PLD-4: A Multi-Task Framework for Detecting and Attributing LLM-Generated Paraphrases

Anonymous ACL submission

Abstract

LLMs make distinguishing human from machine text challenging, particularly via paraphrasing used for evasion, impacting academic integrity, IP, and misinformation. We introduce the novel Paraphrase-based LLM Detection Framework (PLD-4), formalizing four tasks to evaluate detection in nuanced scenarios, including identifying layered AI text. Using MRPC and HLPC datasets, we employ a dual approach with feature-based and transformer models (XGBoost, DeBERTa-v3, RoBERTa). While achieving high accuracy on tasks like Sentence Pair Paraphrase Source Detection (XGBoost 96%) and Single Sentence Authorship Attribution (RoBERTa 93.9%), distinguishing original vs. paraphrased LLM output proved significantly challenging (RoBERTa 83.28%), highlighting limitations in detecting layered AI generation. PLD-4 provides a critical foundation for developing more robust detection techniques.

1 Introduction

The rapid advancements in Large Language Models (LLMs) have fundamentally transformed natural language processing, enabling the generation of text that often rivals human content in fluency and coherence (Wu et al., 2025; Brown et al., 2020). This remarkable progress, while offering substantial benefits, has introduced a significant challenge: reliably distinguishing between human-authored and machine-generated text (Huang et al., 2024; Fariello et al., 2024). The increasing proliferation and seamless integration of high-quality LLM-generated content across digital mediums raise critical concerns regarding academic integrity (e.g., plagiarism), intellectual property (e.g., paraphrased code), and the fight against misinformation (Hunt et al., 2019; Park et al., 2025; Goldstein et al., 2023). Consequently, the imperative need for effective and robust AI-generated text detection mechanisms has become more pronounced than ever before (Huang et al., 2024), especially as human-based detection is often unreliable (Yu et al., 2024).

A particularly complex facet of this challenge lies in the detection of content that has been paraphrased by LLMs (Park et al., 2025). LLMs possess sophisticated capabilities to produce paraphrased iterations of existing text, often mimicking human writing styles (Wei et al., 2023), potentially with the intent to obscure the original source or circumvent detection (Park et al., 2025). While paraphrase detection has a well-established history (Park et al., 2025), the emergence of LLMs introduces new complexities (Tripto et al., 2023). Traditional methodologies, often prioritizing semantic similarity (Wu et al., 2025), may prove inadequate as LLM paraphrases maintain semantic equivalence but exhibit distinct stylistic characteristics that evade detection (Tripto et al., 2023). Indeed, a significant limitation of many current AI text detection systems (Jawahar et al., 2020) is their vulnerability to such paraphrase attacks (Weber-Wulff et al., 2023), with notable performance drops observed in studies (Krishna et al., 2023). This underscores a critical gap and the urgent need for more robust approaches.

To address this underexplored challenge, we introduce the Paraphrase-based LLM Detection framework (PLD-4 framework). This framework systematically defines and facilitates the evaluation of detection mechanisms across four core subtasks representing nuanced real-world scenarios with varying difficulty and contextual information, thereby filling a crucial gap in research overlooking paraphrase-driven evasion and fine-grained authorship attribution.

Building upon this framework, we develop and rigorously evaluate a dual-pronged detection approach tailored to address these tasks. This approach leverages both extensive feature engineering to train interpretable models like XGBoost (Chen and Guestrin, 2016), capturing linguistic and stylistic signals, and fine-tuned state-of-the-art transformer architectures such as DeBERTa-v3 (He et al., 2021). This dual strategy allows for a comprehensive evaluation considering both performance and explainability. Our initial experimental results on the PLD-framework demonstrate the effectiveness of this approach, with the feature-based pipeline achieving 97% accuracy on Task 3, and the fine-tuned DeBERTa-v3 model attaining 92.7% accuracy on the same task.

The remainder of this paper is structured as follows: Section 2 reviews related work. Section 3 describes our paraphrase-based LLM detection framework. Section 4 outlines the original data source and dataset construction. Section 5 presents the experimental setup. Section 6 reports empirical results and feature analysis. Section 7 discusses key insights derived from the results. Section 8 concludes the paper with a summary of contributions and directions for future research. Section 9 highlights the limitations of the current study.

2 Background

Distinguishing AI-generated paraphrases from human-written text lies at the intersection of paraphrase detection and AI-generated text detection. We briefly review both areas and highlight the gap addressed by the PLD-4 framework.

2.1 Paraphrase Detection

Paraphrase detection determines whether two texts express the same meaning, regardless of wording. Applications include plagiarism detection, question answering, and summarization (Bhagat and Hovy, 2013). Early work relied on n-gram overlap, Jaccard similarity, and parse tree comparison (Madnani et al., 2012; Qiu et al., 2006; Das and Smith, 2009). These approaches struggled with semantically similar texts with low lexical overlap. Traditional machine learning methods used engineered features (e.g., WordNet, syntax, semantics) and classifiers like SVMs, often evaluated on datasets like MRPC (Ji and Eisenstein, 2013; Filice et al., 2015). Neural models such as Siamese LSTMs (Mueller and Thyagarajan, 2016), pretrained embeddings (Word2Vec, GloVe (Mikolov et al., 2013; Pennington et al., 2014)), and transformer-based architectures like BERT, RoBERTa, and DeBERTa (Devlin et al., 2019; Liu et al., 2019; He et al., 2021) have achieved state-of-the-art performance on datasets like MRPC and QQP (Wang et al., 2018). However, they typically focus on semantic equivalence, not the source of the paraphrase—an essential distinction in Tasks 1 and 2 of PLD-4.

2.2 AI-Generated Text Detection

The task of identifying text generated by AI models has gained prominence with the rise of increasingly sophisticated LLMs. The goal is to determine whether a given piece of text was authored by a human or a machine. Early detectors used linguistic cues such as n-gram frequencies, perplexity, POS distributions, and readability metrics (Gehrmann et al., 2019). These methods often struggled to generalize to new LLMs. More recent work fine-tunes transformer-based models (e.g., BERT, RoBERTa) for binary classification (Solaiman et al., 2019; Zellers et al., 2020). Zero-shot detection approaches, such as DetectGPT (Mitchell et al., 2023) and Fast-DetectGPT (Bao et al., 2024), aim to identify AI-generated text without task-specific training. Watermarking methods embed detectable signals within generated text to support attribution (Kirchenbauer et al., 2024). Our Task 3 aligns with this domain.

Despite progress, AI text detection faces significant challenges. Detectors often exhibit performance degradation when applied to texts from LLMs not seen during training or when confronted with out-of-domain content (Weber-Wulff et al., 2023). A major vulnerability, highlighted in our introduction, is the susceptibility of detectors to adversarial attacks, particularly through paraphrasing. As demonstrated by prior work (Krishna et al., 2023; Chakraborty et al., 2023; Sadasivan et al., 2025), paraphrasing LLM-generated text, especially using another LLM like DIPPER, can drastically reduce the effectiveness of existing detectors. Human editing of LLM outputs further complicates detection.

2.3 Bridging the Gap: Detecting Paraphrased AI-Generated Text

While prior work explores paraphrase detection and AI-authorship detection independently, few address the intersection—identifying paraphrased LLM outputs. Existing detectors assume direct AI output, and standard paraphrase tasks ignore authorship. Our PLD-4 framework aims to fill this gap by comprising four unique subtasks, each carefully designed to represent a specific real-world detection scenario. These scenarios are characterized by differing levels of contextual information and varying degrees of inherent difficulty, allowing for a thorough evaluation of LLM paraphrase detection methods across several applications.

