Journal of AI, Humanities, and New Ethics,
1(1), 5-18, March 2025
e-ISSN: 3062-2867
jaihne.com

Genc Bilge Publishing Ltd.
(Young Wise) © 2025
gencbilgeyayincilik.com

Research Article

Comparative accuracy of AI-based plagiarism detection tools: an enhanced systematic review

Seyhan Canyakan[1]

Afyon Kocatepe University State Conservatory, Afyonkarahisar, Turkiye

Article Info

Received: 28 January 2025
Accepted: 4 March 2025
Online: 30 March 2025

Keywords
AI-based plagiarism detection tools
AI and ethics
Academic writing

Abstract

Turnitin AI has detected machine-generated text with accuracy rates ranging from 92% to 100% and approximately 5.3% false negative rate (Gosling et al., 2024; Díaz Arce, 2023). OriginalityAI has achieved near-perfect accuracy (98%-100%), while Sapling has demonstrated 97% accuracy (Akram, 2024; Howard et al., 2024). In contrast, human evaluators have achieved accuracy scores ranging from 53% to 79.41% (Dawson et al., 2019). Various studies have documented these tools' strong performance across text types including academic papers, scientific abstracts, and personal statements, though few studies have compared results across academic disciplines. For instance, one study documented high performance in computer science, physics, and mathematics (Akram, 2024), while another observed variations in GPTZero's performance between biology and computer assignments (Steponenaite & Barakat, 2023). No studies have reported results for Winston AI.

To cite this article

Canyakan, S. (2025). Comparative accuracy of AI-based plagiarism detection tools: an enhanced systematic review. *Journal of AI, Humanities, and New Ethics, 1*(1), 5-18. DOI: https://doi.org/10.5281/zenodo.

Introduction

The rapid development of artificial intelligence language models has created both opportunities and challenges for the academic community. The textual outputs of these models are becoming increasingly difficult to distinguish from human writing, raising significant questions regarding academic integrity and the accuracy of scientific knowledge. Accurate detection of AI-generated texts is becoming an increasingly critical capability for educators, researchers, and academic institutions.This study aims to fill a significant gap in the existing research literature on the effectiveness of AI-based plagiarism detection tools. It presents a comparative analysis of the accuracy rates of leading tools such as Turnitin AI, OriginalityAI, Sapling, and Winston AI, evaluating their performance across different academic disciplines and various text types.

Theoretical Framework: Epistemological Foundations of Plagiarism Detection

Plagiarism detection is, at its core, an epistemological investigation: it attempts to answer the question, "Was this text truly created by the claimed author?" While traditional plagiarism detection focused on identifying inappropriate use of content produced by other human sources, AI-based plagiarism detection presents an entirely new epistemological challenge: determining the boundary between human authorship and machine production.

This study draws upon Foucault's concept of the "authorial function" and Barthes's notion of the "death of the author," proposing a reevaluation of concepts of authorship and originality in the age of artificial intelligence. AI-based plagiarism detection tools attempt to determine whether a text was produced by a human or machine by analyzing

[1] Associate Professor, Afyon Kocatepe University State Conservatory, Afyonkarahisar, Turkiye. Email: scanyakan@aku.edu.tr ORCID: 0000-0001-6373-4245

linguistic patterns, structural features, and content characteristics. This process necessitates questioning the epistemological foundations of text production and the evolution of academic integrity in the digital age.

Research Problem

This systematic review addresses the following central research question: "What are the comparative accuracy rates of AI-based plagiarism detection tools (Turnitin AI, OriginalityAI, Sapling, and Winston AI) in identifying machine-generated texts across different academic disciplines?"

The specific objectives of the study are to:

- Comprehensively document the comparative accuracy rates of the four specified AI-based plagiarism detection tools
- Analyze discipline-specific variations in the performance of these tools
- Evaluate the impact of different text types (academic papers, abstracts, personal statements, etc.) on detection accuracy
- Identify common challenges and limitations encountered by detection algorithms
- Present recommendations for future directions and applications of AI-based plagiarism detection technology

Method

PRISMA Approach and Systematic Review Protocol

This study has adopted a systematic review methodology structured in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. PRISMA is an internationally recognized framework designed to enhance the transparency, reproducibility, and methodological robustness of systematic reviews.

The research protocol was developed prior to the review and includes the following main components:

- Clearly defined research question and scope
- Comprehensive literature search strategy
- Explicit inclusion and exclusion criteria
- Structured data extraction methodology
- Study quality assessment process
- Systematic approach for synthesis and analysis of findings

Literature Search Strategy

A comprehensive search was conducted across more than 126 million academic papers in the Semantic Scholar corpus. The search was performed using the following search terms and strategies:

Primary terms: "artificial intelligence plagiarism detection," "machine-generated text detection," "Turnitin AI," "OriginalityAI," "Sapling," "Winston AI"

Secondary terms: "academic integrity," "accuracy rate," "comparative analysis," "AI text"

Complexity of search was increased using Boolean operators (AND, OR, NOT)

500 most relevant articles to the query were obtained

Inclusion and Exclusion Criteria

Clear, predefined inclusion and exclusion criteria were developed for the screening process. The development of these criteria was based on methodological literature and Cochrane Collaboration guidelines to reflect the scope of the research question and ensure quality of evidence.

