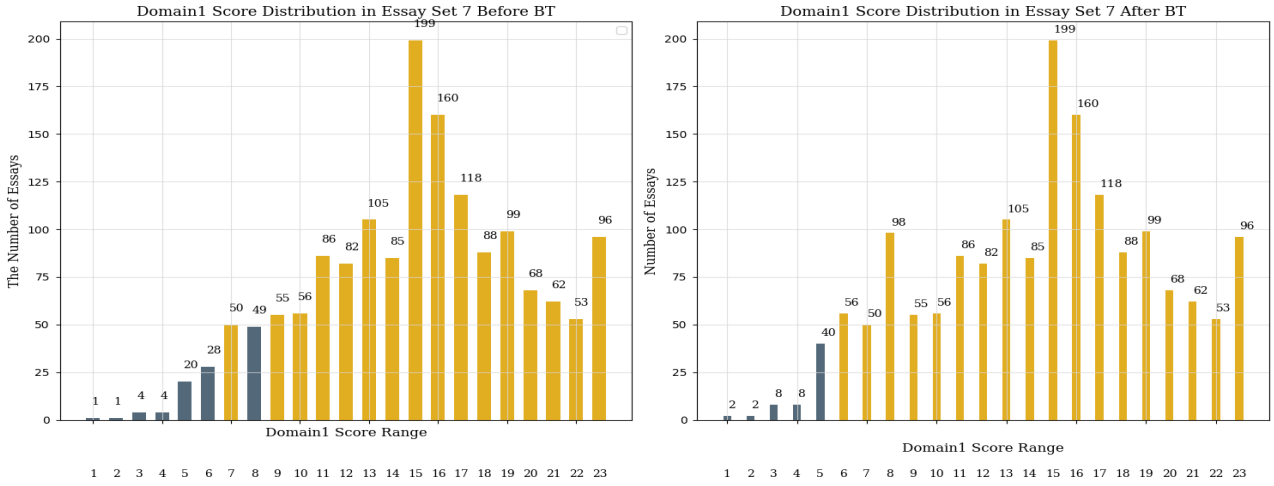
For Essay Set 7, score classes (2, 3, 4, 5, 6, 7) had fewer than 50 samples, thus essays in these classes were doubled, contributing a total of +107 artificial data points.

In Essay Set 8, despite having 34 different score classes, most of these classes had fewer than 50 examples. Therefore, data augmentation was applied to all score classes with fewer than 50 examples. Only score classes 18 and 22, which had more than 50 examples (+65 and +161 respectively), did not undergo back translation, as indicated by the blue markers in the graph. The remaining score classes had their sample sizes doubled, resulting in a cumulative increase of +497 examples.

Consequently, the updated dataset for Essay Set 1 and 7 expanded by +92 and +80 samples respectively, while the new dataset for Essay Set 8 increased by 497 samples. The revised score distribution for Essay sets following the "back-translation" process is illustrated in Figure 14. Scores with fewer than 50 essays have been shown in blue.

Figure 14

Number of Essays in Score Classes more than six *Before and After Back-Translation*



Before Back-Translation, a scaled score range was used to measure the performance of the GPT-2 model on the dataset. After Back-Translation, data augmentation was first applied to the initial data using the back-translation method to evaluate the model's performance, and then the data was converted to an ordinal scale. The score ranges for the classes are shown in Table 17.

Table 17

Score Classes in Set 1,7, and 8 Before and After Back-Translation

	Essay Set					
	1		**7**		**8**	
Domain Score Range	2-12		0-24		0-60	
Rater's Score Range	1-6		0-12		0-30	
Scaled Score Range	1-6		0-3		1-6	
Class 1	10	20	10	20	0	0
Class 2	18	36	258	328	2	4
Class 3	127	144	717		44	88
Class 4	822		584		370	675
Class 5	650		-		291	421
Class 6	156	203	-		16	32
Discrepancy	+92		+80		+497	

The data augmentation process has been thoroughly detailed in Table 18, showing the increase in data for each essay set. It's crucial to emphasize that the extent of data augmentation varies for each distinct essay set. Essay Set 8 exhibited the most substantial increase in the data size compared to the other sets, while Essay Set 5 also contains a significant increase in sample size. Essay Sets 2a and 2b, however, demonstrated a nearly identical level of data increase. More specifically, Essay Set 8 underwent the most significant data augmentation, with the addition of 497 samples, making it the most prominent increase in the difference between QWK scores. Following that, Essay Set 4 experienced an increase of 352 samples, ranking second. In contrast, Essay Set 2b achieved the least data augmentation, with only a 29-sample increase compared to the others.

Table 18

Comparison of Dataset Size Before and After Back-Translation for Each Essay Set

Model	Essay Set 1	2a	2b	3	4	5	6	7	8
Training Set Size Before BT	1783	1800	1800	1726	1772	1805	1800	1569	723
Training Set Size After BT	1875	1831	1829	1765	2124	1829	1844	1676	1220
Discrepancy	+92	+31	+29	+39	+352	+24	+44	+80	+497

Impact of Augmentation on Model Performance

Table 19 provides clear evidence that the data augmentation method has substantially enhanced the model's performance based on three different kinds of performance metrics. Notably, significant performance improvements are discernible in Essay Sets 3, 4, and 8, where a substantial amount of data augmentation has been applied. For instance, in Essay Set 4, the Quadratic Weighted Kappa (QWK) increased by 0.07 as a result of the augmentation process. In Essay Set 8, the QWK score exhibited an even more impressive increase of 0.15 due to augmentation, with the highest improvement observed among all sets. In Essay Set 2a, the QWK score saw a modest increase of 0.01, indicating a slight enhancement in the model's performance for this set. The improvements in Essay Sets 1, 2a, 2b, 3, 5, 6, and 7 are also evident but are characterized by smaller differences. When looking at the average increase values, the back-translation method has improved the performance of the GPT-2 model, showing an increase in the QWK score of +0.06.

Table 19

Comparison of Model Performance Before and After Back-Translation

Model	Performance Metrics	Essay Set 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	0.48	0.41	0.49	0.42	**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	0.69	0.68	0.64	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	+0.07	+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**

When the precision, recall, and F1 scores for the essay sets after Back-Translation (BT) with the GPT-2 model presented in Tables 20, 21, and 22 were analyzed, some classes and performance metrics have shown remarkable improvements.

