Testing the Limits: Evaluating AI Detectors' Accuracy and the Impact of Obfuscation Techniques on AI-Generated Text

Alfira Makhmutova [1,*], Batyr Sharimbayev [2], Altynbek Amirzhanov [3], and Ardak Shalkarbay-uly [4]

[1] Department of General Education, New Uzbekistan University, Tashkent, Uzbekistan
[2] Department of Information Systems, SDU University, Kaskelen, Kazakhstan
[3] Department of Mathematics and Natural Sciences, SDU University, Kaskelen, Kazakhstan
[4] Institute of Digital Transformation and Artificial Intelligence, Narxoz University, Almaty, Kazakhstan
Email: a.makhmutova@newuu.uz (A.M.); batyr.sharimbayev@sdu.edu.kz (B.S.); altynbek.amirzhanov@sdu.edu.kz (A.A.); ardak.shalkar@gmail.com (A.S.)
*Corresponding author

***Abstract*—The rise of Artificial Intelligence (AI)-generated text has led to the development of numerous detection tools to distinguish between human and machine-authored content. However, the effectiveness of these tools, especially against manipulated texts, remains uncertain. This study evaluates nine widely used AI detection tools—Turnitin, ZeroGPT, Detecting-AI.com, GPTZero, QuillBot, Grammarly, Sapling, Copyleaks, and Originality.ai—using texts from four large language models—ChatGPT, DeepSeek, Gemini, and Grok—as well as human-written samples. Initial findings indicate that commercial tools, such as Copyleaks and Originality.ai, achieved near-perfect detection rates, while free tools, including Grammarly and QuillBot, performed less reliably, with some as low as 63.0%. On the other hand, paraphrasing and Non-Native English Speakers (NNES)-style rewriting techniques reduced detection accuracy across most detectors. Turnitin dropped to 45.7%, while Grammarly fell to 19.0% in some cases. Only Copyleaks, GPTZero, and Sapling maintained strong performance under obfuscation. The study highlights three issues: inconsistent detector performance, the impact of obfuscation, and ethical risks, including bias and false positives. The study suggests that while some detectors offer robust baseline performance, combining them with pedagogical strategies and policies is essential to uphold academic integrity.**

***Keywords*—AI detection tools, large language models, obfuscation techniques, academic integrity, ethical implications**

I. Introduction

The rapid evolution of Large Language Models (LLMs) such as ChatGPT, DeepSeek, Gemini, and Grok has revolutionized text generation. These models produce human-like writing, creating serious challenges to academic integrity. The rise of such technologies has spurred the development of AI detection tools, such as Turnitin, Copyleaks, and GPTZero, that analyze linguistic and statistical patterns to identify machine-generated content. Obfuscation methods such as paraphrasing, translation, or Non-Native English Speakers (NNES)-style rewriting can significantly reduce detection accuracy. Prior studies, such as Krishna *et al.* [1], have shown that paraphrasing can dramatically reduce detection accuracy (e.g., from 70.3% to 4.6%). Akram [2] and Chaka [3] highlight inconsistent performance across tools and LLMs, with limited focus on newer models like DeepSeek and Gemini. These vulnerabilities raise concerns about ensuring fairness in academic assessments and preventing the erosion of critical thinking skills.

This study evaluates the reliability of nine widely used AI detection tools (Turnitin, ZeroGPT, Detecting-AI.com, GPTZero, QuillBot, Grammarly, Sapling, Copyleaks, and Originality.ai) in identifying texts generated by four LLMs (ChatGPT, DeepSeek, Gemini, and Grok) and human controls, with and without obfuscation. Objectives are to: (1) assess whether AI-generated texts from diverse LLMs can be reliably detected across platforms, and (2) examine how obfuscation techniques impact detection accuracy. By testing baseline and modified texts (paraphrased, translated, NNES-style), we address gaps in understanding tool performance and discuss their ethical and educational implications.

II. Literature Review

The advancement of LLMs such as ChatGPT, DeepSeek, Gemini, and Grok has transformed text generation, enabling outputs that closely mimic human writing. However, in educational settings, the misuse of AI-generated text threatens academic integrity through plagiarism and misinformation. AI text detection tools

Manuscript received July 24, 2025; revised August 27, 2025; accepted October 31, 2025; published March 10, 2026.

have emerged to distinguish human-written from AI-generated content. Yet, their effectiveness, particularly against obfuscation techniques such as paraphrasing, machine translation, and stylistic rewriting, remains under scrutiny. This literature review synthesizes recent studies to evaluate the capabilities of generative AI tools, the efficacy of detection tools, the impact of obfuscation on detection reliability across LLMs, and existing research gaps, with a focus on their implications for educational contexts.

A. Generative AI Tools and Their Role in Text Generation

Advanced LLMs now generate coherent and contextually relevant text. Their outputs often rival human writing. ChatGPT, developed by OpenAI, gained prominence for its "potential to act as a supportive scaffold". At the same time, GPT-4, its successor, offers "personalized recommendations" that depend on the "depth of the information provided" [4, 5]. On the other hand, according to Xiong *et al.* [6], DeepSeek has gained global recognition as an open-source Chinese LLM noted for its combination of low cost and high performance, positioning it as a competitive alternative to proprietary models. Gemini, developed by Google, emphasizes multimodal capabilities, integrating text and visual processing, while Grok, created by xAI, focuses on truth-seeking and concise responses, often tailored for specific domains. These models, which leverage transformer architectures and vast training datasets, produce outputs that challenge traditional notions of authorship, particularly in education, where students may use them to generate assignments, raising concerns about plagiarism [7]. The implications of these tools in academic settings are profound. Chaka [3] notes that LLMs, such as ChatGPT and YouChat, generate responses that are indistinguishable from human writing in controlled settings, complicating the detection of AI misuse. Sadasivan *et al.* [7] found that recursive paraphrasing attacks can dramatically reduce the effectiveness of AI-text detectors, with the actual positive rate of a watermarking scheme dropping from 99.3% to just 9.7%. In comparison, human evaluators still rated 77%–89% of the paraphrased passages as high quality in terms of content and grammar.

However, the accessibility of these models varies: DeepSeek and Grok offer open-source or subscription-based access. At the same time, Gemini and ChatGPT operate under proprietary frameworks that may limit research into their outputs. This variability underscores the need for detection tools that can identify texts generated by diverse LLMs. This challenge remains underexplored for newer models, such as DeepSeek and Gemini, compared to ChatGPT and GPT-4.

B. Efficacy of AI Text Detection Tools

To counter the proliferation of AI-generated texts, various detection tools have been developed, including Originality, Copyleaks, GPTZero, and others, with performance varying significantly across datasets and methodologies. Akram [2] evaluated 6 tools—GPTKit, GPTZero, Originality, Sapling, Writer, and Zylab—using Arslan's Human and AI Text Database (AH&AITD), which comprises 11,580 samples of human and AI-generated texts from models such as ChatGPT and GPT-4. The study found that "Originality is the most reliable alternative for AI text identification", achieving 97% accuracy, 98% precision, and 96% recall. In contrast, GPTKit achieved 55.29% accuracy, while GPTZero demonstrated balanced performance at 63.7% (p. 52). Kar *et al.* [8] tested ten free tools, including ZeroGPT, Copyleaks, and QuillBot, on a ChatGPT-generated scientific article, reporting that "Sapling and Undetectable AI were the most effective, detecting all three paraphrased contents with 100% accuracy". However, Dupli Checker consistently failed, misidentifying paraphrased texts as 0% AI-generated, revealing stark inconsistencies. Similarly, Said [9] demonstrated that deep neural network models, such as BL-CNN with Word2Vec and BL-CNN-BL with FastText, achieve high accuracy (up to 99.32%) in news text classification, suggesting potential applications for detecting AI-generated text using advanced neural architectures.

