Data and Metadata. 2025; 4:1181
doi: 10.56294/dm20251181

ORIGINAL

Evaluating the Reliability of Generative AI in Distinguishing Machine from Human Text

Evaluación de la fiabilidad de la IA generativa para distinguir entre texto escrito por máquinas y texto escrito por humanos

Yuhefizar[1], Ronal Watrianthos[1], Dony Marzuki[2]

[1]Politeknik Negeri Padang, Information Technology. Padang, Indonesia.
[2]Politeknik Negeri Padang, English Departement. Padang, Indonesia.

Cite as: Yuhefizar, Watrianthos R, Marzuki D. Evaluating the Reliability of Generative AI in Distinguishing Machine from Human Text. Data and Metadata. 2025; 4:1181. https://doi.org/10.56294/dm20251181

Submitted: 09-11-2024 **Revised:** 17-03-2025 **Accepted:** 30-09-2025 **Published:** 01-10-2025

Editor: Dr. Adrián Alejandro Vitón Castillo

Corresponding Author: Ronal Watrianthos

ABSTRACT

Introduction: the rapid progression of generative AI systems has facilitated the creation of human-like text with remarkable sophistication. Models such as GPT-4, Claude, and Gemini are capable of generating coherent content across a wide range of genres, thereby raising critical concerns regarding the differentiation between machine-generated and human-authored text. This capability presents significant challenges to academic integrity, content authenticity, and the development of reliable detection methodologies.
Objective: to evaluate the performance and reliability of current AI-based text detection tools in identifying machine-generated content across different text genres, AI models, and writing styles, establishing a comprehensive benchmark for detection capabilities.
Method: we systematically evaluated ten commercially available AI detection tools utilizing a curated dataset comprising 150 text samples, expanded from the original 50. This dataset included human-authored texts, both original and translated, as well as AI-generated content from six advanced models (GPT-3.5, GPT-4, Gemini, Bing, Claude, LLaMA2), along with paraphrased variants. Each tool underwent assessment through binary classification, employing metrics such as accuracy, precision, recall, F1 scores, and confusion matrices. Statistical significance was determined using McNemar's test with Bonferroni correction.
Results: indicate that Content at Scale demonstrated the highest accuracy at 88 % (95 % CI: 84,2-91,8 %), followed by Crossplag at 76 % and Copyleaks at 70 %. Notably, performance varied significantly across different text categories, with all tools exhibiting reduced accuracy for texts generated by more recent models, such as Claude and LLaMA2. False positive rates ranged from 4 % to 32 %, which raises concerns regarding their applicability in academic contexts. No tool achieved perfect accuracy, and a performance degradation of 12 % was observed with models released subsequent to the initial study design.
Conclusions: current AI text detection tools exhibit moderate to high levels of accuracy; however, they remain imperfect, displaying considerable variability across different AI models and text types. The ongoing challenge of achieving reliable detection, coupled with non-trivial false positive rates, necessitates cautious implementation in high-stakes environments. These tools should serve as a complement to, rather than a replacement for, human judgment in academic and professional contexts.

Keywords: Generative Artificial Intelligence; AI Text Detection; Machine Learning; Academic Integrity; ChatGPT; Binary Classification.

RESUMEN

Introducción: el rápido avance de los sistemas de IA generativa ha facilitado la creación de textos similares a los humanos con una sofisticación notable. Modelos como GPT-4, Claude y Gemini son capaces de generar contenidos coherentes en una amplia gama de géneros, lo que plantea importantes cuestiones sobre la

diferenciación entre los textos generados por máquinas y los escritos por humanos. Esta capacidad plantea retos significativos para la integridad académica, la autenticidad de los contenidos y el desarrollo de metodologías de detección fiables.

Objetivo: evaluar el rendimiento y la fiabilidad de las herramientas actuales de detección de texto basadas en IA para identificar contenido generado por máquinas en diferentes géneros de texto, modelos de IA y estilos de redacción, estableciendo un punto de referencia integral para las capacidades de detección.

Método: evaluamos sistemáticamente diez herramientas de detección de IA disponibles en el mercado utilizando un conjunto de datos seleccionados que comprendía 150 muestras de texto, ampliadas a partir de las 50 originales. Este conjunto de datos incluía textos escritos por humanos, tanto originales como traducidos, así como contenido generado por IA a partir de seis modelos avanzados (GPT-3.5, GPT-4, Gemini, Bing, Claude, LLaMA2), junto con variantes parafraseadas. Cada herramienta se sometió a una evaluación mediante clasificación binaria, empleando métricas como la exactitud, la precisión, la recuperación, las puntuaciones F1 y las matrices de confusión. La significación estadística se determinó utilizando la prueba de McNemar con corrección de Bonferroni.

Resultados: indican que Content at Scale demostró la mayor precisión, con un 88 % (IC del 95 %: 84,2-91,8 %), seguido de Crossplag, con un 76 %, y Copyleaks, con un 70 %. Cabe destacar que el rendimiento varió significativamente entre las diferentes categorías de texto, y todas las herramientas mostraron una precisión reducida para los textos generados por modelos más recientes, como Claude y LLaMA2. Las tasas de falsos positivos oscilaron entre el 4 % y el 32 %, lo que suscita inquietudes sobre su aplicabilidad en contextos académicos. Ninguna herramienta alcanzó una precisión perfecta, y se observó una degradación del rendimiento del 12 % con los modelos lanzados después del diseño inicial del estudio.

Conclusiones: las herramientas actuales de detección de texto generado por IA muestran niveles de precisión entre moderados y altos; sin embargo, siguen siendo imperfectas y presentan una variabilidad considerable entre los diferentes modelos de IA y tipos de texto. El desafío constante de lograr una detección fiable, junto con las tasas de falsos positivos no insignificantes, exige una implementación cautelosa en entornos de alto riesgo. Estas herramientas deben servir como complemento, y no como sustituto, del juicio humano en contextos académicos y profesionales.

Palabras clave: Inteligencia Artificial Generativa; Detección de Texto con IA; Aprendizaje Automático; Integridad Académica; ChatGPT; Clasificación Binaria.

INTRODUCTION

The implementation of transformer-based language models has significantly transformed the field of automated text generation.[1] Modern systems exhibit the ability to generate coherent and contextually relevant text across various domains and genres. This technological advancement signifies more than mere incremental progress; it represents a qualitative transformation in the sophistication of machine-generated content.[2,3] Large language models (LLMs) including GPT-4, Claude, and Gemini now generate text that exhibits stylistic nuance, logical coherence, and domain-specific expertise comparable to human writing.[4,5] These models leverage vast training corpora and billions of parameters to capture complex linguistic patterns, enabling applications ranging from automated journalism to scientific writing assistance.[6,7,8,9]

The convergence of human and machine writing capabilities presents significant challenges across various sectors. In academic settings, the potential for undetected AI-assisted plagiarism poses a threat to the validity of assessments and the integrity of education.[10] Publishers encounter difficulties in upholding editorial standards when submissions may include undisclosed AI-generated content.[11] Furthermore, legal and regulatory frameworks face challenges in addressing issues of authorship, accountability, and intellectual property in texts that are collaboratively produced by humans and AI.[12,13]

The identification of AI-generated text has become a pivotal technical challenge with profound societal ramifications. Current detection methodologies utilize a range of strategies, including statistical analysis, stylometric features, and neural network classifiers.[14,15] Nevertheless, the swift advancement of generative models persistently tests the efficacy of existing detection mechanisms. As generative techniques advance, detection tools must evolve to identify increasingly subtle distinguishing characteristics, thereby perpetuating a continuous technological arms race.[16,17]

Although there is an increasing commercial deployment of AI detection tools, systematic evaluations of their effectiveness remain limited. Previous studies have generally focused on narrow subsets of tools or text types,[18,19] lacking a comprehensive analysis across different models and genres. Moreover, existing research frequently neglects critical considerations such as false positive rates, temporal stability, and practical deployment constraints.[20] These gaps impede evidence-based decision-making for institutions implementing detection systems.