The most remarkable precision score was 0.86 for Class 1 in Essay Set 6 + BT. This indicates that the model's predictions for this class were highly accurate, as the model correctly classified the elements of this class. In terms of sensitivity, the score of 0.90 for Class 4 in Essay Set 2b+BT shows that the model has a very high ability to successfully detect instances of this class. In terms of F1 scores, a score of 0.83 for Class 4 in Essay Set 2b+BT is noteworthy, indicating a very balanced and strong performance of the model for this class in terms of both precision and sensitivity.

These results show that the data augmentation process, and in particular the back-translation method, can significantly improve the classification performance of the model.

Table 20

Average Accuracy of the GPT-2 Model with Back-Translation at the Essay Sets with 4 classes

	Essay Set								
	2b+BT			3+BT			4+BT		
Performance Level	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Class 1	1.00	0.80	0.89	0.80	0.50	0.62	0.73	0.63	0.68
Class 2	0.56	0.60	0.58	0.58	0.86	0.69	0.59	0.70	0.64
Class 3	0.71	0.55	0.62	0.55	0.54	0.55	0.64	0.66	0.65
Class 4	0.78	0.90	0.83	0.79	0.46	0.58	0.34	0.43	0.45
Weighted Average	0.74	0.74	0.74	0.63	0.61	0.60	0.67	0.62	0.63

Table 21

Average Accuracy of the GPT-2 Model with Back-Translation at Essay Sets with 5-6 classes

	Essay Set								
	2a+BT			5+BT			6+BT		
Performance Level	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Class 1	0.00	0.00	0.00	0.80	0.67	0.73	0.86	0.77	0.81
Class 2	0.63	0.53	0.58	0.57	0.75	0.65	0.36	0.40	0.38
Class 3	0.59	0.70	0.64	0.72	0.66	0.69	0.54	0.57	0.55
Class 4	0.64	0.66	0.65	0.57	0.63	0.60	0.69	0.72	0.71
Class 5	0.00	0.00	0.00	0.66	0.45	0.53	0.67	0.57	0.62
Class 6	0.00	0.00	0.00	-	-	-	-	-	-
Weighted Average	0.58	0.62	0.59	0.65	0.64	0.64	0.64	0.64	0.64

Table 22

Average Accuracy of the GPT-2 Model with Back-Translation at Essay Sets 1, 7, and 8

	Essay Set								
	1			7			8		
Performance Level	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Class 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Class 2	1.00	0.10	0.18	0.62	0.59	0.60	0.00	0.00	0.00
Class 3	0.50	0.81	0.62	0.64	0.66	0.64	0.75	0.38	0.50
Class 4	0.77	0.74	0.75	0.72	0.71	0.72	0.80	0.74	0.77
Class 5	0.69	0.58	0.63	-	-	-	0.59	0.79	0.68
Class 6	0.46	0.69	0.55	-	-	-	1.00	0.14	0.25
Weighted Average	0.68	0.66	0.65	0.66	0.67	0.66	0.73	0.71	0.71

Chapter 5: Discussion

In this chapter, firstly, the objectives of the study and the research questions are presented. Then, the main findings are summarized in detail. Finally, the limitations of the study are evaluated and suggested directions for future research are discussed.

Purpose of the Study

The purpose of my study is to provide a comprehensive evaluation of the performance of GPT-2, one of the transformer models used in AES systems. More specifically, I investigated how the meanings obtained by GPT-2 through tokenization are processed with the added classification layer and how this process affects the scoring of essays in the ASAP dataset. Previous studies have shown that existing AES systems fail to detect a large proportion of errors in essays, especially errors involving long-distance relationships (Ng et al., 2014). In this study, I evaluate the potential advantages and limitations of using the GPT-2 model in AES, which incorporates an attention mechanism that can effectively capture and annotate such long-distance errors. In this study, I first analyze the model's performance in AES, and second, I detail the impact of back-translation-based data augmentation techniques on model performance.

Main Findings of the Study

The current study is designed to address the following questions:

(1) How accurately and reliably does the GPT-2 model score essays on the ASAP dataset?

The GPT-2 model performed within an acceptable range of reliability in the experiments on the ASAP dataset, except for the essay sets with small sample sizes. Furthermore, the fine-tuned GPT-2 small model has shown the potential to exceed the performance of the non-fine-tuned GPT-3.5 and GPT-4 models, as observed in the study by Gunduz and Gierl (2024). This

underscores the significance of fine-tuning as a crucial factor in enhancing the model performance.

The GPT-2 tokenized essays were trained on the model with the classification layer added to the GPT-2 architecture. The performance of the model was evaluated using QWK, Cohen's Kappa, accuracy, precision, recall and F1 scores. These evaluation measures determine the overall accuracy of score classification as well as the accuracy of each class being detected and classified.

When comparing the QWK performance scores of the sets, I used the criteria recommended by Williamson, Xi, and Breyer (2012) for a more comprehensive analysis. Based on the QWK results for the GPT-2 model, it achieved an average QWK score of 0.68 across all essay sets. This is slightly below the average QWK score of 0.74 achieved by human raters, indicating a low performance of 0.06. Notably, the GPT-2 model outperformed human raters in Essay Sets 1, 5, and 7. The QWK values between 0.60 and 0.80 indicate the lower limit of acceptable reliability, as stated by Williamson et al. (2012), and all other sets except Essay Set 8 are in this range.

When I compared the Cohen's Kappa values of each essay set with the QWK values, the Cohen's Kappa values were quite low in all essay sets. QWK is a weighted scoring method used to overcome some of the shortcomings of kappa, and due to the uneven distribution across classes, Cohen's Kappa values are usually lower than QWK. According to the Landis and Koch (1977) scale, Cohen's Kappa values between 0.41 and 0.60 represent a "moderate" level of agreement.

Score predictions using the GPT-2 model on various essay sets were also evaluated for accuracy. Essay Sets 1 and 5 had the highest accuracy rates (67% and 66%, respectively). For

example, approximately 66% of the essays in Essay Set 5 were scored correctly. Essay Set 5 is divided into different performance categories, and in the largest class, class 3 (with 649 essays), the GPT-2 model performs best with 83% accuracy, while it performs significantly worse in classes with fewer essays. This outcome suggests that the model learns better in classes with more examples and needs to be improved in classes with fewer examples. In short, sample size clearly affected the classification performance of the GPT-2 model.

(2) How do the data augmentation techniques applied with back-translation affect the performance of GPT-2 and what are the characteristics of these effects?

My findings show that the prediction accuracy and coverage of the GPT-2 model improve significantly after a 20% data augmentation, especially for score categories with less than 50 essays. Back-translation as a method of data augmentation has improved the model's performance, increasing the average QWK score across all essay sets by 0.06 (from an average QWK of 0.68 before back-translation to 0.74 after back-translation). The data augmentation with back-translation resulted in a 33% improvement in the QWK score from 0.45 to 0.60, especially in Essay Set 8. These are the most important findings of our study.

