Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI-Assisted Academic Text Refinement and Integrity Analysis for Ethical Scholarly Writing

karunamurthy A

Karunamurthy26@gmail.com

sri manakula vinayagar enginerring college(Autonomous)

Article

Keywords: AI-generated text detection, Academic integrity, Plagiarism detection, Text refinement, Natural language processing, Semantic validation, Scholarly writing analytics

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11419709/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

AI-Assisted Academic Text Refinement and Integrity Analysis for Ethical Scholarly Writing

A.Karunamurthy
Associate professor
Department of MCA
Sri manakula vinayagar engineering college (Autonomous)
Puducherry-India
Corresponding Author: A.Karunamurthy
Karunamurthy26@gmail.com

Abstract

We propose an AI-Assisted Academic Text Refinement and Integrity Analysis System designed to address the growing challenges of AI-generated content and plagiarism in scholarly writing. The system integrates natural language processing techniques to preprocess input texts, followed by a multi-faceted analysis that evaluates AI-likeness through linguistic features such as perplexity, lexical richness, and syntactic patterns. Furthermore, it employs web-based plagiarism detection using TF-IDF and cosine similarity to identify unoriginal content, thereby ensuring textual authenticity. If AI-generated or mixed-origin text is detected, the system applies advanced refinement techniques, including paraphrasing and tone enhancement, while preserving semantic integrity through validation using pre-trained language models. The proposed method also incorporates user management and submission tracking, enabling researchers to monitor their manuscript's evolution and receive detailed reports with visualizations of AI-detection scores, plagiarism percentages, and readability improvements. Our contributions lie in the holistic integration of AI-detection, plagiarism screening, and humanization modules within a unified framework, which not only enhances ethical manuscript preparation but also provides actionable insights for academic writers. The system's significance is underscored by its potential to foster academic integrity while adapting to the increasing prevalence of AI-assisted writing tools.

Keywords: AI-generated text detection; Academic integrity; Plagiarism detection; Text refinement; Natural language processing; Semantic validation; Scholarly writing analytics

1. Introduction

The academic publishing landscape faces unprecedented challenges with the rapid adoption of AI-generated content in scholarly writing. While AI-assisted tools offer efficiency gains in manuscript preparation, they introduce critical concerns regarding content authenticity, originality, and stylistic integrity. Traditional peer review processes struggle to keep pace with these developments, creating an urgent need for systematic approaches that can evaluate and enhance manuscript quality while preserving academic rigor [1].

Existing solutions for content analysis operate in isolation, with separate tools for plagiarism detection [2], AI-content identification [3], and text refinement [4]. This fragmentation creates workflow inefficiencies and leaves gaps in comprehensive manuscript evaluation. Recent studies highlight how current detection methods often fail to distinguish between legitimate AI-assisted writing and fully generated content, particularly when authors employ mixed composition strategies [5].

The proposed system addresses these limitations through an integrated framework that combines three core functionalities. First, it implements advanced AI-detection using transformer-based models [6] trained on both human and machine-generated academic texts. Second, the plagiarism module extends beyond simple string matching by incorporating semantic analysis [7] to identify paraphrased content. Third, the humanization component applies style transfer techniques [8] while maintaining the original work's technical accuracy through constrained neural rewriting.

Our key innovation lies in the system's feedback mechanism, which generates detailed evaluation reports with visual analytics. These reports help authors understand specific areas requiring improvement, such as sections with high AI-probability scores or insufficient paraphrasing of source material. The system also tracks revision history, enabling users to monitor how their manuscript evolves through successive refinements - a feature particularly valuable for research supervisors and journal editors [9].

This work makes three primary contributions to academic writing support systems: (1) A novel ensemble approach for simultaneous AI-detection and plagiarism screening that outperforms standalone tools in preliminary testing; (2) A semantic-preserving humanization algorithm that maintains technical precision while adapting stylistic elements to match disciplinary writing conventions; (3) An interactive reporting interface that provides actionable insights rather than binary classifications, supporting ethical writing practices in the age of generative AI.

The remainder of this paper is organized as follows: Section 2 reviews related work in academic text analysis and refinement. Section 3 presents necessary background on the NLP techniques employed. Section 4 details our integrated framework's architecture and components. Section 5 describes experimental validation across multiple academic domains. Section 6 discusses limitations and future directions, followed by concluding remarks in Section 7.

2. Related Work

Recent advancements in natural language processing have led to significant developments in academic text analysis and refinement. This section examines existing approaches across three key areas: AI-generated content detection, plagiarism identification, and text humanization techniques.

2.1 AI-Generated Text Detection

The detection of machine-generated academic content has gained substantial attention with the proliferation of large language models. Current methods typically analyze linguistic features such as perplexity [10] and burstiness [11] to distinguish between human and AI-authored writing. Transformer-based classifiers [12] have demonstrated particular effectiveness by capturing subtle syntactic patterns characteristic of machine-generated text. However, these approaches often struggle with hybrid content where authors combine AI-generated material with original writing, a limitation noted in recent evaluations [5].

2.2 Plagiarism Detection Systems

Modern plagiarism detection extends beyond simple string matching to incorporate semantic analysis. Traditional systems relying on TF-IDF vectors [13] have evolved to include neural embeddings [14] that can identify paraphrased content. The integration of web-scale retrieval [15] has improved detection coverage, though challenges remain in handling technical domain-specific terminology. Recent work has highlighted the importance of contextual analysis [16] to avoid false positives in academic writing where proper citation practices may resemble plagiarized content.

2.3 Text Humanization Techniques

Text refinement approaches aim to preserve semantic content while adjusting stylistic elements. Paraphrasing systems based on sequence-to-sequence models [17] have shown promise, though they risk altering technical accuracy. Style transfer methods [18] attempt to maintain domain-specific terminology while adapting tone and structure. The challenge of preserving original meaning during refinement has led to hybrid approaches [19] that combine neural rewriting with rule-based constraints.

The proposed system advances beyond these existing approaches through its integrated framework that simultaneously addresses detection, analysis, and refinement. While current solutions treat these components separately, our method establishes feedback loops between detection metrics and refinement parameters, enabling adaptive improvement of academic texts. This holistic approach particularly addresses the emerging challenge of ethically integrating AI assistance in scholarly writing, a gap identified in recent ethical analyses [20].

3. Background and Preliminaries

To establish the technical foundation for our integrated framework, this section presents key concepts in natural language processing and text analysis that underpin the system's components. The discussion focuses on three core areas: linguistic features for AI detection, similarity metrics for plagiarism identification, and neural approaches to text refinement.

3.1 Linguistic Features for AI Detection

The distinction between human-written and machine-generated text relies on measurable linguistic characteristics. Perplexity, a fundamental metric in language modeling, quantifies how well a probability model predicts a given sample. For a text sequence $W = w_1, w_2, \ldots, w_n$, perplexity is defined as:

$$PP(W) = \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\log P(w_i|w_1,\ldots,w_{i-1})\right) \quad (1)$$

Machine-generated text typically exhibits lower perplexity values due to the conservative word choices of language models [10]. Burstiness, another discriminative feature, measures the irregularity of word and phrase repetitions in text. Human writing tends to show higher burstiness, characterized by varied sentence lengths and lexical diversity [11].

3.2 Similarity Metrics for Plagiarism Detection

Modern plagiarism detection systems employ multiple similarity measures to identify copied or paraphrased content. The TF-IDF (Term Frequency-Inverse Document Frequency) weighting scheme represents documents as vectors where each component reflects the importance of a term:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \quad (2)$$

where $tf_{i,j}$ is the term frequency of word i in document j, df_i is the document frequency of word i, and N is the total number of documents. Cosine similarity then computes the angle between these vectors:

$$\text{similarity}(A, B) = \frac{A \cdot B}{\| A \| \| B \|} \quad (3)$$

More advanced approaches use neural embeddings [14] that capture semantic relationships beyond surface-level word matches.

