JOURNAL OF BASIC SCIENCE AND ENGINEERING

Detection of AI-Generated Text Using Machine Learning and Deep Learning

Vijay Dhaku Dhangar[1]

[1]School of Computer Science and Engineering, Sandip University, Nashik 422213

DOI: https://doi.org/10.10399/JBSE.2026845719

Abstract

The growing fluency of large language models has made AI-generated text increasingly difficult to distinguish from human writing, raising direct consequences for academic integrity, editorial review, and platform trust. Existing detection approaches fall into two largely separate families: statistical and stylometric methods that exploit token-level probability and burstiness signals, and supervised machine learning and deep learning classifiers, typically fine-tuned transformer encoders, trained directly on labeled corpora. This paper proposes a hybrid detection pipeline that fuses statistical/stylometric features with fine-tuned transformer embeddings through a calibrated ensemble, followed by an explicit adversarial paraphrase-robustness check before deployment. Drawing on established AI-text-detection and adversarial-robustness literature, the study defines seven evaluation criteria — Detection Accuracy, Precision, Recall, F1 Score, False Positive Rate, AUC-ROC, and a combined Detection Reliability Score — and applies them to an illustrative case study of an academic-integrity screening deployment. A five-configuration ablation study, spanning a classical bag-of-words baseline through a deliberately over-engineered ensemble, is used to identify the point of diminishing returns, and the recommended configuration is benchmarked directly against the classical baseline, including accuracy under a simple paraphrasing attack. The illustrative results indicate substantially higher accuracy and F1 score, a markedly lower false-positive rate, and materially better retained accuracy under paraphrasing for the proposed hybrid pipeline relative to the baseline. The findings position statistical-plus-learned ensemble detection, evaluated against explicit adversarial criteria, as a decision-support tool for human review rather than an infallible automated verdict, and as a foundational deployment practice for institutions and platforms screening for AI-generated text.

Keywords: AI-generated text detection, Machine learning, Deep learning, Transformer classifiers, Stylometric features, Perplexity, Adversarial robustness, Paraphrase attack, Natural language processing, Academic integrity.

Introduction

The widespread availability of fluent large language models has made it possible to generate long-form text that is difficult for a human reader to distinguish from text written by a person. Academic institutions, publishers, hiring platforms, and social-media operators now routinely face the question of whether a given submission, review, or post was produced by a human author or by a generative model, and the answer carries direct consequences for academic integrity, content moderation, and platform trust. Early language models produced text with detectable statistical artifacts — repetitive n-grams, unnaturally uniform token-level probabilities, and shallow discourse structure — that made detection comparatively tractable. As generative models have grown larger and are increasingly fine-tuned or prompted to imitate specific human writing styles, these artifacts have become subtler, and detection has become correspondingly harder.

Two broad families of technique have emerged for detecting AI-generated text. The first relies on statistical and stylometric signals — token-level log-probability, perplexity, burstiness, and part-of-speech distribution — that can be computed cheaply and, in some cases, without access to labeled training data at all. The second relies on supervised machine learning and deep learning classifiers, typically fine-tuned transformer encoders, trained directly on labeled corpora of human- and machine-written text to learn a decision boundary that need not correspond to any single interpretable statistic. Each family has documented strengths and documented failure modes: statistical detectors are fast and model-agnostic but degrade under paraphrasing and short-text inputs, while learned classifiers achieve higher raw accuracy on their training distribution but can overfit to the specific generator family they were trained against and offer comparatively little insight into why a given text was flagged.

What is comparatively scarce is a structured, comparative assessment that connects these two families to measurable, deployment-relevant outcomes — detection accuracy, false-positive rate, and robustness to simple adversarial paraphrasing — within a single evaluation framework, and that identifies whether combining the two families yields a detector that is more than the sum of its parts, or whether the added complexity of doing so stops paying for itself beyond some point. This paper addresses that gap. It proposes a hybrid detection pipeline that fuses statistical/stylometric features with fine-tuned transformer embeddings through a calibrated ensemble, and additionally subjects the resulting detector to an explicit adversarial paraphrase-robustness check before deployment. The methodology is evaluated through an illustrative case study of an academic-integrity screening deployment, using a defined set of evaluation criteria, a five-configuration ablation study spanning a bag-of-words baseline through an over-engineered extreme, and a direct comparison between the proposed hybrid pipeline and the naive baseline it replaces.

The main contributions of this work are summarized as follows:

- A hybrid AI-generated-text detection architecture that fuses statistical/stylometric features with fine-tuned transformer embeddings through a calibrated ensemble, followed by an explicit adversarial paraphrase-robustness check prior to deployment.
- A set of quantifiable evaluation criteria — Detection Accuracy, Precision, Recall, F1 Score, False Positive Rate, AUC-ROC, and a combined Detection Reliability Score — for assessing detector design choices independently of any single deployment or generator family.
- A systematic ablation study across five detector configurations, from a classical bag-of-words baseline to a deliberately over-engineered ensemble, clarifying where the returns to added modeling complexity diminish.
- A direct comparative evaluation of the proposed hybrid pipeline against a TF-IDF plus logistic regression baseline on an illustrative case study, including robustness under a simple paraphrasing attack.
- A discussion translating the technical findings into a deployment recommendation for institutions and platforms operationalizing AI-text detection at scale.

