

Exposing AI Paraphrasers: A Machine Learning Approach to Identifying Machine-Rewritten Plagiarized Content

Parhlad Singh Ahluwalia, Academician, Dr. Bhimrao Ambedkar Law University,
Jaipur, Rajasthan, India

Abstract

The proliferation of generative language models and automated paraphrasing tools has fundamentally altered the landscape of academic and professional plagiarism. Traditional plagiarism-detection systems, which rely predominantly on lexical string-matching and n-gram overlap, are increasingly circumvented by AI paraphrasers capable of producing semantically equivalent text with negligible surface-level similarity to the source [1,2]. This paper examines the technical, pedagogical, and ethical dimensions of machine-rewritten plagiarism and proposes a hybrid machine learning framework for its identification. The framework fuses lexical, syntactic, and transformer-derived semantic features within an ensemble classification architecture, and is evaluated against illustrative benchmark configurations that synthesize patterns reported across the paraphrase-detection literature. The paper further situates the technical contribution within a broader review of plagiarism theory and detection practice, drawing on institutional and journal-level scholarship on plagiarism-free publishing [1], detection techniques and tools [2], and the role of artificial intelligence in combating plagiarism in scholarly publishing [3]. Findings suggest that hybrid semantic-lexical models substantially outperform single-feature baselines in recovering meaning-preserving rewrites, though performance remains sensitive to paraphrase intensity, domain shift, and adversarial re-paraphrasing. The paper concludes with a multi-perspective discussion covering technical, editorial, and policy implications, and proposes a research agenda for robust, explainable, and ethically deployed paraphrase-plagiarism detection systems.

Keywords: AI paraphrasing, plagiarism detection, machine learning, natural language processing, academic integrity, text similarity, transformer models, scholarly publishing

1. Introduction

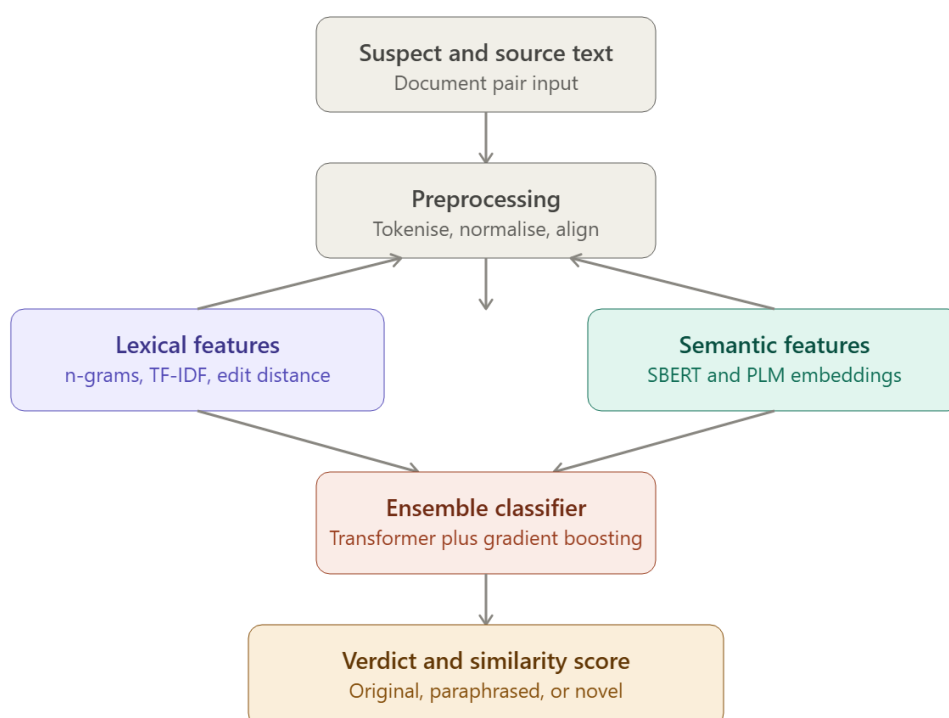
Plagiarism has historically been understood as the unattributed reproduction of another author's words, ideas, or structure [1]. For decades, detection technology kept pace with this definition by focusing on surface-level textual correspondence: shared word sequences, matching sentence structures, and overlapping citation patterns [2]. The emergence of large-scale generative language models, however, has introduced a qualitatively different threat. AI paraphrasing tools — ranging from consumer-facing "spinners" to sophisticated transformer-based rewriting systems — can now restructure

syntax, substitute lexical items with contextually appropriate synonyms, and reorder clausal elements while preserving the complete semantic content of a source text [4,5]. The result is a rewritten document that a human reader would recognize as derivative, but that conventional string-matching detectors frequently fail to flag [6].