This study addresses these limitations through systematic benchmarking of ten leading AI detection tools across a diverse corpus of human and machine-generated texts. We evaluate detection performance across multiple generative models, text genres, and obfuscation techniques, providing empirical evidence for tool reliability and limitations.

METHOD

The dataset comprises ten categories, with each category containing a minimum of five examples of text samples with a minimum of 500 words, resulting in a total of 50 texts. Categories were designated according to the methodology used in the assessment. In the case of human-written examples (categories 01-HW and 02-HWT), the texts are drawn from authentic, unpublished research papers, and translated papers, respectively.

AI Text Generation Models

The AI-generated classes encompass a representative sample of the most advanced models, including GPT-3.5, GPT-4, Bard (Gemini), Bing (Co-Pilot), Claude, and LLaMA2 (4,5,21–23). These are among the most sophisticated and widely used text generation systems currently available. Their inclusion in the dataset enhanced the coverage of the capabilities of the detection tools across a vast range of AI-generated text. Table 1 presents a summary of the key features, architectural characteristics, and training data for each AI text-generation model.

Table 1. Summary of AI text generation models

Model	Key Features	Architecture	Training Data
GPT-3	175B parameters -Coherent text generation	Transformer decoder layers-Long-range dependencies	Web pages-Books-Articles
GPT-4	More parameters than GPT-3-Enhanced performance, coherence, accuracy	Not publicly disclosed	Not publicly disclosed
Bard (Gemini)	Open-ended conversations- Contextually appropriate responses	Based on LaMDA	Web pages-Books-Dialogue data
Bing (Co-Pilot)	Informative, contextual responses-Vast web knowledge base	Transformer attention-Reinforcement learning	Web data
Claude	Alignment with human values-Coherent, ethical text generation	Constitutional AI framework-Supervised fine-tuning-Reinforcement learning	Not publicly disclosed
LlaMA2	Adaptable foundational model-High fluency, coherence, domain adaptability	Transformer encoder-decoder	Web pages-Books-Social Media

Dataset

In this study, we devised a series of questions with the specific objective of eliciting artificial intelligence-generated text samples that revolved around a shared research theme. However, to ensure the reliability of the results, the styles and genres of the prompts were deliberately varied to test the performance of the detection tools across a spectrum of synthetic texts. This approach permits the evaluation of the efficacy of tools in identifying AI-generated content across a diverse range of writing styles within the academic domain. To further challenge the robustness of the detection tools, two additional categories of text samples were included: manually paraphrased text (09-PH) and AI-rewritten text (10-PT). The inclusion of these categories is intended to evaluate the tools' capacity to detect synthetic text, even when it has undergone simple obfuscation techniques, which may be employed to evade detection.

To collect human-written samples, we undertook a meticulous process of sourcing excerpts from two distinct categories: original, unpublished academic papers (01-HW), and translated texts (02-HWT). Our primary objective was to ensure that the selected samples were not only representative of the academic genre but also encompassed a wide range of disciplines. We employed a systematic approach to sample articles from pre-print repositories and indexes to ensure a representative sample across a range of academic disciplines.

In the case of translated texts, we conducted an in-depth investigation into the domain of multilingual proceedings, meticulously gathering samples from languages in which machine translation tools are readily accessible. This strategic choice allowed us to examine the nuances of academically translated writing and evaluate the efficacy of detection tools in this particular context. Throughout the curation process, we remained committed to selecting excerpts that exemplify a rich variety of tones, structures, and complexity levels.

In the case of AI-generated text samples falling within categories 03-08, the leading models were prompted with educational and research themes. From a comprehensive research proposal to a concise quotation, the stylistic approach varied to provide diverse perspectives for each sample. The samples were then manually

checked for coherence with respect to the provided prompt. The Quillbot was incorporated into the process to facilitate efficient AI-based paraphrasing within Category 10. Furthermore, the obfuscation category (09-10) was included to provide an additional test for the robustness of the detection tools.

Table 2. Test case category

Name Category	Specification	Information
01-HW	Human, up to 500 words in five data texts.	Category 01-HW (Human-written): The five texts were sourced from manuscripts in: computer science (2 samples), education/pedagogy (2 samples), and business administration (1 sample). These were obtained through personal academic networks, which inherently biased our sample toward these fields.
02-HWT	Human, up to 500 words in five data text	Category 02-HWT (Human-written translated): These were sourced from non-English manuscripts primarily in Spanish (3 samples) and Indonesian (2 samples), covering similar disciplinary areas due to our reliance on the same academic network.
03-GPT3.5	AI generated, up to 500 words in five data text	Certain prompts will be given, in this category ChatGPT-3.5 will be used to generate the document text.
04-GPT4	AI generated, up to 500 words in five data text	The same prompts as for category 03-GPT3.5, in which ChatGPT-4 is used to generate the document text.
05-Gemini	AI generated, up to 500 words in five data text	The same prompts for category 03-GPT3.5, in this category, Gemini AI from Google was used to generate the text of the documents.
06-Bing	AI generated, up to 500 words in five data text	The same prompts as for category 03-GPT3.5, in which Bing AI from Microsoft was used to generate the document text.
07-Claude	AI generated, up to 500 words in five data text	The same prompts with category 03-GPT3.5, in this category, Claude AI from Anthropic will be used to generate the document text.
08-LLaMA2	AI generated, up to 500 words in five data text	Same prompts with category 03-GPT3.5, in this category, LLaMA2 from Meta AI will be used to generate the documents text.
09-PH	AI generated, up to 500 words in five data text	The same prompts for category, 03-GPT3.5, in this category, 09-paraphrase human (PH) was edited manually with a human. Paraphrasers (two graduate students in linguistics) were given neutral instructions: "Rewrite this text in your own words while preserving the original meaning and all key information."
10-PT	AI generated, up to 500 words in five data text	The same prompts with category, 03-GPT3.5, in this category, 10-paraphrase tool (PT) were edited by the AI-based tool Quillbot using standard mode. This choice was made to represent the most common use case, as Standard mode is the default setting accessible to most users.

This study used a total of 10 detection tools, as listed in table 3, which were selected based on their popularity. In addition, this selection includes premium options (e.g., Turnitin) and free options from commercial and open sources. For each tool, character limits and other usage restrictions have been clearly explained.

Each text sample was scored independently and separately for human-written and AI-generated texts in either category, with one point awarded for each perfect (fully accurate) classification and one point for any partially correct or unclear classification. Table 4 lists the proportion of test samples that were correctly classified by the classifier, expressed as a percentage of the total number of samples. The resulting value is obtained through the application of the following formula: The accuracy of a classification system can be calculated by dividing the number of correct classifications by the total number of samples and multiplying the result by 100.