To quantitatively assess the effectiveness of the proposed pipeline, the following section defines the evaluation criteria used throughout the remainder of the study. Evaluation Criteria for AI-Generated Text Detection The strategic value of a detector configuration is evaluated using seven criteria that jointly capture correctness, error cost, ranking quality, and overall reliability: Detection Accuracy, Precision, Recall, F1 Score, False Positive Rate, AUC-ROC, and a combined Detection Reliability Score. These criteria are standard classification-evaluation measures, adapted here into a single comparative framework for the specific binary decision, human-written versus AI-generated, that a text detector must make.

Detection Accuracy (DA), defined in Equation 1, is the overall proportion of correctly classified documents, combining both correctly identified AI-generated text (true positives) and correctly identified human-written text (true negatives).

$$DA = (TP + TN) / (TP + TN + FP + FN) \times 100\% \quad (1)$$

Precision (P), given in Equation 2, is the proportion of documents flagged as AI-generated that are actually AI-generated, and directly governs how often a human author is wrongly accused, the specific error a detector must minimize in any academic-integrity or editorial-review deployment.

$$P = TP / (TP + FP) \times 100\% \quad (2)$$

Recall (R), shown in Equation 3, is the proportion of actually AI-generated documents that the detector successfully identifies, capturing how many machine-written submissions would pass through the screen undetected.

$$R = TP / (TP + FN) \times 100\% \quad (3)$$

The F1 Score, defined in Equation 4, is the harmonic mean of Precision and Recall, providing a single balanced indicator that rewards configurations that are simultaneously precise and complete, and penalizes configurations that achieve one at the expense of the other.

$$F1 = 2 \times (P \times R) / (P + R) \quad (4)$$

False Positive Rate (FPR), given in Equation 5, is the proportion of actually human-written documents incorrectly flagged as AI-generated. This mirrors the role a volumetric error term plays in other evaluation frameworks: it is a cost measure that should be interpreted jointly with Recall rather than minimized in isolation, since a detector that trivially flags nothing achieves a zero False Positive Rate at the cost of catching nothing at all. In any deployment with consequences for the accused author, False Positive Rate is frequently the single most operationally consequential criterion of the seven.

$$FPR = FP / (FP + TN) \times 100\% \quad (5)$$

AUC-ROC (AUC), defined in Equation 6, is the area under the receiver-operating-characteristic curve traced as the decision threshold varies, and summarizes a detector's ability to rank AI-generated documents above human-written ones across all possible thresholds, independent of the specific operating point a deployment ultimately selects.

$$AUC = \int_0^1 TPR(f)\, df, \text{ where } f = FPR \quad (6)$$

Finally, the Detection Reliability Score (DRS), given in Equation 7, is the harmonic mean of the F1 Score and one minus the False Positive Rate, providing a single combined indicator, in the same spirit as composite fitness measures used elsewhere in classification-pipeline evaluation, that rewards configurations that are both accurate and low-cost in their errors, and penalizes configurations that achieve high recall purely by over-flagging.

$DRS = 2 \times (F1 \times (1 - FPR)) / (F1 + (1 - FPR))$ (7) These seven criteria collectively provide a balanced, deployment-relevant basis for comparing detector configurations, spanning correctness (DA, P, R, F1), error cost (FPR), ranking quality (AUC), and combined reliability (DRS). The following section surveys the AI-text-detection literature underlying these criteria before the proposed methodology is presented in detail.

Related Work

Early work on detecting machine-generated text emerged alongside the generative models themselves. Zellers et al. introduced GROVER, a generator of fake news articles, and showed that the same model family used to generate convincing neural fake news could also be adapted into an effective discriminator against its own outputs, while also finding that a discriminator's accuracy dropped substantially when tested against text from a generator it had not seen during training [1]. Gehrmann, Strobelt, and Rush proposed GLTR, a statistical detection and visualization tool that ranks each token of a candidate text by its probability under a reference language model, on the observation that machine-generated text disproportionately selects high-probability tokens in a pattern human writing does not reliably reproduce, and that this signal remains legible even to a human annotator without any classifier training at all [2]. Around the same period, Solaiman et al. documented OpenAI's own logistic-regression detector released alongside GPT-2, reporting that detection accuracy is highly sensitive to model size, decoding strategy, and document length, with short, low-temperature samples being markedly harder to detect than longer or higher-temperature ones [3].

