

Towards Fine-grained and Robust Detection of Human–AI Co-authorship

Ziruo Zhao Jie Fu Wendy Hui Wang Zining Zhu

Stevens Institute of Technology

Hoboken, New Jersey, USA

{zzhao83, jfu13, hwang4, zzhu41}@stevens.edu

Abstract

The widespread adoption of large language models (LLMs) is reshaping how textual content is produced, complicating text provenance, and raising concerns for misinformation, plagiarism, and intellectual property. While prior work on AI-generated text detection has progressed, one-dimensional human-AI co-authorship scores lack signals directly relevant to “AI involvement”. It remains unclear how adversarial paraphrasing distorts detection signals. In this work, we formulate a discrete-class text provenance detection framework. This framework allows an abstraction of realistic co-authorship scenarios into distinct degrees of AI involvement, differentiating the AI polishing on the surface level from the AI updating the discourse structures. This framework reveals the mechanisms of how adversarial paraphrasing blurs the decision boundaries for co-authorship detectors. To improve robustness of the detectors, we augment training data with adversarial paraphrases generated by the state-of-the-art paraphrasing attack model, exposing detectors to challenging meaning-preserving rewrites. Extensive experiments across multiple datasets and generation models demonstrate that adversarial paraphrasing substantially degrades detectors trained only on benign data, while adversarial data augmentation leads to consistent robustness gains and good cross-generator transferability. Overall, our findings highlight the value of discrete-class formulation and principled data augmentation for building reliable detectors of human-AI co-authorship.

1 Introduction

Large language models (LLMs) have become widely used as writing assistants, supporting a broad range of tasks such as drafting, editing, and paraphrasing text (Liang et al., 2025; Bhagat and Hovy, 2013). As a result, many real-world documents are no longer purely human-written or fully

machine-generated, but instead reflect varying degrees of human-AI co-authorship (Paredes et al., 2025; Law et al., 2025). The shift introduces novel challenge, including the academic integrity, the tracing of information and the content provenance.

Therefore, many recent works have sought to detect AI involvement in the writing. In practice, AI involvement in writing falls into diverse categories. Texts may undergo light AI-assisted editing, such as grammatical correction or stylistic refinement, or be substantially rewritten by language models while still originating from human-authored content. These intermediate cases blur the boundary between human-written and AI-generated text, recent work has started moving beyond binary detection toward fine-grained classification of AI involvement (Abassy et al., 2024; Cheng et al., 2025; Su et al., 2025; Thai et al., 2026).

There are two unaddressed challenges.

First, defining a continuous “AI involvement” score is inherently difficult, as the degree of AI involvement is ultimately a property of the authoring process, not directly recoverable from post-hoc detector outputs. While detectors like Thai et al. (2026) may leverage various signals from stylistic and geometric features to semantic and discourse-level cues, their output scores are not designed to map onto a one-dimensional “AI involvement” level. Content provenance requires a more direct signal showing the level of “AI involvement”.

Second, adversarial paraphrasing poses a particularly realistic threat to AI-involved text detection (Sadasivan et al., 2025; Tripto et al., 2024). Recent paraphrasing methods, such as DIPPER (Krishna et al., 2023), are explicitly designed to generate meaning-preserving rewrites that substantially perturb lexical and syntactic patterns. As illustrated in Figure 1, adversarial paraphrasing can cause texts with low AI involvement to be misclassified as human-written. This is a critical vulnerability of detectors, and the mechanisms behind this vulnera-

bility remains largely unknown.

In this work, we tackle both challenges through a unified and controlled framework. Instead of trying to propose a continuous score that capture AI involvement, we take a step back and propose a four-class task formulation, distinguishing between multiple levels of AI involvements: “pure-human”, “low AI-involved”, “high AI-involved” and “fully-AI”. This four-class setting corresponds to realistic scenarios of AI-assisted writing: after humans write “pure-human” class texts, one may use LLMs to polish the texts. This polish stays on the surface level, correcting typos and grammatical errors. This scenario is reflected by the “low AI-involved” class. The human user may use LLMs to develop the article, changing the argument structures and the discourse patterns. This scenario is reflected by the “high AI-involved” class. The human user may use LLMs to fill in the contents after briefly describing an idea or topic at the high level. The details in this scenario are generated by AI, and is reflected by the “fully-AI” class. Therefore, this four-class setting provides signals that are directly relevant to the content provenance. The contents in “low AI-involved” texts should be attributed to the humans, while one could argue that the content details of the “fully-AI” texts can be attributed to the AI. For “high AI-involved” texts, one could argue that, for example, the contents attributed to humans while the discourse structures to the AIs.

This formulation allows us to stress-test the detector’s robustness when faced with adversarial paraphrasing. Under this setting, we study the impacts of adversarial paraphrasing on the co-authorship detectors. We validate that the adversarial paraphrasing obscures the syntactic and semantic signals used for detection, blurring the detection boundaries for traditional detection settings. To enhance the robustness, we augment the adversarially paraphrased data. We systematically evaluate robustness under increasing rewriting strength and cross-generator evaluation.

Our contributions are summarized as follows:

- We propose a four-class formulation for human–AI co-authorship detection, where we carefully control the different degrees of AI involvement.
- We construct a benchmark dataset for fine-grained human–AI co-authorship detection under adversarial paraphrasing. We further validate that the dataset conforms to the intended AI-involved settings through quantitative analysis.

- We systematically analyze the impact of adversarial paraphrasing on both syntactic and semantic similarity, showing how meaning-preserving rewrites distort surface cues and exacerbate class overlap in fine-grained detection, thereby directly addressing the challenges posed by adversarial paraphrasing.
- We conduct comprehensive robustness evaluations across multiple datasets, generation models, and paraphrasing attacks, demonstrating consistent improvements in detection robustness under realistic rewriting scenarios.

Overall, our findings highlight the importance of fine-grained task formulation and data augmentation with adversarially paraphrased samples for building human-AI co-authorship detectors. Our framework and benchmark make a concrete step towards the data provenance of human-AI co-authored articles.

2 Related Works

AI-generated Text Detection Methods. Existing methods for AI-generated text detection can be broadly categorized into three paradigms. (1) Zero-shot methods leverage model-internal statistics or generation artifacts such as likelihood-based signals or watermarking, but often require access to specific model internals, limiting applicability to third-party text (Mitchell et al., 2023; Su et al., 2023; Zhao et al., 2023; Bao et al., 2024). (2) Feature-based methods combine explicit statistical or linguistic features with lightweight classifiers, but remain sensitive to feature design and degrade under paraphrasing or stylistic variation (Joulin et al., 2017; Gehrmann et al., 2019). (3) PLM-based methods fine-tune pretrained language models on labeled data for text classification, achieving strong in-distribution performance but exhibiting reduced robustness in out-of-distribution data (Salazar et al., 2020; Rodriguez et al., 2022). To handle longer inputs, long-document transformers such as Longformer (Beltagy et al., 2020) extend standard PLMs with sparse attention mechanisms for efficient long-form text processing.

Overall, while existing detection methods perform well under benign conditions, they remain vulnerable to paraphrasing and distribution shifts, motivating the need for detection frameworks that explicitly consider adversarial rewriting and varying degrees of AI involvement. (Pu et al., 2023; Guo et al., 2023; Li et al., 2024a,b)

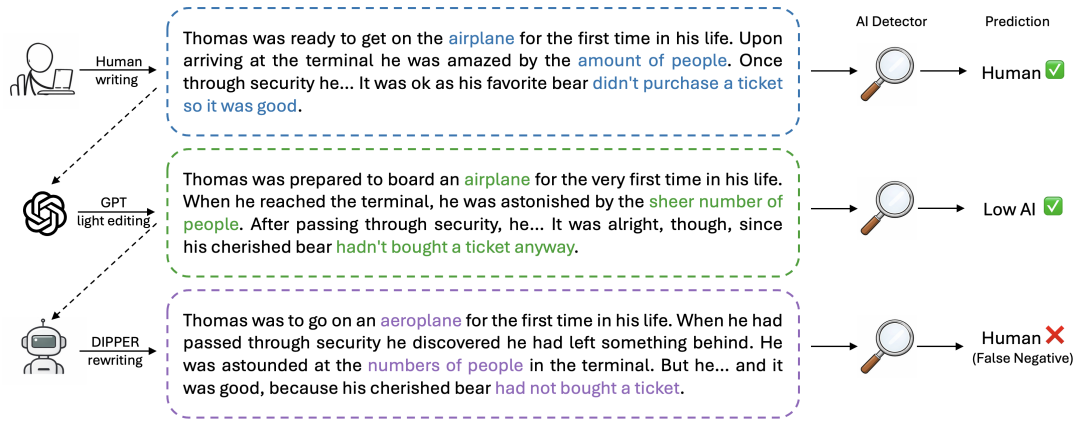


Figure 1: A ROCStories example showing that light GPT editing is correctly identified, while adversarial paraphrasing (DIPPER) preserves meaning but induces a false negative in an existing detector.