3 PLD-4: The Paraphrase-based LLM Detection Framework

To address the growing challenge of detecting LLM-generated paraphrases, this paper introduces the Paraphrase-based LLM Detection Framework (PLD-4), as illustrated in Figure 1. It offers a structured approach for evaluating detection methods across varied scenarios by defining four subtasks that reflect real-world conditions. These subtasks differ in contextual information and difficulty, enabling a nuanced assessment of detection performance in diverse settings.

3.1 Task 1: Sentence Pair Paraphrase Detection

Definition: Given two input sentences, determine if the second sentence in the pair is paraphrased by Human or a Large Language Model (LLM). **Scenario**: This task assumes access to both the original and paraphrased sentences, common in legal, patent, and academic contexts. The focus is on identifying the paraphraser human or LLM by comparing stylistic and linguistic features. **Example**: Detecting AI-generated paraphrases in legal or patent documents that may obscure prior art.

3.2 Task 2: Single Sentence Paraphraser Detection

Definition: Given a single input sentence known to be a paraphrase, determine if the paraphrasing was generated by an LLM or by human. **Scenario**: This task models situations where the original source is unavailable and aims to attribute authorship based solely on the paraphrase's linguistic and stylistic features. It is relevant in contexts such as plagiarism detection or content moderation, where only the paraphrased text is accessible. Unlike traditional paraphrase identification focused on semantic similarity, this task requires detecting subtle cues—such as vocabulary diversity, syntactic complexity, or LLM-specific linguistic patterns—that distinguish human and AI-generated paraphrases. **Example**: Spotting AI-generated content used in "content spinning" for SEO or bulk article generation.

3.3 Task 3: Single Sentence Authorship Attribution

Definition: Determine whether a given sentence-whether original or paraphrased-was authored by a human or generated by an LLM. **Scenario**: This represents the general, often challenging task of AI-generated text detection at the sentence level. It is broadly applicable in contexts where the origin of any given piece of text needs to be ascertained without prior knowledge of its nature (original vs. paraphrase). **Example**: Detecting AI-written content in online reviews, news, or social media to combat misinformation.

3.4 Task 4: Single Sentence AI Authorship Attribution

Definition: Given a machine-generated sentence, determine whether it was directly generated from a LLM or is a paraphrased version of an LLM output produced by another LLM. **Scenario**: This task isolates the challenge of detecting layered or iterative AI generation. It's relevant for understanding how LLMs modifies their own output and for identifying content that has been specifically manipulated by paraphrasing tools applied to pre-existing AI-generated text. **Example**: Analyzing stylistic shifts and information loss from paraphrasing to understand AI-to-AI transformations.

4 Dataset

We evaluated the four tasks defined in the PLD-4 framework using two benchmark datasets: the Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) and the Human-LLM Paraphrase Collection (HLPC) (Lau and Zubiaga, 2024). For our paraphrase source identification task, we created the MRPC-Source Identification (MRPC-SI) dataset from MRPC, containing 5,801 sentence pairs labeled by paraphraser origin (Human or LLM). Human paraphrases (for original MRPC pairs) were retained, while LLM paraphrases (for original non-paraphrase pairs) were generated using GPT-4o, resulting in data formatted as (`original sentence`, `paraphrased_sentence`, `label`). The HLPC dataset is designed to capture a diverse range of human and LLM paraphrases. It includes human paraphrases (H-PP) of human-authored documents (H-DOC) from MRPC and other datasets, alongside LLM-generated paraphrases (LLM-PP) produced through iterative paraphrasing of GPT2-XL

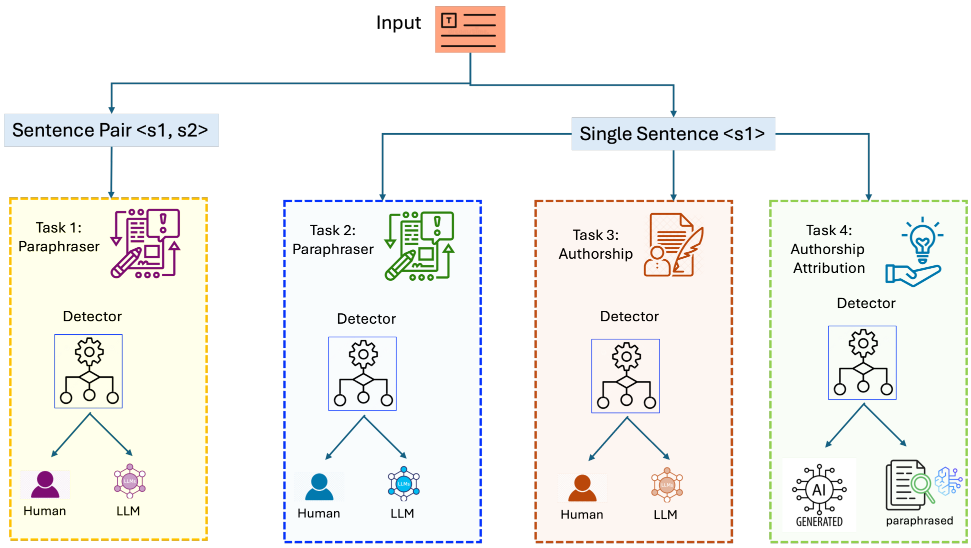


Figure 1: Paraphrase-based LLM Detection Framework (PLD-4)

and OPT-1.3B generated texts (LLM-DOC) using DIPPER and BART. While HLPC contains 600 base documents across these categories, our study focused on GPT2-XL generated base documents and their non-watermarked paraphrases, in order to analyze linguistic cues indicative of AI authorship.

5 Experiment

This section presents the empirical evaluation conducted to assess the effectiveness of various computational approaches across the four PLD-4 tasks. For each task, we examined both traditional feature-based machine learning methods—specifically using XGBoost classifiers—and deep learning models based on transformer architectures, namely RoBERTa and DeBERTa-v3. The choice of which model's results to emphasize for each task was informed by preliminary performance assessments, alignment between model characteristics and task requirements, and considerations such as interpretability and susceptibility to overfitting.

5.1 Experiment Setup: Task 1

For **Task 1** (Sentence Pair Paraphrase Source Detection), we primarily focused on a feature-based XGBoost classifier. This choice was motivated by its strong performance in sentence-pair classification and its interpretability—crucial for analyzing linguistic cues that differentiate human and LLM-generated paraphrases when the original sentence is provided. Although we initially experimented with deep learning models, they showed signs of overfitting due to the limited data and paired input format.

XGBoost was trained on a set of handcrafted lexical, syntactic, and semantic features extracted from each sentence pair. To address class imbalance, we applied SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002) to the training portion of each fold during cross-validation. We evaluated Task 1 using: **Adapted MRPC:** 5,801 sentence pairs, with 3,900 human and 1,901 LLM paraphrases; and **HLPC:** 1,800 sentence pairs, including 600 human and 1,200 LLM paraphrases.

To examine the impact of paraphrasing depth on detectability, we conducted experiments using both **once-paraphrased** and **five-times-paraphrased** LLM outputs from HLPC. This allowed us to assess whether deeper paraphrasing reduces detectable signals and increases similarity to human-written paraphrases.

