Developing a Transformer-Based Automated Essay Scoring System with Explainability Considerations

Oscar Dos Santos Nunes

March 2025 - September 2025

Contents

Chapter 1

Introduction

This is a placeholder for the real introduction. will likely require being shortened down and restructured to better encompas the final projects outcomes and completed work

1.1 Background and Context

Automated Essay Scoring (AES) systems have been in development for over fifty years, evolving from statistical and rule-based approaches into increasingly sophisticated machine learning and deep learning technologies. Early developments such as Project Essay Grade (PEG) (Ellis Batten Page 1966) and e-rater (**attali2006e˙rater**) relied on handcrafted features, including high-level syntactic and lexical features. While these methods demonstrated some success and saw real-world application, they often struggled to capture the semantic and contextual depth found in student writing (M. D. Shermis and Burstein 2013).

With the growth of more data-focused approaches in natural language processing (NLP), AES has increasingly adopted machine learning techniques. Traditional models such as support vector machines (SVM), random forests, and ridge regression were trained on handcrafted features to predict human-marked essay scores. However, these models were still limited by their reliance on manually created inputs, which creates constraints with scalability and generalisation across writing prompts and student population groups.

The growth of deep learning, particularly with transformer-based models, has significantly advanced AES research. Pre-trained language models such as BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019) have demonstrated impressive results across various NLP tasks, thanks to their ability to model contextual relationships within text. When compared to more typical machine learning techniques, studies using BERT for AES have shown substantial gains in accuracy, especially when using benchmark datasets such as the ASAP (Automated Student Assessment Prize) corpus (Taghipour and H. T. Ng 2016a).

Despite these advancements, a number of obstacles still exist. Deep learning models frequently act as "black boxes", providing little to no room for interpretation. This is a major drawback for AES, as user trust and educational integration rely on the reasoning behind a score. In recent years, there has also been growing concern about algorithmic bias, particularly when scoring responses from under-represented student groups or non-native English speakers (Blodgett et al. 2020). These issues raise questions about fairness, generalisability, and the ethical use of AES in high-stakes contexts.

Furthermore, much of current AES work focuses on holistic scoring without exploring the formative aspects of feedback, rubric-based scoring, or the integration of multi-modal inputs. Expandability methods such as SHAP (Lundberg and Lee 2017) or attention-based visualisations have been actively re-

searched as a means of improving model transparency. Hybrid systems that combine neural and symbolic components have also shown promise in enhancing interpretability and control.

This project builds on these advancements by developing an AES model using deep learning, evaluating its performance and fairness, and investigating potential pathways for expansion. Through comparing results against traditional baselines and considering explainability and ethical implications, this work aims to contribute to research seeking to make AES systems more robust, transparent, and educationally useful.

1.2 Problem Statement

The rapid growth of digital education and online learning platforms has brought a fresh interest in an efficient, scalable and consistent method for assessing students' work. Essay writing remains one of the vital methods for evaluating students' understanding of content as well as their ability to communicate ideas and think critically. However, this method requires significant time and resources from educators when marked manually. Human grading is not only more labour-intensive but it is also prone to consistency issues due to marker variability, fatigue and personal bias (Ellis B. Page and Petersen 2003). In large-scale settings, the demand for a scalable assessment has led to the exploration of AES systems which are capable of mimicking human judgement in evaluating written answers.

Despite the decades of research and development within the field (ranging from early rule-based systems to more modern neural linguistic approaches), there are still significant limitations. Traditional AES systems such as e-rater (Attali and Burstein 2006) and Project Essay Grade (PEG) (Ellis B. Page and Petersen 2003) relied on shallow linguistic features (e.g. word count, spelling, grammar) that do not capture deeper semantic and argument quality (Dikli 2006). More recent advancements with transformer-based models such as BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019) offer promising alternatives through leveraging contextual embeddings and attention mechanisms to better model the meaning and coherence of a given text (Devlin et al. 2019).However, these models still struggle with generalisation across writing prompts, transparency within their decision-making, and potential still hold bias against certain writing styles or demographic groups Beigman Klebanov, Madnani, et al. 2021.