My analysis shows that a 20% data increase in score categories with less than 50 essays improves the model's coverage and prediction accuracy. Comparing before and after data augmentation, the average QWK score increased by +0.07 points and the average Cohen's Kappa score increased by +0.04 points. This outcome suggests that the variety and depth of language structure provided by back-translation helps the model to classify more accurately.

More specifically, I found that in Essay Set 8, the QWK score increased by +0.15 and the Cohen's Kappa score increased by +0.11 as a result of the data augmentation with back-translation, with the highest performance improvement observed in this set at 33% for score

category 1. This reflects the overall improvement in QWK, Cohen's Kappa, and accuracy values observed across all essay sets. These results show that data augmentation through back-translation significantly improves the overall performance of the model and increases its reliability and accuracy, especially in the score categories with very small samples.

As a result of the data augmentation in Essay Set 5, the number of compositions in Class 1, which contains few examples, was doubled, and significant improvements in model performance were observed in this class. Significant increases in the accuracy, recall and F1 scores of Class 1 were observed before and after back-translation, indicating significant improvements in the model's ability to recognize and correctly classify this class. The increase in precision and recall indicates that the model has improved both in correctly recognizing compositions belonging to Class 1 (fewer false positives) and in not missing those that belong to Class 1 (fewer false negatives).

To summarize, the results from my research demonstrate the critical role that data augmentation techniques, and in particular back-translation, play in the training and classification success of language models and their potential to improve model performance.

The GPT-2 model achieved high QWK scores on all essay sets except the 8th set, with an average performance of 0.68, very close to the human raters' average QWK score of 0.73. These results show that despite the simple structure of the model, its accuracy is comparable to previous research. In particular, Mizumoto and Eguchi's (2023) study showed that the performance of models such as GPT-3 can improve when language measures are integrated (the QWK increased from 0.39 in their base model to 0.61 when language measures were added). Similarly, in this study, the GPT-2 model performed impressively without linguistic measures, and I would expect that this performance can be further improved by integrating language

measures. Fang et al. (2023) showed how data augmentation using GPT-4 improves model performance. In our study, we also observe that data augmentation using back translation yields positive results for our GPT-2 model. This finding shows the potential of data augmentation to improve model performance.

In conclusion, my study contributes to the research literature by providing an in-depth investigation of the performance of AES systems in recognizing and evaluating long-distance relations in text using GPT-2. Using GPT-2's attention mechanisms, I demonstrated that the shortcomings identified in previous studies can be overcome and accurate ratings can be achieved within acceptable reliability limits, even on low sample size datasets. The fine-tuning of GPT-2 has shown to outperform GPT-3.5 and GPT-4, thereby substantiating the significance of fine-tuning for the enhancement of a model's performance (Gunduz &Gierl, 2024). Data augmentation techniques, in particular back-translation, were shown to significantly improve the model's classification performance, allowing it to score more accurately and reliably with unbalanced distributions between classes. This allowed the model to perform well even in underrepresented classes. In addition, my study contributes significantly to advancing the field by providing methods for correcting language interpretation capacity and data imbalances for developing AES tools and their effective use in education.

Study Limitations and Directions for Future Research

This study presents a comprehensive analysis of the use of GPT-2 in AES systems, highlighting the potential of GPT-2, a transformer model with a decoder structure, in AES. The research has shown that language models can perform impressively in this domain, even under certain limitations. Nevertheless, the applicability and extensibility of the results to a wider audience should be evaluated, taking into account some limitations of the study.

The first limitation is the small sample size of some of the essay sets used and the uneven distribution of score categories within each set. The use of this data structure may limit the model's capacity to learn scores with small samples and make it difficult to generalize the results to a wider population. This raises some questions about whether the model can perform consistently across different classes. Therefore, it is recommended that future studies examine more diverse, large-sample, and balanced datasets to assess the generalizability and performance of the model in a more balanced manner.

The second limitation is that only a small model of GPT-2 was used in this study. This may have limited the model's capacity to understand intertextual relationships and more complex language structures. More advanced and larger models like GPT-3 or GPT-4 could analyze these complex elements in greater detail due to their increased parameter count. However, an evaluation by Gunduz and Gierl (2024) has revealed that when used without fine-tuning, the ChatGPT 3.5 and ChatGPT 4.0 models performed worse on essay scoring than the fine-tuned GPT-2 model. This finding highlights the significant impact of the fine-tuning process on model performance. Future research should focus on the fine-tuned versions of larger and more advanced models, exploring how these models handle complex linguistic structures and intertextual relations.

As a final limitation, this study focused specifically on one data augmentation technique, back-translation. While this technique may improve the performance of the model, especially in underrepresented categories, the effects of other potential data augmentation methods (e.g. synonym replacement, sentence rearrangement) were not examined. Future work should examine in detail the potential effects of these techniques on model performance and make important

contributions to the development of AES tools and the advancement of language modelling technology in education.

References

Ali, S., Nasir, A., Samad, A., Basser, S., & Irshad, A. (2022). An automated approach for the prediction of the severity level of bug reports using GPT-2. *Security and Communication Networks*, *2022*.

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. *arXiv preprint arXiv:1606.04289*.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of cross-cultural psychology*, *1*(3), 185-216.

Chernyavskiy, A., Ilvovsky, D., & Nakov, P. (2021). Transformers:"the end of history" for natural language processing?. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21* (pp. 677-693). Springer International Publishing.

Chollet, F. (2018). Keras: The python deep learning library. *Astrophysics source code library*, ascl-1806.

Cooper, C. R. (1977). Holistic evaluation of writing. *Evaluating writing: Describing, measuring, judging*, 3-31.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ebel R.L., Frisbie D.A. (1986). *Essentials of education measurement*. Englewood Cliffs, N.J: Prentice Hall.

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Fang, L., Lee, G. G., & Zhai, X. (2023). Using gpt-4 to augment unbalanced data for automatic scoring. *arXiv preprint arXiv:2310.18365*.

Firoozi, T., & Gierl, M. J. (2024). WRITTEN IN PERSIAN USING A TRANSFORMER-BASED MODEL. *The Routledge International Handbook of Automated Essay Evaluation*, 55.

Firoozi, T., Mohammadi, H., & Gierl, M. J. (2023). Using Active Learning Methods to Strategically Select Essays for Automated Scoring. *Educational Measurement: Issues and Practice*, *42*(1), 34-43.