Chaka [3] evaluated five AI content detection tools—GPTZero, OpenAI Text Classifier, Writer.com's AI Content Detector, Copyleaks, and Giant Language Model Test Room (GLTR) —using responses generated by ChatGPT, YouChat, and Chatsonic. Among these, Copyleaks was the top-performing detector, particularly for English texts and some Google-translated outputs. However, the study concluded that none of the tools is yet fully reliable across different contexts. Bhattacharjee and Liu [10] evaluated the performance of ChatGPT (GPT-3.5) and GPT-4 as detectors of AI-generated text using the TuringBench dataset, which contains outputs from 19 different language models. They found that ChatGPT struggled, correctly identifying less than 50% of AI-generated samples for most generators.

In contrast, GPT-4 achieved very high detection rates (97%–100%) across the generators but frequently misclassified human-written text as AI-generated. These results highlight the trade-off between sensitivity to AI-generated text and specificity for human-authored text. To address challenges with black-box models, Shimada and Kimura [11] proposed a detection method combining log likelihoods and sentence embeddings, achieving improved accuracy for AI-generated texts, such as those from ChatGPT, in academic contexts, including essay evaluation. Similarly, deep learning techniques, as reviewed by Alduhailan *et al.* [12] in biometric authentication, highlight the challenges of achieving robust pattern recognition and offer invaluable insights for improving the accuracy of AI text detection systems. To consolidate these findings and highlight both the progress and the limitations of current detection tools, Table I provides a comparative summary of key evaluations, top-performing tools, accuracy ranges, and their most significant shortcomings.

TABLE I. SUMMARY OF AI TEXT DETECTION TOOL PERFORMANCE

No	Study	Tools Evaluated	Top Performer	Accuracy Range	Key Limitation
1	Akram [2]	GPTKit, GPTZero, Originality, Sapling, Writer, Zylalab	Originality	55.29%–97.00%	Struggles with paraphrased text
2	Kar *et al.* [8]	ZeroGPT, Copyleaks, QuillBot, GPTZero, Sapling, Undetectable AI, etc.	Sapling, Undetectable AI	0%–100%	Variability in paraphrased text detection
3	Bhattacharjee and Liu [10]	ChatGPT, GPT-4	GPT-4	<50%–100%	GPT-4 misclassified human text
4	Chaka [3]	GPTZero, OpenAI Text Classifier, Writer.com, Copyleaks, GLTR	Copyleaks	Varies	Fails with translated texts
5	Perkins *et al.* [13]	Turnitin	Turnitin	54.8%–91%	Vulnerable to prompt engineering
6	Elkhatat *et al.* [14]	OpenAI Classifier, Writer, Copyleaks, GPTZero, CrossPlag	Copyleaks	Varies	Inconsistent with GPT-4 and human text
7	Walters [15]	16 detectors	Copyleaks	Varies	Inconsistent with obfuscated texts

C. Impact of Obfuscation Techniques and Detection Reliability Across LLMs

The reliability of detecting AI-generated text depends strongly on the detector model used. Bhattacharjee and Liu [10] demonstrated that GPT-4, when employed as a detector on the TuringBench dataset, achieved very high accuracy (97%–100%) in identifying AI-generated outputs from 19 different generators; however, it frequently misclassified human-written texts as AI-generated. In contrast, ChatGPT (GPT-3.5) struggled to detect AI-generated texts, correctly identifying less than half of the samples for most generators. Walters [15] compared the effectiveness of 16 publicly available AI text detectors on undergraduate essays generated by ChatGPT-3.5, ChatGPT-4, and human students. The study found that while many detectors could reliably distinguish GPT-3.5 texts from human-written texts, most struggled with GPT-4 outputs, often misclassifying them as human-written. Only three detectors—Copyleaks, Turnitin, and Originality.ai—achieved consistently high accuracy across all three document sets, underscoring the need for broader, more systematic testing of detectors as LLMs advance.

Obfuscation techniques, such as paraphrasing, machine translation, and stylistic rewriting, significantly undermine detection accuracy. Krishna *et al.* [1] show that even modest paraphrasing can sharply reduce detector performance: their DIPPER model lowered DetectGPT's accuracy from 70.3% to just 4.6%, despite preserving semantic meaning. Amirzhanov *et al.* [16] further highlight the challenges of detecting paraphrased and cross-lingual plagiarism, noting that advanced techniques, such as semantic similarity models and multilingual embeddings, improve detection but struggle with AI-generated content and sophisticated obfuscation methods. Malik and Amjad [17] evaluated the effectiveness of four AI-detection tools—Turnitin, ZeroGPT, GPTZero, and Writer AI- against text generated by ChatGPT, Perplexity, and Gemini. They also examined how adversarial techniques such as Grammarly edits, QuillBot paraphrasing, and human revisions affect detection accuracy. Turnitin demonstrated the highest and most consistent performance, while QuillBot paraphrasing significantly reduced detection accuracy across other tools, highlighting the inconsistency and unreliability of current AI-detection systems.

Chaka [3] reported that GPTZero misidentified all machine-translated texts as human-produced, underscoring its limitations in detecting AI-generated content across non-English languages. Weber-Wulff *et al.* [18] found that machine-translated texts reduced detection accuracy by approximately 20%, as translation traces often led tools to misclassify human-written text as AI-generated. Fishchuk and Braun [19] demonstrated that while commercial detectors, such as Copyleaks and GPTZero, remained generally robust, paraphrasing attacks, particularly with tools like QuillBot, could still lower detection scores, highlighting that no system is entirely invulnerable. Additionally, Perkins *et al.* [20] found that simulating NNES writing styles, incorporating minor grammatical errors and reduced coherence, decreased detection accuracy by 12.0%. Tools like Copyleaks misclassified 50.0% of human-written NNES texts as AI-generated. This highlights a critical bias against NNES writers, raising ethical concerns about unfair accusations and the need for more robust, inclusive detection methods.

Theoretical limits further complicate detection. Perkins *et al.* [13] found that although Turnitin's detector flagged 91% of AI-generated submissions, it only identified 54.8% of the actual content as AI, underscoring the difficulty of relying on current tools in real assessment contexts. However, Chakraborty *et al.* [21] argue that detection remains feasible with sufficient data, since even when the total variation distance between human- and AI-generated texts is small, increasing the sample size allows AUROC scores to approach 1.

Ethical concerns are significant in educational contexts. False positives, where human-written text is misclassified, risk unfair plagiarism accusations, especially for non-native English speakers. Bhattacharjee and Liu [10] found that "GPT-4 struggles to identify human-written text, and misclassified over 95% as AI-generated", a limitation that could disproportionately affect non-native English-speaking students by increasing the risk of false plagiarism accusations. Rafiq *et al.* [22] caution that false allegations from unreliable detection methods may unfairly penalize students and undermine confidence in assessment systems. He further notes that reliance on cloud-based tools can raise concerns under regulations such as GDPR.

D. Research Gaps in AI Text Detection

Despite rapid advances in AI text detection, several research gaps persist. First, the absence of standardized evaluation metrics complicates cross-study comparisons: Akram [2] reported accuracy, precision, and recall values, Kar *et al.* [8] emphasized percentage-based detection outcomes, and Bhattacharjee and Liu [10] focused on misclassification rates, while Elkhatat *et al.* [14] highlighted inconsistencies across five commercial detectors, cautioning that no single metric or tool is sufficient for academic decision-making. Second, existing benchmark datasets such as AH&AITD [2] and TuringBench [10] remain limited in scope, with most studies focusing on ChatGPT or GPT-4, leaving models such as DeepSeek, Gemini, and Grok underexplored [3, 6]. Third, while paraphrasing attacks have been well-studied [1, 7], more sophisticated adversarial strategies, such as embedding-based perturbations [23], demonstrate that even state-of-the-art detectors like Fast-DetectGPT can be compromised, underscoring the fragility of current systems. Fourth, multilingual detection remains underdeveloped: GPTZero failed on machine-translated texts [3], and Weber-Wulff *et al.* [18] reported accuracy drops of 20% on translated samples, while Gundersen [24] found that LLM-based detectors achieved only ~50% accuracy on Norwegian texts, illustrating severe limitations in low-resource language contexts. Finally, domain-specific detection challenges remain largely overlooked. Perkins *et al.* [20] demonstrated that simulating NNES writing styles reduces accuracy and biases detectors against certain student groups. Pan *et al.* [25] showed that detectors perform poorly on AI-generated code, raising integrity concerns in programming education. Collectively, these gaps highlight the urgent need for more systematic evaluations across diverse models, languages, and domains, with stronger defenses against obfuscation and adversarial attacks, and for fairer approaches that mitigate bias against NNES writers and students in technical disciplines.