The core problem of this project seeks to address, is how to design and evaluate an AES model that offers both high predictive accuracy and with interpretability, while remaining robust across different prompts. In particular This project will examine:

- the efficiency of fine-tuned transformers models for essay scoring tasks;
- the generalisability of models across prompt types and writing levels;
- the trade off between model complexity, performance and interoperability;

This problem is significant both in an academia and practical context. From the perspective of research, it advances the state-of-the-art in educational NLP and model explainability. From a practical standpoint, it offers potential improvements in the fairness and scalability of the assessment tools used in modern educational institutions, especially focusing on the needs of large-scale online learning platforms such as Coursera and edX.

1.3 Aims and Objectives

The aim of this project is to develop a transformer-based AES system of which achieves high predictive performance while improving upon current work in interpretability and fairness. Through the use of state-

of-the-art deep learning models such as BERT and RoBERTa, this work hopes to contribute towards more accurate, trustworthy and scalable alternatives to the more traditional essay marking systems.

The primary objectives of this project are as follows:

1. **Review and evaluate existing work:** Conduct a comprehensive literature review of traditional and modern AES systems, focusing on their performance, limitations and explainability.

2. **Data selection and Preprocessing:** Identify appropriate datasets for training and evaluation. Preprocess essays through tokenization, cleaning and normalize score variations to enable effective modeling at later stages.

3. **Model development and comparisons:** Implement and fine-tune a transformer-based model for essay scoring. Develop a performance baseline off traditional techniques and previous AES systems for comparison.

4. **Performance and robustness evaluation:** Asses the accuracy of each models on standard performance metrics such as Quadratic Weighted Kappa (QWK) and Mean Squared Error. Evaluation the models across different prompts and writing levels.

5. **Explainability and transparency:** Apply interpretability techniques such as SHAP or attention-based visualisetion to provide insights into model behaviors and decision making processes.

6. **Bias and fairness analysis:** investigate potengial demographic or linguitic biases within the models predictions. Expolore the impact of these biases and evaluate potential solutions.

7. **Explore extensibility:** Discuss and develop potential pathways for extending model performances. Such as feedback generation, rubric-based scoring or the addaption of multi model inputs (Audio and handwriten answers).

1.4 Research Questions

This research is guided by the following key questions, which collectively shape the development, evaluation, and potential impact of the proposed Automated Essay Scoring (AES) system:

1. **To what extent can transformer-based models outperform traditional machine learning approaches in essay scoring tasks?** This question examines the predictive performance of deep learning models such as BERT and RoBERTa in comparison to classical baselines like ridge regression or support vector machines.

2. **How well do transformer-based AES systems generalise across different essay prompts and writing levels?** Here, the focus is on evaluating model robustness and adaptability across diverse input types, including prompt-specific and cross-prompt scenarios.

3. **What trade-offs exist between predictive accuracy, model complexity, and interpretability in AES?** This question investigates whether performance gains from complex neural models justify potential losses in transparency and ease of understanding.

4. **Can model explainability techniques make AES systems more interpretable and trustworthy in educational settings?** This explores the practical value of applying tools such as SHAP or attention visualisations to clarify how model predictions are generated.

5. **What forms of algorithmic bias are present in AES outputs, and how can they be detected and mitigated?** This question addresses the ethical dimension of AES systems, particularly their fairness across student demographics and linguistic backgrounds.

6. **Is it feasible to extend the AES system with additional features, such as feedback generation or support for handwritten inputs?** This final question considers the future potential of AES models to go beyond scoring, towards becoming comprehensive tools for formative assessment.

1.5 Significance of the Study

AES have an ever-increasing importance within education, particularly with the rise of online learning environments such as MOOCs (Massive Open Online Courses), remote assessments, and digital feedback systems. Traditional human methods are heavily time-consuming, expensive, and prone to inconsistency. As a result, the development of a scalable, fair, and explainable AES system is a growing concern for educational tech.