Firoozi, T., Bulut, O., Epp, C. D., Naeimabadi, A., & Barbosa, D. (2022). The effect of fine-tuned word embedding techniques on the accuracy of automated essay scoring systems using neural networks. *Journal of Applied Testing Technology*, 21-29.

Gaddipati, S. K., Nair, D., & Plöger, P. G. (2020). Comparative evaluation of pretrained transfer learning models on automatic short answer grading. *arXiv preprint arXiv:2009.01303.*

Giacaglia, G. (2023, May 11). *Transformers*. Medium. https://towardsdatascience.com/transformers-141e32e69591

Gierl, M. J., Latifi, S., Lai, H., Boulais, A. P., & Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical education*, 48(10), 950-962.

Goulden, N. R. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *Journal of Research & Development in Education*.

Gunduz, A., & Gierl, M. (2023, April). Automated essay scoring with ChatGPT 3.5 and 4.0. In *UAlberta Graduate Student Research in Education Conference 2024*. University of Alberta. https://doi.org/10.13140/RG.2.2.28282.09924

Hartford, J. S., Wright, J. R., & Leyton-Brown, K. (2016). Deep learning for predicting human strategic behavior. *Advances in neural information processing systems*, *29*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Henkel, O., Hills, L., Roberts, B., & McGrane, J. (2023). Can LLMs Grade Short-answer Reading Comprehension Questions: Foundational Literacy Assessment in LMICs. *arXiv preprint arXiv:2310.18373*.

Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.

Irissappane, A. A., Yu, H., Shen, Y., Agrawal, A., & Stanton, G. (2020). Leveraging GPT-2 for classifying spam reviews with limited labeled data via adversarial training. *arXiv preprint arXiv:2012.13400.*

Kingma, D.P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Klebanov, B. B., & Madnani, N. (2022). *Automated Essay Scoring*. Springer Nature.

Kumar, V., & Boulanger, D. (2020, October). Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in education* (Vol. 5, p. 572367). Frontiers Media SA.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Liu, Q., Kusner, M. J., & Blunsom, P. (2020). A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.

Liu, Q., Kusner, M. J., & Blunsom, P. (2020). A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050.

Nugaliyadde, A., Somaratne, U., & Wong, K. W. (2019). Predicting electricity consumption using deep recurrent neural networks. *arXiv preprint arXiv:1909.08182*.

Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W. T. (2018). Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*.

Plakans, L., & Gebril, A. (2015). *Assessment myths: Applying second language research to classroom teaching*. University of Michigan Press.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495-2527.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Salton, G., Yang, C. S., & Yu, C. T. (1975). A theory of term importance in automatic text analysis. *Journal of the American society for Information Science*, *26*(1), 33-44.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53-76.

Shin, J., & Gierl, M. J. (2021). More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing*, *38*(2), 247-272.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.

Şahin, N. (2021). *Malware detection using transformers-based model GPT-2* (Master's thesis, Middle East Technical University).

Tay, Y., Luu, A. T., & Hui, S. C. (2018). Recurrently controlled recurrent networks. *Advances in neural information processing systems*, *31*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, *31*(1), 2-13.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xu, B., Gil-Jardiné, C., Thiessard, F., Tellier, E., Avalos, M. F., & Lagarde, E. (2019). Neural language model for automated classification of electronic medical records at the emergency room. the significant benefit of unsupervised generative pre-training.

Yan, D., Rupp, A. A., & Foltz, P. W. (Eds.). (2020). *Handbook of automated scoring: Theory into practice*. CRC Press.

Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023, July). 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)* (pp. 576-584).

Yenduri, G., Srivastava, G., Maddikunta, P. K. R., Jhaveri, R. H., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2023). Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *arXiv preprint arXiv:2305.10435*.

Zhang, W., Li, X., Jia, X. D., Ma, H., Luo, Z., & Li, X. (2020). Machinery fault diagnosis with imbalanced data using deep generative adversarial networks. *Measurement*, *152*, 107377.

Zhu, X., Liu, Y., Li, J., Wan, T., & Qin, Z. (2018). Emotion classification with data augmentation using generative adversarial networks. In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22* (pp. 349-360). Springer International Publishing.

Appendix: Essay Sets

A1: Essay Set 1

Prompt

More and more people use computers, but not everyone agrees that this benefits society. Those who support advances in technology believe that computers have a positive effect on people. They teach hand-eye coordination, give people the ability to learn about faraway places and people, and even allow people to talk online with other people. Others have different ideas. Some experts are concerned that people are spending too much time on their computers and less time exercising, enjoying nature, and interacting with family and friends.

Write a letter to your local newspaper in which you state your opinion on the effects computers have on people. Persuade the readers to agree with you.

A2: Essay Set 2

Prompt

Censorship in the Libraries

"All of us can think of a book that we hope none of our children or any other children have taken off the shelf. But if I have the right to remove that book from the shelf -- that work I abhor -- then you also have exactly the same right and so does everyone else. And then we have no books left on the shelf for any of us." --Katherine Paterson, Author

Write a persuasive essay to a newspaper reflecting your vies on censorship in libraries. Do you believe that certain materials, such as books, music, movies, magazines, etc., should be removed from the shelves if they are found offensive?

Support your position with convincing arguments from your own experience, observations, and/or reading.

A3: Essay Set 3

Source

ROUGH ROAD AHEAD: Do Not Exceed Posted Speed Limit

by Joe Kurmaskie

FORGET THAT OLD SAYING ABOUT NEVER taking candy from strangers. No, a better piece of advice for the solo cyclist would be, "Never accept travel advice from a collection of old-timers who haven't left the confines of their porches since Carter was in office." It's not that a group of old guys doesn't know the terrain. With age comes wisdom and all that, but the world is a fluid place. Things change.

At a reservoir campground outside of Lodi, California, I enjoyed the serenity of an early summer evening and some lively conversation with these old codgers. What I shouldn't have done was let them have a peek at my map. Like a foolish youth, the next morning I followed their advice and launched out at first light along a "shortcut" that was to slice away hours from my ride to Yosemite National Park.

They'd sounded so sure of themselves when pointing out landmarks and spouting off towns I would come to along this breezy jaunt. Things began well enough. I rode into the morning with strong legs and a smile on my face. About forty miles into the pedal, I arrived at the first "town." This place might have been a thriving little spot at one time—say, before the last world war—but on that morning it fit the traditional definition of a ghost town. I chuckled, checked my water supply, and moved on. The sun was beginning to beat down, but I barely noticed it. The cool pines and rushing rivers of Yosemite had my name written all over them. Twenty miles up the road, I came to a fork of sorts. One ramshackle shed, several rusty pumps, and a corral that couldn't hold in the lamest mule greeted me. This sight was troubling. I had been hitting my water bottles pretty regularly, and I was traveling through the high deserts of California in June.

