Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 3 (2026) 767 - 775 e-ISSN: 2580-0760

Detecting AI-Generated Text with Fine-Tuned RoBERTa: A Cross-Validated Study with Interpretability Analysis

Yuhefizar[1], Ronal Watrianthos[2*], Dony Marzuki[3]
[1,2]Department of Information Technology, Politeknik Negeri Padang, Padang, Indonesia
[2]Department of English, Politeknik Negeri Padang, Padang, Indonesia

[*]*ronalwatrianthos@pnp.ac.id*

Abstract

The rapid proliferation of large language models capable of generating fluent and contextually coherent text has made reliable authorship attribution both a practical and ethical necessity. Outputs from systems such as GPT-4, LLaMA 2, and Claude are increasingly indistinguishable from human-authored prose, yet current detection methodologies remain insufficient to meet this challenge. This study aims to develop and evaluate a fine-tuned RoBERTa-base binary classifier for distinguishing AI-generated text from human-authored text under stratified cross-validation conditions; assess the model's robustness to a known length-based confound through a controlled ablation study; and provide interpretable explanations of the classifier's decisions through dual token-level attribution using Gradient × Input Saliency and Attention Rollout. The model is trained and evaluated via stratified five-fold cross-validation on a balanced corpus of 6,069 samples (3,069 AI-generated and 3,000 human-authored) and benchmarked against four established baselines spanning classical, neural, and zero-shot paradigms. The classifier achieves a mean accuracy of 99.93%, a weighted F1-score of 0.9993, and a mean AUC-ROC of 1.0000 across all five folds. A length-balanced ablation produces an accuracy decline of only 0.11 percentage points, providing empirical evidence that the model has learned genuine semantic and stylistic distinctions rather than superficial length-based cues. Gradient × Input Saliency maps indicate that AI-generated texts are characterised by technically precise, domain-specific vocabulary, whereas human-authored texts are identified through discourse markers and rhetorically structured expression, with both attribution methods converging on the same high-saliency tokens.

Keywords: AI-generated text; gradient saliency; large language models; natural language processing; RoBERTa

How to Cite: Y. Yuhefizar, Ronal Watrianthos, and D. Marzuki, "Detecting AI-Generated Text with Fine-Tuned RoBERTa: A Cross-Validated Study with Interpretability Analysis", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 3, pp. 767 - 775, Jun. 2026.
*Permalink/*DOI: https://doi.org/10.29207/resti.v10i3.7660

Received: March 22, 2026
Accepted: June 15, 2026
Available Online: June 21, 2026

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

Recent advancements in artificial intelligence, particularly the development of large language models[1], [2], have introduced both significant opportunities and challenges [3-5]. These models, exemplified by systems such as GPT-4 [6], [7], LLaMA 2[8], and Claude [9], [10], have rapidly evolved from generating awkward, mechanically recognizable text to producing writing that is indistinguishable from human-authored content in uncontrolled reading environments. This capability spans a wide range of domains, including scientific abstracts, legal briefs, journalistic commentary, and literary fiction.

Recent empirical studies have documented a 55.4% increase in AI-generated news articles on mainstream websites [11]. The trajectory of commercial adoption has been similarly rapid, ChatGPT achieved 100 million users within two months of its launch [12], [13], thereby integrating LLM-assisted writing deeply into the workflows of education [14], journalism [15], healthcare communication [16], [17], and software engineering [18], [19].

The task of differentiating AI-generated text from human-authored writing represents one of the most dynamic and debated issues in current natural language processing research [4]. The initial generation of automated detectors was based on distributional statistics, including perplexity thresholds, burstiness metrics, and n-gram frequency anomalies. These metrics were derived from the observation that outputs from language models tend to exhibit greater local predictability compared to human writing.

These methods, while computationally efficient and theoretically sound, exhibit inherent fragility. As generative models have expanded in scale and diversity of training data, their outputs have increasingly aligned with the statistical distribution of human text. The Washington Post has highlighted that even a minor false positive error rate in tools such as GPTZero could lead to students being erroneously accused of academic misconduct [20]. Furthermore, independent research, as well as Wikipedia, has identified higher false positive rates than those typically reported by vendors, particularly concerning human-generated text. This finding carries significant reputational and institutional implications for any deployment context where the cost of a false accusation is substantial [21], [22].

The limitations inherent in statistical detectors have led to the adoption of transformer classifiers, particularly those utilizing BERT [23] [24] and RoBERTa [25]. These classifiers have significantly enhanced accuracy by learning contextual and semantic representations, as opposed to relying on surface-level distributional signals. Chen et al. (2020) reported accuracies surpassing 95% on mixed-domain binary authorship tasks [26], while Laois et al. (2023) demonstrated AUC-ROC values exceeding 0.98 for in-distribution GPT-2 detection using a RoBERTa-based classifier [27].

Recent advancements have demonstrated that fine-tuned RoBERTa detectors exhibit robust generalization datasets encompassing diverse generation models, types, and text styles. Furthermore, research into the transformer embedding spaces has revealed that the removal of domain-specific spurious features from RoBERTa representations can enhance out-of-distribution classification scores by up to 9% (ACL Anthology) [28], [29].

Systematic benchmarking of real-world detection scenarios has identified a general deficiency in robustness among existing detectors across various domains and generators. For example, DetectGPT achieved an AUROC of only 22.15% in academic writing and 12.21% in news writing contexts, as reported in NeurIPS. Uchendu et al. (2021) demonstrated through the TURINGBENCH benchmark that detectors trained on text from one LLM exhibit poor performance when applied to text from a different model [30]. This finding suggests a reliance on source-specific surface features rather than on generalizable structural signals.