This study contributes to the field through advancing the application of transformer-based deep learning models to the domain of essay evaluation. While prior research has demonstrated the performance benefits of these models, relatively little work has been examined for ts transparency, fairness, and practical applicability. This project directly addresses that gap by incorporating explainability and fairness analysis into the model's development.

Furthermore, the study explores the trade-offs between a model's accuracy and its interpretability, a critical consideration for its potential use within education. Through this comparison, this research will aid with guiding future work to mitigate biases and provide additional baselines for performance across a range of demographics.

Practically, the findings may support the development of a more robust essay evaluation tool for institutions seeking a new approach to assessment marking methods. Academically, it aids with the broader research into explainable AI and bias mitigation within educational NLP research.

1.6 Dissertation Structure Overview

This section will be written at a later date once later section structure is completed

Chapter 2

Literature Review

2.1 Historical Evolution of Automated Essay Scoring

The first notable milestone within AES was Ellis Page's development of Project Essay Grade (PEG) in the 1960s, pioneering the idea of using computers to assess writing quality. PEG worked through the identification of measurable surface features (such as word count, spelling, and grammar) as measures of writing quality, then fitting a regression model in order to predict human-assigned scores. Somewhat surprisingly, with such simple processes, Page reported correlations of around 0.7 between the PEG's scores and that of human raters. While this early work was far lacking in depth compared to that of today's, it demonstrated the "imminence of grading essays by computer" (Ellis Batten Page 1966) and generated excitement about the opportunities of algorithms judging complex writing.

After a significant break in research, in the 1990s many more AES systems began being developed. This was largely due to the significant advancements made within computing power and NLP research. One of the more notable advancements was e-rater, developed by the Educational Testing Service (ETS), which was first introduced in 1999, utilising a large set of linguistic features in order to mimic the scoring of a human marker (Attali and Burstein 2006). Around the same time, other paradigms emerged: the Intelligent Essay Assessor (IEA) applied latent semantic reference tests (LSA) to assess content's relevance by comparing the semantic similarity of an essay to a corpus of reference texts. This content-focused approach complemented the earlier surface-feature methods (Landauer, Laham, and Foltz 2003), demonstrating the shift towards capturing more of an essay's substance and meaning in the automated scoring process.

In the early 2000s, research into AES increasingly included machine learning algorithms capable of leveraging both surface and content features. The more traditional statistical models were augmented and/or replaced by techniques such as support vector machines (SVM) and Bayesian classifiers, which were trained on large datasets of essays to optimise their scoring accuracy (M. D. Shermis and Burstein 2013). A key event for research was the **Hewlett Foundation's Automated Student Assessment Prize (ASAP)** competition in 2012; this released a corpus of 13,000 scored essays across eight prompts and accelerated progress and benchmarking of what was possible with handcrafted features. In the mid-2010s a new breakthrough was found with the development of deep learning. Neural network-based AES models (beginning with Taghipour and Ng's 2016 RNN model) learned essay representations from the raw text, rather than relying on hand-created features. These models achieved scoring accuracies that were comparable to human raters, often outperforming the previously developed algorithm (Taghipour and H. T. Ng 2016a). This rise of deep learning also renewed work on cross-domain scoring (training the model on one prompt but assessing work on another prompt), which had largely been abandoned in the early 2000s due to the lack of generalisability of the models (Beigman Klebanov, Burstein, et al. 2021).

Alongside the improvements in accuracy, more recent work has brought attention towards AES systems' fairness, validity, and explainability. It is increasingly recognised that AES systems must not only just be reliable but also fair and transparent with its actions. Recent work has analysed for algorithmic favouring or penalising of certain groups. For example, studies have checked for biases in scores related to demographics (e.g. native language or gender), finding that in some cases performance can be significantly skewed (Blodgett et al. 2020). In parallel, research also stresses the importance of making the "black box" of deep learning models more interpretable. The latest systems attempt to provide feedback on the writing tasks, such as breaking the score down into a series of sub-scores looking at organisation, grammar, content and style. This push for explainability is driven by the need for trust and accountability within an educational context, where students and educators require an understanding of how the system arrived at its scores (Lundberg and Lee 2017).

