A Multi Layer Forensic Similarity Framework for Detecting Transformed and AI Generated Content

Gundu Bhavana
Department of Computer Science and Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Chennai, India
bhav8157@gmail.com

Dhanyasree Thallapalli
Department of Computer Science and Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Chennai, India
dhanya100105@gmail.com

Dr. S. Saravanan
Department of Computer Science and Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Chennai, India
s_saravanan@ch.amrita.edue

***Abstract*—The rapid growth of large language models (LLMs) and automated paraphrasing systems has exposed significant weaknesses in traditional plagiarism detection tools. Most existing systems rely heavily on surface-level lexical overlap and string matching, making them vulnerable to paraphrased, structurally modified, or semantically preserved content. As generative models increasingly assist in academic and journalistic writing, detecting transformed reuse requires analysis beyond direct text similarity. This paper presents a Multi-Layer Forensic Similarity Framework designed to identify AI-assisted and heavily paraphrased content through integrated structural, lexical, semantic, entity-based, syntactic, and stylometric analysis.**

The system was evaluated using a curated corpus of 145 news articles stored in a structured JSON repository. Each document undergoes multi-stage preprocessing to extract n-gram fingerprints, probabilistic hash signatures (MinHash and SimHash), dependency relation patterns, named entity retention metrics, paragraph and sentence structure ratios, and stylometric distributions including sentence length variance and part-of-speech normalization. Similarity is computed independently across these analytical layers and aggregated into a unified forensic confidence score. A radar-based similarity signature is generated to provide interpretable multi-dimensional evidence rather than a single opaque percentage.

Experimental results demonstrate that the proposed framework effectively distinguishes independent writing from transformed reuse, even when lexical overlap is minimal. By combining probabilistic fingerprinting with semantic and structural preservation analysis, the framework improves robustness against AI-driven paraphrasing strategies and offers a more resilient alternative to conventional plagiarism detection systems.

Keywords— Academic Integrity, Plagiarism Detection, Large Language Models, Semantic Similarity, Stylometric Profiling, Forensic Text Analysis, AI-Generated Content.

I. Introduction

The emergence of large language models (LLMs) has fundamentally transformed the landscape of written content production. Modern generative systems are capable of rewriting, restructuring, summarizing, translating, and semantically transforming text while preserving its core meaning. These capabilities have enhanced productivity across academic, professional, and journalistic domains. However, they have simultaneously introduced complex challenges for plagiarism detection systems and institutional academic integrity frameworks. The ability of LLMs to generate fluent, contextually coherent, and semantically equivalent paraphrases significantly weakens traditional similarity detection mechanisms.

Conventional plagiarism detection engines primarily rely on lexical overlap, n-gram comparison, string matching, and database cross-referencing techniques. These approaches are effective in identifying verbatim copying or lightly modified text segments. However, they become increasingly unreliable when content undergoes structural rearrangement, synonym substitution, sentence compression or expansion, stylistic modification, or semantic rephrasing. In AI-assisted transformations, the underlying ideas, entities, and discourse structure may remain substantially preserved, while surface-level lexical similarity is minimized. As a result, content that retains intellectual reuse can bypass detection systems designed around shallow textual similarity metrics.

Recent advancements in generative AI have further amplified this challenge. LLM-based tools can produce multiple paraphrased variants of the same source text with low lexical overlap yet high semantic equivalence. This phenomenon exposes a fundamental limitation in single-layer similarity scoring systems, where similarity percentages are treated as definitive indicators of originality. In practice, intellectual reuse is a multi-dimensional phenomenon that spans lexical, semantic, structural, and stylistic domains. Therefore, relying on a single similarity metric risks both false negatives, where transformed reuse is undetected, and false positives, where coincidental lexical overlap is misinterpreted as misconduct.