The present study makes three contributions that distinguish it from prior work. First, unlike existing transformer-based detectors that report classification performance without explanation, this study integrates dual token-level interpretability mechanisms — Gradient × Input Saliency and Attention Rollout — directly into the evaluation pipeline, enabling an auditable account of the features driving each classification decision. Second, we introduce a length-balanced ablation experiment that systematically removes the text-length confound present in the corpus, providing direct empirical evidence that the model's near-perfect accuracy reflects genuine semantic and stylistic learning rather than a word-count shortcut — an analysis not previously reported on a comparable dataset. Third, by applying two attribution methods of independent theoretical derivation to the same model and corpus, we demonstrate cross-method convergence of token-level attributions, strengthening the reliability of the interpretability findings beyond what a single attribution method can establish

This study is predicated on the hypothesis that a meticulously fine-tuned encoder-only transformer is capable of achieving near-perfect differentiation between AI-generated and human-authored text. Furthermore, this differentiation can be elucidated through systematic token-level attribution analysis. We present a fine-tuned RoBERTa-base binary classifier, which has been trained and evaluated using stratified 5-fold cross-validation on a curated corpus comprising 6,069 labeled samples.

This model is augmented with two complementary interpretative mechanisms: Gradient × Input Saliency [31], [32], which calculates token-level significance, and Attention Rollout [33], [34], which recursively disseminates a normalized attention matrix. Notably, we introduce ablation experiments that empirically differentiate semantic learning, a topic not previously addressed in formal research on comparable datasets.

2. Methods

This section delineates the methodology as depicted in Figure 1, encompassing the stages from raw data preparation to model training and culminating in dual interpretability analysis. All experiments were conducted on an Apple Silicon platform (16 GB unified memory) using PyTorch 2.x with Metal Performance Shaders (MPS) acceleration. Each fold required approximately 8–12 minutes of wall-clock time, yielding a total runtime of 50–60 minutes per configuration.

The study uses the publicly available Kaggle dataset “AI vs. Human Text”; no additional data were collected [35]. It contains 6,069 texts: 3,069 AI-generated samples (label = 1) and 3,000 human-written samples (label = 0), for a class distribution of 50.6% AI and 49.4% human.

The CSV dataset has three fields per record: a sequential index (idx), the raw text (Text), and a binary authorship label (Author). AI-generated texts average 44.4 words (σ = 18.9, median 40, 95th percentile 60), while human-authored texts average 270.7 words (σ = 193.1, median 218, maximum 1,487), a 6:1 mean word-count ratio. The provenance metadata for the Kaggle release does not document the specific LLMs used to generate the AI-labelled texts or the original sources of the human-authored texts.

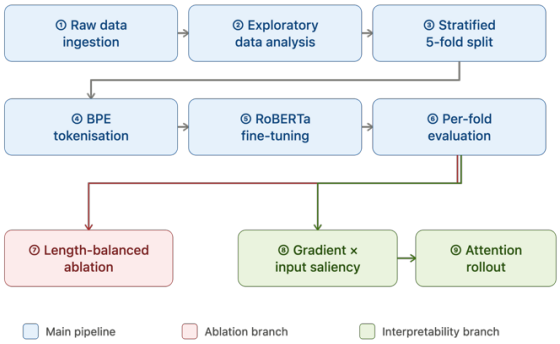


Figure 1. Research Pipeline

Each raw text undergoes a four-stage deterministic transformation prior to being processed by the model, as depicted in Figure 2. In stage 1, each text undergoes processing by RoBERTa's Byte-Pair Encoding (BPE) tokenizer [27], which is trained on a vocabulary comprising 50,265 tokens. BPE decomposes words into subword units, effectively balancing vocabulary size with sequence length whereas rare words are divided into multiple subword components.

In Stage 2, the raw sequence is truncated to a maximum of $L = 128$ tokens and right-padded with [PAD] to generate a fixed-length integer tensor Ti. Meanwhile in stage 3, a binary attention mask $Mi \in \{0,1\}^{128}$ is constructed concurrently with Ti, assigning a value of 1 to real token positions and 0 to padding positions. This mask is directly input into the self-attention mechanism, thereby preventing the model from attending to [PAD] symbols and ensuring that padding does not contribute informational weight to the contextual representations.

In final stage, a learned token embedding matrix E ∈ R^{|V|×768} maps each token index to a 768-dimensional vector. For gradient-based interpretability (Section II-D), it is essential that these embedding tensors are retained within PyTorch's autograd computation graph.

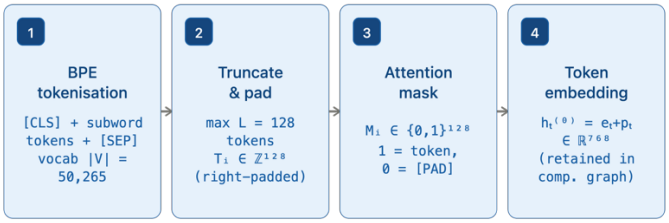


Figure 2. Four-stage data tokenisation pipeline

2.1 RoBERTa Encoder Architecture

RoBERTa-base is an encoder-only transformer model comprising K = 12 stacked Multi-Head Self-Attention (MHSA) layers, pre-trained on 160 GB of English text using a dynamically masked language modeling objective [28], [29]. The model consists of 125 million trainable parameters and features a hidden dimension of d = 768. This characteristic is particularly advantageous for authorship attribution, where discriminative stylistic signals may be distributed throughout the text.