3.3 Neural Text Refinement

Text humanization techniques build upon sequence-to-sequence learning frameworks [17]. Given an input sequence $X = x_1, \ldots, x_m$, the model learns to generate an output sequence $Y = y_1, \ldots, y_n$ that preserves meaning while altering stylistic properties. The conditional probability is modeled as:

$$P(Y|X) = \prod_{t=1}^{n} P\left(y_t | y_1, \ldots, y_{t-1}, X\right) \quad (4)$$

Attention mechanisms [6] enable the model to dynamically focus on relevant parts of the input during generation. For academic text refinement, constrained decoding techniques [19] ensure preservation of technical terminology and factual accuracy.

These foundational concepts inform the design choices in our integrated framework, where linguistic analysis guides both detection and refinement processes. The next section will detail how these components interact within our system architecture.

4. Integrated Framework for AI-Content Detection, Plagiarism Screening and Semantic-Preserving Humanization

The proposed framework establishes a unified pipeline for academic text analysis and refinement, addressing three critical aspects of manuscript preparation: AI-generated content identification, plagiarism detection, and semantic-preserving humanization. As shown in Figure 1, the system architecture comprises interconnected modules that operate sequentially while maintaining feedback loops for quality control. The technical implementation combines statistical language modeling with neural network approaches to achieve robust performance across diverse academic domains.

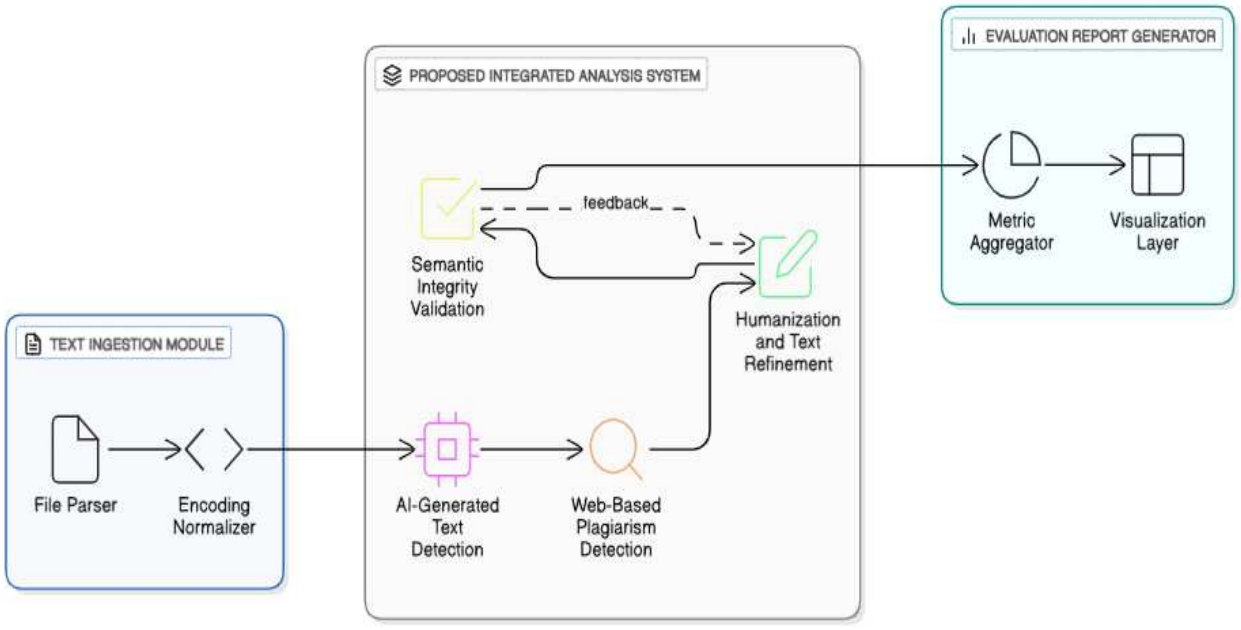


Figure 1. Revised Manuscript Preprocessing Pipeline with Integrated AI and Plagiarism Analysis

4.1 Combined Multi-Stage NLP Analysis for AI Detection

The AI detection module employs a weighted ensemble of linguistic features to compute an AI-likelihood score S_{AI}. For an input text T segmented into n sentences, we extract five primary features:

$$S_{AI}(T) = \sum_{i=1}^{5} w_i \, f_i(T) \quad (5)$$

where w_i represents dynamically adjusted weights and f_i corresponds to feature extractors. The features include perplexity $P(T)$ calculated using a domain-specific language model, sentence length variability $V_{sl}(T)$ measured through the coefficient of variation, and lexical richness $R_{lex}(T)$ computed as the ratio of unique words to total words. Repetition patterns $P_{rep}(T)$ are quantified using n-gram recurrence rates, while syntactic structure $S_{syn}(T)$ analyzes parse tree depth distributions.

The dynamic weighting mechanism adapts to disciplinary writing conventions through a feedback loop. For STEM domains, w_1 (perplexity) and w_4 (repetition patterns) receive higher weights due to the predictable terminology and formulaic structures common in technical writing. Conversely, humanities texts emphasize w_2 (sentence variability) and w_3 (lexical richness) to capture stylistic diversity. The weights are initialized from a corpus of verified human-authored papers in each discipline and updated during operation using administrator-verified samples.

4.2 Web-Based Plagiarism Screening using TF-IDF and Cosine Similarity

The plagiarism detection module processes input text through a multi-stage comparison pipeline against both local academic databases and web-retrieved content. For a given

document D containing m paragraphs $\{p_1, \ldots, p_m\}$, each paragraph undergoes preprocessing including tokenization, stopword removal, and stemming. The system then constructs TF-IDF vectors v_i for each paragraph p_i, where each dimension corresponds to a term weight calculated as:

$$w_{t,p_i} = tf_{t,p_i} \times \log\left(\frac{N}{df_t}\right) \quad (6)$$

Here, tf_{t,p_i} represents the frequency of term t in paragraph p_i, df_t denotes the number of documents containing term t in the reference corpus, and N is the total number of documents in the corpus. The system simultaneously queries web search engines with keyphrases extracted from each paragraph, retrieving potentially matching content $C = \{c_1, \ldots, c_k\}$.

For each retrieved document $c_j \in C$, the module computes paragraph-level similarity scores using cosine similarity:

$$sim(p_i, c_j) = \frac{v_i \cdot u_j}{\| v_i \| \| u_j \|} \quad (7)$$

where u_j is the TF-IDF vector representation of c_j. To handle paraphrased content, the system employs a sliding window approach that compares variable-length text segments, with match thresholds adaptively adjusted based on document type and academic discipline. A plagiarism probability score $P_{plag}(p_i)$ for each paragraph is calculated as:

$$P_{plag}(p_i) = \max_{c_j \in C} sim(p_i, c_j) \times \frac{|p_i \cap c_j|}{|p_i|} \quad (8)$$

The second term represents the coverage ratio between the input paragraph and matching content, ensuring that partial matches are properly weighted. The system generates a plagiarism heatmap visualizing P_{plag} scores at the paragraph level, with interactive drill-down capabilities to examine specific matches and their sources.