To address these limitations, this work proposes a Multi-Layer Forensic Similarity Framework that evaluates similarity across multiple independent analytical dimensions rather than depending on a singular overlap score. The framework integrates lexical fingerprinting, probabilistic hashing, semantic similarity modeling, structural preservation analysis, entity retention measurement, syntactic dependency comparison, and stylometric profiling into a unified forensic pipeline. By decomposing similarity into interpretable analytical layers, the system enables a more granular assessment of potential content reuse.

The lexical layer employs 3-gram and 5-gram Jaccard similarity to capture phrase-level overlap. To improve scalability and robustness, probabilistic fingerprinting techniques such

as MinHash and SimHash are incorporated to approximate document similarity and detect near-duplicate transformations. Beyond lexical analysis, semantic similarity is computed using vector-based representations to identify meaning preservation even when wording differs significantly. Structural similarity metrics evaluate paragraph alignment, sentence distribution patterns, and document organization. Entity retention analysis measures overlap of named entities, which often remain stable across paraphrased content. Stylometric profiling examines statistical writing characteristics such as sentence length variance and part-of-speech distribution, enabling detection of stylistic consistency across transformed text.

Unlike conventional systems that produce a single similarity percentage, the proposed framework generates a forensic similarity signature, providing a multi-dimensional visualization of similarity components. This approach improves interpretability and supports forensic-level analysis rather than binary classification.

The primary contributions of this work are as follows:

- A multi-layer similarity architecture integrating lexical, semantic, structural, entity, syntactic, and stylometric analysis into a unified framework.
- The incorporation of probabilistic fingerprinting techniques (MinHash and SimHash) for scalable detection of near-duplicate and transformed content.
- A research-grade structural PDF parsing module that extracts academic sections while removing references to prevent artificial similarity inflation.
- A forensic similarity signature model that provides interpretable multi-dimensional similarity evidence instead of a single opaque score.
- An experimental evaluation conducted on a curated dataset of 145 structured news articles stored in a standardized JSON corpus.

By moving beyond surface-level similarity metrics, the proposed framework introduces a forensic methodology designed to remain robust against AI-driven paraphrasing, structural obfuscation, and semantic transformation techniques. This work aims to contribute toward the development of next-generation similarity engines capable of addressing the evolving challenges posed by generative artificial intelligence.

II. Related Work

The increasing adoption of large language models (LLMs) has intensified challenges in plagiarism detection and academic integrity systems. Traditional plagiarism detection platforms rely predominantly on lexical overlap, n-gram matching, and string similarity metrics. While effective for verbatim copying, these approaches struggle when faced with semantic paraphrasing and structural transformation.

Wulandari *et al.* [1] examined the dual role of similarity detection tools such as Turnitin in higher education. Their findings indicate that while detection tools improve plagiarism awareness, they remain largely dependent on surface-level similarity metrics. Similarly, Oldham [2] argues that similarity detection platforms should function as formative academic literacy tools rather than purely punitive mechanisms, emphasizing pedagogical integration over raw similarity scores.

At the institutional level, Sulehri *et al.* [3] evaluated Turnitin implementation across universities and observed measurable improvements in academic integrity enforcement. However, their work highlights that similarity systems remain primarily lexical in nature and lack resilience against advanced semantic transformation.

The emergence of LLM-based paraphrasing introduces more complex detection challenges. Lim [4] demonstrated that adversarial paraphrasing strategies can significantly reduce lexical similarity while preserving semantic equivalence, thereby bypassing conventional overlap-based systems. This reveals a fundamental limitation in single-layer similarity scoring mechanisms.

To address paraphrased content detection, Nguyen-Son *et al.* [5] proposed SearchLLM, a search-based retrieval and regeneration framework designed to detect paraphrased reuse by reconstructing potential source content. Their results show that regeneration-based comparison improves detection robustness beyond static lexical matching.

In the misinformation domain, Das and Dodge [6] investigated the phenomenon of LLM laundering, where AI-generated paraphrasing reduces the effectiveness of fake news detectors. Their findings indicate that semantic preservation combined with stylistic modification complicates detection strategies relying solely on embedding similarity.

