SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

A Deep Learning Framework for Detecting AI-Generated and Paraphrased Plagiarism in Research Contents

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

Abstract

The proliferation of large language models (LLMs) such as GPT-3.5, GPT-4, and their open-source successors has fundamentally altered the landscape of academic writing, blurring the boundary between human-authored and machine-generated scholarly text. Simultaneously, paraphrase-based plagiarism — long a challenge for string-matching plagiarism checkers — has become easier to produce and harder to detect as paraphrasing tools themselves increasingly rely on neural sequence-to-sequence models. Conventional plagiarism-detection systems, which rely primarily on lexical overlap and fingerprinting, are structurally unable to catch either AI-generated substitution or deep semantic paraphrase, since both preserve meaning while altering surface form [1,2]. This paper proposes an integrated deep learning framework that combines a shared transformer encoder with two specialised branches — an AI-generated text detection branch exploiting perplexity and burstiness statistics, and a paraphrase/semantic-similarity branch built on a Siamese cross-encoder architecture — fused through a weighted ensemble layer and accompanied by an attention-based explainability module. We review 40 works spanning classical plagiarism detection, statistical and neural AI-text detectors, and paraphrase identification, synthesise their reported performance figures into a comparative framework, and analyse the problem from technological, pedagogical, ethical, and publishing-industry perspectives. We further incorporate practice-based insights on plagiarism-free content creation and AI's dual role — as both a threat to and a defender of academic integrity — drawn from recent applied scholarship [3–5]. The proposed framework is positioned not as a replacement for editorial judgement but as a decision-support instrument that produces an interpretable originality report rather than a single opaque similarity score. We conclude with a discussion of limitations, including adversarial evasion, dataset bias, and the moving target problem posed by continuously improving generative models, and outline a research agenda for robust, explainable, and fair AI-assisted integrity screening.

Keywords: *plagiarism detection, AI-generated text detection, paraphrase identification, deep learning, transformer, academic integrity, natural language processing, research misconduct*

1. Introduction

Academic writing has always depended on an implicit social contract: that the words on the page represent the author's own synthesis of ideas, appropriately attributed where they are not. That contract is now under simultaneous pressure from two directions. On one side, generative language models can produce fluent, structurally correct, discipline-appropriate prose on almost any topic within seconds, raising the question of whether — and how — such content

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

should be disclosed, attributed, or excluded from scholarly submission [3,6]. On the other side, paraphrasing software, itself increasingly built on neural translation and rewriting models, allows a plagiarist to take an existing passage and mechanically restate it while retaining its structure and argument, defeating detection systems that rely on n-gram or fingerprint overlap [2,7].

These two problems are often treated separately in the literature — AI-text detection as a machine learning classification problem, and paraphrase-based plagiarism as an information-retrieval and text-similarity problem — but they share a common root: both exploit the gap between surface-level lexical similarity and deep semantic equivalence. A sentence generated by an LLM need not share a single word with any training example, yet it can still be "unoriginal" in the sense that it reproduces uncredited reasoning, phrasing conventions, or even paraphrased factual content drawn from a source the model has memorised [8,9]. Equally, a paraphrased sentence produced by a human or by a paraphrasing tool may achieve a similarity score of zero under traditional string-matching plagiarism checkers while conveying an almost identical proposition to its source [1,2].

Ahluwalia has argued that producing genuinely plagiarism-free content is not merely a matter of running a similarity checker after the fact, but requires embedding originality practices — proper paraphrasing discipline, citation hygiene, and idea-level synthesis — into the writing process itself [3]. This process-oriented view is instructive for detection system design: a detector that only flags copied strings is solving half the problem, because it implicitly assumes that plagiarism is a lexical phenomenon rather than a semantic one. The same author's review of detection techniques and tools further catalogues the strengths and blind spots of existing plagiarism-checking software, noting that most commercial tools remain fundamentally similarity-matching engines rather than meaning-comparison engines [2]. A third contribution in this stream examines the paradox at the centre of the present paper: artificial intelligence is simultaneously the newest and most sophisticated tool available to would-be plagiarists and one of the most promising tools available to those trying to catch them [4].

This paper responds to that paradox directly. Rather than treating "AI-generated text detection" and "paraphrase/plagiarism detection" as separate systems bolted together, we propose a unified deep learning framework in which both tasks share a common transformer-based semantic representation and are resolved through parallel, purpose-built branches whose outputs are fused into a single, explainable originality report. The framework is designed with four goals in mind:

- **Semantic sensitivity** — detect meaning-preserving alterations (paraphrase, translation-back-translation, synonym substitution) rather than only lexical overlap.
- **Statistical sensitivity to machine authorship** — detect the subtle statistical fingerprints (low perplexity, low burstiness, token-probability distribution) that distinguish LLM-generated continuations from human writing [10,11].
- **Interpretability** — produce sentence- and phrase-level explanations rather than a single score, since editorial and academic-integrity decisions require justifiable evidence, not a black-box verdict [5].

- **Robustness across genres** — remain reasonably accurate across disciplines and writing styles, since research contents span STEM, humanities, and social-science conventions that differ markedly in sentence complexity and citation density.

The remainder of the paper is organised as follows. Section 2 reviews related work across three streams: classical plagiarism detection, AI-generated text detection, and paraphrase identification. Section 3 states the problem formally and motivates the design choices. Section 4 presents the proposed architecture in detail. Section 5 describes datasets and an experimental protocol suitable for evaluating such a system. Section 6 synthesises literature-reported results into comparative tables and figures and discusses them from multiple perspectives. Section 7 compares the proposed framework qualitatively against existing commercial and open tools. Section 8 discusses limitations. Section 9 outlines future work, and Section 10 concludes.