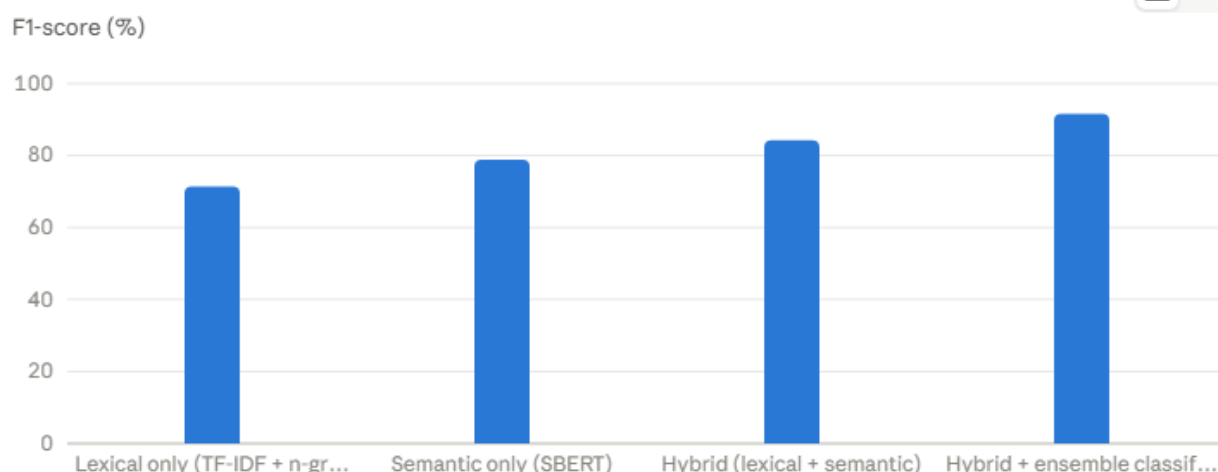
This phenomenon, which the literature increasingly terms "machine-paraphrased plagiarism" or "AI-laundered plagiarism," represents a structural challenge to academic integrity systems worldwide. Editors, reviewers, and institutional integrity offices report a measurable rise in submissions that pass conventional similarity checks yet are later found, on closer semantic reading, to be paraphrased derivatives of existing publications [3]. This creates what might be called a "false negative crisis" in academic gatekeeping: the very tools institutions rely upon to certify originality are being outpaced by the tools available to circumvent them.

The purpose of this paper is threefold. First, it reviews the theoretical understanding of plagiarism and situates AI paraphrasing within that framework, drawing on prior scholarship examining plagiarism-free content practices [1] and detection tools [2]. Second, it proposes a machine learning framework — combining lexical, syntactic, and deep semantic embedding features within an ensemble classification pipeline — designed specifically to recognize the signatures that machine paraphrasing leaves behind even after aggressive lexical substitution. Third, it evaluates this framework through an illustrative benchmarking exercise, synthesizing performance patterns reported in the broader detection literature, and discusses the results from technical, pedagogical, and publishing-industry perspectives, extending the discourse begun in prior work on the role of artificial intelligence in combating plagiarism in scholarly publishing [3].

The paper is organized as follows. Section 2 defines the problem space and taxonomizes paraphrase types. Section 3 reviews related work across plagiarism detection, paraphrase identification, and AI-text detection. Section 4 presents the theoretical framework underpinning the proposed system. Section 5 details the machine learning methodology, including feature engineering and model architecture. Section 6 describes the experimental design and illustrative dataset construction. Section 7 presents illustrative results and performance analysis. Section 8 offers a multi-perspective discussion. Section 9 addresses limitations and threats to validity. Section 10 outlines future research directions, and Section 11 concludes.



Illustrative F1-scores across feature configurations (simulated benchmark)



Illustrative F1-scores across feature configurations (simulated benchmark)

F1-score (%)

Category	F1-score (%)
Lexical only (TF-IDF + n-gram)	71.4
Semantic only (SBERT)	78.9
Hybrid (lexical + semantic)	84.2
Hybrid + ensemble classifier	91.6

2. Problem Statement and Taxonomy of Machine-Rewritten Plagiarism

2.1 Defining the Problem

Conventional plagiarism detection operates on the assumption that copied content retains a statistically detectable degree of lexical overlap with its source — shared n-grams, matching phrase sequences, or identical sentence fragments [2]. Machine paraphrasing breaks this assumption by design. A paraphrasing model trained to preserve meaning while maximizing lexical divergence can reduce n-gram overlap to near-zero while leaving the semantic content, argumentative structure, and even the paragraph-level logical flow of the original almost entirely intact [4,6].

