Automated Essay Grading Using Transformer Models

Othman Bensouda Koraichi
Stanford University
othmanb@stanford.edu

Elena Pittarokoili
Stanford University
elenapit@stanford.edu

Abstract

Automated essay grading is an emerging area in education technology that can revolutionize the educational experience for students and teachers. In this study, we leverage state-of-the-art transformer models to assess essay quality. We also try ensemble models that incorporate diverse transformer architectures, including RoBERTa (Liu et al. (2019)), ELECTRA (Clark et al. (2020)), and ALBERT (Lan et al. (2020)), and we also consider GPT-3's unique prompting capabilities to enrich the models with the essay scoring rubrics. Our objective is to test whether our ensemble can enhance the overall grading performance by blending these models' strengths, using the ASAP Essay Grading Dataset by the William and Flora Hewlett Foundation. Our results indicate that RoBERTa, Electra,and ALBERT perform significantly better than GPT-3, and that RoBERTa alone outperforms all other models, including the ensembles.

1 Introduction

Automated essay grading has emerged as a promising solution to the challenges posed by manual grading in education. With assessment arising as one of the fundamental components of contemporary educational systems and digitalization populating schools and classrooms, there has been a substantial increase in the number of tests administered to students and the amount of assessment data to measure learning outcomes. Nevertheless, the traditional process of grading essays is time-consuming, costly, and subject to human biases and inconsistencies. Furthermore, it limits the scalability and effectiveness of assessments, hindering the measurement of critical thinking and analytical skills in students. To address these limitations, there is a need for fast, effective, and affordable automated grading systems.Thus, the automation of the process is in great need for the sake of efficiency and reliability. Meanwhile, the task of automated scoring is considered more high-stakes compared to some other automation tasks as it can have significant individual ramifications, thus requiring careful research and design.

There are, in general, two classes of assessment scoring tasks that may be automated. One class can be sufficiently solved using exact matching techniques while the other requires more sophisticated scoring "models". Our study refers to a task where exact matching may not suffice. Still, automated scoring can be applied to a large collection of constructed-response tasks, thus encompassing a variety of approaches and systems. One dimension of diversity lies in the subject of the constructed-response task.

To solve this task on the ASAP Essay Grading Dataset,we propose an approach that leverages state-of-the-art transformer models to tackle automated essay grading. Transformer models such as RoBERTa, Electra, ALBERT, and GPT-3, have demonstrated remarkable capabilities in capturing linguistic patterns, contextual information, and structural features of text. By harnessing the power of these models, we aim to enhance the accuracy, consistency, and efficiency of essay grading.

Our approach involves fine-tuning each model to our specific task, and building an ensemble model that combines the predictions of these transformer models. We hypothesize that ensemble model benefits from the diverse perspectives and strengths of each individual model, resulting in a more comprehensive and robust assessment of essay quality. Additionally, we consider GPT-3 with its unique prompting capabilities, which could potentially provide valuable insights into the grading process by leveraging its language understanding and generation capabilities.

We believe that automated essay scoring models based on transformer architectures can accurately assess essay quality, provide consistent and reliable feedback, identify and evaluate key elements

such as grammar and organization, save time and resources, enhance objectivity, and adapt and improve over time through machine learning techniques. By leveraging the advancements in transformer models and ensemble-based approaches, we can improve the educational assessment process, promote fairness in grading, and provide valuable feedback to students, ultimately enhancing the overall educational experience.

We evaluate these hypotheses by training and evaluating the individual transformer models and comparing their performance against traditional methods. Furthermore, we will construct an ensemble model that combines the predictions of the individual models and analyze its impact on grading accuracy. Additionally, we investigate the effectiveness of GPT-3 in providing grading assessments through its prompting mechanism.

All in all, this research aims to contribute to the development of automated essay grading systems that are reliable, efficient, and scalable, thereby enhancing the educational assessment process and promoting fairness and objectivity in grading practices.

