

## Detecting Machine-Generated Text in Academic Writing: Stylometric Fingerprinting of Humans and Large Language Models for Authorship Verification

Abdollah Givechi

PhD Student in Islamic Philosophy and Theology, Imam Sadiq University, Tehran, Iran

[a.givechi@isu.ac.ir](mailto:a.givechi@isu.ac.ir)

### Abstract

The proliferation of Large Language Models (LLMs) such as GPT-4 and ChatGPT has introduced significant challenges to maintaining academic integrity, necessitating robust methods for distinguishing human-written texts from machine-generated content. This paper provides a comprehensive review of contemporary techniques for detecting machine-generated text in academic writing, with a particular focus on stylometric fingerprinting and machine learning approaches. We systematically analyze key methodologies, including stylometric analysis (lexical diversity, syntactic complexity, and punctuation patterns), psycholinguistic mapping, and the trigram-cosine delta metric. Furthermore, we examine advanced machine learning models such as supervised classification, ensemble learning, and Graph Neural Networks (GNNs) integrated with pre-trained language models. The review also explores multilingual and cross-domain detection strategies, benchmark datasets, and performance evaluation metrics. Despite high accuracy rates reported in recent studies (up to 98%), significant challenges remain regarding generalizability across different LLMs and domains, computational efficiency, and ethical considerations related to privacy. The paper concludes that integrating stylometric analysis with advanced machine learning offers a promising pathway for safeguarding academic integrity, while emphasizing the need for continued research to address existing limitations.

**Keywords:** Machine-Generated Text Detection; Academic Writing; Stylometry; Large Language Models; Authorship Verification.

## 1. Introduction

The rapid advancement and widespread accessibility of Large Language Models (LLMs), most notably OpenAI's GPT-4 and ChatGPT, have marked a transformative era in natural language generation. These models are capable of producing coherent, contextually relevant, and stylistically varied text that is often indistinguishable from human writing. While this technological leap offers numerous benefits across various sectors, including education, research, and content creation, it simultaneously presents profound challenges to the foundational principles of academic integrity. The potential for students, researchers, and writers to undetectably generate essays, research papers, and scholarly articles using AI threatens to undermine the very essence of authorship, originality, and intellectual merit that form the cornerstone of academia (Maaloul, 2024; Przystaliski et al., 2026).

Consequently, the task of accurately discriminating between human-written and machine-generated text has emerged as a critical area of research. The stakes are particularly high in academic writing, where the verification of authorship is paramount to ensuring fair assessment and maintaining the credibility of scholarly publications. The traditional methods of plagiarism detection, which rely on matching text against existing databases, are largely ineffective against AI-generated content, which is often original in its specific string of words, even if unoriginal in concept (Ramnial, Panchoo, & Pudaruth, 2016). This inadequacy has spurred a surge in research dedicated to developing robust detection methodologies.

Among the various approaches proposed, stylometric analysis has proven to be a particularly promising avenue. Stylometry, the quantitative study of literary style, operates on the premise that every writer possesses a unique, subconscious stylistic fingerprint. This fingerprint can be quantified through features such as lexical diversity, syntactic complexity, sentence length distribution, and the frequency of function words and punctuation marks (Maaloul, 2024; Przystaliski et al., 2026; Ramnial et al., 2016). The central hypothesis in this domain is that current LLMs, despite their sophistication, exhibit stylistic patterns that are distinct from those of humans. These patterns may manifest as an unnatural consistency, a lack of creative idiosyncrasy, or specific statistical regularities in word and structure choice (Opara, 2025).

Early detection efforts have focused on supervised machine learning models trained on these stylometric features, achieving high accuracy in controlled settings (Maaloul, 2024; Berriche & Larabi-Marie-Sainte, 2024). More recent advancements include the integration of psycholinguistic theories to map writing patterns to cognitive processes (Opara, 2025), the development of lightweight and interpretable metrics like the trigram-cosine delta (Salnikov & Bonch-Osmolovskaya, 2025), and the application of complex architectures such as Graph Neural Networks (GNNs) combined with pre-trained language models (Valdez-Valenzuela & Gómez-Adorno, 2024). Furthermore, the challenge has expanded to encompass multilingual and cross-domain contexts, requiring detection systems that can generalize beyond a single language or genre, such as academic essays or social media posts (Agrahari et al., 2025; Chhatwal & Zhao, 2024; Al-Shaibani & Ahmed, 2026).

Despite these significant strides, several challenges persist. A primary concern is the generalizability of detection models. A system trained to detect text from one specific LLM may perform poorly on text

generated by a newer or different model (Salnikov & Bonch-Osmolovskaya, 2025; Al-Shaibani & Ahmed, 2026). Similarly, a model trained on academic texts may fail when applied to creative writing. Another critical issue is computational efficiency; for practical, large-scale deployment in educational institutions or publishing houses, detection tools must be both accurate and lightweight (Yan, Zhao, & Guo, 2025). Finally, the deployment of stylometric techniques raises significant ethical considerations, particularly concerning privacy, data protection, and the potential for misclassification, which could wrongly accuse students of academic dishonesty (Patergianakis & Limnietis, 2022).

