Literature Survey Report (Minimum 10 Research Papers)

1. Paperwise Comparison Table

Title	Author	Year	Objective / Purpose	Methodology / Technique	Findings
RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors	Dugan, L., Hwang, A., Trhlik, F., et al.	2024	To create a large-scale shared benchmark for evaluating AI text detectors across multiple generators, domains, and adversarial attacks.	Compiled 6M+ generated documents from 11 language models across 8 domains with 11 adversarial attack strategies. Evaluated existing detectors on this unified benchmark.	Most existing detectors showed significant performance drops under adversarial attacks. The benchmark revealed that no single detector generalizes well across all generators and domains, highlighting the need for ensemble approaches.
M4: Multi-generator, Multi-domain, Multi-lingual Black-Box Machine-Generated Text Detection	Wang, Y., et al. (MBZUAI)	2024	To build a multilingual, multi-generator dataset and detection framework for cross-lingual AI text detection.	Collected machine-generated text from multiple LLMs across multiple languages and domains. Trained and evaluated detection models in both monolingual and cross-lingual settings.	Detectors trained exclusively on English data perform poorly on other languages. Cross-lingual transfer learning with multilingual transformers like XLM-RoBERTa significantly improves detection across languages.
HC3: Human ChatGPT Comparison Corpus	Guo, B., et al. (Hello-SimpleAI)	2023	To create a paired dataset of human-written and ChatGPT-generated answers for training and evaluating AI text detectors.	Collected question-answer pairs from multiple domains where both human experts and ChatGPT provided answers. Analyzed linguistic differences between human and AI responses.	ChatGPT responses tend to be more verbose, use more formal language, and exhibit less variability than human answers. Simple classifiers trained on HC3 achieve high accuracy but may not generalize to other generators.
On the Reliability of AI-Text Detectors	Sadasivan, V.S., et al.	2023	To investigate the fundamental limitations of AI text detection methods and assess their robustness against evasion attacks.	Tested multiple detection methods (watermarking, neural classifiers, statistical methods) against paraphrasing attacks using models like DIPPER. Provided theoretical analysis of detection limits.	Simple paraphrasing attacks can reduce detection accuracy to near-random levels. As language models improve, the statistical gap between human and machine text shrinks, making detection inherently harder.
Detecting Machine-Generated Text: A Critical Survey	Tang, R., et al.	2023	To provide a comprehensive survey of machine-generated text detection methods, categorizing existing approaches and identifying open	Systematic review of detection methods categorized into statistical approaches, neural classifiers, and watermarking-based techniques.	No single detection method dominates across all settings. Statistical methods work for older generators but struggle with newer fluent models. Neural classifiers achieve

			challenges.	Comparative analysis across different evaluation settings.	higher accuracy but are brittle under distribution shift. Ensemble methods offer the best generalization.
Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks	Reimers, N. & Gurevych, I.	2019	To derive semantically meaningful sentence embeddings from BERT that can be compared using cosine similarity for efficient similarity search.	Modified BERT with Siamese and triplet network structures. Trained on NLI and STS datasets using contrastive and triplet loss functions to produce fixed-size sentence embeddings.	Sentence-BERT reduces the computational cost of finding similar sentence pairs from quadratic (65 hours for 10K sentences with BERT cross-encoder) to linear (5 seconds with cosine similarity), while maintaining competitive accuracy on STS benchmarks.
SimCSE: Simple Contrastive Learning of Sentence Embeddings	Gao, T., Yao, X., & Chen, D.	2021	To improve sentence embedding quality through a simple contrastive learning framework using both unsupervised and supervised approaches.	Unsupervised: uses dropout as minimal data augmentation, passing the same sentence through the encoder twice with different dropout masks. Supervised: leverages NLI pairs with contrastive objectives.	SimCSE achieves significantly better alignment and uniformity in the embedding space compared to previous methods. The supervised variant outperforms Sentence-BERT on multiple STS benchmarks.
Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval is an Effective Defense (DIPPER)	Krishna, K., et al.	2023	To demonstrate that discourse-level paraphrasing can evade AI text detectors and to propose retrieval-based detection as a robust countermeasure.	Developed DIPPER, an 11B parameter T5-XXL paraphraser with controllable lexical diversity and order diversity knobs. Tested evasion against multiple detectors and proposed retrieval-based defense.	Even moderate paraphrasing with DIPPER fools most existing detectors. However, retrieval-based methods that compare against a database of known AI outputs remain effective even after paraphrasing.
DeBERTa: Decoding-enhanced BERT with Disentangled Attention	He, P., Liu, X., Gao, J., & Chen, W.	2021	To improve upon BERT by introducing disentangled attention that separately encodes content and position information, along with an enhanced mask decoder.	Proposed two novel techniques: disentangled attention mechanism using separate vectors for content and position, and an enhanced mask decoder that incorporates absolute position information for token prediction.	DeBERTa achieved state-of-the-art results on multiple NLU benchmarks including SuperGLUE. The disentangled attention provides better fine-grained token-level understanding compared to standard BERT attention.
Fine-tuned LLMs for Multilingual Machine-Generated Text Detection (SemEval-2024 Task 8)	Hlavnova, E. & Pikuliak, M.	2024	To explore fine-tuning strategies for detecting machine-generated text across multiple languages using parameter-efficient methods.	Fine-tuned RoBERTa and XLM-RoBERTa models using LoRA (Low-Rank Adaptation) for multilingual AI text detection. Used majority voting across language-specific classification heads.	LoRA-adapted RoBERTa with majority voting achieves competitive multilingual detection performance without training separate models per language. Parameter-efficient fine-tuning is practical for multilingual deployment.
Fine-grained AI-generated Text Detection using Multi-task	Multiple authors	2025	To introduce a fine-grained detection framework that classifies text as fully	Multi-task learning with auxiliary sentence-level detection alongside document-level	The fine-grained three-class approach (human/AI/mixed) outperforms binary

Auxiliary and Multi-level Contrastive Learning (FAIDSet)			human, fully AI, or mixed content, with a large-scale multilingual dataset.	classification. Multi-level contrastive learning to separate human, AI, and mixed text representations.	classification on real-world data where documents often contain both human and AI-written sections. The multilingual dataset enables cross-lingual fine-grained detection.
A Survey on Plagiarism Detection	Foltýnek, T., Meuschke, N., & Gipp, B.	2019	To provide a comprehensive survey of plagiarism detection methods covering string matching, citation analysis, and semantic approaches.	Systematic review categorizing plagiarism detection into string-based, syntax-based, semantic-based, and citation-based methods. Analysis of commercial and academic detection tools.	String-based methods detect verbatim copying effectively but miss paraphrased plagiarism. Semantic approaches using embeddings are the most promising for detecting meaning-level plagiarism but were the least mature at the time of the survey.
Authorship Style Transfer with Policy Optimization	Multiple authors	2024	To develop policy optimization techniques for transferring authorship style while preserving content meaning.	Used reinforcement learning with style classifiers as reward signals to guide a language model in adopting target writing styles. Evaluated on style transfer accuracy and content preservation.	Policy optimization enables controlled style transfer that preserves semantic content while successfully changing authorship characteristics. The approach outperforms supervised fine-tuning on style transfer benchmarks.