Human-AI Collaborative Text Detection. Recent work has begun to move beyond binary AI detection. Some studies treat human-AI collaboration as a distinct category (Abassy et al., 2024; Cheng et al., 2025; Su et al., 2025), and another line of work considers humanized machine-generated text (Zhang et al., 2024; Masrour et al., 2025). A related direction focuses on span or boundary detection, which identifies localized human-AI transitions within a document (Kushnareva et al., 2024; Li et al., 2024b). Most related to our work, Thai et al. (2026) propose EDITLENS, a regression model that quantifies the degree of AI editing using similarity-based metrics as a proxy for AI involvement. In contrast, our work operationalizes AI involvement directly through the generation process itself, using controlled rewriting scenarios as ground-truth supervision, and targets global involvement estimation under realistic full-document editing scenarios.

Paraphrasing attack. Recent studies have demonstrated that paraphrasing-based attacks pose a significant threat to AI-generated text detection. Paraphrasing-based attacks aim to preserve the original semantic content while altering surface-level lexical and syntactic patterns, thereby directly targeting the stylistic cues relied upon by many detectors. Sadasivan et al. (2025); Tripto et al. (2024) show that paraphrasing-based attacks can consistently evade a wide range of detection approaches.

Among existing methods, DIPPER (Krishna et al., 2023) is specifically designed to generate high-quality, meaning-preserving rewrites by jointly controlling semantic similarity and lexical diversity. By selectively rewriting watermark-

related tokens and adjusting sentence structure at the paragraph level, DIPPER generates natural and fluent text that more closely resembles human writing. While originally proposed to bypass watermarking-based detectors, our experiments show that DIPPER is also effective against other detection methods.

Several works further explore adversarial settings that explicitly model interactions between detectors and paraphraser, including adversarial training or multi-round attack scenarios (Hu et al., 2023; Zhou et al., 2024; CHENG et al., 2025; Shi et al., 2024). While effective, such frameworks typically require iterative detector-paraphraser interaction, resulting in substantial computational overhead and limiting scalability. Our work differs from their settings, focusing on meaning-preserving paraphrasing as a realistic and scalable threat model for studying detection robustness.

3 Problem Setup

This section defines the task formulation used throughout this work and characterizes how adversarial paraphrasing interacts with fine-grained human-AI co-authorship detection.

3.1 4-Class Task Formulation

Instead of framing text detection as a binary classification problem, we formulate a four-class task to capture different degrees of AI involvement in writing: Human-written, Low AI-involved, High AI-involved, and Fully AI-generated. Throughout this work, AI involvement refers to the use of practically deployed large language models in real-world applications, such as ChatGPT and similar systems.

Generator	Data	Threshold	Low AI (%)	High AI (%)
GPT	XSum	0.233	95.00	97.67
	arXiv	0.247	94.61	97.06
	ROCStory	0.605	94.99	97.83
Claude	XSum	0.298	90.00	92.50
	arXiv	0.366	94.61	91.18
	ROCStory	0.678	94.99	98.83

Table 1: Evaluation of surface-level separation on AI-paraphrased data using normalized Levenshtein distance. The percentages indicate the proportion of samples classified as low/high-AI involvement under each threshold.

Human-written text is authored entirely by humans. **Low AI-involved** text corresponds to scenarios where AI assistance is limited to surface-level editing, such as refining lexical choice or adjusting local sentence form, while preserving the original discourse organization, and factual content. This setting aligns with common real-world applications such as grammar checking or light stylistic polishing tools. **High AI-involved** text reflects a stronger form of assistance in which AI is entrusted with higher-level linguistic organization (Namuduri et al., 2025; Huber and Niklaus, 2025). Given a human-authored draft or factual input, the model is allowed to organize discourse, expression, and narrative structure while preserving the underlying factual content. In this setting, AI contributes to structuring and re-articulating the text, rather than merely editing its surface form, which reflects common real-world scenarios in which large language models assist humans in drafting and reorganizing text. For the **Fully AI-generated** class, a text is produced entirely by an AI model given only a writing direction or topic, with the model independently determining both the organization of content and its linguistic realization.

In this work, to simulate the defined AI-involved classes, labels are assigned based on the prompt-guided generation process rather than post-hoc similarity thresholds. Detailed descriptions of the data generation procedures are provided in Appendix A. To validate that the resulting categories exhibit the intended surface-level distinctions, we compute surface similarity measures between human-written text and its Low and High AI-involved counterparts. Table 1 reports the proportion of samples categorized as Low or High AI involvement under different similarity thresholds across datasets, showing that the two classes remain consistently separable in terms of surface form.

In addition, Appendix C presents visualizations

of both syntactic and semantic similarity metrics, further confirming that the generated data aligns with the intended distinctions between AI involvement levels. By explicitly modeling different degrees of AI involvement, our formulation moves beyond a binary authorship setting and enables a finer-grained analysis of detection behavior under varying levels of human–AI co-authorship. This distinction is particularly important in realistic writing scenarios, where human and AI contributions are often intertwined at different intensities.

3.2 Impact of Adversarial Paraphrasing on Detection

To analyze adversarial paraphrasing within the proposed four-class formulation, we first clarify its role with respect to class semantics. Adversarial paraphrasing operates as a meaning-preserving transformation applied to an existing document, designed to perturb surface-level cues exploited by detectors while preserving the underlying content and writing process. In our formulation, class labels are determined by the degree of AI involvement in the original generation process rather than by surface similarity or representation proximity. Consequently, although adversarial paraphrasing can substantially perturb lexical, syntactic, and stylistic patterns, it is treated within our formulation as preserving the original level of AI involvement, rather than introducing a new source of authorship.

To examine how adversarial paraphrasing affects detection behavior, we analyze document representations from a Longformer-based four-class detector trained solely on benign data, projected into two dimensions via PCA. Figure 2 presents three settings with increasing adversarial rewriting strength: (a) benign only, (b) combined with moderate adversarial samples, and (c) combined with stronger adversarial samples. Under the benign setting, representations form relatively compact and ordered clusters reflecting a continuous human–AI spectrum. As adversarial samples are introduced, this structure is progressively disrupted. In particular, Low and High AI-involved representations become widely dispersed along the "human-to-high AI" continuum rather than shifting monotonically. This dispersion leads to substantial overlap between neighboring classes. In contrast, the Fully AI-generated class primarily exhibits cluster expansion, resulting in a broader but still coherent distribution. This asymmetric behavior indicates that adversarial paraphrasing induces non-directional

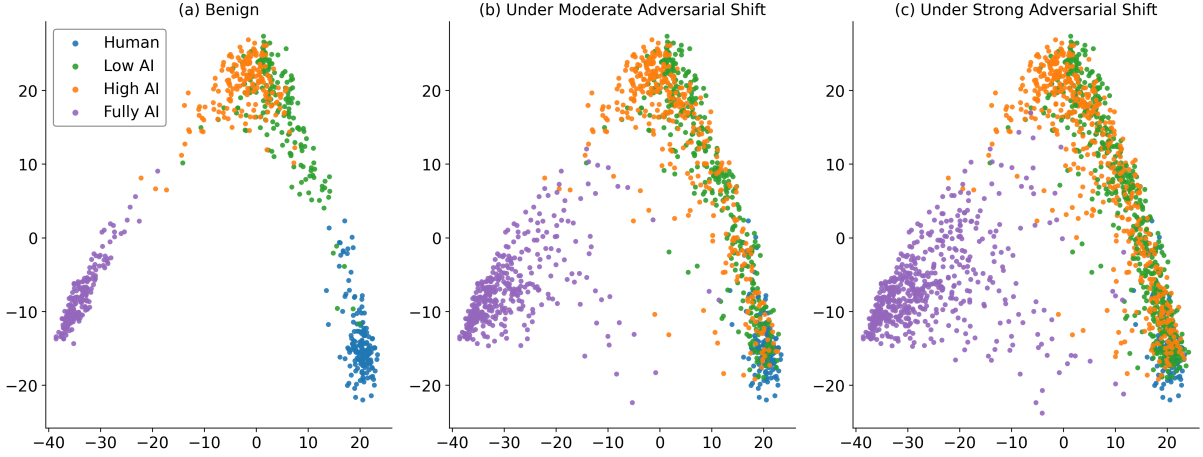


Figure 2: PCA visualization of document representations under increasing adversarial paraphrasing, with stronger paraphrases inducing shifts and overlap among AI-involved classes in representation space.

variation within the human-authored regime rather than producing a new, more AI-involved category.