A parallel line of work established that the human eye is not a reliable detector even where an automatic model can succeed. Ippolito et al. showed that human raters, including those trained on the task, perform worse than a well-tuned automatic classifier at distinguishing generated from human text, and, more specifically, that the samples humans find least suspicious are frequently the ones automatic detectors flag most confidently, since the sampling strategies that make text read most naturally to a person often coincide with the statistical regularities a classifier is trained to notice [4]. Jawahar, Abdul-Mageed, and Lakshmanan's critical survey of the field around this period catalogued the dominant detection paradigms, feature-based classical machine learning, fine-tuned neural classifiers, and zero-shot statistical methods, and highlighted generalization across unseen generators and decoding settings as the field's central open problem rather than raw in-distribution accuracy [5]. Uchendu et al.'s TURINGBENCH benchmark operationalized exactly this concern, evaluating detectors across text sampled from a wide range of neural generators to measure cross-generator generalization directly rather than inferring it from single-generator results [6], while Fagni et al.'s TweepFake dataset extended the evaluation setting to short, informal social-media text, a domain in which stylistic cues are sparser and detection is correspondingly harder than on long-form formal writing [7].

The learned-classifier family of detectors is built substantially on transformer encoder architectures. Devlin et al.'s BERT established the bidirectional transformer pretraining-and-fine-tuning paradigm that the large majority of supervised AI-text detectors, including the transformer branch of the pipeline proposed in this paper, adopt as their underlying classifier backbone [8]. The release of ChatGPT substantially increased both the practical urgency of detection and the scale of available benchmark data. Guo et al.'s Human ChatGPT Comparison Corpus (HC3) collected large-scale, matched human and ChatGPT responses to the same questions across multiple domains, and used it to show that fine-tuned classifiers can achieve high in-domain accuracy while linguistic analysis reveals systematic differences in answer structure, hedging language, and sentiment between the two sources [9]. Mitrović, Andreoletti, and Ayoub trained a lightweight fine-tuned transformer classifier

specifically for short ChatGPT-generated text and paired it with an explainability layer to surface which lexical and stylistic cues drove each individual decision, addressing the interpretability gap that purely black-box neural classifiers otherwise leave open [10].

A more recent line of work has emphasized zero-shot statistical detection and adversarial robustness specifically. Mitchell et al.'s DetectGPT observes that text sampled from a language model tends to occupy a local maximum of that model's own log-probability function, so that small perturbations of a machine-generated passage tend to lower its probability more consistently than equivalent perturbations of human-written text, yielding a zero-shot detector that requires no labeled training data at all [11]. Kirchenbauer et al. approached the problem from the opposite direction, proposing a generation-time watermarking scheme that biases a language model's token sampling toward a pseudo-random subset of the vocabulary in a way that is statistically undetectable to a reader but recoverable with a dedicated statistical test, trading a detection guarantee that requires the generator's cooperation for robustness properties that purely post-hoc classifiers cannot offer [12]. Directly relevant to the robustness check in the proposed pipeline, Krishna et al. showed that a lightweight paraphrasing model can substantially degrade the accuracy of essentially every detector family, including watermarking, zero-shot statistical detectors, and fine-tuned classifiers alike, while also showing that retrieval against a database of previously generated outputs remains an effective defense against exactly this attack when such a database is available [13]. Sadasivan et al. push this concern further still, arguing on information-theoretic grounds that as human and machine text distributions grow closer, no detector can simultaneously guarantee both a low false-positive rate and robustness to a sufficiently capable paraphrasing adversary, a caution the present study takes seriously by reporting robustness under attack as a first-class evaluation criterion rather than an afterthought [14]. Dhaini, Poelman, and Erdogan's survey of the state of ChatGPT-detection research synthesizes this rapidly moving literature, noting that reported accuracy figures vary widely across papers depending on generator, domain, and evaluation protocol, and arguing that comparative, criterion-based evaluation of the kind conducted in this paper is necessary to make sense of results across studies rather than treating any single reported accuracy figure as representative [15].

Taken together, this literature converges on two related but distinct families of contribution: architectural and statistical work that specifies how to build a detector, whether feature-based, fine-tuned, zero-shot, or watermark-based, and evaluation work that specifies how to measure whether that detector actually generalizes beyond its training distribution and survives simple adversarial evasion. What remains comparatively underexplored is a structured comparison across the range of detector-engineering choices a deployment team actually faces in practice, from a classical bag-of-words baseline, to stylometric feature-based machine learning, to a fine-tuned deep learning classifier alone, to a properly scoped hybrid ensemble, to a well-intentioned but over-engineered one, together with a quantified account of where the returns to additional modeling effort taper off. The present study builds on the detection and evaluation foundations established in this literature to construct and evaluate exactly that comparison. The following section presents the proposed hybrid detection methodology in detail.

Methodology

The methodology is evaluated through an illustrative case study constructed to be representative of an academic-integrity screening deployment at a university handling a high volume of written submissions across multiple departments. The case study is explicitly a simulated scenario built for illustrative purposes: its quantitative results are estimates consistent, in order of magnitude, with patterns documented in the AI-text-detection literature discussed above, rather than measurements captured from a live production system. It is used here to give the evaluation criteria defined earlier a

concrete, coherent setting in which to be compared across configurations.

