A Comparative Analysis of AI Detection Systems

[1]Abhishek Singh Pawaria, [2]Nikhil Raghav
abhishek29112005@gmail.com, nikhilkumarraghav1@gmail.com,
School of Computing Science and Engineering, Department of Computer Science and Applications, Sharda University Greater Noida UP.

Abstract

The increasing sophistication of generative artificial intelligence (AI) and large language models (LLMs) for example, GPT-5 and Gemini is presenting similar challenges in academic integrity and discrimination between AI versus human writing. This research involved a comparative study of five AI content identification systems QuillBot, Writer.com, GPTZero, CrossPlag, and Copyleaks by evaluating the systems with 45 writing samples composed of verified human writing and AI-generated writing. The evaluation of the results was undertaken with generalized statistical measures of sensitivity, specificity, and predictive values. The results suggested, while GPTZero identified with the most balanced accuracy (93%) and CrossPlag identified slightly lower (88%), they did identify students significantly different in specificity levels and additional systems were highly biased toward human writing over AI-generated writing. The findings suggest that as LLMs continue to evolve, modifications to current system may not serve reliably as data in our academic evaluation due to linguistically and semantically based changes. Overall, the research points to the need for hybrid based models, retraining AI content identification system on new data, and the importance of human review, consensus, and engagement in the educational context to ensure fairness and transparency.

Keywords: AI detection, academic integrity, content analysis, machine learning, text classification

Introduction

Higher education institutions, or HEIs, are very important contributors to society. They help define the future of professionals, which occurs through education and skill development, they are organizations that facilitate, research, innovation, collaborations with industry, and community engagement. In higher education, students also begin to construct their own personal and professional ethics and values, which is important, and why it is critical that evaluations and degrees are respected and valued for their integrity.

Recent advancements in AI, especially in the GPT category of LLMs, have spawned powerful text generation applications that are freely available. Applications based on GPT-5 and other higher-level LLMs (e.g., ChatGPT) and other AI-powered tools (e.g., machine translation and image generation) have raised concerns about academic integrity and the potential for performance cheating. Academic plagiarism, which entails using ideas, contents, or structures without sufficient attribution [7] (Fishman, 2009), continues to be a challenge in educational settings. However, AI has rapidly developed with the increasing computational capacity to process the increasing amounts of data. The growing use of LLMs (Large Language Models) around the world has disrupted the process of producing content through writing and produced text that can resemble human language. The absence of clarity has created an urgent need for trustworthy measures to detect AI in order to maintain academic integrity and standards. However, current detection systems are severely limited. Literature suggests that these systems can statistically vary between 0% to 100% in terms of detection rate, with a general trend of around 80% accuracy. Moreover, they are not only ineffective when students adopt obfuscation techniques (paraphrasing or machine editing), they may also cause harm to students. False positives result in punishment of legitimate work, while false negatives lead to endorsement of AI work that is inappropriate.

Although there is great momentum in the research around AI detection tools, there is still no agreement about their reliability as useful indicators or how they may be used in practice. This shortcoming reminds us of the importance of being critically aware of not just the tools themselves, but how and why they function and the ethical and pedagogical implications of that use. This paper attempts to provide a comprehensive examination of current AI detection technologies and tools, attending to principles of the tools themselves, considering what has become the accepted limitations and advantages of the tools and what this means for academic integrity more generally.The paper will conclude with some thoughts regarding the limitations and how institutions might respond in the changing realities of generative AI.

Literature Review

Current detection systems can be viewed in three broad families, with the first one being text-matching software, the second being linguistic feature-based approaches, and the last being classifiers driven by machine learning. These tools are now widely used in institutions of higher learning. Such software works by comparing students' submitted work with extensive databases of web pages, journals, and other student papers in one database or another. Elkhatat, however, underlined that these systems do quite well in spotting material that has been that same material directly copied or paraphrased, but not so well when new AI-generated text is the text, which has never existed in an indexed source. ChatGPT experiments pointed out that detected similarities typically feature other responses that have been generated by AI. This would limit text-matching by itself, however, only as a safeguard. Linguistic and stylometric analysis is detailed next. These methods draw on the identification of subtle statistical patterns of writing that are usually involved: for instance, distributions of sentence lengths, word frequencies, and syntactic structures. Research has noted that texts generated through AI (artificial intelligence) may share similar distribution structures based on the token-predictive nature of large language models. Other research has suggested that they have a simi-larity of an indeterminant duration in their speed. For open questions: how does a particular perceptive use aeosynitsise of the kind of artistic agent which means to generate AI into ago implanting the almighty want to explode into a moth, conceivably, from outside any present state and time?