I got down on my hands and knees, working the handle of the rusted water pump with all my strength. A tarlike substance oozed out, followed by brackish water feeling somewhere in the neighborhood of two hundred degrees. I pumped that handle for several minutes, but the water wouldn't cool down. It didn't matter. When I tried a drop or two, it had the flavor of battery acid.

I got back on the bike, but not before I gathered up a few pebbles and stuck them in my mouth. I'd read once that sucking on stones helps take your mind off thirst by allowing what spit you have left to circulate. With any luck I'd hit a bump and lodge one in my throat. It didn't really matter. I was going to die and the birds would pick me clean, leaving only some expensive outdoor gear and a diary with the last entry in praise of old men, their wisdom, and their keen sense of direction. I made a mental note to change that paragraph if it looked like I was going to lose consciousness for the last time.

Somehow, I climbed away from the abandoned factory of juices and dreams, slowly gaining elevation while losing hope. Then, as easily as rounding a bend, my troubles, thirst, and fear were all behind me.

GARY AND WILBER'S FISH CAMP—IF YOU WANT BAIT FOR THE BIG ONES, WE'RE YOUR BEST BET!

"And the only bet," I remember thinking.

As I stumbled into a rather modern bathroom and drank deeply from the sink, I had an overwhelming urge to seek out Gary and Wilber, kiss them, and buy some bait—any bait, even though I didn't own a rod or reel.

An old guy sitting in a chair under some shade nodded in my direction. Cool water dripped from my head as I slumped against the wall beside him.

"Where you headed in such a hurry?"

"Yosemite," I whispered.

"Know the best way to get there?"

I watched him from the corner of my eye for a long moment. He was even older than the group I'd listened to in Lodi.

"Yes, sir! I own a very good map."

And I promised myself right then that I'd always stick to it in the future.

Prompt

Write a response that explains how the features of the setting affect the cyclist. In your response, include examples from the essay that support your conclusion.

The old guys had sworn the next town was only eighteen miles down the road. I could make that! I would conserve my water and go inward for an hour or so—a test of my inner spirit. Not two miles into this next section of the ride, I noticed the terrain changing. Flat road was

replaced by short, rolling hills. After I had crested the first few of these, a large highway sign jumped out at me. It read: ROUGH ROAD AHEAD: DO NOT EXCEED POSTED SPEED LIMIT.

The speed limit was 55 mph. I was doing a water-depleting 12 mph. Sometimes life can feel so cruel.

I toiled on. At some point, tumbleweeds crossed my path and a ridiculously large snake—it really did look like a diamondback—blocked the majority of the pavement in front of me. I eased past, trying to keep my balance in my dehydrated state.

The water bottles contained only a few tantalizing sips. Wide rings of dried sweat circled my shirt, and the growing realization that I could drop from heatstroke on a gorgeous day in June simply because I listened to some gentlemen who hadn't been off their porch in decades, caused me to laugh.

It was a sad, hopeless laugh, mind you, but at least I still had the energy to feel sorry for myself. There was no one in sight, not a building, car, or structure of any kind. I began breaking the ride down into distances I could see on the horizon, telling myself that if I could make it that far, I'd be fi ne.

Over one long, crippling hill, a building came into view. I wiped the sweat from my eyes to make sure it wasn't a mirage, and tried not to get too excited. With what I believed was my last burst of energy, I maneuvered down the hill.

In an ironic twist that should please all sadists reading this, the building—abandoned years earlier, by the looks of it—had been a Welch's Grape Juice factory and bottling plant. A sandblasted picture of a young boy pouring a refreshing glass of juice into his mouth could still be seen.

I hung my head.

That smoky blues tune "Summertime" rattled around in the dry honeycombs of my deteriorating brain.

A4: Essay Set 4

Source

Winter Hibiscus by Minfong Ho

Saeng, a teenage girl, and her family have moved to the United States from Vietnam. As Saeng walks home after failing her driver's test, she sees a familiar plant. Later, she goes to a florist shop to see if the plant can be purchased.

It was like walking into another world. A hot, moist world exploding with greenery. Huge flat leaves, delicate wisps of tendrils, ferns and fronds and vines of all shades and shapes grew in seemingly random profusion.

"Over there, in the corner, the hibiscus. Is that what you mean?" The florist pointed at a leafy potted plant by the corner.

There, in a shaft of the wan afternoon sunlight, was a single blood-red blossom, its five petals splayed back to reveal a long stamen tipped with yellow pollen. Saeng felt a shock of recognition so intense, it was almost visceral.1

"Saebba," Saeng whispered.

A saebba hedge, tall and lush, had surrounded their garden, its lush green leaves dotted with vermilion flowers. And sometimes after a monsoon rain, a blossom or two would have blown into the well, so that when she drew the well water, she would find a red blossom floating in the bucket.

Slowly, Saeng walked down the narrow aisle toward the hibiscus. Orchids, lanna bushes, oleanders, elephant ear begonias, and bougainvillea vines surrounded her. Plants that she had not even realized she had known but had forgotten drew her back into her childhood world. When she got to the hibiscus, she reached out and touched a petal gently. It felt smooth and cool, with a hint of velvet toward the center—just as she had known it would feel. And beside it was yet another old friend, a small shrub with waxy leaves and dainty flowers with purplish petals and white centers. "Madagascar periwinkle," its tag announced. How strange to see it in a pot, Saeng thought. Back home it just grew wild, jutting out from the cracks in brick walls or between tiled roofs.

And that rich, sweet scent—that was familiar, too. Saeng scanned the greenery around her and found a tall, gangly plant with exquisite little white blossoms on it. "Dok Malik," she said,

savoring the feel of the word on her tongue, even as she silently noted the English name on its tag, “jasmine.”

One of the blossoms had fallen off, and carefully Saeng picked it up and smelled it. She closed her eyes and breathed in, deeply. The familiar fragrance filled her lungs, and Saeng could almost feel the light strands of her grandmother’s long gray hair, freshly washed, as she combed it out with the fine-toothed buffalo-horn comb. And when the sun had dried it, Saeng would help the gnarled old fingers knot the hair into a bun, then slip a dok Malik bud into it. Saeng looked at the white bud in her hand now, small and fragile. Gently, she closed her palm around it and held it tight. That, at least, she could hold on to. But where was the fine-toothed comb? The hibiscus hedge? The well? Her gentle grandmother?