2 Prior Literature

Prior literature offers diverse perspectives on automated scoring and grading. It covers various aspects, including response types, challenges, interpretability, and fairness. While some papers provide comprehensive reviews, others propose novel approaches and highlight specific challenges. The contrasting perspectives on interpretability and the focus on fairness underscore the evolving nature of this field. Moving forward, it is crucial to address these challenges, refine methodologies, and develop models that promote transparency, fairness, and accuracy in automated scoring and grading. The following papers highlight the main challenges of automated scoring.

Erickson et al. (2020) highlight that teachers who grade open-ended responses often have their own criteria and requirements for assessment, which can vary widely. Factors such as articulation, knowledge demonstration, effort, grammar, and completeness may influence teachers' grading decisions. Detecting and normalizing these variations in grading practices to establish a common scale for assessment is complex. Hence, to gain insights into the variations in grading policies among teachers, Erickson et al. (2020) conducted a pilot study with 14 teachers who regularly use the educational platform ASSISTments. The study involved presenting the teachers with a subset of student responses to assess, including their own students' responses as well as responses from other teachers. The inter-rater agreement among the teachers revealed low levels of agreement in grading open questions. Moreover, the study found variations in the internal consistency of teachers' grades for their own students. Teacher surveys indicated that contextual factors beyond the content itself may have influenced grading decisions. Thus, the wide variation in grades and the presence of contextual factors pose challenges in developing automated models generalizable across teachers and students. As a solution to these challenges, the researchers added a teacher-level factor. The dataset used for developing models to assess student open responses was collected directly from ASSISTments. The final approach involved tokenization and the creation of a numeric representation of the parsed words. In this study, researchers used two tokenization approaches: standard count vectorizer splitting and the Stanford Tokenizer. They developed their models with supervised machine learning techniques such as XGBoost and Random forest, and with more complex deep learning algorithms combined with NLP approaches. Overall, they found that tree-based models offer interpretability, but deep learning models allow the use of embeddings to understand the semantics of words and equations in the student's response. Finally, they explored the impact of data quantity on performance and concluded that the model reaches its maximum potential in its current form at just under 55 training points per problem, meaning that additional data does not significantly enhance its ability to predict student grades.

Doshi-Velez and Kim (2017) warns that the European Union recently implemented regulations requiring algorithms that make decisions based on user-level predictors to provide explanations ("right to explanation"). This highlights the urgency and importance of interpretability in the field. Moreover, the volume of research on interpretability is rapidly increasing. The authors' main contribution is to give a formal taxonomy of interpretability evaluation in the context of Machine Learning. According to the authors, interpretability is not required for all machine learning systems. For example, ad servers, postal code sorting, and air-

craft collision avoidance, all operate without human intervention and do not require explanations. In such cases, there are no significant consequences for incorrect outputs, or the problems have been well-studied and validated, establishing trust in the system's decision-making. However, the need for interpretability arises when there is an incompleteness in the problem formalization, which creates a fundamental barrier to optimizing and evaluating ML systems. Hence, interpretability becomes necessary to bridge this gap and provide insights into the system's decision-making process. The taxonomy developed by the researchers aims to the quality of an explanation in the context of its end-task, such as whether it results in better identification of errors, new facts, or less discrimination. The authors emphasize the importance of matching the evaluation approach with the claimed contribution and suggest categorizing applications and methods using a common taxonomy. This taxonomy could be divided into 3 approaches : Application-grounded evaluation, which involves conducting human experiments within a real application, where the model is evaluated with respect to its intended task. This approach aligns with evaluation methods used in human-computer interaction and visualization communities. Human-grounded evaluation, which focuses on conducting simpler human-subject experiments that maintain the essence of the target application. These experiments can be performed with lay humans and are useful for testing more general notions of explanation quality. Functionally-grounded evaluation, which does not involve human experiments, but instead, uses a formal definition of interpretability as a proxy for explanation quality. This approach is appropriate when models have already been validated or when human experiments are impractical or unethical.