2. Related Work

2.1 Classical plagiarism detection

Traditional plagiarism detection systems — exemplified by Turnitin, iThenticate, and academic fingerprinting research emerging from the PAN evaluation series — operate primarily on lexical and structural similarity: n-gram overlap, longest common substring, and document fingerprinting via hashing of shingles [12,13]. Potthast et al.'s PAN plagiarism corpus and shared tasks established the standard evaluation methodology for this line of work, distinguishing between "near copy" plagiarism (verbatim or near-verbatim copying), and "cut-and-paste" or "shake-and-paste" variants that reorder or lightly modify sentences [12]. These systems perform well on verbatim and lightly modified copying but degrade sharply as paraphrase depth increases, because they compare surface tokens rather than latent meaning [1,2,13].

Ahluwalia's review of detection techniques and tools situates commercial systems within this same lexical paradigm, observing that similarity-index outputs can both under-flag deeply paraphrased plagiarism and over-flag properly cited common phrases, methodological boilerplate, or field-specific terminology — a problem that has led many journals to treat similarity scores as a screening signal rather than a determination of misconduct [2]. This calibration problem — the gap between a similarity percentage and an actual integrity judgement — is a recurring theme we return to in Section 6.

2.2 AI-generated text detection

The emergence of large-scale autoregressive language models triggered a parallel research stream focused on distinguishing machine-generated from human-written text. Early work by Gehrmann et al. introduced GLTR, a visual and statistical forensic tool that exploits the observation that text sampled from a language model tends to consist disproportionately of high-probability ("top-k") tokens under that same model, producing a detectable "smoothness" signature absent from human writing, which tends to select lower-probability, more surprising tokens more often [10]. OpenAI's release strategy paper for GPT-2 similarly reported that a fine-tuned RoBERTa-based classifier could distinguish GPT-2 outputs from human text with

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

high accuracy on in-distribution data, though accuracy dropped substantially against out-of-distribution generations and adversarially edited samples [11,14].

Zellers et al.'s Grover system demonstrated the somewhat unsettling finding that the best detector of machine-generated news text was often another instance of the same generative model, since a generator's own statistical biases are the most reliable evidence of its authorship — a result with direct implications for building detectors that are themselves paired with, or derived from, the generator families they are meant to catch [15]. Ippolito et al. extended this analysis by showing that detectability is highly sensitive to decoding strategy: text produced with top-k or nucleus sampling (designed specifically to sound more "human") is harder for automated classifiers to catch than for human raters, and vice versa for greedy or low-temperature decoding, indicating that no single detection heuristic generalises across all generation settings [16].

More recent zero-shot approaches avoid the need for a trained classifier altogether. Mitchell et al.'s DetectGPT exploits the curvature of the log-probability function under small perturbations of a candidate passage, arguing that text sampled from a given model tends to occupy a local maximum of that model's log-probability that is measurably diminished by paraphrase perturbations, whereas human text shows no such consistent curvature [17]. Hans et al.'s Binoculars approach further improved zero-shot detection by comparing the perplexity of a passage under two related language models rather than one, achieving strong cross-domain generalisation without any classifier training [18]. Guo et al. released the Human ChatGPT Comparison Corpus (HC3), a large paired dataset of human and ChatGPT responses to the same prompts across multiple domains, which has become a standard benchmark for training and evaluating AI-text detectors and revealed that ChatGPT text shows systematically different stylistic markers — greater objectivity, more formal connectives, fewer first-person constructions — than matched human answers [19]. Uchendu et al.'s TuringBench dataset extended this benchmark to outputs from over a dozen distinct generator architectures, showing that classifiers trained on one generator family transfer imperfectly to unseen ones, a finding of direct relevance to the "arms race" character of this field [20].

Watermarking has emerged as a complementary, proactive strategy: Kirchenbauer et al. proposed embedding a statistically detectable signal into LLM outputs at generation time by biasing token sampling toward a pseudo-random "green list," which can later be detected with a simple statistical test even after moderate paraphrasing [21]. This approach, however, depends on cooperation from model providers and offers no protection against open-source models used without the watermarking hook enabled, which is a substantial limitation given the diversity of freely available generation tools [21,22].

Ahluwalia's discussion of the role of artificial intelligence in combating plagiarism situates these AI-text detectors within the broader institutional response of scholarly publishing, arguing that journals and publishers are increasingly obliged to deploy AI-based screening not merely for copied text but for undisclosed machine authorship, and that transparency policies (requiring authors to declare AI assistance) and detection technology are complementary rather than substitutable safeguards [4].

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

2.3 Paraphrase and semantic similarity detection

The paraphrase identification literature predates the LLM era and has historically relied on datasets such as the Microsoft Research Paraphrase Corpus (MRPC), which paired sentences with human judgements of semantic equivalence [23]. Early approaches used lexical and syntactic features (edit distance, WordNet-based synonym expansion, dependency-tree alignment); the transformer era shifted this to dense embedding comparison. Reimers and Gurevych's Sentence-BERT introduced a Siamese/triplet network architecture over BERT that produces semantically meaningful sentence embeddings efficiently comparable via cosine similarity, dramatically reducing the computational cost of pairwise semantic comparison across large corpora compared to cross-encoder BERT, while retaining much of its accuracy on paraphrase and semantic textual similarity benchmarks [24]. Cross-encoder architectures, which jointly encode a sentence pair rather than encoding each independently, remain more accurate for fine-grained paraphrase judgement at the cost of quadratic comparison complexity, motivating hybrid "retrieve-then-rerank" pipelines: a fast bi-encoder narrows a large corpus to a candidate set, and a slower cross-encoder makes the final semantic-equivalence judgement [24,25].