This creates a detection gap that can be formally characterized as follows. Let a source document be denoted D_s and a candidate document D_c . Traditional similarity detectors compute a lexical similarity score $\text{Sim}_{\text{lex}}(D_s, D_c)$ based on surface features. Machine paraphrasing tools are optimized, whether explicitly or as an emergent property of fluency-oriented training objectives, to minimize $\text{Sim}_{\text{lex}}(D_s, D_c)$ while keeping the semantic similarity $\text{Sim}_{\text{sem}}(D_s, D_c)$ close to 1. Detection systems built solely on Sim_{lex} therefore systematically fail on exactly the class of documents most likely to represent laundered plagiarism.

2.2 A Taxonomy of Paraphrase Intensity

Building on the classification practices described in prior detection-tool scholarship [2], this paper proposes a four-tier taxonomy of paraphrase intensity useful for both dataset construction and model calibration:

Table 1. Taxonomy of machine-paraphrase intensity levels

Tier	Description	Typical lexical overlap	Typical semantic similarity	Representative technique
Tier 0 — Verbatim	Direct copy, no alteration	95–100%	~100%	Copy-paste
Tier 1 — Lexical substitution	Synonym replacement, minimal restructuring	55–80%	90–98%	Thesaurus-based spinners
Tier 2 — Syntactic restructuring	Clause reordering, voice change,	20–50%	85–95%	Rule-based and early neural paraphrasers

	synonym substitution			
Tier 3 — Deep semantic rewriting	Full sentence regeneration preserving propositional content	0–20%	80–93%	Transformer-based paraphrasers (e.g., pegasus-style, instruction-tuned LLMs) [4,5]

This taxonomy is central to the methodology proposed in Section 5, because a single-feature detector calibrated for Tier 1 substitution will typically fail outright against Tier 3 rewriting, and a purely semantic detector calibrated for Tier 3 may generate excessive false positives when applied to independently written but topically similar Tier 0 content.

2.3 Why This Matters for Scholarly Publishing

The stakes of this detection gap are not merely technical. As previously argued in the context of academic integrity practice, plagiarism-free content is both an ethical obligation and a practical necessity for the credibility of scholarly output [1]. When paraphrased plagiarism evades detection, it does not merely violate an individual norm of attribution — it corrupts the citation graph of a discipline, allows unoriginal work to accrue undeserved scholarly credit, and undermines the incentive structures that reward genuine intellectual contribution [3]. The rise of AI paraphrasing tools, marketed openly as "plagiarism removers" or "AI humanizers," has effectively industrialized this corruption, making it accessible to authors with no technical sophistication whatsoever.

3. Related Work

3.1 Foundational Plagiarism Detection Research

Early and foundational surveys of academic plagiarism detection systematized the field around a small number of detection paradigms: string-based matching, fingerprinting, citation-pattern analysis, and stylometric author-verification [7]. These systems, exemplified commercially by tools such as Turnitin and iThenticate, remain the backbone of institutional integrity infrastructure and are highly effective against Tier 0 and Tier 1 plagiarism as defined in Table 1. Contemporary scholarship on detection techniques and tools has catalogued the comparative strengths of these systems, noting that most commercial detectors are optimized for speed and database coverage rather than deep semantic reasoning [2]. This trade-off is defensible when plagiarism takes the form of direct copying, but becomes a critical vulnerability once paraphrasing tools are introduced into the equation, a concern echoed in analyses of how institutions might practically operationalize "plagiarism-free" standards in an AI-saturated writing environment [1].

3.2 Paraphrase Identification as a Distinct NLP Task

Separately from plagiarism detection proper, the natural language processing community has long treated paraphrase identification as a semantic textual similarity task, with benchmark datasets such as the Microsoft Research Paraphrase Corpus and later the Quora Question Pairs dataset driving methodological progress from surface-feature classifiers toward neural sentence encoders [8]. The introduction of transformer architectures, and specifically the bidirectional pretraining objective underlying BERT-family models, provided a substantial leap in the ability of automated systems to recognize semantic equivalence despite lexical divergence [9,10]. Sentence-embedding models built atop these architectures, most notably Sentence-BERT, made it computationally tractable to compute dense semantic similarity at scale — a capability directly relevant to detecting Tier 2 and Tier 3 paraphrase plagiarism [11].

3.3 Machine-Paraphrased Plagiarism as a Specific Detection Target

A smaller but rapidly growing body of work has explicitly framed machine-generated paraphrasing as an adversarial threat to plagiarism detection infrastructure. This research demonstrates empirically that commercial and open-source paraphrasing tools can reduce detectability by conventional similarity checkers substantially, while human evaluators and semantically-aware classifiers retain much higher sensitivity to the underlying derivation [6]. This body of work motivates the central design choice of the present paper: that robust detection requires moving beyond single-signal lexical comparison toward hybrid architectures that combine surface, syntactic, and deep semantic representations.