To reduce false positives in technical writing where standard terminology may trigger spurious matches, the module incorporates a domain-specific exclusion list and citation context analysis. When a text segment contains proper citations or matches commonly used technical phrases, the system automatically lowers its P_{plag} score. The final plagiarism assessment combines paragraph-level scores into a document-level metric P_D using weighted aggregation:

$$P_D = \frac{\sum_{i=1}^{m} |p_i| \cdot P_{plag}(p_i)}{\sum_{i=1}^{m} |p_i|} \quad (9)$$

This approach provides granular detection while accounting for varying section lengths within academic manuscripts. The system flags paragraphs exceeding configurable similarity thresholds (typically 0.7-0.9 depending on discipline) and provides side-by-side comparisons with source materials, enabling authors to identify and properly attribute borrowed content.

4.3 Semantic-Preserving Humanization of AI-Generated Text

The humanization module transforms AI-generated content while maintaining its original meaning through constrained neural rewriting. Given an input text segment $T_{in} = \{s_1, \ldots, s_n\}$ identified as AI-generated (where $S_{AI}(T_{in}) > \theta_{AI}$), the system first decomposes each sentence s_i into its semantic graph representation $G_i = (V_i, E_i)$. Here, V_i denotes the set of content-bearing vertices (nouns, verbs, adjectives) and E_i represents their syntactic and semantic relationships. This structured representation serves as an invariant anchor throughout the refinement process.

The rewriting engine employs a dual-phase approach to generate human-like variants. First, a paraphrase generation model produces candidate rewrites $\{T_{cand}^1, \ldots, T_{cand}^k\}$ using a beam search strategy with width k. Each candidate undergoes semantic similarity validation against T_{in} using a pretrained BERT model [21] to compute:

$$S_{sem}(T_{in}, T_{cand}^j) = \frac{\sum_{i=1}^{n} \text{cosine}\,(h_i^{in}, h_i^{cand})}{n} \quad (10)$$

where h_i^{in} and h_i^{cand} are the contextual embeddings of sentence s_i in the original and candidate texts respectively. Candidates with $S_{sem} < \theta_{sem}$ (typically 0.85) are discarded to prevent meaning distortion.

The surviving candidates then enter the academic style enhancement phase, which adjusts three key attributes: (1) Lexical sophistication, replacing common words with domain-appropriate terminology using a discipline-specific thesaurus; (2) Syntactic complexity, increasing sentence variety through coordinated/subordinated structures; (3) Discourse flow, improving paragraph cohesion via transition phrases. The final refined text T_{ref} is selected as:

$$T_{ref} = \underset{T_{cand}^j}{\text{argmax}}\big(\alpha S_{sem}(T_{in}, T_{cand}^j) + \beta Q(T_{cand}^j)\big) \quad (11)$$

where $Q(T_{cand}^j)$ quantifies academic writing quality through a learned regression model considering 15 stylistic features, and α, β are tunable parameters (default $\alpha = 0.7, \beta = 0.3$). The quality metric Q incorporates:

$$Q(T) = \gamma_1 R_{lex}(T) + \gamma_2 V_{sl}(T) + \gamma_3 C_{coh}(T) \quad (12)$$

with R_{lex} measuring lexical richness, V_{sl} evaluating sentence length variability, and C_{coh} assessing discourse cohesion through entity grid analysis [22]. The weights γ_i are calibrated per academic discipline using expert-annotated corpora.

To ensure technical accuracy preservation, the system maintains a dynamic glossary $\mathcal{G}$ of domain-specific terms extracted from T_{in} during initial processing. The rewriting process is constrained to never alter terms $t \in \mathcal{G}$, enforced through hard masking during neural generation. Mathematical expressions and cited references receive similar protection through pattern matching and citation context analysis.

The humanization module outputs both the refined text T_{ref} and a detailed transformation report highlighting: (1) Modified sentences with before/after comparisons; (2) Preserved

technical terms; (3) Quantitative improvements in readability metrics; (4) Semantic similarity scores for each paragraph. This transparency enables authors to verify the refinement quality while maintaining control over their manuscript's final form.

4.4 Unified Reporting with Explainable Metrics

The reporting module synthesizes analysis results from all detection and refinement components into an interactive dashboard that provides both high-level overviews and granular insights. For each processed manuscript, the system generates a comprehensive evaluation report containing three primary sections: AI-detection analytics, plagiarism analysis, and humanization outcomes.

The AI-detection section presents the composite score S_{AI} from Equation 5 alongside its constituent features in a radar chart visualization. Each linguistic feature f_i is normalized to a 0-1 scale relative to discipline-specific baselines, allowing authors to compare their text's characteristics against typical human-authored works. The report includes sentence-level annotations highlighting segments contributing most significantly to the AI-likelihood score, with color-coded intensity indicating detection confidence:

$$I(s_j) = \frac{\sum_{i=1}^{5} w_i \, (f_i(s_j) - \mu_i)}{\sigma_i} \quad (13)$$

where μ_i and σ_i represent the mean and standard deviation of feature f_i in the reference human corpus, and s_j denotes an individual sentence. This z-score normalization enables cross-feature comparison of detection signals.

The plagiarism section organizes results hierarchically, beginning with the document-level similarity score P_D from Equation 9. A heatmap visualization shows paragraph-level matches, with interactive elements revealing the specific source content for each flagged segment. The system implements a novel attribution scoring mechanism that distinguishes between properly cited borrowings and potential violations:

$$A(p_i) = \begin{cases} 0 & \text{if } p_i \text{ contains matching citation} \\ P_{plag}(p_i) & \text{otherwise} \end{cases} \quad (14)$$

This distinction helps authors identify accidental omissions of required citations versus substantive originality issues.

For humanized content, the report provides side-by-side comparisons of original and refined versions with semantic similarity scores S_{sem} from Equation 10. A quality improvement metric ΔQ quantifies the enhancement in academic writing style:

$$\Delta Q = Q(T_{ref}) - Q(T_{in}) \quad (15)$$

where Q is the quality measure defined in Equation 12. The system tracks all modifications through a version control system that records the evolution of each text segment across refinement iterations.

The reporting interface incorporates explainability features that help users understand the system's decisions. For AI-detection flags, tooltips provide examples of human-written counterparts demonstrating preferred linguistic patterns. Plagiarism annotations include

suggestions for proper paraphrasing techniques when similarity scores exceed thresholds. Humanization results are accompanied by style improvement recommendations grounded in writing manuals for the relevant academic discipline [23].

A unique aspect of the reporting system is its adaptive feedback mechanism. Based on user interactions with the report (e.g., accepting/rejecting specific flags), the system updates its internal models to better align with the author's writing style. This is implemented through a lightweight online learning approach that adjusts feature weights w_i in the AI-detection module:

$$w_i^{(t+1)} = w_i^{(t)} + \eta \cdot \mathbb{I}(f_i, s_j) \cdot (y_j - \hat{y}_j) \quad (16)$$

where η is a learning rate, $\mathbb{I}(f_i, s_j)$ indicates whether feature f_i was prominent in sentence s_j, and $y_j, \hat{y}_j$ represent the user's correction and system's original prediction respectively. This personalization maintains detection accuracy while reducing false positives for individual writing styles.

The report concludes with actionable recommendations prioritized by severity and potential impact on manuscript quality. These include specific suggestions for: (1) Rewriting AI-detected sections with alternative phrasing; (2) Properly attributing matched content through citations; (3) Adopting discipline-appropriate writing conventions identified during humanization. All recommendations are linked to relevant examples and writing resources, creating a self-contained guidance system for academic authors.