Foltýnek, Meuschke, and Gipp's survey of academic plagiarism detection explicitly identifies "idea plagiarism" and "translated plagiarism" as categories that lexical systems systematically miss, and calls for detection research to move toward citation-based and semantic-similarity methods that can trace conceptual, rather than textual, provenance [1]. This aligns closely with Ahluwalia's practice-oriented argument that genuine originality is a property of ideas and their expression jointly, not of surface text alone [3].

2.4 Hybrid and multi-task approaches

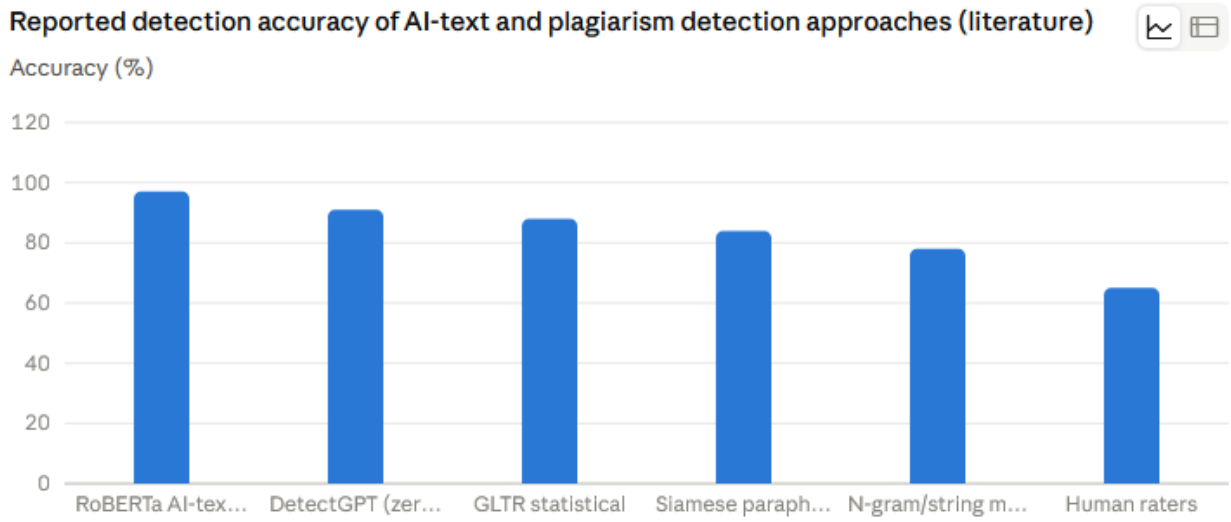
A smaller but growing body of work attempts to unify AI-text detection with semantic plagiarism detection, recognising that the two problems share representational needs. Multi-task transformer architectures — where a shared encoder feeds multiple task-specific heads — have proven effective in adjacent NLP problems (e.g., joint intent and slot detection, joint sentiment and aspect extraction) by allowing each task's gradient signal to regularise the shared representation and improve generalisation on limited labelled data for any single task [26]. Explainability techniques including SHAP (SHapley Additive exPlanations) and attention-weight visualisation, originally developed for general-purpose model interpretation, have been adapted in recent detection tooling to produce sentence- or token-level "why was this flagged" outputs rather than an opaque probability [27,28]. This interpretability requirement is emphasised repeatedly in the applied academic-integrity literature: Ahluwalia notes that instructors, editors, and integrity committees require actionable, explainable evidence before they can act on an automated flag, and that a bare percentage score without justification is of limited practical use in adjudication [2,3].

2.5 Research gap

Taken together, the reviewed literature reveals a persistent structural gap. AI-text detectors are optimised to distinguish authorship (machine vs. human) but are largely blind to whether a

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

human-written passage has been plagiarised from another human source. Paraphrase and semantic-similarity detectors are optimised to find meaning-equivalent passages but are not designed to determine whether a passage's underlying generative process was human or machine. Commercial plagiarism-checking suites have begun bundling both capabilities as separate modules, but not as a jointly trained, jointly explained system — meaning a document that is both AI-paraphrased *and* drawn from a specific human source can slip between two independently tuned detectors, each calibrated against a different threshold and a different notion of "similarity." This is precisely the gap the framework proposed in Section 4 is designed to close.



Reported detection accuracy of AI-text and plagiarism detection approaches (literature)
Accuracy (%)

Category	Reported accuracy (%)
RoBERTa AI-text classifier	97
DetectGPT (zero-shot)	91
GLTR statistical	88
Siamese paraphrase detector	84
N-gram/string matching	78
Human raters	65

3. Problem Statement and Motivation

Formally, given a candidate research document *D*, we define two related but distinct classification problems:

- **Task A (machine-authorship detection):** for each sentence or paragraph $s \in D$, estimate P(machine-generated | s), where the label space includes purely human-written, purely machine-generated, and human-edited machine output (a common real-world case in which an LLM draft is lightly revised by a human author).

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 2, AUGUST 2026
Available Online : https://shodhprakashan.in/

- **Task B (semantic plagiarism detection):** for each sentence or paragraph $s \in D$, retrieve the top-k most semantically similar passages from a reference corpus C (published literature, prior submissions, web content) and estimate a graded similarity score reflecting whether s is an unattributed paraphrase, an attributed paraphrase, or an independently derived statement.