In summary, over the last 50 years, AES has developed from simple rule-based systems to complex deep learning models. The field has shifted from a focus on surface features to a more nuanced understanding of content and context, with an ever growing focus on fairness and explainability. This evolution reflects the broader trends in NLP and machine learning, as well as the increasing importance of ethical considerations in AI applications.

2.2 Rule-Based and Traditional ML Methods

2.3 Deep Learning and Transformer-Based Models

The introduction of deep learning marked a key turning points for AES systems, allowing for models to learn features directly from raw text over the previous reliance on hand-crafted inputs. Early neural networks such as taghipour and Ng's(2016) RNN-rule based approach achived near-human scoring through modeling sequential dependencies in the texts. Heirachical and attention-based networks followed this, which captured higher-level essay structures and enabled better interpretability. These architectures offered improved generalisation compared to traditional methods, however the performance gains where stil limited by their sequential processes.

Transformer based models such as BERT and RoBERTa brought a significant boost to AES system capabilities. Thes models utilise self attantion mechanisms to encome complex relationships within the text, allowing for an understanding of semantics, coherance and broad patters more effectively. Fine-tuning transformers on AES datassets such as ASAP has consistently outperformed prior approahces across both accuracy and generalisability benchmarks.

However, while transformers has provided strong predictive performance, they introduce many challengees with regards to explanability and fairness. These models act as "black boxes" with very limited transparency as to how a decision is made. This is highly problematic within the setting of education as stakeholders will require an insight into scoring raionales. Furthermore, models trained on specific prompts may not generalise well, and there are many concerns about potential algorithmic-biases that may favour or penalise certain groups of students.

Regardless of these drawbacks, transformer based models are currently the state-of-the-art within AES systems. The ever ongoing research into explainable AI and cross-prompt generalisation is beggining to adress these shortcomings, suggesting that more robust and fair systems are on the horizon.

2.4 Datasets for AES

AES research has been fueled by the availability of large, annotated essay datasets. Commonly the Automated Student Assessment Prize (ASAP) corpus is used, released by the Hewlett Foundation in 2012. Consisting of over 12,000 essays across eight prompts, each with prompt-specific scoring rubrics, ASAP has become the benchmark for AES systems. It enabled both prompt-specific and propt-agnostic evaluations and remains as a standard for comparative performative analysis.

An alternative to ASAP is the TOEFL11 corpus, containing essays written by non-native english speakers. This data set is often used for evaluating a systems ability to handle second-language writing. The FCE dataset by Cambridge learner Corpus also supports trait-based scoring with fine-grained annotations on vocabulary, cohesion and grammar. These datasets combined with others aid in research into both holistic and analytic scoring, aiding with the expansion of AES to a broader range of writing tasks.

More recently, datasets such as ASAP++ has added multi-trait labels, enabling models to be move beyond holistic scoring towards more detailed feedback generation. Along side this there are also corpus' tailor for specific writing dimensions such as Argument Annotated Essays (AAE) which evaluate the persuasiveness and structure of arguments within text. Resources such as these have supported the development of systems capable of evaluating across different genres and to comprehend a wider range of instruction goals.

Despite these advancements, much of the AES datasets are still limited to English and typed responses, with very few multi-lingual datasets or multi-modal datasets available. This significantly restrict our models ability to generalize across languages and formats. Expanding this landscape of datasets to be more inclusive remains a key step in developing a Generic AES system.

2.5 Evaluation Metrics in AES

For AES system the standard evaluation metric is the Quadratic Weighted Kappa (QWK), this measures the agreement between a models predictions and human scores. QWK is a quadratic weighted version of Cohen's Kappa, which accounts for the ordinal nature of essay scores. It ranges between -1 and 1, where 1 indicates perfect agreement, 0 indicates no agreement, and negative values indicate worse than random agreement. QWK is particularly useful in AES as it penalises larger discrepancies between predicted and actual scores more heavily than smaller ones.