A wave of loss so deep and strong that it stung Saeng’s eyes now swept over her. A blink, a channel switch, a boat ride into the night, and it was all gone. Irretrievably, irrevocably gone. And in the warm moist shelter of the greenhouse, Saeng broke down and wept. It was already dusk when Saeng reached home. The wind was blowing harder, tearing off the last remnants of green in the chicory weeds that were growing out of the cracks in the sidewalk. As if oblivious to the cold, her mother was still out in the vegetable garden, digging up the last of the onions with a rusty trowel. She did not see Saeng until the girl had quietly knelt down next to her.

Her smile of welcome warmed Saeng. “Ghup ma laio le? You’re back?” she said cheerfully. “Goodness, it’s past five. What took you so long? How did it go? Did you—?” Then she noticed the potted plant that Saeng was holding, its leaves quivering in the wind. Mrs. Panouvong uttered a small cry of surprise and delight. “Dok faeng-noi!” she said. “Where did you get it?”

“I bought it,” Saeng answered, dreading her mother’s next question.

“How much?”

For answer Saeng handed her mother some coins.

“That’s all?” Mrs. Panouvong said, appalled, “Oh, but I forgot! You and the Lambert boy ate Bee-Maags.....”

“No, we didn’t, Mother,” Saeng said.

"Then what else—?"

"Nothing else. I paid over nineteen dollars for it."

"You what?" Her mother stared at her incredulously. "But how could you? All the seeds for this vegetable garden didn't cost that much! You know how much we—" She paused, as she noticed the tearstains on her daughter's cheeks and her puffy eyes.

"What happened?" she asked, more gently.

"I—I failed the test," Saeng said.

For a long moment Mrs. Panouvong said nothing. Saeng did not dare look her mother in the eye. Instead, she stared at the hibiscus plant and nervously tore off a leaf, shredding it to bits. Her mother reached out and brushed the fragments of green off Saeng's hands. "It's a beautiful plant, this dok faeng-noi," she finally said. "I'm glad you got it."

"It's—it's not a real one," Saeng mumbled.

"I mean, not like the kind we had at—at—" She found that she was still too shaky to say the words at home, lest she burst into tears again. "Not like the kind we had before," she said. "I know," her mother said quietly. "I've seen this kind blooming along the lake. Its flowers aren't as pretty, but it's strong enough to make it through the cold months here, this winter hibiscus. That's what matters."

She tipped the pot and deftly eased the ball of soil out, balancing the rest of the plant in her other hand. "Look how root-bound it is, poor thing," she said. "Let's plant it, right now." She went over to the corner of the vegetable patch and started to dig a hole in the ground. The soil was cold and hard, and she had trouble thrusting the shovel into it. Wisps of her gray hair trailed out in the breeze, and her slight frown deepened the wrinkles around her eyes. There was a frail, wiry beauty to her that touched Saeng deeply.

"Here, let me help, Mother," she offered, getting up and taking the shovel away from her. Mrs. Panouvong made no resistance. "I'll bring in the hot peppers and bitter melons, then, and start dinner. How would you like an omelet with slices of the bitter melon?"

"I'd love it," Saeng said.

Left alone in the garden, Saeng dug out a hole and carefully lowered the "winter hibiscus" into it. She could hear the sounds of cooking from the kitchen now, the beating of eggs against a bowl, the sizzle of hot oil in the pan. The pungent smell of bitter melon wafted out, and Saeng's mouth watered. It was a cultivated taste, she had discovered—none of her classmates or friends, not even Mrs. Lambert, liked it—this sharp, bitter melon that left a golden aftertaste on the tongue. But she had grown up eating it and, she admitted to herself, much preferred it to a Big Mac.

The "winter hibiscus" was in the ground now, and Saeng tamped down the soil around it. Overhead, a flock of Canada geese flew by, their faint honks clear and—yes—familiar to Saeng now. Almost reluctantly, she realized that many of the things that she had thought of as strange before had become, through the quiet repetition of season upon season, almost familiar to her now. Like the geese. She lifted her head and watched as their distinctive V was etched against the evening sky, slowly fading into the distance.

When they come back, Saeng vowed silently to herself, in the spring, when the snows melt and the geese return and this hibiscus is budding, then I will take that test again.

Prompt

Read the last paragraph of the story.

"When they come back, Saeng vowed silently to herself, in the spring, when the snows melt and the geese return and this hibiscus is budding, then I will take that test again."

Write a response that explains why the author concludes the story with this paragraph. In your response, include details and examples from the story that support your ideas.

A5: Essay Set 5

Source

Narciso Rodriguez

from *Home: The Blueprints of Our Lives*

My parents, originally from Cuba, arrived in the United States in 1956. After living for a year in a furnished one-room apartment, twenty-one-year-old Rawedia Maria and twenty-seven year-old Narciso Rodriguez, Sr., could afford to move into a modest, three-room apartment I would soon call home.

In 1961, I was born into this simple house, situated in a two-family, blond-brick building in the Ironbound section of Newark, New Jersey. Within its walls, my young parents created our traditional Cuban home, the very heart of which was the kitchen. My parents both shared cooking duties and unwittingly passed on to me their rich culinary skills and a love of cooking that is still with me today (and for which I am eternally grateful). Passionate Cuban music (which I adore to this day) filled the air, mixing with the aromas of the kitchen. Here, the innocence of childhood, the congregation of family and friends, and endless celebrations that encompassed both, formed the backdrop to life in our warm home.

Growing up in this environment instilled in me a great sense that "family" had nothing to do with being a blood relative. Quite the contrary, our neighborhood was made up of mostly Spanish, Cuban, and Italian immigrants at a time when overt racism was the norm and segregation prevailed in the United States. In our neighborhood, despite customs elsewhere, all of these cultures came together in great solidarity and friendship. It was a close-knit community of honest, hardworking immigrants who extended a hand to people who, while not necessarily their own kind, were clearly in need.

Our landlord and his daughter, Alegria (my babysitter and first friend), lived above us, and Alegria graced our kitchen table for meals more often than not. Also at the table were Sergio and Edelmira, my surrogate grandparents who lived in the basement apartment. (I would not know my "real" grandparents, Narciso the Elder and Consuelo, until 1970 when they were allowed to leave Cuba.) My aunts Bertha and Juanita and my cousins Arnold, Maria, and Rosemary also all lived nearby and regularly joined us at our table.