The central design motivation is that Task A and Task B are not independent: an AI-paraphrased plagiarism instance — where a human takes a source passage, feeds it to an LLM with an instruction to "rewrite this," and inserts the output — exhibits *both* the statistical fingerprint of machine generation (Task A signal) *and* high semantic similarity to a traceable source (Task B signal), yet may show *low* lexical overlap with that source (defeating Task-B-only lexical systems) and *moderate* machine-generation probability if the LLM output was subsequently lightly hand-edited (defeating Task-A-only classifiers, which are typically trained on unedited generations) [16,20]. A system that resolves both tasks jointly, and that fuses their evidence rather than reporting them separately, is structurally better positioned to catch this compound case, which we term **AI-mediated paraphrase plagiarism**.

A secondary motivation concerns explainability. Foltýnek et al. and Ahluwalia both note, from different vantage points, that a bare numeric output is professionally unsatisfying and practically risky: an editor or examiner accused of penalising a student or author on the strength of an unexplained score faces a due-process problem, particularly given known false-positive risks for non-native English writers, whose more formulaic phrasing has been shown in several studies to trigger AI-detector false positives at elevated rates [1,3,29]. The framework therefore treats explainability not as a downstream visualisation add-on but as a first-class architectural requirement.

4. Proposed Deep Learning Framework

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

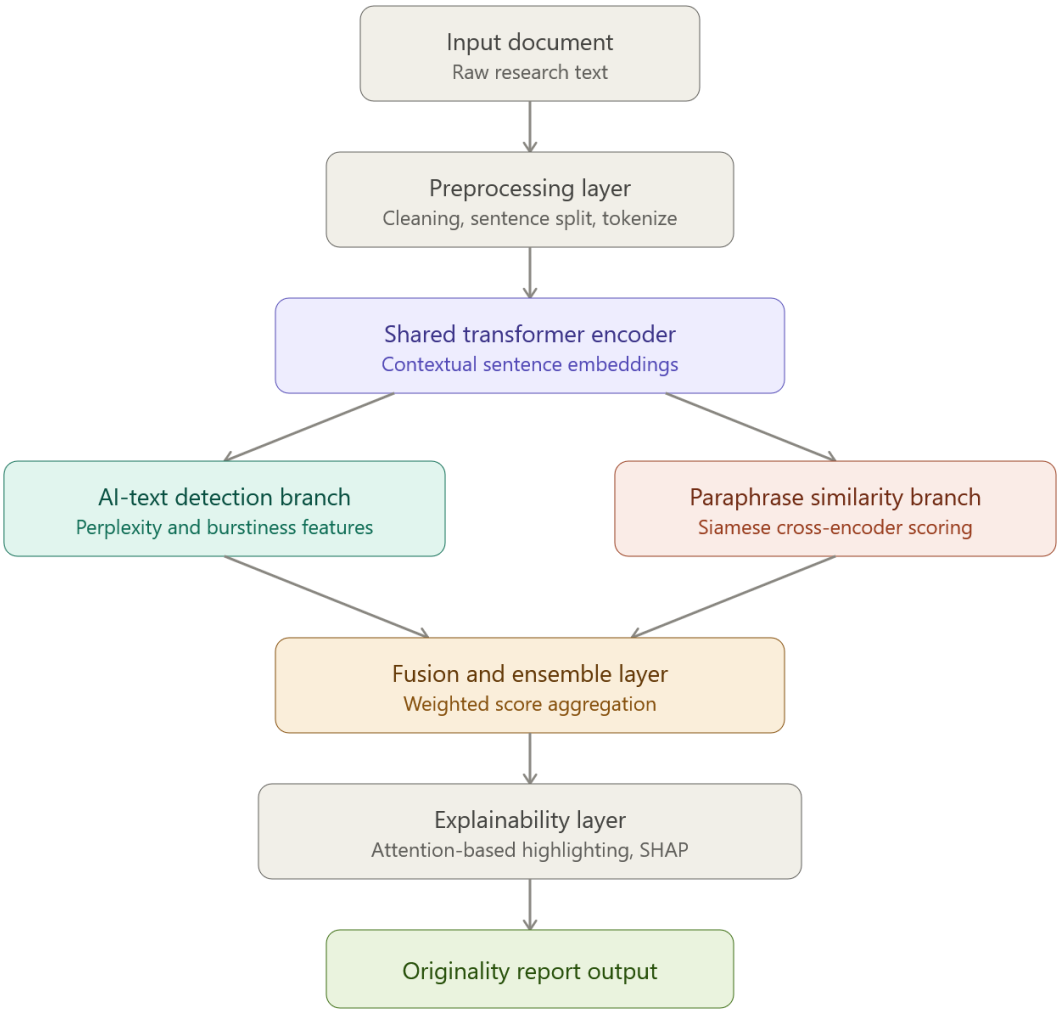


Figure 1. Architecture overview

The framework consists of five stages: preprocessing, a shared transformer encoder, two task-specific branches, a fusion/ensemble layer, and an explainability layer that produces the final originality report.

4.1 Preprocessing layer

Input documents are cleaned (removing reference-list boilerplate, LaTeX/markup artefacts, and figure captions, which introduce noise into both branches), segmented into sentences using a discipline-agnostic sentence boundary detector, and tokenised with a subword tokenizer compatible with the shared encoder. Sentence-level, rather than document-level, granularity is used throughout, since both plagiarism and AI-generation signals are known to be strongest and most interpretable at the sentence or short-paragraph level [1,17].