Lipton (2018) discusses and criticizes the notion of interpretability in Machine Learning. He highlights that linear models are not strictly more interpretable than deep neural networks, since it depends on the notion of interpretability being considered. For example, with respect to algorithmic transparency, this claim seems uncontroversial, but given high-dimensional or heavily engineered features, linear models lose simulatability or decomposability, respectively. The author stresses that any assertion regarding interpretability should specify a clear definition and provide evidence that the offered interpretation achieves the desired objective. Furthermore, transparency may not always align with the broader objectives of artificial intelligence. For instance, the short-term goal of building trust with doctors by developing transparent models might clash with the longer-term goal of improving health care. Lastly, optimizing algorithms to present plausible explanations can lead to misleading interpretations. It is essential to be mindful of the potential for reproducing biased or discriminatory behavior at scale. The article suggests future research directions, such as developing richer loss functions and performance metrics to bridge the gap between real-life objectives and machine learning objectives.

Kumar and Boulanger (2020) assess the possibility of deep learning in automated essay scoring and aims to address the dilemma of accuracy-interpretability trade-off in AES models by providing XAI (explainable AI) solutions for future neural network models. The authors introduce a feature-based multi-layer perception (MLP) deep neural network as its predictive model with a huge pool of linguistic indices (n=1592), followed by semi-automatic feature selection through pruning and regularization. A SHapley Additive exPlanations (SHAP) explanation model is then trained on top of the predictive model to inspect feature contributions. The authors also additionally probe the trustworthiness of their explanation models as well as provide additional tools for teachers to leverage the system for formative feedback.

Bridgeman et al. (2012) tackle the fairness aspect of algorithms, and highlight that although there is a strong overall relationship between human and machine scores, data scientists should not overlook the possibility of significant differences for specific gender, ethnic, or country groups when designing NLP models. The study of Bridgeman et al. (2012) examined essay data from high-stakes testing programs like the TOEFL or the GRE. Across most subgroups, human and machine scores were highly similar, but there were notable exceptions that could not be ignored. Moreover, the researchers showed that essays from groups with different native languages were scored differently by humans and machines. This issue has been further demonstrated in Burstein and Chodorow (1999), who evaluated essays from the Test of Written English (TWE) that were scored both by humans and by a machine. They found a significant interaction between type of scoring (human or machine)

and language group such that Arabic and Spanish speakers appeared to receive relatively higher scores from humans than from the machine while Chinese speakers received higher scores from the machine. This differential performance for subgroups clearly has fairness implications for individuals in high- to moderate-stakes assessments, and we should take it into account while developing our model. All in all, Bridgeman et al. (2012) highlight that studying fairness involves more than just examining differences; it requires understanding the underlying causes. For instance, the researchers have shown in an experiment on TOEFL essays that results for Korean and Chinese students might suggest that e-rater favors Asian languages. However, the lack of a difference for Japanese students suggested more subtle explanations might be needed. Furthermore, although Hindi and Telugu are both languages spoken in India, human scoring provided a relative advantage only for speakers of Hindi, which means that language may simply be a proxy for other meaningful distinctions. In this case, the contrast between the findings for Hindi and for Telugu may be because of cultural differences between northern and southern India rather than due to linguistic differences. This suggests that fairness is an iterative process that requires to formulate and verify multiple hypotheses.

In addition to the aforementioned challenges and corresponding attempts on the task, given the task is an application in the field of education, copious literature has further explored the ramification of applying AI-based models to the scoring tasks. Two themes have emerged: interpretability and fairness. On interpretability, Doshi-Velez and Kim (2017); Lipton (2018) present contrasting perspectives. Doshi-Velez and Kim (2017) develop a taxonomy for evaluating interpretability, while Lipton (2018) criticizes the notion, emphasizing potential misleading interpretations and the need for alignment with broader objectives. These papers underscore the ongoing debate and the importance of defining and evaluating interpretability in the context of automated scoring. Kumar and Boulanger (2020) contribute to addressing the dilemma of accuracy-interpretability trade-off in automated essay scoring models by applying an XAI framework on deep learning models. The authors propose a feature-based MLP deep neural network and utilize SHAP explanations to enhance interpretability. The authors highlight the trustworthiness of their explanation models and provide additional tools for teachers. This paper exemplifies efforts to strike a balance between accuracy and interpretability, fostering transparency in the scoring process.