5.1.1 Feature Engineering

To support the XGBoost classifier in Task 1, we engineered **39 features** designed to capture lexical, syntactic, semantic, and stylistic differences between each sentence pair (`sentence1`, `paraphrased_sentence`). These features include sentence-level statistics (e.g., length, readability), overlap measures, POS and tense distributions, named entity counts, and a semantic similarity score based on RoBERTa embeddings. Detailed categories and representative examples are shown in Table 1.

Table 1: Overview of Engineered Features for Task 1

Category	**Count**	**Example Features**
Basic Properties	7	Word count, SMOG index and Flesch Reading Ease for each sentence, length difference
Lexical Overlap	3	Unigram Jaccard similarity, bigram overlap, trigram overlap
Readability & Diversity	4	Gunning Fog index, lexical diversity for each sentence
Syntactic Complexity	2	Dependency parse tree depth for each sentence
Sentiment Analysis	2	Sentiment polarity score for each sentence (TextBlob polarity)
Part-of-Speech Ratios	8	Fraction of nouns, verbs, adjectives, and adverbs in each sentence
Verb Tense Ratios	6	Past tense ratio, present tense ratio, modal verb ratio (per sentence)
Named Entity Counts	6	Count of PERSON, ORG, and LOC entities in each sentence
Semantic Similarity	1	Cosine similarity of sentence embeddings from RoBERTa
Total	39	

5.2 Experiment Setup: Task 2, 3 and 4

For the single-sentence classification challenges presented in Task 2 (Single Sentence LLM Paraphrase Detection), Task 3 (Single Sentence Authorship Attribution), and Task 4 (Original LLM vs. Paraphrase Model Output), which require capturing subtle intrinsic linguistic cues, we primarily utilized fine-tuned RoBERTa and DeBERTa-v3 transformer models. These models significantly outperformed our feature-based methods on these nuanced single-sentence tasks; for instance, preliminary feature-based experiments on Task 2 yielded considerably lower performance (76% Accuracy) compared to transformers. The transformer models were fine-tuned for each task using standard settings, including cross-entropy loss, the AdamW optimizer, early stopping based on validation ROC-AUC, and 5-fold cross-validation. Evaluation metrics included Accuracy, F1 score, ROC-AUC, and TPR@1%FPR. As a baseline, we also report the performance of OpenAI's RoBERTa-based classifier (Solaiman et al., 2019).

Evaluation for these tasks was based on the Human-LLM Paraphrase Collection (HLPC) dataset. Task-specific datasets, each comprising 1,800 instances, were independently constructed from HLPC content for training, validation, and testing. These datasets were formulated to align with the specific classification goals of each task: Task 2's dataset included instances of Human-paraphrased and LLM-paraphrased sentences, aiming to classify the source of the paraphrase; Task 3's dataset comprised original Human-authored sentences and LLM-generated paraphrases, focusing on general authorship attribution; and Task 4 evaluated the ability, given a machine-generated sentence, to determine whether it was originally generated by an LLM or is a paraphrased version produced by an LLM, with its dataset consisting of original LLM outputs and sentences generated by paraphrasing those LLM outputs using models like BART and Dipper.

5.2.1 Feature Engineering

For the preliminary XGBoost experiments conducted on the single-sentence Tasks 2, 3, and 4 (with results briefly referenced in the Task 2 performance comparison), a reduced set of 17 features was employed. These features were derived by applying relevant single-sentence metrics—such as length, readability, lexical diversity, dependency depth, sentiment, part-of-speech ratios, verb tense ratios, and named entity counts—to each input sentence individually, without using any comparative or relational features.

6 Result Analysis

This section presents the empirical results of the experiments described in Section 5. We report the performance of both the traditional machine learning approach (XGBoost) and the deep learning models on the PLD-4 tasks. Additionally, we provide findings from the feature analysis of the XGBoost model, including feature importance rankings and statistical comparisons between human- and LLM-generated paraphrases for Task 1.

6.1 Result: Task 1

As presented in Table 2, the XGBoost model demonstrated strong performance on the adapted MRPC dataset, achieving an accuracy of 0.96 and macro-averaged precision, recall, and F1-score of 0.95. It reached an AUC-ROC of 0.9876 on the test set, with 5-fold cross-validation confirming robustness (mean AUC-ROC = 0.9889 ± 0.0021). These results highlight the discriminative power of the engineered linguistic features. For the HLPC dataset, we used the paraphrased outputs of LLM-

Table 2: Task 1 Performance Results (MRPC and HLPC Datasets)

Metric	MRPC (XGBoost)	HLPC (1st Para)	HLPC (5th Para)
Accuracy	0.96	0.88	0.88
Precision	0.95	0.87	0.86
Recall	0.95	0.85	0.86
F1-score	0.95	0.86	0.86
AUC-ROC	0.9876	0.9504	0.9517
CV AUC-ROC	0.9889 ± 0.0021	0.9472 ± 0.0074	0.9511 ± 0.0087
TPR@1%FPR	0.6927	0.4917	0.4917

Note: All results are reported on the test set partition, SMOTE was applied after the train/test split for the XGBoost model. The reported AUC-ROC scores represent the mean ± standard deviation from 5-fold cross-validation.

DOC (GPT2-XL) as the paraphrased sentences (1st para and 5th Para).

Figure 2 presents the corresponding ROC curve, illustrating the trade-off between true positive rate (TPR) and false positive rate (FPR) across decision thresholds. The area under the curve (AUC) reaches 0.9876, indicating excellent discriminative performance. Notably, at the low-FPR region ($FPR \leq 1\%$), the model achieves a TPR of 0.6927 using a threshold of 0.9349, demonstrating practical reliability in high-precision scenarios.

Distribution of true positives and true negatives for both datasets are provided in Appendix A (Figure 6, Figure 7 and Figure 8).

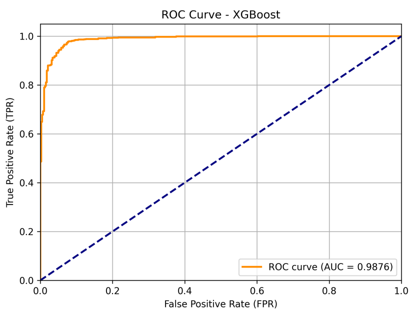


Figure 2: Task 1: ROC curve for the XGBoost model on the adapted MRPC.

6.1.1 Feature Analysis

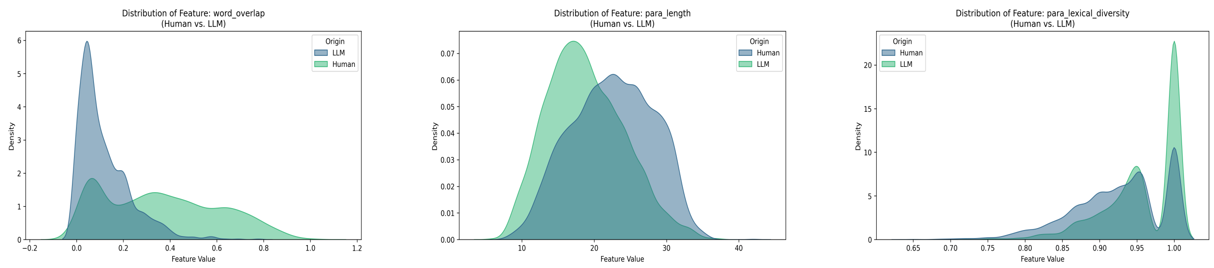
To understand which linguistic characteristics drive the XGBoost model's performance and how human/LLM paraphrases differ, we analyzed the engineered features as follows:

Feature Importance Figures 4 and 5 illustrate feature importance for the XGBoost model trained on adapted MRPC. Figure 4, a SHAP summary plot, shows both the magnitude and direction of each feature's impact, ranking features by mean absolute SHAP value and using color to indicate feature value (red = high, blue = low). This highlights key linguistic cues for distinguishing human and LLM paraphrases.