Countless extended family members came and went — and there was often someone staying with us temporarily until they were able to get back on their feet. My parents always kept their arms and their door open to the many people we considered family, knowing that they would do the same for us. My mother and father had come to this country with such courage, without any knowledge of the language or the culture. They came selflessly, as many immigrants do, to give their children a better life, even though it meant leaving behind their families, friends, and careers in the country they loved. They struggled both personally and financially, braving the harsh northern winters while yearning for their native tropics and facing cultural hardships. The barriers to work were strong and high, and my parents both had to accept that they might not be able to find the kind of jobs they deserved. In Cuba, Narciso, Sr., had worked in a laboratory and Rawedia Maria had studied chemical engineering. In the United States, they had to start their lives over entirely, taking whatever work they could find. The faith that this struggle would lead them and their children to better times drove them to endure these hard times.

I will always be grateful to my parents for their love and sacrifice. I've often told them that what they did was a much more courageous thing than I could have ever done. I've often told them of my admiration for their strength and perseverance, and I've thanked them repeatedly. But, in reality, there is no way to express my gratitude for the spirit of generosity impressed upon me at such an early age and the demonstration of how important family and friends are. These are two lessons that my parents did not just tell me. They showed me with their lives, and these teachings have been the basis of my life.

It was in this simple house that my parents welcomed other refugees to celebrate their arrival to this country and where I celebrated my first birthdays. It was in the warmth of the kitchen in this humble house where a Cuban feast (albeit a frugal Cuban feast) always filled the air with not just scent and music but life and love. It was here where I learned the real definition of "family." And for this, I will never forget that house or its gracious neighborhood or the many things I learned there about how to love. I will never forget how my parents turned this simple house into a home.

— Narciso Rodriguez, Fashion designer

Hometown: Newark, New Jersey

Prompt

Describe the mood created by the author in the memoir. Support your answer with relevant and specific information from the memoir.

A6: Essay Set 6

Source

The Mooring Mast

by Marcia Amidon Lüsted

When the Empire State Building was conceived, it was planned as the world's tallest building, taller even than the new Chrysler Building that was being constructed at Forty-second Street and Lexington Avenue in New York. At seventy-seven stories, it was the tallest building before the Empire State began construction, and Al Smith was determined to outstrip it in height.

The architect building the Chrysler Building, however, had a trick up his sleeve. He secretly constructed a 185-foot spire inside the building, and then shocked the public and the media by hoisting it up to the top of the Chrysler Building, bringing it to a height of 1,046 feet, 46 feet taller than the originally announced height of the Empire State Building.

Al Smith realized that he was close to losing the title of world's tallest building, and on December 11, 1929, he announced that the Empire State would now reach the height of 1,250 feet. He would add a top or a hat to the building that would be even more distinctive than any other building in the city. John Tauranac describes the plan:

> [The top of the Empire State Building] would be more than ornamental, more than a spire or dome or a pyramid put there to add a desired few feet to the height of the building or to mask something as mundane as a water tank. Their top, they said, would serve a higher calling. The Empire State Building would be equipped for an age of transportation that was then only the dream of aviation pioneers.

This dream of the aviation pioneers was travel by dirigible, or zeppelin, and the Empire State Building was going to have a mooring mast at its top for docking these new airships, which would accommodate passengers on already existing transatlantic routes and new routes that were yet to come.

The Age of Dirigibles

By the 1920s, dirigibles were being hailed as the transportation of the future. Also known today as blimps, dirigibles were actually enormous steel-framed balloons, with envelopes of cotton fabric filled with hydrogen and helium to make them lighter than air. Unlike a balloon, a dirigible could be maneuvered by the use of propellers and rudders, and passengers could ride in the gondola, or enclosed compartment, under the balloon.

Dirigibles had a top speed of eighty miles per hour, and they could cruise at seventy miles per hour for thousands of miles without needing refueling. Some were as long as one thousand feet, the same length as four blocks in New York City. The one obstacle to their expanded use in New York City was the lack of a suitable landing area. Al Smith saw an opportunity for his Empire State Building: A mooring mast added to the top of the building would allow dirigibles to anchor there for several hours for refueling or service, and to let passengers off and on. Dirigibles were docked by means of an electric winch, which hauled in a line from the front of the ship and then tied it to a mast. The body of the dirigible could swing in the breeze, and yet passengers could safely get on and off the dirigible by walking down a gangplank to an open observation platform.

The architects and engineers of the Empire State Building consulted with experts, taking tours of the equipment and mooring operations at the U.S. Naval Air Station in Lakehurst, New Jersey. The navy was the leader in the research and development of dirigibles in the United States. The navy even offered its dirigible, the Los Angeles, to be used in testing the mast. The architects also met with the president of a recently formed airship transport company that planned to offer dirigible service across the Pacific Ocean.

> When asked about the mooring mast, Al Smith commented:
>
> [It's] on the level, all right. No kidding. We're working on the thing now. One set of engineers here in New York is trying to dope out a practical, workable arrangement and the Government people in Washington are figuring on some safe way of mooring airships to this mast.

Designing the Mast

The architects could not simply drop a mooring mast on top of the Empire State Building's flat roof. A thousand-foot dirigible moored at the top of the building, held by a single cable tether, would add stress to the building's frame.

The stress of the dirigible's load and the wind pressure would have to be transmitted all the way to the building's foundation, which was nearly eleven hundred feet below. The steel frame of the Empire State Building would have to be modified and strengthened to accommodate this new situation. Over sixty thousand dollars' worth of modifications had to be made to the building's framework.

Rather than building a utilitarian mast without any ornamentation, the architects designed a shiny glass and chrome-nickel stainless steel tower that would be illuminated from inside, with a stepped-back design that imitated the overall shape of the building itself. The rocket-shaped mast would have four wings at its corners, of shiny aluminum, and would rise to a conical roof that would house the mooring arm. The winches and control machinery for the dirigible mooring would be housed in the base of the shaft itself, which also housed elevators and stairs to bring passengers down to the eighty-sixth floor, where baggage and ticket areas would be located.

The building would now be 102 floors, with a glassed-in observation area on the 101st floor and an open observation platform on the 102nd floor. This observation area was to double as the boarding area for dirigible passengers.

Once the architects had designed the mooring mast and made changes to the existing plans for the building's skeleton, construction proceeded as planned. When the building had been framed to the 85th floor, the roof had to be completed before the framing for the mooring mast could take place. The mast also had a skeleton of steel and was clad in stainless steel with glass windows. Two months after the workers celebrated framing the entire building, they were back to raise an American flag again—this time at the top of the frame for the mooring mast.

