Proceedings of the 89th Annual International Meeting of the Psychometric Society, Prague, Czech Republic (2024).

Comparative Analysis of Machine Learning and LLM Approaches for Detecting ChatGPT-Written Essays Under Revision Conditions

Haowei Hua [*†] and Jiayu Yao [‡]

†Princeton University, Princeton, 08544, New Jersey, USA
‡Anhui Polytechnic University, Wuhu, 241000, Anhui, China
*Corresponding author. Email: jack.hua@princeton.edu

Abstract

ChatGPT, a powerful generative artificial intelligence (AI), can play a significant role in enhancing K-12 education by offering support with various tasks, such as answering questions, solving math problems, and generating content like essays, code, and presentation slides. While it represents an invaluable resource for learning, concerns have arisen regarding its potential misuse by students when completing school assignments. Current commercial detectors, like Grammarly and GPTZero, are designed to identify general text generated by AI, lacking specificity for high-stakes assessments. This study addresses the challenge of detecting the potential use of ChatGPT for academic cheating in high-stakes assessments. Classical machine learning methods, including logistic regression, naïve Bayes, and decision trees, were employed to distinguish between essays generated by ChatGPT and those authored by students. Additionally, pre-trained language models like Roberta and BERT were compared against traditional machine learning approaches. The analysis focused on prompt 1 from the Kaggle Automated Student Assessment Prize (ASAP) competition. To evaluate the effectiveness of the detection methods, four approaches were applied to revise ChatGPT-generated essays: Grammarly Premium, revisions by eighth-grade students, revisions by ninth-grade or above students, and further modifications by ChatGPT with additional prompting to humanize and naturalize the essays by introducing grammatical mistakes. For detecting unmodified ChatGPT essays, Electra, a pre-trained language model, demonstrated a high quadratic weighted Kappa (QWK) score of 97%, while support vector machine (SVM) outperformed the large language models with a remarkable QWK score of 99%. The modification methods significantly influence the detection rate crossing various models. This research addresses concerns about academic integrity in high-stakes assessments involving generative AI technologies.

Keywords: classification, prediction, statistical and machine learning, Generative AI

1. Introduction

The advancement of generative artificial intelligence (AI) models, such as OpenAI's ChatGPT, has significantly impacted various fields, including education, by enabling the rapid production of high-quality written content (Strzelecki, 2024). While these models offer new opportunities for academic assistance, they also present challenges in maintaining academic integrity, as students increasingly rely on AI-generated text for essay assignments (Borenstein & Howard, 2021). Concerns about AI-assisted plagiarism and the authenticity of student writing have prompted a growing demand for reliable detection mechanisms (Pudasaini et al., 2024). As generative AI evolves, educators and researchers face the pressing challenge of distinguishing between human-authored and machine-generated

content, particularly as AI-generated text becomes more coherent and stylistically indistinguishable from human writing (Alasadi & Baiz, 2023).

A growing number of machine learning-based detectors have been developed to differentiate between human-written and AI-generated text, including tools such as GPTZero and large language model (LLM)-based classifiers (Elkhatat et al., 2023). These detectors flag AI-generated content, primarily relying on linguistic features, perplexity scores, and other text-based attributes. However, the effectiveness of these models is limited by the continuous advancements in generative AIs, which can produce increasingly human-like text, often bypassing existing detection mechanisms (Weber-Wulff et al., 2023). Studies have also shown that adversarial techniques—such as minor paraphrasing, deliberate grammatical errors, or structural modifications—can significantly reduce the accuracy of AI detectors, making detection a constantly evolving challenge (Zhou et al., 2024).

The present study explored the efficacy of feature-based machine learning models and LLMs in detecting AI-generated essays in the context of writing assessments. This research focused primarily on comparing the performance of traditional machine learning classifiers, such as logistic regression (LaValley, 2008), SVMs (Hearst et al., 1998), and random forests (Breiman, 2001), against more advanced deep learning-based language models like BERT (Devlin, 2018), ELECTRA (Clark, 2020), and RoBERTa (Liu, 2019). The goal was to identify the strengths and weaknesses of different approaches in accurately distinguishing between human-authored and AI-generated texts. Given that generative AIs continue to improve in coherence and contextual understanding, this study also examined whether classical approaches are still relevant in the context of AI text detection, especially when AI-generated text has been manually or algorithmically revised (Akram, 2023).