The present study responds to these shortcomings by systematically comparing nine widely used detection tools across four LLMs and a set of human-authored texts. It expands evaluation to under-studied models, applies three distinct obfuscation strategies in a controlled design, and standardizes performance metrics to allow comparability across detectors. Importantly, it also provides empirical evidence on how current detection systems may disadvantage NNES writers, thereby highlighting the need for fairer and more reliable approaches in educational and professional contexts.

III. Methodology

This study employs a multi-dataset experimental design to address two primary research questions: (1) Can AI-generated texts from various Large Language Models (LLMs) be reliably detected across different detection platforms? and (2) How do obfuscation techniques, such as paraphrasing, back-translation, and stylistic rewriting, impact detection accuracy? To address these research questions, the study first identified and selected nine representative AI detection tools through a structured, multi-source review process, followed by the construction of multiple datasets for comparative evaluation. Next, we generated five datasets comprising baseline AI-generated texts from four LLMs (ChatGPT-4, DeepSeek, Gemini, and Grok), human-authored texts, and three sets of obfuscated texts (paraphrased, back-translated, and rewritten to simulate non-native English speaker writing at the International English Language Testing System (IELTS) (Band 6 level). Nine widely used AI detection tools were evaluated for accuracy, precision, recall, and F1 score using a within-subjects non-parametric approach. Statistical analyses were performed in Python to compare performance across datasets and LLMs, ensuring robust and reproducible findings.

A. Systematic Process for Shortlisting Tools

The study employed a systematic, multi-source review process to identify and select nine representative AI detection tools, aiming to answer the research questions. The selection process included (1) a thorough review of the literature, (2) focused web searches, and (3) consulting resources that are commonly used in institutional and educational settings. This triangulated approach ensured the inclusion of both theoretically supported and practically applicable resources available to researchers and educators.

The shortlisted tools represent a variety of both free and paid detectors currently in use in academic and professional settings. The source and justification for the inclusion of each detector are compiled in Table II. Originality.ai, QuillBot, Sapling, and Copyleaks are among the tools mentioned in scholarly literature. In contrast, others, such as Grammarly and Turnitin, were chosen due to their institutional relevance and ease of integration into educational processes. To reflect the diversity of the market and research today, new tools such as Detecting-ai.com and well-known platforms such as ZeroGPT and GPTZero were added.

TABLE II. Systematic Process for Shortlisting AI Detection Tools

Tool	Source of Inclusion	Rationale
Originality.ai	Ref. [2]	Top performer in AI–human text differentiation studies
Copyleaks	Refs. [3, 13, 14]	Frequently reported as accurate and widely adopted
Turnitin	Educational consultation [13]	Established use in universities for plagiarism and AI detection
Sapling	Ref. [8]	High accuracy for paraphrased and AI-generated text
QuillBot	Ref. [8]	Relevant for paraphrasing and text rephrasing detection
ZeroGPT	Web + Refs. [2, 3, 8, 14, 17]	Popular, cited in comparative studies, broadly accessible
GPTZero	Web + Refs. [2, 3, 8, 14, 17]	Widely adopted by educators and researchers
Grammarly	Educational consultation	Integrated AI detection; institutional relevance
Detecting-ai.com	Web search	Recent, free, emerging tool with limited prior research

Tools were retained if they were functional, online-accessible, and produced interpretable text-level reports.

Google Scholar and specific web queries (such as "AI text detector", "AI content detection tool", and "AI-generated text identification in education") were used for the searches. Only tools specifically mentioned in scholarly publications, institutional records, or professional forums focused on education were included between 2023 and 2024. Tools were retained if they were usable and accessible online, and generated reports at the text level that could be understood.

B. Selection and Evaluation Criteria

Each shortlisted detector was evaluated based on three main educational usability principles: Accessibility, Usability, and Prevalence, which were adapted from technology evaluation frameworks commonly used in digital learning research. The operational definitions and indicators for each criterion are summarized in Table III. To ensure a balanced representation of accessibility (free vs. paid, inclusivity of use), usability (simple interface, clarity of output), and prevalence (degree of academic and institutional adoption), tools were qualitatively compared across these dimensions rather than being assigned numerical scores.

Table IV lists the quantitative parameters of each tool, including usage restrictions, availability, and input size limits, to support these qualitative standards. Table IV lists usage limits and input restrictions, such as Copyleaks' credit-based system (1 credit = 250 words) and Turnitin's 300–30,000 word limit with no daily restrictions. These requirements were taken into account to ensure uniformity in the experimental design and to put the accessibility and usability results into context. The practicality and scalability of each detector for use in research and educational settings are determined by these technical features taken together.

TABLE III. EVALUATION CRITERIA FOR AI DETECTION TOOLS

Criterion	Operational Definition	Measurement/Indicator
Accessibility	Availability and inclusivity of tool access	Free/paid status, browser compatibility, registration requirements, accessibility features
Usability	Ease of use and interface efficiency	Time to analyze text, clarity of output, and user interface intuitiveness
Prevalence	Level of adoption and visibility	Mentions in educational studies, institutional use, and web search frequency

TABLE IV. MINIMUM AND MAXIMUM INPUT SIZES FOR FREE AND PAID AI DETECTION TOOLS

No	Tool Name	Available Online	Minimum Size	Maximum Size	Limits
1	Turnitin	Free	300 words	30,000 words	Unlimited
2	ZeroGPT	Free	~250 characters	15,000 characters	Unlimited
3	Detecting-ai.com V2	Free	1 character	5000 characters	100 detections/day
4	Copyleaks	Paid	350 characters	25,000 characters	1200 credits (1 credit = 250 words), Cost: $18.99
5	Originality.ai	Paid	300 words	7500 words	2000 credits (1 credit = 100 words), Cost: $14.95
6	GPTZero	Free	250 characters	5000 characters	10,000 words/month
7	QuillBot AI Detector	Free	80 words	1200 words	Unlimited
8	Grammarly AI Detector	Free	200 characters	10,000 characters	300 documents or 150,000 words/month
9	Sapling	Free	2000 characters	100,000 characters	Unlimited

C. Dataset Generation

This study constructed five datasets to evaluate AI detection tools, as detailed in the publicly available dataset on Kaggle [25]. An overview of the datasets is provided in Table V. Dataset 1 consists of 24 baseline AI-generated texts produced by four LLMs: ChatGPT-4, DeepSeek, Gemini, and Grok. Each model generated responses to 6 academic prompts covering Mathematics, Biology, History, Economics, Computer Science, and an IELTS-style essay, with formatting restrictions in place to ensure consistency (approximately 400 words, excluding formulas, symbols, lists, or special formatting). The exact prompts are provided in the Kaggle dataset metadata to ensure replicability [26]. Dataset 2 includes paraphrased versions of Dataset 1, reworded using QuillBot's default settings to preserve meaning while altering lexical and syntactic structures. Dataset 3 comprises back-translated texts, where the texts from Dataset 1 were translated into Russian using Yandex Translate and then back into English using Google Translate to introduce stylistic distortions. Dataset 4 contains rewritten versions of the texts from Dataset 1, modified to simulate Non-Native English Speaker (NNES) writing at the IELTS Band 6 level by incorporating minor grammatical errors, vocabulary misuse, and awkward phrasing while maintaining clarity, achieved through targeted rewriting prompts. This method, adopted from [13], was shown to be the most effective humanizer approach in their study. To ensure consistency, each text in Dataset 4 was generated by the same LLM that produced the corresponding baseline text in Dataset 1. Finally, Dataset 5 comprises 20 human-written excerpts (approximately 400 words each) from peer-reviewed articles on arXiv.org published before ChatGPT (November 2022) in the same five academic subjects, serving as a control group to assess false positives.