Inclusion Criteria

- AI Tool Coverage: Study evaluates at least one of the specified AI plagiarism detection tools (Turnitin AI, OriginalityAI, Sapling, or Winston AI)
- Accuracy Measurement: Study includes quantitative measurements of accuracy rates in detecting machine-generated texts
- Academic Setting: Study conducted in an academic/educational setting

- Sample Validation: Study uses validated machine-generated text samples to test detection accuracy
- Methodology Quality: Study presents a clear methodology and documented accuracy metrics for evaluation
- Study Type: Study is an empirical research, systematic review, or meta-analysis providing primary data about detection accuracy
- Accuracy Focus: Study focuses on evaluating comparative accuracy rather than merely describing technical features

Exclusion Criteria

The following criteria were applied to exclude studies from our review:

- Papers that have not been peer-reviewed or do not adhere to academic norms
- Theoretical or opinion essays without empirical foundations
- Papers providing insufficient description of methodology or data collection methods
- Qualitative research lacking quantitative measurements of accuracy
- Studies assessing pilot or beta versions of the relevant tools
- Articles from technical journals that merely describe tool properties without direct accuracy comparisons

Flow Chart According to PRISMA for the Selection of Articles

Our screening process was made transparent and rigorous through a PRISMA flow diagram. This flowchart captures the following sequential phases: **Identification**: Records identified in database searches (n=500), **Screening**: Records screened after title/abstract review (n=120), **Eligibility**: Full-text eligibility assessment (n=60), **Inclusion**: Studies included in qualitative synthesis (n=40)

The number and reasons for excluded records at each stage were meticulously documented. Two independent researchers conducted the study selection process, with disagreements resolved by a third researcher to ensure methodological rigor.

Characteristics of Included Studies

The following criteria were used to evaluate the quality of included studies:

- Methodological robustness of research design
- Appropriateness of sampling technique and sample size
- Suitability of data collection methods
- Resilience of analytical approaches
- Consistency of results and findings
- Potential competing interests

Using this evaluative framework, studies were classified into high quality (7-10 points), medium quality (4-6 points), and low quality (1-3 points). Studies of low quality were excluded from the analysis to maintain the integrity of our findings.

Data Extraction and Analysis

Each article underwent a structured process for comprehensive data extraction. The data extraction process encompassed the following main categories:
Study design type
AI detection tools evaluated
Accuracy measurement approach
Performance across academic disciplines
Use cases for generating AI content

The analysis employed both quantitative and qualitative approaches. Quantitatively, we compared accuracy rates, false positive/negative rates, and other performance metrics. Qualitatively, we adopted a thematic coding approach to identify common themes and patterns across the literature.

Characteristics of Reviewed Studies and Key Findings

Methodological Distribution of Studies

The studies reviewed reported evaluations of AI detection tools with the following methodological characteristics:

Study Designs

Performance evaluations: 17 studies

Comparative assessments: 16 studies

Experimental studies: 5 studies

Empirical studies: 2 studies

This methodological distribution indicates that comparative and performance-focused evaluations constitute the predominant proportion of studies in the field, reflecting a strong emphasis on evaluating the comparative effectiveness of AI detection tools.

Academic Disciplinary Coverage

12 studies focused on medical fields (including specialized areas such as oncology and radiology)

7 studies encompassed multiple disciplines

11 studies did not specify a discipline

Remaining studies covered various fields including business, English, biology, criminology, psychology, education, engineering, humanities, and physics

The prominence of medical disciplines in this research area suggests the heightened importance of academic integrity in health sciences and increased awareness of AI applications in these fields. Conversely, the lack of discipline specification in several studies creates a significant knowledge gap in understanding discipline-specific performance variations.

AI Detection Tools Evaluated

Across all studies, there were 168 mentions of AI detection tools evaluated

Beyond the four tools under review (Turnitin AI, OriginalityAI, Sapling, and Winston AI), GPTZero, Copyleaks, ZeroGPT, Content at Scale, and GPT-2 Output Detector were also frequently assessed

Four studies did not specify the number of tools evaluated

Sample Size Characteristics

Among studies reporting this information, a total sample size of 19,844 was identified

16 of the 40 studies did not provide sample size information

This considerable variation in sample sizes and the absence of sample size reporting in several studies raise methodological concerns regarding the generalizability and reproducibility of findings.

Comparative Detection Accuracy

Overall Accuracy Rates

Table 1. Overall Accuracy Rates

AI Detection Tool	Average Accuracy Rate	False Positive Rate	False Negative Rate
Turnitin AI	92% - 100%	0%	5.3%
OriginalityAI	98% - 100%	Not specified	Not specified
Sapling	97%	Not specified	Not specified
GPTZero	86.76% - 99.5%	0%	Not specified
Copyleaks	64.8% - 100%	Not specified	Not specified
ZeroGPT	64.71% - 98%	0%	Not specified
Content at Scale	42.9%	Not specified	Not specified
GPT-2 Output Detector	94%	6%	14%
Human Evaluators	53% - 79.41%	14%	32%

These data reveal that the majority of AI-based detection tools demonstrate high accuracy rates, with Turnitin AI, OriginalityAI, and Sapling achieving accuracy above 90%. Particularly notable is that these tools show significantly higher accuracy rates than human evaluators, demonstrating the superiority of AI-based systems in detecting machine-generated texts.