Overall, these observations show that adversarial paraphrasing causes systematic distributional shifts in representation space while preserving class semantics. This distinction explains why paraphrased samples retain their original labels yet remain challenging for detectors trained on benign data, motivating the need for training strategies that explicitly account for adversarial rewriting.

4 Methodology

Based on the problem formulation and challenges described above, we propose a supervised detection framework that incorporates adversarial paraphrasing during training to improve robustness under a fine-grained multi-class formulation.

4.1 Augmenting Data with Adversarial Paraphrasing

To enhance robustness against adversarial rewriting, we augment the training data with adversarially paraphrased samples. We adopt DIPPER (Krishna et al., 2023) as our primary paraphrasing method, as it produces high-quality, meaning-preserving rewrites that are designed to disrupt surface-level and structural cues exploited by pretrained language model-based detectors. Other paraphrasing approaches were explored but found to be less effective in generating reliable adversarial examples under our multi-class formulation.

Adversarial paraphrasing introduces substantial representational variation beyond the base four-class dataset, reducing over-reliance on generator-specific or surface-level features. While para-

phrased documents may appear more human-like at the surface level, the rewriting process is a post-hoc transformation applied to existing content rather than a new instance of AI-assisted generation.

Consistent with our formulation, adversarially paraphrased samples retain the same class labels as their original inputs and are not treated as a separate category. This design preserves the semantics of AI involvement defined by the writing process, enabling systematic evaluation of detector robustness under increasing levels of adversarial rewriting.

4.2 Training Detection Classifier

Our study focuses on robustness and co-authorship detection under adversarial rewriting. Prior work (Krishna et al., 2023) has shown that paraphrasing-based attacks can effectively evade zero-shot and heuristic detectors; and such approaches are not designed to support controlled multi-class formulations. Therefore, we adopt a supervised PLM-based detection framework that enables explicit multi-class modeling and controlled robustness evaluation under adversarial rewriting.

We use Longformer (Beltagy et al., 2020) as the backbone encoder. Longformer replaces full self-attention with sparse attention patterns, allowing efficient modeling of long-range dependencies in long text. Given an input document, the encoder produces contextualized token representations in the final hidden layer.

We obtain a fixed-length document representation by applying mean pooling over token embeddings weighted by the attention mask. This representation is passed to a lightweight classification head consisting of a fully connected layer followed

by a softmax function to predict the target label. Detection is formulated as a 4-class classification task and trained by minimizing cross-entropy loss.

5 Benchmark Dataset Construction

To enable systematic evaluation under the proposed four-class formulation, we construct a benchmark dataset using controlled, prompt-guided generation.¹ This design supports fine-grained analysis by disentangling surface-level rewriting from underlying writing processes and enabling systematic examination of detector behavior.

5.1 Human-written Data Sourcing

We collect human-written texts from three data sources covering diverse writing domains:

1. News writing, including 600 BBC articles from the XSum dataset (Narayan et al., 2018);
2. Academic writing, consisting of 204 paper introductions collected from arXiv computer science categories published between 2020 and 2021;
3. Story writing, including 599 five-sentence stories from the ROCStories corpus (Mostafazadeh et al., 2016).

5.2 Generation of AI-involved Classes

For AI-involved classes generation, we use two large language models: GPT-4o (OpenAI, 2024) and Claude Sonnet 4 (Anthropic, 2025). For each human-written document, we generate machine-written variants using three prompting strategies that reflect increasing levels of AI involvement. Low-AI prompts restrict the model to sentence-by-sentence rewriting with minimal surface-level edits. High-AI prompts provide the full original text and allow more flexible rewriting, including sentence restructuring and stylistic variation. Fully-AI prompts supply only a brief prefix of the original document and require the model to generate a complete continuation. Detailed prompts are provided in Appendix A.

5.3 Adversarial Paraphrasing

To further construct adversarial examples, we apply DIPPER (Krishna et al., 2023) to generate paraphrased versions of machine-written texts. Following the analysis in the original DIPPER study, we control paraphrasing strength using lexical diversity $L=60$ and order diversity $O=60$, which provides a balanced trade-off between surface-level

variation and semantic preservation. DIPPER-based paraphrasing is applied single and two times to samples from three AI-involved classes. In addition, we construct held-out adversarial test sets using GPT-5.1 (OpenAI, 2025). This model is not used during training or data augmentation and serves solely for robustness evaluation.

6 Experiments

6.1 Experimental Setup

We evaluate our approach under a controlled experimental setup designed to study human-AI co-authorship and robustness under adversarial rewriting. All experiments follow the proposed 4-class classification formulation.

Training Settings. We consider three training settings with increasing levels of adversarial data augmentation. In the **Benign** setting, the detector is trained on the original 4-class dataset. **Level-1** augments the Benign setting with single-step DIPPER paraphrases. **Level-2** further extends Level-1 with a cumulative mixture of single-step and two-step DIPPER paraphrases. Adversarial paraphrasing is applied only to AI-involved samples, with paraphrased samples assigned the same labels as their original texts. All models are trained using class-weighted cross-entropy loss to address class imbalance.

Testing Settings. For each training setting, we evaluate the detector on test sets constructed under the same three adversarial levels, enabling cross-level robustness evaluation.

Metrics. We report Accuracy, Macro F1 score and ROC-AUC for all experiments, with per-class F1 scores for detailed analysis. Predictions are obtained using a default decision threshold of 0.5. Accuracy reflects performance at a fixed operating point, while ROC-AUC provides a threshold-independent measure of detection performance.

6.2 Performance Evaluation

We first establish baseline performance under benign conditions to contextualize the human-AI co-authorship detection task. We evaluate a set of representative supervised baselines, including feature-based classifiers and pretrained language model (PLM) classifiers, under the same 4-class formulation as our main approach.

For feature-based methods, we consider a Linguistic baseline implemented with FastText (Joulin et al., 2017) and a perplexity-based detector

¹<https://github.com/zhaoziruo/CoAuthDetect>

Data	GPT-4o				Claude Sonnet 4			
	Method	Acc	Macro-F1	AUC	Method	Acc	Macro-F1	AUC
XSum	Linguistic	0.8694	0.8682	0.9694	Linguistic	0.9472	0.9471	0.9904
	Perplexity	0.5250	0.5054	0.7836	Perplexity	0.5722	0.5561	0.8253
	BERT	0.9387	0.9388	0.9942	BERT	0.9526	0.9525	0.9946
	RoBERTa	0.9716	0.9716	0.9995	RoBERTa	0.9638	0.9638	0.9974
	Ours	0.9805	0.9805	0.9998	Ours	0.9749	0.9729	0.9979
ROCStory	Linguistic	0.8500	0.8489	0.9723	Linguistic	0.8806	0.8808	0.9858
	Perplexity	0.5944	0.5914	0.8531	Perplexity	0.6778	0.6733	0.8968
	BERT	0.9605	0.9605	0.9995	BERT	0.9944	0.9945	0.9999
	RoBERTa	0.9577	0.9577	0.9992	RoBERTa	0.9861	0.9861	0.9999
	Ours	0.9666	0.9665	0.9996	Ours	0.9972	0.9972	1.0000
arXiv	Linguistic	0.5200	0.4438	0.8059	Linguistic	0.7600	0.7464	0.9351
	Perplexity	0.5650	0.5657	0.7988	Perplexity	0.6300	0.6253	0.8751
	BERT	0.6131	0.5862	0.8774	BERT	0.7337	0.7182	0.9520
	RoBERTa	0.5980	0.5791	0.8461	RoBERTa	0.8844	0.8800	0.9847
	Ours	0.7286	0.7252	0.9382	Ours	0.9296	0.9287	0.9913

Table 2: Detection performance of baseline methods and ours.

that uses LLaMA-2 to compute token-level perplexity, followed by a Random Forest classifier trained on aggregated perplexity statistics. We also evaluate supervised PLM-based classifiers, including BERT (uncased) (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and our Longformer-based model (Beltagy et al., 2020), all fine-tuned under identical data splits and evaluation protocols to ensure fair comparison.