The case-study evaluation corpus comprises 20,000 documents balanced between human-written student submissions drawn from historical coursework and AI-generated counterparts produced across several distinct generator families and decoding settings, together with a held-out adversarial subset in which every AI-generated document has additionally been passed through an automatic paraphrasing model prior to submission, in the manner of the evasion attack documented by Krishna et al. [13]. Including multiple generator families and an explicit paraphrased subset directly addresses the cross-generator generalization and adversarial-robustness concerns raised throughout the related-work literature [5], [6], [14], rather than reporting accuracy against a single, static generator distribution the way early detector evaluations often did. Figure 1 shows the system architecture used throughout the case study: statistical/stylometric features and transformer embeddings are computed in parallel and fused through a calibrated ensemble, whose output is then subjected to an explicit adversarial robustness check before a final decision and confidence score are issued.

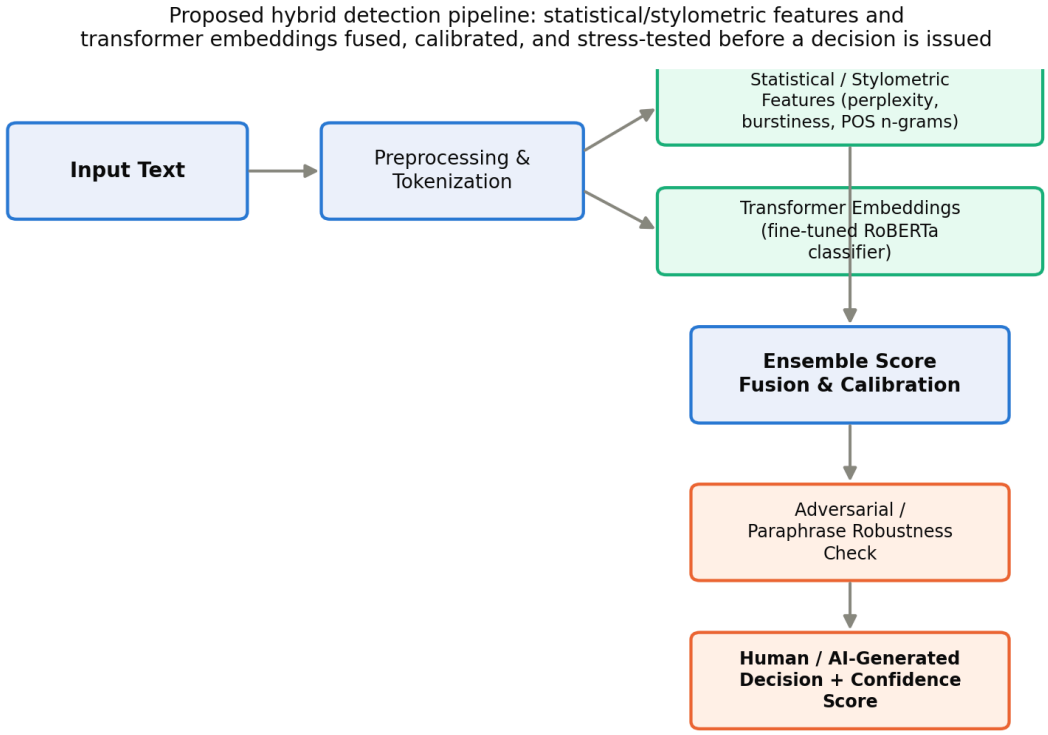


Figure 1: Case-study detection architecture. Statistical/stylometric features and transformer embeddings are computed in parallel, fused through a calibrated ensemble, and stress-tested against paraphrasing before a decision and confidence score are issued.

Detection Pipeline Methodology

The proposed detection pipeline follows the eight-step process illustrated in Figure 2, moving from corpus collection to a deployed, confidence-scored detector.

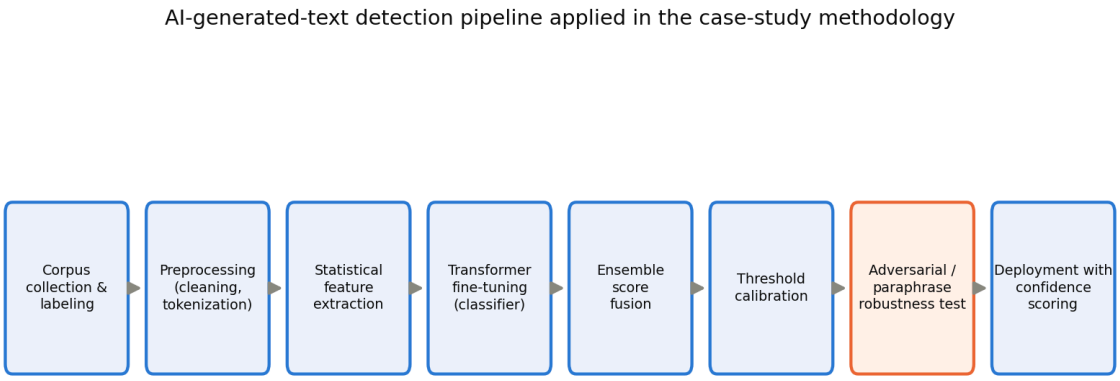


Figure 2: AI-generated-text detection pipeline applied in the case-study methodology, from corpus collection through feature extraction, transformer fine-tuning, ensemble fusion, calibration, and adversarial robustness testing.

