

# Evaluating Machine Learning Models and Large Language Models for Detecting ChatGPT-Generated Essays in Writing Assessments: The Impact of Revision Techniques on Detection Performance"

Haowei Hua Author<sup>\*†</sup> and Jiayu Yaor<sup>‡</sup>

<sup>†</sup>The Culver Academies, Organization, City, Pincode, State, Country

<sup>‡</sup>Anhui Polytechnic University, Organization, City, Pincode, State, Country

\*Corresponding author. Email: haowei.hua@culver.org

## Abstract

ChatGPT, a powerful generative AI, holds significant role for enhancing K-12 education by offering support on various tasks such as answering questions, solving math problems, and generating content like essays, code, and presentation slides. While it presents an invaluable resource for learning, concerns arise regarding its potential misuse by students for completing school assignments. Current commercial detectors, like Grammarly and GPTZero, are designed for general text generated by AI, lacking specificity for high-stakes assessments. This study addresses the challenge of detecting potential academic cheating using ChatGPT in high-stakes assessments. Classical machine learning methods, including logistic regression, naive Bayes, and decision trees, were employed to identify distinctions between essays generated by ChatGPT and those authored by students. Additionally, pretrained language models such as Roberta and BERT were compared against traditional machine learning approaches. The analysis focused on the prompt 1 from the ASAP Kaggle 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. In the detection of unmodified ChatGPT essays, Electra, a pretrained 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%. This research contributes to addressing concerns around academic integrity in high-stakes assessments involving generative AI technologies.

**Keywords:** classification, prediction, statistical and machine Learning, ChatGPT

## 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 [strzelecki2024chatgpt](#). 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 [borenstein2021emerging](#). Concerns about AI-assisted plagiarism and the authenticity of student writing have prompted a growing demand for reliable detection mechanisms [pudasaini2024survey](#). As generative AI continues to

evolve, 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 [alasadi2023generative](#).

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 [selkhatat2023evaluating](#). These detectors primarily rely on linguistic features, perplexity scores, and other text-based attributes to flag AI-generated content. 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 [weber2023testing](#). Furthermore, studies have 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 [zhou2024humanizing](#).

The present study explores the efficacy of both feature based machine learning models and large language models in detecting AI-generated essays in the context of writing assessments. The research is particularly focused on comparing the performance of traditional machine learning classifiers, such as logistic regression [lavalley2008logistic](#), support vector machines (SVM) [hearst1998support](#), and random forests [breiman2001random](#), against more advanced deep learning-based language models like BERT [devlin2018bert](#), ELECTRA [clark2020electra](#), and RoBERTa [liu2019roberta](#). The goal is 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 examines whether classical approaches still hold relevance in AI text detection, especially under conditions where AI-generated text has been manually or algorithmically revised [akram2023empirical](#).

To construct a robust evaluation framework, the Kaggle Automated Student Assessment Prize (ASAP) dataset is utilized [hamner2012asap](#), supplemented with AI-generated essays using ChatGPT-3.5 and ChatGPT-4.0. In total, 1,500 AI-written essays based on predefined prompts are generated and evaluated the performance of detection models across various scenarios, including essays modified using Grammarly, student revisions, and ChatGPT-based rewrites. These modifications were introduced to assess the impact of text alterations on detection accuracy and robustness [brown2020language](#). 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 [jawahar2019does](#).

A key component of our study is the evaluation of 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. In order to maintain the generability and consistency of the AI-driven detectors, QWK score serves as a basic benchmark to evaluate its performance crossing different texts [cohen1968weighted](#).

This study contributes to the growing challenge on 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, the research aims to inform the development of more resilient AI detection methodologies in educational settings.

## 2. Methods

### 2.1 Dataset

### 2.1.1 Overview

This study employs a dataset comprising **3,285 essays**, which include both human-written and AI-generated texts. The dataset is designed to assess the effectiveness of machine learning and deep learning models in distinguishing between human-authored and AI-generated essays. Additionally, a subset of AI-generated essays underwent various modification techniques to examine their impact on detection accuracy.