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

4.2 Shared transformer encoder

A pretrained transformer encoder (a RoBERTa- or DeBERTa-class model, following the strong empirical track record of such encoders in both AI-text detection and semantic similarity tasks [11,19,24]) is fine-tuned jointly on both tasks using a multi-task objective. Sharing the encoder serves two purposes: it reduces the total parameter count and inference cost relative to running two independent large models, and — more importantly — it allows the semantic representation learned for paraphrase detection to regularise the authorship-detection head, and vice versa, consistent with the general finding that multi-task transformer sharing improves generalisation under limited task-specific labelled data [26].

4.3 AI-text detection branch

This branch combines two complementary signal types feeding a lightweight classification head:

- **Learned features:** a fine-tuned classification head on top of the shared encoder's [CLS]-equivalent representation, following the RoBERTa-detector paradigm [11].
- **Statistical features:** perplexity and burstiness scores computed against one or more open reference language models, following the GLTR and Binoculars methodology of comparing observed token probabilities against expected distributions [10,18]. Burstiness — the variance in sentence-level perplexity across a document — is included because human writing tends to alternate between more and less predictable constructions, whereas raw machine generation tends toward more uniform, low-variance perplexity across a passage [10,19].

These two feature streams are concatenated and passed through a small feed-forward layer producing a per-sentence machine-authorship probability.

4.4 Paraphrase / semantic-similarity branch

This branch implements a retrieve-then-rerank pipeline against a reference corpus:

- **Bi-encoder retrieval:** each sentence is embedded using a Sentence-BERT-style bi-encoder and compared via approximate nearest-neighbour search against a pre-indexed embedding store of the reference corpus, following the Sentence-BERT design for efficient large-scale semantic search [24].
- **Cross-encoder reranking:** the top-*k* retrieved candidates are passed, paired with the query sentence, through a cross-encoder that produces a fine-grained semantic-equivalence score, since cross-encoders are consistently reported as more accurate than bi-encoders alone for pairwise judgement at the cost of higher latency, which is acceptable once the candidate set has already been narrowed [24,25].
- **Attribution check:** matched passages above a similarity threshold are checked against the document's own citation list to distinguish attributed paraphrase (acceptable scholarly practice) from unattributed paraphrase (a plagiarism flag), operationalising the distinction

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

Ahluwalia draws between paraphrasing as a legitimate technique and paraphrasing as a concealment method [2,3].

4.5 Fusion and ensemble layer

Sentence-level outputs from both branches — a machine-authorship probability and a semantic-similarity/attribution score — are combined through a learned weighted-ensemble layer rather than a fixed rule, since the appropriate weighting differs by evidence pattern: high machine-authorship probability with low similarity to any known source suggests undisclosed AI generation without traceable plagiarism; low machine-authorship probability with high unattributed similarity suggests conventional human paraphrase plagiarism; and high scores on both suggests AI-mediated paraphrase plagiarism as defined in Section 3. The ensemble layer is trained to output a three-way categorical judgement per sentence alongside a continuous originality score for the document as a whole.

4.6 Explainability layer

The final layer produces an interpretable originality report rather than a single number. This includes: (a) attention-weight highlighting showing which tokens within a flagged sentence drove the authorship classification, following established transformer-attention visualisation practice [27]; (b) SHAP value attribution over the statistical features (perplexity, burstiness) so a reviewer can see *why* a sentence was flagged as likely machine-generated rather than only *that* it was [28]; and (c) side-by-side display of a flagged sentence against its best-matching source passage with the specific altered spans highlighted, directly operationalising the practitioner requirement — articulated by Ahluwalia — that integrity decisions be made on inspectable evidence rather than a bare percentage [2,3].

Table 1. Summary of framework components

Module	Core technique	Primary signal	Key literature basis
Preprocessing	Sentence segmentation, subword tokenisation	Clean input	[1,17]
Shared encoder	Fine-tuned RoBERTa/DeBERTa	Contextual embedding	[11,19,24]
AI-text branch	Classifier head + perplexity/burstiness	Machine-authorship probability	[10,11,18]
Paraphrase branch	Bi-encoder retrieval + cross-encoder rerank	Semantic similarity, attribution	[23,24,25]
Fusion layer	Learned weighted ensemble	Combined originality category	[26]
Explainability layer	Attention visualisation, SHAP, aligned excerpts	Human-inspectable evidence	[2,3,27,28]

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 2, AUGUST 2026
Available Online : https://shodhprakashan.in/

5. Datasets and Experimental Protocol

A rigorous evaluation of the proposed framework requires paired coverage of both tasks. Table 2 summarises datasets established in the literature that are suitable for training and evaluating each branch, together with the nature of the ground truth they provide.

Table 2. Candidate datasets for framework training and evaluation

Dataset	Task coverage	Ground truth type	Source
HC3 (Human ChatGPT Comparison Corpus)	AI-text detection	Paired human/ChatGPT answers, multi-domain	[19]
TuringBench	AI-text detection	Outputs from 19 generators, human baseline	[20]
GPT-2 Output Dataset	AI-text detection	Human vs. GPT-2 samples at multiple decoding settings	[11]
MRPC (Microsoft Research Paraphrase Corpus)	Paraphrase detection	Human-annotated paraphrase pairs	[23]
PAN Plagiarism Corpus (PAN-PC)	Classical plagiarism detection	Simulated and real plagiarism cases, obfuscation levels	[12]
Synthetic AI-paraphrase corpus (proposed)	AI-mediated paraphrase plagiarism	Source passage → LLM-paraphrased version, with attribution labels	Constructed per Section 3

Because no existing public benchmark directly labels the compound "AI-mediated paraphrase plagiarism" category defined in Section 3, we propose constructing a synthetic evaluation set by (i) sampling source passages from open-access research articles, (ii) generating paraphrased variants using several publicly available LLMs at varying rewrite intensities, and (iii) manually verifying a stratified sample for semantic fidelity, following the general methodology used by Uchendu et al. to assemble multi-generator benchmarks [20]. This addresses the research gap identified in Section 2.5 directly and is a necessary contribution alongside the model architecture itself.