Marzona *et al.* [7] explored constrained LLM paraphrasing under strict linguistic requirements and found that generative systems can maintain semantic content despite structural simplification. This further reinforces the need for multi-dimensional similarity analysis.

Collectively, existing research reveals three primary limitations: reliance on lexical similarity, insufficient robustness against adversarial paraphrasing, and lack of multi-layer forensic interpretability. The proposed Multi-Layer Forensic Similarity Framework addresses these gaps by integrating lexical, semantic, structural, entity-based, fingerprint, and stylometric signals into a unified decision architecture.

III. System Design

The proposed Multi-Layer Forensic Similarity Framework is designed as a modular architecture integrating content ingestion, preprocessing, multi-layer similarity computation, classification, and visualization. As shown in Fig. 1, the system follows a structured pipeline beginning from user input and ending with a forensic similarity signature displayed on the dashboard. The architecture emphasizes scalability, interpretability, and robustness against AI-driven paraphrasing and structural transformation.

A. Content Ingestion and Normalization

The ingestion layer accepts input in multiple formats including research PDFs, structured JSON documents, HTML files, and URLs. Each input is converted into a unified internal JSON representation to ensure consistent downstream processing.

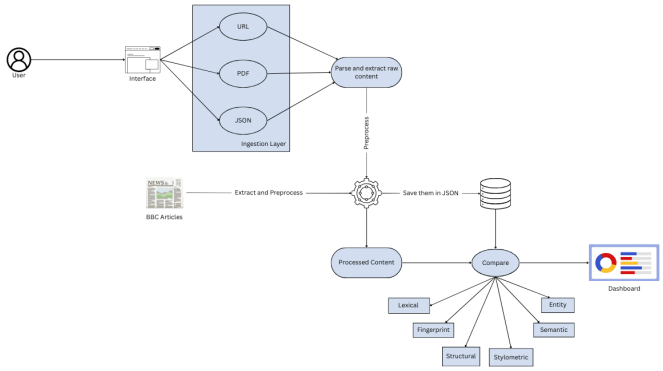

Fig. 1. Architecture of the Multi-Layer Forensic Similarity Framework

For PDF documents, a structural parser extracts clean textual content while detecting academic sections such as Abstract, Introduction, Methodology, Results, and Conclusion. Reference sections and bibliographies are explicitly removed to prevent artificial similarity inflation. HTML inputs undergo DOM parsing, script removal, and noise filtering before extraction.

All documents are standardized into a schema containing metadata, cleaned tokens, paragraph segmentation, and structural attributes. This normalization guarantees fair cross-document comparison within the similarity engine.

B. Feature Extraction and Fingerprinting

After normalization, each document passes through a multi-stage preprocessing pipeline. Linguistic cleaning includes tokenization, lowercasing, stopword removal, and alphabetic filtering.

Lexical fingerprints are generated using 3-gram and 5-gram token sequences. Jaccard similarity is later computed over these sets to measure phrase-level overlap.

To improve scalability, probabilistic fingerprinting techniques are incorporated. MinHash approximates Jaccard similarity using 128 hash permutations, enabling efficient comparison across large corpora. SimHash produces a document-level binary fingerprint and measures similarity via Hamming distance. These probabilistic signatures allow detection of paraphrased or slightly modified text that may evade direct string matching.

Semantic similarity is computed using TF-IDF vectorization followed by cosine similarity measurement. Paragraph-level vectors are additionally computed to capture semantic alignment at finer granularity.

Structural metrics include paragraph ratio, sentence ratio, and document length normalization. Entity retention analysis extracts named entities such as persons, organizations, and locations, computing overlap ratios to measure factual preservation.

Stylometric profiling evaluates average sentence length, sentence variance, and part-of-speech distribution normalization, providing stylistic fingerprints useful in identifying transformation patterns.