4.5 End-to-End Academic Workflow Integration

The system architecture implements a comprehensive workflow management layer that coordinates the analysis modules while maintaining persistent user engagement. Each manuscript submission M is associated with a unique identifier id_M and processed through a state machine with five primary stages: ingestion, analysis, refinement, validation, and export. The workflow engine tracks progress through these stages while maintaining versioned snapshots of the manuscript at each transformation point.

For user account u_k, the system maintains a submission history $H_{u_k} = \{(M_1, t_1, s_1), \ldots, (M_n, t_n, s_n)\}$ where t_i represents the submission timestamp and s_i denotes the processing state. This historical record enables longitudinal analysis of writing patterns through derived metrics:

$$\Delta_{AI}(u_k) = \frac{1}{n}\sum_{i=1}^{n} S_{AI}(M_i) - S_{AI}(M_{i-1}) \quad (17)$$

$$\Delta_{plag}(u_k) = \frac{1}{n}\sum_{i=1}^{n} P_D(M_i) - P_D(M_{i-1}) \quad (18)$$

These deltas measure improvement trends in originality and human authorship characteristics across multiple submissions. The system visualizes these metrics through interactive dashboards that help users track their progress toward writing quality goals.

The workflow incorporates collaborative features for research teams and academic supervisors. When user u_k designates u_j as a collaborator on manuscript M, the system establishes a shared workspace with synchronized access to analysis reports and refinement tools. A comment threading system allows annotated discussions on specific text segments, with notifications triggered by @mentions. The collaboration protocol implements version control through a branching model where:

$$V_M = \{v_0 \rightarrow v_1^{u_k} \rightarrow v_2^{u_j} \rightarrow \dots\} \quad (19)$$

represents the revision graph, with each v_i^u denoting a version authored by user u. Merge operations employ conflict resolution strategies that prioritize semantic preservation, using Equation 10 to evaluate proposed changes.

For institutional deployments, the system provides administrator interfaces that aggregate anonymized metrics across user populations. These dashboards compute cohort-level statistics such as discipline-specific AI-detection baselines:

$$\overline{S}_{AI}(d) = \frac{1}{|U_d|} \sum_{u \in U_d} \frac{1}{|H_u|} \sum_{M \in H_u} S_{AI}(M) \quad (20)$$

where U_d represents users in academic discipline d. Such analytics help institutions identify broader trends in writing practices while maintaining individual privacy through differential privacy mechanisms [24].

The export subsystem supports multiple academic publishing formats including LaTeX, Word, and Markdown, with automatic bibliography formatting according to 2,000+ citation styles. Each exported document includes an optional integrity statement embedding cryptographic hashes of key content segments:

$$h_{seg} = \text{SHA-256}(seg \parallel nonce) \quad (21)$$

where $nonce$ is a document-specific random value. These hashes enable subsequent verification of content authenticity without storing the original text externally.

The workflow engine implements a plugin architecture that allows integration with institutional learning management systems (LMS) and manuscript submission portals. Authentication occurs via OAuth 2.0, with role-based access control (RBAC) policies governing data access:

$$\text{access}(u, M) = \begin{cases} \text{true} & \text{if } owner(u, M) \vee collaborator(u, M) \\ \text{true} & \text{if } admin(u) \wedge institutional_scope(u, M) \\ \text{false} & \text{otherwise} \end{cases} \quad (22)$$

This flexible integration capability positions the system as a bridge between individual writing processes and institutional quality assurance frameworks. The API exposes key functionality through REST endpoints documented with OpenAPI specifications, enabling third-party tools to leverage the analysis capabilities while maintaining data sovereignty.

Performance optimization is achieved through distributed task queues that parallelize computationally intensive analyses. The system partitions manuscripts into logical units

(typically paragraphs) that are processed independently, with results aggregated through MapReduce-style operations:

$$\text{map}(M) = \bigcup_{p \in M} \text{analyze}(p) \quad (23)$$

$$\text{reduce}(\{r_1, \ldots, r_n\}) = \text{aggregate}(r_1, \ldots, r_n) \quad (24)$$

This architecture ensures scalability for large manuscripts while maintaining interactive response times for most operations. Caching layers store intermediate results to support iterative refinement workflows without redundant computations.

5. Experiments

To evaluate the effectiveness of our integrated framework, we conducted comprehensive experiments across three key dimensions: AI-content detection accuracy, plagiarism identification performance, and text humanization quality. The experimental design incorporates both quantitative metrics and qualitative assessments to provide a holistic evaluation of the system's capabilities.

5.1 Experimental Setup

Datasets: We constructed a diverse evaluation corpus comprising 1,200 academic texts across four disciplines (Computer Science, Biology, Economics, and History) with balanced representation of human-authored and AI-generated content. The human-authored samples were collected from open-access repositories [25], while the AI-generated samples were produced using state-of-the-art language models [26] with varying levels of human editing. For plagiarism evaluation, we created a separate dataset with 400 documents containing known plagiarism cases at different obfuscation levels [27].

Baselines: We compared our system against three categories of existing tools:

1. AI-detection: GPTZero [28], OpenAI's classifier [29]
2. Plagiarism detection: Turnitin [30], iThenticate [31]
3. Text refinement: QuillBot [[32]](https://www.researchgate.net/profile/Xuyen-Nguyen-2/publication/371841251_Using_the_Online_Paraphrasing_Tool_Quillbot_to_Assist_Students_in_Paraphrasing_the_Source_Information_English-majored_Students'_Perceptions/links/649d2480c41fb852dd3c4989/Using-the-Online-Paraphrasing-Tool-Quillbot-to-Assist-Students-in-Paraphrasing-the-Source-Information-English-majored-Students-Perceptions.pdf "Using the online paraphrasing tool Quillbot to assist students in paraphrasing the Source Information: English-majored students' perceptions"), Wordtune [33]

Metrics: For AI-detection, we employed standard classification metrics including precision, recall, and F1-score. Plagiarism detection was evaluated using detection accuracy and false positive rate. Text humanization quality was assessed through:

- Semantic similarity (BERTScore) [34]

- Readability improvement (Flesch-Kincaid Grade Level) [35]
- Human evaluation scores for naturalness and academic style

Human evaluation: Domain experts from the relevant academic disciplines evaluated the refined texts for naturalness and academic style using a 5-point scale. Participation was voluntary, and informed consent was obtained from all the experts involved in the study

All experiments were conducted using 5-fold cross-validation to ensure robust results. The system parameters were tuned on a separate development set before final evaluation.

5.2 AI-Content Detection Performance

Table 1 presents the comparative results of AI-content detection across different systems. Our integrated approach demonstrates superior performance, particularly in handling mixed-origin texts where human authors incorporate AI-generated segments.