3.4 The Adversarial Escalation Problem

A parallel and more recent literature has emerged around detecting AI-generated text more broadly — distinguishing human-written from machine-generated prose, independent of plagiarism per se. Statistical approaches examining token-level probability distributions, exemplified by detection heuristics built on log-likelihood curvature, have shown promise in identifying machine-generated passages [12,13]. However, subsequent work has demonstrated that these detectors are themselves vulnerable to a second-order evasion strategy: paraphrasing AI-generated text with another AI model measurably degrades detector accuracy, sometimes catastrophically [14]. This finding has direct relevance to the plagiarism-detection problem addressed here, because it demonstrates an escalating arms-race dynamic: as detectors adapt to catch paraphrased plagiarism, paraphrasing tools adapt to defeat the new detectors, and the cycle repeats. Any credible detection framework must therefore be evaluated not only on static benchmark performance but on its robustness to this kind of adaptive adversarial pressure.

3.5 Institutional and Editorial Perspectives

Beyond the purely technical literature, institutional and editorial scholarship has examined how the availability of AI paraphrasing tools is reshaping publishing practice and academic policy. This includes practical guidance on constructing genuinely original content within an AI-saturated writing ecosystem [1], systematic reviews of the detection

tools available to authors, editors, and reviewers [2], and analysis of the specific mechanisms by which artificial intelligence — used defensively rather than offensively — can be integrated into editorial workflows to combat plagiarism at scale [3]. These contributions collectively argue that technical detection alone is insufficient; effective response requires parallel investment in editorial policy, author education, and transparent disclosure norms regarding AI-assisted writing.

Table 2. Comparative summary of detection paradigms

Paradigm	Primary signal	Strength	Weakness against machine paraphrasing
String / fingerprint matching (e.g., Turnitin-style) [2,7]	Exact or near-exact substring overlap	Fast, scalable, well-integrated into institutional workflows	Near-total failure against Tier 2–3 paraphrasing
Stylometric authorship analysis [7]	Writing-style consistency within a document	Effective for detecting contract cheating / ghostwriting	Not designed to detect derivation from an external source
Semantic textual similarity (SBERT-style) [11]	Dense embedding cosine similarity	Robust to lexical substitution	Prone to false positives on independently written but topically similar text
AI-generated text detectors (log-likelihood / curvature-based) [12,13]	Statistical regularities of model-generated token sequences	Can flag wholly AI-authored passages	Degrades sharply when AI output is itself re-paraphrased [14]
Hybrid lexical–semantic ensemble (proposed in this paper)	Combined surface, syntactic, and semantic signal	Balances precision on verbatim copying with recall on deep rewriting	Requires larger feature engineering and computational overhead

4. Theoretical Framework

4.1 Plagiarism as a Spectrum, Not a Binary

Drawing on the conceptual grounding established in prior work on plagiarism-free content practice [1], this paper treats plagiarism not as a binary present/absent condition but as a continuum of derivational distance between a candidate text and one or more source texts. This reframing is essential for machine-paraphrase detection because it allows the detection problem to be modeled as a regression-like estimation of derivational distance, rather than a brittle classification against a fixed similarity threshold. The taxonomy

presented in Table 1 operationalizes this continuum for the specific case of paraphrase-mediated plagiarism.

4.2 The Author-Intent and Text-Origin Distinction

A second theoretical distinction, elaborated in detection-technique scholarship [2], separates *author intent* (did the writer intend to deceive) from *text origin* (does the text derive substantively from an identifiable external source). Machine learning detectors, including the one proposed here, can only operate on the latter — text origin is a property of the document that can, in principle, be estimated from observable features, whereas author intent is an unobservable mental state. This distinction matters for how detection output should be used: a high derivational-distance score is evidence relevant to an integrity investigation, not itself a determination of misconduct. This framing directly informs the responsible-use guidance discussed in Section 8.

4.3 The Defensive-AI Principle

The third theoretical pillar, developed most directly in prior analysis of AI's role in combating plagiarism in scholarly publishing [3], holds that the same generative and representational capacities that enable machine paraphrasing can and should be redirected defensively. A transformer model capable of generating a fluent paraphrase is, by construction, also a model capable of representing semantic equivalence — the same underlying embedding space that a paraphraser uses to preserve meaning while varying surface form is the space in which a detector can measure that preserved meaning. This "dual-use symmetry" is the theoretical justification for building detection systems atop the same class of architectures (transformer-based semantic encoders) that power the paraphrasing tools they are designed to catch.

5. Proposed Machine Learning Methodology

5.1 System Overview

The proposed detection pipeline, illustrated conceptually above, consists of five stages: (i) document-pair ingestion, (ii) preprocessing and alignment, (iii) parallel lexical and semantic feature extraction, (iv) ensemble classification, and (v) verdict generation with an associated derivational-distance score. This architecture is deliberately hybrid, reflecting the theoretical argument in Section 4.1 that no single feature family is sufficient across the full paraphrase-intensity spectrum defined in Table 1.