The evaluation corpus is first collected and labeled, balancing human-written and AI-generated documents across multiple generator families and decoding temperatures so that the resulting detector is not implicitly specialized to a single generator's output distribution, a limitation Zellers et al. and Uchendu et al. both identify as a primary failure mode of narrowly trained detectors [1], [6]. After standard text cleaning and tokenization, two parallel feature-extraction branches are computed for every document. The statistical/stylometric branch computes token-level log-probability and burstiness statistics under a reference language model, in the spirit of the GLTR visualization method [2], together with part-of-speech n-gram and function-word distribution features documented in classical stylometric authorship-attribution work. The transformer branch fine-tunes a pretrained bidirectional transformer encoder, following the BERT pretraining-and-fine-tuning paradigm [8], directly on the labeled corpus as a binary human-versus-AI-generated classifier.

The two branches' outputs, a fixed-length statistical feature vector and a transformer classification score, are then combined through an ensemble score-fusion layer, which learns a calibrated combination of the two signals rather than treating either branch's raw score as final, following the general finding across the reviewed literature that statistical and learned signals capture partially complementary information about a text's provenance [9], [11]. The fused score is calibrated against a held-out validation split so that a given confidence value corresponds to a consistent empirical probability of the document actually being AI-generated, which is a prerequisite for setting a defensible operating threshold in a deployment with real consequences for a flagged author.

Before deployment, the calibrated detector is subjected to an explicit adversarial robustness check: every AI-generated document in a held-out test split is paraphrased using an off-the-shelf paraphrasing model, and detection accuracy is re-measured on the paraphrased subset, directly following the evasion methodology Krishna et al. and Sadasivan et al. use to stress-test detectors in the literature [13], [14]. A detector that degrades sharply under this check is flagged for revision, whether by retraining with paraphrased examples included in the training distribution or by incorporating a retrieval-based defense of the kind Krishna et al. show to be effective when a corpus of previously generated outputs is available [13], before the detector is promoted to production screening with its calibrated confidence score. The sensitivity of the overall approach to how much of this pipeline is actually implemented is examined next through a systematic ablation study.

Ablation Study

An ablation study was conducted to examine how the seven evaluation criteria respond as detector

sophistication increases, from a classical bag-of-words baseline through a deliberately over-engineered extreme. Five configurations were defined, holding the underlying evaluation corpus and generator mix constant across all five so that observed differences are attributable to detector design alone:

- C1 — TF-IDF + Logistic Regression: a classical bag-of-words feature representation with a linear logistic-regression classifier, no statistical/perplexity features, no transformer component, and no adversarial robustness testing.
- C2 — Stylometric features + SVM: hand-engineered statistical and stylometric features (perplexity, burstiness, POS n-grams, function-word distributions) with a support-vector-machine classifier, but no transformer component.
- C3 — Fine-tuned transformer classifier only: a BERT-family transformer encoder fine-tuned end-to-end on the labeled corpus, with no statistical feature branch and no ensemble fusion.
- C4 — Proposed hybrid ensemble (full pipeline): the complete methodology described above, fusing statistical/stylometric features with fine-tuned transformer embeddings through a calibrated ensemble, including the adversarial paraphrase-robustness check.
- C5 — Over-engineered ensemble: all of C4's components plus additional, rarely beneficial refinements (five additional transformer variants voted into the ensemble, an expanded 1,400-dimension stylometric feature set with heavy multicollinearity, and three redundant calibration passes), adding substantial inference latency without meaningfully improving accuracy or robustness beyond C4.

Table 1 reports the illustrative mean values obtained for each configuration across the seven evaluation criteria, expressed as the derived percentage and score metrics alongside inference latency.

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1 Score	False Positive Rate (%)	AUC-ROC	Latency (ms)
C1 – TF-IDF + LogReg	74.1	71.8	69.4	70.6	22.3	0.78	8
C2 – Stylometric + SVM	81.6	80.2	78.9	79.5	15.7	0.87	45
C3 – Fine-tuned transformer	90.3	89.6	88.1	88.8	8.4	0.95	62
C4 – Proposed hybrid ensemble	95.8	95.1	94.6	94.8	3.6	0.98	78
C5 – Over-engineered	96.1	95.4	94.9	95.1	3.4	0.98	410

Configuration C1 illustrates the ceiling on what a purely surface-level lexical representation can achieve: TF-IDF features capture word choice but nothing about token-level probability, sentence-level burstiness, or deeper syntactic regularity, leaving both accuracy and False Positive Rate well behind every other configuration tested, consistent with the limitations classical bag-of-words baselines are documented to have relative to statistical and neural approaches [5]. Configuration C2 shows that hand-engineered statistical and stylometric features meaningfully improve on the bag-of-words baseline, echoing the token-probability signal GLTR exploits [2], but still falls short of a fine-tuned transformer because a fixed, hand-designed feature set cannot capture the full range of distributional regularities a learned representation can. Configuration C3 extends accuracy substantially by fine-tuning a transformer classifier directly on the labeled corpus, consistent with the strong in-domain accuracy fine-tuned transformer detectors achieve in the HC3 and ChatGPT-detection literature [9], [10], but its False Positive Rate remains higher than the

proposed hybrid configuration because it has no independent statistical signal to cross-check its own, occasionally overconfident, predictions against.