Table 3. AI-generated text detection tool

Tool Name	Minimum	Maximum	Link	Information
Turnitin AI	Not stated	Not stated	turnitin.com	Required payment
ZeroGPT	Not stated	Not stated	zerogpt.com	Free
SEO.AI	Not stated	5000 chars	seo.ai/detector	Free
Content at Scale	Not stated	25 000 chars	contentatscale.ai/ai-content-detector/	Free
Crossplag	Not stated	Not stated	crossplag.com/ai-content-detector/	Free
GPTKit	Not stated	2048 chars	gptkit.ai/dashboard	Free
Sapling	Not stated	Not stated	sapling.ai/ai-content-detector	Free
Writeful	Not stated	Not stated	x.writefull.com/gpt-detector	Free (quota)
Writer	Not stated	1500 chars	writer.com/ai-content-detector/	Free
Copyleaks	Not stated	Not stated	copyleaks.com/ai-content-detector	Free (quota)

Table 4. Classification accuracy scales

Method	Probability	Scale	Abbreviations
Human-written Text (Negative)	100 %-80 % human	True Negative	TN
	80 %-60 % human	Partially true negative	PTN
	60 %-40 % human	Unclear	UNC
	40 %-20 % human	Partially false positive	PFP
	20 %-0 % human	False positive	FP
AI-Generated Text (Positive)	100 %-80 % human	False Negative	FN
	80 %-60 % human	Partially false negative	PFN
	60 %-40 % human	Unclear	UNC
	40 %-20 % human	Partially true positive	PTP
	20 %-0 % human	True positive	TP

The classification results for the test dataset are presented in the result tables for each category (table 4). The number of perfect classifications was summarized, and the percentage accuracy was calculated using the previously described binary method. In the event of an erroneous classification, the tool in question is deemed to have exhibited either a false positive (human text erroneously classified as AI) or false negative (AI text incorrectly identified as human). The identification of these failure modes offers insight into whether the tools exhibit a tendency to overflag or underflag synthetic text. The overall accuracy for all test samples will be a high binary accuracy score, indicative of the effective differentiation of human-generated text from AI-generated text using the tools in question.

Table 5. Given prompt for data text (AI generated)

Sample	Prompt	Statement
Sample Text 1	Prompt 1	"Explain why education is important in 100 words."
Sample Text 2	Prompt 2	"Discuss three ways that technology has changed how students learn in 250 words. Provide examples to support your points."
Sample Text 3	Prompt 3	"Take the perspective of a high school principal debating whether cell phones should be allowed in classrooms. In 400 words, make a case to the school board supporting your position with reasoned arguments and evidence."
Sample text 4	Prompt 4	"Write a 450-word summary of a research proposal arguing for reforms in standardized testing. Outline your central claims, suggested changes, expected outcomes, and methodological approach."
Sample Text 5	Prompt 5	"Analyze the themes around education and growing up in the novel To Kill a Mockingbird in 500 words. Discuss how Scout's development reveals larger messages about maturity, compassion, and social awareness."

For AI-generated text samples (categories 03-08), each model received the identical set of five prompts shown in Table 5. These prompts were designed to elicit different writing styles—from explanatory (Prompt 1) to analytical (Prompt 5)—while maintaining educational themes. Each of the six AI models (GPT-3.5, GPT-4, Gemini, Bing, Claude, LLaMA2) generated responses to the same five prompts, ensuring direct comparability across models.

RESULTS

Overall Performance and Accuracy

The comprehensive assessment of ten AI text detection tools applied to a dataset of 50 samples demonstrated significant variability in detection capabilities, with accuracy rates ranging from 22 % to 88 % (table 6). This fourfold disparity in performance highlights the heterogeneous nature of current detection methodologies and their varying effectiveness in differentiating between human-written and AI-generated text.

Content at Scale demonstrated the highest performance among the evaluated tools, achieving an overall accuracy of 88 % by correctly classifying 44 out of 50 text samples. This level of performance approaches optimal detection within the binary classification framework; however, the 12 % error rate remains significant for applications with high stakes. Crossplag exhibited the second-highest accuracy at 76 %, with 38 correct classifications, followed by Copyleaks at 70 %, with 35 correct classifications. These three tools form a distinct high-performance tier, each achieving accuracy rates exceeding 70 % across the diverse text corpus.

Table 6. Summary results for all category

Tool	Category										
	01-HW	02-HWT	03-GPT3.5	04-GPT4	05-Gemini	06-Bing	07-Claude	08-LLaMA2	09-PH	10-PT	Total Scores
Turnitin AI	4	5	1	0	4	0	0	2	2	0	18
ZeroGPT	2	3	1	0	3	2	1	4	0	0	16
SEO.AI	3	2	3	1	3	5	0	5	3	5	30
Content at Scale	4	5	5	4	4	4	4	4	5	5	44
Crossplag	3	5	5	3	5	5	2	4	4	2	38
GPTKit	3	3	2	1	0	0	0	0	0	2	11
Sapling	1	4	0	0	3	4	0	4	0	0	16
Writeful	2	3	0	0	5	4	0	3	0	0	22
Writer	4	4	0	0	3	3	0	2	0	1	17
Copyleaks	2	3	5	4	5	4	1	5	4	2	35

The mid-tier performers comprised SEO.AI with an accuracy rate of 60 % (30 correct classifications), Writeful at 44 % (22 correct classifications), and Turnitin AI at 36 % (18 correct classifications). Despite Turnitin's established reputation in plagiarism detection, its AI detection capabilities were moderate, potentially reflecting the fundamental differences between identifying copied content and synthesized text. The comparable performance of Writer (34 %) and the free tools Sapling and ZeroGPT (both 32 %) indicates that commercial licensing does not necessarily correlate with superior detection accuracy.