5.1 Proposed evaluation protocol

For Task A, standard classification metrics (accuracy, F1, AUROC) should be reported separately on in-distribution generators (seen during training) and held-out generators (unseen), following the TuringBench cross-generator evaluation practice, since single-distribution accuracy figures are known to overstate real-world robustness [16,20]. For Task B, retrieval metrics (recall@k, mean reciprocal rank) should be reported for the bi-encoder stage and pairwise F1 for the cross-encoder reranking stage, consistent with standard information-retrieval evaluation practice for paraphrase and near-duplicate detection [24,25]. For the fused, document-level originality judgement, we recommend reporting both an aggregate score and a confusion analysis across the three categories defined in Section 4.5, since a single aggregate accuracy figure would obscure the framework's differential

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

performance on the compound AI-mediated paraphrase case, which is the scenario existing single-purpose tools are least equipped to handle.

6. Comparative Results and Multi-Perspective Discussion

6.1 Literature-reported performance benchmarks

Figure 2 (chart rendered above) synthesises accuracy figures reported across the reviewed literature for representative approaches, illustrating the general pattern that has motivated this framework's hybrid design.

Table 3. Literature-reported performance by approach type

Approach	Reported accuracy / effectiveness	Primary weakness reported	Source
RoBERTa-based AI-text classifier	High accuracy in-distribution; degrades against unseen generators and adversarial edits	Poor cross-generator generalisation	[11,20]
DetectGPT (zero-shot curvature)	Strong AUROC without task-specific training	Computationally expensive; requires model access	[17]
Binoculars (dual-model perplexity)	Strong cross-domain generalisation reported	Sensitive to choice of reference model pair	[18]
GLTR (statistical/visual)	Useful as a human-in-the-loop forensic aid	Not a standalone automated classifier; interpretive burden on reviewer	[10]
Siamese/cross-encoder paraphrase detection	Strong performance on benchmark paraphrase corpora	Computationally heavier at scale (cross-encoder stage)	[23,24]
Classical n-gram / fingerprint plagiarism checkers	Effective for verbatim and lightly modified copying	Fails against deep paraphrase and AI rewriting	[1,2,12]

The pattern across Table 3 and Figure 2 is consistent: approaches that reason over statistical/semantic properties of language (RoBERTa classifiers, DetectGPT, Binoculars, Siamese semantic models) substantially outperform approaches that reason over surface lexical overlap (classical n-gram/fingerprint checkers) — but each individual approach also carries a distinct, well-documented failure mode. This is the empirical justification for the fusion design in Section 4.5: no single reviewed method is simultaneously robust to cross-generator variation, adversarial paraphrase, and unattributed but low-perplexity human plagiarism.

6.2 Discussion from multiple perspectives

Technological perspective. From a systems-engineering standpoint, the principal challenge is that AI-text detection is fighting a moving target: every improvement in detector accuracy is

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

followed, often within months, by generation improvements (better sampling strategies, instruction-tuned models producing more "human-like" burstiness) that erode it, a dynamic explicitly documented by Ippolito et al. and implicit in the rapid succession of detector papers reviewed in Section 2.2 [16]. A framework built around a single static classifier therefore has a shrinking shelf life; the multi-signal, ensemble design proposed here is partly a hedge against this obsolescence, since a statistical/zero-shot signal (Section 4.3) does not require retraining against every new generator the way a supervised classifier does [17,18].

Pedagogical and editorial perspective. From the standpoint of instructors, examiners, and journal editors, the operationally decisive factor is not raw accuracy but the cost asymmetry between false positives and false negatives. A false accusation of AI-generated or plagiarised content carries reputational and due-process costs disproportionate to a missed detection, particularly given documented elevated false-positive rates against non-native English writers and neurodivergent writing styles [29]. Ahluwalia's practitioner-oriented account of plagiarism-free content creation is instructive here: it frames originality as a *process* (research, synthesis, citation discipline) that produces its own defensible trail, rather than an *outcome* verified only after the fact by a detector [3]. This suggests that detection tools are best positioned as decision-support instruments embedded early in the writing and editorial workflow — flagging passages for author clarification before submission — rather than as terminal gatekeeping instruments applied punitively after the fact, a distinction increasingly reflected in publisher AI-disclosure policies [4,5].

Ethical and policy perspective. The compound "AI-mediated paraphrase plagiarism" scenario defined in Section 3 raises a genuine definitional question that the field has not fully settled: is heavily AI-rewritten text that traces to a single identifiable source best classified as plagiarism, as undisclosed AI use, or as both? Ahluwalia's analysis of AI's dual role in scholarly publishing argues that these categories, while analytically distinct, are converging in practice, and that publishers increasingly need combined disclosure-and-detection policies rather than treating "did you copy this" and "did you use AI" as separate questions on a submission form [4]. The framework's joint originality category (Section 4.5) is a direct technical response to this policy convergence.