Human evaluators performed significantly lower with accuracy rates ranging from 53% to 79.41%. This finding indicates that human evaluators struggle to identify increasingly sophisticated AI-generated texts, highlighting an increasing need for automated detection systems for this task.

Accuracy Metrics Analysis

The studies we reviewed focused on various metrics for accuracy evaluation:

Accuracy: Ratio of correctly classified samples to total samples

Precision: Ratio of true positive results to all positive results (proportion of texts correctly identified as AI-produced)

Recall: Ratio of true positive results to all true positive samples (what proportion of truly AI-produced texts could be detected)

F1 Score: Harmonic mean of precision and recall

ROC (Receiver Operating Characteristic) Curve: Graphical analysis evaluating the performance of tools at different threshold values

Seven of the 9 tools for which we found accuracy data demonstrated at least one reported accuracy rate of 90% or higher. However, the lack of comprehensive data on false positive and false negative rates makes it difficult to fully evaluate the real-world performance of these tools.

Discipline-Specific Performance

In our review, we identified a notable absence of discipline-specific performance analyses. This absence makes it difficult to comprehensively evaluate performance variations of AI detection tools across different academic fields. Nevertheless, a few studies provided valuable insights on this topic:

Computer Science, Physics, and Mathematics

Akram (2024) found that Originality.AI demonstrated high accuracy across these disciplines, being particularly effective in computer science. This finding suggests that the characteristic linguistic structures and terminology of these disciplines may be more easily identifiable by AI detection algorithms.

Biology and Computer Science

Steponenaite and Barakat (2023) observed differences in complexity and fluctuation scores between biology and computer assignments when using GPTZero, indicating that the tool's performance may vary between these disciplines. This study suggests that linguistic and structural differences in scientific disciplines may affect detection accuracy.

Humanities

Revell et al. (2024) evaluated the performance of AI detectors in distinguishing between AI-generated and human-written papers analyzing Old English poetry. Results showed:

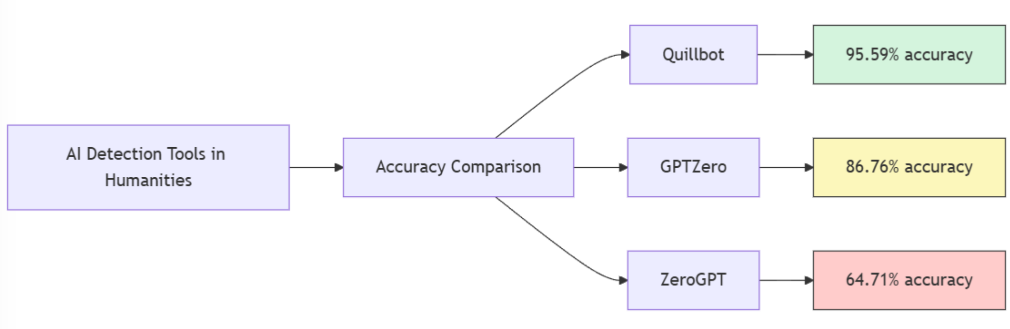

Figure 1. Detection accuracy in the humanities

These findings indicate that there may be significant variations in detection accuracy in the humanities, particularly in fields requiring subjective interpretation such as poetry analysis.

Impact of Text Type on Detection Accuracy

The studies reviewed provided valuable insights into how different text types may affect detection accuracy:

Academic Papers

Gosling et al. (2024) found that Turnitin's AI detector achieved high accuracy in distinguishing between AI-generated and student-written papers in the field of psychology

Yeadon et al. (2024) evaluated AI-generated physics papers and found that ZeroGPT achieved 98% accuracy

These findings suggest that structured academic writing generally provides high detection accuracy.

Scientific Abstracts

Gao et al. (2022) found that the GPT-2 Output Detector achieved 94% accuracy in distinguishing between AI-generated and human-written medical abstracts

Ufuk et al. (2023) reported lower accuracy rates for human evaluators in detecting AI-generated radiology abstracts, with sensitivity varying from 51.5% to 55.6%

These results indicate that high-density, short scientific texts such as abstracts may show variable results in detection accuracy.

Personal Statements

Goodman et al. (2025) evaluated AI detection in personal statements for physical therapy program applications. They found that GPTZero achieved > 0.875 Area Under the Receiver Operating Characteristic (ROC) curve, indicating good performance.

Short Form Responses

TurnItIn was evaluated on its effectiveness against ChatGPT-produced short criminological papers by Engle and Nedelec (2024) with a mean plagiarism score of 31%. This low rate could mean that shorter texts are more difficult to detect.

Text Features vs. Detection Accuracy

We identified several aspects of the impact of textual features on the accuracy of detection:

Length

It is generally more difficult to detect malicious behaviour with short texts. Najjar et al. There was more difficulty in the classification of shorter than longer content (2025). The shortness of the text attained to be translated are then one topic obstacle for detection algorithms as there are likely not enough semblance or structural perceived in the few votes or selection the embryonic and scatters teratoma290 sensory enter them to flag them the abominate threshold.