Table 2 reports performance of different methods under the benign in-distribution condition, with results over six settings formed by three datasets and two generators. Overall, PLM-based classifiers achieve higher performance than feature-based methods. Among PLM models, RoBERTa consistently improves over BERT, and our model achieves the strongest overall performance across metrics, showing its effectiveness in modeling co-authorship cue information across diverse domains.

6.3 Adversarial Training Evaluation

Table 3(a) presents detection performance under different combinations of training and testing adversarial levels. We analyze how adversarial paraphrasing affects model robustness as the rewriting strength increases.

First, models trained only on benign data perform well under benign testing conditions but exhibit substantial performance degradation when evaluated on adversarial augmented test sets Level-1 and Level-2. This behavior indicates that detectors trained without adversarial exposure rely heavily on surface-level cues of the specific gen-

erator, which become unreliable once adversarial paraphrasing is introduced.

In contrast, incorporating adversarial paraphrases during training yields consistent robustness gains. Models trained with Level-1 data maintain performance across all testing levels, suggesting exposure to single-step adversarial rewriting enables the detector to move beyond superficial lexical patterns and learn more stable signals that persist under paraphrasing.

To assess whether stronger augmentation is necessary, we further train models with Level-2 data including two-step DIPPER paraphrases. While Level-2 training preserves the robustness achieved by Level-1, improvements in accuracy and AUC are marginal, suggesting diminishing returns from increasing paraphrasing strength beyond a single step. Additional results on other datasets, reported in Appendix D (Table 9 and 10) exhibit the same qualitative trends across domains.

6.4 Transferability Study

We further evaluate cross-generator generalization to assess whether detectors trained on one generation model can transfer to text generated by another. As summarized in Table 3(b), out-of-model performance is slightly lower than in-model evaluation, as expected given the distribution shift but remains competitive overall.

We then evaluate robustness under cross-adversarial attacks in Table 3(c). Although performance decreases on out-of-distribution adversarial samples generated by GPT-5.1, models trained with

Settings		Benign Data Generator: GPT-4o			Settings		Benign Data Generator: Claude		
Train	Test	Acc	F1	AUC	Train	Test	Acc	F1	AUC
(a) Same generator and same attack.									
Benign	Benign	0.9805	0.9805	0.9998	Benign	Benign	0.9749	0.9729	0.9979
Level-1		0.9861	0.9861	0.9990	Level-1		0.9694	0.9693	0.9948
Level-2		0.9638	0.9636	0.9973	Level-2		0.9610	0.9609	0.9960
Benign	Level-1	0.7695	0.7560	0.9386	Benign	Level-1	0.6789	0.6796	0.8999
Level-1		0.9475	0.9514	0.9947	Level-1		0.9491	0.9504	0.9927
Level-2		0.9396	0.9413	0.9920	Level-2		0.9332	0.9356	0.9926
Benign	Level-2	0.6652	0.6367	0.8988	Benign	Level-2	0.5373	0.5426	0.8326
Level-1		0.9321	0.9398	0.9921	Level-1		0.9321	0.9367	0.9890
Level-2		0.9244	0.9292	0.9890	Level-2		0.9143	0.9197	0.9913
(b) Transfer across generators									
Benign	Claude	0.9387	0.9386	0.9960	Benign	GPT-4o	0.8997	0.8991	0.9933
Level-1		0.9062	0.9043	0.9898	Level-1		0.9078	0.9084	0.9825
Level-2		0.8877	0.8849	0.9833	Level-2		0.8954	0.9010	0.9816
(c) Transfer across adversarial paraphrasing attacks									
Benign	GPT-5.1	0.6212	0.5377	0.8413	Benign	GPT-5.1	0.4903	0.3694	0.7882
Level-1		0.8691	0.8682	0.9662	Level-1		0.8134	0.8038	0.9433
Level-2		0.8719	0.8680	0.9554	Level-2		0.7911	0.7756	0.9324

Table 3: Results on the XSum dataset under different training and testing conditions. Rows indicate increasing adversarial training levels based on DIPPER-based paraphrasing attack. Columns compare detectors trained on benign data from different generators. GPT-5.1 serves as a held-out adversarial paraphrasing attack method.

Augmentation (%)	Accuracy	Macro-F1	AUC
0%	0.7695	0.7560	0.9386
25%	0.9078	0.9127	0.9887
50%	0.9348	0.9378	0.9893
75%	0.9396	0.9427	0.9919
100%	0.9475	0.9514	0.9947

Table 4: Ablation study on the amounts of augmented data. All models are evaluated on the Level-1 test set.

Level-1 augmentation consistently improve robustness in both training settings. AUC exceeds 0.94 in both cases, indicating strong cross-adversarial generalization.

6.5 Ablation Study

Next, we analyze the effect of mixing benign and adversarial samples during training. As shown in Table 4, increasing the proportion of adversarial samples leads to consistent performance improvements, confirming the benefit of adversarial exposure. Notably, the best results are obtained under the 100% (Level-1) setting, where 100% adversarial data is added on top of the benign data, combining benign samples with a diverse set of single-step adversarial paraphrases. This balanced training strategy prevents over-specialization to ei-

ther purely benign or purely adversarial distributions, enabling the model to learn more robust and broadly applicable representations that generalize well across both benign and adversarial evaluation settings.

7 Conclusions

In this work, we study human–AI co-authorship detection under adversarial rewriting and propose a unified framework that explicitly models different degrees of AI involvement. By formulating detection as a four-class classification task, our framework moves beyond binary human-AI distinctions and enables fine-grained analysis of realistic AI-assisted writing scenarios. Extensive experiments show that adversarial data augmentation consistently improves robustness and generalization to out-of-distribution settings. We further find that a balanced mixture of benign and adversarial samples is sufficient to achieve robust performance. Overall, our results highlight the importance of principled task formulation, data augmentation design, and evaluation protocols for building reliable detectors in increasingly prevalent human–AI collaborative writing settings.

Limitations

Our experiments involve a limited set of LLM generators and adversarial paraphraser due to practical constraints such as API availability, cost, and output stability. Although we include cross-generator evaluation and held-out models to reduce overfitting to a single generator, fully covering the rapidly evolving landscape of large language models remains challenging.

In addition, real-world human–AI writing can involve more complex interaction patterns, such as multiple rounds of revision and interleaved human and AI contributions, which are difficult to simulate at scale. Our adversarial paraphrasing setup introduces diverse, meaning-preserving surface variations, but it does not fully capture interactive or human-driven editing dynamics. Exploring richer simulation frameworks for human–AI collaboration remains an important direction for future work.

Ethics Statement

This work investigates the detection of human–AI co-authored text under varying degrees of AI involvement, including benign rewriting and adversarial paraphrasing. All experiments are conducted using publicly available datasets derived from human-written and model-generated text, and all data usage complies with the original dataset licenses. No new data involving human subjects is collected, and the datasets do not contain private or personally identifiable information.

This research has wide potential societal impact. First, our discrete AI involvement level classification framework provides support for text provenance detection, potentially informing discussions on copyright and authorship attribution. Second, we aim to improve the robustness and transparency of characterization of human–AI co-authored texts, a type of text that has become increasingly popular on the internet. At the same time, we acknowledge that detection and paraphrasing techniques could be misused by malicious users to, for example, evade detection systems, and the scope of this project does not include the defense to every potential attack. We emphasize that the proposed methods are intended for research purposes only and encourage any future research building upon this work to continue to consider the ethical implications.

References

- Mervat Abassy, Kareem Elozeiri, Alexander Aziz, Minh Ngoc Ta, Raj Vardhan Tomar, Bimarsha Adhikari, Saad El Dine Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, Jonibek Mansurov, Ekaterina Artemova, Vladislav Mikhailov, Rui Xing, Jiahui Geng, Hasan Iqbal, Zain Muhammad Mujahid, Tarek Mahmoud, Akim Tsvigun, and 5 others. 2024. [LLM-DetectAIve: a tool for fine-grained machine-generated text detection](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 336–343. Association for Computational Linguistics.
- Anthropic. 2025. [Introducing claude 4](#).
- 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](#). In *The Twelfth International Conference on Learning Representations*.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. [Longformer: The long-document transformer](#). *CoRR*, abs/2004.05150.
- Rahul Bhagat and Eduard Hovy. 2013. [Squibs: What is a paraphrase?](#) *Computational Linguistics*, 39(3):463–472.
- Yize CHENG, Vinu Sankar Sadasivan, Mehrdad Saberi, Shoumik Saha, and Soheil Feizi. 2025. [Adversarial paraphrasing: A universal attack for humanizing AI-generated text](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Zihao Cheng, Li Zhou, Feng Jiang, Benyou Wang, and Haizhou Li. 2025. [Beyond binary: Towards fine-grained LLM-generated text detection via role recognition and involvement measurement](#). In *Proceedings of the ACM on Web Conference 2025*, WWW ’25, pages 2677–2688. Association for Computing Machinery.
- 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.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. [GLTR: statistical detection and visualization of generated text](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*, pages 111–116. Association for Computational Linguistics.

- Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. [How close is chatgpt to human experts? comparison corpus, evaluation, and detection](#). *Preprint*, arXiv:2301.07597.
- Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. [RADAR: Robust AI-text detection via adversarial learning](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Thomas Huber and Christina Niklaus. 2025. [CLEAR: A comprehensive linguistic evaluation of argument rewriting by large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19548–19568. Association for Computational Linguistics.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. [Bag of tricks for efficient text classification](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.
- 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](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Laida Kushnareva, Tatiana Gaintseva, German Magai, Serguei Barannikov, Dmitry Abulkhanov, Kristian Kuznetsov, Eduard Tulchinskii, Irina Piontkovskaya, and Sergey Nikolenko. 2024. [AI-generated text boundary detection with RoFT](#). *Preprint*, arXiv:2311.08349.
- Ryan Law, Xibei Guan, and Tim Soulo. 2025. [74% of New Webpages Include AI Content \(Study of 900k Pages\)](#).
- Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024a. [MAGE: machine-generated text detection in the wild](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, *ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 36–53. Association for Computational Linguistics.
- Yafu Li, Zhilin Wang, Leyang Cui, Wei Bi, Shuming Shi, and Yue Zhang. 2024b. [Spotting AI’s touch: Identifying LLM-paraphrased spans in text](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7088–7107. Association for Computational Linguistics.
- Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D. Manning, and James Zou. 2025. [Quantifying large language model usage in scientific papers](#). *Nature Human Behaviour*, 9(12):2599–2609.
- 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](#). *CoRR*, abs/1907.11692.
- Elyas Masrour, Bradley N. Emi, and Max Spero. 2025. [DAMAGE: Detecting adversarially modified AI generated text](#). In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, pages 120–133. International Conference on Computational Linguistics.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. [DetectGPT: Zero-shot machine-generated text detection using probability curvature](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. [A corpus and cloze evaluation for deeper understanding of commonsense stories](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849. Association for Computational Linguistics.
- Ramya Namuduri, Yating Wu, Anshun Asher Zheng, Manya Wadhwa, Greg Durrett, and Junyi Jessy Li. 2025. [QUDsim: Quantifying discourse similarities in LLM-generated text](#). In *Second Conference on Language Modeling*.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807. Association for Computational Linguistics.
- OpenAI. 2024. [GPT-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI. 2025. [GPT-5.1: A smarter, more conversational ChatGPT](#).
- Jose Luis Paredes, Ethan Smith, Gregory Druck, and Bevin Benson. 2025. [More Articles Are Now Created by AI Than Humans](#).
- Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhatlacharya, Mobin Javed, and Bimal Viswanath. 2023. [Deepfake text detection: Limitations and opportunities](#). In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*, pages 1613–1630. IEEE.

- Juan Diego Rodriguez, Todd Hay, David Gros, Zain Shamsi, and Ravi Srinivasan. 2022. [Cross-domain detection of GPT-2-generated technical text](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1213–1233. Association for Computational Linguistics.
- Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2025. [Can AI-generated text be reliably detected? stress testing AI text detectors under various attacks](#). *Transactions on Machine Learning Research*.
- Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. [Masked language model scoring](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712. Association for Computational Linguistics.
- Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. 2024. [Red teaming language model detectors with language models](#). *Transactions of the Association for Computational Linguistics*, 12:174–189.
- Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023. [DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12395–12412. Association for Computational Linguistics.
- Zhixiong Su, Yichen Wang, Herun Wan, Zhaohan Zhang, and Minnan Luo. 2025. [HACo-Det: A study towards fine-grained machine-generated text detection under human-AI coauthoring](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22015–22036. Association for Computational Linguistics.
- Katherine Thai, Bradley Emi, Elyas Masrour, and Mohit Iyyer. 2026. [EditLens: Quantifying the extent of AI editing in text](#). In *The Fourteenth International Conference on Learning Representations*.
- Nafis Irtiza Tripto, Saranya Venkatraman, Dominik Macko, Robert Moro, Ivan Srba, Adaku Uchendu, Thai Le, and Dongwon Lee. 2024. [A ship of theseus: Curious cases of paraphrasing in LLM-generated texts](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6608–6625. Association for Computational Linguistics.
- Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. 2024. [LLM-as-a-Coauthor: Can mixed human-written and machine-generated text be detected?](#) *Preprint*, arXiv:2401.05952.
- Xuandong Zhao, Yu-Xiang Wang, and Lei Li. 2023. [Protecting language generation models via invisible watermarking](#). *Preprint*, arXiv:2302.03162.
- Ying Zhou, Ben He, and Le Sun. 2024. [Humanizing machine-generated content: Evading ai-text detection through adversarial attack](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 8427–8437. ELRA and ICCL.

A Prompt Examples

We provide representative prompt examples used to construct light and heavy paraphrases, as well as fully generated text. For clarity, we present abstracted prompt templates and illustrate them using news articles (XSum), which are the primary domain analyzed in our experiments. Across all datasets, prompts share the same rewriting objectives and constraints, with only minor adaptations for domain-specific writing style. Adversarial paraphrases generated with GPT-5.1 are used solely for evaluation and are excluded from training.

A.1 Low AI-involved

This setting restricts the model to surface realization only, enforcing a fixed sentence plan and a strict one-to-one sentence mapping while preserving all factual content.

PROMPT.

You are a professional news journalist. Rewrite the article sentence by sentence with a strict one-to-one mapping. Preserve all facts, entities, dates, and numerical values. Use a formal news style and keep sentence length within a small deviation from the original. Do not merge or split sentences. Return only the rewritten article.

A.2 High AI-involved

This setting allows the model to perform generative sentence planning and surface realization under a fixed content plan, enabling substantial restructuring of sentences and paragraphs while preserving all factual content.

PROMPT.

You are a creative news journalist. Heavily rewrite the article by restructuring sentences and paragraphs as needed. You may merge or split sentences and vary

phrasing and discourse structure, while preserving all factual content. Keep the overall length within a moderate range of the original. Return only the rewritten article.

A.3 Fully AI-generated

In this setting, the model autonomously performs content planning, determining how to expand a brief introduction into a complete article by selecting, organizing, and elaborating content.

PROMPT.

You are a professional news journalist. Given a brief news introduction or summary, expand it into a complete news article. Add relevant background and context while maintaining a formal news style. Do not repeat the original introduction verbatim. Return only the generated article.

A.4 Adversarial Paraphrasing with GPT-5.1

Adversarial paraphrasing evaluates robustness under a held-out generator by introducing surface variation while preserving meaning.

PROMPT.

You are a professional editor. Paraphrase the following text to sound natural and fluent, as if written by a human, while preserving the original meaning and all factual details. Return only the paraphrased text.

B Representation Analysis

Figure 3 illustrates a PCA visualization of representations extracted from the fine-tuned Longformer model under a 10-class formulation on GPT-generated XSum data. Each class corresponds to a specific combination of AI-involvement level and adversarial strength, including benign samples, Level-1 and Level-2 DIPPER paraphrases, resulting in ten distinct data variants.

As shown in Fig. 3(a), the learned representation space exhibits a clear directional structure. Samples generated directly by GPT and those produced via DIPPER paraphrasing tend to separate along different principal directions, indicating that adversarial paraphrasing induces systematic, directionally consistent shifts rather than random perturbations. This suggests that although DIPPER is designed as a paraphrasing attack model, it introduces

its own stylistic regularities and does not fully emulate the diversity of genuine human-written text. Consequently, DIPPER-based samples form coherent clusters instead of smoothly interpolating within the human-writing distribution.

We further compare two alternative strategies for organizing class boundaries in the representation space. The red boundaries in Fig. 3(a) attempt to reassign adversarially paraphrased samples across AI-involvement levels based on assumed paraphrasing strength (e.g., treating twice-paraphrased high-AI samples as low-AI). However, this strategy does not align well with the intrinsic geometry learned by the model: clusters intersect multiple red regions and fail to exhibit coherent separation. This mismatch indicates that paraphrasing strength alone is insufficient to redefine AI-involvement categories in the learned feature space.