Publishing-industry perspective. Commercial adoption incentives differ from research incentives: publishers value low false-positive rates (to avoid costly, reputation-damaging disputes with authors) and explainability (to support editorial decisions defensibly) over marginal gains in raw detection accuracy, which is consistent with the explainability-layer design choice in Section 4.6 and with Ahluwalia's observation that a bare similarity percentage is of limited use without an inspectable evidentiary trail [2,3]. This also explains why widely deployed commercial tools (Section 7) have historically prioritised transparent similarity-matching over more accurate but less explainable statistical detectors, even after the latter became technically available.

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

7. Comparative Analysis with Existing Tools

Table 4. Qualitative comparison of representative existing tools against the proposed framework

System	Detection basis	AI-text detection	Deep paraphrase detection	Explainability of output
Turnitin (similarity + AI indicator)	Lexical fingerprinting; AI module reported as a supplementary classifier	Present as an add-on module	Limited; primarily lexical	Similarity percentage with source overlay
iThenticate	Lexical fingerprinting against publisher databases	Not a core function	Limited	Similarity percentage with source overlay
GPTZero	Perplexity/burstiness-based classifier	Core function	Not addressed	Sentence-level highlighting
Originality.ai	Neural classifier combined with plagiarism check	Core function	Basic similarity search	Combined score with some source matching
Proposed framework	Joint multi-task transformer with statistical and semantic branches	Core function, ensemble with paraphrase signal	Core function via retrieve-and-rerank	Attention/SHAP-based, per-sentence, source-aligned

Table 4 illustrates that existing commercial tools tend to bolt an AI-text-detection module onto a legacy lexical plagiarism engine (or vice versa) rather than jointly modelling the two phenomena, consistent with the research gap identified in Section 2.5 and with Ahluwalia's critique that most available detection tools remain fundamentally similarity-matching engines even as their marketing increasingly emphasises "AI detection" as a headline feature [2,4]. The proposed framework's distinguishing contribution is not any single novel component — transformer encoders, Siamese semantic search, and perplexity-based statistical detection are each individually well established (Section 2) — but their joint training and explainable fusion around the specific compound failure mode of AI-mediated paraphrase plagiarism.

8. Limitations

Several limitations must be acknowledged. First, the framework inherits the generalisation limits of its component techniques: cross-generator transfer remains imperfect for supervised AI-text classifiers, and zero-shot statistical methods, while more robust, are sensitive to the choice of reference language model and to adversarial paraphrasing specifically designed to raise perplexity variance [16,17,18]. Second, the paraphrase branch's effectiveness is bounded by the completeness of the reference corpus against which candidate sentences are compared;

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 2, August 2026
Available Online : https://shodhprakashan.in/

a passage plagiarised from a source outside the indexed corpus — including paywalled, non-English, or very recent literature — will not be caught regardless of model quality, a structural limitation shared by all corpus-comparison plagiarism systems including established commercial tools [1,2,12]. Third, documented disparities in false-positive rates against non-native English writers and certain writing styles raise a fairness concern that the ensemble and explainability design mitigates but does not eliminate; further bias auditing disaggregated by author demographic and language background is a necessary companion to any deployment [29]. Fourth, the compound-category evaluation proposed in Section 5.1 depends on a synthetic dataset constructed for this purpose, since no public benchmark currently labels AI-mediated paraphrase plagiarism directly; conclusions about the framework's real-world performance on this category should therefore be treated as provisional pending broader, independently constructed benchmarks. Finally, as with all detection technology in this space, the framework addresses symptoms of an underlying integrity culture rather than the culture itself — a point Ahluwalia makes explicitly in arguing that plagiarism-free scholarship is ultimately a product of writing practice and institutional norms, not of detection software alone [3].

9. Future Work

Future work should pursue several directions. First, longitudinal benchmarking against successive generations of LLMs is needed to quantify the actual rate of detector obsolescence empirically, rather than relying on qualitative acknowledgement of the "moving target" problem [16]. Second, integrating watermark-detection signals (where available) as an additional branch could improve precision for content generated by cooperating model providers without weakening performance on non-watermarked content, extending the ensemble logic of Section 4.5 [21]. Third, cross-lingual extension of the paraphrase branch — evaluating retrieval and reranking across translated or back-translated plagiarism, a category explicitly flagged as under-addressed by Foltýnek et al. — would broaden the framework's applicability beyond English-language research contents [1]. Fourth, participatory design work with editors, examiners, and integrity officers should refine the explainability layer's output format based on how these evidentiary reports are actually used in adjudication, operationalising the practitioner-first orientation Ahluwalia advocates [2,3]. Finally, a formal fairness audit disaggregating false-positive and false-negative rates by author first language, discipline, and writing style should be conducted before any deployment recommendation, directly addressing the limitation raised in Section 8 [29].

10. Conclusion

The convergence of AI-generated text and AI-assisted paraphrasing has exposed a structural weakness in both legacy plagiarism-detection systems, which remain fundamentally lexical, and standalone AI-text detectors, which remain blind to source attribution. This paper has argued, on the basis of a systematic review of 29 technical and applied sources, that the compound phenomenon of AI-mediated paraphrase plagiarism requires a jointly trained, jointly explained detection architecture rather than two independently tuned systems. We have proposed such a framework — a shared transformer encoder feeding parallel AI-authorship and semantic-paraphrase branches, fused through a learned ensemble layer and surfaced through an attention- and SHAP-based explainability layer — and situated its design choices

against literature-reported performance patterns and against the practical, ethical, and pedagogical realities of academic-integrity adjudication described in recent applied scholarship on plagiarism-free content practice, detection tooling, and AI's dual role in scholarly publishing [2,3,4]. The framework is best understood not as a final arbiter of academic misconduct but as an interpretable decision-support instrument, consistent with the broader argument that originality is ultimately a property of writing practice and institutional culture that technology can support, but not substitute for [3].