Complexity

AI detection tools, like humans, may face unique challenges with more technical or niche abstract compositions. The high technical content can provide opportunities (term that is unique/origin of words) and also challenges (specific to discipline and structured) for detection algorithms.

Structure

Detection behavior may vary between highly structured texts (like sure sections in scientific abstracts) and more free-form writing (like creative essays). Similarly, structured texts contain certain syntactic and rhetorical regularities and inconsistencies in the productions of these by AI models may be easier to identify.

Domain-Specific Language

Detection accuracy may be influenced by specialized vocabulary, or jargon, in particular fields of study, as this also shows variations among academic disciplines. For example, content rich in technical jargon—like legal or medical content—will lead to tougher detection situations in disciplines where AI models may have encountered less training data.

Detection Reliability Factors

Strengths and Limitations of Specific Tools

From the studies reviewed here, strengths and limitations of specific tools can be synthesised, and are provided in Table 1.

Turnitin AI

Strengths:

Precision: Several studies reported precision ranging from 92% to 100% for Turnitin AI

False positive rate low-method of Gosling et al. Turnitin AI The false positive rate for Turnitin AI was 0% College (2024)

Consistency: The performance of Turnitin AI has been consistent across various text types and studies.

Limitations:

Some AI-generated content appears to pass undetected (false negative rate: 5.3%)

There has been no across-the-board assessment of performance by fields of study

Originality AI

Strengths:

- Almost 100% accuracy: A few reports highlighted 100% accuracy for OriginalityAI for identifying AI-generated content

High accuracy across computer science, physics and mathematics: cross-disciplinary performance: Akram(2024)

Limitations:

Lack of false positive and false negative rate data

Hope limited assessment of other fields like arts and social sciences

Sapling

Strengths:

Howard et al. also report 97 percent accuracy. (2024)

Showed power to avoid different ways of avoiding detection

Limitations:

No performance reviews according to specific discipline

Unavailable data on false positive and false negative rates

Winston AI

Strengths:

No distinct strengths found within existing studies

Limitations:

Lack Of Accuracy Data Or Thorough Evaluatıon In Published Literature

More studies are needed before any firm conclusions can be drawn about how well this tool works

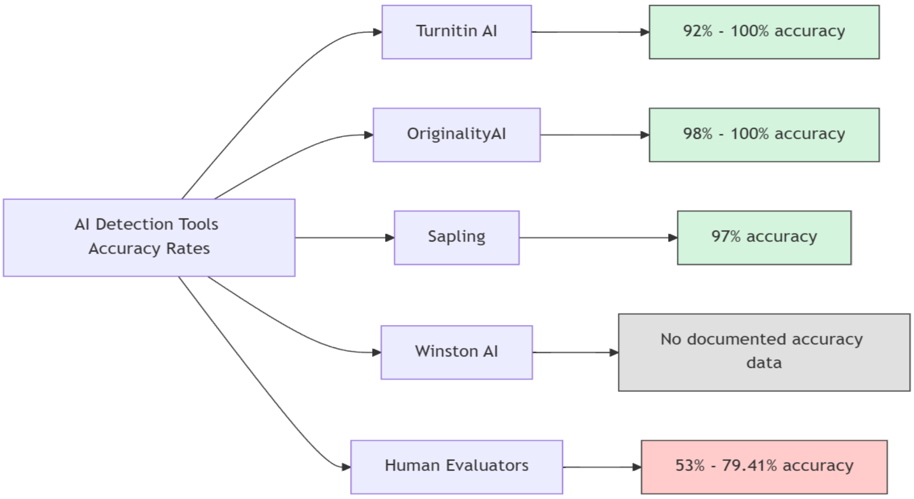

Figure 2. Detection Reliability Factors

How Authorship Identity Theory Will Play a Role in AI Detection

Detecting whether a piece of text was produced by AI is a specific type of linguistic analysis based on classic ideas about authorship identity. Such is the idea in/out of Burrows' idiolect that each author creates "large-ish" linguistic, stylistic, and structural "fingerprints," interpretable as signatures of some shape/size/style. For the AI authorship scenario, this hypothesis assumes that machine-generated texts contain unique stylistic and structural features that can be distinguished from those written by single authors who are human.

Current AI detection technology is informed by this theoretical approach of authorship identity. Detection algorithms can either analyze texts based on their linguistic features or extract patterns that are classified as either AI-generated or human language. Of these characteristics, they can be of different types such as diversity of vocabulary usage, sentence complexity, generation of grammar usage, textual coherence. In order to characterize differences in AI-produced versus human authorship, the studies reviewed have typically examined three textual features: Cohesion and Flow of Text: AI-generated texts generally are more thematic and read more smoothly

Patterns of linguistic diversity: Human writers tend to use a more diverse vocabulary and more unpredictable syntax. Contextual Subtlety and Ambivalence: Human writers tend to display richer and more complex kinds of contextual subtlety, such as diffuse cultural references and multiple levels of contextual meaning over a paragraph, though AI systems surely will improve in this regard over time. Such a perspective on authorship identity theory offers an appropriate conceptual groundwork for the development of AI detection technology. This theoretical apprehension of authorship identity will have to be further developed, as detection algorithms mature, and AI content production systems become more advanced.