Finally, Bridgeman et al. (2012) explore the fairness aspect of automated scoring. They examine differences in human and machine scores for specific subgroups, emphasizing the need to understand underlying causes. These findings prompt researchers to address potential biases and ensure equitable treatment for all students.

3 Data

The following table summarizes the size of the training and validation sets for the 8 types of essay.

Essay set	**Type of essay**	**Grade level**	**Set size**
1	persuasive / narrative / expository	8	1783
2	persuasive / narrative / expository	10	1800
3	source dependent responses	10	1726
4	source dependent responses	10	1772
5	source dependent responses	8	1805
6	source dependent responses	10	1800
7	persuasive / narrative / expository	7	1569
8	persuasive / narrative / expository	10	723

Table 1: Overview of Essay Sets

The data provided for the Hewlett Foundation Automated Student Assessment Prize (ASAP) consists of hand-scored essays, which are used for building, training, and testing scoring engines. Each essay was independently graded by multiple human raters, and their consensus was taken as the final grade.

The essays range in length from approximately 150 to 550 words, with an average length falling within this range. Some essays may be more dependent on source materials than others, indicating variations in the essay type. The number of training essays varies depending on the specific prompt. Each essay in the training data is accompanied by one or more human scores. In some cases, multiple

human scores are provided to assess the reliability of the human scorers. However, predictions are to be made to the resolved score, indicating the final score assigned to the essay.

4 Model

Our approach encompasses several phases. In the first phase, we fine-tune three transformer-based language models: RoBERTa, ELECTRA, and ALBERT. These models, pre-trained on large-scale text corpora, have shown remarkable capabilities in understanding and generating human-like text. By fine-tuning these models on our specific task - essay grading - we adapt their generalized language understanding to our particular domain.

Alongside these models, we utilize OpenAI's GPT-3, employing few-shot learning prompting. Rather than fine-tuning GPT-3 on the grading task, we craft prompts that instruct the model to assess an essay, capitalizing on GPT-3's ability to understand and follow complex instructions in natural language. This enables us to harness GPT-3's impressive language understanding and generation capabilities without needing to fine-tune the model, which can be prohibitively expensive or even unfeasible.

After training each of these models, we use their predictions as inputs for two ensemble techniques: linear regression and gradient boosting via XGBoost. The rationale behind this approach is that while each model might make different errors in grading, by learning from all their predictions, we can compensate for individual weaknesses and accentuate strengths, improving overall grading performance.

In particular, we test two variations of our ensemble approach. In the first variant, we use the predictions from the three fine-tuned transformer models (RoBERTa, ELECTRA, and ALBERT) as inputs to the ensemble models. In the second variant, we add the grades predicted by GPT-3 to the input, exploring whether the addition of GPT-3's 'opinion' can enhance the ensemble's performance.

AlL in all, our approach tests the strengths of multiple powerful language models and ensemble techniques to build a robust, accurate automated essay grading system.

5 Methods

The models for this study include individual transformer models such as RoBERTa, Electra, and ALBERT, as well as GPT-3. These models have been chosen due to their proven success in numerous NLP tasks. We then investigate whether an ensemble model that aggregates predictions from these individual models to generate final grades.

We then test ensemble approaches, leveraging the strengths of multiple transformer models, in an attempt to improve performance in automated essay grading. Each model may capture different linguistic and structural features of the essays, and by combining their predictions, we aim to obtain a more holistic and accurate assessment of essay quality.

5.1 Metrics

The primary metric for evaluating our models' performance is the Quadratic Weighted Kappa (QWK) score, which was used in the Automated Student Assessment Competition. This metric assesses the agreement between the grades assigned by our models and the actual grades given by human raters. The QWK score is especially well-suited to this task because it takes into account the possibility of agreement occurring by chance, providing a more robust measure of performance than simple accuracy.