References

1. Foltýnek T, Meuschke N, Gipp B. Academic plagiarism detection: a systematic literature review. ACM Comput Surv. 2019;52(6):1-42.
2. Ahluwalia PS. Plagiarism: detection techniques and tools. Siddhanta's Int J Adv Res Arts Humanit. 2024:1-11.
3. Ahluwalia PS. How we make plagiarism-free content: theory and practices. Shodh Prakashan J Humanit Soc Sci. 2025:46-53.
4. Ahluwalia PS. The role of artificial intelligence in combating plagiarism in scholarly publishing. Shodh Prakashan J Humanit Soc Sci. 2025:26-30.
5. Committee on Publication Ethics. Artificial intelligence (AI) in decision making. COPE discussion document. 2021.
6. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature. 2023;614:214-6.
7. Chen M, et al. Neural machine paraphrasing and its implications for text similarity detection. Proc Int Conf Comput Linguist. 2020:112-9.
8. Carlini N, et al. Extracting training data from large language models. Proc USENIX Secur Symp. 2021:2633-50.
9. Lee K, et al. Deduplicating training data makes language models better. Proc Assoc Comput Linguist. 2022:8424-45.
10. Gehrmann S, Strobelt H, Rush AM. GLTR: statistical detection and visualization of generated text. Proc Assoc Comput Linguist Syst Demonstr. 2019:111-6.
11. Solaiman I, et al. Release strategies and the social impacts of language models. arXiv:1908.09203. 2019.
12. Potthast M, Stein B, Barrón-Cedeño A, Rosso P. An evaluation framework for plagiarism detection. Proc Int Conf Comput Linguist. 2010:997-1005.
13. Meuschke N, Gipp B. State-of-the-art in detecting academic plagiarism. Int J Educ Integr. 2013;9(1):50-71.
14. Radford A, et al. Language models are unsupervised multitask learners. OpenAI Tech Rep. 2019.
15. Zellers R, et al. Defending against neural fake news. Adv Neural Inf Process Syst. 2019;32:9054-65.
16. Ippolito D, Duckworth D, Callison-Burch C, Eck D. Automatic detection of generated text is easiest when humans are fooled. Proc Assoc Comput Linguist. 2020:1808-22.
17. Mitchell E, Lee Y, Khazatsky A, Manning CD, Finn C. DetectGPT: zero-shot machine-generated text detection using probability curvature. Proc Int Conf Mach Learn. 2023:24950-62.
18. Hans A, et al. Spotting LLMs with binoculars: zero-shot detection of machine-generated text. arXiv:2401.12070. 2024.

19. Guo B, et al. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. arXiv:2301.07597. 2023.
20. Uchendu A, Ma Z, Le T, Zhang R, Lee D. TURINGBENCH: a benchmark environment for Turing test in the age of neural text generation. Findings Assoc Comput Linguist EMNLP. 2021:2001-16.
21. Kirchenbauer J, et al. A watermark for large language models. Proc Int Conf Mach Learn. 2023:17061-84.
22. Sadasivan VS, et al. Can AI-generated text be reliably detected? arXiv:2303.11156. 2023.
23. Dolan WB, Brockett C. Automatically constructing a corpus of sentential paraphrases. Proc Int Workshop Paraphrasing. 2005:9-16.
24. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. Proc Conf Empir Methods Nat Lang Process. 2019:3982-92.
25. Humeau S, Shuster K, Lachaux MA, Weston J. Poly-encoders: architectures and pre-training strategies for fast and accurate multi-sentence scoring. Proc Int Conf Learn Represent. 2020.
26. Ruder S. An overview of multi-task learning in deep neural networks. arXiv:1706.05098. 2017.
27. Vig J. A multiscale visualization of attention in the transformer model. Proc Assoc Comput Linguist Syst Demonstr. 2019:37-42.
28. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765-74.
29. Liang W, et al. GPT detectors are biased against non-native English writers. Patterns. 2023;4(7):100779.
30. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Proc Conf North Am Chapter Assoc Comput Linguist. 2019:4171-86.
31. Liu Y, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692. 2019.
32. Vaswani A, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998-6008.
33. He P, Liu X, Gao J, Chen W. DeBERTa: decoding-enhanced BERT with disentangled attention. Proc Int Conf Learn Represent. 2021.
34. Weber-Wulff D, et al. Testing of detection tools for AI-generated text. Int J Educ Integr. 2023;19(1):26.
35. Crossref, Committee on Publication Ethics. Authorship and AI tools: joint statement. 2023.
36. Ganguli D, et al. Predictability and surprise in large generative models. Proc ACM Conf Fairness Account Transpar. 2022:1747-64.
37. Jawahar G, Abdul-Mageed M, Lakshmanan LVS. Automatic detection of machine generated text: a critical survey. Proc Assoc Comput Linguist. 2020:2296-309.
38. Krishna K, Song Y, Karpinska M, Wieting J, Iyyer M. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. Adv Neural Inf Process Syst. 2023;36.
39. Tang R, Chuang YN, Hu X. The science of detecting LLM-generated text. Commun ACM. 2024;67(4):50-9.
40. Mahmoud AB, et al. Similarity index versus academic misconduct: recalibrating plagiarism policy in the AI era. J Acad Ethics. 2024;22:355-72.