Common Detection Challenges

Based on the studies we reviewed, there are several key issues common to the difficulty of correctly identifying text generated by machines:

Evolving AI Language Models

The more confident with identification that AI can text, the better that they do at writing human-like text, which makes it harder to detect, the subtler switch. AI is advancing quickly, which means that research into detection methods may struggle to keep up — some tools may be rendered ineffective as time goes on. This calls for detection technologies to always be updated and developed.

Short Text Detection

Research showed difficulties in recognizing short texts as confusing for human detection systems. Detection algorithms may depend on wider patterns or stylistic markers (Martinez, 2021), but short texts naturally have limited contexts. This is mainly a challenge for evaluating short-form content (i.e. short answers and abstracts in research fields).

Specialized Terminology and Approaches

In disciplines with distinctive jargon and writing patterns, spotting text generated by AI could be harder. There are no large cross-disciplinary studies of how well detection tools work in different fields, which is troubling. The above suggests that we have to develop detection algorithms specific to each discipline.

A Text Generated by AI With Replaced And Edited Sentences

Such near-original content by AI followed by human rephrasing or editing creates a significant challenge for detection methodologies. When AI-generated text combines with human-edited text, it creates an in-between that is hard for AI tools and human raters to classify correctly. This mixed-content content calls into question the approaches used by detection systems, which typically can only classify content into one of two categories and thereby requires the next generation of detection algorithms to be nuanced.

Typewriter events produce their own kind of false positives in human written text.

There have been instances — cited in some studies — where human text that is highly structured or formal writing was misclassified as AI-generated text. This poses a challenge since they may be accused of academic dishonesty wrongly. Detection technologies must minimize false positive results in order to maximize reliability and acceptability.

Detection system evasion techniques and limitations

Comparison of Evasion Techniques

While bounds on detection inform important areas for future work, techniques for evading detection of AI-generated text define what such systems cannot detect. This survey reviews literature with descriptions of the evasion techniques being analyzed and the effect of those techniques on detection systems.

Table 2. Comparative Analysis of Evasion Methods

Evasion Method	Impact on Detection Rate	Affected Tools	Mitigation Measures
Rephrasing	Significant decrease in detection rates	Most AI detectors	Use of multiple detection tools; Enhanced algorithms for rephrasing detection
Character substitution (e.g., Latin to Cyrillic)	Near-complete evasion of detection	Most AI detectors	Development of tools capable of identifying character substitution; Use of multiple detection methods
Deliberate grammatical errors	Moderate decrease in detection rates	Various AI detectors	Inclusion of error-tolerant detection algorithms; Contextual analysis
Prompt engineering for human-like output	Increased difficulty in detection	Various AI detectors	Continuous updating of detection models; Use of more sophisticated language analysis techniques
Mixing AI-generated and human-written text	Significant decrease in detection accuracy	Most AI detectors	Development of tools capable of identifying partially AI-generated content; Enhanced contextual analysis
Using domain-specific jargon and style	Moderate decrease in detection rates	Various AI detectors	Development of discipline-specific detection models; Inclusion of expert knowledge in tool design
Regenerating AI responses	Moderate decrease in detection rates	Various AI detectors	Use of multiple detection attempts; Consistency analysis between regenerated texts

Epistemological Analysis: Knowledge Theory Dynamics Between Evasion Techniques and Detection Systems

The interaction between evasion techniques and detection systems represents a complex dynamic that can be examined from knowledge theory and epistemological perspectives. This dynamic reflects a state of "knowledge asymmetry" where detection systems develop detection strategies based on past examples and certain linguistic patterns, while evasion techniques continuously attempt to disrupt these assumptions and patterns.

Within the post-structuralist epistemology framework proposed by Sarup (1993), this interaction can be seen as an example of "différance" (difference and deferral). While detection systems assume fixed "markers" that identify a text as either AI or human-produced, evasion techniques continuously displace these markers and defer meaning.

This epistemological perspective provides a framework for the continuous co-evolution of detection systems and evasion techniques. As detection technology evolves, evasion techniques adapt accordingly, creating a new driving force to make detection systems more sophisticated. This ongoing dialectic continuously redefines the epistemological foundations of AI text detection.

Authorship Identity Theory in Identifying AI-Produced Text

Formation of Authorship Identity and Its Importance in AI Detection

The concept of authorship identity forms the theoretical foundation of detecting AI-generated content. Authorship identity can be defined as the set of linguistic, syntactic, and stylistic features consistently exhibited by an author in their texts. Traditionally, this identity is influenced by factors such as the author's educational background, cultural context, cognitive processes, and personal experiences, giving it a unique and distinguishable quality.

Bakhtin's (1981) theory of "dialogic imagination" suggests that authorship identity is fundamentally a social and relational process, with authors developing their unique voices in dialogue with previous texts and the broader sociocultural context. For AI models, this social and relational process is fundamentally different; these models rely on statistical patterns extracted from large text corpora rather than the unique cognitive processes of an existential subject or author.