Figure 5 presents traditional feature importance based on XGBoost's gain metric, ranking the top 10 features by their contribution to decision splits, offering a global view of feature utility. Despite different theoretical bases, both SHAP and gain-based methods showed broadly consistent results in identifying the most impactful features. SHAP additionally provides local interpretability and the directionality of feature effects. For HLPC results, see Appendix A (Figures 13, 14).

Feature Distribution Comparison Complementing the model-based feature importance, we conducted a direct statistical comparison of the feature distributions distinguishing human- from LLM-generated paraphrases. Welch's t-tests were performed across all 39 engineered features, revealing statistically significant differences ($p<0.05$) between the Human and LLM groups for **21** of these features. Full details for these significant features, including t-statistics, p-values, and effect sizes, are presented in Appendix A (Table 4).

To visually examine the most pronounced differences, we highlight the distributions of the three features with the largest absolute values of Cohen's d: word overlap, paraphrased sentence length, and paraphrased sentence lexical diversity. Figure 3 presents the kernel density estimate plots comparing these features for the Human and LLM classes on the Adapted MRPC dataset. These distributional differences provide visual evidence supporting the statistical significance observed and reflect distinct stylistic tendencies between the two groups. For the HLPC dataset, the top 3 features (by absolute Cohen's d) are also visualized in Figure 15 and Figure 16, with corresponding Welch's t-test results detailed in Table 5 and Table 6 in Appendix.



(a): `word_overlap` (b): `para_length` (c): `para_lexical_diversity`

Figure 3: Comparison of feature distributions for the top 3 most statistically significant features distinguishing human vs. LLM-generated paraphrases on the Adapted MRPC dataset.

6.2 Results: Tasks 2, 3, and 4

The performance of DeBERTa-v3 and RoBERTa on the HLPC-derived sentence-level tasks is summarized in Table 3. Overall, both models demonstrate strong classification performance, though their effectiveness varies across tasks and evaluation metrics. These results are consistent with the findings of (Wu et al., 2024), which highlight the robustness and generalization ability of RoBERTa-based models for detecting LLM-generated text in real-world scenarios.

Task 2 (Single Sentence LLM Paraphrase Detection): Both models perform well in distinguishing between Human- and LLM-paraphrased sentences, with RoBERTa achieving slightly higher accuracy (92.44%) and F1-score (87.62%). DeBERTa-v3, while marginally behind on these metrics, remains competitive, suggesting that both models effectively capture stylistic and lexical cues indicative of LLM paraphrasing. The corresponding ROC curve is presented in Figure 17.

Task 3 (Single Sentence Authorship Attribution): In Task 3.1, which compares original human-written sentences with LLM paraphrases, both models achieve strong performance. RoBERTa again leads slightly in accuracy (93.94%) and F1-score (90.47%), while DeBERTa-v3 records the highest AUROC (0.9866) and TPR@1%FPR (0.8224), indicating superior calibration and sensitivity under low false positive conditions. The ROC curve is shown in Figure 18.

Task 3.2 introduces additional complexity by incorporating human-written paraphrases into the human-authored class and expanding the overall sample size to 2,400 instances, thereby increasing intra-class diversity. Despite this, both models maintain strong performance. Notably, DeBERTa-v3 slightly outperforms RoBERTa in F1-score (93.62% vs. 93.20%), suggesting enhanced robustness under more varied authorship scenarios. In addition, compared to the results of Task 3.1, the TPR@1%FPR shows a slight decrease for the DeBERTa-v3 model and a slight increase for the RoBERTa model, while the changes in AUROC and overall accuracy are minimal. These findings contrast with those reported by (Lau and Zubiaga, 2024), who directly applied OpenAI's RoBERTa-based detector and found that incorporating human-written paraphrases improved TPR@1%FPR but potentially reduced AUROC and overall accuracy. The detailed results and ROC curves are presented in Table 7 and Figure 19.

Task 4 (Distinguishing Original vs. LLM-Paraphrased LLM Output): This task is the most challenging, as it requires distinguishing between two types of LLM-generated text. Both models exhibit performance degradation compared to earlier tasks. RoBERTa outperforms DeBERTa-v3 in accuracy (83.28%) and F1-score (74.35%), while DeBERTa-v3 shows a marked drop in TPR@1%FPR (0.2292). These results suggest that identifying paraphrased variants of machine-generated text may require more fine-grained modeling capabilities than those offered by current transformer architectures. The ROC curve is shown in Figure 20.

Overall, modern transformer models demonstrate strong performance in authorship attribution and paraphrase source detection, especially in settings with clear Human-vs-LLM distinctions. However, more nuanced scenarios—such as distinguishing layered generations within LLM outputs—pose greater difficulty, revealing limitations in current models' sensitivity to subtle semantic or stylistic shifts. These findings highlight a promising yet incomplete path toward fine-grained LLM provenance detection.

Table 3: Average 5-Fold Cross-Validation Performance of Transformer Models on HLPC Dataset Variants

	Task 2		Task 3		Task 4	
Model	**DeBERTa**	**RoBERTa**	**DeBERTa**	**RoBERTa**	**DeBERTa**	**RoBERTa**
Accuracy	0.9111	0.9244	0.9361	0.9394	0.8011	0.8328
F1-score	0.8469	0.8762	0.8970	0.9047	0.6967	0.7435
AU-ROC	0.9837	0.9681	0.9866	0.9893	0.8726	0.9125
TPR@1%FPR	0.7538	0.7667	0.8224	0.7750	0.2292	0.3158

Note: All tasks used variants of the HLPC dataset. Models are DeBERTa-v3 and RoBERTa. Metrics are averaged across 5-fold cross-validation.

7 Discussion

This study evaluated LLM paraphrase detection using feature-based (XGBoost) and deep learning (DeBERTa-v3, RoBERTa) models within our PLD framework, revealing varied performance across tasks.

XGBoost, relying on linguistic features, demonstrated high accuracy (96%, 0.9876 AUC-ROC on adapted MRPC) in sentence-pair paraphrase source detection (Task 1) and maintained robustness even after five paraphrasing rounds on HLPC. Feature importance analysis highlighted word overlap, trigram overlap, and paraphrased sentence lexical diversity as key discriminators, supported by statistical tests across datasets, indicating persistent LLM stylistic signatures related to differences in human and LLM paraphrasing strategies concerning overlap, length, and lexical richness. Notably, LLM paraphrases sometimes exhibited higher global lexical diversity—contrasting with prior observations of LLM text repetition (Gehrmann et al., 2019), suggesting LLMs may show complex, context-dependent stylistic patterns.

Transformer models excelled in single-sentence LLM paraphrase detection (Task 2) and authorship attribution (Task 3), achieving over 90% accuracy. While their overall performance was comparable, specific metrics varied by subtask. Distinguishing between original and LLM-paraphrased LLM outputs (Task 4) proved significantly harder, with lower TPR@1%FPR scores suggesting reduced reliability under strict precision. This indicates that further AI processing can obscure distinct LLM signals.