In addition to QWK, regression metrics are very commonly used such as the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). These metrics provide an insight into the magnitude of predictive erros and a highly beneficial in continuing predictive tasks. Mean Absolute Error (MAE) is also occasionally used due to its resistance to outliers, though it lacks the sensitivity of QWK to ordinal score differences.

Some researchers report accuracy, especially in the form of accuracy (i.e., within ± 1 of the human score), to reflect more lenient scoring expectations similar to inter-rater variability among humans (M. Shermis and Burstein 2013; Taghipour and H. T. Ng 2016b). However, accuracy alone may not fully reflect a model's ranking reliability or nuanced error distribution, making it less preferred in rigorous evaluations (Williamson, Xi, and Breyer 2012; Yannakoudakis, Briscoe, and Medlock 2011).

Ultimately the combination of QWk, MSE and correlation coefficients is typically used to assess various aspects of a models performance. The multi-metric approach aims to ensure that AES models are evaluated not just on pure agreement but additionally on how well they can replicate the scoring patters of their human counterparts across a range of tasks.

2.6 Challenges in AES

Despite the significant progression in AES technology, several significant challenges remain. One major concern is bias and fairness within the scoring. AES models can learn and perpetuate biases present within training data, leading to an unfair scoring for certain groups of students. Recent research has scrutinized wether AES systems favour or penalize essays based on factors unrelated to the writings quality. Studies have found that a number of algorithms have exhibited different types of biases. For example, a comparative evaluation by (Litman, Baffour, and Crossley 2024). found that some AES models' scores varies systematically with students race and gender demographics, even when essay quality was consistent. These biases often occur when a model picks up on linguistic styles or topics more common for certain groups, or if the training essays are not properly balanced across demographics. This poses a risk where if these issues are ignores, AES systems may reinforce educational inequalities by giving lower scores to those essays written by non-native speakers or students from under-represented backgrounds. Addressing these biases is challenging due to the requirement of identifying the subtle correlations and ensuring the models predictions are based purely on the relevant indicators. Ongoing research has including techniques such as bias audits, debiaseing models via data augmentations, fairness constraints and developing new "fairness" metrics to evaluate across subgroups fo students. The insurance of AES fairness and lack of bias is an active concern within research as any real-world deployment of AES systems will require a high level of trust and accountability.

A second persistent challenge is cross-domain generalization (meaning a models ability to maintain its performance across new prompts and varying writing tasks outside of its training data). Traditionally, AES systems have been prompt-specific, an individual model developed trained for a singular essay prompt, trained on specific data. This approach is naturally limited in scalability as, requiring an individual model per task, and leads to issues in model robustness. In a practical ideal, AES systems would be capable of generalizing to a prompt of which it has never seen, without requiring further training. Research has shown that a model trained on one prompt is applied to an unseen prompt can lead to significant drops in performance, with some models only achieving 50% of their original QWK scores (Blanchard et al. 2013). The features and patterns learnt in training may not be applicable to a new prompt, due to the differences in vocabulary, structure and content required for different writing tasks. This problem of prompt dependency lead to a period of stagnation in AES research through the early 2000s, however with modern deep learning, a growing effort to tackle issues has resurfaced. Some approaches treat cross-prompt scoring as a domain adaption problem, where each prompt is a "domain" (Litman, Baffour, and Crossley 2024). Other methods inject prompt information into the model to aid in an AES systems adjusting its expectations for content (Blanchard et al. 2013). Some recent studies have reported progression with training prompt-agnostic models or multi-prompt training to improve generalization. The problem of a reliable cross-prompt AES remains unsolved - most models still perform best when fine-tuned on the target prompt data, to ensure a continent score criteria across problems is an open challenge.