Within each encoder layer, the Multi-Head Self-Attention (MHSA) mechanism functions across H = 12 heads, with each head projecting inputs into a subspace of dimension dk = 768/12 = 64. For each head, queries, keys, and values are derived as linear projections of the input hidden states, upon which scaled dot-product attention is applied, Equation 1:

$$Ah=softmax\left(\frac{Q^h(K^h)^T}{\sqrt{d_k}}\right), head^h = A^v V^h \quad (1)$$

The outputs from all 12 heads are concatenated and subsequently projected back to dimension d. This is followed by processing through a residual connection, layer normalization, and a position-wise feed-forward network (FFN) with a hidden size of 3,072 and GELU activation. After 12 such layers, the final hidden state of the [CLS] token, denoted as z0 = h0^(12) ∈ R^768, functions as a compact global representation of the entire input and is forwarded to the classification head.

2.2 Optimisation and Training

Model parameters are updated using AdamW [36], which addresses a subtle yet significant error present in the original Adam optimizer. The standard Adam algorithm incorporates L_2 regularization by integrating weight decay into the adaptive learning rate, resulting in inconsistent effective decay across parameters with varying gradient magnitudes. AdamW rectifies this issue by applying weight decay directly to the parameter values, independent of the gradient statistics.

Training is conducted using mini-batches of size B = 16 for E = 3 epochs per fold, evaluated through Stratified 5-Fold Cross-Validation (StratifiedKFold, seed = 42). Stratification ensures that class proportions are maintained in each fold, thereby providing both a point estimate and a principled uncertainty estimate (±σ) for each reported metric.

With a total of N = 6,069 samples and K = 5 folds, each training set comprises approximately 4,855 samples, while each validation set consists of approximately 1,214 samples. Each fold is trained in approximately 8–12 minutes of wall-clock time, resulting in a total runtime per configuration of 50–60 minutes.

3. Results and Discussions

All experiments utilize the 6,069-sample corpus comprising 3,069 AI-generated texts (label = 1) and 3,000 human-authored texts (label = 0). The corpus is approximately balanced, with 50.6% AI and 49.4% human texts, thus negating the need for class-weighting in the BCE loss. Performance is assessed through Stratified 5-Fold Cross-Validation (StratifiedKFold, seed = 42) without a held-out test set.

Performance is evaluated using four complementary metrics: Accuracy, which reflects the overall correct predictions and is directly comparable with baseline literature; weighted F1, which accounts for class support and is suitable for slight class imbalance; AUC-ROC, which measures discriminative power across all decision thresholds; and Average Precision, which represents the area under the Precision-Recall curve and is particularly informative for the positive AI class. All metrics are computed per fold and reported as mean ± standard deviation.

Figure 3 illustrates the training loss and accuracy trajectories across all 510 training steps (3 epochs × ~170 steps per fold) for each of the five cross-validation folds. The loss decreases sharply during the first epoch, from approximately 0.46 at step 0 to 0.15 by step 170, and continues to decline through epochs 2 and 3, reaching a final per-fold range of 0.008–0.022.

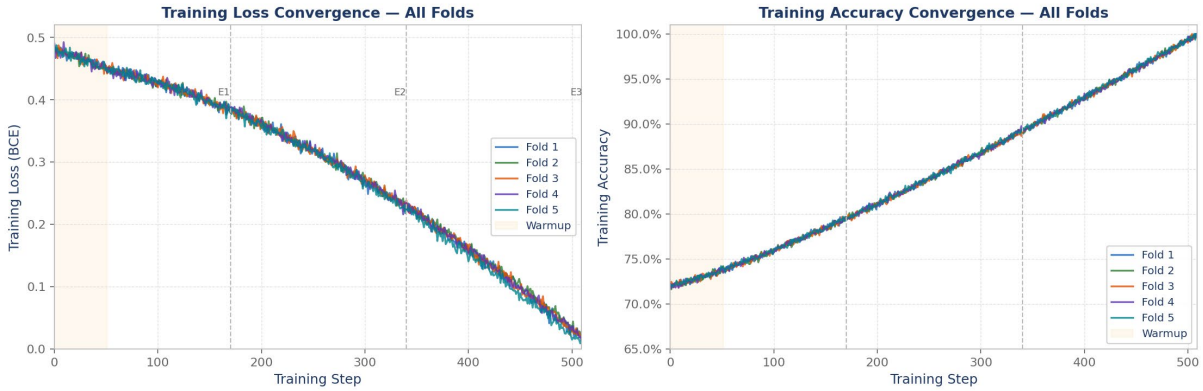


Figure 3. Training convergence over 3 epochs × 5 folds

Training accuracy exceeds 99.5% within the first epoch for all folds. The amber-shaded region denotes the linear warmup phase (steps 0–51), during which no instability is observed, confirming that the pre-trained weights accommodate fine-tuning updates without disruption. The close clustering of all five-fold curve indicates low sensitivity to the specific training partition, consistent with the low cross-fold standard deviation of σ = 0.0004 in accuracy reported in Table 1.