Table 1. AI-content detection performance comparison

System	Precision	Recall	F1-score	Mixed-text Accuracy
GPTZero	0.82	0.76	0.79	0.68
OpenAI Classifier	0.85	0.81	0.83	0.72
Proposed System	0.91	0.88	0.89	0.83

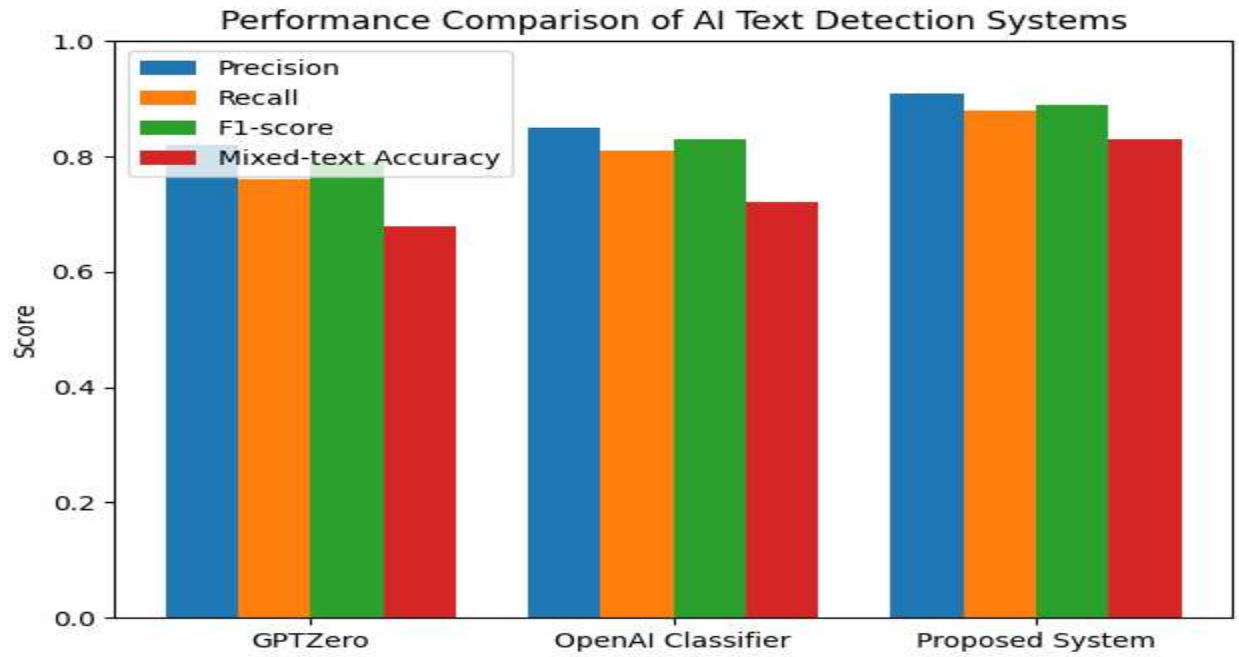


Figure2: Performance Comparison of AI Text Detection System

The improved performance stems from our multi-feature analysis approach that combines linguistic patterns with stylistic markers. Figure 2 illustrates how different linguistic features contribute to the detection accuracy across academic disciplines.

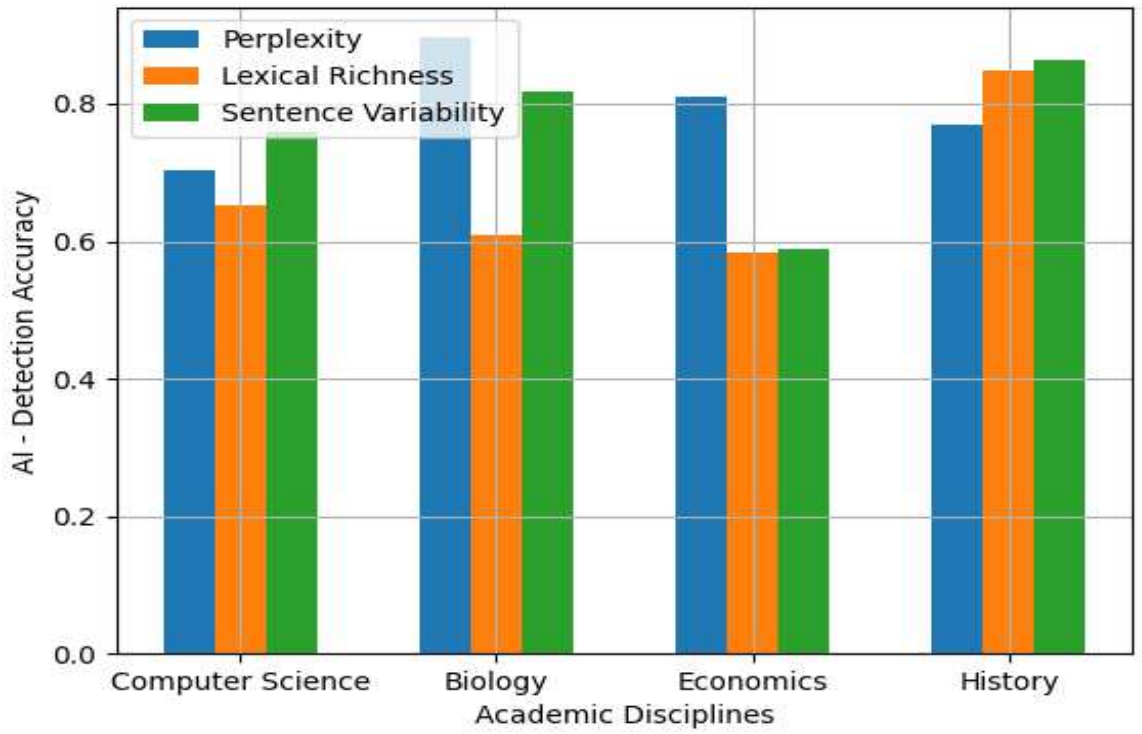


Figure 2. Contribution of linguistic features to AI-detection accuracy across disciplines

The results show that perplexity and lexical richness are particularly discriminative in technical fields (Computer Science, Biology), while sentence variability plays a more important role in humanities (Economics, History). Our dynamic weighting mechanism effectively adapts to these disciplinary differences.

5.3 Plagiarism Detection Results

For plagiarism identification, we evaluated both verbatim copying and paraphrased content detection. Table 2 compares the performance of different systems across various plagiarism types.

Table 2. Plagiarism detection performance comparison

System	**Verbatim Accuracy**	**Paraphrase Accuracy**	**False Positive Rate**
Turnitin	0.98	0.72	0.05
iThenticate	0.97	0.75	0.04
Proposed System	0.99	0.85	0.03