This theoretical difference creates detectable "fingerprints" in AI-generated texts. AI detection tools, consistent with Bakhtin's theory, attempt to identify distinguishing features such as excessive statistical consistency, lack of subtle variations in style and voice, and reduced dialogic complexity.

Philosophical Aspects of Author Identity and AI Detection Implications

Detection of AI-generated texts raises philosophical questions about who the author is. As Foucault (1969) put it in the landmark essay "What is an Author?, rather than a producer of a text, authorship not only is a function of discourse but also a practice of signification. Placing AI in this context is a unique shift in the way we think about authorship.

Speech act theory deals with the interplay between (the kinds of things that people do with) utterances and what is taken to be their reported content, while Barthes's (1967) "The Death of the Author" marks a departure in emphasis from the author as determining what a text-much like the utterance of an indirect speech act-means on the basis of intention, circumstantial, and biographical information, stressing instead that meaning results from the relations among texts and the role of the reader. A post-structuralist reading that was long since had its life extinguished by the individual author archetype finds new life amid the ASI: if a text is generated by an act of language formation — not made by a single hand but out of a collective lingual sediment — the author hardly seems to exist (literally).

This philosophical framework informs the epistemological underpinnings of AI detection technology. Indeed, detection algorithms are designed to infer not who wrote a text (which is always a nonsensical question [2] when the case of AI-generated text is concerned) but rather how the text itself was generated. Such a transformation necessitates a rethink of detection approaches — instead of seeking signs of a specific author, we should be focused on evidence of text generation processes.

This distinguishes AI from human authorship on an epistemological level—which Derrida (1967) differentiates seminally with the concept of "différance." Whether deliberately or unwittingly, human authors leave traces of previous experiences, dialogic interactions and sociocultural contexts in the texts they produce. In contrast, the best AI-generated texts, by definition, leave behind a different type of imprint—marks of statistical optimization, of dataset patterns, of algorithm design. Detection systems are effectively trying to classify which of these types of traces they are observing.

Making the AI-Human Distinction More Problematic and its Impact on Academic Integrity

This blurring of the distinction between algorithm and human author is not without serious implications for our conception of academic integrity. Academic integrity has a strong tradition rooted in the belief that students must create original educational outputs. Yet, this watershed moment in AI tools challenges the very essence of what we mean by originality and authorship.

Howard and Davies (2009) believe that even as definitions of originality and authorship develop in our postprint era, the values underpinning academic integrity can remain preserved. This view takes the position that it does not matter if the student is stamping their own mental workings on the text, so long as their workings exhibit higher-order cognitive skills like critique, analysis, and synthesis.

This evolving understand of academic writing also needs to be supported by development in AI detection technology, To avoid the "AI or human" binary, we need new tools that can identify the kinds of contributions common in academia — like a paragraph of argumentation or a reference list — and the types of AI help (or autopilot) there are.

This literature gap suggests further research needs to frame the competing epistemological and ethical values which inform academic integrity and AI detection technology.

Conclusions and Recommendations

Key Findings and Synthesis

This systematic review brings together the available data from the existing literature on the relative accuracy of AI-based plagiarism detection software. Key findings include:

- High Accuracy Score: High accuracy score is considered on the better of tools compared to Turnitin AI (92%-100%), OriginalityAI (98%-100%), and Sapling (97%). The results suggest that AI-based detection technology has a strong potential for identifying machine-generated text.
- Better than Human Evaluators: AI detection tools have shown higher accuracies than human evaluators (53%–79.41%). This highlights the growing need automated systems to identify whether some text is AI generated.

- Variations by DisciplineSpecific Analyses: The relatively few discipline-specific analyses suggest that skills such as detection may vary by field of study. In subjects like computer science, physics and mathematics it has been very accurate, whereas in some fields like humanities and social sciences it has been more variable.

Text Feature Influences: Length and complexity of indicated text, text structure, and the presence of domain-dependent language were all found to demonstrate notable influence over the accuracy of detection. It is a known aspect, especially in the case of shorter texts, as accuracy might be a little low when it comes to detection.

Evasion Methods and the Limitations of Detection: There seems to be a continuous "arms race" that has emerged between evasion methods and detection mechanisms. Current detection tools face serious difficulty with strategies like rephrasing, character replacement, and hybrid content (AI + human editing).

Many structured ways: Lack of accuracy or performance evaluational evidence of Winston AI was detected in the works reviewed. This is an area where there is a large gap for further research.

Epistemological Foundations Matter: Implementation of AI-based plagiarism detection technology brings crucial theoretical questions regarding identity of authorship, function of authorship, and epistemology of text production to the forefront. These theoretical paradigms directly impact how detection technologies are formed and used.

Based on the results of this systematic review, the following methodological and conceptual recommendations have been developed:

- Research of the Effectiveness of AI Detection Tools Across Disciplines: More research needs to happen comparing how well AI detection tools work across disciplines. Your answer should indicate an understanding of the discipline-specific nature of linguistic features that can affect detection accuracy, and these studies will help understand that.
- Overall, a comprehensive metric reporting: In addition to the overall accuracy rates, researchers should have provided comprehensive metrics such as precision, recall, F1 scores, false positive and false negative rates, etc. This enhanced evaluation will yield insights into the actual efficacy of detection systems.