The QWK ranges from 0 (random agreement) to 1 (complete agreement). If there is less agreement between the raters than expected by chance, the metric may go below 0. The mean of the quadratic weighted kappa is calculated across all sets of essays after applying the Fisher Transformation to the kappa values. The quadratic weighted kappa is calculated based on the comparison of scores between the automated model and the resolved human scores. This performance metric ensures that the automated scoring systems closely match the grading accuracy and reliability of human expert graders.

The formula to calculate the quadratic weighted kappa is as follows:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}$$

Where:

- κ represents the quadratic weighted kappa.
- O_{ij} is the observed agreement frequency between rater A and rater B for rating i and rating j.

- E_{ij} is the expected agreement frequency between rater A and rater B, assuming no correlation between the ratings.
- w_{ij} is a weight assigned to each rating combination (i, j) based on the disagreement between the ratings. The weights are typically calculated using a quadratic function, such as the Fleiss' kappa or Cohen's kappa weights.

The weights w_{ij} contribute to the overall calculation of the quadratic weighted kappa and reflect the degree of disagreement between the raters for each rating combination. These weights help capture the importance of the level of disagreement, with higher weights assigned to larger disagreements and lower weights assigned to smaller disagreements. By comparing the observed agreement frequencies (O_{ij}) and the expected agreement frequencies (E_{ij}) while taking into account the weights (w_{ij}), the QWK provides a comprehensive measure of agreement that considers both the distribution of ratings and the degree of disagreement between the raters.

5.2 Data Preparation and Encoding

We utilize PyTorch, a powerful open-source library for deep learning. For each transformer model, we load the pre-trained base model along with its corresponding tokenizer. The tokenizer transforms the raw text of an essay into a format that the model can process. Each essay is tokenized, and attention masks are created to distinguish relevant content from padding. This results in tensors of input IDs and attention masks that serve as inputs for the model.

5.3 Training

The training procedure follows standard practice for training neural networks with gradient descent. We use the AdamW optimizer along with a learning rate scheduler. For loss calculation, we use Mean Squared Error (MSE) which is a common choice for regression problems. The gradients are computed via backpropagation and then used to update the model's parameters. We also apply gradient clipping to prevent the exploding gradients problem. This process is repeated for a specified number of epochs.

5.4 Evaluation

Evaluation is conducted on a separate validation set. Each model generates grade predictions for the essays in this set, and these predictions are then compared to the actual grades to determine the model's grading performance. We use the Quadratic Weighted Kappa (QWK) score as our performance metric, which measures the agreement between the predicted and actual grades.

5.5 Prediction

After training, the model can be used to generate essay grade predictions. For this, we set the model to evaluation mode, encode the essays in the same way as during training, and load the data into a DataLoader object. We then iterate over the data in mini-batches, feeding each batch into the model and collecting the model's outputs. These outputs are converted into a format suitable for analysis and added to a new dataframe, which is returned by the function.

5.6 Ensemble Approach

The ensemble strategy is a central part of our methodology. After training each transformer model, we gather their predictions on the same validation set and determine the final grade using an aggregation strategy. Simple averaging, weighted averaging based on each model's QWK score, and more sophisticated stacking or boosting methods will be used to generate the final prediction.

5.7 GPT-3 and Prompting Strategy

Alongside transformer models, we also explore GPT-3's prompting capabilities for automatic essay grading. This involves crafting prompts that instruct GPT-3 to assess an essay, leveraging its language understanding and generation capabilities. To overcome prompt length limitation,

6 Results

Set	RoBERTa	Electra	Albert	GPT-3
1	0.78	0.71	0.71	0.02
2	0.74	0.73	0.73	0.25
3	0.72	0.66	0.57	0.23
4	0.86	0.74	0.82	0.31
5	0.74	0.70	0.73	0.15
6	0.84	0.76	0.83	0.54
7	0.84	0.82	0.85	0.11
8	0.19	0.22	0.20	-0.01
Avg	0.71	0.67	0.68	0.20