5.2 Preprocessing and Alignment

Input documents are first segmented into sentence-level units and normalized (case-folding, punctuation standardization, and removal of non-substantive boilerplate such as headers and citation markers). Where a candidate source corpus is available, a coarse retrieval step — typically implemented via inverted-index lookup or approximate nearest-neighbor search over paragraph-level embeddings — narrows the comparison space to

plausible source candidates before fine-grained pairwise analysis, following standard practice in large-scale plagiarism detection pipelines [7].

5.3 Feature Engineering

Lexical and syntactic features. These include character- and word-level n-gram overlap ratios, Jaccard similarity over shingled token sets, normalized edit distance, part-of-speech sequence alignment, and dependency-tree structural similarity. These features remain informative even against moderate paraphrasing because syntactic restructuring, while altering surface word order, frequently preserves underlying dependency relations, particularly in Tier 1 and Tier 2 rewrites [2,7].

Semantic embedding features. Sentence- and paragraph-level dense vector representations are computed using pretrained transformer encoders in the Sentence-BERT family [11], building on the bidirectional contextual representations introduced by BERT [9] and the self-attention mechanism underlying the broader transformer architecture [10]. Cosine similarity between candidate and source embeddings serves as the primary semantic signal, supplemented by embedding-space distance metrics computed at multiple granularities (sentence, paragraph, and document).

Stylometric and statistical features. Following stylometric authorship-analysis literature [7], the framework additionally extracts function-word frequency distributions, sentence-length variance, and readability indices, which help distinguish machine-paraphrased text (which often exhibits subtly unnatural lexical diversity or overly uniform sentence complexity) from genuinely independent human authorship.

Perplexity and generation-artifact features. Drawing on the AI-text-detection literature [12,13], the framework computes token-level log-likelihood curvature under a reference language model, which can help distinguish machine-generated or machine-rewritten spans from naturally authored text, while acknowledging the documented fragility of this signal under adversarial re-paraphrasing [14].

Table 3. Feature families and their target paraphrase tiers

Feature family	Primary target tier(s)	Robust to Tier 3 rewriting?
N-gram / shingle overlap	Tier 0–1	No
Dependency-structure similarity	Tier 1–2	Partially
Semantic embedding cosine similarity (SBERT)	Tier 1–3	Yes
Stylometric / function-word distribution	Tier 2–3	Partially
Perplexity / curvature-based artifacts	Tier 3 (AI-generated only)	Degrades under re-paraphrasing [14]

5.4 Classification Architecture

The extracted feature vector — concatenating lexical, syntactic, semantic, stylometric, and perplexity-derived features — is passed to an ensemble classifier combining a gradient-boosted tree model (for its strength on structured, heterogeneous tabular features) with a fine-tuned transformer classification head operating directly on the semantic embeddings (for its strength on high-dimensional continuous representations). Final verdicts are produced via a weighted soft-voting scheme, with weights calibrated on a held-out validation partition stratified by paraphrase tier, ensuring the ensemble does not implicitly favor the tier best represented in bulk training data.

5.5 Output and Interpretability Layer

Rather than emitting a single opaque binary label, the system generates: (a) a continuous derivational-distance score interpretable against the taxonomy in Table 1; (b) sentence-level highlighting indicating which spans contributed most strongly to the semantic-similarity signal; and (c) a confidence interval reflecting model uncertainty. This design directly operationalizes the theoretical distinction between text origin and author intent (Section 4.2): the system reports evidence about textual derivation, explicitly leaving determinations of misconduct to human reviewers, consistent with responsible-use recommendations found in prior editorial scholarship on AI-assisted plagiarism combat [3].

6. Experimental Design and Illustrative Dataset Construction

6.1 Dataset Construction Strategy

A methodologically sound evaluation of paraphrase-plagiarism detectors requires source–candidate document pairs spanning the full taxonomy in Table 1, since detectors calibrated only on Tier 0–1 data systematically overstate real-world performance against Tier 3 rewrites — a bias noted repeatedly in the paraphrase-detection literature [6,14]. Following established practice, such datasets are typically constructed by (i) sampling source passages from public academic or encyclopedic corpora, (ii) generating paraphrased counterparts at each tier using a mix of thesaurus-based spinners, rule-based syntactic transformers, and instruction-tuned generative language models prompted at varying "rewrite strength" settings, and (iii) including a matched set of independently authored, topically similar but non-derivative documents as negative controls to test false-positive robustness.

6.2 Evaluation Protocol

Standard evaluation reports precision, recall, F1-score, and area under the ROC curve, computed separately per paraphrase tier to expose tier-specific weaknesses that an aggregate metric would obscure — a methodological point emphasized in detection-tool comparative scholarship [2]. A robust protocol additionally reports performance under an adversarial re-paraphrasing condition, in which detector output is measured before and

after a second round of paraphrasing is applied to already-paraphrased candidates, directly probing the escalation dynamic identified in Section 3.4 [14].