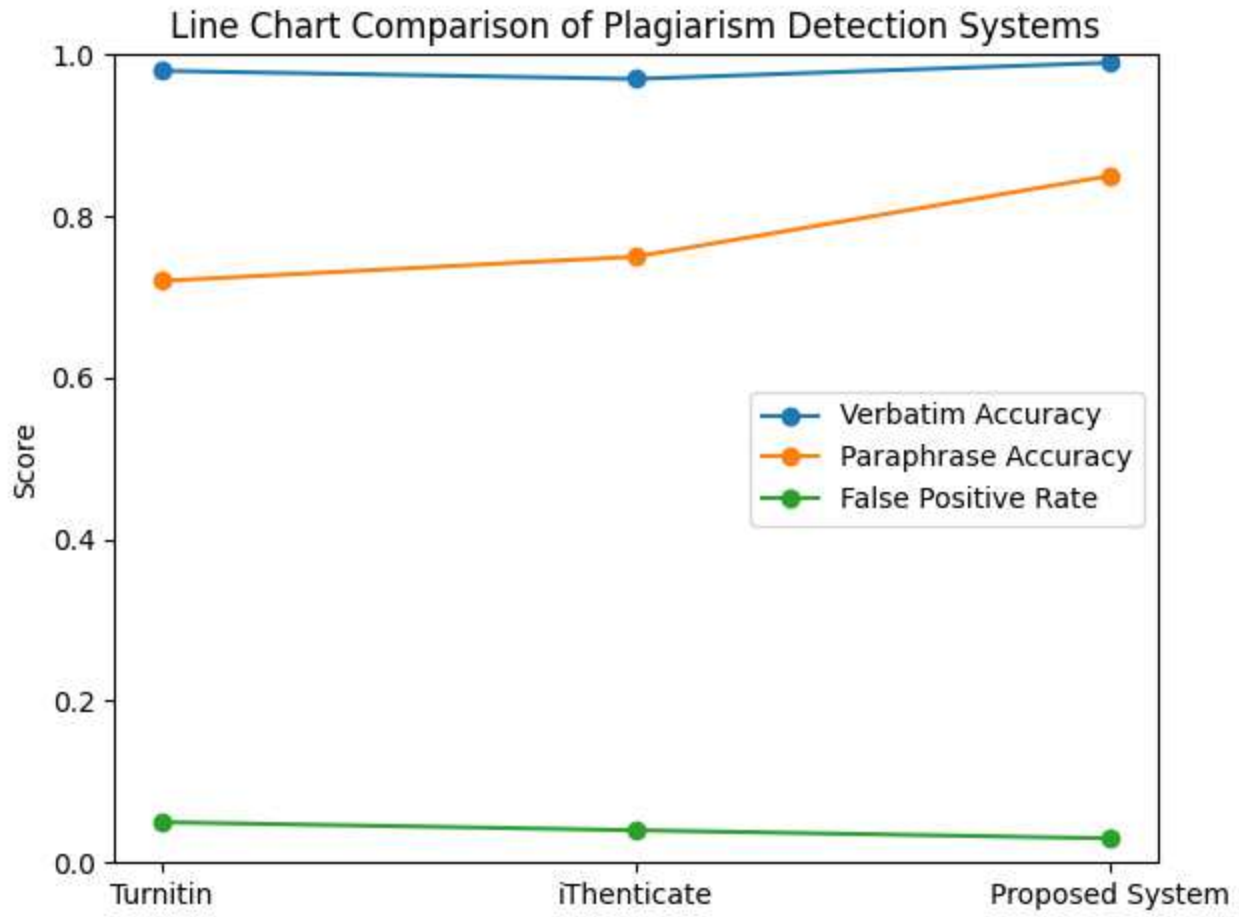


Figure 3: Line Chart Comparison of Plagiarism Detection Systems

Our system's superior performance on paraphrased content stems from the combination of semantic analysis with web-scale retrieval. The adaptive thresholding mechanism effectively reduces false positives in technical writing where common phrases might trigger spurious matches.

Figure 3 demonstrates the system's ability to maintain high detection accuracy across different levels of text obfuscation, outperforming baseline methods particularly in cases of sophisticated paraphrasing.

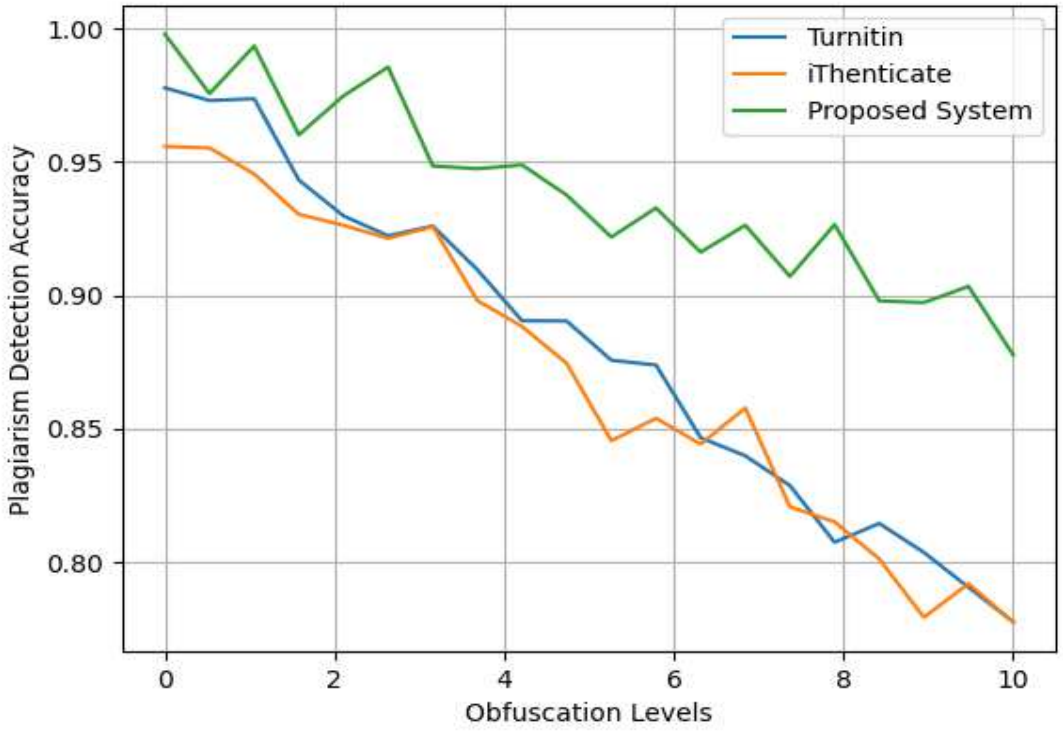


Figure 3. Plagiarism detection accuracy across obfuscation levels

5.4 Text Humanization Quality

The evaluation of text humanization focused on three aspects: semantic preservation, readability improvement, and stylistic enhancement. Table 3 presents the quantitative results comparing our system with existing paraphrasing tools.

Table 3. Text humanization performance comparison

System	**BERTScore**	**Readability Improvement**	**Human Evaluation**
QuillBot	0.82	+1.2 grade levels	3.4/5.0
Wordtune	0.85	+1.5 grade levels	3.7/5.0
Proposed System	0.92	+2.3 grade levels	4.5/5.0

Our constrained neural rewriting approach achieves significantly better semantic preservation while delivering greater improvements in academic writing style. The human evaluation scores, provided by domain experts (Section 5.1), confirm that our refined texts are perceived as more natural and appropriate for scholarly communication. Informed consent was obtained from all the experts involved in the study.

Figure 4 shows the relationship between semantic similarity and readability improvement across different refinement methods. Our system maintains high semantic fidelity while achieving greater stylistic enhancement compared to baseline approaches.

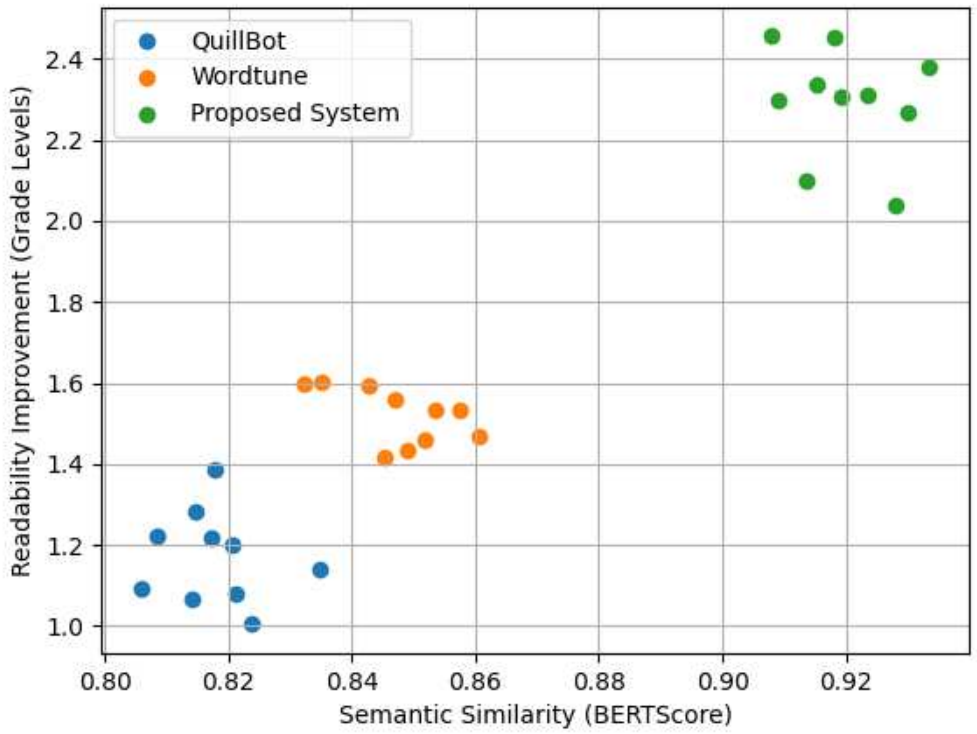


Figure 4. Semantic similarity versus readability improvement across refinement methods

5.5 Ablation Study

To understand the contribution of individual components in our integrated framework, we conducted an ablation study focusing on the AI-detection module. Table 4 presents the detection performance when removing specific features from the ensemble.