Table 2: Model Performances - Individual Models

Set	LinReg	XGB	LinRegGPT	XGBGPT
1	0.76	0.64	0.76	0.57
2	0.74	0.60	0.75	0.59
3	0.71	0.74	0.71	0.72
4	0.84	0.77	0.84	0.80
5	0.70	0.63	0.70	0.63
6	0.81	0.78	0.81	0.80
7	0.83	0.81	0.83	0.82
8	0.25	0.45	0.25	0.38
Avg	0.71	0.68	0.71	0.66

Table 3: Model Performances - Ensemble Models

7 Analysis

Table 2 shows the performance of individual models, RoBERTa, Electra, Albert, and GPT-3, across eight different essay sets, as well as their average performance.

Looking at the average scores, RoBERTa leads with a score of 0.71, followed closely by Albert at 0.68, and then Electra at 0.67. This suggests that RoBERTa and Albert have slightly better performance in general across different essay sets compared to Electra.

On the other hand, GPT-3 stands out with a significantly lower average score of 0.20, indicating its underperformance in this grading task. This aligns with our expectations as GPT-3, being an autoregressive model, is more suited to text generation tasks rather than grading tasks.

The scores for the individual essay sets display a degree of variation, with set 8 displaying the lowest scores for all models. This suggests that essay set 8 might be more challenging to evaluate for all models.

Overall, it is clear that different models have varying strengths when applied to different essay sets, emphasizing the importance of model selection and ensemble techniques in real-world grading applications.

Table 3 presents the performance of ensemble models, specifically Linear Regression (LinReg), XGBoost (XGB), Linear Regression with GPT-3 (LinRegGPT), and XGBoost with GPT-3 (XGBGPT), on the same eight different essay sets, along with their average performance.

Examining the average scores, Linear Regression models (both with and without GPT-3) demonstrate a slight edge with a score of 0.71, followed closely by XGBoost at 0.68. When GPT-3 is included in the ensemble (XGBGPT), the average performance slightly decreases to 0.66. This might suggest that adding GPT-3 to the ensemble does not necessarily enhance the performance for this grading task, which is in line with our earlier observations about GPT-3's performance.

	Feature	With GPT-3	Without GPT-3
1	RoBERTa	0.502645	0.508146
2	Electra	0.064120	0.061515
3	Albert	0.272255	0.273370
4	GPT-3	0.017151	NaN

Table 4: Linear Regression Coefficients

Table 4 presents coefficients of various transformer models (RoBERTa, Electra, Albert, GPT-3) as computed by two separate linear regression ensembles - one with GPT-3 and the other without GPT-3. In both ensembles, the coefficients represent the weightage or the importance given to each model's predictions in determining the final prediction of the ensemble.

All in all, we conclude that RoBERTa, ELECTRA, and ALBERT have a good individual performance on each essay set, except the last one. Moreover, they perform significantly better than GPT-3. We also observe that ensemble models do not perform necessarily better than RoBERTa on average, suggesting that most of their predictive power might come from RoBERTa. We believe that the poor performance on the 8th essay set comes from its high complexity, but also from its set size. As mentioned in the Data section, this essay set has much less datapoints than the other sets, but it is also inherently more complicated. One very interesting point to notice is the ability of the XGBoost ensemble to perform much better than RoBERTa on the 8th essay set.

We also notice that when GPT-3 is included in the ensemble, it receives the least weight (0.017151) compared to other models. It's interesting to note that despite the hype around GPT-3's capabilities, in this specific context, the ensemble model doesn't seem to rely heavily on its predictions. Given that GPT-3 is an autoregressive model, it's primarily designed for generating text, making it an exceptionally powerful tool for tasks that involve creating human-like, coherent narratives. However, the task at hand here involves predicting the grade of an essay, a decidedly different chal-

lenge that doesn't explicitly require text generation capabilities.