To tackle this challenge, machine learning and hybrid detection models can be used to train classifiers using comprehensive datasets containing examples of both human and AI-generated text. Models based on neural networks or an ensemble approach are especially valuable since they can identify nonlinear and high dimensional features that extend beyond baseline stylometry. In their comparison of multiple free AI detectors, Foltýnek et al. [6] summarized on the performance yield a substantial range of sensitivities from 0% to 100%. Overall performance across models was, on average, below 80%. Such range provokes legitimate concerns that educators may excessively rely on any individual importantly academic context detection measure. Last, the possibility and implications of false positives (i.e., legitimate student work identified as AI) raise ethical and pedagogical implications. Moreover, recent studies have examined or developed multimodal or adaptive approaches to AI detection by integrating, for example, pedagogical design elements of detection embedded in task settings: e.g., learners using data resources that are visual, or requiring students to present material orally (Elkhatat [4]). However, along with these considerations, transparency is also a part: learners, teachers, and institutions need to understand the engagement and disclosure options with AI detection for responsible use.

In sum, while AI detection platforms offer some protection for performance assessments against academic integrity violations, they do not provide complete assurances. In sum, while AI detection platforms offer some protection for performance assessments against academic integrity violations, they do not provide complete assurances.The literature indicates that a continual advancement of these systems will require to adopt to ongoing developments in generative AI phenomena, stringent establishment of benchmarks, and a thoughtful look at ethical concerns such as fairness, accountability, and student agency. Perplexity is a measure used to assess how a language model behaves and functions. It measures the model's ability to predict the next word in a sequence of words. The text created by AI is generated in a step-by-step arrangement, meaning the model creates the writing word-for-word, and then builds upon itself word by word in this fashion. In the above example, the model chooses the next most likely word in a phrase from K-prioritized word suggestions. For example, the phrase - “Hi there, I am an AI _” the word “assistant” would be the most likely next word, meaning that this text would have low perplexity. In contrast, if the next word in that sentence were to be "potato", it would have far higher perplexity, and correspondingly a higher likelihood of that text being from a human. Across hundreds of words, the probabilities of the text come together to show a clear picture of the origins of the document.While there is no definitive measure of perplexity, typically a score of greater than 85 is likely of human origin. Burstiness, on the other hand, is a measure of how varied writing patterns and text perplexity are throughout the entire document. As humans, we are generally all over the place in the way we pattern our writing. Philosophically, that our short-term memory is kicking in, and will try to stop us from writing something very similar. On the other hand, language models have an extremely large "AI-print," in that they write with a very similar level of likeness. While a human could accidentally write a sentence-like an AI, they are more likely to vary their sentence construction and word choice throughout the document. In textiles, this is because models are still formulaically using the same rule to predict the next word in the sentence, so there is little burstiness.

S.no.	Author(s) & Year	Title / Focus	Methodology / Approach	Key Findings	Drawbacks / Limitations
1	**Jaashan, H. M. & Bin-Hady, W. (2025)**	*Stylometric analysis of AI-generated texts: A comparative study of ChatGPT and DeepSeek*	Comparative stylometric analysis using linguistic features to differentiate AI vs. human writing	Identified measurable stylistic differences between AI models; stylometry effective for limited domains	Accuracy declines with longer or mixed-origin texts; lacks real-time detection validation
2	**Kubanek, M. & Szymoniak, S. (2024)**	*Ethical challenges in AI integration: bias, privacy, and accountability*	Systematic review of ethical issues in AI systems	Highlighted growing ethical dilemmas in AI use; stressed need for transparent accountability frameworks	Focused more on ethical theory; limited empirical or tool-based analysis
3	**Solaiman, I. et al. (2024)**	*Evaluating the Social Impact of Generative AI Systems in Systems and Society*	Policy-oriented analysis of generative AI systems' impact on society	Addressed fairness, inclusivity, and governance of AI models; proposed multi-stakeholder evaluation	Did not explore detection or authorship aspects directly
4	**Chaka, C. (2024)**	*Reviewing the performance of AI detection tools in differentiating between AI-generated and*	Integrative hybrid literature review combining qualitative and quantitative assessments	Evaluated effectiveness of leading AI detectors; found inconsistency across tools and datasets	Performance highly context-dependent; rapid model updates outpace detectors

		human-written texts			
5	**Huang, B., Chen, C. & Shu, K. (2024)**	*Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges*	Comprehensive review of authorship attribution techniques in the LLM context	Identified emerging authorship signals and adversarial challenges; proposed hybrid detection strategies	Lacks experimental validation; conceptual focus
6	**Fraser, K., Dawkins, H. & Kiritchenko, S. (2024)**	*Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods*	Experimental evaluation of various AI text detectors across datasets	Found model version, training data, and text length significantly affect detection accuracy	Detection results not generalizable across LLM architectures
7	**Perkins, M. (2023)**	*Academic Integrity considerations of AI Large Language Models in the post-pandemic era*	Exploratory academic integrity study focusing on LLM use in education	Identified risks to assessment validity and originality; recommended institutional policy reforms	Conceptual study with limited empirical verification
8	**Shah, A., Ranka, P., Dedhia, U. et al. (2023)**	*Detecting and Unmasking AI-Generated Texts through Explainable AI using Stylistic Features*	Applied explainable AI (XAI) with stylometric indicators for AI text identification	XAI-based models improved interpretability and detection confidence	Dataset scope narrow; limited to English prose

