# Literature Survey

*Multilingual Content Integrity and Authorship Intelligence Platform:*
*An Ensemble Transformer Approach for AI-Generated Text Detection,*
*Semantic Plagiarism Identification, and Adaptive Content Humanization*

## 1. Introduction

Over the past few years, large language models have become remarkably good at producing text that reads as though a person wrote it. Tools like ChatGPT, Gemini, and Claude can draft essays, reports, and even research papers in seconds. This has raised genuine concerns in education, journalism, and publishing, because telling human writing apart from machine output is no longer straightforward. At the same time, plagiarism has evolved beyond simple copy-paste; students and writers now paraphrase or lightly reword existing sources, making traditional string-matching detectors far less effective.

Our project sits at the intersection of three closely related problems: detecting whether a piece of text was generated by an AI model, identifying whether it was plagiarised from existing sources, and transforming flagged content so that it reads naturally. Each of these areas has seen significant research activity in the last two to three years, and this literature survey reviews the key papers that inform our approach. We organise the discussion into four broad themes: benchmarks and datasets for AI text detection, detection methods and their limitations, sentence-level semantic similarity for plagiarism identification, and text rewriting techniques for content humanization.

## 2. Benchmarks and Datasets for AI Text Detection

A recurring challenge in this field has been the lack of standardised evaluation. Early detectors were often tested on small, in-house datasets, making it hard to compare results across studies. Dugan and colleagues addressed this gap with RAID, a shared benchmark containing over six million generated documents spanning eleven different language models, eight subject domains, and eleven adversarial attack strategies [1]. What makes RAID particularly valuable is its scale and diversity; it forces detectors to generalise rather than overfit to a single generator or writing style. The benchmark was presented at ACL 2024 and has since become a reference point for evaluating new detection systems.

While RAID focuses primarily on English, the M4 dataset introduced by Wang et al. tackles the multilingual dimension [2]. M4 covers multiple generators and multiple languages, reflecting the reality that AI-generated content is not an English-only phenomenon. Their work showed that detectors trained exclusively on English data perform poorly when applied to other languages, which motivated our decision to include XLM-RoBERTa in the detection ensemble.

Another widely used resource is HC3, the Human ChatGPT Comparison Corpus, assembled by Guo et al. [3]. HC3 pairs human-written answers with ChatGPT responses across several question-answering domains. The paired structure is useful for training classifiers because it provides direct contrastive examples. However, HC3 is limited to a single generator, so it works best as a complement to broader benchmarks like RAID and M4 rather than as a standalone training set.

More recently, the FAIDSet framework proposed a fine-grained labelling scheme that goes beyond the binary human-or-AI distinction [11]. Instead of treating detection as a two-class problem, FAIDSet introduces a third category for mixed content, where parts of a document are human-written and parts are machine-generated. This is closer to how AI tools are actually used in practice, since many writers use language models to draft certain paragraphs while writing others themselves. The accompanying multi-task learning approach with contrastive objectives showed promising results on this harder classification task.

## 3. AI-Generated Text Detection Methods

The most common approach to detecting AI-generated text is to fine-tune a pre-trained transformer on labelled examples. He et al. introduced DeBERTa, which uses a disentangled attention mechanism that separately encodes content and position information [9]. This architectural choice gives DeBERTa an edge on tasks that require fine-grained token-level understanding, and it has consistently ranked among the top models on natural language understanding benchmarks. In our project, DeBERTa-v3-large serves as the primary classifier in the detection ensemble.

Tang et al. provided a thorough survey of detection methods, grouping them into statistical approaches, neural classifiers, and watermarking-based techniques [5]. Their analysis highlighted that no single method dominates across all settings.

Statistical methods like perplexity scoring work well for older generators but struggle with newer models that produce more fluent text. Neural classifiers achieve higher accuracy but can be brittle when the test distribution shifts away from the training data. This finding reinforced our choice to use an ensemble of four different transformer architectures rather than relying on any single model.

Sadasivan et al. raised important questions about the fundamental reliability of AI text detectors [4]. Their experiments demonstrated that simple paraphrasing attacks can dramatically reduce detection accuracy, sometimes bringing it close to random chance. They argued that as language models improve, the statistical gap between human and machine text will continue to shrink, making detection inherently harder. This work was a wake-up call for the community and directly influenced our decision to include adversarial robustness testing in the evaluation pipeline.

On the multilingual front, Hlavnova and Pikuliak explored fine-tuning strategies for detecting machine-generated text across languages as part of SemEval-2024 Task 8 [10]. They found that LoRA-adapted RoBERTa models combined with majority voting across language-specific heads could achieve competitive performance without training separate models for each language. Their results confirmed that parameter-efficient fine-tuning is a practical path for multilingual detection, which aligns with our use of XLM-RoBERTa-large trained on the M4 corpus.

## 4. Semantic Similarity for Plagiarism Detection

Traditional plagiarism detection relies on n-gram overlap or fingerprinting techniques, which work well for verbatim copying but miss paraphrased content entirely. Foltynek, Meuschke, and Gipp surveyed the landscape of plagiarism detection methods in their comprehensive review [12]. They categorised approaches into string-based, syntax-based, semantic-based, and citation-based methods, noting that semantic approaches were the most promising but also the least mature at the time of writing. Our project builds on this observation by placing sentence-level semantic embeddings at the core of the plagiarism detection module.