In contrast, Fig. 3(b) shows the blue-boundary formulation, which groups samples according to their original AI-involvement labels while allowing adversarial paraphrases to expand the surface realizations within each group. Under this formulation, clusters are more consistently aligned with decision boundaries, suggesting that mixing benign and adversarial samples encourages the model to learn broader, generator-agnostic representations rather than overfitting to paraphrase-specific artifacts. This observation provides qualitative support for our data augmentation strategy and helps explain why single-step adversarial training improves robustness without requiring aggressive multi-step paraphrasing.

C Quantitative Characterization

We conduct quantitative analyses to validate the construction of benign AI-involvement degrees and to characterize how adversarial paraphrasing departs from this benign continuum. The analysis aims to distinguish natural variation induced by increasing AI involvement from additional surface-level perturbations introduced by adversarial rewriting.

Length analysis. We first examine document length to assess whether the generated classes can be trivially distinguished by length. Table 5 reports document-length statistics across datasets and generators. For XSum and arXiv, the generated texts are broadly comparable in length to their corresponding human-written sources, suggesting that length alone is unlikely to provide a dominant class

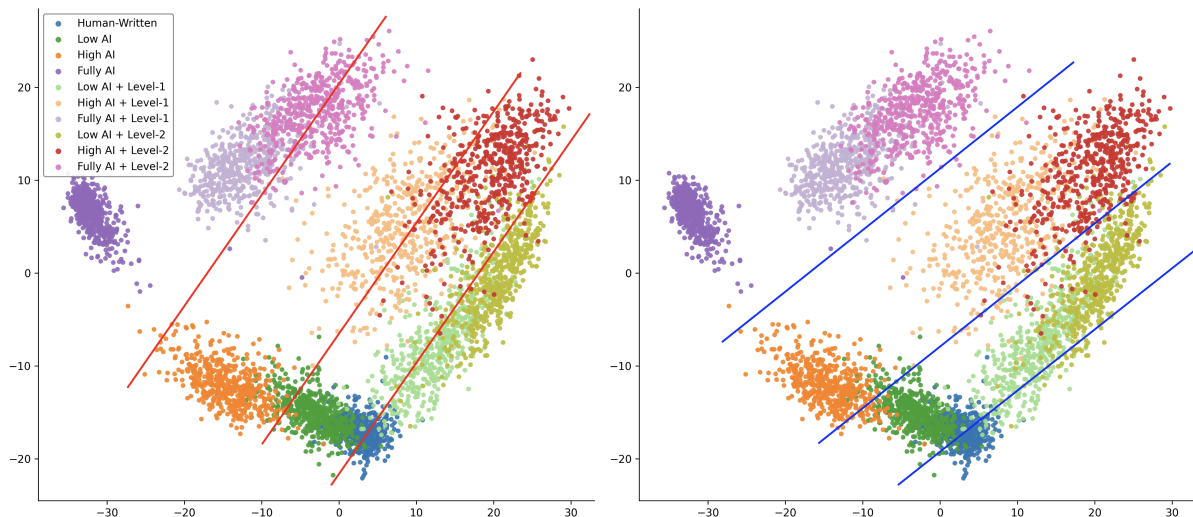


Figure 3: PCA visualization of document representations from a Longformer detector fine-tuned under a 10-class formulation. Each class corresponds to a benign or adversarial data variant treated as a separate label. The separation patterns illustrate that adversarial paraphrases occupy distinct directions in representation space rather than extending the benign AI-involvement continuum.

cue in these domains. ROCStory is the main exception. Because the source documents contain only five sentences, strict word-count constraints led to unnatural truncation during generation. We therefore allowed greater length flexibility for the High AI-involved and Fully AI-generated classes. This results in larger length ratios for these classes, particularly for ROCStory. Values are computed over aligned source-generation pairs.

Similarity analysis. We next examine whether the benign AI-involvement categories exhibit the intended progression in surface form while preserving semantic content, and how adversarial paraphrasing departs from this progression. We report three complementary similarity measures. Jaccard similarity captures lexical overlap based on word-level token sets. BLEU measures n-gram overlap at the sentence level with smoothing. Embedding cosine similarity measures semantic relatedness using representations from a pretrained Longformer encoder. Figure 4 summarizes similarity trends computed on aligned sample pairs drawn from the same underlying source documents, with mean values reported and shaded bands indicating the standard error of the mean (SEM).

Figure 4a compares benign AI-involved texts with their corresponding human-written sources. Across all three metrics, similarity decreases monotonically as AI involvement increases. Lexical and n-gram overlap drop substantially from Low to

High AI involvement, while embedding similarity decreases more gradually. This pattern indicates increasing divergence in surface form while semantic content remains comparatively stable, consistent with the intended progression of AI editing strength.

Figure 4b compares adversarial paraphrases with their benign counterparts within the same AI-involvement level, as well as Adv Moderate versus Adv Strong. Relative to benign text, both adversarial settings lead to pronounced additional drops in Jaccard and BLEU, indicating substantial perturbations at the lexical and local phrasal levels. In contrast, embedding cosine similarity remains close to that of benign samples, with only minor decreases even under multi-step paraphrasing. This pattern is consistent with the design of DIPPER, which aims to evade surface-form-based detection by introducing controlled lexical substitutions and local reordering while preserving the underlying meaning. Moreover, the limited gap between Adv Moderate and Adv Strong in embedding similarity suggests diminishing semantic changes from additional paraphrasing steps. Together, these results indicate that DIPPER paraphrases introduce an extra surface-level shift that is not aligned with the natural progression of AI involvement, supporting our choice to treat DIPPER outputs as adversarial augmentations rather than redefining AI-involvement categories.

Dataset	Gen.	Human words	Low AI/Human	High AI/Human	Fully AI/Human
XSum	GPT	629.6 $\pm$ 285.0	0.967 $\pm$ 0.034	0.838 $\pm$ 0.151	1.012 $\pm$ 0.235
XSum	Claude	629.6 $\pm$ 285.0	0.984 $\pm$ 0.028	0.934 $\pm$ 0.128	1.206 $\pm$ 0.221
arXiv	GPT	758.3 $\pm$ 344.3	0.964 $\pm$ 0.026	0.884 $\pm$ 0.070	1.100 $\pm$ 0.244
arXiv	Claude	758.3 $\pm$ 344.3	0.944 $\pm$ 0.094	0.809 $\pm$ 0.098	0.955 $\pm$ 0.169
ROCStory	GPT	61.9 $\pm$ 4.2	1.079 $\pm$ 0.097	1.910 $\pm$ 0.249	1.659 $\pm$ 0.206
ROCStory	Claude	61.9 $\pm$ 4.2	1.180 $\pm$ 0.115	2.717 $\pm$ 0.388	1.760 $\pm$ 0.218

Table 5: Length statistics across datasets and generators. Human words reports the word count of the source texts, while Low AI/Human, High AI/Human, and Fully AI/Human report generated-to-source word-count ratios. Values are mean $\pm$ standard deviation over aligned source-generation pairs.