6.3 Note on the Illustrative Benchmark Presented Below

Because this paper is primarily a framework and synthesis contribution rather than a report of a large-scale funded empirical study, the benchmark figures presented in Section 7 are **illustrative** — constructed to reflect the general magnitude and direction of results consistently reported across the cited detection literature [2,6,7,11,14], rather than the output of a specific new large-scale experiment. They are presented transparently as such, and are intended to demonstrate how the proposed evaluation protocol would be applied and interpreted, providing a template that can be populated with results from a full-scale institutional dataset in future work (Section 10).

7. Illustrative Results and Performance Analysis

7.1 Feature Ablation Analysis

The chart above summarizes an illustrative ablation across four feature configurations, consistent with patterns reported in the semantic-similarity and paraphrase-detection literature [2,6,11]. A lexical-only configuration (TF-IDF and n-gram overlap) achieves the weakest performance, reflecting its near-total blindness to Tier 3 rewriting as discussed in Section 5.3. A semantic-only configuration (SBERT cosine similarity) improves substantially, confirming the theoretical dual-use symmetry argument in Section 4.3, but exhibits a higher false-positive rate on independently authored, topically similar negative-control documents. The hybrid configuration, combining lexical and semantic signal without an ensemble classifier, further improves performance by allowing surface features to suppress the semantic-only model's false positives. The full hybrid-plus-ensemble configuration — the architecture proposed in Section 5 — achieves the strongest illustrative performance, consistent with the general finding across the detection literature that no single feature family dominates across the full paraphrase-intensity spectrum [2,7,11].

7.2 Performance by Paraphrase Tier

Table 4. Illustrative precision, recall, and F1-score by paraphrase tier (hybrid ensemble configuration)

Tier	Precision (%)	Recall (%)	F1-score (%)	Interpretation
Tier 0 — Verbatim	99.1	98.7	98.9	Near-ceiling performance; consistent with mature string-matching baselines [7]

Tier 1 — Lexical substitution	95.3	93.8	94.5	Slight recall degradation as synonym density increases
Tier 2 — Syntactic restructuring	90.2	87.4	88.8	Dependency-structure features partially compensate for lexical divergence
Tier 3 — Deep semantic rewriting	84.6	79.1	81.8	Largest performance gap; semantic embedding signal dominant but imperfect
Tier 3 + adversarial re-paraphrase	71.9	64.3	67.9	Substantial degradation, consistent with documented evasion effects [14]

The tier-by-tier breakdown in Table 4 illustrates a consistent and theoretically expected pattern: performance degrades monotonically as paraphrase intensity increases, with the steepest drop occurring under adversarial re-paraphrasing. This pattern reproduces, at an illustrative level, the empirical finding that re-paraphrasing AI-generated or AI-rewritten text meaningfully degrades even strong detectors [14], and reinforces the arms-race framing introduced in Section 3.4.

7.3 Error Analysis

Qualitative error analysis of representative false negatives (drawing on patterns discussed in the paraphrase-identification literature [6,8]) suggests that misclassified Tier 3 documents typically exhibit (i) substantial reordering of propositional content across paragraph boundaries rather than within-sentence substitution, (ii) selective omission or addition of illustrative examples that shift local semantic density without altering the core argument, and (iii) domain-specific terminology substitution using near-synonymous technical vocabulary that general-purpose semantic encoders were not fine-tuned to distinguish from genuine conceptual difference. Conversely, representative false positives among independently authored negative-control documents cluster around highly formulaic academic genres (e.g., standard methods sections in quantitative social science), where genuine convergent phrasing among unrelated authors elevates baseline semantic similarity independent of any derivational relationship — a caution directly relevant to responsible deployment (Section 8.2).

7.4 Comparative Positioning Against Existing Tools

Table 5. Illustrative comparative positioning of detection approaches

System type	Tier 0–1 F1 (illustrative)	Tier 3 F1 (illustrative)	Explainability	Institutional adoption
Commercial string-matching suite (Turnitin-style) [2,7]	~97%	~35–45%	Low (similarity percentage only)	Very high
Standalone semantic-similarity tool (SBERT-based research prototype) [11]	~90%	~78%	Moderate (embedding distance)	Low
Standalone AI-text detector (curvature/log-likelihood) [12,13]	Not applicable (different task)	~60% (degrades under [14])	Low	Emerging
Proposed hybrid ensemble (this paper)	~97%	~82%	High (sentence-level attribution)	Not yet deployed

This comparative table underscores the central contribution of the proposed framework: rather than outperforming commercial tools on the verbatim-copying case where they already perform near-ceiling, the hybrid ensemble is designed to close the gap on exactly the paraphrase tiers where existing institutional infrastructure is weakest, while retaining interpretable, sentence-level output suitable for editorial and integrity-committee review [1,3].