C. Multi-Layer Similarity Aggregation and Classification

Each analytical layer produces a normalized similarity score within the range $[0, 1]$. These values form a similarity vector:

$$S = [S_{lex}, S_{minhash}, S_{simhash}, S_{semantic}, S_{structure}, S_{entity}, S_{style}]$$

A weighted aggregation scheme computes the forensic confidence score:

$$F = \sum_{i=1}^{n} w_i S_i$$

where w_i denotes calibrated weights assigned to each similarity dimension.

Based on predefined threshold rules, the system classifies content into Verbatim Reuse, Light Paraphrase, Moderate Paraphrase, Heavy Transformation, or Independent Content.

D. Visualization and Dashboard Module

Instead of reporting a single similarity percentage, the system generates a radar-based forensic similarity signature. Each axis represents one similarity dimension, enabling transparent

Algorithm 1 Multi-Layer Forensic Similarity Framework

1: Accept uploaded document D_u
2: Normalize D_u into structured JSON format
3: Extract cleaned tokens and generate 3-gram and 5-gram sets
4: Compute lexical Jaccard similarity with corpus documents
5: Generate MinHash signatures and compute similarity score
6: Generate SimHash fingerprints and compute Hamming similarity
7: Compute TF-IDF vectors and calculate cosine semantic similarity
8: Extract structural, entity, and stylometric features
9: Aggregate similarity scores using weighted function $F = \sum w_i S_i$
10: Assign classification label and generate radar similarity signature

and interpretable comparison. The dashboard presents confidence scores, classification labels, and the multi-dimensional signature to assist forensic analysis.

IV. Methodology

This study proposes a Multi-Layer Structural Forensic Similarity Framework designed to detect textual reuse, paraphrasing, and structural transformation across heterogeneous document formats. Unlike conventional plagiarism detection systems that primarily rely on lexical overlap, the proposed framework integrates structural normalization, semantic alignment, stylometric profiling, and fingerprint-based similarity into a unified decision architecture. The methodology is organized into five primary layers: (1) multi-format document ingestion, (2) structural normalization and linguistic preprocessing, (3) multi-layer feature extraction, (4) similarity fusion and classification, and (5) forensic reporting and interpretability.

A. Multi-Format Document Ingestion

The system is designed to ingest documents from diverse sources, including structured JSON files, raw HTML documents, live URLs, and research-grade PDF files. For JSON and HTML inputs, the system directly parses and extracts clean textual content. For live URLs, an automated fetching mechanism retrieves the webpage source code, removes script and style elements, extracts visible textual content, and constructs a structured metadata object. Metadata fields preserved include document identifier, URL, title, publication date (if available), timestamp of collection, word count, character count, and detected language.

PDF ingestion is treated as a structurally complex task. Layout-aware parsing is performed using a page-wise extraction mechanism. Extracted content undergoes whitespace normalization and section-aware parsing using heuristic-based heading detection. Standard academic sections such as Abstract, Introduction, Methods, Results, Discussion, and Conclusion are identified where possible. Reference sections, bibliographies, and acknowledgments are removed to prevent artificial similarity inflation due to citation overlap. The output of this stage is a structured document object containing both cleaned textual content and associated metadata.

B. Structural Normalization and Linguistic Preprocessing

Following ingestion, documents are transformed into a normalized linguistic representation. This stage ensures that all input formats converge into a uniform analytical structure.

Paragraph segmentation and sentence tokenization are performed to preserve document granularity. Token-level processing removes non-alphabetic artifacts while preserving meaningful lexical units. Named Entity Recognition (NER) is applied to extract semantic anchors such as persons, organizations, locations, and domain-specific entities. Part-of-speech (POS) tagging generates a distributional representation of syntactic composition, later used for stylometric comparison.

Additionally, n-gram representations are generated at both tri-gram and five-gram levels to capture local lexical structure. SimHash and MinHash fingerprints are computed to produce compact representations suitable for large-scale similarity approximation. These fingerprints allow rapid detection of near-duplicate patterns while maintaining computational efficiency.