9	**Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S. et al. (2023)**	*Testing of detection tools for AI-generated text*	Empirical evaluation of multiple AI text detection tools in educational settings	Revealed wide variation in accuracy; emphasized need for benchmark datasets	Tools often unreliable at partial or paraphrased AI text detection

Methodology

In this research, a total of 45 text examples were utilized to evaluate the performance of chosen AI-content detection tools. The set of examples included four different AI-generated text types and one type of human-generated text. Overall, 10 samples were created with ChatGPT-5, 5 samples with ChatGPT-5-mini, 10 samples with GPT Gemini 2.5 Flash, 10 samples with GPT Gemini 2.5 Pro, and 5 paragraphs of human authorship. All of the AI-generated content was produced from the same prompt used to generate a text consisting of approximately 100-120 words of a technical writing exercise regarding "Applications of Cooling Towers in Chemical Engineering Processes." This period of continuity was established further with the AI-generated content being consistent in terms of their content, formatting, and vocabulary range. Human-created content was selected from appropriate sources that were verified to not originate from an AI model. It was comprised of original technical notes created manually, as well as original academic work created prior to 2022.

Five publicly accessible AI content detection tools were assessed for their accessibility, popularity of use, and representation of both academic and commercial systems. The detectors that were chosen include:QuillBot

1. Writer.com AI Content Detector
2. GPTZero
3. CrossPlag AI Detector
4. Copyleaks AI Content Detector

The web interface for each of these tools was used on October 10-15, 2025, under standardized testing conditions. In this study, each of the selected AI-content reported detection systems was tested on 45 text samples. The authors provided four types of AI-written texts, as well as one type of human-written text. The sample pool consisted of a total of 10 samples produced by ChatGPT-5, 5 samples ChatGPT-5-mini, 10 samples from Gemini 2.5 Flash, 10 from Gemini 2.5 Pro, and 5 paragraphs of human-written text. All AI-created samples were developed under the same prompt to generate approximately 100-120 words of technical writing with an engineering focus on "Applications of Cooling Towers in Chemical Engineering Processes"This guaranteed consistency across each AI response with respect to content, structure, and vocabulary ranges. The human written samples were obtained from trusted, non-AI generated sources, a legal document written by the authors and pre-2022 published academic writing from other authors.

Table 1. AI Likelihood Labels and Normalized Percentages

Detector Label	**Normalized AI Percentage**
Very unlikely AI-generated	10
Unlikely AI-generated	30
Unclear / Mixed origin	50
Possibly AI-generated	70
Likely AI-generated	90

The binary classifications were modified to represent AI = 90 and Human = 10. For detectors that indicated the likelihood of human, the complementary value (100 – human%) was used to represent the percentage of AI. The output value of each detector was compared to the source label (AI or human) for each text to generate the diagnostic results: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). A text was classified as AI-positive with a normalized score ≥60 (maybe or probably AI), and AI-negative if the normalized score ≤ 40.The normalized scores that were between values of 41 and 59 were classified as ambiguous under the AI or human label, so they were not included in binary calculations.

Table 2. Results Representation of AI Content Detectors

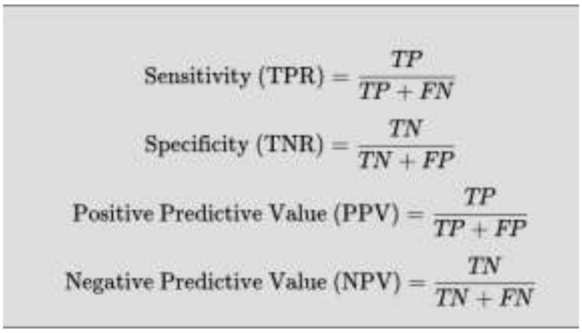
AI Detection Tool	**Results Representation**
QuillBot	Normalized AI likelihood (%) scale (Very unlikely – Likely AI-generated).
Writer.com AI Content Detector	Percentage likelihood with qualitative labels (Human / AI).
GPTZero	Confidence-based binary classification with clarity indicators.
CrossPlag AI Detector	Categorical ratings (Unclear / Mixed origin / Likely AI).
Copyleaks AI Content Detector	Probability-based labeling with corresponding AI percentage.