The foundation for our embedding-based approach comes from Sentence-BERT, proposed by Reimers and Gurevych [6]. By modifying the standard BERT architecture with Siamese and triplet network structures, they produced fixed-size sentence embeddings that can be compared using cosine similarity. This was a significant practical advance because it reduced the computational cost of pairwise comparison from quadratic to linear, making it feasible to compare a query document against a large reference corpus in reasonable time.

Gao, Yao, and Chen later improved on sentence embeddings with SimCSE, a contrastive learning framework that achieves better alignment and uniformity in the embedding space [7]. The unsupervised variant of SimCSE uses dropout as a minimal data augmentation strategy, which is elegant in its simplicity. The supervised variant leverages natural language inference pairs to push entailment examples closer together and contradiction examples further apart. We use SimCSE-enhanced embeddings alongside Sentence-BERT to improve the recall of our plagiarism detection pipeline, particularly for cases where the wording has been substantially changed but the meaning is preserved.

## 5. Text Rewriting and Content Humanization

The idea of rewriting text to evade detection is not new, but Krishna et al. brought it into sharp focus with their work on DIPPER, a discourse-level paraphraser [8]. DIPPER operates at the paragraph level with two controllable knobs: lexical diversity, which governs how much the vocabulary changes, and order diversity, which controls how much the sentence structure is rearranged. Their key finding was that even moderate paraphrasing could fool most existing detectors, but retrieval-based methods offered a more robust defence. For our project, DIPPER serves as one of several rewriting models in the humanization module, and the retrieval insight informed our decision to combine detection with semantic similarity search.

Beyond paraphrasing, the problem of authorship style transfer is relevant to understanding how writing characteristics can be altered while keeping the underlying content intact. Recent work on authorship style transfer using policy optimization techniques [13] showed that reinforcement learning signals can guide a language model to adopt a target writing style. While our project does not directly implement style transfer, the iterative feedback loop in our humanization module draws on a similar principle: the AI detection score acts as a reward signal that guides the rewriting process toward more human-like output.

Our humanization approach uses three sequence-to-sequence models in a fallback chain. Flan-T5-XL handles the initial rewriting attempt, PEGASUS-large provides an alternative with its abstractive summarization strengths, and Mistral-7B fine-tuned with QLoRA offers the most capable but resource-intensive option. The iterative feedback loop checks each rewritten version against the detection ensemble, and if the AI score remains above the threshold, the system increases the diversity parameters and tries again. This closed-loop design is what distinguishes our approach from one-shot paraphrasing tools.

## 6. Summary of Findings

Several themes emerge from this review. First, the field of AI text detection is maturing rapidly, with large-scale benchmarks like RAID and M4 enabling more rigorous evaluation than was possible even two years ago. Second, no single detection method is sufficient on its own; ensemble approaches that combine multiple architectures and training strategies offer the best generalisation. Third, semantic similarity methods have advanced to the point where paraphrase-level plagiarism can be reliably detected, though computational cost remains a concern for very large corpora. Fourth, text rewriting techniques have become sophisticated enough to evade most detectors, which creates both a challenge for detection and an opportunity for humanization applications.

Our project draws on all four of these threads. The detection module uses an ensemble of four transformer classifiers informed by the benchmarking work of Dugan et al. and the multilingual insights of Wang et al. The plagiarism module combines fast retrieval with the semantic embedding techniques of Reimers and Gurevych and Gao et al. The humanization module builds on the rewriting capabilities demonstrated by Krishna et al. and adds an iterative feedback mechanism that treats detection as a quality gate. Together, these components form a unified platform that addresses content integrity from multiple angles.

## References

[1] Dugan, L., Hwang, A., Trhlik, F., et al., “RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors,” Proceedings of the Association for Computational Linguistics (ACL), 2024. arXiv:2405.07940.

[2] Wang, Y., et al., “M4: Multi-generator, Multi-domain, Multi-lingual Black-Box Machine-Generated Text Detection,” 2024. arXiv:2305.14902.

[3] Guo, B., et al., “HC3: Human ChatGPT Comparison Corpus,” Hello-SimpleAI, Hugging Face, 2023.

[4] Sadasivan, V.S., et al., “On the Reliability of AI-Text Detectors,” 2023. arXiv:2304.02819.

[5] Tang, R., et al., “Detecting Machine-Generated Text: A Critical Survey,” 2023. arXiv:2303.07205.

[6] Reimers, N. and Gurevych, I., “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” Proceedings of EMNLP, 2019. arXiv:1908.10084.

[7] Gao, T., Yao, X., and Chen, D., “SimCSE: Simple Contrastive Learning of Sentence Embeddings,” Proceedings of EMNLP, 2021. arXiv:2104.08821.

[8] Krishna, K., et al., “Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval is an Effective Defense,” (DIPPER), 2023. arXiv:2303.13408.

[9] He, P., Liu, X., Gao, J., and Chen, W., “DeBERTa: Decoding-enhanced BERT with Disentangled Attention,” Proceedings of ICLR, 2021. arXiv:2006.03654.

[10] Hlavnova, E. and Pikuliak, M., “Fine-tuned LLMs for Multilingual Machine-Generated Text Detection,” SemEval-2024 Task 8. arXiv:2402.13671.

[11] “Fine-grained AI-generated Text Detection using Multi-task Auxiliary and Multi-level Contrastive Learning,” (FAIDSet), 2025. arXiv:2505.14271.

[12] Foltynek, T., Meuschke, N., and Gipp, B., “A Survey on Plagiarism Detection,” ACM Computing Surveys, 2019. arXiv:1703.05546.

[13] “Authorship Style Transfer with Policy Optimization,” 2024. arXiv:2403.08043.