The Kaggle Automated Student Assessment Prize (ASAP) dataset (Hamner et al., 2012) was utilized to construct a robust evaluation framework supplemented with essays generated using ChatGPT-3.5 and ChatGPT-4.0. In total, 1,500 AI-written essays were generated based on predefined prompts and were used to evaluate the performance of detection models across various scenarios, including essays modified using Grammarly, revised by students, and rewritten by ChatGPT. These modifications were introduced to assess the impact of text alterations on detection accuracy and robustness (Brown et al., 2020). Prior research has indicated that simple post-processing techniques can make AI-generated text significantly harder to detect, emphasizing the need for more sophisticated detection strategies (Jawahar et al., 2019).

A key component of our study was evaluating detection models using multiple performance metrics, including accuracy, precision, recall, F1 score, and Quadratic Weighted Kappa (QWK). QWK, in particular, provides a nuanced assessment of model agreement with human raters and is essential for understanding the reliability of AI-driven detectors in practical applications. The QWK score serves as a basic benchmark to evaluate the performance of AI-driven detectors across different texts to maintain generalizability and consistency (Cohen, 1968).

This study contributes to addressing the growing challenge of AI-generated text detection by offering insights into the limitations of current detection models and highlighting the challenges posed by evolving AI capabilities. By systematically analyzing different detection approaches under various conditions, this research aims to inform the development of more resilient AI detection methodologies in educational settings.

2. Methods

2.1 Dataset

This study employed a dataset comprising 3,285 essays, including human-written and AI-generated texts. The dataset was designed to assess machine learning and deep learning models' effectiveness in distinguishing between human-authored and AI-generated essays. Additionally, a subset of AI-generated essays underwent various modification techniques to evaluate their impact on detection accuracy.

2.1.1 Human-Written Essays

A total of 1,785 essays were sourced from prompt 1 of the publicly available *Kaggle ASAP* dataset. These essays were written by 8th grade students in response to a standardized persuasive writing prompt, which required them to articulate and defend their opinions regarding the societal impact of computers. The full prompt is provided below:

> 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.

2.1.2 Machine-Generated Essays

To analyze the detectability of AI-generated text, this study incorporated a total of 1,500 machine-generated essays. These texts were produced using ChatGPT-3.5 and ChatGPT-4.0 under different configurations. The first subset included 800 essays generated by ChatGPT-3.5 with varied word counts (300, 500, and 600 words). This number was selected to ensure sufficient representation across multiple word count ranges and to align with the typical lengths of human-written essays in the dataset. Another subset comprised 350 essays produced using a scoring-guided generation approach, in which ChatGPT-3.5 was instructed to create essays aligning with predefined scores of 8 or 12. The decision to generate 350 essays per score point was intended to maintain a balanced distribution of moderate and high-scoring responses, facilitating comparative analysis against human-written essays scored similarly. Additionally, 200 essays were generated using ChatGPT-4.0, all corresponding to a score of 12. This subset was included to assess the impact of the newer model's enhanced language capabilities and to provide high-quality essays for the test dataset. The allocation of 200 essays was based on practical constraints, as ChatGPT-4.0 was released during the study period and had resource limitations that restricted the volume of generated content. Consequently, the total number of machine-generated essays was set at 1,500, slightly lower than the 1,785 human-written essays, to account for resource limitations, distribution constraints, and the need to maintain a balanced dataset for effective comparative analysis.

2.1.3 Modification Methods

A subset of the AI-generated essays was systematically modified to examine the impact of multiple revision strategies on detection accuracy. These modifications included both human and AI-driven interventions. Specifically, essays were revised using Grammarly Premium, which introduced sentence structure, word selection, and paraphrasing modifications.

Additionally, 8th-grade and 9th-grade students from a local high school in Indiana, USA, participated in the revision process. The participants included both native and non-native English speakers who had successfully passed an English proficiency exam, ensuring that all were adequately capable of producing coherent written content. Due to time constraints, 10 students from each grade level participated, resulting in a total of 20 student revisers. The 8th-grade students were instructed to revise AI-generated essays from the perspective of a student completing a homework assignment, maintaining a conversational and informal tone. In contrast, the 9th-grade and older students were explicitly directed to refine the essays further, enhancing overall readability while retaining a natural

student-like voice. This dual approach simulated varying levels of academic revision commonly encountered in educational settings.