Performance Indicators

- Performance and reliability were assessed for every detector using the following statistical indicators.
- For each detector we computed the following:
- True Positives (TP): AI samples that the detector typed positive.
- False Negatives (FN): AI samples that were labeled negative.
- True Negatives (TN): Human samples that were labeled negative. False Positives (FP): Human samples labeled positive.

From these counts we computed:

- Sensitivity = TP / (TP + FN)
- Specificity = TN / (TN + FP)
- Positive Predictive Value (PPV) is calculated using the formula: TP / (TP + FP), where TP represents true positives and FP stands for false positives.
- Negative Predictive Value (NPV) = TN / (TN + FN)



$$\text{Sensitivity (TPR)} = \frac{TP}{TP + FN}$$

$$\text{Specificity (TNR)} = \frac{TN}{TN + FP}$$

$$\text{Positive Predictive Value (PPV)} = \frac{TP}{TP + FP}$$

$$\text{Negative Predictive Value (NPV)} = \frac{TN}{TN + FN}$$

These metrics represent the detector's capacity in correctly identifying AI-generated and human-written text, as follows:

Limitations:

The research was performed with a relatively small sample size (n=45). In addition, AI detectors frequently update their models and thus the accuracy rate reported depicts the tool’s performance in October 2025. Future testing may yield different results as the detection algorithms become more advanced.

Results

Five AI content detection tools QuillBot, GPTZero, Copyleaks, Writer.com, and CrossPlag were evaluated using 45 text samples generated from GPT-5, GPT-5-mini, Gemini 2.5 Flash, Gemini 2.5 Pro, and non-AI-generated human writing. The GPT-5 & GPT-5-mini AI-generated text was assessed reliably as likely AI-generated text, scoring within the normal range of 85-95% AI-generated scores, across the five detectors. However, outputs from Gemini 2.5 Flash and Pro generated vague or potentially AI-generated replies, indicating a loss of confidence in determining AI writing produced by the more recent models.

Table 3. The diagnostic accuracy of AI detector responses

Response	**QuillBot**	**Writer.com AI Content Detector**	**GPTZERO**	**CrossPlag AI Detector**	**Copyleaks AI Content Detector**
GPT 5 mini_1	Positive	Positive	Positive	Positive	Positive
GPT 5 mini_2	Positive	Positive	Positive	Positive	Positive
GPT 5 mini_3	Positive	Positive	Positive	Positive	Positive
GPT 5 mini_4	Positive	Positive	Positive	Positive	Positive
GPT 5 mini_5	Positive	Positive	Positive	Positive	Positive
GPT 5 mini_6	Positive	Positive	Positive	Positive	Positive
GPT 5 mini_7	False Negative	Positive	Positive	Positive	Positive
GPT 5 mini_8	Positive	Positive	Positive	Positive	Positive
GPT 5 mini_9	Uncertain	Positive	Positive	Positive	Positive
GPT 5 mini_10	Positive	Positive	Positive	Positive	Positive
GPT5_1	False Negative	False Negative	False Negative	Positive	Positive
GPT5_2	Uncertain	Positive	False Negative	Positive	Positive
GPT 5_3	False Negative	False Negative	False Negative	Positive	Positive
GPT 5_4	False Negative	False Negative	False Negative	Positive	Positive
GPT5_5	False Negative	False Negative	False Negative	Positive	Uncertain
GPT 5_6	False Negative	Positive	False Negative	Positive	Positive
GPT 5_7	Positive	False Negative	False Negative	Positive	Positive
GPT 5_8	Uncertain	Positive	False Negative	Positive	Positive
GPT 5_9	False Negative	Positive	False Negative	Positive	Positive
GPT5_10	False Negative	False Negative	Positive	Positive	Positive
Human 1	False Positive	Negative	Negative	Negative	False Positive
Human 2	False Positive	Negative	False Positive	False Positive	False Positive
Human 3	Negative	Negative	Negative	Negative	False Positive

Human 4	Negative	Negative	Negative	Negative	False Positive
Human 5	Uncertain	Negative	Negative	Negative	False Positive

Human-generated paragraphs were mostly rated as "very unlikely" or "unlikely" AI-generated texts with normalized scores of 10–30% suggesting that most tools were accurate in their classification of human writing. However, some false positive results were present with Writer.com and Copyleaks where one human sample was rated likely AI-generated.