Similarly there is the issue of a models ability to generalize across different proficiency levels and contexts of writing. Models are typically trained on a somewhat narrow band of essay types. Therefore when presented with an essay differing from their profile, their performance suffers. For instance, a system trained on high school essays may struggle when presented with elementary-level writing, or vice versa. Similarly as mentioned, scoring essays written by English language learners (ELLs/ESL students) is a challenge if the model has seen mostly native-level writing. Studies have noted that the widely-used ASAP corpus consists of essays by native English speakers in a timed exam context, which means models tuned to it may not handle non-native grammar errors or unusual expressions well. In contrast to this, real-world applications must deal with a full range of second-language writing issues. Another

important aspect is the robustness to varied essay lengths or formats, in some cases essay length can impact grade (Li and V. Ng 2024), and a model may unfairly rely on length if not carefully regularized. All these factors underline the need for AES systems of which are adaptable to different populations and conditions. researchers are exploring techniques such as adversarial training, or building ensemble/hybrid models which combine neural scoring and rule-based checks to handle out of scope cases. Nevertheless, ensuring an AES remains accurate and robust for any writer and any prompt including across grade levels, genres, and languages - is a long-term challenge. In multilingual settings, this challenge is even greater: an AES model trained for English cannot directly score, say, French essays without re-training on French data. Cross-lingual AES (using one model for multiple languages) is largely unexplored, due to language-specific differences in syntax and discourse. Progress has been made in creating resources for other languages, but transferring AES models across languages is an open research frontier

2.7 Gaps in Existing Research

Chapter 3

Data Description and Preprocessing

3.1 Dataset Overview

The primary dataset being used for this project is the **Automated Student Assessment Prize (ASAP)** dataset, released by the Hewlett Foundation in 2012 as part of a public kaggle competition with the goal of advancing AES research. The dataset consists of almost 13000 essays written by students in response to eight different prompts, each with a unique scoring rubric. The prompts vary in genre, include argumentative, narrative and source based tasks thus providing a diverse set of writing styles and topics. The essays are scored on a scale of 1-6, with multiple human raters providing scores for each essay. This dataset has become a standard benchmark for AES systems, allowing for the comparison of different models and approaches.

The Essay lengths range from approximately 150 to 600 words, with each response being graded by 2 independent human raters. The scoring scales differ across different prompts, with some using a 1-6 scale and others using a 0-3 scale for example **IS THIS TRUE**, depending on the complexity of the task.

Each record includes the following attributes:

- Unique essay ID
- Prompt ID
- Essay text
- Human scores
- Additional features such as the rater ID and scoring rubrics

The ASAP dataset has become a benchmark in the AES research community due to its relatively large size, accessibility, and variety of prompt types. It supports both prompt-specific modelling (where separate models are trained for each prompt) and prompt-agnostic modelling (where a single model is trained across all prompts). This flexibility makes it suitable for evaluating model generalisability — a key concern in the development of fair and scalable AES systems.

3.2 Dataset Attributes and Scoring Schemes

3.3 Dataset Attributes and Scoring Schemes

Each essay in the ASAP dataset is associated with one of eight distinct prompts, each corresponding to a unique writing task. These prompts vary in genre—ranging from narrative to persuasive and source-based

writing—and differ in complexity, structure, and rubric emphasis. As a result, each prompt is scored according to a specific holistic rubric designed to evaluate key aspects such as organisation, grammar, coherence, creativity, or use of supporting evidence.

The scoring schemes vary significantly across prompts. Some prompts use narrow ordinal scales (e.g., 0–3 or 0–6), while others have broader ranges such as 0–12 or 0–60. These scales are not directly comparable, as the same numeric score may imply different proficiency levels depending on the prompt. This discrepancy necessitates label normalisation strategies to enable cross-prompt training and fair evaluation, which are discussed in Section **??**.

Most essays in the dataset have been scored by two trained human raters. Where both scores are available, the average is typically used as the final training label. For essays with only one rater score, the decision to include or exclude them is made based on the modelling strategy, particularly in experiments focused on consistency or rater agreement.

Each essay record also contains additional metadata such as the prompt ID, essay set identifier, and number of raters. While these attributes are not used as direct model inputs, they are essential for organising the data, constructing prompt-specific training splits, and interpreting performance across different writing tasks.