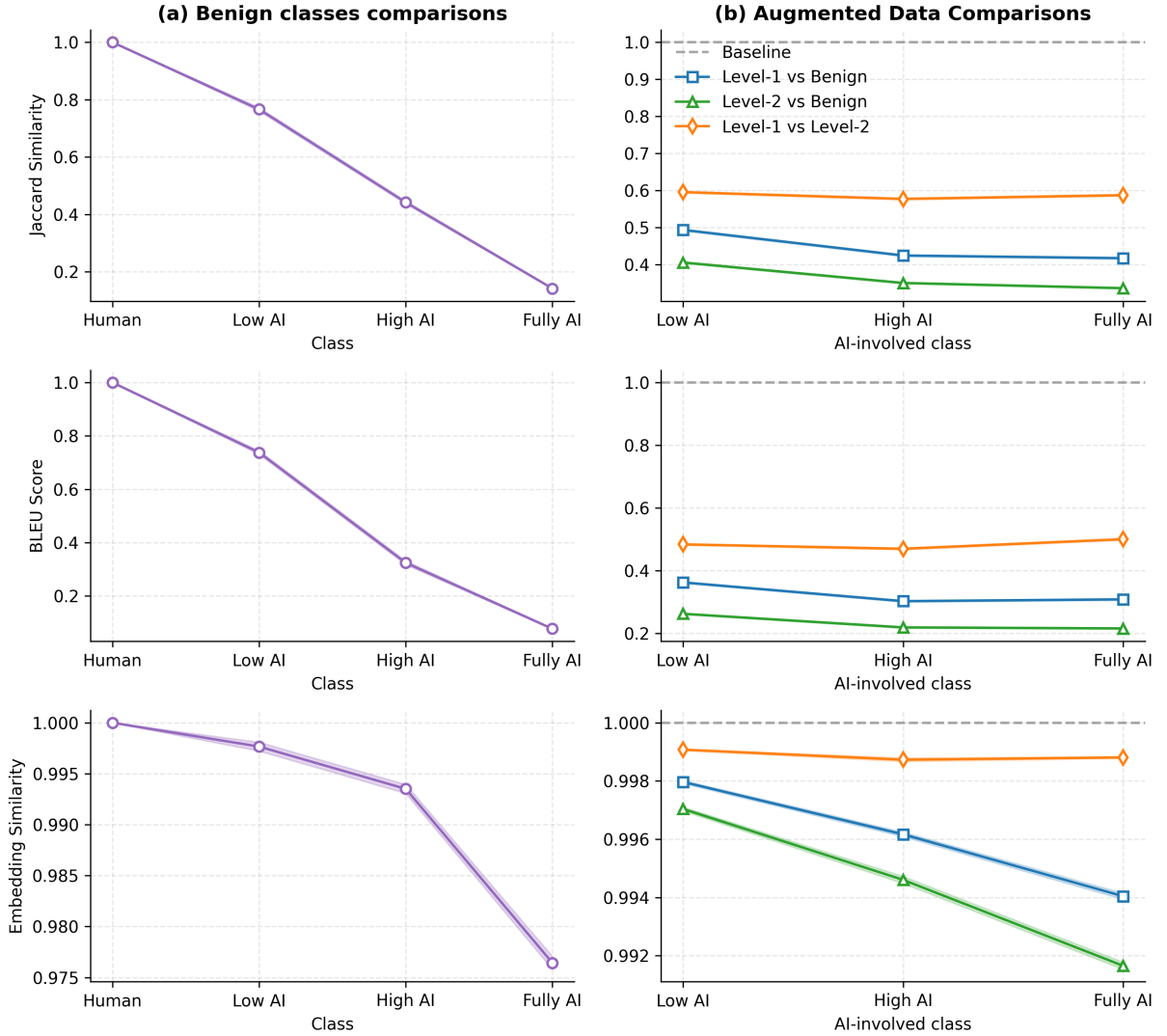


Figure 4: Quantitative similarity analysis using lexical (Jaccard), n-gram (BLEU), and semantic (embedding cosine) metrics. (a) Similarity between benign AI-involved texts and their corresponding human-written sources decreases monotonically with increasing AI involvement. (b) Adversarial paraphrases introduce substantial additional surface-level divergence from their benign counterparts while largely preserving semantic similarity.

Data	Setting	Original (%)	Recalibrated (%)
XSum-GPT-4o	Benign	47.8	95.8
XSum-GPT-4o	Level-1	51.9	84.1
XSum-Claude	Benign	48.6	92.8
XSum-Claude	Level-1	51.9	78.9

Table 6: Alignment between EditLens similarity-based categories and our process-defined AI-involvement labels using the original and recalibrated thresholds.

Data	Test	EditLens			Ours		
		Acc.	F1	AUC	Acc.	F1	AUC
XSum-GPT-4o	Benign	0.950	0.950	0.991	0.978	0.978	0.998
XSum-GPT-4o	Level-1	0.879	0.877	0.979	0.938	0.943	0.991
XSum-GPT-4o	GPT-5.1	0.808	0.802	0.955	0.872	0.872	0.964
XSum-Claude	Benign	0.861	0.858	0.984	0.975	0.975	0.999
XSum-Claude	Level-1	0.778	0.792	0.940	0.943	0.944	0.996
XSum-Claude	GPT-5.1	0.747	0.721	0.935	0.803	0.785	0.953

Table 7: Controlled comparison with EditLens under Level-1 training using the same RoBERTa-large backbone.

D Additional Experiment Results

Class-wise sensitivity and where robustness improves. Per-class F1 shows that the robustness gains are driven mainly by the ambiguous Low AI and High AI classes. On XSum (Table 8), benign-only training degrades sharply under paraphrased testing (e.g., on GPT-4o-generated XSum, High AI drops from 0.9888 under benign testing to 0.6567/0.4916 under Level-1/Level-2 testing), and transfer to GPT-5.1 can be especially brittle. After incorporating Level-1 adversarial data during training, Low AI and High AI become markedly more stable across testing levels and the held-out generator, indicating improved robustness to surface variation.

In contrast, the Fully AI class is consistently the easiest: its F1 remains near-saturated across conditions, suggesting that strong fully-synthetic signals persist even after paraphrasing. The Human class also benefits from adversarial exposure, but the dominant robustness gains come from reducing confusion between Low AI and High AI under distribution shift.

Similar patterns hold across domains: improvements on ROCStory concentrate in Human and Low AI, whereas arXiv again improves most for Low AI and High AI. Overall, single-step adversarial training primarily strengthens the detector

on the most confusable intermediate classes, with diminishing returns from additional multi-step paraphrasing.

Comparison with continuous scoring. We further compare our formulation with EditLens (Thai et al., 2026), a representative continuous estimation approach for fine-grained human-AI detection, on XSum using GPT-4o and Claude generations. We focus on XSum to avoid input-length truncation effects and use the released RoBERTa-large implementation of EditLens.

Since EditLens defines its categories using similarity-based thresholds, we recalibrate the thresholds on our training data to improve alignment with our process-defined labels (Table 6). We then compare the two methods under the same RoBERTa-large backbone. As shown in Table 7, our formulation achieves consistently higher performance across the evaluated benign, adversarial, and held-out rewriting settings.

Dataset: XSum								
Settings		F1 per Class				Overall Metrics		
Train	Test	Human	Low AI	High AI	Fully AI	Acc	Macro F1	AUC
(a) Same generator and same attack.								
Part I: Benign AI-involved Data generated by GPT-4o								
Benign	Benign	0.9775	0.9721	0.9888	0.9836	0.9805	0.9805	0.9998
	Level-1	0.6824	0.7151	0.6567	0.9697	0.7695	0.7560	0.9386
	Level-2	0.4833	0.6131	0.4916	0.9586	0.6652	0.6367	0.8988
	GPT-5.1	0.6905	0.4767	0.0000	0.9836	0.6212	0.5377	0.8413
Level-1	Benign	0.9773	0.9780	0.9945	0.9944	0.9861	0.9861	0.9990
	Level-1	0.9773	0.9143	0.9223	0.9916	0.9475	0.9514	0.9947 ↑
	Level-2	0.9773	0.8931	0.8980	0.9907	0.9321	0.9398	0.9921 ↑
	GPT-5.1	0.9829	0.7089	0.7864	0.9944	0.8691	0.8682	0.9662 ↑
Level-2	Benign	0.9551	0.9379	0.9724	0.9890	0.9638	0.9636	0.9973
	Level-1	0.9551	0.9070	0.9194	0.9835	0.9396	0.9413	0.9920
	Level-2	0.9551	0.8863	0.8936	0.9817	0.9244	0.9292	0.9890
	GPT-5.1	0.9770	0.6892	0.8056	1.0000	0.8719	0.8680	0.9554
Part II: Benign AI-involved Data generated by Claude Sonnet 4								
Benign	Benign	0.9663	0.9667	0.9831	0.9836	0.9749	0.9729	0.9979
	Level-1	0.4873	0.6689	0.6567	0.9053	0.6789	0.6796	0.8999
	Level-2	0.3127	0.5217	0.4916	0.8445	0.5373	0.5426	0.8326
	GPT-5.1	0.4886	0.0000	0.0000	0.9890	0.4903	0.3694	0.7882
Level-1	Benign	0.9609	0.9663	0.9718	0.9783	0.9694	0.9693	0.9948
	Level-1	0.9609	0.9395	0.9282	0.9730	0.9491	0.9504	0.9927 ↑
	Level-2	0.9609	0.9151	0.8995	0.9711	0.9321	0.9367	0.9890 ↑
	GPT-5.1	0.9829	0.5865	0.7032	0.9424	0.8134	0.8038	0.9433 ↑
Level-2	Benign	0.9565	0.9480	0.9659	0.9730	0.9610	0.9609	0.9960
	Level-1	0.9565	0.9235	0.9076	0.9549	0.9332	0.9356	0.9926
	Level-2	0.9514	0.9002	0.8780	0.9490	0.9143	0.9197	0.9913 ↑
	GPT-5.1	0.9944	0.5231	0.6761	0.9091	0.7911	0.7756	0.9324
(b) Transfer across generators								
Train: GPT → Test: Claude								
Benign	Claude	0.9255	0.9176	0.9551	0.9560	0.9387	0.9386	0.9960
Level-1		0.8912	0.8673	0.8989	0.9600	0.9062	0.9043	0.9898
Level-2		0.8718	0.8429	0.8662	0.9586	0.8877	0.8849	0.9833
Train: Claude → Test: GPT								
Benign	GPT-4o	0.9198	0.9070	0.8750	0.8945	0.8997	0.8991	0.9933
Level-1		0.9198	0.8703	0.8864	0.9570	0.9078	0.9084	0.9825
Level-2		0.9362	0.8750	0.8544	0.9385	0.8954	0.9010	0.9816

Table 8: Results on the XSum dataset. Upward arrows (↑) indicate improved AUC relative to the immediately preceding training level under the same testing condition.