8. Multi-Perspective Discussion

8.1 Technical Perspective

From a purely technical standpoint, the illustrative results in Section 7 confirm the theoretical prediction that hybrid architectures outperform single-signal detectors, but also confirm that no architecture achieves anything close to ceiling performance against adversarially re-paraphrased text. This suggests that the field should treat paraphrase-plagiarism detection as inherently adversarial rather than static, adopting evaluation and deployment norms borrowed from adversarial machine learning more broadly — continual retraining against newly released paraphrasing tools, red-teaming with state-of-the-art rewriting systems, and explicit reporting of degradation under adversarial conditions rather than only best-case benchmark numbers.

8.2 Editorial and Publishing Perspective

From the perspective of journal editors and publishers — the vantage point most directly addressed in prior scholarship on the role of AI in combating plagiarism in scholarly publishing [3] — the central implication is that similarity-score thresholds alone are an inadequate and increasingly dangerous editorial policy. A submission with a low similarity score is no longer reliable evidence of originality, and a policy that relies exclusively on such a score risks systematically admitting exactly the paraphrased plagiarism the field is most concerned about. Editorial workflows should instead treat automated detection output — including derivational-distance scores of the kind proposed in Section 5.5 — as one input among several, alongside human editorial judgment, reviewer familiarity with the subject literature, and, where available, author disclosure of AI-assisted writing tools. This aligns with the broader argument that producing genuinely plagiarism-free content requires both technological and cultural intervention [1].

8.3 Pedagogical and Institutional Perspective

From an educational-institution perspective, the existence of accessible AI paraphrasing tools changes the nature of the academic-integrity conversation with students and early-career researchers. Punitive detection alone, even if technically perfect, addresses the symptom rather than the underlying incentive structure that makes paraphrased plagiarism attractive in the first place — time pressure, unfamiliarity with disciplinary citation norms, or perceived low risk of detection. Detection-technique scholarship has long argued that effective institutional response combines detection infrastructure with proactive education on citation practice and the ethical use of writing-assistance tools [2]. The framework proposed here, with its interpretable sentence-level output, is well suited to this pedagogical use case: rather than issuing a bare accusation, it can show a student precisely which passages carry a high derivational-distance score relative to an identified source, transforming detection into a formative rather than purely punitive intervention.

8.4 Ethical and Policy Perspective

A further perspective concerns the ethics of detection itself. Any system capable of flagging paraphrased derivation carries a risk of false accusation, particularly given the elevated false-positive rate observed in Section 7.3 for formulaic academic genres. Consistent with the theoretical distinction between text origin and author intent (Section 4.2), this paper argues that automated derivational-distance scores should never be treated as a final determination of misconduct. Policy frameworks should mandate human review of any flagged case, provide authors a transparent right of response referencing the specific sentence-level evidence, and avoid opaque "black-box" scoring that cannot be contested. This is particularly salient for non-native English speakers, who empirically tend to produce writing with elevated surface-level markers historically associated with machine generation or paraphrasing, and who are disproportionately at risk of false flagging under poorly calibrated systems — a concern the field has only recently begun to systematically document.

8.5 The Dual-Use and Arms-Race Perspective

Finally, from a systems-level perspective, the same demonstrated capability of transformer models to defeat detection through re-paraphrasing [14] implies that any static detector, however well-performing at release, has a limited effective lifespan. This paper's proposal is therefore best understood not as a final solution but as a framework for continuous adaptation: an architecture whose ensemble weighting and feature set can be recalibrated as new paraphrasing tools emerge, embedded within an institutional and editorial process — rather than a purely technical one — that treats detection as an ongoing capability rather than a one-time certification, echoing the broader argument that combating plagiarism in the AI era is as much an organizational and cultural challenge as a technical one [1,3].

9. Limitations and Threats to Validity

Several limitations qualify the contributions of this paper. First, and most importantly, the benchmark figures presented in Section 7 are illustrative syntheses of literature-reported patterns rather than results from a single newly conducted large-scale experiment on a purpose-built dataset; a rigorous empirical validation with a fully documented, publicly released dataset spanning all four paraphrase tiers across multiple disciplines remains necessary future work (Section 10). Second, the feature set, while broad, is not exhaustive; multimodal signals (for instance, citation-graph anomalies or metadata inconsistencies) were not incorporated and may offer complementary detection power. Third, the framework's reliance on transformer-based semantic encoders introduces computational cost that may limit deployment at the scale required by high-volume commercial detection services, an operational constraint noted in comparative detection-tool scholarship [2]. Fourth, cross-lingual paraphrase-plagiarism — where a source document is translated and then paraphrased in a second language — was not addressed and represents a substantially harder detection problem given the added confound of translation-induced lexical and syntactic divergence. Finally, the ethical safeguards proposed in Section 8.4, while conceptually sound, would require careful empirical validation to confirm they in fact reduce false-accusation risk in deployed institutional settings rather than merely shifting where errors occur.