3.4 Data Selection and Filtering

3.5 Tokenisation and Text Preprocessing

3.6 Label Normalisation

3.7 Limitations and Considerations

3.8 Additional datasets

Bibliography

Attali, Yigal and Jill Burstein (2006). "Automated essay scoring with e-rater® V.2". In: *The Journal of Technology, Learning and Assessment* 4.3. URL: https://ejournals.bc.edu/index.php/jtla/article/view/1647.

Beigman Klebanov, Beata, Jill Burstein, et al. (2021). "Limitations of automated essay scoring when evaluating narrative writing". In: *Assessing Writing* 47, p. 100511. DOI: 10.1016/j.asw.2020.100511. URL: https://doi.org/10.1016/j.asw.2020.100511.

Beigman Klebanov, Beata, Nitin Madnani, et al. (2021). "Limitations of automated essay scoring when evaluating narrative writing". In: *Assessing Writing* 47, p. 100511. DOI: 10.1016/j.asw.2020.100511.

Blanchard, Daniel et al. (2013). "TOEFL11: A corpus of non-native English". In: *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 45–52.

Blodgett, Su Lin et al. (2020). "Language (Technology) is Power: A Critical Survey of "Bias" in NLP". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476. URL: https://aclanthology.org/2020.acl-main.485/.

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805*. URL: https://arxiv.org/abs/1810.04805.

Dikli, Semire (2006). "An Overview of Automated Scoring of Essays". In: *The Journal of Technology, Learning and Assessment* 5.1. URL: https://ejournals.bc.edu/index.php/jtla/article/view/1646.

Landauer, Thomas K, Darrell Laham, and Peter W Foltz (2003). "Automated scoring and annotation of essays with the Intelligent Essay Assessor". In: *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 87–112. URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.3597&rep=rep1&type=pdf.

Li, Shengjie and Vincent Ng (2024). "Automated Essay Scoring: A Reflection on the State of the Art". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. URL: https://aclanthology.org/2024.emnlp-main.991.

Litman, Diane, Paul Baffour, and Scott Crossley (2024). "Fairness in Automated Essay Scoring: A Comparative Analysis of Shallow and Deep Learning Algorithms". In: *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. Association for Computational Linguistics, pp. 184–194. URL: https://aclanthology.org/2024.bea-1.18.

Liu, Yinhan et al. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv preprint arXiv:1907.11692*. URL: https://arxiv.org/abs/1907.11692.

Lundberg, Scott M. and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems* 30, pp. 4765–4774. URL: https://arxiv.org/abs/1705.07874.

Page, Ellis B. and Keith N. Petersen (2003). *The Computer Moves into Essay Grading: Updating the Ancient Test*. New York: Pearson.

Page, Ellis Batten (1966). “The Imminence of Grading Essays by Computer”. In: *Phi Delta Kappan* 47.5, pp. 238–243. URL: https://wac.colostate.edu/docs/books/usu/machine/chapter17.pdf.

Shermis, Mark and Jill Burstein (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge.

Shermis, Mark D. and Jill Burstein (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge. URL: https://www.routledge.com/Handbook-of-Automated-Essay-Evaluation-Current-Applications-and-New-Directions/Shermis-Burstein/p/book/9780415810968.

Taghipour, Kaveh and Hwee Tou Ng (2016a). “A Neural Approach to Automated Essay Scoring”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1882–1891. URL: https://aclanthology.org/D16-1193/.

— (2016b). “A Neural Approach to Automated Essay Scoring”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1882–1891.

Williamson, David M., Xiaoming Xi, and F. Julius Breyer (2012). “A Framework for Evaluation and Use of Automated Scoring”. In: *Educational Measurement: Issues and Practice* 31.1, pp. 2–13.

Yannakoudakis, Helen, Ted Briscoe, and Ben Medlock (2011). “A New Dataset and Method for Automatically Grading ESOL Texts”. In: *ACL*, pp. 180–189.