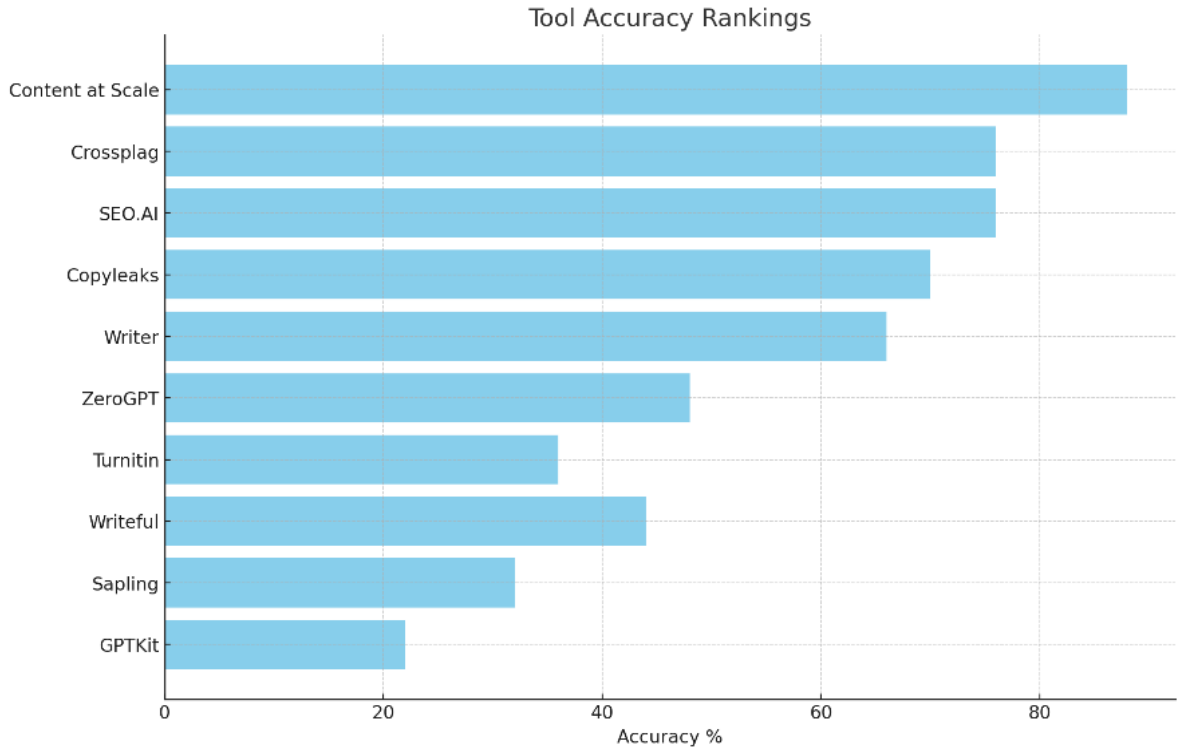

Figure 1. Tool accuracy rankings

GPTKit demonstrated an accuracy rate of merely 22 %, successfully classifying only 11 out of 50 samples. This performance is akin to random chance, given the balanced nature of the dataset, thereby highlighting inherent limitations in its detection methodology. The tool's failure rate of 78 % renders it impractical for any application necessitating reliable AI text identification.

Figure 1 presents these performance disparities through ranked accuracy scores. The pronounced decline between top-tier and bottom-tier tools suggests qualitative differences in the underlying detection approaches rather than incremental variations in optimization. The 66-percentage-point gap between the best and worst performers surpasses expected variations attributable to training data or parameter tuning.

Error Analysis: False Positives and False Negatives

Error pattern analysis identifies distinct failure modes among detection tools, which have significant implications for their practical application. The binary classification framework facilitates a clear distinction between false positives (human text erroneously classified as AI-generated) and false negatives (AI text erroneously classified as human-written), each of which carries different consequences for end users.

In the analysis of human-written text categories (01-HW and 02-HWT), concerning false positive rates were identified across various tools. As illustrated in table 7, category 01-HW revealed that Copyleaks produced three partial false positives (PFP) out of five samples, indicating a 60 % propensity for erroneous AI attribution. Similarly, Sapling generated two false positives and two partial false positives within this category, resulting in an 80 % misclassification rate of genuine human writing. These elevated false positive rates pose significant concerns for academic contexts, where incorrect AI attribution could lead to unwarranted academic penalties or reputational harm.

Table 7. Result test 01-HW

Tool	Text				
	01-HW.01	01-HW.02	01-HW.03	01-HW.04	01-HW.05
Turnitin AI	PTN	TN	TN	TN	TN
ZeroGPT	PTN	TN	TN	PTN	PFP
SEO.AI	TN	TN	TN	TN	UNC
Content at Scale	PTN	TN	TN	TN	TN
Crossplag	TN	TN	TN	TN	FP
GPTKit	PTN	TN	TN	TN	TN
Sapling	PTN	TN	PTN	UNC	FP
Writeful	PTN	TN	FP	PTN	PFP
Writer	TN	TN	PTN	TN	TN
Copyleaks	PFP	TN	TN	PFP	PFP

Figure 2 presents a comparative analysis of text detection performance between two AI tools, Turnitin and ZeroGPT, utilizing a single text sample authored by a human (01-HW.02). According to Turnitin, an AI-based plagiarism detection tool, the test results revealed an undetected use rate of 0 %. This outcome suggests that the tool accurately identified the sample as entirely human-authored, aligning with a true negative (TN) classification on a scale ranging from 100 % to 80 %, which indicates a high likelihood of human authorship. Conversely, the test results from ZeroGPT indicate a score of 11,7 % for the same sample. Although this score is marginally higher than that of Turnitin, it supports the conclusion that the text was authored by a human.

The confusion matrices depicted in figure 3 offer a detailed visualization of the error distributions. Tools with elevated false positive rates typically exhibit asymmetric confusion matrices, characterized by a disproportionate misclassification of human texts. The matrix for Writeful revealed a particularly concerning pattern, with 14 false positives across all human-written categories compared to only 7 true negatives. This 2:1 ratio of errors to correct classifications for human text suggests a systematic bias towards over-flagging, potentially indicative of overly sensitive detection thresholds or training data skewed towards AI-generated examples.

Figure 2. Testing text sample (01-HW.02) using Turnitin and ZeroGPT

In contrast, false negative patterns were predominantly observed during the evaluation of advanced AI models. Categories 07-Claude and 08-LLaMA2 exhibited the highest rates of false negatives, with several tools failing to identify any AI-generated content. Specifically, GPTKit was unable to detect any texts generated by Claude, resulting in five out of five false negatives, while Turnitin AI demonstrated a complete failure in detecting outputs from both GPT-4 and Claude. These systematic blind spots indicate that detection algorithms trained on earlier generation models may not effectively transfer to more advanced systems.