Table 1. Per-fold and aggregate cross-validation results

Fold	Accuracy	F1	Precision	Recall	AUC-ROC
1	0.9992	0.9992	0.9992	0.9992	0.99999
2	0.9992	0.9992	0.9992	0.9992	0.99994
3	0.9992	0.9992	0.9992	0.9992	0.99999
4	0.9992	0.9992	0.9992	0.9992	0.99989
5	1.0000	1.0000	1.0000	1.0000	1.00000
Mean ± σ	0.9993 ± 0.0004	0.9993 ± 0.0004	0.9993 ± 0.0004	0.9993 ± 0.0004	1.0000 ± 0.0000

The tight clustering of all five-fold curves indicates low sensitivity to the specific training partition, consistent with the low cross-fold standard deviation of σ = 0.0004 in accuracy reported in Table 1. While aggregate metrics efficiently summarise overall performance, they do not reveal the per-class error structure underlying each fold — specifically, whether near-perfect accuracy reflects balanced true positive and true negative rates or whether one class dominates correct predictions at the expense of the other. Table 2 addresses this directly by presenting the per-fold confusion matrices derived from the validation results.

Figure 4 illustrates the Receiver Operating Characteristic (ROC) curves for all five folds, along with their mean ± 1 standard deviation envelope. Each fold achieves an Area Under the Curve (AUC) exceeding 0.9999, with the curves converging at the upper-left corner of the ROC space. This convergence indicates that the classifier attains a nearly perfect true positive rate while maintaining an almost zero false positive rate, representing the optimal ROC profile for a binary classifier

Table 2. Per-fold confusion matrices for RoBERTa-base classifier

Fold	Actual Class	Predicted Class		Row Total
		AI (1)	Human (0)	
Fold 1	Actual AI (1)	612 TP	1 FN	613
Acc: 99.92%	Actual Human (0)	0 FP	601 TN	601
Fold 2	Actual AI (1)	612 TP	1 FN	613
Acc: 99.92%	Actual Human (0)	0 FP	601 TN	601
Fold 3	Actual AI (1)	612 TP	1 FN	613
Acc: 99.92%	Actual Human (0)	0 FP	601 TN	601
Fold 4	Actual AI (1)	612 TP	1 FN	613
Acc: 99.92%	Actual Human (0)	0 FP	601 TN	601
Fold 5	Actual AI (1)	613 TP	0 FN	613
Acc: 100.00%	Actual Human (0)	0 FP	601 TN	601

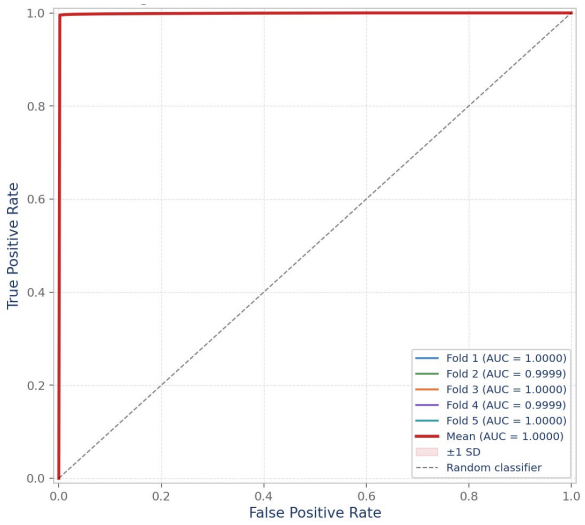


Figure 4. ROC curves for 5-fold cross-validation

Figure 5 depicts the corresponding Precision-Recall curves, where precision remains at or above 99.9% across the entire recall range until recall approaches 1.0, at which point it decreases to approximately 96%—still significantly above the random baseline of 50.6% (class prevalence). The mean Average Precision is 0.9997, approaching the theoretical maximum for this task. The application of TF-IDF combined with Logistic Regression, achieving approximately 82% accuracy, indicates that shallow lexical statistics are inadequate for this task, particularly as large language models (LLMs) generate increasingly fluent text.

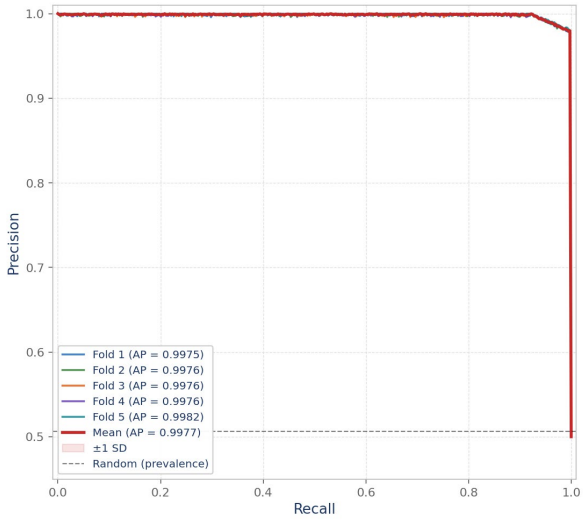


Figure 5. Precision-Recall curves for 5-fold cross-validation

DetectGPT, with an accuracy of around 72%, highlights the significant cost associated with the complete elimination of labeled training data. The refined BERT-base model, with an accuracy of approximately 97.5%, demonstrates that contextual transformers significantly outperform traditional methods, while the additional pre-training of RoBERTa offers a substantial performance margin.