As such, it's not entirely surprising that GPT-3 received the lowest weight in the ensemble with GPT-3 included. In contrast, RoBERTa, Electra, and Albert, which are transformer models more oriented towards understanding and encoding the nuances in the input text, received higher weights. This is because they are likely better equipped to discern the underlying quality of the essays, thus providing more accurate grade predictions.

This analysis highlights the importance of empirical validation of model performance, as opposed to relying solely on theoretical expectations.

Also, the findings above underscore the importance of carefully matching the characteristics and strengths of a model to the requirements of the task at hand. While GPT-3 is a groundbreaking model in its own right, the context-specific nature of model effectiveness becomes evident in this case. It showcases the crucial insight that the most sophisticated or complex model may not always be the most suitable or effective for a particular task, especially when the model's strengths don't directly align with the task's requirements.

8 Conclusion

In conclusion, our exploration into automated essay grading using multiple transformer models and ensemble techniques provides insightful findings and lessons. Despite the powerful capabilities of each individual model, including RoBERTa, ELECTRA, ALBERT, and GPT-3, their effectiveness varies across different essay sets. In particular, the performance on the eighth essay set, marked by its high complexity and smaller size, was notably lower than the other sets. Although our ensemble model, notably the XGBoost variant, managed to exceed the performance of RoBERTa in this specific scenario, the result illuminates the challenge and need for creating models that perform consistently across diverse data characteristics.

Interestingly, the inclusion of GPT-3, an acclaimed autoregressive language model, in our ensemble did not significantly contribute to the grading process, as reflected in its lowest weight assignment. This observation, contrasting with the conventional hype around GPT-3, signals the importance of aligning a model's core competencies with the task's specific requirements. In the context of essay grading, understanding and encoding the subtleties in the input text, something the other transformer models excel at, proved more critical than GPT-3's prowess in text generation.

Our study underscores the necessity of empirical assessment of model performance and highlights the value of ensemble methods, where multiple models with complementary strengths work together to boost overall performance. Importantly, it serves as a reminder that the choice of a model should be informed by its suitability to the task at hand and not merely by its complexity or popularity.

9 Project Limitations

Despite our encouraging findings, our study is not without limitations. A significant constraint is the absence of fairness assessment. Fairness, in the context of automated grading, pertains to the ability of the model to grade essays impartially, not influenced by factors like the student's demographic or socioeconomic background. Regrettably, our dataset did not incorporate these aspects, hence, we were unable to evaluate our model's performance in this critical aspect. Future work should consider this facet to ensure that automated grading systems do not inadvertently perpetuate or exacerbate existing inequities.

Secondly, we must address the lack of interpretability. Despite our models' performance, it's inherently challenging to discern exactly how it's making its decisions. This opacity could limit our ability to understand and further improve the system or diagnose and rectify potential mistakes or biases. As we continue to develop and refine automated grading systems, it will be vital to consider and incorporate methods that improve interpretability without compromising performance.

In conclusion, while our approach offers an intriguing pathway for the development of robust automated essay grading systems, it also highlights crucial areas for improvement and further exploration. We remain hopeful that future research will address these limitations, paving the way for more effective, fair, and interpretable automated grading solutions.

Authorship Statement

Each author acknowledges that this final version was read, and all authors had the opportunity to contribute to its content and agreed to the order of their names in the author list. Any conflicts of

interest, financial or otherwise, that may affect the authors' ability to present data and interpretations honestly have been disclosed.

References

Brent Bridgeman, Catherine Trapani, and Yigal Attali. 2012. Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, 25(1):27–40.

Jill Burstein and Martin Chodorow. 1999. Automated essay scoring for nonnative English speakers. In *Computer Mediated Language Assessment and Evaluation in Natural Language Processing*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

J. A. Erickson, A. F. Botelho, S. McAteer, A. Varatharaj, and N. T. Heffernan. 2020. The automated grading of student open responses in mathematics. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 615–624.

Vivekanandan Kumar and David Boulanger. 2020. Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education*, 5.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, et al. 2019. Roberta: A robustly optimized bert pretraining approach.