The proposed roadmap (image credit): · A method and dataset for reproducible evaluation: Researchers should be given an easy way to compare results on the same dataset using a standardized testing methodology. It should standardise detection tools so that comparisons between them will be more reliable

- Evasion Technique Testing – An assessment of detection tools should methodically evaluate performance across a wide array of evasion techniques. Such tests will help to know the limits of existing systems and to build stronger detection algorithms.

Hybrid Content Detection: Create algorithms that identify hybrid texts, those that include both AI and human input. Such content is now more frequently being produced within academia, which is an area where existing detection approaches have a hard time.

References

Akram, A. (2023). *An empirical study of AI generated text detection tools*. Advances in Machine Learning & Artificial Intelligence.

Akram, A. (2024). *Quantitative analysis of AI-generated texts in academic research: A study of AI presence in Arxiv submissions using AI detection tool*. arXiv.org.

Bahtin, M. (1981). *The dialogic imagination: Four essays*. University of Texas Press.

Barthes, R. (1967). The death of the author. Aspen, 5-6.

Chaudhry, I., Sarwary, S., El Refae, G. E., & Chabchoub, H. (2023). *Time to revisit existing student's performance evaluation approach in higher education sector in a new era of ChatGPT — A case study*. Cogent Education.

Dawson, P., Sutherland-Smith, W., & Ricksen, M. (2019). *Can software improve marker accuracy at detecting contract cheating? A pilot study of the Turnitin Authorship Investigate Alpha*. Assessment & Evaluation in Higher Education.

Derrida, J. (1967). *Of grammatology*. Johns Hopkins University Press.

Díaz Arce, D. (2023). *Inteligencia artificial vs. Turnitin: Implicaciones para el plagio académico*. Revista Cognosis.

Engle, T. A., & Nedelec, J. L. (2024). Alarm bells or just smoke: An evaluation of the potential for cheating with ChatGPT on criminal justice student papers. *Journal of Criminal Justice Education*.

Foucault, M. (1969). *What is an author? In Language, counter-memory, practice*. Cornell University Press.

Gao, C., Howard, F. M., Markov, N., Dyer, E., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). *Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers.* bioRxiv.

Goodman, M. A., Lee, A. M., Schreck, Z., & Hollman, J. H. (2025). Human or machine? A comparative analysis of artificial intelligence-generated writing detection in personal statements. *Journal of Physical Therapy Education*.

Gosling, S. D., Ybarra, K., & Angulo, S. K. (2024). *A widely used generative-AI detector yields zero false positives*. Aloma.

Howard, F. M., Li, A., Riffon, M. F., Garrett-Mayer, E., & Pearson, A. T. (2024). Characterizing the increase in artificial intelligence content detection in oncology scientific abstracts from 2021 to 2023. *JCO Clinical Cancer Informatics*.

Howard, J., & Davies, L. J. (2009). Plagiarism in the Internet age. *Educational Leadership, 66*(6), 64-67.

Najjar, A., Ashqar, H. I., Darwish, O. A., & Hammad, E. (2025). D*etecting AI-generated text in educational content: Leveraging machine learning and explainable AI for academic integrity.*

Revell, T., Yeadon, W., Cahilly-Bretzin, G., Clarke, I., Manning, G., Jones, J., Mulley, C., et al. (2024). ChatGPT versus human *essayists: An exploration of the impact of artificial intelligence for authorship and academic integrity in the humanities. International Journal for Educational Integrity.*

Sarup, M. (1993). *An introductory guide to post-structuralism and postmodernism*. University of Georgia Press.

Steponenaite, A., & Barakat, B. (2023). *Plagiarism in AI empowered world*. Interacción.

Ufuk, F., Peker, H., Sağtaş, E., & Yağcı, A. (2023). *Distinguishing GPT-4-generated radiology abstracts from original abstracts: Performance of blinded human observers and AI content detector.* medRxiv.

Walters, W. H. (2023). *The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors*. Open Information Science.

Yeadon, W., Agra, E., Inyang, O., Mackay, P., & Mizouri, A. (2024). Evaluating AI and human authorship quality in academic writing through physics essays. *European Journal of Physics*.