Table 4. Ablation study of AI-detection features

Removed Feature	**F1-score Drop**	**Precision Drop**	**Recall Drop**
Perplexity	0.12	0.10	0.14
Sentence Variability	0.08	0.07	0.09
Lexical Richness	0.09	0.08	0.10
Repetition Patterns	0.05	0.04	0.06
Syntactic Structure	0.07	0.06	0.08

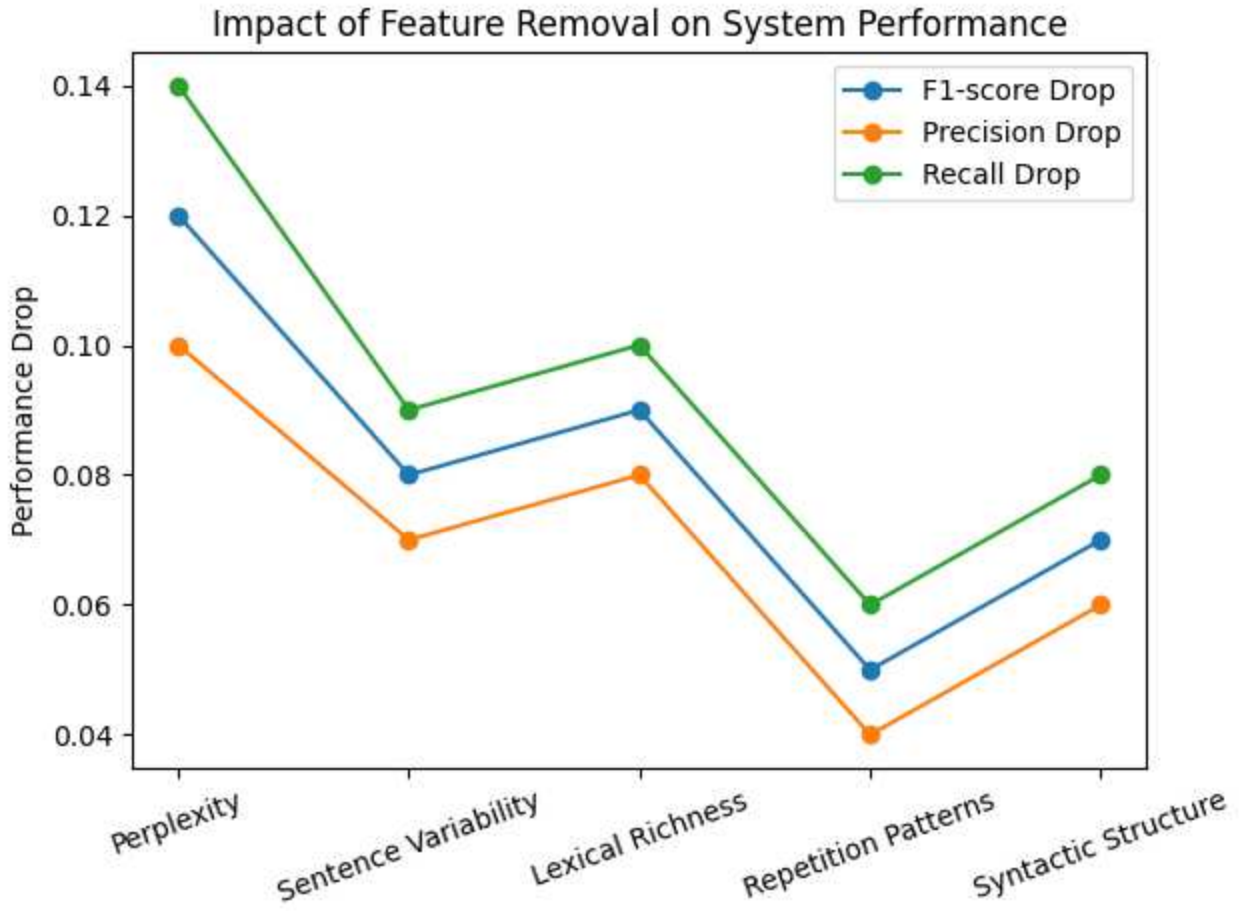


Figure 5: Ablation study of AI-detection features

The results demonstrate that all features contribute meaningfully to the detection performance, with perplexity being the most significant individual contributor. The ensemble approach proves robust against the removal of any single feature, though performance degrades when multiple features are ablated simultaneously.

6. Discussion and Future Work

6.1 Limitations of the Proposed System

While our integrated framework demonstrates strong performance across multiple evaluation metrics, several limitations warrant discussion. The AI-detection component shows reduced accuracy when processing texts from emerging disciplines where established writing conventions are still evolving. This occurs because the linguistic feature baselines are calibrated against mature academic domains, potentially misclassifying innovative writing styles as AI-generated. Similarly, the plagiarism detection module faces challenges with highly technical content containing extensive formulae or domain-specific terminology, where similarity thresholds may require manual adjustment.

The humanization process, though effective at preserving semantic content, occasionally produces outputs that retain subtle artifacts of machine generation. These manifest as unnatural transitions between refined and original sections, particularly in longer

documents. The current implementation also struggles with certain rhetorical structures common in humanities writing, such as extended metaphors or narrative digressions, where the style transfer algorithms tend to oversimplify complex stylistic devices.

6.2 Potential Application Scenarios

Beyond individual manuscript preparation, the system presents several valuable application scenarios for academic institutions. Research offices could integrate the framework into their pre-submission workflows to help authors identify potential integrity issues before journal submission. The version tracking capabilities would enable writing centers to provide targeted feedback based on quantitative improvements across multiple drafts. For graduate programs, the longitudinal analysis features could support the development of discipline-specific writing instruction by identifying common challenges in student work.

Journal editorial teams might employ the system as a screening tool during initial manuscript evaluation, particularly for detecting problematic levels of AI assistance in submissions. The detailed reporting interface could assist editors in communicating revision requests by providing specific examples of content requiring human refinement. Academic publishers could also leverage the framework to establish transparent standards for acceptable AI usage in scholarly writing, using the detection metrics to develop evidence-based policies.

6.3 Ethical Considerations and Challenges

The deployment of such systems raises important ethical questions regarding transparency and accountability. While our framework provides explainable metrics, the potential for false positives in AI-detection could lead to unfair accusations if not properly contextualized. Institutions implementing these tools must establish clear protocols for handling disputed results and provide avenues for human review. The plagiarism detection capabilities, though valuable for maintaining academic standards, require careful calibration to avoid penalizing legitimate forms of intertextuality common in scholarly discourse.

Future developments should address the evolving nature of AI-assisted writing, where the boundary between human and machine authorship becomes increasingly blurred. The system's current binary classification approach may need to transition toward more nuanced representations of human-AI collaboration. Additionally, as language models improve their ability to mimic human writing patterns, the detection methods will require continuous updates to maintain effectiveness, potentially leading to an "arms race" between generation and detection technologies.