Dataset: ROCStory								
Settings		F1 per Class				Overall Metrics		
Train	Test	Human	Low AI	High AI	Fully AI	Acc	Macro F1	AUC
(a) Same generator and same attack.								
Part I: Benign AI-involved Data generated by GPT-4o								
Benign	Benign	0.9467	0.9468	0.9834	0.9889	0.9666	0.9665	0.9996
	Level-1	0.8122	0.8604	0.8829	0.9019	0.8712	0.8644	0.9793
	Level-2	0.6452	0.7649	0.8238	0.8631	0.7953	0.7743	0.9583
	GPT-5.1	0.8163	0.5000	0.4310	0.6496	0.6323	0.5992	0.9062
Level-1	Benign	0.8944	0.9026	0.9832	0.9836	0.9415	0.9410	0.9977
	Level-1	0.8889	0.9305	0.9435	0.9457	0.9332	0.9272	0.9943 ↑
	Level-2	0.8780	0.9355	0.9242	0.9307	0.9255	0.9171	0.9892 ↑
	GPT-5.1	0.8090	0.5139	0.5426	0.6742	0.6518	0.6349	0.9249 ↑
Level-2	Benign	0.9157	0.9043	0.9497	0.9622	0.9331	0.9330	0.9931
	Level-1	0.9048	0.9286	0.9226	0.9284	0.9237	0.9211	0.9904
	Level-2	0.8994	0.9375	0.8992	0.9069	0.9132	0.9108	0.9876
	GPT-5.1	0.8172	0.4511	0.4667	0.6380	0.6212	0.5933	0.8983
Part II: Benign AI-involved Data generated by Claude Sonnet 4								
Benign	Benign	1.0000	1.0000	0.9945	0.9944	0.9972	0.9972	1.0000
	Level-1	0.8357	0.8869	0.9069	0.8987	0.8871	0.8821	0.9848
	Level-2	0.6642	0.7872	0.8251	0.8275	0.7920	0.7760	0.9611
	GPT-5.1	0.8357	0.1443	0.9480	0.7574	0.7437	0.6714	0.9250
Level-1	Benign	0.9770	0.9783	1.0000	1.0000	0.9889	0.9888	1.0000
	Level-1	0.9605	0.9694	0.9749	0.9587	0.9666	0.9659	0.9973 ↑
	Level-2	0.9290	0.9497	0.9605	0.9324	0.9455	0.9429	0.9951 ↑
	GPT-5.1	0.8854	0.5079	0.9831	0.8072	0.8189	0.7959	0.9453 ↑
Level-2	Benign	0.9405	0.9474	0.9890	0.9888	0.9666	0.9664	1.0000
	Level-1	0.9349	0.9565	0.9699	0.9551	0.9571	0.9541	0.9966
	Level-2	0.9294	0.9510	0.9617	0.9318	0.9466	0.9435	0.9944
	GPT-5.1	0.8927	0.5401	0.9545	0.7719	0.8022	0.7898	0.9483 ↑
(b) Transfer across generators								
Train: GPT → Test: Claude								
Benign	Claude	0.9467	0.9362	0.9836	0.9888	0.9638	0.9638	0.9990
Level-1		0.8944	0.8920	0.9091	0.9171	0.9046	0.9031	0.9902
Level-2		0.9157	0.9049	0.8946	0.9062	0.9032	0.9054	0.9849
Train: Claude → Test: GPT								
Benign	GPT-4o	0.9368	0.8953	0.9286	0.9362	0.9248	0.9242	0.9944
Level-1		0.8808	0.8939	0.8528	0.8871	0.8792	0.8786	0.9820
Level-2		0.8977	0.8981	0.8421	0.8454	0.8665	0.8708	0.9785

Table 9: Results on the ROCStory dataset. Upward arrows (↑) indicate improved AUC relative to the immediately preceding training level under the same testing condition.

Dataset: arXiv								
Settings		F1 per Class				Overall Metrics		
Train	Test	Human	Low AI	High AI	Fully AI	Acc	Macro F1	AUC
(a) Same generator and same attack.								
Part I: Benign AI-involved Data generated by GPT-4o								
Benign	Benign	0.8235	0.6000	0.6226	0.8547	0.7286	0.7252	0.9382
	Level-1	0.5344	0.3938	0.4405	0.8544	0.5673	0.5558	0.8343
	Level-2	0.3196	0.2958	0.3363	0.7509	0.4329	0.4257	0.7437
	GPT-5.1	0.8333	0.5741	0.6602	0.9709	0.7538	0.7596	0.9285
Level-1	Benign	0.6111	0.3218	0.5645	0.8696	0.6080	0.5918	0.8870
	Level-1	0.6111	0.3467	0.5537	0.8547	0.6160	0.5916	0.8923 ↑
	Level-2	0.6111	0.4071	0.5674	0.8721	0.6393	0.6144	0.8964 ↑
	GPT-5.1	0.6197	0.2791	0.6364	0.9174	0.6332	0.6132	0.8761
Level-2	Benign	0.7595	0.4884	0.6720	0.9259	0.7186	0.7115	0.9230 ↑
	Level-1	0.7595	0.4868	0.6218	0.8734	0.6905	0.6854	0.9148 ↑
	Level-2	0.7595	0.5022	0.6041	0.8451	0.6794	0.6777	0.9133 ↑
	GPT-5.1	0.7595	0.4000	0.6667	0.9524	0.7035	0.6947	0.9050 ↑
Part II: Benign AI-involved Data generated by Claude Sonnet 4								
Benign	Benign	0.9091	0.8842	0.9505	0.9709	0.9296	0.9287	0.9913
	Level-1	0.4265	0.5316	0.6358	0.8090	0.5931	0.6007	0.8274
	Level-2	0.2521	0.4038	0.4776	0.6552	0.4228	0.4472	0.7279
	GPT-5.1	0.6429	0.1231	0.0392	0.7042	0.5025	0.3774	0.8934
Level-1	Benign	0.8172	0.7312	0.8829	0.9901	0.8593	0.8554	0.9653
	Level-1	0.8172	0.7778	0.8341	0.9346	0.8481	0.8409	0.9625 ↑
	Level-2	0.8172	0.7927	0.7948	0.9102	0.8337	0.8287	0.9571 ↑
	GPT-5.1	0.7917	0.4691	0.6909	0.9009	0.7286	0.7132	0.9139 ↑
Level-2	Benign	0.8660	0.7778	0.8727	0.9703	0.8744	0.8717	0.9771 ↑
	Level-1	0.8660	0.8023	0.8372	0.9474	0.8653	0.8632	0.9698 ↑
	Level-2	0.8660	0.8090	0.8050	0.9177	0.8477	0.8494	0.9646 ↑
	GPT-5.1	0.7925	0.4500	0.6796	0.8991	0.7236	0.7053	0.9224 ↑
(b) Transfer across generators								
Train: GPT → Test: Claude								
Benign	Claude	0.7692	0.5349	0.6847	0.8727	0.7236	0.7154	0.9361
Level-1		0.5946	0.4403	0.6276	0.8761	0.6619	0.6346	0.9042
Level-2		0.7059	0.5067	0.6590	0.8772	0.7034	0.6872	0.9245
Train: Claude → Test: GPT								
Benign	GPT-4o	0.8738	0.6218	0.2456	0.7731	0.6784	0.6286	0.9230
Level-1		0.8172	0.7128	0.6912	0.8500	0.7593	0.7678	0.9251
Level-2		0.8485	0.6714	0.6287	0.8211	0.7174	0.7424	0.9084

Table 10: Results on the Arxiv dataset. Upward arrows (↑) indicate improved AUC relative to the immediately preceding training level under the same testing condition.