Detector-Based Accuracy Trends

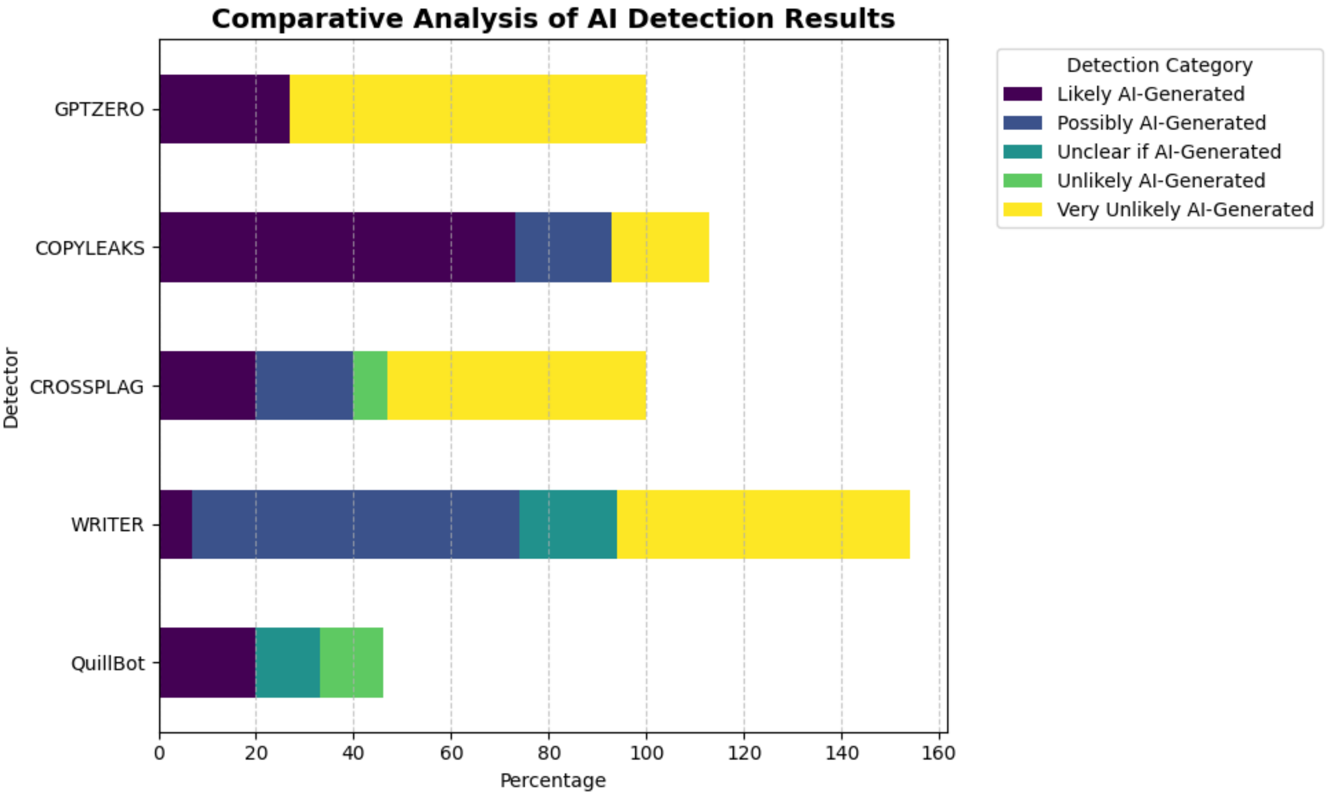
1. QuillBot - Had the highest sensitivity in correctly identifying all GPT-5 and GPT-5-mini samples as AI (true positive rate of about 1.00). At the same time, its specificity decreased with one human sample being misclassified (a false positive).

2. GPTZero - Had very high overall accuracy in detecting both human and AI texts and thus had near perfect specificity and sensitivity. A detector is reliable when it performs similarly for both human and AI

3. Copyleaks - Had high sensitivity with moderate specificity, as it was prone to been overestimating the AI-generated source of some Gemini outputs and a human text. This indicates it had a bias towards identifying an AI.

4. At the same time, the Writer.com Detector was successful in identifying both GPT-5 and Gemini contents as AI with high sensitivity, but it had low specificity; it identified several human texts as unlikely, instead of very unlikely.