D. Evaluation Procedure

The evaluation involved submitting 116 texts (24 AI-generated texts per each of four datasets and 20 human-written texts) to nine AI detection tools, resulting in 1044 manual text submissions. Each text was copied and pasted into each detector's interface, adhering to the input constraints (Table V). Detection outputs were recorded as probability scores, with a 50% threshold used to classify texts as AI-generated or human-written. Performance was assessed using accuracy, precision, recall, and F1-Score, calculated across all datasets and large language models to ensure a comprehensive comparison.

TABLE V. OVERVIEW OF DATASETS USED IN THE STUDY

Dataset	Source	LLMs Used	Obfuscation Technique	Dataset Size
Dataset 1: Baseline AI Texts	Prompts in 5 academic domains + 1 IELTS essay	ChatGPT-4 Turbo, DeepSeek, Gemini, Grok	None	24 texts
Dataset 2: Paraphrased Texts	Derived from Dataset 1	ChatGPT-4 Turbo, DeepSeek, Gemini, Grok	Paraphrasing via QuillBot	24 texts
Dataset 3: Translated Texts	Derived from Dataset 1	ChatGPT-4 Turbo, DeepSeek, Gemini, Grok	Back-translation (Russian ↔ English)	24 texts
Dataset 4: NNES-Style Texts	Derived from Dataset 1	ChatGPT-4 Turbo, DeepSeek, Gemini, Grok	Manual rewriting to simulate IELTS Band 6 non-native writing	24 texts
Dataset 5: Human-Written Texts	Excerpts from academic articles (arXiv.org, pre-Nov 2022)	N/A	None	20 texts

E. Statistical Analysis

A within-subjects non-parametric approach evaluated the performance of seven AI detection tools (Turnitin, ZeroGPT, Detecting-ai, GPTZero, QuillBot, Grammarly, Sapling) across four datasets: Baseline AI-Generated Texts, Paraphrased Texts, Translated Texts, and NNES-Style Texts, with detection accuracy (%) as the dependent variable. Copyleaks and Originality.ai were excluded because they showed zero variance (100% accuracy), precluding statistical analysis. The Shapiro-Wilk test confirmed non-normal distributions for most detectors ($p < 0.05$), rendering parametric Two-Way ANOVA inappropriate. The Friedman test was applied to compare accuracy across detectors within each dataset, treating large language models as subjects. The data were reshaped into a long format, with missing scores imputed using column means, and missing LLM identifiers assigned temporary unique IDs. Significant Friedman test results ($p < 0.05$) prompted post hoc Wilcoxon signed-rank tests with Bonferroni correction to identify specific detector differences, maintaining a family-wise error rate of 0.05. Descriptive statistics (means, standard deviations) and boxplots summarized and visualized the performance. Analyses were conducted in Python (version 3.11) using pandas, scipy, pingouin, and seaborn libraries.

IV. RESULTS AND DISCUSSION

A. Detection Accuracy Across LLMs and Platforms

Table VI presents the detection accuracy of nine AI content detectors across baseline AI-generated texts from four LLMs—ChatGPT-4, DeepSeek, Gemini, and Grok (Dataset 1)—and human-written texts (Dataset 5). Copyleaks and Originality.ai achieved perfect accuracy (100%) across all LLMs and human texts, demonstrating robust performance. Sapling also performed strongly, with accuracies ranging from 88.2% (human texts) to 100% (for ChatGPT), while GPTZero showed high reliability, particularly for Grok (100%) and human texts (98.4%). In contrast, Grammarly exhibited the lowest accuracy, notably struggling with Grok texts (38.2%) despite near-perfect performance on human texts (99.4%). Turnitin and QuillBot displayed variability, with Turnitin scoring 80.0% on Grok texts and QuillBot dropping to 79.5% for the same LLM. Detecting-ai and ZeroGPT maintained moderate to high accuracies, though ZeroGPT's performance dipped to 74.5% for Grok texts.

TABLE VI. DETECTION ACCURACY (%) OF AI DETECTORS BY INDIVIDUAL LLM AND HUMAN TEXTS

No	Tool Name	ChatGPT	DeepSeek	Gemini	Grok	Human
1	Turnitin	83.3	97.7	100.0	80.0	97.5
2	ZeroGPT	98.5	99.8	99.0	74.5	94.8
3	Detecting-ai.com V2	93.8	91.3	84.8	80.3	79.6
4	Copyleaks	100.0	100.0	100.0	100.0	100.0
5	Originality.ai	100.0	100.0	100.0	100.0	100.0
6	GPTZero	97.3	86.7	93.8	100.0	98.4
7	QuillBot AI Detector	97.8	100.0	81.7	79.5	100.0
8	Grammarly AI Detector	75.8	68.0	71.8	38.2	99.4
9	Sapling	100.0	99.2	99.2	90.7	88.2

Following the presentation of detection accuracy data in Table VI, we conducted a non-parametric Friedman test to evaluate whether there were statistically significant differences in detector performance across the various large language models and human-written texts. The test revealed a significant effect of detector type on accuracy scores ($Q = 61.26$, $p < 0.001$), indicating that not all detectors performed equally. To further explore these differences, post-hoc pairwise Wilcoxon signed-rank tests were conducted with both Bonferroni and False Discovery Rate (FDR) corrections. The results showed that Grammarly's detection accuracy was significantly lower than that of several other detectors (e.g., Sapling, ZeroGPT, GPTZero), confirming its comparatively poor performance. Sapling consistently outperformed most competitors, although its advantage over GPTZero and ZeroGPT was not always statistically significant after correction for multiple comparisons. These findings highlight substantial variability in AI detection accuracy across tools, underscoring the importance of selecting robust detectors to identify AI-generated content.

Table VII presents aggregated performance metrics (accuracy, precision, recall, and F1 score) for each AI detection tool, evaluated across a combined dataset comprising LLM-generated texts and human-written texts from Datasets 1 and 5. This comprehensive assessment,

visualized in Fig. 1, highlights each tool's ability to distinguish between AI-generated and human content, complementing the individual LLM accuracies presented in Table VI. The results reveal that Copyleaks, Originality.ai, and GPTZero achieved perfect scores across all metrics, indicating highly consistent and accurate performance across all input types. Tools like Detecting-ai and Sapling also performed strongly, with perfect recall and near-perfect F1-Scores, reflecting their reliability in identifying all AI-generated texts while minimizing false positives. Meanwhile, Grammarly and QuillBot showed lower recall values, suggesting they were more conservative and occasionally failed to detect AI-generated texts despite high precision.

Across baseline LLM outputs and human texts (Table VI), Copyleaks and Originality.ai achieved 100% accuracy for every category. GPTZero also performed strongly (e.g., 100.0% on Grok, 98.4% on human texts; baseline mean 95.24%), and Sapling performed well overall (baseline mean 95.46%). In contrast, Grammarly showed the lowest baseline reliability (mean 70.64%), with particularly weak performance on Grok (38.2%). These differences were statistically significant (Friedman $Q = 61.26$, $p < 0.001$). Aggregated metrics across baseline LLM + human texts (Table VII) echo this pattern: Copyleaks, Originality.ai, and GPTZero achieved 100% accuracy/precision/recall/F1, while Detecting-ai and Sapling posted perfect recall and near-perfect F1, and Grammarly/QuillBot showed lower recall (missed AI cases).