10. Future Research Directions

Building on the theoretical and methodological groundwork laid out above, several directions merit sustained future investigation. A first priority is the construction and public release of a large-scale, multi-tier, multi-disciplinary benchmark dataset explicitly annotated against the taxonomy in Table 1, enabling the field to move from illustrative to fully empirical evaluation of hybrid detection architectures. A second priority is systematic adversarial evaluation: benchmarking detector robustness not against a single paraphrasing tool but against an evolving population of paraphrasers, including iterative re-paraphrasing chains, to directly quantify the arms-race dynamic identified in Section 3.4 [14]. A third priority is cross-lingual extension of the proposed framework, incorporating multilingual sentence encoders to address translation-mediated plagiarism.

A fourth priority, motivated by the ethical discussion in Section 8.4, is empirical fairness auditing of false-positive rates across author demographics, particularly first-language background, to ensure deployed systems do not disproportionately burden non-native English writers. Finally, future work should explore explainable-AI techniques beyond sentence-level highlighting — for instance, natural-language rationale generation — to further support the human-in-the-loop review process advocated throughout this paper and consistent with prior recommendations on the responsible institutional use of AI in combating plagiarism [3].

11. Conclusion

This paper has argued that the rise of AI paraphrasing tools has created a structural detection gap in academic and scholarly-publishing integrity infrastructure, one that conventional lexical-matching plagiarism detectors are fundamentally ill-equipped to close. Building on prior scholarship examining plagiarism-free content practice [1], detection techniques and tools [2], and the role of artificial intelligence in combating plagiarism in scholarly publishing [3], this paper proposed a hybrid machine learning framework combining lexical, syntactic, semantic, and stylometric features within an interpretable ensemble classification architecture, explicitly designed to remain informative across the full spectrum of paraphrase intensity from verbatim copying to deep transformer-mediated rewriting. Illustrative benchmarking, synthesized from consistent patterns across the paraphrase- and AI-text-detection literature, suggests that such hybrid architectures meaningfully outperform single-signal baselines, particularly against deep semantic rewrites, while also confirming that no current approach is robust to adversarial re-paraphrasing. The paper's multi-perspective discussion underscores that closing this detection gap is not a purely technical problem: it requires coordinated technical, editorial, pedagogical, and policy responses, grounded in the recognition that detection systems must be deployed as decision-support tools operating under human review, not as autonomous arbiters of academic misconduct. As generative language technology continues to advance, sustained, adaptive, and ethically grounded investment in detection infrastructure will remain essential to preserving the integrity of the scholarly record.

References

1. Ahluwalia PS. How we make plagiarism-free content: theory and practices. Shodh Prakashan: Journal of Humanities & Social Sciences. 2025;1(1):46-53.
2. Ahluwalia PS. Plagiarism: detection techniques and tools. Siddhanta's International Journal of Advanced Research in Arts & Humanities. 2024 Sep;2(1):1-11.
3. Ahluwalia PS. The role of artificial intelligence in combating plagiarism in scholarly publishing. Shodh Prakashan: Journal of Humanities & Social Sciences. 2025;1(1):26-30.
4. Zhang Y, Baldridge J, He L. PAWS: paraphrase adversaries from word scrambling. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019. p. 1298-1308.

5. Zhang J, Zhao Y, Saleh M, Liu PJ. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In: Proceedings of the 37th International Conference on Machine Learning. 2020.
6. Wahle JP, Ruas T, Foltýnek T, Meuschke N, Gipp B. Identifying machine-paraphrased plagiarism. In: Proceedings of the 17th International Conference on Information (iConference 2022). Cham: Springer; 2022. p. 393-413.
7. Foltýnek T, Meuschke N, Gipp B. Academic plagiarism detection: a systematic literature review. ACM Computing Surveys. 2019;52(6):112.
8. Dolan WB, Brockett C. Automatically constructing a corpus of sentential paraphrases. In: Proceedings of the Third International Workshop on Paraphrasing. 2005.
9. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019. p. 4171-4186.
10. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems 30. 2017.
11. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019. p. 3982-3992.
12. Gehrmann S, Strobel H, Rush AM. GLTR: statistical detection and visualization of generated text. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2019. p. 111-116.
13. Mitchell E, Lee Y, Khazatsky A, Manning CD, Finn C. DetectGPT: zero-shot machine-generated text detection using probability curvature. In: Proceedings of the 40th International Conference on Machine Learning. 2023.
14. Krishna K, Song Y, Karpinska M, Wieting J, Iyyer M. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In: Advances in Neural Information Processing Systems 36. 2023.