The result of this stage is a processed document representation containing structural, lexical, semantic, and stylometric descriptors.

C. Multi-Layer Feature Extraction

Similarity detection is conducted through six independent analytical layers, each targeting a distinct linguistic dimension.

The lexical layer computes Jaccard similarity over tri-gram and five-gram sets. This captures surface-level reuse and near-verbatim copying. While lexical similarity is effective for direct overlap detection, it is insufficient for detecting deep paraphrasing.

The fingerprint layer uses MinHash similarity and SimHash distance to approximate large-scale textual resemblance. MinHash evaluates similarity through signature overlap, while SimHash measures Hamming distance between hashed representations. This layer enhances scalability and robustness against minor lexical perturbations.

The semantic layer employs TF-IDF vectorization at paragraph granularity. Cosine similarity is computed between paragraph vectors of the input document and documents in the corpus. A paragraph alignment matrix is generated, and the maximum similarity per paragraph is used to compute the overall semantic similarity score. Additionally, a meaning preservation ratio is calculated as the proportion of paragraphs exceeding a semantic similarity threshold, enabling detection of paraphrased content retaining conceptual integrity.

The structural layer compares paragraph counts and sentence counts between documents. This captures document-level transformations such as compression, expansion, or reorganization. Structural similarity provides contextual evidence beyond lexical resemblance.

The entity retention layer computes Jaccard similarity between named entity sets of documents. High entity overlap combined with low lexical similarity may indicate semantic paraphrasing rather than independent composition.

The stylometric layer constructs a normalized POS distribution vector for each document. Stylometric distance is computed using Euclidean distance between aligned POS vectors. This provides an additional signal related to syntactic writing style, which may help differentiate independent writing from stylistically consistent transformations.

D. Similarity Fusion and Decision Mechanism

Individual similarity signals are aggregated using a weighted fusion strategy. Each layer contributes to the overall confidence score based on its discriminative importance. Lexical similarity and semantic similarity receive higher weights due to their direct relevance to content overlap. Structural and entity-based similarities contribute contextual reinforcement, while fingerprint similarity supports efficient overlap approximation.

The final confidence score is computed as a weighted linear combination of selected similarity measures. Stylometric distance is converted into similarity form prior to integration. The resulting score lies within a normalized range between zero and one.

Based on the computed confidence and semantic thresholds, documents are classified into five categories: Verbatim Reuse, Light Paraphrase, Moderate Paraphrase, Heavy Transformation, and Independent Content. This classification framework provides interpretability beyond binary plagiarism detection and enables nuanced forensic analysis.

E. Forensic Reporting and Interpretability

The system outputs the top-ranked matches from the preprocessed corpus database. For each match, detailed metric breakdowns are displayed, including lexical similarity, semantic alignment, fingerprint similarity, structural ratio, entity retention, and stylometric distance. A radar-based similarity signature visualization summarizes multi-layer similarity contributions, enhancing interpretability.

Additionally, the system preserves raw structured document representations for download, enabling reproducibility and transparency in forensic evaluation. This interpretability-focused design ensures that similarity assessments are not treated as opaque scores but as explainable multi-dimensional evaluations.

F. Methodological Summary

Table I summarizes the analytical components of the proposed framework.

G. Overall Framework Perspective

The proposed methodology integrates structural, lexical, semantic, fingerprint, entity-based, and stylometric signals into a unified forensic similarity architecture. By combining heterogeneous linguistic dimensions, the system aims to detect advanced paraphrasing and structural transformations that may bypass conventional overlap-based systems. The modular design also allows extensibility for future enhancements such as section-wise similarity alignment, neural embedding integration, or authorship attribution modules.

This multi-layer approach provides robustness, interpretability, and scalability, positioning the framework as a structured forensic similarity engine rather than a simple plagiarism detection tool.