In comparing approaches, XGBoost offered interpretability and strong performance on structured sentence-pair tasks with informative handcrafted features. Transformer models were more accurate on nuanced single-sentence tasks but less transparent. The choice depends on task complexity, interpretability needs, and resources.

Overall, our findings highlight that LLM paraphrasing leaves detectable traces, even after multiple iterations, with implications for AI detection in various content creation contexts. These results also inform future LLM development regarding the potential need to address or exploit these persistent stylistic artifacts.

8 Conclusions and Future Directions

Detecting LLM-generated paraphrases is a growing challenge in NLP, with implications for academic integrity, IP protection, and content authenticity. This work introduces the Paraphrase-based LLM Detection Framework (PLD-4), which breaks down detection into four subtasks, each with varying context and difficulty.

Using the MRPC and HLPC datasets, we evaluate models across these tasks. For Task 1 (Sentence Pair Source Detection), XGBoost achieved high performance (Accuracy: 0.96, F1-score: 0.95, AUC-ROC: 0.9876), and interpretability analysis revealed key linguistic features useful for detection. For Tasks 2–4, transformer models (DeBERTa-v3, RoBERTa) performed well but struggled most on Task 4—distinguishing between LLM-generated and LLM-paraphrased text. RoBERTa achieved 83.28% accuracy and 74.35% F1-score, but performance dropped significantly in fine-grained metrics like TPR@1%FPR, underscoring the challenge of layered AI paraphrasing.

Overall, PLD-4 provides a structured, interpretable foundation for paraphrase detection by combining feature-based and transformer models. It reveals current limitations and paves the way for future work involving more advanced LLMs, adversarial robustness, hybrid modeling, and mixed-authorship document detection.

9 Limitations

Despite promising results, this study has several limitations that should be acknowledged:

- **Scope of Language Models:** The Large Language Models (LLMs) utilized for generating original content (specifically, GPT2-XL for the LLM-DOC component within the Human-LLM Paraphrase Collection, HLPC) and for paraphrasing tasks (BART and Dipper for HLPC) are potent; however, newer and potentially more sophisticated models are continuously emerging (e.g., GPT-4, Claude 3, and their successors as of early 2025). Consequently, the findings herein might not fully generalize to text generated or paraphrased by these state-of-the-art models, which could produce paraphrases that are even more human-like or challenging to detect.

- **Dataset and Framework Specificity:** The Paraphrase-based LLM Detection framework (PLD-framework) introduced and employed in this research was constructed using specific datasets: an adapted version of the Microsoft Research Paraphrase Corpus (MRPC) and the Human-LLM Paraphrase Collection (HLPC). While this setup offers a controlled environment for evaluation, the framework's performance can be influenced by the inherent characteristics of these datasets. Factors such as the source and domain of the original texts, the specific human paraphrasing styles represented, and the LLMs chosen for generation and paraphrasing within HLPC invariably affect the outcomes. Therefore, the generalizability of the findings may be constrained by the particular nature and data distribution within these PLD-framework components.

- **Sample Size for HLPC-Derived Tasks:** Tasks 2, 3, and 4, which utilized datasets derived from the HLPC, were conducted with a sample size of 1,800 instances. Although mitigation strategies such as 5-fold cross-validation, the use of early stopping callbacks during model training, and weight decay were implemented to enhance generalization and reduce potential overfitting, this relatively modest sample size may still restrict the broader applicability or statistical robustness of the findings for these specific transformer-based tasks. More definitive insights could be gained from larger and more diverse datasets for these tasks.

- **Focus on Sentence-Level Detection:** This study primarily concentrates on sentence-level detection. The detection of LLM-paraphrased content at the document level, or the identification of AI-generated segments within documents containing a mix of human and AI authorship, introduces additional complexities and challenges that were not addressed in this work. The efficacy of the proposed features and models when applied to longer texts remains an area for future investigation.

- **Nature of the Paraphrasing Task:** The research utilized paraphrases generated by specific models (BART and Dipper within the HLPC) using their default or predetermined configurations. Different paraphrasing strategies, variations in the level of abstraction in paraphrasing prompts (where applicable), or outputs from alternative paraphrasing models could produce text with distinct characteristics, potentially affecting the performance of the detection methods.

Overall, this work contributes to the understanding and detection of LLM-generated paraphrases, highlighting both the progress made and the challenges that lie ahead in accurately identifying AI-modified text.

10 Ethics Statement

We use only publicly available datasets and pretrained models in this study, all of which are accessed and utilized strictly for research purposes. The use of these resources complies with their original licenses and terms of access. No personally identifiable or sensitive information is present in any of the data used.

Our code will be released under the MIT license to support transparency and reproducibility.

References

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *Preprint*, arXiv:2310.05130.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39:463–472.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the possibilities of ai-generated text detection. *Preprint*, arXiv:2304.04736.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM.

Dipanjan Das and Noah A. Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 468–476, Suntec, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*. Asia Federation of Natural Language Processing.

Serena Fariello, Giuseppe Fenza, Flavia Forte, Mariacristina Gallo, and Martina Marotta. 2024. Distinguishing human from machine: A review of advances and challenges in ai-generated text detection. *International Journal of Interactive Multimedia and Artificial Intelligence*, In press(In press):1–13.

Simone Filice, Giovanni Da San Martino, and Alessandro Moschitti. 2015. Structural representations for learning relations between pairs of texts. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1003–1013, Beijing, China. Association for Computational Linguistics.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. Gltr: Statistical detection and visualization of generated text. *Preprint*, arXiv:1906.04043.

Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *Preprint*, arXiv:2301.04246.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *arXiv preprint arXiv: 2408.08946*.

Ethan Hunt, Ritvik Janamsetty, Chanana Kinares, Chanel Koh, Alexis Sanchez, Felix Zhan, Murat Ozdemir, Shabnam Waseem, Osman Yolcu, Binay Dahal, Justin Zhan, Laxmi Gewali, and Paul Oh. 2019. Machine learning models for paraphrase identification and its applications on plagiarism detection. In *2019 IEEE International Conference on Big Knowledge (ICBK)*, pages 97–104.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey. *Preprint*, arXiv:2011.01314.

Yangfeng Ji and Jacob Eisenstein. 2013. Discriminative improvements to distributional sentence similarity. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 891–896, Seattle, Washington, USA. Association for Computational Linguistics.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2024. A watermark for large language models. *Preprint*, arXiv:2301.10226.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Preprint*, arXiv:2303.13408.

Hiu Ting Lau and Arkaitz Zubiaga. 2024. Understanding the effects of human-written paraphrases in llm-generated text detection. *Preprint*, arXiv:2411.03806.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012. Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 182–190, Montréal, Canada. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *Preprint*, arXiv:2301.11305.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2786–2792. AAAI Press.

Shinwoo Park, Hyundong Jin, Jeong won Cha, and Yo-Sub Han. 2025. Detection of llm-paraphrased code and identification of the responsible llm using coding style features. *Preprint*, arXiv:2502.17749.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Long Qiu, Min-Yen Kan, and Tat-Seng Chua. 2006. Paraphrase recognition via dissimilarity significance classification. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 18–26, Sydney, Australia. Association for Computational Linguistics.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2025. Can ai-generated text be reliably detected? *Preprint*, arXiv:2303.11156.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *Preprint*, arXiv:1908.09203.