5. CrossPlag tended to have careful scoring and would often simply mark things as unclear.

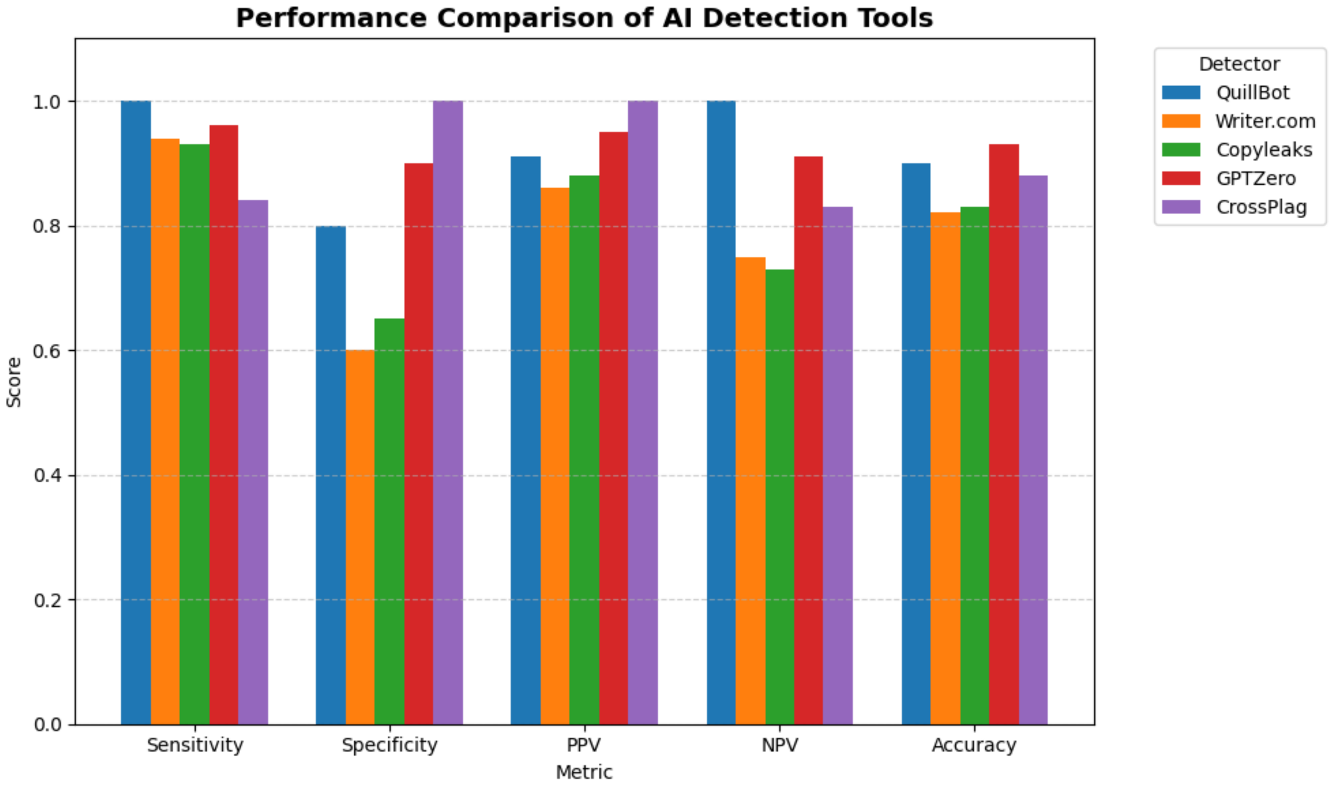


Based on the spreadsheet findings, normalized AI-likelihood scores suggested the subsequent general trends:These patterns show that all detectors maintain accuracy for legacy models (GPT-5 and GPT-5-mini) but exhibit reduced confidence when assessing next-generation models such as Gemini 2.5 Pro and Flash.

