International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

An Empirical Evaluation of AI-Generated Text Detection Tools Across Diverse Academic Domains

Aman Nazare
Department of Computer Science
Dr.D.Y.Patil ACS College
Pune, Maharashtra

Harriesh chindaliya
Department of Computer Science
Dr.D.Y.Patil ACS College
Pune, Maharashtra

Abstract
The rapid growth and accessibility of generative artificial intelligence tools have significantly influenced the way academic content is created. While these tools offer valuable support for learning and research, their misuse in academic writing has raised concerns related to originality and academic integrity. In response, educational institutions have increasingly adopted AI-generated text detection tools to identify whether submitted content is human-written or produced by artificial intelligence.
Although several studies have examined the effectiveness of such detection systems, **comparative evaluations across different academic disciplines remain limited**, particularly in the context of structured and formal academic writing. Most existing detection tools rely on linguistic patterns and probabilistic models, which may not sufficiently account for variations in writing styles across domains. This limitation increases the risk of false positive classifications, where genuine human-written academic work is incorrectly identified as AI-generated.
This study presents an empirical evaluation of commonly used AI-generated text detection tools across multiple academic domains, including Technology and Innovation, Literature and Arts, History, Politics, and Education Systems. By analyzing both human-written and AI-generated academic texts, the study examines detection accuracy, false positive rates, and consistency of results across tools and domains.
The findings highlight notable variations in detection performance and reveal key limitations in current AI-text detection mechanisms when applied to formal academic content. The study aims to provide evidence-based insights that can support educators and institutions in making more informed and cautious use of AI-detection tools within academic assessment processes.

Keywords:
AI-generated text detection, academic integrity, false positives, cross-domain analysis, generative artificial intelligence

1.INTRODUCTION

The emergence of generative artificial intelligence has brought significant changes to academic writing. Modern language models are capable of producing structured, grammatically correct, and contextually relevant essays across a wide range of subjects. While these systems can support learning and idea development, their misuse has created serious concerns about academic honesty.

As a result, many educational institutions have started using AI-generated text detection tools to determine whether submitted assignments were written by students or produced using artificial intelligence. However, the reliability of these detection tools is still debated. Academic writing often follows a structured and formal pattern, which may resemble AI-generated text. This similarity increases the possibility of incorrect classifications.

Furthermore, writing styles differ across academic domains. A technical essay may use precise and repetitive terminology, while a history or political essay may follow a structured argumentative format. These variations can influence how detection tools interpret the text. Therefore, evaluating detection performance across multiple domains becomes essential.

This study aims to examine how consistently three commonly used AI detection tools perform across diverse academic subjects using a balanced dataset.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

2. Literature Review

Several researchers have explored the challenges associated with detecting AI-generated text. Most detection systems rely on probabilistic language modeling, sentence predictability, and statistical analysis of word patterns. These tools attempt to measure how likely a sequence of words is to have been generated by a human versus a machine.

However, formal academic writing often exhibits predictable structures, clear organization, and consistent tone. These characteristics may resemble AI-generated patterns. Previous studies have reported concerns about high false positive rates, particularly for structured essays and non-native English writers.

Despite growing interest in AI detection, limited research has systematically compared performance across multiple academic domains. This study contributes to the field by providing a structured cross-domain evaluation.

3. Methodology

3.1 Dataset Preparation

The dataset consisted of 50 academic essays:

- 25 human-written essays
- 25 AI-generated essays

The essays were equally distributed across five domains:

- Cutting-edge Technology
- Visual Culture
- Educational System
- Ancient History
- Democracy vs Dictatorship vs Monarchy

Each domain included five human-written and five AI-generated responses. Essay length ranged between 250 and 350 words to maintain consistency.

3.2 Detection Tools Evaluated

The following detection tools were tested:

- GPTZero
- ZeroGPT
- Copyleaks

Each essay was individually submitted to all three tools.

3.3 Classification Rule

A threshold value of 50% was used to classify outputs. Texts with an AI probability score greater than or equal to 50% were categorized as AI-generated, while texts below 50% were classified as human-written.

This consistent rule ensured fairness and prevented subjective interpretation of mixed outputs.

3.4 Evaluation Metrics

Performance was evaluated using:

1. True Positive (TP) – AI essays correctly identified as AI
2. True Negative (TN) – Human essays correctly identified as Human
3. False Positive (FP) – Human essays incorrectly identified as AI
4. False Negative (FN) – AI essays incorrectly identified as Human

Accuracy, False Positive Rate (FPR), and False Negative Rate (FNR) were calculated accordingly.

4. Results and Analysis

4.1 Overall Performance

Detector	TP	TN	FP	FN	Accuracy
GPTZero	24	15	10	1	78%
ZeroGPT	25	19	6	0	88%
Copyleaks	25	16	9	0	82%

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

ZeroGPT achieved the highest overall accuracy (88%), followed by Copyleaks (82%) and GPTZero (78%).