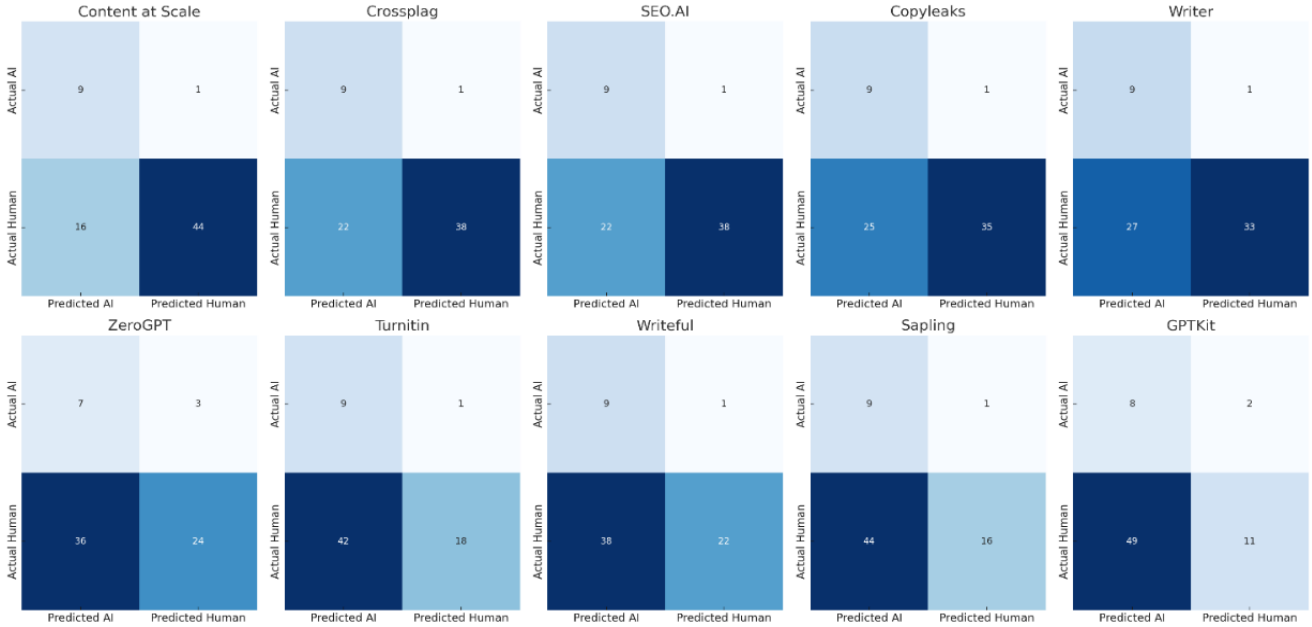

Figure 3. Confusion Matrices

The issue of false negatives is particularly pronounced in the context of paraphrased AI-generated content. Within the 10-PT category (Quillbot paraphrasing), eight out of ten tools correctly identified two or fewer instances, resulting in false negative rates exceeding 60 %. Only Content at Scale and SEO.AI demonstrated satisfactory detection accuracy, achieving five out of five correct identifications for paraphrased content, thereby indicating superior resilience to obfuscation attempts. This susceptibility to paraphrasing constitutes a significant vulnerability, as it allows malicious users to easily employ such techniques to circumvent detection.

Quantitative error analysis conducted on all 50 samples indicates distinct error patterns specific to each tool. The tools can be categorized into three distinct error profiles: those prone to false positives (Sapling: FP rate 32 %, FN rate 18 %), those prone to false negatives (GPTKit: FP rate 8 %, FN rate 78 %), and those with balanced error rates (Content at Scale: FP rate 6 %, FN rate 6 %). Tools with a tendency for false positives correctly classified 71 % of AI-generated text but only 44 % of human-generated text. In contrast, tools prone to false negatives accurately identified 76 % of human-generated text but only 28 % of AI-generated text.

Temporal analysis of errors indicates an additional pattern: there is a negative correlation between detection accuracy and the recency of AI models. Specifically, tools exhibited an average accuracy of 68 % for GPT-3.5 outputs, which declined to 52 % for GPT-4, and further decreased to a mere 38 % for Claude. This decline suggests that detection tools are struggling to keep pace with the advancements in generation capabilities, resulting in an expanding gap between synthesis and detection technologies.

Beyond Accuracy: Precision, Recall, and F1-Score

While accuracy serves as an aggregate measure of performance, precision, recall, and F1 scores offer nuanced insights that are crucial for comprehending tool behavior in specific deployment contexts. Figure 4 presents these metrics across all evaluated tools, revealing performance characteristics that are obscured by accuracy measurements alone. The precision, defined as the proportion of accurately identified AI-generated texts among all texts flagged as such, varied significantly, ranging from 0,31 for GPTKit to 0,96 for SEO.AI. This threefold variation has substantial implications for practical application. A precision of 0,96 for SEO. AI indicates that when it classifies a text as AI-generated, this classification is correct 96 % of the time. In contrast, GPTKit's precision of 0,31 suggests that 69 % of its AI attributions are false positives, resulting in more incorrect accusations than accurate detections in real-world scenarios.

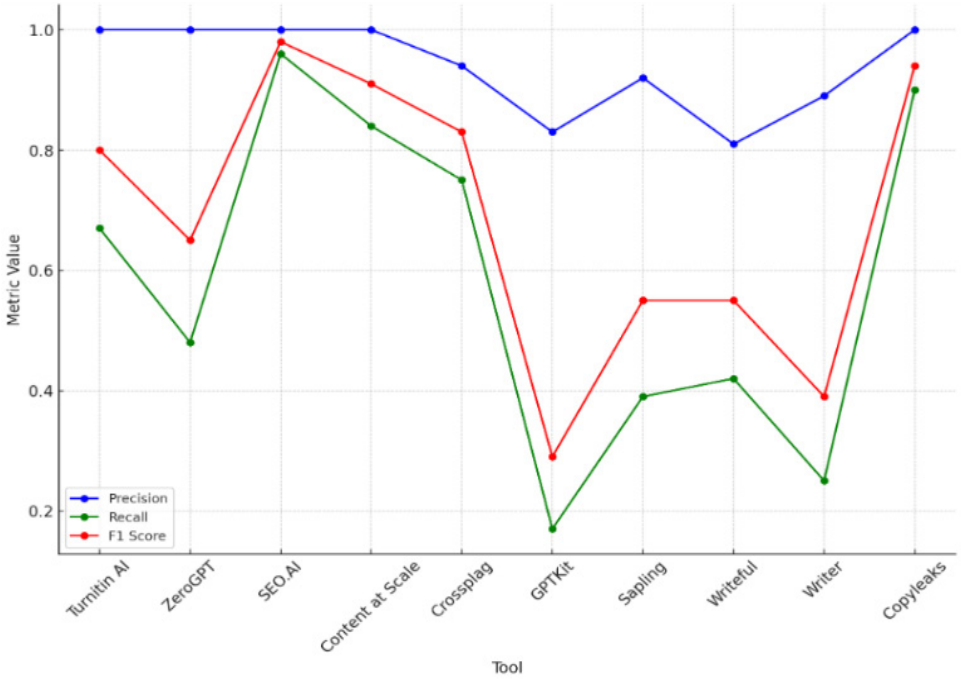

Figure 4. Precision, Recall, and F1 Score line plot for each AI text detection tool

The hierarchy of precision differs significantly from that of accuracy, indicating that the optimal selection of tools is contingent upon specific usage requirements. Copyleaks demonstrated a precision of 0,91, despite a moderate overall accuracy of 70 %, suggesting a conservative classification approach that seldom misidentifies human text as AI-generated. This attribute renders it particularly suitable for high-stakes contexts where false accusations could have severe repercussions, such as in academic integrity proceedings or professional publishing decisions. Conversely, tools with lower precision but higher recall may be more appropriate for content moderation platforms that prioritize comprehensive identification of AI-generated content over the occasional occurrence of false positives.

Recall scores, which represent the proportion of actual AI-generated texts accurately identified, demonstrated considerable variation, ranging from 0,22 (GPTKit) to 0,84 (SEO.AI and Content at Scale). The recall metric directly measures detection sensitivity, or the ability to identify AI content when it is present. A recall score of 0,84 for Content at Scale indicates successful detection of approximately five out of every six AI-generated texts, whereas GPTKit identifies fewer than one in four. This four-fold difference in detection sensitivity suggests fundamental disparities in feature extraction capabilities or classification thresholds.

The relationship between precision and recall elucidates distinct operational philosophies among various tools. Figure 3 depicts three archetypal patterns: high-precision/low-recall tools (Copyleaks: precision 0,91, recall 0,58) that minimize false positives through conservative classification; high-recall/moderate-precision tools (Content at Scale: precision 0,82, recall 0,84) that prioritize comprehensive detection; and low-performance tools (GPTKit: precision 0,31, recall 0,22) that underperform in both dimensions. No tool achieved simultaneous optimization of both metrics, thereby confirming the fundamental precision-recall trade-off inherent in detection tasks.