Nafis Irtiza Tripto, Saranya Venkatraman, Dominik Macko, Róbert Móro, Ivan Srba, Adaku Uchendu, Thai Le, and Dongwon Lee. 2023. A ship of theseus: Curious cases of paraphrasing in llm-generated texts. *ArXiv*, abs/2311.08374.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, 19(1).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

Junchao Wu, Runzhe Zhan, Derek F. Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S. Chao. 2024. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. In *Advances in Neural Information Processing Systems*, volume 37, pages 100369–100401. Curran Associates, Inc.

Xiao Yu, Kejiang Chen, Qi Yang, Weiming Zhang, and Nenghai Yu. 2024. Text fluoroscopy: Detecting LLM-generated text through intrinsic features. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15838–15846, Miami, Florida, USA. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2020. Defending against neural fake news. *Preprint*, arXiv:1905.12616.

A Appendix

A.1 Task 1 Additional Results

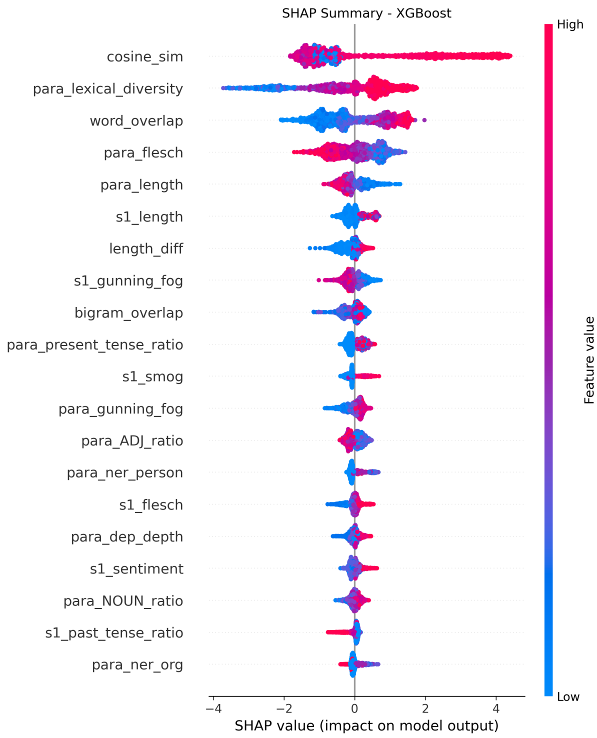


Figure 4: Task 1: SHAP summary plot illustrating feature importance and impact for the XGBoost model on the Adapted MRPC. The plot displays the top 20 features ranked by mean absolute SHAP value.

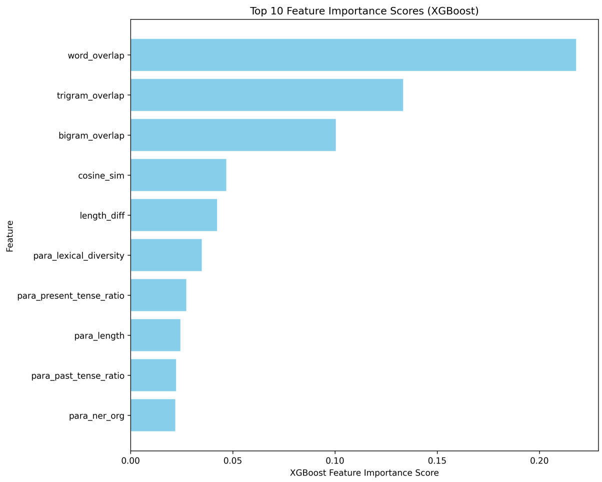


Figure 5: Task 1: XGBoost Feature Importance scores for the top 10 features, derived from the model trained on the Adapted MRPC.

A.1.1 Statistical Comparison of Features (Human vs. LLM)

Tables 4, 5, and 6 present the results of Welch's t-tests comparing the means of linguistic features between human-generated and LLM-generated paraphrase pairs. Table 4 details findings for the Adapted MRPC dataset. Tables 5 and 6 detail findings for the HLPC dataset, using the first and fifth iterative paraphrase outputs, respectively. Only features showing a statistically significant difference ($p<0.05$) are listed, sorted by their original ascending p-value. Significance levels are indicated using standard asterisk notation (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$). Cohen's d is included as a measure of effect size, indicating the magnitude of the difference (positive values indicate the mean is higher for the Human group, negative values indicate the mean is higher for the LLM group based on the Cohen's d calculation order used in this study).

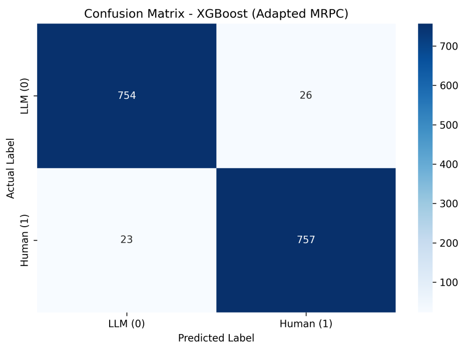


Figure 6: Task 1: Confusion matrix for the XGBoost model predictions on the Adapted MRPC.

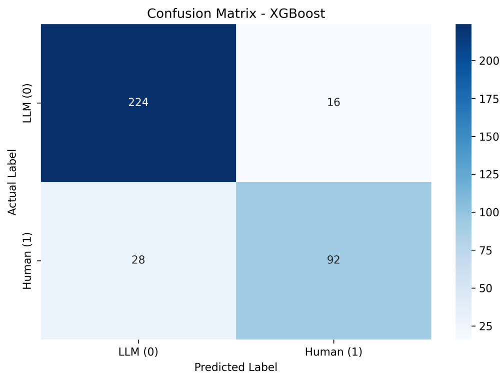


Figure 7: Confusion matrix for the XGBoost model predictions on HLPC (1st Para).

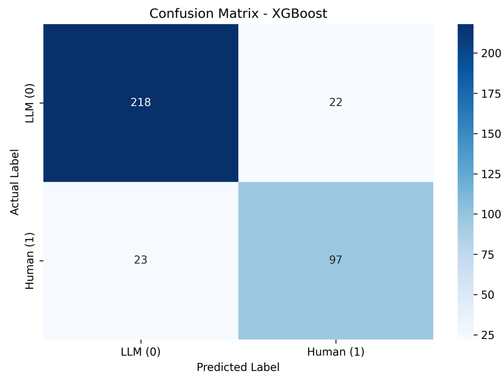


Figure 8: Confusion matrix for the XGBoost model predictions on HLPC (5th Para).

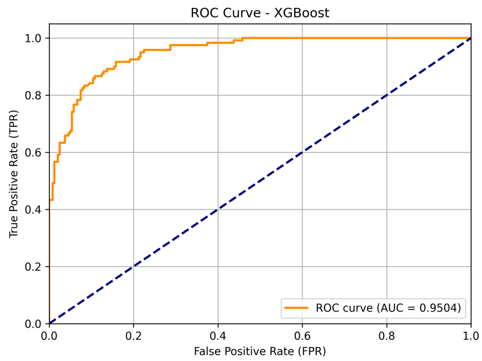


Figure 9: ROC curve for the XGBoost model on HLPC (1st Para).