TABLE VII. AGGREGATED PERFORMANCE METRICS (%) ACROSS COMBINED LLM AND HUMAN TEXTS

No	Tool Name	Accuracy	Precision	Recall	F1-Score
1	Turnitin	93.2	95.7	91.7	93.6
2	ZeroGPT	95.5	95.8	95.8	95.8
3	Detecting-ai.com V2	97.7	96.0	100.0	98.0
4	Copyleaks	100.0	100.0	100.0	100.0
5	Originality.ai	100.0	100	100.0	100.0
6	GPTZero	100.0	100.0	100.0	100.0
7	QuillBot AI Detector	95.5	100.0	91.7	95.7
8	Grammarly AI Detector	90.9	100.0	83.3	90.9
9	Sapling	97.7	96.0	100.0	98.0

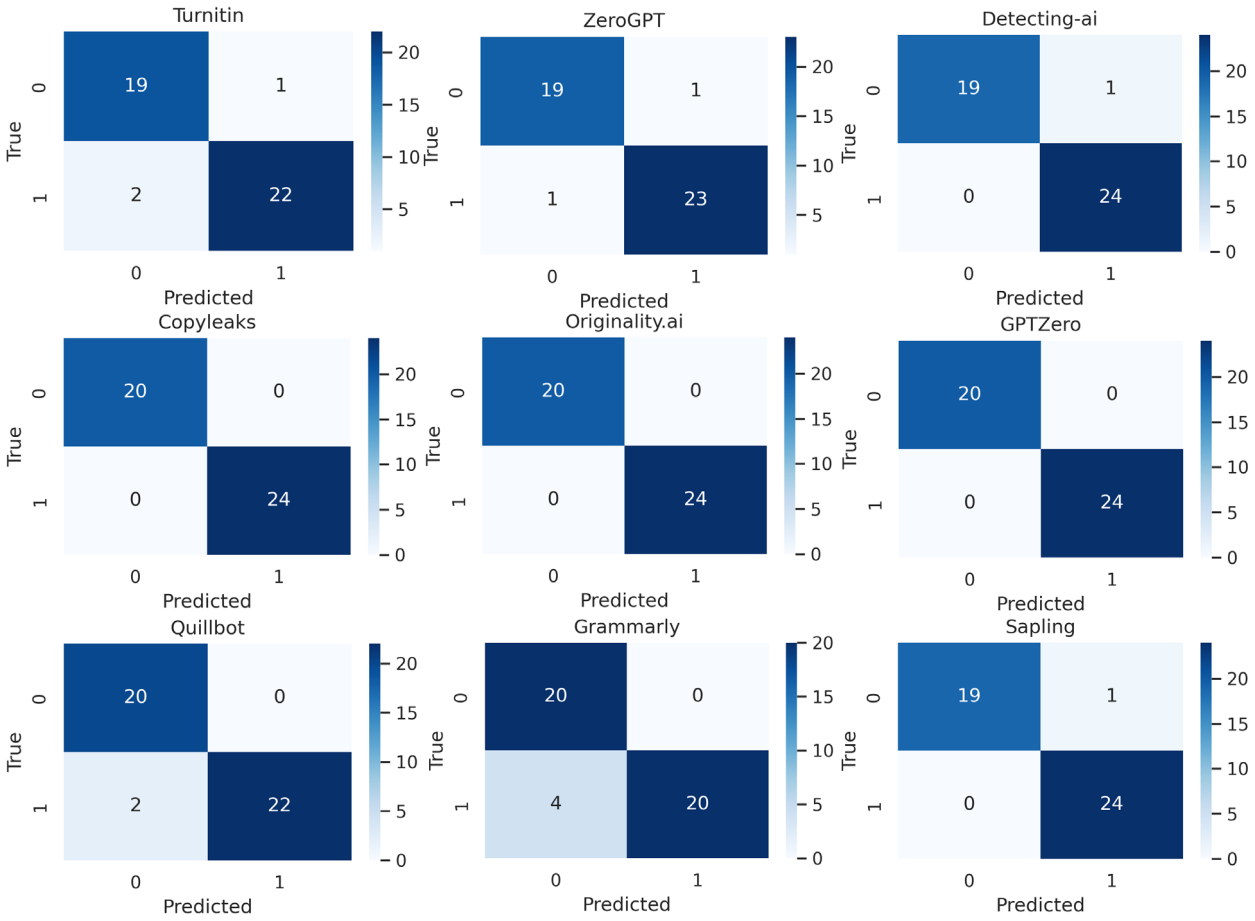

Fig. 1. Confusion matrices for AI detection tools across combined LLM (Dataset 1) and human texts (Dataset 5).

B. Effect of Obfuscation Techniques on Detection Accuracy

Table VIII presents the detection accuracy (%) of various AI detectors when applied to paraphrased versions of LLM-generated texts. This experiment assesses each detector's ability to identify AI-generated content after it has been modified using paraphrasing techniques, a common strategy used to evade detection. The texts were initially generated by four large language models (ChatGPT, DeepSeek, Gemini, and Grok) and then paraphrased before being evaluated by each detection tool. The results reveal substantial variability in how detectors respond to paraphrased input. Tools like Copyleaks and Sapling maintained extremely high accuracy (near or at 100%) across all models, indicating strong robustness to paraphrasing. GPTZero and Originality.ai also performed well, though with some variance across LLMs. In contrast, detectors such as Turnitin, QuillBot, and, especially, Grammarly experienced significant drops in performance, with accuracies as low as 19.0% on Grok-generated texts.

Sapling achieved the highest mean accuracy (M = 96.50, SD = 10.71), followed by GPTZero (M = 93.38, SD = 15.49). Grammarly had the lowest mean accuracy (M = 40.00, SD = 22.23), with high variability, suggesting poor, inconsistent performance. Turnitin (M = 71.21, SD = 40.45) and QuillBot (M = 63.88, SD = 43.07) showed substantial variability, with minimum scores of 0, indicating challenges in detecting certain translations. ZeroGPT (M = 80.21, SD = 22.11) and Detecting-ai (M = 80.75, SD = 11.74) had moderate performance.

TABLE VIII. PARAPHRASED TEXT DETECTION ACCURACY (%) OF AI DETECTORS

No	Tool Name	ChatGPT	DeepSeek	Gemini	Grok
1	Turnitin	71.5	75.3	92.3	45.7
2	ZeroGPT	90.0	87.5	78.5	64.8
3	Detecting-ai.com V2	82.0	85.2	80.8	75.0
4	Copyleaks	100.0	100.0	100.0	100.0
5	Originality.ai	83.7	100.0	86.7	100.0
6	GPTZero	100.0	93.8	88.5	91.2
7	QuillBot AI Detector	79.2	59.2	62.7	54.5
8	Grammarly AI Detector	46.3	54.0	40.7	19.0
9	Sapling	100.0	99.0	99.8	87.2

A Friedman test was conducted to assess whether statistically significant differences existed among the detectors when evaluating paraphrased LLM-generated texts. The test revealed a highly significant effect of detector type on accuracy scores ($Q = 70.14$, $p < 0.001$), indicating that paraphrasing had a differential impact across tools. Subsequent post hoc Wilcoxon signed-rank tests with Bonferroni and FDR corrections confirmed that Grammarly consistently underperformed relative to nearly all other detectors, with statistically significant differences from Sapling, GPTZero, ZeroGPT, and Originality.ai (all $p < 0.001$, FDR-corrected). Sapling demonstrated the most consistent superiority, significantly outperforming Grammarly, QuillBot, Turnitin, and ZeroGPT in multiple pairwise comparisons. These findings highlight the importance of detector robustness in adversarial contexts such as paraphrasing, where only a few tools, particularly Sapling and Copyleaks, demonstrated consistent detection reliability.

To provide a more aggregated perspective, Fig. 2 illustrates the average accuracy of each AI detector across all paraphrased LLM outputs using a bar chart. This visual representation highlights the overall effectiveness of each tool under obfuscation conditions. It reinforces the table's finding that some detectors remain highly accurate even when the text has been significantly altered, while others are more susceptible to manipulation.