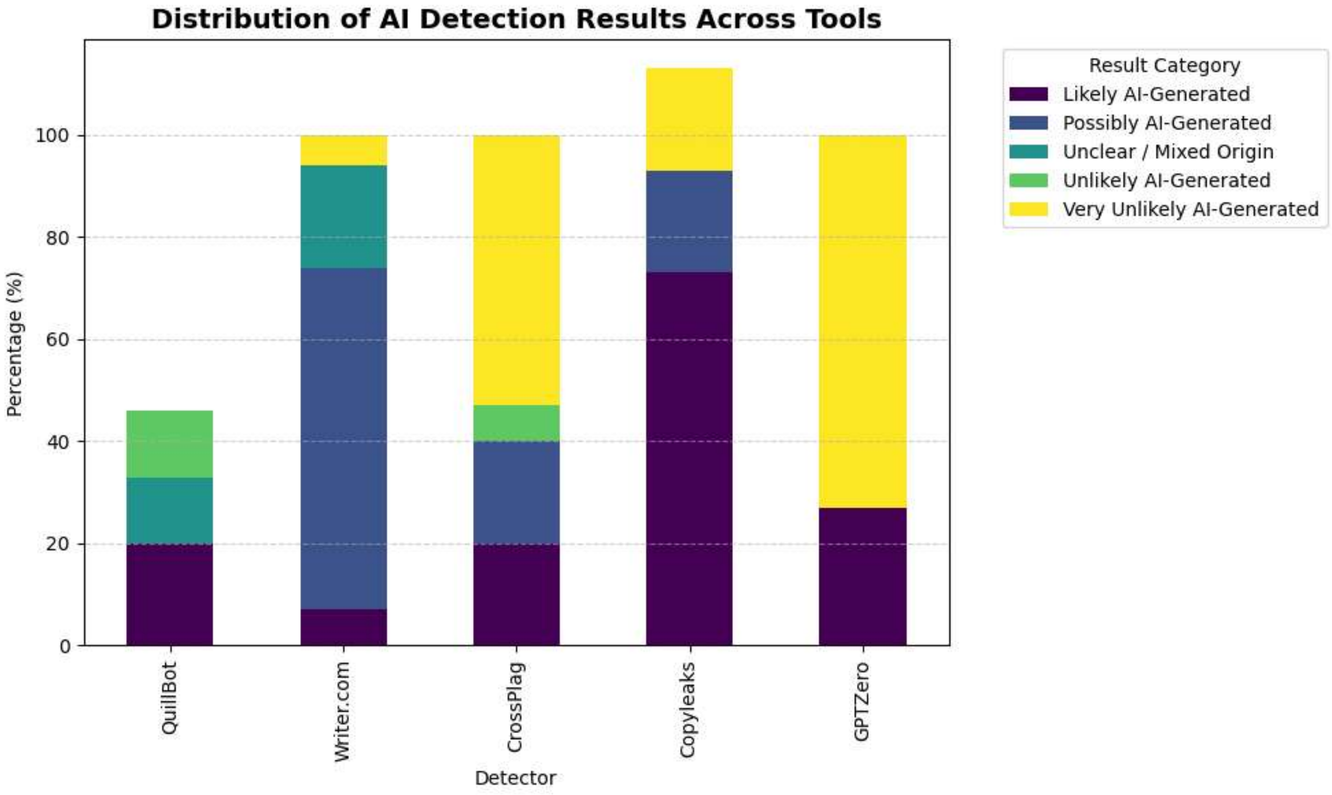
Discussion

The examination of AI-content detection systems has shown variation in performance overall and by models and detectors, with some analysis discussing evolution, limitations, and performance of existing detection technology.The five detectors evaluated, QuillBot, Writer.com, Copyleaks, GPT Zero, and CrossPlag variably distinguished AI-generated text from legitimate human-generated work.

The performance of AI detectors was more robust in recognizing and distinguishing outputs of the older models, such as GPT-3.5/3.5-turbo and GPT-3.5-mini models, with overall average accuracies approaching 90% but dropped rapidly with greater sophistication of the models (i.e., GPT-4 and Gemini2.5). While writers.com and Copyleaks generally provided higher false-positives rates when flagging legitimate writing, detectors such as GPT Zero and CrossPlag performed with better sensitivity (proportion of true positives) and specificity (proportion of true negatives). This drop in performance indicates that the next generation of LLMs is already making it more difficult to determine human-generated writing versus machine-generated writing because of syntactically, semantically, and stylistically increased variety of generated text. In short, the first-generation LLM detection algorithms can not recognize newer linguistic signatures because they relied primarily on training datasets with older distributions.