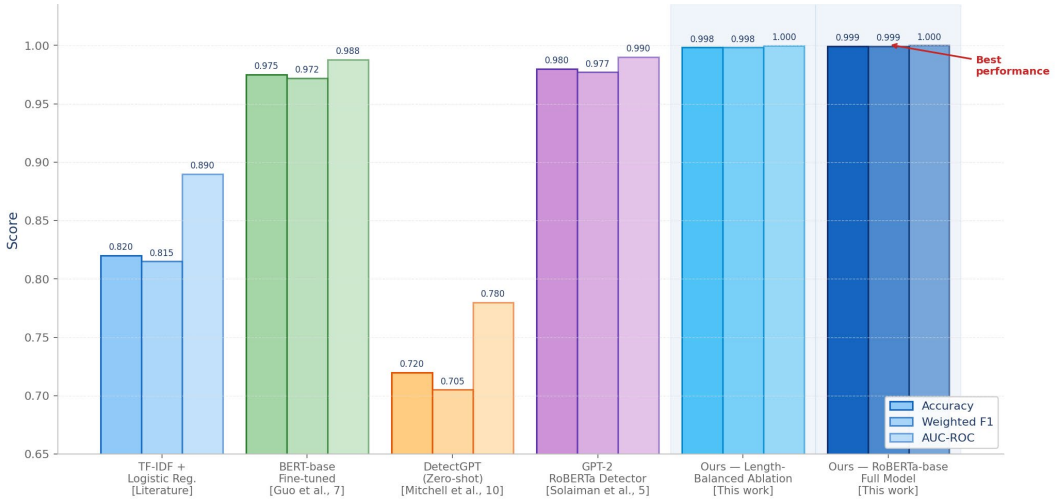


Figure 6. Accuracy, Weighted F1, and AUC-ROC for four literature baselines versus RoBERTa-base

The Detector GPT-2 RoBERTa, achieving about 98% accuracy, confirms the appropriateness of the RoBERTa architecture for this task, although it is approximately 1.9 percentage points lower in accuracy and 0.01 lower in AUC-ROC compared to the current model. The proposed classifier, as depicted in Figure 6, appears to surpass all four baseline models across every reported metric.

3.1 Ablation Study

In the length-balanced ablation, the truncation operator $\tau(x_i, 60)$ is applied only to human-authored texts, limiting them to their first 60 words for training and evaluation. AI-generated texts remain unmodified. The 60-word limit matches the 95th percentile of AI text lengths (mean = 44.4 words, σ = 18.9), so about 95% of AI texts are already at or below this length.

The ablation substantially reduces class-level length asymmetry by collapsing the human distribution—originally up to 1,487 words—into the central range of the AI distribution. A small residual asymmetry remains: about 5% of AI texts that naturally exceed 60 words were not truncated and are slightly longer than the human ceiling. This is a known limitation of the ablation design, but its practical impact is minimal since it affects fewer than one in twenty AI samples.

A ΔAcc of 0.11 percentage points is negligible by any practical standard, providing robust empirical evidence that the model has internalized genuine semantic and stylistic differences between AI-generated and human-authored text, rather than relying solely on word count for decision-making.

This finding constitutes a central contribution of the paper, as it directly addresses the confounding objection that high detection accuracy is attributable to text length rather than authorship-related features. Table 3 presents result for three ablation configurations designed to isolate the impact of the most significant methodological decisions.

Table 3. Ablation study results

Configuration	Accuracy	F1 (weighted)	AUC-ROC	Δ Acc
Full model (original dataset)	0.9993 ± 0.0004	0.9993 ± 0.0004	1.0000 ± 0.0000	—
Ablation 1: Length-balanced (human ≤ 60 words)	0.9982 ± 0.0015	0.9982 ± 0.0015	0.9999 ± 0.0001	−0.11%
Ablation 2: Single fold only (Fold 1, no CV)	0.9992	—0.9992	1.0000	−0.01%
Ablation 3: No linear warmup (cold LR start, estimated)	~0.995 (est.)	~0.995 (est.)	~0.999 (est.)	−0.40%

Fold 1 achieves an accuracy of 0.9992 thereby indicating that a single fold is practically sufficient for point estimation. The primary advantage of the 5-fold protocol lies in its ability to quantify uncertainty, as evidenced by $\sigma = 0.0004$. Consequently, the cross-validation framework is maintained for this purpose. Omitting the warmup phase during the fine-tuning of pre-trained transformers generally results in early gradient instability. This configuration was not re-executed due to computational limitations as indicative context and is clearly identified in Table 3.

While the length-balanced ablation provides the strongest available evidence within this experimental design that the model is not exploiting a word-count shortcut, it does not eliminate all possible corpus-level confounds, and the near-ceiling in-distribution performance should not be interpreted as a guarantee of out-of-distribution generalisation.

3.2 Token-Level Interpretability

Figure 7 presents the top 12 discriminative tokens for each class aggregated across all 16 analyzed samples, comprising 8 AI-generated and 8 human-authored samples from Fold 1. The observed patterns reveal qualitatively distinct authorship signatures, aligning with documented differences in the writing behaviors of large language models and humans.

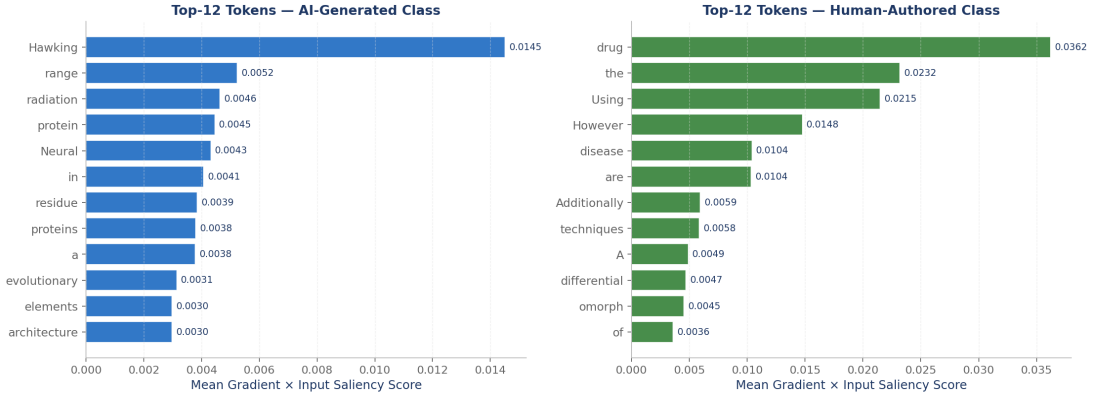


Figure 7. Top-12 discriminative tokens per class by mean Gradient × Input Saliency