Furthermore, ChatGPT was prompted to revise essays by deliberately incorporating grammatical imperfections and stylistic elements characteristic of human writing. The specific prompt used for this revision was:

```
Please humanize this AI-generated essay and include grammatical mistakes
or other linguistic features that make it appear more natural, simulating
the writing style of an eighth grade student
```

In total, 200 essays were modified using these techniques, forming five distinct evaluation datasets, including the GPT-4.0 generated high score essays and four additional revision strategies.

2.2 Models

To explore the effectiveness of automated approaches in short-answer scoring, a range of traditional machine learning algorithms and modern transformer-based deep learning models was employed. The classical machine learning models used include Logistic Regression, Support Vector Machines (SVM), Naïve Bayes, and Random Forest. Logistic Regression serves as a strong linear baseline due to its simplicity and interpretability. SVM is known for its effectiveness in high-dimensional spaces, making it suitable for sparse text data. Naïve Bayes, while based on strong independence assumptions, is computationally efficient and often surprisingly competitive in text classification tasks. Random Forest, an ensemble of decision trees, introduces non-linear decision boundaries and robustness to overfitting. In addition to these, ensemble-based learners such as the Passive-Aggressive Classifier (PAC), XGBoost, and LightGBM (LGBM) were utilized as well. PAC is an online learning algorithm particularly well-suited for large-scale, high-dimensional datasets. XGBoost and LGBM are gradient-boosted decision tree frameworks that have demonstrated success across a wide range of classification tasks, offering strong performance through feature importance and boosting techniques. To harness recent advances in natural language understanding, transformer-based models: BERT, ELECTRA, and RoBERTa were investigated. These models are pre-trained on large-scale corpora and fine-tuned on our classification task. BERT (Bidirectional Encoder Representations from Transformers) introduces bidirectional context modeling, while RoBERTa is a robustly optimized variant of BERT with improved training dynamics. ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) adopts a replaced token detection objective that often yields better sample efficiency and performance with fewer computational resources.

2.3 Preprocessing

The preprocessing stage of this project was meticulously designed to ensure high-quality data for model training and evaluation. A detailed and academic description of the process is presented as follows: Text normalization aimed to unify the text format and reduce noise. All characters were converted to lowercase to eliminate case-related inconsistencies. Punctuation, special characters, and numerals were removed as they often do not contribute semantically to the content, especially in essay-focused analyses. Tokenization used advanced natural language processing (NLP) libraries like NLTK or spaCy. Text was split into tokens (words or phrases). This facilitated granular analysis and feature extraction, forming the basis for subsequent linguistic and syntactic analyses. Stopword removal reduced dimensionality and computational load while focusing on content-bearing words. The predefined list of stop words (e.g., "the," "is, "and") was compiled, and these words were removed. Lemmatization was conducted which words were normalized to their base forms. Feature Extraction allows following models to better process the data and classify. For traditional machine learning models, term frequency-inverse document frequency (TF-IDF) was employed to quantify word importance, transforming text into numerical features that reflect term frequency and inverse

document frequency. Pre-trained embeddings like BERT and ELECTRA were utilized for deep learning models, offering rich semantic and contextual information.

2.4 Training

The training process for the automated scoring models involved two primary approaches: traditional machine learning models and language models such as BERT. Initially, the dataset without revision essays was divided into training, validation, and test sets using an 80-10-10 ratio to maintain data consistency across different scoring levels. In the second stage of the study, the revised essays were added into the whole dataset under same split procedure evaluating the performance of pretrained automated scoring system. The traditional machine learning models, including LightGBM and XGBoost, were trained using a range of handcrafted features such as word count, lexical diversity, and syntactic complexity. Hyperparameter tuning was performed using cross-validation to optimize model performance and prevent overfitting.

For the LM models, text inputs were tokenized and embedded using pre-trained BERT representations, capturing both semantic and contextual nuances in the essays. The BERT-based models were then fine-tuned using the training set, with learning rates and batch sizes adjusted based on validation set performance. The test set was reserved for the final evaluation of both traditional ML and LM models, assessing predictive accuracy and consistency using metrics. This training structure allowed for a comprehensive comparison of traditional feature-based models and transformer-based language models in automated scoring tasks.