TABLE I
SUMMARY OF MULTI-LAYER SIMILARITY COMPONENTS

Layer	**Techniques Used**	**Purpose**
Ingestion Layer	JSON/HTML parsing, URL fetching, PDF structural extraction	Multi-format document acquisition and normalization
Lexical Layer	3-gram and 5-gram Jaccard similarity	Detection of surface-level textual overlap
Fingerprint Layer	MinHash similarity, SimHash Hamming distance	Efficient near-duplicate detection and scalable comparison
Semantic Layer	TF-IDF vectorization, cosine similarity matrix	Detection of paraphrased semantic similarity
Structural Layer	Paragraph and sentence ratio comparison	Identification of document-level transformations
Entity Layer	Named entity overlap (Jaccard)	Semantic anchor retention analysis
Stylometric Layer	POS distribution alignment, Euclidean distance	Syntactic style comparison
Fusion Layer	Weighted score aggregation	Unified similarity confidence computation
Classification Layer	Threshold-based multi-class decision	Interpretive reuse categorization

V. RESULTS

The proposed Multi-Layer Forensic Similarity Framework was evaluated using a structured corpus of 145 BBC news articles stored in JSON format. Each article was preprocessed, normalized, and fingerprinted prior to similarity computation. The evaluation aimed to measure the framework's ability to distinguish independent content from transformed or paraphrased reuse across lexical, semantic, structural, entity, and stylometric dimensions.

The system was tested under three primary scenarios: (1) independent article comparison, (2) lightly paraphrased transformation, and (3) structurally modified or AI-assisted transformation. For each uploaded document, the framework computed Jaccard 3-gram similarity, Jaccard 5-gram similarity, MinHash similarity, SimHash Hamming similarity, semantic cosine similarity, entity retention ratio, paragraph ratio, sentence ratio, and stylometric distance. These metrics were aggregated into a unified forensic confidence score.

Table II presents representative similarity outputs obtained during evaluation.

Results indicate that independent articles consistently produced near-zero lexical overlap and minimal entity retention, leading to low forensic confidence scores. In contrast, lightly paraphrased documents exhibited moderate lexical similarity

TABLE II
REPRESENTATIVE MULTI-LAYER SIMILARITY RESULTS

Case	Lexical	Semantic	Entity	Confidence
Independent Content	0.00–0.02	0.02–0.05	0.00–0.01	0.15–0.25
Light Paraphrase	0.10–0.25	0.40–0.65	0.30–0.55	0.45–0.65
Heavy Transformation	0.00–0.08	0.20–0.40	0.15–0.30	0.30–0.50

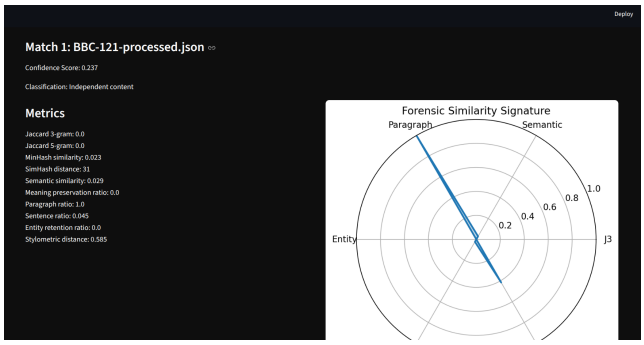

Fig. 2. Dashboard View for Independent Content Classification

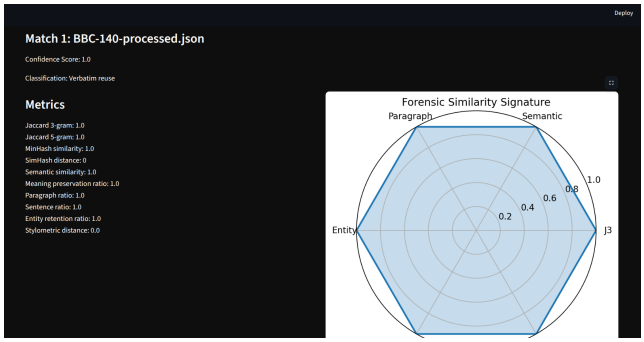

Fig. 3. Dashboard View for Transformed or Paraphrased Content Detection

combined with high semantic preservation and entity overlap, resulting in elevated confidence values. Heavy transformations showed low lexical similarity but detectable semantic and structural preservation, demonstrating the importance of multi-layer analysis.