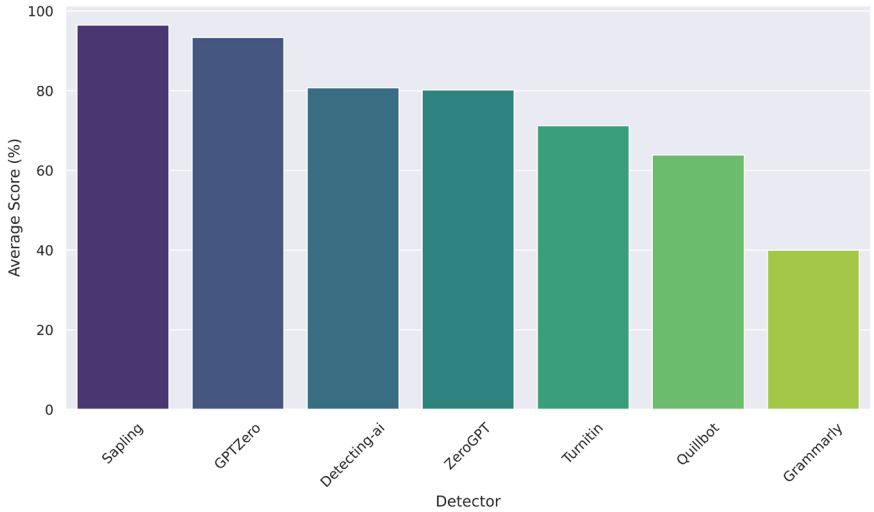

Fig. 2. Average AI detector accuracy on paraphrased text.

TABLE IX. TRANSLATED TEXT DETECTION ACCURACY (%) OF AI DETECTORS

No	Tool Name	ChatGPT	DeepSeek	Gemini	Grok
1	Turnitin	77.2	94.8	98.5	38.2
2	ZeroGPT	97.3	99.0	97.3	75.2
3	Detecting-ai.com V2	90.8	91.8	79.8	81.5
4	Copyleaks	100.0	100.0	100.0	100.0
5	Originality.ai	100.0	100.0	100.0	88.0
6	GPTZero	94.7	91.5	87.5	91.0
7	Quillbot AI Detector	94.0	86.3	72.3	40.3
8	Grammarly AI Detector	48.7	36.3	37.8	30.3
9	Sapling	100.0	99.2	100.0	87.0

Table IX presents the detection accuracy (%) of nine AI detectors evaluating LLM-generated texts subjected to back-translation (from Russian to English), simulating an obfuscation technique that alters the surface structure while preserving semantic meaning. Sapling achieved the highest mean accuracy (M = 96.54, SD = 10.55), followed by GPTZero (M = 91.17, SD = 14.93) and ZeroGPT (M = 92.21, SD = 12.79). Grammarly had the lowest mean accuracy (M = 38.29, SD = 19.18), with high variability, suggesting poor, inconsistent performance. QuillBot (M = 70.75, SD = 36.43) and Turnitin (M = 77.17, SD = 37.13) showed substantial variability, with minimum scores of 0, reflecting challenges in detecting certain translated texts.

The Friedman test revealed significant differences in the performance of AI detectors on translated LLM-generated texts ($\chi^2 = 69.52$, $p < 0.001$), indicating that at least some detectors differed in accuracy. Post-hoc Wilcoxon tests (with FDR correction) showed that Sapling significantly outperformed several detectors, including Detecting-ai ($p = 0.0033$), Grammarly ($p < 0.0001$), QuillBot ($p = 0.0020$), and Turnitin ($p = 0.0150$), confirming its robustness in handling translated text. Conversely, Grammarly demonstrated the weakest performance, significantly underperforming compared to all others (all $p < 0.001$), consistent with its lowest mean accuracy and high variability. GPTZero, ZeroGPT, and Detecting-ai are clustered in the upper-middle tier, with some significant pairwise differences but generally comparable performance. These findings highlight that while some detectors are resilient to obfuscation through translation, others, particularly Grammarly, struggle to maintain reliability across languages. Fig. 3, a bar chart, displays the mean detection accuracy across detectors, highlighting the superior performance of Sapling and GPTZero, as well as Grammarly's vulnerability to translation-based obfuscation.

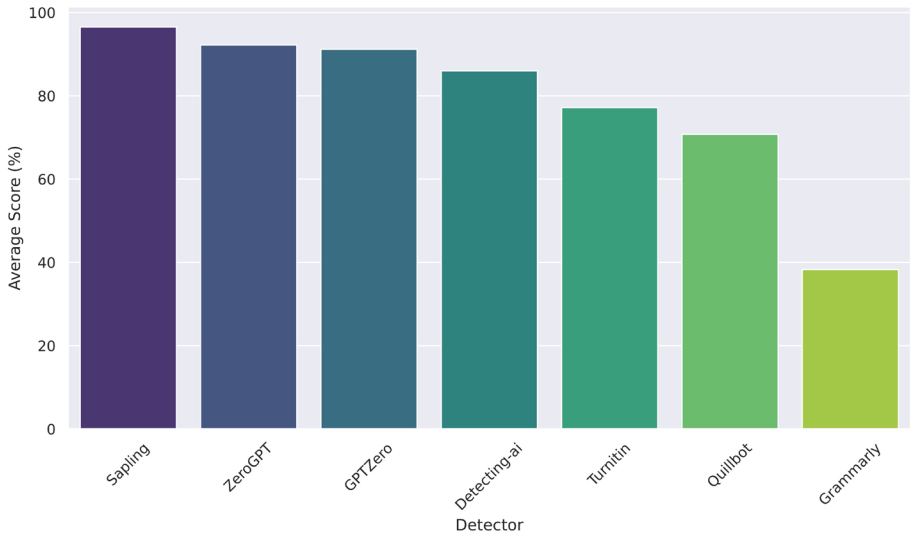

Fig. 3. Average AI detector accuracy on translated text

TABLE X. DETECTION ACCURACY (%) OF AI DETECTORS ON TEXTS SIMULATING NNES.

No	Tool Name	ChatGPT	DeepSeek	Gemini	Grok
1	Turnitin	0.0	14.7	0.0	0.0
2	ZeroGPT	18.8	34.7	0.0	0.0
3	Detecting-ai.com V2	17.2	21.7	1.3	1.5
4	Copyleaks	83.3	100	83.3	83.3
5	Originality.ai	95.2	99.7	39.3	80.8
6	GPTZero	99.8	93.8	66.4	100.0
7	QuillBot AI Detector	18.0	41.0	11.2	19.2
8	Grammarly AI Detector	0.0	8.3	0.0	0.0
9	Sapling	54.5	45.7	0.0	0.0

Table X presents the detection accuracy (%) of nine AI detectors on texts rewritten to simulate the NNES writing style at IELTS Band 6 proficiency, introducing minor grammatical errors, awkward phrasing, and vocabulary misuse to reduce detectability while maintaining coherence, see also Fig. 4. Across NNES-simulated texts, GPTZero achieved the highest mean accuracy (90.0%). Copyleaks (87.5%) and Originality.ai (79.0%) also performed strongly, whereas Sapling (25.0%), QuillBot (22.3%), ZeroGPT (13.4%), Detecting-ai (10.4%), Turnitin (3.7%), and Grammarly (2.1%) lagged, often scoring 0% on Gemini and Grok cases. These results indicate that while GPTZero, Copyleaks, and Originality.ai retain comparatively high reliability under this obfuscation, most other detectors are fragile.