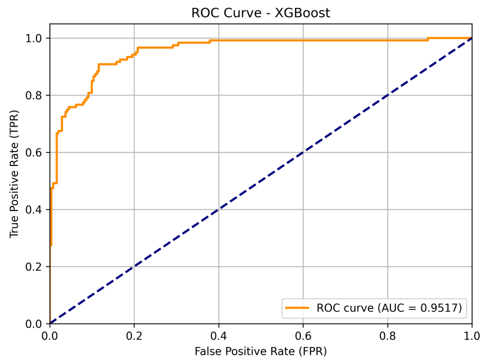


Figure 10: ROC curve for the XGBoost model on HLPC (5th Para).

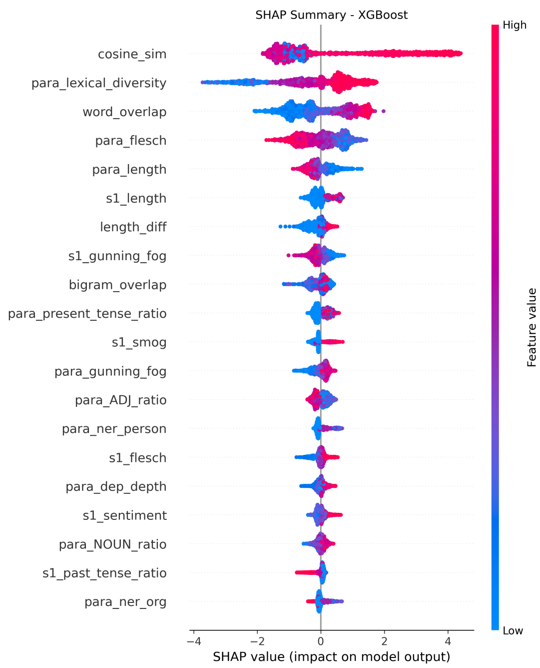


Figure 11: SHAP summary plot for the XGBoost model on HLPC (1st Para), showing the top 20 features ranked by mean absolute SHAP value.

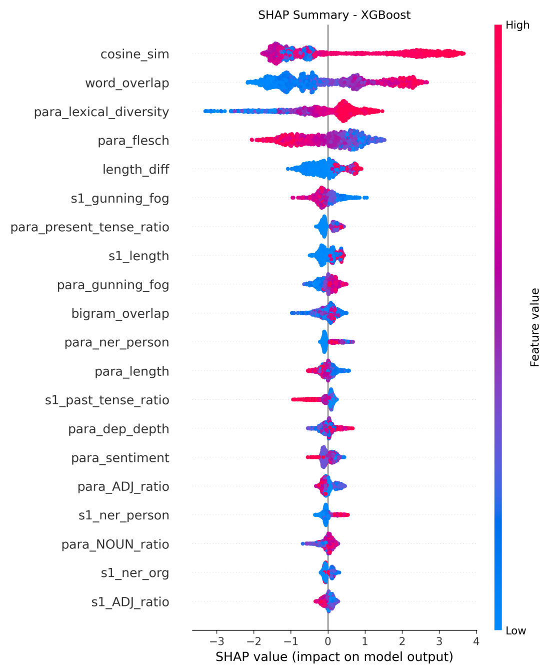


Figure 12: SHAP summary plot for the XGBoost model on HLPC (5th Para), showing the top 20 features ranked by mean absolute SHAP value.

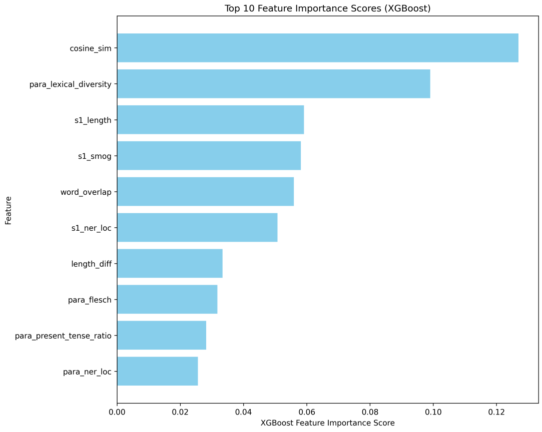


Figure 13: Top 10 feature importance scores from the XGBoost model on HLPC (1st Para).

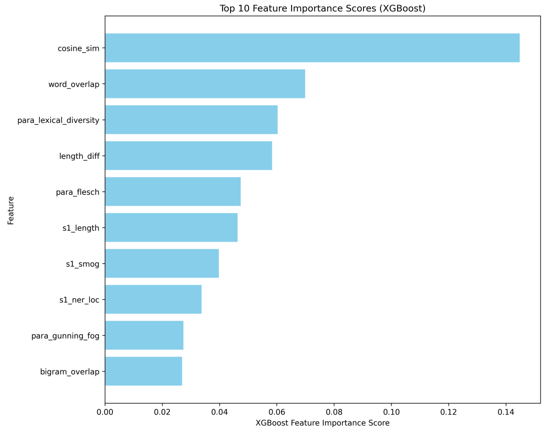
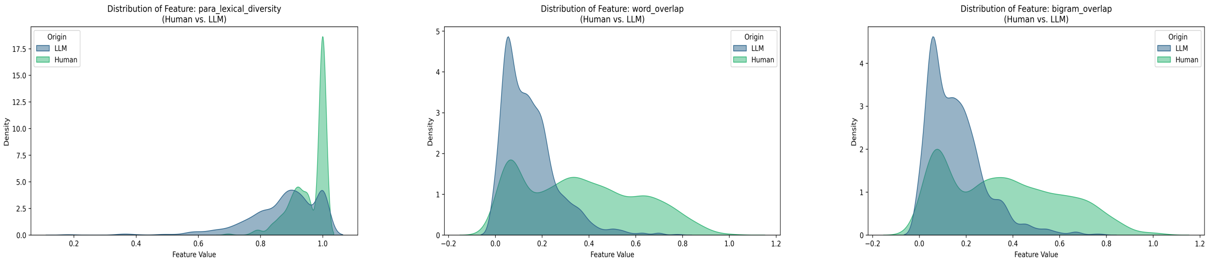
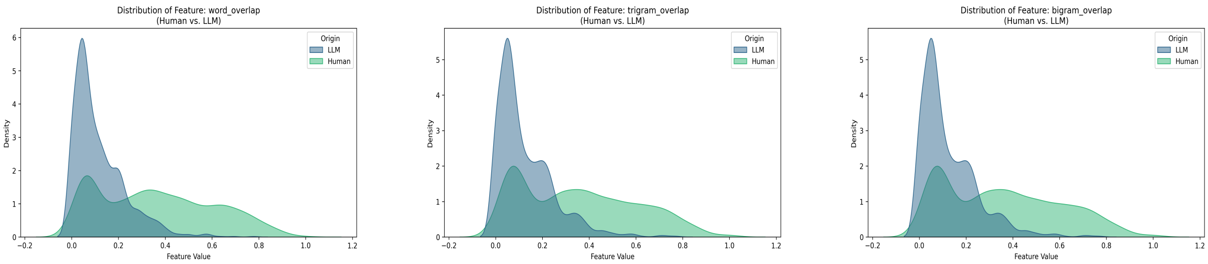


Figure 14: Top 10 feature importance scores from the XGBoost model on HLPC (5th Para).



(a): `para_lexical_diversity` (b): `word_overlap` (c): `bigram_overlap`

Figure 15: Comparison of feature distributions for the top 3 most statistically significant features distinguishing human vs. LLM-generated paraphrases on the Adapted HLPC dataset (First Paraphrase).