2.5 Evaluation

This study employed a comprehensive set of evaluation metrics to assess the performance of various detection models in identifying AI-generated essays. These metrics allowed us to analyze the effectiveness and reliability of both classical machine learning and large language models under different conditions. Precision, Recall, F1, and Accuracy scores are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

Precision represents the model's ability to correctly identify positive instances (AI-generated essays) out of all instances predicted as positive. Precision is crucial for academic integrity assessments where minimizing false positives is essential. Models with high precision, such as SVM and Electra, are particularly effective in ensuring that essays predicted as AI-generated are indeed AI-generated. This reduces the risk of falsely accusing students of cheating.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

Recall (sensitivity) represents the model's capacity to detect all actual positive instances. Recall is important for ensuring that as many AI-generated essays are detected as possible. In our study, models like Electra and SVM exhibited high recall values, indicating their ability to identify a large proportion of AI-generated essays. This is crucial for maintaining academic integrity, as it minimizes the number of undetected AI-generated essays.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

The F1 score represents the harmonic mean of precision and recall, providing a balanced measure of these two metrics. It offers a more nuanced view of model performance by considering both false positives and false negatives. Models with high F1 scores, such as Electra and SVM, demonstrate a

good balance between precision and recall, making them robust choices for detecting AI-generated essays.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Accuracy represents the proportion of correctly classified instances out of the total instances. Accuracy provides a general overview of model performance. In our study, models like SVM and Electra demonstrated high accuracy scores, indicating their strong overall performance in classifying essays as human-written or AI-generated. However, accuracy alone is not sufficient, especially when dealing with imbalanced datasets or when the cost of false positives and false negatives differs significantly.

The QWK score is defined as follows:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}, \tag{5}$$

where $w_{i,j}$ denotes the quadratic weights, $O_{i,j}$ is the observed frequency, and $E_{i,j}$ is the expected frequency.

The QWK is a metric that measures the agreement between two raters, accounting for the magnitude of disagreement. QWK evaluates the consistency between the detection models and human raters. In our study, QWK scores were particularly useful in understanding the practical reliability of AI-driven detectors. Models like Electra and SVM showed high QWK scores, indicating strong agreement with human raters, which is essential for real-world applications.

3. Results

This study comprehensively evaluated the performance of various machine learning models and LLMs in detecting ChatGPT-generated essays, particularly in scenarios involving revised or modified texts. The findings, assessed by way of accuracy, precision, recall, F1 score, and QWK metrics, highlight the strengths and vulnerabilities of detection methodologies in academic integrity enforcement.

As presented in Table 1, generally, BERT, Roberta, Robert-A perform better than traditional machine learning models when no modification method is applied to the original context. The SVM and ELECTRA models demonstrated exceptional detection capabilities for unmodified ChatGPT-generated essays. The SVM model achieved a QWK score of 0.934, while the ELECTRA model attained the highest QWK score of 0.964, reflecting near-perfect agreement with human evaluators. These results suggest that both models are highly effective at detecting unaltered AI-generated content.

Additionally, both models excelled in other classification metrics. SVM achieved an accuracy of 99.3%, making it the most accurate model among those evaluated. Its precision (99.5%) and recall (99.1%) indicate that it not only identifies AI-generated text correctly but also minimizes false positives and negatives. Similarly, the ELECTRA model achieved an accuracy of 98.2%, with precision (98.3%) and recall (98.1%) scores that reinforce its reliability.

These results highlight the robustness of SVM and ELECTRA in identifying AI-generated content, particularly in educational settings where maintaining academic integrity is critical. Compared to other models, such as naïve Bayes (QWK = 0.837) and XGBoost (QWK = 0.816), SVM and ELECTRA show superior consistency with human evaluations. While traditional machine learning models like logistic regression and random forest performed well (QWK = 0.872 and 0.867, respectively), deep learning-based architectures such as ELECTRA and BERT (QWK = 0.938) offered a more refined understanding of AI-generated text patterns.

However, the effectiveness of detection models significantly diminished when essays underwent revisions. For Grammarly Premium edits, post-revision essays showed a notable decline in detection

accuracy. For instance, SVM's QWK score dropped to 89%, and Electra's to 85%, likely due to Grammarly's optimization of syntax and vocabulary, which obscured subtle AI-generated patterns. Under revisions by eighth-grade students Human revisions, even by younger students, further reduced model performance (SVM: QWK 82%; Electra: QWK 78%), suggesting that minor stylistic or structural changes introduced by humans can disrupt detection algorithms. Under revisions by advanced students (Ninth Grade and Above), more sophisticated revisions exacerbated the decline (SVM: QWK 76%; Electra: QWK 72%), highlighting the challenge of distinguishing AI-generated content refined by human intervention. By ChatGPT-Based rewrites, the most significant performance drop occurred when ChatGPT reprocessed its own essays to introduce grammatical errors and human-like phrasing. SVM and Electra QWK scores plummeted to 68% and 63%, respectively, illustrating how iterative AI modifications can effectively bypass detection systems.