## 2.2 Human-Written Essays

A total of **1,785 essays** were sourced from the publicly available *Kaggle Automated Student Assessment Prize (ASAP)* dataset, prompt 1. These essays were written by 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.

## 2.3 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). Another subset consisted of **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. Additionally, 200 essays were generated using ChatGPT-4.0, all corresponding to a score of 12.

To ensure robust evaluation, a separate dataset was allocated exclusively for testing purposes, allowing for a more comprehensive assessment of model performance beyond the training set.

## 2.4 Modification Methods

A subset of the AI-generated essays was systematically modified using multiple revision strategies to examine their impact on detection accuracy. These modifications included both human and AI-driven interventions. Specifically, essays were revised using Grammarly Premium, which introduced modifications related to sentence structure, word selection, and paraphrasing. Additionally, 8th-grade students were tasked with revising AI-generated essays while preserving coherence, whereas 9th-grade and older students were provided explicit instructions to refine the texts to enhance their natural readability. 8th grade students mimics the life situation where the prompt designated grade group (grade 8) reviews the machine generated prompt and revise while the 9th grade or elder symbolizes the situtaion the students seeks for additional assistance to further revise the result. Furthermore, ChatGPT was prompted to revise essays by deliberately incorporating grammatical imperfections and stylistic elements characteristic of human writing. In total, 200 essays were modified using these techniques, forming five distinct evaluation datasets, including the GPT-4.0 modified essays and four additional revision strategies.

This dataset design facilitates a comprehensive investigation into the detectability of AI-generated essays, the effectiveness of various machine learning and deep learning models, and the extent to which modification techniques impact detection accuracy. The inclusion of both unaltered and modified AI-generated texts allows for an in-depth exploration of the evolving challenges associated with distinguishing human-authored writing from AI-assisted compositions.

## 2.5 Preprocessing

In this project, the preprocessing stage was meticulously designed to ensure high – quality data for model training and evaluation. Here’s a detailed and academic description of the process:

**Text Normalization:** This step 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 don’t contribute semantically to the content, especially in essay – focused analyses.

**Tokenization:** Using advanced 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.

**Stop Word Removal:** A predefined list of stop words (e.g., "the", "is", "and") was compiled and these words were removed. This step reduced dimensionality and computational load while focusing on content – bearing words. **Lemmatization and Stemming:** Words were normalized to their base forms. Lemmatization, which requires morphological analysis, was preferred over stemming in most cases to ensure morphological correctness, though stemming was used when speed was prioritized.

**Feature Extraction:** For traditional machine learning models, TF-IDF was employed to quantify word importance, transforming text into numerical features that reflect term frequency and inverse document frequency. For deep learning models, pretrained embeddings like BERT and ELECTRA were utilized, offering rich semantic and contextual information.

**Handling Paraphrased Text:** A comprehensive synonym list was developed through manual analysis of paraphrased essays. This list was applied during preprocessing to standardize synonyms, countering the effects of paraphrasing tools and ensuring text consistency.

**Data Splitting:** The dataset was divided into training, validation, and test sets following an 80:10:10 ratio. Stratified sampling was used to maintain class distribution, especially crucial given the binary classification nature of the task.

**Class Imbalance Handling:** Techniques like SMOTE or ADASYN were considered but not ultimately applied due to the relatively balanced nature of the dataset. However, the project maintained flexibility to incorporate such methods if imbalance issues arose during model training.

Each step was carefully implemented in Python, with libraries like pandas for data manipulation and scikit – learn for feature processing. This rigorous preprocessing pipeline laid the foundation for robust model training and evaluation, ensuring that the data fed into the models was clean, standardized, and optimally formatted for detecting AI – generated essays.

## 2.6 Evaluation

In this study, we 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 models and large language models under different conditions. Precision, Recall, F1, and Accuracy scores are defined as follows:

## 3. Equations

Sample equations. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Lorem ipsum dolor sit amet, consectetur<sup>1</sup> adipiscing

elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Below is a detailed analysis using these evaluation metrics: Accuracy: 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 either 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.