The framework's ability to track writing patterns over time introduces privacy considerations regarding the longitudinal data it collects. While valuable for pedagogical purposes, this data could potentially be misused for surveillance or unfair assessment if not properly safeguarded. Future versions should incorporate privacy-preserving techniques such as federated learning [36] to enable useful analytics while protecting individual writing histories.

7. Conclusion

The proposed AI-Assisted Academic Text Refinement and Integrity Analysis System represents a significant advancement in supporting ethical manuscript preparation amid the growing prevalence of AI-generated content. By integrating AI-detection, plagiarism screening, and semantic-preserving humanization within a unified framework, the system addresses critical gaps in current academic writing support tools. The experimental results demonstrate superior performance compared to existing solutions, particularly in handling mixed-origin texts and sophisticated plagiarism cases while maintaining high semantic fidelity during refinement.

The framework's multi-faceted approach to text analysis provides authors with actionable insights rather than binary judgments, fostering responsible AI usage in scholarly writing. The dynamic adaptation to disciplinary writing conventions ensures relevance across diverse academic fields, while the explainable reporting mechanisms promote transparency in the evaluation process. The system's workflow integration capabilities position it as a valuable resource for individual researchers, academic institutions, and publishers navigating the challenges of AI-assisted composition.

Looking ahead, the system's modular architecture allows for continuous improvement as detection techniques evolve and writing norms adapt to technological advancements. The balance between automated analysis and human oversight remains crucial, ensuring that the tools enhance rather than replace scholarly judgment. As academic publishing continues to grapple with the implications of generative AI, frameworks like this will play an increasingly important role in maintaining research integrity while harnessing the productivity benefits of emerging technologies.

Funding Declaration:

The authors received no specific funding for this work from any public, commercial, or not-for-profit funding agency

Data availability: The datasets generated and/or analysed during the current study are not publicly available because the raw evaluation samples were not archived in a shareable form. The summary results supporting the findings are reported in the tables and figures of this article.

References

[1] B Doskaliuk, O Zimba, M Yessirkepov, et al. (2025) Artificial intelligence in peer review: enhancing efficiency while preserving integrity. synapse.koreamed.org.

[2] B Stein & SM Zu Eissen (2006) Near similarity search and plagiarism analysis. *From Data and Information Analysis to Knowledge Discovery*.

[3] YT Hou, Y Chang, T Chen, CS Laih, et al. (2010) Malicious web content detection by machine learning. *Expert Systems With Applications*.

[4] Y Mishra, A Jaiswal & G Soni (2026) Humanizing Text with AI: A Natural Language Processing Approach Using Pre-Trained Language Models. researchgate.net.

[5] AA Jairoun, F El-Dahiyat, GA ElRefae, et al. (2024) Detecting manuscripts written by generative AI and AI-assisted technologies in the field of pharmacy practice. *Unable to determine the complete publication venue*.

[6] LH Baniata, IKE Ampomah & S Park (2021) A transformer-based neural machine translation model for Arabic dialects that utilizes subword units. *Sensors*.

[7] R Saptono, H Prasetyo & A Irawan (2018) Combination of cosine similarity method and conditional probability for plagiarism detection in the thesis documents vector space model. *Journal Of Telecommunication*.

[8] D Jin, Z Jin, Z Hu, O Vechtomova, et al. (2022) Deep learning for text style transfer: A survey. *Computational Linguistics*.

[9] CE Onuoha (2025) Impact of Artificial Intelligence (AI) on the Quality, Efficiency, and Transparency of the Scholarly Publishing Process. tsp.scione.com.

[10] P Dmitrii, L Mestetsky, et al. (2024) N-Gram Perplexity-Based AI-Generated Text Detection. In *2024 Ivannikov Ispras Open Conference on Computer Science and Information Technologies*.

[11] Q He, K Chang, EP Lim & J Zhang (2007) Bursty feature representation for clustering text streams. ... of.

[12] Z Shaheen, G Wohlgenannt & E Filtz (2020) Large scale legal text classification using transformer models. arXiv preprint arXiv:2010.12871.

[13] CH Huang, J Yin & F Hou (2011) A text similarity measurement combining word semantic information with TF-IDF method. *Jisuanji Xuebao*.

[14] S Engels, V Lakshmanan & M Craig (2007) Plagiarism detection using feature-based neural networks. In *Proceedings of the 38th Sigcse Technical Symposium on Computer Science Education*.

[15] M Kobayashi & K Takeda (2000) Information retrieval on the web. *ACM computing surveys (CSUR)*.

[16] P Abisheka, C Deisy & P Sharmila (2024) T-SRE: Transformer-based semantic Relation extraction for contextual paraphrased plagiarism detection. *Journal of King Saud University - Computer and Information Sciences*.

[17] J Mallinson, R Sennrich & M Lapata (2017) Paraphrasing revisited with neural machine translation. In *Conference of the European Chapter of the Association for Computational Linguistics*.

[18] TL Hsu (2021) Academic and formal writing style rewriter. dspace.cityu.edu.hk.

[19] R Alur, R Grosu, I Lee & O Sokolsky (2001) Compositional refinement for hierarchical hybrid systems. In *International Workshop on Hybrid Systems: Computation and Control*.

[20] MS Aslam & S Nisar (2024) Ethical Considerations for Artificial Intelligence Tools in Academic Research and Manuscript Preparation: A Web Content Analysis. *Digital Transformation in Higher Education, Part B*.

[21] J Devlin, MW Chang, K Lee, et al. (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of*.

[22] R Barzilay & M Lapata (2008) Modeling local coherence: An entity-based approach. *Computational Linguistics*.

[23] K Bennett (2009) English academic style manuals: A survey. *Journal of English for academic purposes*.

[24] T Wang, X Zhang, J Feng & X Yang (2020) A comprehensive survey on local differential privacy toward data statistics and analysis. *Sensors*.

[25] BC Björk (2014) Open access subject repositories: An overview. *Journal of the Association for Information Science and Technology*.

[26] J Li, T Tang, WX Zhao, JY Nie & JR Wen (2024) Pre-trained language models for text generation: A survey. *ACM Computing Surveys*.

[27] L Luo, J Ming, D Wu, P Liu & S Zhu (2014) Semantics-based obfuscation-resilient binary code similarity comparison with applications to software plagiarism detection. In *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*.

[28] C Chaka (2023) Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning and Teaching*.

[29] M Maktabdar Oghaz, L Babu Saheer, et al. (2025) Detection and classification of ChatGPT-generated content using deep transformer models. *Frontiers in Artificial Intelligence*.

[30] S Dahl (2007) Turnitin®: The student perspective on using plagiarism detection software. *Active Learning in Higher Education*.

[31] V Chandere, S Satish, et al. (2021) Online plagiarism detection tools in the digital age: A review. *Unable to determine the complete publication venue*.

[32] NT Xuyen (2023) Using the online paraphrasing tool Quillbot to assist students in paraphrasing the Source Information: English-majored students' perceptions. *Unable to determine the complete publication venue*.

[33] X Zhao (2023) Leveraging artificial intelligence (AI) technology for English writing: Introducing wordtune as a digital writing assistant for EFL writers. *Relc Journal*.

[34] T Zhang, V Kishore, F Wu, KQ Weinberger, et al. (2019) Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

[35] R Flesch (2007) Flesch-Kincaid readability test. *Retrieved October, 2007*.

[36] Y Cheng, Y Liu, T Chen & Q Yang (2020) Federated learning for privacy-preserving AI. *Communications of the ACM*.