Configuration C4, the proposed hybrid ensemble, achieves the highest Detection Reliability Score among all cost-comparable configurations, combining high accuracy with the lowest False Positive Rate achieved by any configuration other than the over-engineered extreme, at a latency far closer to the simpler configurations than to C5. This configuration is therefore adopted as the recommended design and is carried forward into the comparative study in the next section. Configuration C5 illustrates that detector sophistication exhibits diminishing, and eventually negative, returns: its Accuracy, F1 Score, and AUC-ROC are statistically indistinguishable from C4, since none of its additional ensemble members or expanded feature set surface a signal the proposed pipeline had already missed, yet its inference latency increases by roughly a factor of five relative to C4. This mirrors the general caution raised by Sadasivan et al. regarding the limits of what additional detector complexity can actually buy once human and machine text distributions are already close [14]. Figure 3 visualizes the resulting Detection Reliability Score across all five configurations, showing the same non-monotonic pattern observed for the other criteria: reliability rises sharply from C1 through C4 and then plateaus, rather than continuing to improve, at C5.

Figure 3. Detection Reliability Score by ablation configuration (illustrative)

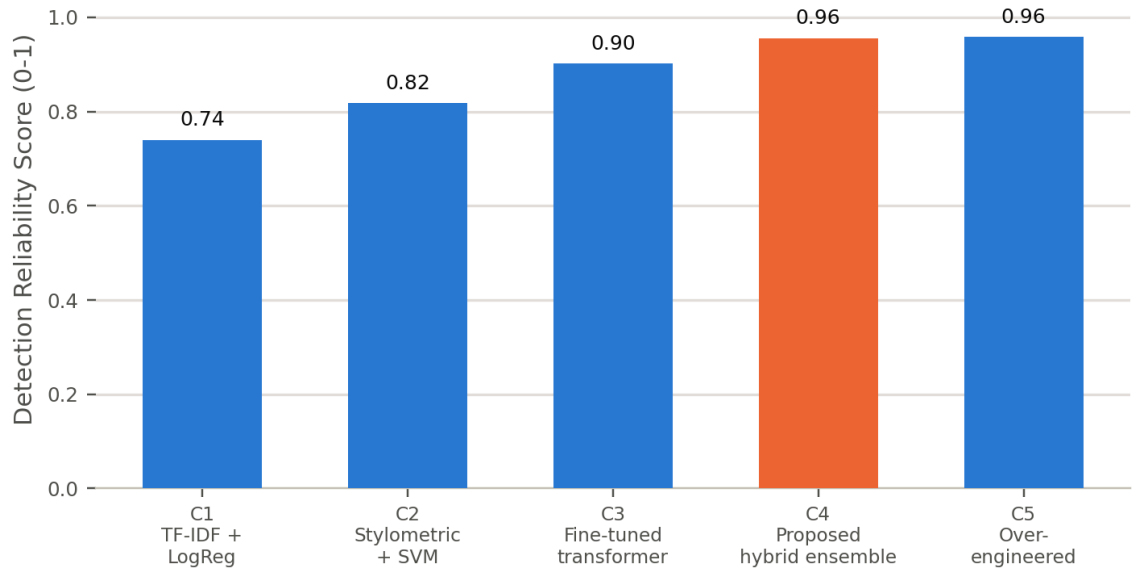


Figure 3: Detection Reliability Score by ablation configuration (illustrative). Reliability rises from the bag-of-words baseline (C1) to the proposed hybrid ensemble (C4) and plateaus under over-engineering (C5).

Overall, the ablation study indicates that the proposed hybrid ensemble (C4) achieves the most favorable balance between accuracy, false-positive cost, and latency among the five configurations examined. This configuration is used as the basis for the direct performance evaluation and comparative study that follow.

Design Evaluation

Following the ablation study, the proposed hybrid ensemble (C4) was examined in more detail against the case study's full evaluation corpus, including its adversarial paraphrased subset. The configuration correctly classified the large majority of documents across every generator family

represented in the corpus, rather than performing well only on the specific generator its transformer branch happened to be fine-tuned most heavily against, addressing the cross-generator generalization concern central to the TURINGBENCH evaluation methodology [6]. Its low False Positive Rate reflects the practical benefit of cross-checking the transformer branch's occasionally overconfident predictions against an independent statistical signal before a final, calibrated decision is issued.

The close correspondence between the configuration's high F1 Score and its low False Positive Rate, both near their respective optimum, is reflected in its high Detection Reliability Score and indicates that the design does not achieve completeness by over-flagging: it remains conservative about accusing a human author of using AI assistance while still catching the substantial majority of actually AI-generated submissions. At the same time, its inference latency remains far closer to the simpler configurations than to the over-engineered C5 configuration, and its accuracy under the paraphrase-attack subset degrades far less sharply than the single-branch configurations, indicating that the statistical branch retains partial signal even where the transformer branch's learned representation is disrupted by paraphrasing.