The Friedman test confirmed significant differences in detection accuracy across the nine systems, $\chi^2(8) = 134.42$, $p < 0.001$. Post-hoc Wilcoxon tests (Bonferroni- and FDR-corrected) revealed that GPTZero, Copyleaks, and Originality.ai consistently outperformed the weaker detectors. GPTZero did not differ significantly from Copyleaks ($p = 0.72$) or Originality.ai ($p = 0.24$), suggesting comparable top-tier performance. Still, it significantly surpassed QuillBot, Sapling, Turnitin, ZeroGPT, Detecting-ai, and Grammarly (all $p < 0.001$ after correction). Copyleaks also outperformed most low-performing detectors, though its advantage over Originality.ai was not significant ($p = 0.05$, ns after correction). Originality.ai showed clear superiority to QuillBot, Sapling, Turnitin, and ZeroGPT (all corrected $p < 0.01$). Among the weaker systems, Grammarly, Turnitin, and Detecting-ai performed worst, with no reliable differences between them. These results reinforce the descriptive pattern: GPTZero, Copyleaks, and Originality.ai maintained high robustness against NNES-style obfuscation, while the remaining detectors exhibited substantial vulnerabilities.

These pairwise comparisons highlight relative detector strengths under NNES-style obfuscation, but they do not capture performance across all obfuscation types. To address this, we aggregated results from all five datasets and applied Principal Component Analysis (PCA) to visualize overall detector behavior (Fig. 5). By aggregating performance, the analysis captures detectors' overall consistency rather than results tied to a single obfuscation type. Detectors that appear close together exhibit similar response patterns across all datasets, while the marker size and color indicate the average accuracy. GPTZero and Originality.ai emerge as the most reliable tools, clustering in the high-performance region, whereas Turnitin, Grammarly, and Detecting-ai occupy low-accuracy areas with limited robustness. These visual trends are supported by the Friedman test ($\chi^2(8) = 346.53$, $p < 0.001$), which confirmed significant overall differences among detectors, and post-hoc Wilcoxon analyses that highlighted GPTZero and Originality.ai significantly outperforming weaker systems. Taken together, the PCA and statistical tests reveal clear performance stratification, demonstrating that some detectors generalize well across diverse text manipulations while others fail consistently.

Paraphrasing (Table VIII; Fig. 2) distinguishes between robust and fragile detectors. Copyleaks remained at 100% across all LLMs; Sapling stayed very high (e.g., 87.2% on Grok; near-100% otherwise), and GPTZero was consistently strong (e.g., 88.5% on paraphrased Gemini). In contrast, Turnitin fell to 45.7% (Grok) and Grammarly to 19.0% (Grok). Differences among detectors were significant (Friedman Q = 70.14, $p < 0.001$).

Back-translation (Table IX; Fig. 3) produced a similar pattern: Copyleaks again at 100%; Sapling remained high (87.0%–100.0% across LLMs); GPTZero/ZeroGPT generally performed well; Grammarly was weakest (e.g., 30.3% on Grok). Differences were significant ($\chi^2 = 69.52$, $p < 0.001$).

NNES-style rewrites (Table X; Fig. 4) were the most challenging. GPTZero retained the highest mean (90.0%), with Copyleaks (87.5%) and Originality.ai (78.8%) also strong. Most other tools collapsed (Sapling: 25.0%, QuillBot: 22.3%, ZeroGPT: 13.4%, Detecting-ai: 10.4%, Turnitin: 3.7%, Grammarly: 2.1%). Statistical tests confirmed marked differences (Friedman $\chi^2(8) = 134.42$, $p < 0.001$), with GPTZero, Copyleaks, and Originality.ai significantly outperforming weaker systems in post-hoc analyses.

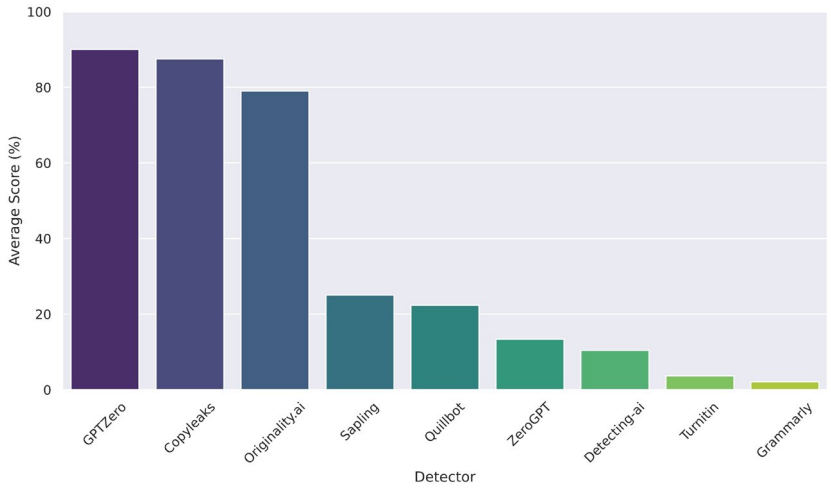

Fig. 4. Average detection accuracy of AI detectors on simulated NNES.

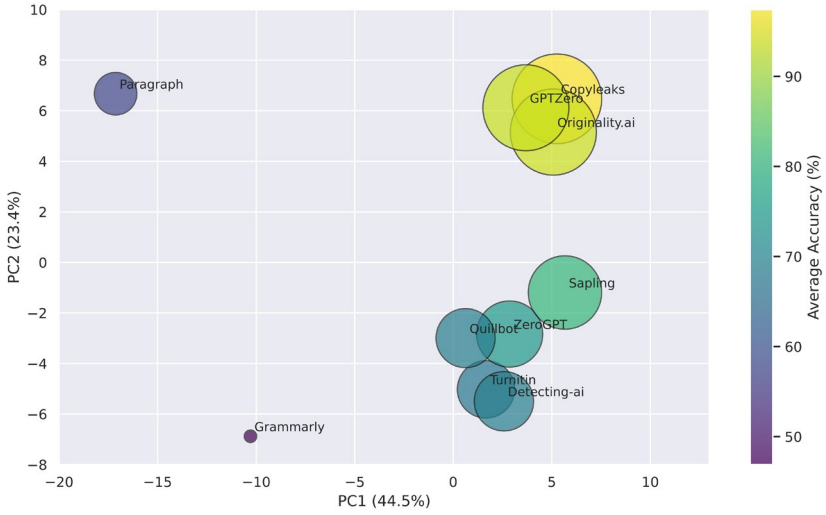

Fig. 5. PCA visualization of AI detector performance across different text categories.

A global view that aggregates all datasets and visualizes detector behavior via PCA (Fig. 5) reinforces this stratification; a global Friedman test also indicated significant overall differences ($\chi^2(8) = 346.53$, $p < 0.001$).

Beyond overall performance trends, two key risks emerge from the results. First, false positives: several detectors occasionally misclassified human writing (see Table VI and Fig. 1), which is especially concerning for NNES students given the severe accuracy drops for many tools on NNES-style text (Table X). Second, evasion risk: accessible obfuscation (paraphrasing/translation) can degrade performance for some detectors (e.g., Grammarly, 19.0%, and Turnitin, 45.7%, on paraphrased Grok; Table VIII), making sole reliance on any single detector unwise.

V. Conclusion

This study assessed the reliability of nine widely used AI content detectors on texts generated by four large language models (ChatGPT-4, DeepSeek, Gemini, and Grok) alongside human-authored samples. The findings revealed a clear performance divide: Copyleaks, Originality.ai, and GPTZero consistently achieved near-perfect accuracy across baseline and obfuscated texts, while Grammarly and Turnitin were substantially less reliable, in some cases detecting as little as 2.1% and 3.7% of NNES-style texts. Obfuscation strategies such as paraphrasing, back-translation, and particularly stylistic rewriting to mimic non-native English speaker patterns significantly reduced the effectiveness of most detectors, underscoring their vulnerability to commonly accessible evasion techniques.

The implications of these results extend beyond technical performance. Ethical concerns arise in the form of false positives and disproportionate risks for non-native English writers, who are more likely to be misclassified by less accurate tools. The variability observed across detectors underscores the lack of standardized evaluation metrics and raises concerns about equity and fairness when institutions rely on a single system for high-stakes decisions.