(a): `word_overlap` (b): `trigram_overlap` (c): `bigram_overlap`

Figure 16: Comparison of feature distributions for the top 3 most statistically significant features distinguishing human vs. LLM-generated paraphrases on the Adapted HLPC dataset (Fifth Iterative Paraphrase).

A.2 Task 2, 3, and 4 Additional Results

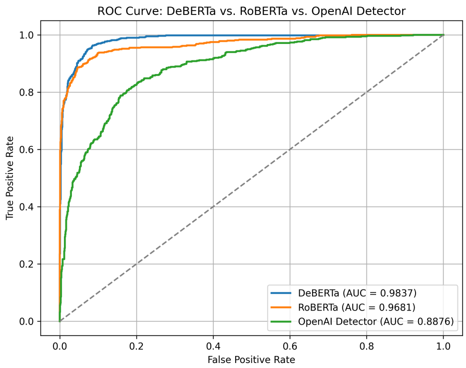


Figure 17: Task 2: ROC Curve — DeBERTa vs. RoBERTa vs. OpenAI Detector.

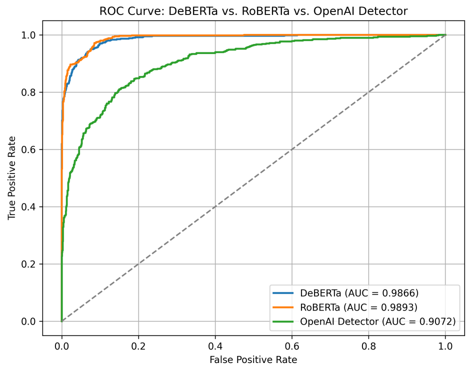


Figure 18: Task 3: ROC Curve — DeBERTa vs. RoBERTa vs. OpenAI Detector.

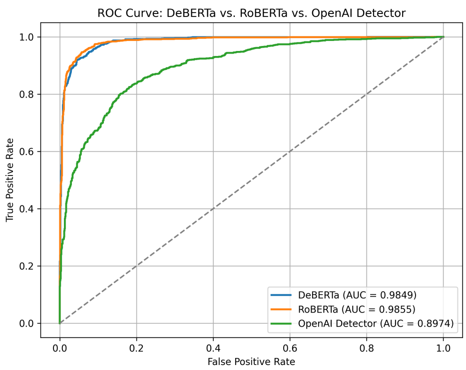


Figure 19: Task 3.2: ROC Curve — DeBERTa vs. RoBERTa vs. OpenAI Detector.

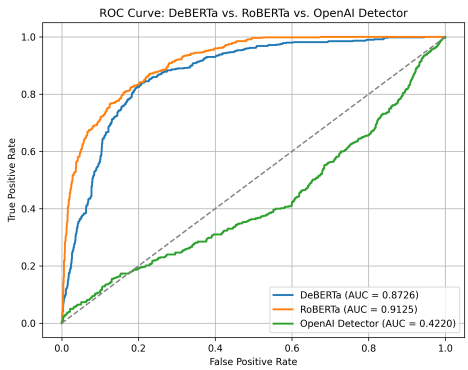


Figure 20: Task 4: ROC Curve — DeBERTa vs. RoBERTa vs. OpenAI Detector.

Table 4: Significant Features Distinguishing Human vs. LLM Paraphrases on Adapted MRPC Dataset (Welch's t-test, p<0.05, sorted by p-value)

Feature	t-statistic	Cohen's d	Signif.
word_overlap	66.76	1.719	***
para_length	24.47	0.671	***
para_lexical_diversity	-21.50	-0.561	***
cosine_sim	19.99	0.557	***
s1_length	12.58	0.349	***
bigram_overlap	12.20	0.316	***
trigram_overlap	12.20	0.316	***
s1_gunning_fog	11.20	0.320	***
s1_dep_depth	10.68	0.290	***
length_diff	9.15	0.248	***
s1_flesch	-8.73	-0.249	***
para_smog	-6.76	-0.212	***
s1_smog	-5.83	-0.175	***
s1_past_tense_ratio	-4.75	-0.160	***
para_flesch	4.65	0.135	***
para_dep_depth	3.95	0.109	***
s1_lexical_diversity	-3.69	-0.103	***
para_VERB_ratio	-3.36	-0.095	***
s1_modal_verb_ratio	-2.61	-0.097	**
s1_VERB_ratio	2.45	0.071	*
para_sentiment	2.19	0.060	*

Note: Features sorted by original ascending p-value. Welch's t-test results comparing feature means between Human and LLM groups. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Cohen's d indicates effect size.

Table 5: Significant Features Distinguishing Human vs. LLM Paraphrases on the HLPC Dataset (First Paraphrase Outputs). Welch's t-test, p<0.05, sorted by p-value.

Feature	t-statistic	Cohen's d	Signif.
para_lexical_diversity	20.6839	0.8353	***
word_overlap	20.5785	1.2773	***
bigram_overlap	18.9607	1.1658	***
trigram_overlap	18.9607	1.1658	***
para_flesch	-11.2761	-0.5374	***
para_gunning_fog	9.4867	0.4764	***
cosine_sim	8.0593	0.4285	***
para_length	-5.8078	-0.2960	***
para_sentiment	-4.9107	-0.2400	***
para_ner_person	4.2959	0.2216	***
para_smog	-4.0234	-0.1696	***
para_present_tense_ratio	3.3096	0.1700	***
para_ADJ_ratio	-2.9596	-0.1450	**

Note: Features sorted by original ascending p-value. Welch's t-test results comparing feature means between Human and LLM groups. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Cohen's d indicates effect size. Data based on analysis of HLPC first paraphrase outputs.

Table 6: Significant Features Distinguishing Human vs. LLM Paraphrases on the HLPC Dataset (Fifth Iterative Paraphrase Outputs). Welch's t-test, p<0.05, sorted by p-value.

Feature	t-statistic	Cohen's d	Signif.
word_overlap	23.5635	1.4736	***
para_lexical_diversity	14.6123	0.5948	***
para_flesch	-14.2612	-0.6830	***
para_gunning_fog	13.2602	0.6602	***
bigram_overlap	21.9905	1.3550	***
trigram_overlap	21.9905	1.3550	***
cosine_sim	10.0935	0.5328	***
para_ner_person	6.4044	0.3414	***
para_dep_depth	4.6873	0.2335	***
para_ner_org	4.3942	0.2302	***
para_present_tense_ratio	4.1584	0.2130	***
para_length	3.9678	0.1989	***
para_sentiment	-3.5301	-0.1771	***
para_NOUN_ratio	3.0496	0.1442	**
para_ner_loc	2.5881	0.1449	**

Note: Features sorted by original ascending p-value. Welch's t-test results comparing feature means between Human and LLM groups. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Cohen's d indicates effect size. Data based on analysis of HLPC fifth iterative paraphrase outputs.

Table 7: Average 5-Fold Cross-Validation Performance on Task 3.2 (HLPC Dataset)

Metric	DeBERTa	RoBERTa
Accuracy	93.71%	93.46%
F1-score	93.62%	93.20%
AU-ROC	0.9887	0.9896
TPR@1%FPR	0.7950	0.8048

Note: Results reflect average performance on Task 3.2 using the HLPC dataset. Metrics were computed via 5-fold cross-validation.