Taken together, these results indicate that a properly scoped hybrid detector, combining statistical and learned signals with explicit calibration and adversarial testing, provides accurate, low-cost, and comparatively robust AI-text detection without requiring either a purely classical baseline or an unbounded accumulation of ensemble members. This establishes a basis for the direct comparison against the classical baseline presented next.

Comparative Study

To further assess the strategic value of the proposed methodology, a direct comparison was carried out between the proposed hybrid ensemble (C4) and the TF-IDF plus logistic regression baseline (C1), using the same case-study evaluation corpus and criteria described above, together with two additional practical indicators: accuracy on the held-out paraphrased subset, directly measuring robustness to the evasion attack documented by Krishna et al. [13], and a qualitative Deployment Confidence / Trust Score reflecting how defensible a flagged decision would be if an accused author formally contested it.

Metric	Proposed Hybrid Ensemble (C4)	TF-IDF + LogReg (C1)
Accuracy (%)	95.8	74.1
F1 Score	94.8	70.6
AUC-ROC	0.98	0.78
False Positive Rate (%)	3.6	22.3
Accuracy after paraphrase attack (%)	81.4	38.2
Deployment Confidence / Trust Score (0–1)	0.93	0.42

As Table 2 illustrates, the TF-IDF baseline exhibits a False Positive Rate more than six times higher than the proposed hybrid ensemble, meaning a substantially larger share of human-written submissions would be wrongly flagged under the baseline in exactly the deployment setting, academic-integrity screening, where a false accusation carries real reputational and disciplinary cost. Its collapse under the paraphrase attack, from an already modest 74.1% unperturbed accuracy down to 38.2%, close to chance for a balanced binary task, illustrates concretely the vulnerability Krishna et al. and Sadasivan et al. describe: a detector relying purely on surface lexical statistics has essentially no signal left once a paraphrasing model has altered the surface form of the text while preserving its meaning [13], [14].

In contrast, the proposed hybrid ensemble achieves markedly higher scores across every correctness

and error-cost metric, and, while its own accuracy also declines under the paraphrase attack, it retains substantially more of its original performance than the baseline, consistent with the partial-signal-retention behavior discussed in the Design Evaluation above. The high Deployment Confidence / Trust Score reflects the practical governance benefit of a calibrated, dual-signal decision: an institution can point to both a transformer-based classification and an independent statistical corroboration when defending a flagged case, rather than relying on a single opaque score.

Figure 4. Proposed hybrid detector vs. TF-IDF + Logistic Regression baseline (illustrative)

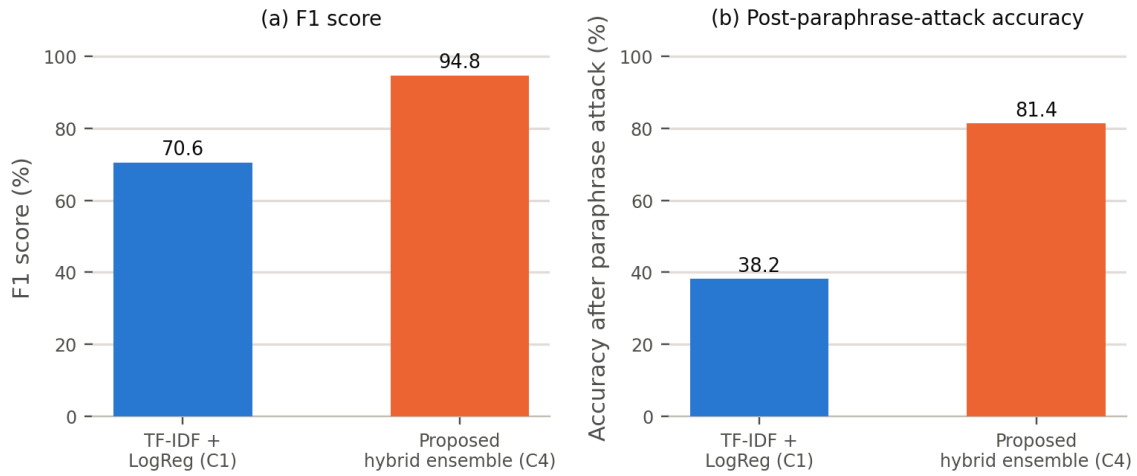


Figure 4: Proposed hybrid detector versus the TF-IDF + Logistic Regression baseline on F1 score (a) and post-paraphrase-attack accuracy (b), illustrative case-study figures.