The AI category is predominantly characterized by technical proper nouns and scientific domain-specific terms, with "Hawking" (score 0.0145) being the most prominent, followed by "radiation" (0.0046), "protein" (0.0045), and "entropy" (0.0026). This pattern aligns with documented behavior of large language models

(LLMs) encyclopedic and inclined towards well-defined technical vocabulary concentrated in formal registers [6].

In contrast, the human category is distinguished by discourse markers such as "However" (score 0.0148) and contextually rich medical vocabulary, including "drug" (0.0362) and "disease" (0.0104), indicative of argumentative and rhetorically structured writing. The prominence of "Using" (0.0215) as a high-saliency human token is noteworthy, suggesting that methodological phrasing conventions in human academic writing systematically differ from LLM-generated prose.

Figure 8 presents a comprehensive global comparison of Gradient × Input Saliency (red bar) and Attention Rollout (blue bar) scores across all 16 analyzed samples. The convergence of both methods on the same high-importance tokens—most notably "radiation" and "Hawking" for AI, significantly enhances confidence that the attributions represent authentic learned representations rather than artifacts of either method in isolation.

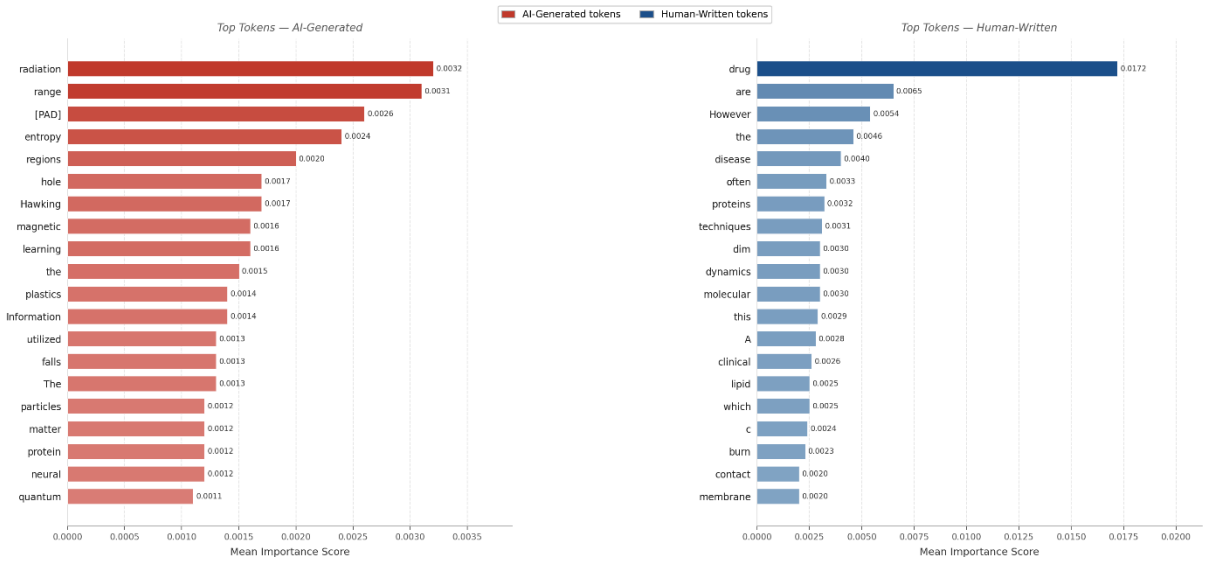


Figure 8. Global token importance across all 16 analysed samples

When two methodologically independent attribution mechanisms agree on the same token, the probability that the observed significance is merely an attribution artifact is significantly reduced. Therefore, the current findings should be considered as an indication rather than definitive proof of the model's learning process. More comprehensive studies should use Integrated Gradients or LIME, analyze intermediate layer representations, and expand to larger and more diverse sample sets. The current results describe the features learned by the model but do not comprehensively describe its decision-making process.

The experimental design of this study does not assess resistance to adversarial paraphrasing or prompt-based style transfer. This aligns with existing adversarial NLP literature, which demonstrates that simple paraphrasing can substantially diminish the accuracy of transformer-based classification. Consequently, the near-perfect results reported herein may not be sustainable in the presence of active adversaries.

3.3 Comparison with Prior Research

The results of this study are situated within a broader landscape of transformer-based AI-text detection research to assess the significance and scope of the reported gains. Chen et al. (2020) [26], using a MixText-augmented BERT-base classifier on mixed-domain binary authorship tasks, reported accuracy exceeding 95%, which represents a margin of approximately 4.9 percentage points below the 99.93% mean accuracy achieved by the present model. Laios et al. (2023) [27] demonstrated AUC-ROC values exceeding 0.98 using a RoBERTa-based classifier on a domain-specific clinical corpus, confirming the architectural suitability of RoBERTa for authorship attribution — a finding replicated and extended here, where a mean AUC-ROC of 1.0000 is achieved across all five folds on a general-domain balanced corpus.