4. Discussion

This study's findings reveal the complexities of detecting AI-generated text, particularly within academic writing assessments. As generative AI technologies—especially chatbots—continue to evolve, the challenge of distinguishing between human and machine-generated essays increases. Our analysis reveals that while LLMs such as BERT, Electra, and RoBERTa outperform traditional machine learning models in AI-generated text detection, the effectiveness of these models is significantly impacted by modification techniques such as human revision, Grammarly adjustments, and ChatGPT rewriting.

Our results align with the findings of previous research that highlight the superior performance of transformer-based models over classical approaches. Previous studies (Elkhatat et al., 2023; Weber-Wulff et al., 2023) have confirmed that tools leveraging deep learning methodologies demonstrate higher precision and recall rates when distinguishing AI-generated content from human-written text. These models, trained on vast corpora, can identify intricate textual features indicative of machine-generated outputs. However, research (Zhang et al., 2024) has elucidated an emerging phenomenon where advanced AI systems generate increasingly human-like text, reducing the efficacy of current detection mechanisms. Our study corroborates this trend, showing that paraphrasing, tools-based modifications, and human revisions significantly lower detection accuracy.

Modification techniques and the increasing convenience of utilizing online paraphrasing tools for students introduce another significant challenge in the field of AI text detection. The inclusion of Grammarly revisions, manual edits by students, and AI-generated rewrites obfuscate original AI markers, diminishing the performance of detection models. This corroborates the findings of (Zhang et al., 2024), who explored how anti-detection strategies influence classifier robustness. The QWK metric used in our study further demonstrates a substantial decline in agreement between raters when modifications are applied, emphasizing the need for adaptive detection methodologies that can cope with iterative text refinements. The potential reasoning behind is the paraphrasing tools outperform the embedding utilized during the model training session. Once the texts are revised to certain extent, the pretrained models failed to identify the synonyme as features for machine generated texts.

The decreasing effectiveness of AI detectors due to modification raises concerns about the reliability of existing academic integrity enforcement strategies. With tools such as Chat-GPT, Gemini, Claude, and other AI chatbots, students can subtly alter AI-generated essays, evading detection systems currently employed by educational institutions. The research by (Elkhatat et al., 2023) suggests that while detection tools serve as a deterrent, a more holistic approach that integrates pedagogical interventions is necessary. Educators must incorporate AI literacy training, ensuring that students understand the ethical implications of AI-assisted writing while promoting authentic authorship.

Future research should explore the integration of ensemble learning methods, combining multiple

detection strategies to enhance robustness against modifications. Additionally, developing forensic linguistic techniques focusing on coherence, argument structure, and syntactic variation may provide alternative means of distinguishing AI-assisted writing. Zhang et al. (2024) (Zhang et al., 2024) point toward the necessity for a dynamic, continuously evolving detection framework that adapts to advancements in generative AI capabilities. Improving the word embedding to better represent the semantic features behind words may contribute to the performance of detection mechanism. Finally, comparative studies between proprietary detectors, such as GPTZero, and open-source models will offer valuable insights into their respective strengths and limitations in real-world applications.

5. Tables

Table 1. Performance comparison of different models

Model	Precision	Recall	F1-score	Accuracy	QWK
Logistic Regression	0.988	0.990	0.989	0.989	0.872
SVM	0.995	0.991	0.993	0.993	0.934
Naive Bayes	0.936	0.939	0.938	0.938	0.837
Random Forest	0.985	0.983	0.984	0.984	0.867
PAC	0.979	0.979	0.979	0.979	0.834
XGBoost	0.973	0.967	0.970	0.970	0.816
LGBM	0.976	0.966	0.971	0.970	0.823
Bert	0.959	0.960	0.960	0.960	0.938
ELECTRA	0.983	0.981	0.982	0.982	0.964
Robert-A	0.932	0.945	0.939	0.939	0.867

Table 2. Performance of SVM model on different types of modified AI-generated essays.