The Fate of the Mast

The mooring mast of the Empire State Building was destined to never fulfill its purpose, for reasons that should have been apparent before it was ever constructed. The greatest reason was one of safety: Most dirigibles from outside of the United States used hydrogen rather than helium, and hydrogen is highly flammable. When the German dirigible Hindenburg was destroyed by fire in Lakehurst, New Jersey, on May 6, 1937, the owners of the Empire State Building realized how much worse that accident could have been if it had taken place above a densely populated area such as downtown New York.

The stress of the dirigible's load and the wind pressure would have to be transmitted all the way to the building's foundation, which was nearly eleven hundred feet below. The steel frame of the Empire State Building would have to be modified and strengthened to accommodate this new situation. Over sixty thousand dollars' worth of modifications had to be made to the building's framework.

Rather than building a utilitarian mast without any ornamentation, the architects designed a shiny glass and chrome-nickel stainless steel tower that would be illuminated from inside, with a stepped-back design that imitated the overall shape of the building itself. The rocket-shaped mast would have four wings at its corners, of shiny aluminum, and would rise to a conical roof that would house the mooring arm. The winches and control machinery for the dirigible mooring would be housed in the base of the shaft itself, which also housed elevators and stairs to bring passengers down to the eighty-sixth floor, where baggage and ticket areas would be located.

The building would now be 102 floors, with a glassed-in observation area on the 101st floor and an open observation platform on the 102nd floor. This observation area was to double as the boarding area for dirigible passengers.

Once the architects had designed the mooring mast and made changes to the existing plans for the building's skeleton, construction proceeded as planned. When the building had been framed to the 85th floor, the roof had to be completed before the framing for the mooring mast could take place. The mast also had a skeleton of steel and was clad in stainless steel with glass windows. Two months after the workers celebrated framing the entire building, they were back to raise an American flag again—this time at the top of the frame for the mooring mast.

The Fate of the Mast

The mooring mast of the Empire State Building was destined to never fulfill its purpose, for reasons that should have been apparent before it was ever constructed. The greatest reason was one of safety: Most dirigibles from outside of the United States used hydrogen rather than helium, and hydrogen is highly flammable. When the German dirigible Hindenburg was destroyed by fire in Lakehurst, New Jersey, on May 6, 1937, the owners of the Empire State Building realized how much worse that accident could have been if it had taken place above a densely populated area such as downtown New York.

Left alone in the garden, Saeng dug out a hole and carefully lowered the “winter hibiscus” into it. She could hear the sounds of cooking from the kitchen now, the beating of eggs against a bowl, the sizzle of hot oil in the pan. The pungent smell of bitter melon wafted out, and Saeng’s mouth watered. It was a cultivated taste, she had discovered—none of her classmates or friends, not even Mrs. Lambert, liked it—this sharp, bitter melon that left a golden aftertaste on the tongue. But she had grown up eating it and, she admitted to herself, much preferred it to a Big Mac.

The “winter hibiscus” was in the ground now, and Saeng tamped down the soil around it. Overhead, a flock of Canada geese flew by, their faint honks clear and—yes—familiar to Saeng now. Almost reluctantly, she realized that many of the things that she had thought of as strange before had become, through the quiet repetition of season upon season, almost familiar to her now. Like the geese. She lifted her head and watched as their distinctive V was etched against the evening sky, slowly fading into the distance.

When they come back, Saeng vowed silently to herself, in the spring, when the snows melt and the geese return and this hibiscus is budding, then I will take that test again.

“Winter Hibiscus” by Minfong Ho, copyright © 1993 by Minfong Ho, from Join In, Multiethnic Short Stories, by Donald R. Gallo, ed.

Prompt

Read the last paragraph of the story.

"When they come back, Saeng vowed silently to herself, in the spring, when the snows melt and the geese return and this hibiscus is budding, then I will take that test again."

Write a response that explains why the author concludes the story with this paragraph. In your response, include details and examples from the story that support your ideas.

The greatest obstacle to the successful use of the mooring mast was nature itself. The winds on top of the building were constantly shifting due to violent air currents. Even if the dirigible were tethered to the mooring mast, the back of the ship would swivel around and around the mooring mast. Dirigibles moored in open landing fields could be weighted down in the back with lead weights, but using these at the Empire State Building, where they would be dangling high above pedestrians on the street, was neither practical nor safe. The other practical reason why dirigibles could not moor at the Empire State Building was an existing law against airships flying too low over urban areas. This law would make it illegal for a ship to ever tie up to the building or even approach the area, although two dirigibles did attempt to reach the building before the entire idea was dropped. In December 1930, the U.S. Navy dirigible Los Angeles approached the mooring mast but could not get close enough to tie up because of forceful winds. Fearing that the wind would blow the dirigible onto the sharp spires of other buildings in the area, which would puncture the dirigible's shell, the captain could not even take his hands off the control levers.

Two weeks later, another dirigible, the Goodyear blimp Columbia, attempted a publicity stunt where it would tie up and deliver a bundle of newspapers to the Empire State Building. Because the complete dirigible mooring equipment had never been installed, a worker atop the mooring mast would have to catch the bundle of papers on a rope dangling from the blimp. The papers were delivered in this fashion, but after this stunt the idea of using the mooring mast was shelved. In February 1931, Irving Clavan of the building's architectural office said, "The as yet unsolved problems of mooring air ships to a fixed mast at such a height made it desirable to postpone to a later date the final installation of the landing gear." By the late 1930s, the idea of using the mooring mast for dirigibles and their passengers had quietly disappeared. Dirigibles, instead of becoming the transportation of the future, had given way to airplanes. The rooms in the Empire State Building that had been set aside for the ticketing and baggage of dirigible passengers were made over into the world's highest soda fountain and tea garden for use by the sightseers who flocked to the observation decks. The highest open observation deck, intended for disembarking passengers, has never been open to the public.

Prompt
Based on the excerpt, describe the obstacles the builders of the Empire State Building faced in attempting to allow dirigibles to dock there. Support your answer with relevant and specific information from the excerpt.

A7: Essay Set 7

Prompt
Write about patience. Being patient means that you are understanding and tolerant. A patient person experience difficulties without complaining. Do only one of the following: write a story about a time when you were patient OR write a story about a time when someone you know was patient OR write a story in your own way about patience.

A8: Essay Set 8

Prompt
We all understand the benefits of laughter. For example, someone once said, “Laughter is the shortest distance between two people.” Many other people believe that laughter is an important part of any relationship. Tell a true story in which laughter was one element or part.