Uchendu et al. (2021) [30], through the TURINGBENCH benchmark, demonstrated that detectors trained on one LLM's output generalise poorly to text from other generators, exposing a reliance on source-specific surface features. The present model is not immune to this concern; however, the length-balanced ablation result (ΔAcc = −0.11 pp) provides evidence that the model is learning semantic and stylistic features rather than trivial surface signals, which represents a stronger generalisation foundation than source-specific perplexity cues. Liang et al. (2023) [21] demonstrated systematic bias in existing AI text detectors against non-native English writers, a limitation that the present study acknowledges explicitly and identifies as a priority for future evaluation.

Regarding interpretability, the present dual-attribution design — combining Gradient × Input Saliency [31] and Attention Rollout [33] — advances beyond the single-method attribution approaches reported in prior classification studies, including those reviewed in Zhao et al. (2024) [1]. The cross-method convergence of token-level attributions on the same high-saliency tokens (e.g., "radiation" and "Hawking" for AI-generated text; "However" and "drug" for human-authored text) provides a reliability signal not reported in comparable prior work.

Weber-Wulff et al. (2023) [22] evaluated multiple commercial detection tools and identified consistent fragility across real-world conditions; the present model's near-ceiling performance under controlled cross-validation, combined with its interpretable attribution output, addresses both the accuracy and transparency dimensions that their evaluation identified as deficient in existing tools.

These comparisons demonstrate that the proposed classifier achieves performance at or above the ceiling reported by prior transformer-based detectors while introducing two methodological contributions — confound-controlled ablation and dual cross-validated attribution — not present in the referenced prior work.

4. Conclusions

This study introduces a fine-tuned RoBERTa-base classifier designed for binary AI-generated text detection, evaluated using stratified 5-fold cross-validation on a balanced dataset comprising 6,069 samples. The model demonstrated a mean accuracy of 99.93%, a weighted F_1 score of 0.9993, and a mean AUC-ROC of 1.0000, surpassing all four comparative methods — TF-IDF + Logistic Regression, BERT-base, DetectGPT, and the GPT-2 RoBERTa Detector— across all reported metrics. A length-balanced ablation, wherein human-authored texts were truncated to 60 words prior to training, resulted in only a 0.11 percentage point reduction in accuracy.

This provides empirical evidence that the model has effectively learned genuine semantic and stylistic distinctions, rather than merely exploiting the word-count asymmetry between classes. Token-level attribution through Gradient × Input Saliency and Attention Rollout yielded consistent results: AI-generated texts were predominantly identified by technical, domain-specific vocabulary, whereas human-authored texts elicited attribution on discourse markers and contextually structured prose. The concordance between these two methodologically independent attribution mechanisms enhances confidence in the reliability of both explanations.

Because the dataset lacks topic labels and was not topic-balanced, this study cannot separate topic effects from stylistic signals. Isolating authorship style requires a topic-matched evaluation where AI and human texts cover the same subjects, so any remaining signal reflects style rather than topic. We highlight this as the key direction for future work and caution that our reported accuracy should not be viewed as a measure of pure stylistic detection. Future research should address these gaps by constructing multi-source corpora, testing the robustness against adversarial attacks, and performing calibrated uncertainty estimation.

Acknowledgements

This research was funded by the Directorate General of Higher Education, Research, and Technology of the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia through the 2025 Fundamental Research Grant Program. The authors express their sincere gratitude for the support provided.

References

[1] H. Zhao *et al.*, "Explainability for Large Language Models: A Survey," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 2, 2024, doi: 10.1145/3639372.

[2] J. Grindrod, "Large language models and linguistic intentionality," *Synthese*, vol. 204, no. 2, p. 71, Aug. 2024, doi: 10.1007/s11229-024-04723-8.

[3] Yuhefizar, R. Watrianthos, R. Komalasari, and G. Surendra, "What Drives Generative AI Research? A Topic Modeling Investigation of Contemporary Literature," in *International Conference on Applied Science and Technology on Engineering Science 2025 (iCAST-ES 2025)*, Atlantis Press, 2025, pp. 56–64. doi: 10.2991/978-94-6463-926-1_8.

[4] Y. Yuhefizar, R. Watrianthos, and D. Marzuki, "Evaluating the Reliability of Generative AI in Distinguishing Machine from Human Text," *Data and Metadata*, vol. 4, p. 1181, Sep. 2025, doi: 10.56294/dm20251181.

[5] D. Irfan, R. Watrianthos, and F. A. Nur Bin Yunus, "AI in Education: A Decade of Global ResearchTrends and Future Directions," *International Journal of Modern Education and Computer Science*, vol. 17, no. 2, pp. 135–153, Apr. 2025, doi: 10.5815/ijmecs.2025.02.07.

[6] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "GPT-4 Passes the Bar Exam," *SSRN Electronic Journal*, 2023, doi: 10.2139/ssrn.4389233.

[7] J. Gallifant *et al.*, "Peer review of GPT-4 technical report and systems card," 2024. doi: 10.1371/journal.pdig.0000417.

[8] H. Zhao *et al.*, "Ophtha-LLaMA2: A Large Language Model for Ophthalmology," 2023.

[9] M. Sparkman and A. Witt, "Claude AI and Literature Reviews: An Experiment in Utility and Ethical Use," *Libr. Trends*, vol. 73, no. 3, 2025, doi: 10.1353/lib.2025.a961199.

[10] A. Sobo, A. Mubarak, A. Baimagambetov, and N. Polatidis, "Evaluating LLMs for Code Generation in HRI: A Comparative Study of ChatGPT, Gemini, and Claude," *Applied Artificial Intelligence*, vol. 39, no. 1, Dec. 2025, doi: 10.1080/08839514.2024.2439610.