Model	Precision	Recall	F1-score	Accuracy	QWK
8th grade modified	0.985	0.970	0.977	0.953	0.986
9th grade modified	0.881	0.826	0.853	0.828	0.794
Grammarly Premium	0.789	0.777	0.783	0.753	0.736
Chat-GPT remodified	0.892	0.836	0.863	0.828	0.813
GPT 4.0	0.923	0.891	0.906	0.873	0.854

6. Conclusion

The rapid evolution of generative AI technologies presents both opportunities and challenges for academic integrity. While AI-generated text detection has made significant advancements, this study illuminates the challenges and weaknesses of existing models when faced with modified AI-generated content. Our findings suggest that an exclusive reliance on automated detection tools is insufficient for maintaining academic integrity. Instead, a multifaceted approach integrating advanced machine learning techniques with educational awareness will be critical in addressing this growing challenge.

One key aspect of this research was to investigate the role of text modifications in diminishing detection accuracy. Modifications such as Grammarly-based revisions, human refinements, and

Table 3. Performance of ELECTRA model on different types of modified AI-generated essays.

Model	Precision	Recall	F1-score	Accuracy	QWK
8th grade modified	0.938	0.910	0.924	0.900	0.965
9th grade modified	0.830	0.789	0.809	0.785	0.754
Grammarly Premium	0.747	0.775	0.761	0.738	0.714
Chat-GPT remodified	0.840	0.803	0.821	0.788	0.763
GPT 4.0	0.887	0.878	0.882	0.850	0.834

reprocessing through generative AI models obscure detectable AI markers, reducing the effectiveness of current detection frameworks. This finding is consistent with previous studies suggesting that AI-assisted writing tools are becoming increasingly adept at mimicking human stylistic patterns, making it more difficult to distinguish between AI-generated and human-authored content.

Furthermore, our analysis highlights that while transformer-based models such as BERT, Electra, and RoBERTa perform well in detecting AI-generated content, they are not immune to the constraints imposed by evolving generative AI strategies. The decline in QWK scores after the application of modifications reinforces the need for adaptive detection systems that incorporate linguistic analysis alongside machine learning methodologies.

Future research should focus on refining detection algorithms to account for the evolving sophistication of generative AI. Hybrid detection approaches that leverage forensic linguistic analysis, contextual examination, and deep learning-based methodologies may offer a more robust framework for detecting AI-assisted writing. Additionally, integrating detection models within educational policies and raising awareness about ethical AI usage among students will be crucial in mitigating the misuse of AI in academic settings.

Acknowledgement

I would like to express my sincere gratitude to Professor Hong Jiao at the University of Maryland for her invaluable support and guidance throughout this study. Her insightful instructions, constructive feedback, and encouragement have been instrumental in shaping my understanding and approach to this research.

Competing Interests None

References

Akram, A. (2023). An empirical study of ai generated text detection tools. *arXiv preprint arXiv:2310.01423*.

Alasadi, E. A., & Baiz, C. R. (2023). Generative ai in education and research: Opportunities, concerns, and solutions. *Journal of Chemical Education*, *100*(8), 2965–2971.

Borenstein, J., & Howard, A. (2021). Emerging challenges in ai and the need for ai ethics education. *AI and Ethics*, *1*, 61–65.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Clark, K. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, *70*(4), 213.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of ai content detection tools in differentiating between human and ai-generated text. *International Journal for Educational Integrity*, *19*(1), 17.

Hamner, B., Morgan, J., lynnvandev, Shermis, M., & Ark, T. V. (2012). The hewlett foundation: Automated essay scoring. https://kaggle.com/competitions/asap-aes

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18–28.
Jawahar, G., Sagot, B., & Seddah, D. (2019). What does bert learn about the structure of language? *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.
LaValley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395–2399.
Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, *364*.
Pudasaini, S., Miralles-Pechuán, L., Lillis, D., & Llorens Salvador, M. (2024). Survey on ai-generated plagiarism detection: The impact of large language models on academic integrity. *Journal of Academic Ethics*, 1–34.
Strzelecki, A. (2024). Chatgpt in higher education: Investigating bachelor and master students' expectations towards ai tool. *Education and Information Technologies*, 1–25.
Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, *19*(1), 26.
Zhang, Y., Ma, Y., Liu, J., Liu, X., Wang, X., & Lu, W. (2024). Detection vs. anti-detection: Is text generated by ai detectable? *International Conference on Information*, 209–222.
Zhou, Y., He, B., & Sun, L. (2024). Humanizing machine-generated content: Evading ai-text detection through adversarial attack. *arXiv preprint arXiv:2404.01907*.