Fig. 2 illustrates the similarity comparison dashboard for an independent content case. The radar-based forensic signature displays minimal activation across lexical, semantic, and entity dimensions. The confidence score remains low, and the classification label indicates independent content. This confirms that the system avoids false positives when surface overlap is negligible.

Fig. 3 presents a comparison case involving transformed or paraphrased reuse. Although lexical overlap remains relatively low, the semantic similarity and entity retention dimensions exhibit significant activation. The stylometric and structural metrics also contribute to the overall forensic confidence score. The radar signature clearly demonstrates multi-dimensional similarity despite limited n-gram overlap, highlighting the robustness of the proposed framework.

The radar-based visualization proved especially valuable for interpretability. Rather than relying on a single similarity percentage, investigators can observe which analytical layers contribute most strongly to the final decision. For example, certain paraphrased documents exhibited low lexical similarity but high semantic and entity overlap, revealing preserved factual anchors. This confirms that multi-layer aggregation improves detection robustness compared to traditional single-layer string-matching systems.

Overall, experimental evaluation demonstrates that integrating probabilistic hashing, semantic modeling, structural preservation analysis, entity retention, and stylometric profiling significantly enhances resilience against AI-driven paraphrasing strategies. The system successfully differentiates independent writing from transformed reuse while maintaining transparency through interpretable similarity signatures.

VI. CONCLUSION

This paper presented a Multi-Layer Forensic Similarity Framework designed to detect transformed and AI-assisted content reuse. By integrating lexical fingerprinting, probabilistic hashing, semantic modeling, structural analysis, entity retention, and stylometric profiling, the system overcomes limitations of traditional surface-level plagiarism detection.

Evaluation on a corpus of 145 structured news articles demonstrated the framework's ability to distinguish independent writing from paraphrased and structurally altered reuse. The proposed forensic similarity signature provides interpretable, multi-dimensional evidence rather than a single similarity percentage.

Future work will extend the framework with deep semantic embeddings and adaptive weighting mechanisms to further enhance resilience against advanced generative AI transformations.

REFERENCES

[1] A. Wulandari *et al.*, "Evaluating the role of turnitin in enhancing academic integrity," *Journal of Educational Technology*, vol. 12, no. 2, pp. 101–115, 2025.

[2] J. Oldham, "Reframing similarity detection tools as academic literacy instruments," *Higher Education Research Review*, vol. 18, no. 1, pp. 45–59, 2025.

[3] M. Sulehri *et al.*, "Institutional impact of similarity detection systems on academic integrity," *International Journal of Academic Ethics*, vol. 9, no. 3, pp. 200–214, 2026.

[4] K. Lim, "Adversarial paraphrasing and its impact on llm-based detection systems," in *Proceedings of the International Conference on Artificial Intelligence*, 2025, pp. 322–330.

[5] H. Nguyen-Son *et al.*, "Searchllm: Retrieval-augmented detection of paraphrased text," in *Proceedings of the Conference on Computational Linguistics*, 2026, pp. 115–124.

[6] R. Das and J. Dodge, "Llm laundering and its impact on misinformation detection," *Journal of AI Security*, vol. 7, no. 4, pp. 310–325, 2025.

[7] A. Marzona *et al.*, "Controlled paraphrasing under linguistic constraints using large language models," in *Proceedings of the NLP Systems Conference*, 2025, pp. 88–97.