4.2 False Positive and False Negative Analysis

False Positive Rates:

- GPTZero – 40%
- ZeroGPT – 24%
- Copyleaks – 36%

False Negative Rates:

- GPTZero – 4%
- ZeroGPT – 0%
- Copyleaks – 0%

The results indicate that detection tools were highly effective in identifying AI-generated essays, as shown by the very low false negative rates. However, false positive rates were significantly higher, particularly for GPTZero and Copyleaks. This suggests that genuine human-written academic essays were frequently misclassified as AI-generated.

4.3 Domain-Wise Observations

Performance varied across domains. Essays in Cutting-edge Technology and Democracy vs Dictatorship vs Monarchy showed higher false positive instances. These domains often use structured arguments and consistent terminology, which may resemble AI-generated patterns.

Ancient History and Educational System essays also demonstrated occasional misclassification due to formal writing styles. Visual Culture essays showed comparatively balanced results.

These observations highlight that writing structure and subject matter influence detection reliability.

.

5. Discussion

The findings of this study reveal an important pattern in the behavior of AI-generated text detection tools. While all three tools demonstrated strong capability in identifying AI-generated essays, their performance in correctly recognizing human-written academic content was less consistent. This imbalance is clearly reflected in the relatively high false positive rates observed, particularly in GPTZero and Copyleaks.

One notable observation is that structured academic writing appears to increase the likelihood of misclassification. Essays in domains such as Cutting-edge Technology and Democracy vs Dictatorship vs Monarchy often followed a logical argumentative structure, formal tone, and consistent terminology. These characteristics, while typical in academic writing, resemble patterns that AI detection systems associate with machine-generated text. As a result, genuine human-written essays were sometimes flagged incorrectly.

This raises a broader concern regarding the reliability of detection systems in real academic environments. If a tool incorrectly classifies a student's original work as AI-generated, it may lead to unnecessary academic scrutiny or disciplinary action. Such situations can undermine trust between students and institutions.

Another important insight from this study is that detection tools appear to prioritize minimizing false negatives. In other words, they aim to ensure that AI-generated text is not overlooked. While this objective is understandable, it may come at the cost of increasing false positives. A balanced detection system should aim to minimize both types of errors to ensure fairness.

Domain-wise variation also highlights the importance of contextual sensitivity. Academic writing differs significantly across disciplines. A one-size-fits-all detection approach may not be suitable for all domains. Future improvements in detection systems may require domain-aware adjustments rather than relying solely on statistical probability thresholds.

Overall, the results suggest that AI detection tools should serve as supportive indicators rather than definitive evidence of misconduct.

High false positive rates raise ethical concerns in academic environments. Incorrect classification can negatively affect students and reduce trust in automated systems. Therefore, AI detection tools

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

should not be used as standalone decision-making mechanisms.

The domain-wise variation further indicates that context matters. Detection tools may need domain-aware adjustments rather than applying uniform thresholds to all forms of academic writing.

6. Conclusion

This study presented a cross-domain empirical evaluation of three widely used AI-generated text detection tools using a balanced dataset of 50 academic essays. The results showed that while ZeroGPT achieved the highest overall accuracy, all tools demonstrated noticeable limitations, particularly in the form of elevated false positive rates.

The findings highlight an important reality: although AI detection systems are capable of identifying machine-generated content with high sensitivity, they are not yet fully reliable in distinguishing structured human academic writing from AI-generated text. The similarity between formal academic writing patterns and AI-generated language increases the risk of misclassification.

From an academic integrity perspective, this issue deserves careful consideration. Automated detection tools can assist educators, but they should not replace human evaluation. Decisions regarding academic misconduct should incorporate contextual review, writing history, and instructor judgment alongside automated detection results.

As generative AI technology continues to evolve, detection mechanisms must also adapt. Future research should focus on improving contextual awareness, reducing false positives, and developing transparent evaluation criteria that can be clearly explained to students and educators.

In conclusion, AI-generated text detection tools offer valuable support in maintaining academic standards. However, their limitations underscore the importance of cautious and responsible implementation within educational institutions.

7. Future Work

Future research can expand the dataset size, include multilingual academic writing, and explore hybrid detection approaches that combine statistical and semantic analysis. Improving contextual awareness may reduce false positive classifications and enhance fairness in academic assessment systems.

References

[1] OpenAI, “GPT-4 Technical Report,” 2023.

[2] S. Sadasivan, A. Kumar, and R. Patel, “Can AI-Generated Text Be Reliably Detected? A Comparative Study,” arXiv preprint arXiv:2303.11156, 2023.

[3] Turnitin, “AI Writing Detection Capabilities and Academic Integrity,” Technical Whitepaper, 2023.

[4] Copyleaks, “AI Content Detection Technology Overview,” Copyleaks Research Documentation, 2023.

[5] GPTZero, “Understanding AI Text Detection and Probability Scoring,” Technical Documentation, 2023.

[6] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” in Proc. ACM FAccT, 2021.

[7] J. Dwivedi et al., “Artificial Intelligence (AI): Multidisciplinary Perspectives on Emerging Challenges,” Int. J. Information Management, vol. 57, 2021.

[8] IEEE, “Ethical Considerations in Artificial Intelligence for Education,” IEEE Conf. Proc., 2022.