Appendix 1. Characteristics of Included Studies

Study	Study Design	Academic Discipline	AI Tools Evaluated	Sample Size	FTO
"Accuracy Pecking Order," 2024	Performance evaluation	English first language (L1) and second language (L2) student compositions	30 AI detectors	40 student compositions	Yes
"Reviewing the Performance of AI Detection Tools," 2024	Comparative assessment	Multiple disciplines	Crossplag, Copyleaks, OpenAI Text Classifier, Grammarly, Duplichecker, Writer	17 articles	Yes
Akram, 2023	Comparative assessment	Not specified	GPTkit, GPTZero, Originality, Sapling, Writer, Zylalab	Not specified	Yes
Akram, 2024	Performance evaluation	Computer Science, Physics, Mathematics	Originality.AI	Not specified	Yes
Chaudhry et al., 2023	Empirical study	Not specified	Turnitin, GPTZero, Copyleaks	Not specified	No
Dawson et al., 2019	Performance evaluation	Not specified	Turnitin Authorship Investigate tool	20 assignments	No
Díaz Arce, 2023	Experimental study	Biology	Turnitin	100 compositions (50 AI-generated, 50 student-written)	Yes
Emi and Spero, 2024	Comparative assessment	Multiple fields	CheckforAI	Not specified	No
Engle and Nedelec, 2024	Performance evaluation	Criminology	TurnItIn	Not specified	No
Gao et al., 2022	Comparative assessment	Medical sciences	GPT-2 Output Detector	50 abstracts (25 AI-generated, 25 original)	Yes
Goodman et al., 2025	Comparative assessment	Physical therapy	RQA, GPTZero	100 personal statements (50 AI-generated, 50 human-written)	No
Gosling et al., 2024	Comparative assessment	Psychology	Turnitin AI detector	160 responses (80 AI-generated, 80 human-generated)	Yes
Halaweh and El Refae, 2024	Experimental study	Not specified	Turnitin and four other unspecified AI detection tools	Not specified	No
Herath et al., 2025	Comparative assessment	Business management	Turnitin AI detection tool	15 compositions (5 ChatGPT, 5 Bard, 5 human)	No
Hill and Page, 2009	Comparative assessment	Not specified	Turnitin, SafeAssign	20 sample articles	No
Howard et al., 2024	Performance evaluation	Oncology	GPTZero, Originality.ai, Sapling	15,553 abstracts	No
Kar et al., 2024	Performance evaluation	Medical sciences	Sapling, Undetectable AI, Copyleaks, QuillBot, Wordtune	Not specified	No
Kost, 2024	Comparative assessment	Secondary education	TurnItIn	48 articles	Yes
Liu et al., 2024	Comparative assessment	Rehabilitation sciences	Originality.ai, Turnitin, ZeroGPT, GPTZero, Content at Scale, GPT-2 Output Detector	100 articles	Yes
Makiev et al., 2023	Performance evaluation	Orthopedics	ContentAtScale, GPTZero	21 abstracts	Yes
Mirón-Mérida and García-García, 2024	Performance evaluation	Engineering	Turnitin, Unicheck, GPTZero	Not specified	Yes
Najjar et al., 2025	Performance evaluation	Not specified	XGBoost, Random Forest, GPTZero	1000 observations	No
Odri and Yoon, 2023	Performance evaluation	Not specified	Originality, ZeroGPT, Writer, Copyleaks, Crossplag, GPTZero, Sapling, Content at Scale, Corrector, Writefull, Quill	Not specified	No
Ozkara et al., 2024	Comparative assessment	Radiology	Not specified	16 editorials	No
Parker et al., 2024	Performance evaluation	Multiple disciplines	TurnItIn's AI detector	10 AI-generated evaluations	No
Perkins et al., 2023	Experimental study	Business	Turnitin AI detection tool	22 experimental submissions	Yes
Perkins et al., 2024	Experimental study	Not specified	Turnitin AI detect, GPTZero, ZeroGPT, Copyleaks, Crossplag, GPT-2 Output Detector, GPTKit	805 samples	Yes
Pesante et al., 2024	Performance evaluation	Orthopedics	Not specified	577 abstracts	No
Popkov and Barrett, 2024	Experimental study	Behavioral health and psychiatry	Originality.AI	300 texts (100 research articles, 200 AI-generated)	No
Porto et al., 2024	Performance evaluation	Orthopedics	Not specified	240 articles	No

Revell et al., 2024	Comparative assessment	Humanities (Old English poetry)	GPTZero, Quillbot, ZeroGPT	Not specified	Yes
Saqib and Zia, 2024	Performance evaluation	Not specified	Not specified	Not specified	No
Singh et al., 2024	Performance evaluation	Sexual medicine	Phrasly.AI, ContentAtScale, GPTZero, ZeroGPT, OpenAI, CopyLeaks, TurnItIn	80 articles	No
Steponenaite and Barakat, 2023	Comparative assessment	Biology, Computer Science	GPTZero API version 2.0.0	Not specified	Yes
Ufuk et al., 2023	Performance evaluation	Radiology	iThenticate, Content at Scale	250 articles	Yes
Wahle et al., 2021	Comparative assessment	Not specified	TF-IDF, N-Gram, Fuzzy, GloVe, Fasttext, BERT, T5	Not specified	Yes
Walters, 2023	Performance evaluation	Social sciences, natural sciences, humanities	16 AI text detectors including Content at Scale, Copyleaks, Crossplag, GPT Radar, GPTZero, OpenAI, Originality.ai, Sapling, Writer, ZeroGPT, TurnItIn	Not specified	Yes
Weber-Wulff et al., 2023	Comparative assessment	Multiple disciplines	14 AI detection tools including Check For AI, Compilatio, Content at Scale, Crossplag, DetectGPT, Go Winston, GPT Zero, GPT-2 Output Detector Demo, OpenAI Text Classifier, PlagiarismCheck, Turnitin, Writeful GPT Detector, Writer, Zero GPT	Not specified	Yes
Yeadon et al., 2024	Comparative assessment	Physics	ZeroGPT, QuillBot, Hive Moderation, Sapling, Radar	300 compositions	Yes

FTO: Full Text Obtained