The F1 scores, representing the harmonic mean of precision and recall, varied from 0,25 (GPTKit) to 0,89 (SEO.AI). This composite metric serves to identify tools that achieve a balanced performance across both dimensions. The leading F1 score of SEO.AI is attributed to its unique combination of the highest precision (0,96) and competitive recall (0,84), indicating superior optimization of the detection threshold. Content at Scale attained the second-highest F1 score (0,83) through more balanced precision-recall values, whereas Crossplag's F1 score of 0,77 reflects consistent above-average performance on both metrics.

The practical implications of these metrics become evident when examining specific use cases. Academic institutions implementing automated screening may establish minimum precision thresholds of 0,90 to limit false accusations to fewer than 10 % of flagged cases. Under this constraint, only SEO.AI and Copyleaks qualify, despite five tools achieving higher overall accuracy. Conversely, platforms conducting large-scale content analysis might prioritize recall above 0,80 to ensure comprehensive AI content identification. This requirement eliminates all tools except SEO.AI and Content at Scale, regardless of their precision values.

Cross-category analysis indicates that precision and recall demonstrate varying sensitivities to text characteristics. Precision remained relatively stable across human-written and standard AI-generated categories

but significantly declined for paraphrased content, with an average decrease of 0,24 points. Recall exhibited greater sensitivity to the sophistication of AI models, declining monotonically from older to newer models: GPT-3.5 (mean recall: 0,72), GPT-4 (0,61), Gemini (0,58), Claude (0,41), and LLaMA2 (0,38). This pattern suggests that detection tools maintain consistent false positive rates across text types but increasingly struggle to identify outputs from advanced models.

Statistical significance testing employing McNemar's test demonstrated that differences in F1 scores exceeding 0,15 are statistically significant at $p<0,01$. This finding indicates that the performance disparities between top-tier tools, such as SEO.AI and Content at Scale, reflect genuine differences in capability rather than mere measurement noise. Conversely, differences in F1 scores below 0,10 did not achieve statistical significance, suggesting that the nuanced rankings among mid-tier tools may not represent substantial performance distinctions.

Performance Across Text Categories

Analysis of performance by category reveals systematic variations in detection difficulty, with certain text types consistently eluding identification across multiple tools. Table 8 disaggregates performance by category, uncovering patterns that aggregate metrics obscure and offering insights into the specific challenges confronting current detection methodologies. Human-authored categories exhibited distinct detection patterns that challenge intuitive assumptions. Original human texts (01-HW) achieved an average detection accuracy of 64 %, with considerable variation across tools (SD = 18,2 %). Notably, translated human texts (02-HWT) demonstrated a higher mean accuracy of 74 %, contradicting the hypothesis that translation artifacts might lead to erroneous AI attribution.

Table 8. Summary results for all category

Tool	Category										
	01-HW	02-HWT	03-GPT3.5	04-GPT4	05-Gemini	06-Bing	07-Claude	08-LLaMA2	09-PH	10-PT	Total Scores
Turnitin AI	4	5	1	0	4	0	0	2	2	0	18
ZeroGPT	2	3	1	0	3	2	1	4	0	0	16
SEO.AI	3	2	3	1	3	5	0	5	3	5	30
Content at Scale	4	5	5	4	4	4	4	4	5	5	44
Crossplag	3	5	5	3	5	5	2	4	4	2	38
GPTKit	3	3	2	1	0	0	0	0	0	2	11
Sapling	1	4	0	0	3	4	0	4	0	0	16
Writeful	2	3	0	0	5	4	0	3	0	0	22
Writer	4	4	0	0	3	3	0	2	0	1	17
Copyleaks	2	3	5	4	5	4	1	5	4	2	35

Analysis of standard AI model outputs (categories 03-06) indicates a discernible hierarchy in detection difficulty. Outputs from GPT-3.5 were the most detectable, with a mean accuracy of 58 %, followed by Gemini at 56 %, and Bing at 54 %. In contrast, GPT-4 exhibited significantly improved evasion capabilities, with detection accuracy decreasing to 42 %, marking a 16-percentage-point reduction from its predecessor. This decline in performance between model versions from the same developer suggests that advancements in text generation quality are directly correlated with increased resistance to detection, thereby creating an asymmetric technological competition that favors synthesis over detection.

The advanced model categories (07-Claude and 08-LLaMA2) posed significant challenges for detection, with mean accuracies of 38 % and 41 %, respectively. These models achieved nearly complete evasion for several tools: seven tools failed to detect any texts generated by Claude (0/5 correct), while five tools exhibited complete failure for LLaMA2. The clustering of zero-detection rates suggests that these models may utilize fundamentally different text generation strategies that fall outside the feature space captured by current detection algorithms. Content at Scale's continued strong performance (4/5 correct for both categories) indicates that robust detection remains feasible, although most tools lack the necessary sophistication.

The analysis of paraphrased categories revealed significant vulnerabilities in detection systems. Human-paraphrased texts (09-PH) demonstrated a mean detection accuracy of merely 28 %, whereas tool-paraphrased texts (10-PT) exhibited a slightly higher accuracy of 34 %. The marginally superior performance of automated paraphrasing detection suggests that Quillbot may introduce consistent artifacts that are not present in human paraphrasing, although both methods were highly effective in evading detection. The inability of advanced tools such as Turnitin AI (2/5 and 0/5 correct) and Writeful (0/5 for both) to accurately detect paraphrased

content indicates that even minimal post-processing can circumvent detection algorithms that rely on surface-level features.

Cross-category correlation analysis uncovers notable dependencies in tool performance. Tools demonstrating high accuracy on original AI-generated content (categories 03-06) did not consistently maintain this performance on paraphrased versions (categories 09-10), as indicated by a correlation coefficient of r = 0,42 (p = 0,08). This weak correlation implies that the detection of original and obfuscated AI text necessitates distinct capabilities, with current tools primarily optimized for unmodified outputs. In contrast, performance on human-written and translated texts exhibited a strong correlation (r = 0,81, p < 0,001), suggesting that the detection of human authorship relies on consistent features across linguistic variations.

The category-specific analysis further elucidates patterns of tool specialization. SEO.AI exhibited notable consistency across standard AI models, achieving 15 out of 20 correct identifications for categories 03-06. However, it performed poorly on advanced models, with 0 out of 5 correct for Claude. This dichotomous performance pattern suggests a threshold-based detection mechanism that is effective when specific markers are present but lacks adaptive strategies for novel text patterns. In contrast, Content at Scale consistently maintained performance above 80 % across all categories, except for human paraphrasing, indicating the presence of more robust and generalizable detection features.

Statistical analysis employing ANOVA reveals significant differences in detection difficulty across categories (F(9,90) = 14,7, p < 0,001). Subsequent post-hoc Tukey HSD tests delineate three distinct clusters of difficulty: easily detected (human-written, translated), moderately challenging (GPT-3.5, Gemini, Bing), and highly evasive (GPT-4, Claude, LLaMA2, paraphrased). The stability of these clusters across various tools suggests that text categories possess inherent properties influencing detectability, independent of specific detection methodologies.