Taken together, the results suggest that robust tools such as Copyleaks, GPTZero, and Originality.ai can be valuable components of academic integrity practices, but they should not be used in isolation. Automated detection must be complemented with pedagogical strategies that emphasize process-oriented assessment, such as in-class writing, oral defenses, and iterative drafting, which offer alternative ways to authenticate student work. At the institutional level, transparent policies and safeguards are necessary to mitigate the risk of false positives and to ensure fair treatment of students, particularly those writing in a second language.

As shown in the Results and Discussion section, the data reveal two main risks: false positives and evasion through obfuscation. False positives, when human writing is incorrectly flagged as AI-generated, raise serious ethical concerns, particularly for NNES students. The accuracy drops observed for NNES-style texts (Table X) suggest that these detectors may unintentionally penalize linguistic diversity, raising concerns about fairness and potential bias in academic evaluation.

Evasion risk adds another layer of complexity to the educational use of AI detectors. The apparent decline in performance under paraphrasing and translation (Table VIII) shows how easily AI-generated texts can escape detection. This situation creates tension between maintaining academic integrity and supporting students who use language tools to improve their writing. Because of these challenges, educational institutions should treat detector results with caution and use them alongside other assessment methods, rather than relying entirely on automated tools.

While the study spans nine detectors, four LLMs, and three obfuscation types, the dataset scope (116 texts) constrains generalizability. Future work should expand text volumes, languages, and domains, and include additional/open-source detectors to triangulate performance and accessibility.

Future research should expand evaluation datasets, examine detector performance across additional languages and disciplines, and advance algorithms that can handle obfuscation while minimizing bias. Ultimately,

maintaining academic integrity in the era of generative AI requires not only technical solutions but also thoughtful integration of policy, pedagogy, and ethics.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

AM conducted the study and contributed through formal and statistical analyses, and by participating in the original draft and subsequent revisions; BS conducted experiments, contributed to the original draft, and participated in revisions; AS and AA were responsible for data curation, including dataset collection, and contributed to drafting the initial manuscript; they also participated in revisions; all authors had approved the final version.

FUNDING

This research was funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan within the framework of the project AP23487777.

REFERENCES

[1] K. Krishna, Y. Song, M. Karpinska *et al.* "Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 27469–27500, 2023.

[2] A. Akram, "An empirical study of AI-generated text detection tools," *Advances in Machine Learning & Artificial Intelligence*, vol. 4, no. 2, pp. 44–55, 2023.

[3] C. Chaka, "Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools," *Journal of Applied Learning and Teaching,* vol. 6, no. 2, 2023. doi: 10.37074/jalt.2023.6.2.12

[4] V. Liu and L. B. Chilton, "Design guidelines for prompt engineering text-to-image generative models," in *Proc. the 2022 CHI Conf. on Human Factors in Computing Systems*, 2022, pp. 1–23.

[5] A. Bhattacharjee, Y. Zeng, S. Y. Xu *et al.*, "Understanding the role of large language models in personalizing and scaffolding strategies to combat academic procrastination," in *Proc. the 2024 CHI Conf. on Human Factors in Computing Systems*, 2024, pp. 15–15

[6] L. Xiong, H. Wang, X. Chen *et al.* "DeepSeek: Paradigm shifts and technical evolution in large AI models," *IEEE/CAA J. Automatica Sinica*, vol. 12, no. 5, pp. 841–858, 2025.

[7] V. S. Sadasivan, A. Kumar, S. Balasubramanian *et al.* "Can AI-generated text be reliably detected?" arXiv preprint, arXiv:2303.11156, 2023.

[8] S. K. Kar, T. Bansal, S. Modi *et al.* "How sensitive are the free AI-detector tools in detecting AI-generated texts? A comparison of popular AI-detector tools," *Indian Journal of Psychological Medicine*, 2024. doi: 10.1177/02537176241247934

[9] N. A. Said, "Advancing news text classification: A comparative analysis of deep neural network models," *Journal of Advances in Information Technology*, vol. 16, no. 5, pp. 686–695, 2025. doi: 10.12720/jait.16.5.686-695

[10] A. Bhattacharjee and H. Liu, "Fighting fire with fire: Can ChatGPT detect AI-generated text?" *ACM SIGKDD Explorations Newsletter*, vol. 25, no. 2, pp. 14–21, 2024. doi: 10.1145/3655103.3655106

[11] H. Shimada and M. Kimura, "A method for distinguishing model generated text and human written text," *Journal of Advances in Information Technology*, vol. 15, no. 6, pp. 714–722, 2024. doi: 10.12720/jait.15.6.714-722

[12] A. Alduhailan, N. H. Kamarudin, S. N. H. S. Abdullah *et al.,* "Deep learning in biometric authentication: Challenges, recent advancements, and future trends," *Journal of Advances in Information Technology*, vol. 16, no. 4, pp. 458–477, 2025. doi: 10.12720/jait.16.4.458-477

[13] M. Perkins, J. Roe, D. Postma *et al.* "Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse," *Journal of Academic Ethics*, vol. 22, 2023. doi: 10.1007/s10805-023-09492-6

[14] A. M. Elkhatat, K. Elsaid, and S. Al-Meer, "Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text," *International Journal for Educational Integrity*, vol. 19, no. 1, 2023. doi: 10.1007/s40979-023-00140-5

[15] W. H. Walters, "The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors," *Open Information Science,* vol. 7, no. 1, 2023. doi: 10.1515/opis-2022-0158

[16] A. Amirzhanov, C. Turan, and A. Makhmutova, "Plagiarism types and detection methods: A systematic survey of algorithms in text analysis," *Frontiers in Computer Science*, vol. 7, 1504725, 2025. doi: 10.3389/fcomp.2025.1504725

[17] M. A. Malik and A. I. Amjad, "AI vs AI: How effective are Turnitin, ZeroGPT, GPTZero, and Writer AI in detecting text generated by ChatGPT, Perplexity, and Gemini?" *Journal of Applied Learning and Teaching*, vol. 8, no. 1, pp. 91–101, 2025.

[18] D. Weber-Wulff, A. Anohina-Naumeca, S. Bjelobaba *et al.*, "Testing of detection tools for AI-generated text," *International Journal for Educational Integrity*, 2023. doi: 10.1007/s40979-023-00146-z

[19] V. Fishchuk and D. Braun, "Robustness of generative AI detection: Adversarial attacks on black-box neural text detectors," *International Journal of Speech Technology*, vol. 27, no. 4, pp. 861–874, 2024. doi: 10.1007/s10772-024-10144-2

[20] M. Perkins, J. Roe, B. H. Vu *et al.* "Simple techniques to bypass GenAI text detectors: Implications for inclusive education," *International Journal of Educational Technology in Higher Education*, vol. 21, no. 1, 53, 2024. doi: 10.1186/s41239-024-00487-w

[21] S. Chakraborty, A. S. Bedi, S. Zhu *et al.* "On the possibilities of AI-generated text detection," arXiv preprint, arXiv:2304.04736, 2023.

[22] S. Rafiq, Q. U. Ain, and D. A. Afzal, "The role of AI detection tools in upholding academic integrity: An evaluation of their effectiveness," *Contemporary Journal of Social Science Review,* vol. 3, no. 1, pp. 901–915, 2025.

[23] A. K. Kadhim, L. Jiao, R. Shafik *et al.,* "Adversarial attacks on AI-generated text detection models: A token probability-based approach using embeddings," arXiv preprint, arXiv:2501.18998, 2025.

[24] J. Gundersen, "AI vs. AI: Exploring synthetic text detection of large language models in a low-resource language," M.S. thesis, Dept. Info., Sec. and Comm. Tech., Norwegian Univ., of Sci. and Tech., 2024.

[25] W. H. Pan, M. J. Chok, J. L. S. Wong *et al.,* "Assessing AI detectors in identifying AI-generated code: Implications for education," in *Proc. the 46th International Conf. on Software Engineering: Software Engineering Education and Training,* 2024, pp. 1–11.

[26] B. Sharimbayev. (2025). AI-generated vs. human-written text dataset (Version 1) [Dataset]. *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/hardkazakh/ai-generated-vs-human-written-text-dataset