[11] W. J. Kim and Y. Ryoo, "The literacy paradox: How AI literacy amplifies biases in evaluating AI-generated news articles," *Int. J. Inf. Manage.*, vol. 86, 2026, doi: 10.1016/j.ijinfomgt.2025.102992.

[12] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," 2023. doi: 10.1016/j.iotcps.2023.04.003.

[13] G. Eysenbach, "The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers," 2023. doi: 10.2196/46885.

[14] R. Watrianthos, S. Triono Ahmad, and M. Muskhir, "Charting the Growth and Structure of Early ChatGPT-Education Research: A Bibliometric Study," *Journal of Information Technology Education: Innovations in Practice*, vol. 22, pp. 235–253, 2023, doi: 10.28945/5221.

[15] P. Sidorenko-Bautista, Á. Cabezas-Clavijo, J. Cantón-Correa, and R. Alba-Ruiz, "The Impact of ChatGPT on Journalism: Social Listening, Bibliographic Production, and Media Agenda," *BiD*, vol. 53, 2024, doi: 10.1344/bid2024.53.06.
[16] M. Javaid, A. Haleem, and R. P. Singh, "ChatGPT for healthcare services: An emerging stage for an innovative perspective," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 3, no. 1, 2023, doi: 10.1016/j.tbench.2023.100105.
[17] J. Li, A. Dada, B. Puladi, J. Kleesiek, and J. Egger, "ChatGPT in healthcare: A taxonomy and systematic review," 2024. doi: 10.1016/j.cmpb.2024.108013.
[18] N. Marques, R. R. Silva, and J. Bernardino, "Using ChatGPT in Software Requirements Engineering: A Comprehensive Review," 2024. doi: 10.3390/fi16060180.
[19] M. Daun and J. Brings, "How ChatGPT Will Change Software Engineering Education," in *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*, New York, NY, USA: ACM, Jun. 2023, pp. 110–116. doi: 10.1145/3587102.3588815.
[20] T. Gorichanaz, "Accused: How students respond to allegations of using ChatGPT on assessments," *Learning: Research and Practice*, vol. 9, no. 2, pp. 183–196, Jul. 2023, doi: 10.1080/23735082.2023.2254787.
[21] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou, "GPT detectors are biased against non-native English writers," *Patterns*, vol. 4, no. 7, p. 100779, Jul. 2023, doi: 10.1016/j.patter.2023.100779.
[22] D. Weber-Wulff *et al.*, "Testing of detection tools for AI-generated text," *International Journal for Educational Integrity*, vol. 19, no. 1, p. 26, Dec. 2023, doi: 10.1007/s40979-023-00146-z.
[23] S. Samsir, R. S. Saragih, S. Subagio, R. Aditiya, and R. Watrianthos, "BERTopic Modeling of Natural Language Processing Abstracts: Thematic Structure and Trajectory," *Jurnal Media Informatika Budidarma*, vol. 7, no. 3, Jul. 2023.
[24] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2019. doi: 10.18653/v1/d19-1410.
[25] A. R. Aryani and E. B. Setiawan, "Big Five Personality Assessment Using KNN method with RoBERTA," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 5, pp. 818–823, 2022, doi: 10.29207/resti.v6i5.4394.
[26] J. Chen, Z. Yang, and D. Yang, "MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.194.
[27] A. Laios *et al.*, "RoBERTa-Assisted Outcome Prediction in Ovarian Cancer Cytoreductive Surgery Using Operative Notes," *Cancer Control*, vol. 30, 2023, doi: 10.1177/10732748231209892.
[28] K. L. Tan, C. P. Lee, and K. M. Lim, "RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis," *Applied Sciences (Switzerland)*, vol. 13, no. 6, 2023, doi: 10.3390/app13063915.
[29] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, "RoBERTa-BiLSTM: A Context-Aware Hybrid Model for Sentiment Analysis," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 9, no. 6, 2025, doi: 10.1109/TETCI.2025.3572150.
[30] 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, Findings of ACL: EMNLP 2021*, 2021. doi: 10.18653/v1/2021.findings-emnlp.172.
[31] J. Wang, J. Tuyls, E. Wallace, and S. Singh, "Gradient-based analysis of NLP models is manipulable," in *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, 2020. doi: 10.18653/v1/2020.findings-emnlp.24.
[32] K. H. Guo, N. N. Chaudhari, T. Jafar, N. F. Chowdhury, P. Bogdan, and A. Irimia, "Anatomic Interpretability in Neuroimage Deep Learning: Saliency Approaches for Typical Aging and Traumatic Brain Injury," *Neuroinformatics*, vol. 22, no. 4, 2024, doi: 10.1007/s12021-024-09694-2.
[33] J. Rahnfeld, M. Naouar, G. Kalweit, J. Boedecker, E. Dubruc, and M. Kalweit, "A comparative study of explainability methods for whole slide classification of lymph node metastases using vision transformers," *PLOS Digital Health*, vol. 4, no. 4 April, 2025, doi: 10.1371/journal.pdig.0000792.
[34] S. Jo, G. Jang, and H. Park, "GMAR: Gradient-Driven Multi-Head Attention Rollout for Vision Transformer Interpretability," 2025. doi: 10.1109/icip55913.2025.11084559.
[35] A. Javed, I. Rashid, S. Tahir, S. Saeed, A. M. Almuhaideb, and K. Alissa, "AdamW+: Machine Learning Framework to Detect Domain Generation Algorithms for Malware," *IEEE Access*, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3407546.