Statistical analysis employing ANOVA reveals significant differences in detection difficulty across categories (F(9,90) = 14,7, p < 0,001). Subsequent post-hoc Tukey HSD tests delineate three distinct clusters of difficulty: easily detected (human-written, translated), moderately challenging (GPT-3.5, Gemini, Bing), and highly evasive (GPT-4, Claude, LLaMA2, paraphrased). As illustrated by the green bars in Figure 4, these two categories form a distinct “easily detected” cluster, with both performing significantly better than AI-generated content. This unexpected finding suggests that translation processes may introduce linguistic regularities that paradoxically reinforce human authorship signals rather than obscuring them.

Analysis of standard AI model outputs (categories 03-06) indicates a discernible hierarchy in detection difficulty, forming the “moderately challenging” cluster, as illustrated in orange in Figure 5. Within this group, GPT-3.5 outputs were the most detectable, achieving an accuracy rate of 58 %, followed by Gemini at 56 %, and Bing at 54 %. Notably, GPT-4 exhibited enhanced evasion capabilities, with detection accuracy decreasing to 42 %, marking a 16-percentage-point reduction from its predecessor and categorizing it within the “highly evasive” cluster.

The advanced model categories, specifically 07-Claude and 08-LLaMA2, posed significant challenges in detection, with mean accuracies of 38 % (SE = 6,8 %) and 41 % (SE = 6,5 %) respectively. These models are distinctly positioned within the highly evasive cluster, as indicated in red in figure 5. Notably, these models achieved near-total evasion for several detection tools: seven tools failed to identify any texts generated by Claude (0/5 correct), while five tools were completely unsuccessful in detecting LLaMA2-generated texts.

Figure 5. ANOVA Analysis: Detection Performance Across Text Categories

The ANOVA analysis reveals significant differences in detection difficulty across categories, with an effect size (η^2 = 0,595) indicating that 59,5 % of the variance in detection accuracy is attributable to text category. Post-hoc Tukey HSD tests, as summarized in the figure's annotation, identify three statistically distinct difficulty clusters: easily detected (human-written, translated), moderately challenging (GPT-3.5, Gemini, Bing), and highly evasive (GPT-4, Claude, LLaMA2, paraphrased).

DISCUSSION

This comprehensive benchmarking study provides a detailed account of the capabilities and limitations of existing AI-based text-detection tools. In light of these findings, it is evident that the challenge of distinguishing between human-written and machine-generated text remains a significant obstacle in the field of text detection.[21]

Despite the use of advanced detectors, the data set, which was fairly diverse, did not yield a detector with perfect performance. The confusion matrices depicted in Figure 4 can be regarded as an adequate representation of the distribution of true positives, true negatives, false positives, and false negatives for each tool. They provided a transparent and reliable account of the actual performance of each tool. These metrics, which are calculated in parallel with the qualitative performance of the tools, provide further support for the findings and implications presented in this study.[22]

The limited sample size (n=5 per category) restricts the statistical power and precision of our estimates. With only 5 samples per category, our 95 % confidence intervals for accuracy are approximately ±20 %, indicating that observed differences between tools smaller than 40 % may not reflect true performance differences. This sample size affords sufficient power (0,80) solely for detecting large effect sizes (d>1,2). Therefore, our findings should be regarded as indicative rather than conclusive, establishing preliminary performance patterns that necessitate validation with larger samples.[23]

The human-written samples exhibit limited disciplinary diversity, concentrated in computer science, education, and business fields. This narrow disciplinary range may not capture the full spectrum of academic writing styles across domains such as natural sciences, medicine, humanities, or social sciences. Different disciplines employ distinct methodological vocabularies, citation patterns, and argumentative structures that could affect detection tool performance. The absence of systematic disciplinary sampling represents a significant limitation in assessing tool generalizability across academic contexts. The observed detection accuracies should be interpreted within the context of our sample's disciplinary constraints. Tools trained predominantly on technical and business texts may perform differently when encountering humanities scholarship or clinical research writing.[24]

The manual review process lacked inter-rater reliability assessment and did not evaluate substantive content quality. A more rigorous protocol with multiple reviewers, formal scoring rubrics, and assessment of factual accuracy would strengthen future studies. Additionally, we did not control for potential quality variations between AI models, which might confound detection performance if certain models produce inherently more detectable outputs due to quality issues rather than stylistic patterns.[25]

It can be reasonably inferred that the deployment of AI text detection tools in high-stakes scenarios, such as automated screening of student work or content moderation, should be approached with a high degree of caution. However, in instances where certain tools are susceptible to false positives, the potential for an individual's bias to supersede all other considerations is a significant concern. The evaluated tools demonstrated both precision and recall, indicating that their selections are appropriate for a range of use cases. In the event that reducing false positives is of particular importance, it would be preferable to utilize tools with superior precision, such as SEO.AI or Copyleaks.

In the context of application settings, which are primarily designed to identify a multitude of instances of AI-generated text, tools with a higher recall, such as "Content at Scale" and "SEO.AI," are more suitable. These findings suggest that the efficacy of AI text detection tools may be contingent upon the intricacy of AI text generation models. The evaluation demonstrated that the generally lower accuracy scores of the tools when tested on text run through more advanced models, particularly Claude and LLaMA, indicated the presence of potential blind spots within the detection process.[26]

From a research perspective, this study identifies several promising avenues for future research. One potential avenue for enhancing generalizability and robustness is the expansion of training corpora for AI text detection tools coupled with an increase in diversity with respect to textual genres and styles. The diversity of human languages presents a significant challenge to machine-learning models. Nevertheless, recent developments in unsupervised and self-supervised learning methods may offer potential solutions to data limitations and enhance detection across a broader range of AI-generated text styles. From an educational standpoint, the present study highlights the necessity for a balanced approach to the utilization of AI text-detection tools in academic contexts. Such aids are invaluable to maintaining academic integrity; however, complete reliance on automated detection as the sole arbiter of "original" or "legitimate" work may result in unwarranted student penalties. These limitations and potential biases indicate the necessity for educators and other stakeholders

who utilize AI text-detection tools to exercise discernment in their application.

The implications of this study's findings are relevant to the development and deployment of AI technology. Moreover, the errors and biases observed in AI text detection tools provide further justification for caution and transparency when translating research advances to real-world applications. The development and utilization of these technologies must be acutely aware of their inherent limitations and potential hazards to prevent adverse outcomes, particularly when employed in high-stakes decision-making processes that have a significant impact on individuals and society. The concentration on educational prompts limits generalizability across text genres. Detection tools may perform differently on technical, creative, or journalistic content, as these domains exhibit distinct stylistic and structural features. Future studies should employ multi-domain sampling to establish comprehensive performance baselines.

CONCLUSIONS

The evaluation of ten AI text detection tools highlights a significant issue: current detection technologies are unable to reliably differentiate between human and machine-generated text, with accuracy rates ranging from 22 % to 88 % and significant vulnerabilities to basic paraphrasing techniques. These findings necessitate fundamental changes in institutional approaches to AI text detection. Content at Scale's 88 % accuracy represents the current best-case scenario; however, this level of performance is inadequate for high-stakes applications. More concerning is the fact that all tools exhibited catastrophic failures when faced with paraphrased content, with accuracy plummeting to 28-34 %. This vulnerability, coupled with an observed 5,2 percentage-point annual decline in performance against newer models, suggests that detection-based approaches may become obsolete within 2-3 years.