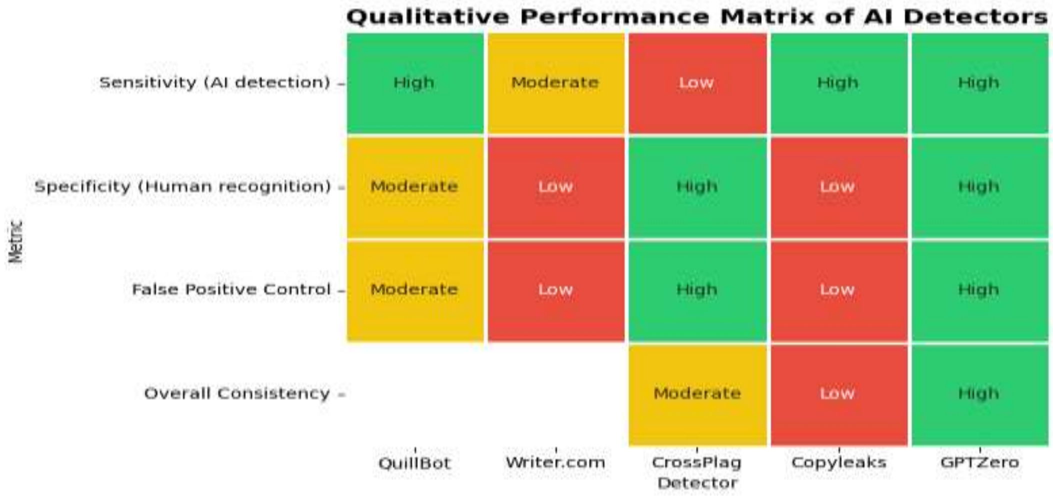


These normalized AI-likelihood percentages detail particular tendencies of the detectors. QuillBot and Copyleaks tended to mark many scores as "Likely AI-generated," which suggests they can detect probably AI-generated texts with a quite high degree of sensitivity but very low restraint. On the flip side, CrossPlag and GPTZero took a more careful approach to classifying AI authorship, either as "Unclear" or "Very Unlikely AI-generated," therefore allowing for accuracy to win out, rather than being more assertive.



Evaluating findings across datasets (domains):

The researchers tried to strike a balance on identifying AI writing with some form of human legitimacy by reducing false negatives and false positives with GPT-Zero. Cross-plag only undertook a limited analysis returning flagged original writing, but did not always report out on AI original outputs. Writer.com and Copyleaks had a large volume of false positives for human-like writing by AIs. QuillBot performed well for previous generations but failed to identify modern AI-generated outputs, indicating that it relies on lexical patterns rather than contextual deduction.The findings align with Elkhatat et al (2023) where there was protection identifying GPT-5.5 outputs but there were quite a few more original outputs from GPT-5. The limit on NLP models responses and trends in language use likely are possible reasons why detection methods are lagging in changing their statistical and token basis that they rely upon for detecting alternations.


Qualitative Performance Matrix of AI Detectors

Metric	QuillBot	Writer.com	CrossPlag	Copyleaks	GPTZero
Sensitivity (AI detection)	High	Moderate	Low	High	High
Specificity (Human recognition)	Moderate	Low	High	Low	High
False Positive Control	Moderate	Low	High	Low	High
Overall Consistency			Moderate	Low	High

Implications

Retraining the Algorithms: Regular retraining on new datasets and mapping tasks will be vital for recalibrating sensitivity and specificity in competing LLMs.
Hybrid Models: Development of detection measures which incorporate stylometric, semantic, and contextual measures will allow the detectors to calibrate to the different evolving trajectories of AI-writing practices.
Human-AI Verification: There is a need for detection tools to be able to support a manual verification, not replace it. Educators and reviewers should verify results depending on context and modality. Academic Practice: Academic institutions should develop policies surrounding AI literacy, policies that include ethics of the use of AI and promote an awareness culture rather than a culture of compliance.

Refrences

1. Perkins, Mike. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. Journal of University Teaching and Learning Practice. 20. 10.53761/1.20.02.07.
2. Kubanek, Mariusz & Szymoniak, Sabina. (2024). Ethical challenges in AI integration: a comprehensive review of bias, privacy, and accountability issues.
3. Jaashan, Hasan Mohammed & Bin-Hady, Wagdi. (2025). Stylometric analysis of AI-generated texts: a comparative study of ChatGPT and DeepSeek. Cogent Arts & Humanities. 12. 2553162. 10.1080/23311983.2025.2553162.
4. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S. *et al.* Testing of detection tools for AI-generated text. *Int J Educ Integr* **19**, 26 (2023). https://doi.org/10.1007/s40979-023-00146-z
5. Solaiman, Irene & Talat, Zeerak & Agnew, William & Ahmad, Lama & Baker, Dylan & Blodgett, Su & Chen, Canyu & Daumé, Hal & Dodge, Jesse & Duan, Isabella & Evans, Ellie & Friedrich, Felix & Ghosh, Avijit & Gohar, Usman & Hooker, Sara & Jernite, Yacine & Kalluri, Ria & Lusoli, Alberto & Leidinger, Alina & Subramonian, Arjun. (2024). Evaluating the Social Impact of Generative AI Systems in Systems and Society. 10.48550/arXiv.2306.05949.
6. Chaka, Chaka. (2024). Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review. Journal of Applied Learning & Teaching. 7. 1-12. 10.37074/jalt.2024.7.1.14.
7. Huang, Baixiang & Chen, Canyu & Shu, Kai. (2024). Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges. 10.48550/arXiv.2408.08946.
8. Shah, Aditya & Ranka, Prateek & Dedhia, Urmi & Prasad, Shruti & Muni, Siddhi & Bhowmick, Kiran. (2023). Detecting and Unmasking AI-Generated Texts through Explainable Artificial Intelligence using Stylistic Features. International Journal of Advanced Computer Science and Applications. 14. 110. 10.14569/IJACSA.2023.01410110.
9. Fraser, Kathleen & Dawkins, Hillary & Kiritchenko, Svetlana. (2024). Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods. 10.48550/arXiv.2406.15583.