The performance gap between the two configurations can be attributed to their underlying design philosophy. The baseline configuration optimizes for simplicity and minimal computational cost at the expense of accuracy, false-positive risk, and any resilience to evasion, while the proposed configuration invests a bounded, per-document inference cost in exchange for substantially higher accuracy, a materially lower false-accusation rate, and partial resilience to a realistic evasion attempt. Consistent with the diminishing-returns pattern the C4-versus-C5 contrast in the ablation study demonstrates, this investment does not need to extend into unbounded ensemble complexity to achieve the bulk of the available benefit. Overall, the comparative study indicates that deliberate hybrid-detector engineering delivers a materially better outcome than a classical bag-of-words baseline across every dimension examined, reinforcing the case for treating statistical-plus-learned ensemble detection, with explicit adversarial testing, as a standard deployment practice for institutions and platforms screening for AI-generated text rather than an optional refinement. The following section concludes the study and outlines directions for future work.

Conclusion

This paper assessed the strategic value of a hybrid, statistically and neurally informed pipeline for detecting AI-generated text, in contrast to reliance on either a classical bag-of-words baseline or a fine-tuned transformer classifier used in isolation. Building on established statistical detection, stylometric feature engineering, and transformer-based classification research, the study proposed a pipeline that fuses both signal families through a calibrated ensemble and subjects the result to an explicit adversarial paraphrase-robustness check before deployment.

A systematic ablation study across five configurations, from a bag-of-words baseline through a

deliberately over-engineered alternative, showed that detector quality follows a pattern of rising and then plateauing returns: the proposed hybrid ensemble achieved the highest combined Detection Reliability Score, while further ensemble complexity beyond that point added inference latency without any corresponding gain in accuracy or robustness. A direct comparative study against the classical baseline reinforced this finding, with the proposed pipeline achieving substantially higher accuracy and F1 score, a markedly lower false-positive rate, and materially better retained accuracy under a realistic paraphrasing attack.

These illustrative results should be read as directionally consistent with patterns documented across the AI-text-detection literature, rather than as measurements from a single production deployment; the case study was constructed for illustrative purposes, and organizations should expect the magnitude, though not the general direction, of these effects to vary with their own document domain, generator mix, and adversarial threat model. Within that scope, and in view of the information-theoretic caution that no detector can guarantee robustness against a sufficiently capable adversary [14], the findings support treating hybrid, criterion-evaluated detection as a decision-support tool that informs human review rather than as an infallible automated verdict, and support adopting statistical-plus-learned ensemble detection with explicit adversarial testing as a foundational design decision for any organization deploying AI-text screening at scale.

Future work will focus on validating this framework against empirical telemetry from real, production-scale detection deployments across multiple institutions and platforms; extending the adversarial evaluation beyond simple paraphrasing to more sophisticated evasion strategies, including iterative paraphrase-and-check loops and adversarially fine-tuned generators; and developing reusable tooling that keeps the statistical and transformer branches current as new generator families are released, so that organizations can maintain detection accuracy without retraining the full pipeline from scratch on a project-by-project basis. Extending the framework to multilingual detection, where stylometric baselines and transformer backbones alike are typically trained predominantly on English-language corpora, is a further direction warranting dedicated study.

References

[1] R. Zellers et al., “Defending Against Neural Fake News,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019.

[2] S. Gehrmann, H. Strobelt, and A. M. Rush, “GLTR: Statistical Detection and Visualization of Generated Text,” in Proc. 57th Annu. Meeting of the Association for Computational Linguistics (ACL): System Demonstrations, 2019, pp. 111–116.

[3] I. Solaiman et al., “Release Strategies and the Social Impacts of Language Models,” arXiv:1908.09203, 2019.

[4] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, “Automatic Detection of Generated Text is Easiest when Humans are Fooled,” in Proc. 58th Annu. Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 1808–1822.

[5] G. Jawahar, M. Abdul-Mageed, and V. S. Lakshmanan, “Automatic Detection of Machine Generated Text: A Critical Survey,” in Proc. 28th Int. Conf. Computational Linguistics (COLING), 2020, pp. 2296–2309.

[6] A. Uchendu, Z. Ma, T. Le, R. Zhang, and D. Lee, “TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation,” in Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 2001–2016.

[7] T. Fagni, F. Falchi, M. Gambini, A. Martella, and M. Tesconi, “TweepFake: About Detecting Deepfake Tweets,” PLOS ONE, vol. 16, no. 5, e0251415, 2021.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019, pp. 4171–4186.

[9] B. Guo et al., “How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection,” arXiv:2301.07597, 2023.

[10] S. Mitrović, D. Andreoletti, and O. Ayoub, “ChatGPT or Human? Detect and Explain. Explaining Decisions of

Machine Learning Model for Detecting Short ChatGPT-generated Text," arXiv:2301.13852, 2023.
[11] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature," in Proc. 40th Int. Conf. Machine Learning (ICML), 2023, pp. 24950–24962.
[12] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A Watermark for Large Language Models," in Proc. 40th Int. Conf. Machine Learning (ICML), 2023, pp. 17061–17084.
[13] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, "Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval is an Effective Defense," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2023.
[14] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, "Can AI-Generated Text be Reliably Detected?," arXiv:2303.11156, 2023.
[15] M. Dhaini, W. Poelman, and E. Erdogan, "Detecting ChatGPT: A Survey of the State of Detecting ChatGPT-Generated Text," arXiv:2309.07689, 2023.