Precision: the model's ability to correctly identify positive instances (AI-generated essays) out of all instances predicted as positive. Precision is crucial in 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.

Recall (sensitivity): the model's capacity to detect all actual positive instances. Recall is important for ensuring that as many AI-generated essays as possible are detected. In our study, models like Electra and SVM showed 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.

The F1 score: The harmonic mean of precision and recall, providing a balanced measure of these two metrics. The F1 score 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.

The QWK score is defined as follows:

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

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

Quadratic Weighted Kappa (QWK): 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.

#### 4. Results

This study comprehensively evaluates the performance of various machine learning models and large language models (LLMs) in detecting ChatGPT-generated essays, particularly under scenarios

involving revised or modified texts. The findings, assessed through metrics including accuracy, precision, recall, F1 score, and Quadratic Weighted Kappa (QWK), highlight both the strengths and vulnerabilities of detection methodologies in academic integrity enforcement.

For unmodified ChatGPT-generated essays, as presented in Table 1, the Support Vector Machine (SVM) and ELECTRA models demonstrated exceptional detection capabilities. 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.

Moreover, 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 false negatives. Similarly, the ELECTRA model achieved an accuracy of 98.2%, with precision (98.3%) and recall (98.1%), reinforcing 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. Four revision strategies were tested:

**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.

**Revisions by Eighth-Grade Students:** Human revisions, even by younger students, reduced model performance further (SVM: QWK 82%; Electra: QWK 78%). This suggests that minor stylistic or structural changes introduced by humans can disrupt detection algorithms.

**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.

**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.

## 5. Discussion

The findings of the study indicate the complexities in 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 large language models (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 Chat-GPT rewriting.

Our results align with previous research highlighting the superior performance of transformer-based models over classical approaches. Previous studies [elkhatat2023evaluating](#) and [weber2023testing](#) confirm 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 **zhang2024detection** discuss 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 for students to utilize online paraphrasing tools introduce another significant challenge in 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 aligns with findings by **zhang2024detection**, who examined how anti-detection strategies influence classifier robustness. The Quadratic Weighted Kappa (QWK) metric 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 against the iterative text refinements.

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 **elkhatat2023evaluating** suggests that while detection tools serve as a deterrent, a more holistic approach integrating pedagogical interventions is necessary. Educators must incorporate AI literacy training, ensuring 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, the development of forensic linguistic techniques focusing on coherence, argument structure, and syntactic variation may provide alternative means of distinguishing AI-assisted writing. The work **zhang2024detection** points toward the necessity of a dynamic, continuously evolving detection framework that adapts to advancements in generative AI capabilities. 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.

6. 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 Model	0.959	0.960	0.960	0.960	0.938
ELECTRA Model	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

**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

7. 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, the study indicate 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 is to investigate the role that text modifications play in diminishing detection accuracy. Modifications such as Grammarly-based revisions, human refinements, and reprocessing through generative AI models obscure detectable AI markers, thereby reducing the effectiveness of current detection frameworks. This finding is consistent with previous studies that indicate 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 limitations posed by evolving generative AI strategies. The decline in Quadratic Weighted Kappa (QWK) scores when modifications were applied 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. Ultimately, while AI-generated text detection has made substantial progress, continuous innovation and interdisciplinary collaboration will be necessary to ensure fairness and originality in academic assessments. By adopting a proactive and adaptive approach, institutions can better safeguard academic integrity in the face of advancing AI technologies.



## **Acknowledgement**

Insert the Acknowledgment text here.

**Funding Statement** This research was supported by grants from the <funder-name> <doi> (<award ID>); <funder-name> <doi> (<award ID>).

**Competing Interests** A statement about any financial, professional, contractual or personal relationships or situations that could be perceived to impact the presentation of the work — or ‘None’ if none exist.

## **Notes**

1 Another footnote/endnote

## **Appendix 1. Example Appendix Section**

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.