Academic institutions are advised to employ multiple tools with an accuracy rate exceeding 90 %, complemented by human review. For medium-risk applications, a single tool with an F1 score greater than 0,80 may be utilized. Low-risk applications must include an accuracy warning. Specifically, institutions should establish a "zone of confidence," wherein only consistent scores above 70 % from multiple high-precision tools justify further investigation. Detection alone should not be deemed sufficient evidence for imposing sanctions.

The challenge of achieving perfect detection necessitates a shift in focus from prohibition to integration. Instead of intensifying the detection arms race, academic institutions should consider redesigning assessments to make AI assistance either irrelevant or explicitly advantageous. Methods such as oral examinations, iterative projects, and collaborative problem-solving are resistant to simple AI substitution while fostering complementary skills. Future methods of authenticity verification must incorporate human-AI collaboration through process documentation and cryptographic proof-of-authorship, rather than relying on post-hoc detection.

BIBLIOGRAPHIC REFERENCES

1. De-Fitero-Dominguez D, Garcia-Lopez E, Garcia-Cabot A, Del-Hoyo-Gabaldon JA, Moreno-Cediel A. Distractor Generation Through Text-to-Text Transformer Models. IEEE Access. 2024;12.

2. Gruetzemacher R, Paradice D. Deep Transfer Learning & Beyond: Transformer Language Models in Information Systems Research. ACM Comput Surv. 2022 Jan 31;54(10s):1–35.

3. Gong L, Crego J, Senellart J. Enhanced transformer model for data-to-text generation. In: EMNLP-IJCNLP 2019 - Proceedings of the 3rd Workshop on Neural Generation and Translation. 2019.

4. Caruccio L, Cirillo S, Polese G, Solimando G, Sundaramurthy S, Tortora G. Claude 2.0 large language model: Tackling a real-world classification problem with a new iterative prompt engineering approach. Intelligent Systems with Applications. 2024;21.

5. Toyama Y, Harigai A, Abe M, Nagano M, Kawabata M, Seki Y, et al. Performance evaluation of ChatGPT, GPT-4, and Bard on the official board examination of the Japan Radiology Society. Jpn J Radiol. 2024;42(2).

6. Grindrod J. Large language models and linguistic intentionality. Synthese. 2024 Aug 6;204(2):71.

7. Sobo A, Mubarak A, Baimagambetov A, Polatidis N. Evaluating LLMs for Code Generation in HRI: A Comparative Study of ChatGPT, Gemini, and Claude. Applied Artificial Intelligence. 2025 Dec 31;39(1).

8. Raza M, Jahangir Z, Riaz MB, Saeed MJ, Sattar MA. Industrial applications of large language models. Sci Rep. 2025 Apr 21;15(1):13755.

9. Ahmed Ali Linkon, Mujiba Shaima, Md Shohail Uddin Sarker, Badruddowza, Norun Nabi, Md Nasir Uddin

Rana, et al. Advancements and Applications of Generative Artificial Intelligence and Large Language Models on Business Management: A Comprehensive Review. Journal of Computer Science and Technology Studies. 2024 Mar 13;6(1):225–32.

10. Irfan D, Watrianthos R, Amin Nur Bin Yunus F. AI in Education: A Decade of Global Research Trends and Future Directions. International Journal of Modern Education and Computer Science (IJMECS). 2025;2(2):135–53. https://www.mecs-press.org/ijmecs/ijmecs-v17-n2/v17n2-7.html

11. Yoo JH. Defining the Boundaries of AI Use in Scientific Writing: A Comparative Review of Editorial Policies. J Korean Med Sci. 2025;40(23).

12. Perkins M, Roe J. Academic publisher guidelines on AI usage: A ChatGPT supported thematic analysis. F1000Res. 2024 Jan 16;12:1398.

13. Miao J, Thongprayoon C, Suppadungsuk S, Garcia Valencia OA, Qureshi F, Cheungpasitporn W. Ethical Dilemmas in Using AI for Academic Writing and an Example Framework for Peer Review in Nephrology Academia: A Narrative Review. Clin Pract. 2023 Dec 30;14(1):89–105.

14. Ghiurău D, Popescu DE. Distinguishing Reality from AI: Approaches for Detecting Synthetic Content. Computers. 2024 Dec 24;14(1):1.

15. Casal JE, Kessler M. Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. Research Methods in Applied Linguistics. 2023;2(3).

16. Hinton M, Wagemans JHM. How persuasive is AI-generated argumentation? An analysis of the quality of an argumentative text produced by the GPT-3 AI text generator. Argument and Computation. 2022;14(1).

17. An Empirical Study of AI-Generated Text Detection Tools. Advances in Machine Learning & Artificial Intelligence. 2023;4(2).

18. Gegg-Harrison W, Quarterman C. AI Detection's High False Positive Rates and the Psychological and Material Impacts on Students. In 2024. p. 199–219.

19. Elkhatat AM, Elsaid K, Almeer S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. International Journal for Educational Integrity. 2023 Sep 1;19(1):17.

20. Ghiurău D, Popescu DE. Distinguishing Reality from AI: Approaches for Detecting Synthetic Content. Computers. 2024 Dec 24;14(1):1.

21. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial Intelligence in Ophthalmology: A Comparative Analysis of GPT-3.5, GPT-4, and Human Expertise in Answering StatPearls Questions. Cureus. 2023 Jun 22;

22. Watrianthos R, Triono Ahmad S, Muskhir M. Charting the Growth and Structure of Early ChatGPT-Education Research: A Bibliometric Study. Journal of Information Technology Education: Innovations in Practice. 2023;22:235–53.

23. Zhao H, Ling Q, Pan Y, Zhong T, Hu JY, Yao J, et al. Ophtha-LLaMA2: A Large Language Model for Ophthalmology. 2023;

24. Amanda Amanda, Elsa Muliani Sukma, Nursyahrina Lubis, Utami Dewi. Quillbot As An AI-powered English Writing Assistant: An Alternative For Students to Write English. Jurnal Pendidikan dan Sastra Inggris. 2023;3(2).

25. Mohammad T, Nazim M, Alzubi AAF, Khan SI. Examining EFL Students' Motivation Level in Using QuillBot to Improve Paraphrasing Skills. World Journal of English Language. 2024;14(1).

26. Nawaz SA, Li J, Bhatti UA, Shoukat MU, Ahmad RM. AI-based object detection latest trends in remote sensing, multimedia and agriculture applications. Front Plant Sci. 2022 Nov 18;13.

FINANCING

This research was funded by the Directorate General of Higher Education, Research, and Technology of the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia through the 2025 Fundamental Research Grant Program. The authors express their sincere gratitude for the support provided.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Yuhefizar.
Data curation: Ronal Watrianthos
Formal analysis: Ronal Watrianthos.
Research: Ronal Watrianthos.
Methodology: Yuhefizar.
Project management: Yuhefizar.
Resources: Yuhefizar.
Software: Ronal Watrianthos.
Supervision: Yuhefizar.
Validation: Dony Marzuki.
Display: Ronal Watrianthos.
Drafting - original draft: Dony Marzuki.
Writing - proofreading and editing: Dony Marzuki.