This paper aims to provide a comprehensive and systematic review of the state-of-the-art in detecting machine-generated text within academic writing. The primary objectives are: to synthesize and analyze the key techniques and approaches, with a specific focus on stylometric fingerprinting and its integration with advanced machine learning models; to evaluate the performance of these methods by examining benchmark datasets and reported metrics from recent studies; to critically discuss the prevailing challenges, including generalizability, computational cost, and ethical implications; and to identify future research directions necessary for developing more robust, fair, and practical detection systems that can effectively safeguard academic integrity in an era of increasingly sophisticated generative AI. By consolidating recent findings and highlighting both the potential and the limitations of current methodologies, this review seeks to offer a valuable resource for researchers, educators, and policymakers navigating the complex landscape of AI in academia.

## 2. Key Techniques and Approaches

### 2.1 Stylometric Analysis

Stylometric analysis, or stylometry, is a foundational technique in the field of authorship attribution and verification. It operates on the core premise that every writer possesses an inherent and relatively stable stylistic fingerprint—a unique set of linguistic habits and patterns that manifest subconsciously in their writing (Maaloul, 2024; Przystalski et al., 2026; Ramnial et al., 2016). By quantifying these patterns, stylometry provides a robust framework for distinguishing between different authors and, more recently, for differentiating human writers from machine-generated text. In the context of academic writing, this approach is particularly valuable, as it delves beyond the superficial content to analyze the underlying structure and statistical regularities of the text itself.

The application of stylometry for detecting machine-generated text typically involves extracting a wide array of linguistic features that capture distinct dimensions of writing style. These features can be broadly categorized into several levels of linguistic analysis:

**Lexical Diversity:** This refers to the range and variety of vocabulary used by an author. Humans tend to exhibit a natural variation in word choice, sometimes employing rare or context-specific terms, while LLMs may display a more statistically average or "safe" vocabulary profile. Common metrics for



measuring lexical diversity include the Type-Token Ratio (TTR), which compares the number of unique words (types) to the total number of words (tokens), and more sophisticated indices that are less sensitive to text length (Maaloul, 2024; Przystalski et al., 2026).

**Syntactic Complexity:** This dimension analyzes the structure of sentences, capturing how writers construct phrases and clauses. Features such as average sentence length, the frequency of subordinate clauses, the use of passive versus active voice, and the distribution of parts of speech can reveal distinct syntactic signatures. Humans often employ a more varied and occasionally irregular syntactic structure, whereas machine-generated text might exhibit a more predictable or uniform syntactic pattern (Przystalski et al., 2026; Ramnial et al., 2016).

**Punctuation Patterns:** The usage of punctuation marks—such as commas, semicolons, dashes, and quotation marks—is a highly idiosyncratic stylistic element. Some authors may favor long, complex sentences with numerous commas, while others prefer shorter, more direct statements. LLMs, trained on vast corpora, may replicate common punctuation patterns but often lack the subtle, author-specific nuances in punctuation that characterize human writing (Maaloul, 2024; Ramnial et al., 2016).

Building upon these traditional feature sets, recent research has integrated more advanced theoretical and methodological frameworks to enhance detection accuracy. One significant advancement is Psycholinguistic Mapping, which seeks to bridge the gap between writing patterns and underlying cognitive processes. This approach, as explored by Opara (2025), involves mapping stylometric features onto psycholinguistic theories of language production. The underlying assumption is that human writing is a product of complex cognitive functions like memory retrieval, reasoning, and emotional state, which leave distinct traces in the text. In contrast, LLMs generate text based on statistical probabilities without any genuine cognitive or intentional states. By identifying features that correlate with cognitive processes—such as the use of self-references, cognitive verbs, or emotional markers—researchers can develop more nuanced and theoretically grounded detectors. This method adds a layer of depth to stylometric analysis, moving beyond mere statistical comparison to a more functional understanding of text generation.

Another innovative and highly practical development is the introduction of the Trigram–Cosine Stylometric Delta. Proposed by Salnikov and Bonch-Osmolovskaya (2025), this metric offers a lightweight, unsupervised, and interpretable approach to distinguishing LLM-generated texts from human-written ones. The method operates by representing texts as vectors based on the frequency of character trigrams (sequences of three consecutive characters). It then calculates the cosine delta—a measure of the angle between these vectors—to quantify stylistic similarity or difference. This trigram-based approach is particularly effective because character-level features are sensitive to subtle stylistic choices in morphology, punctuation, and even spacing, which can be powerful discriminators. The key advantages of this method, as highlighted by the authors, are its computational efficiency and interpretability, making it a practical tool for real-world applications and a robust baseline for comparison against more complex "black box" models. Its effectiveness has been demonstrated across diverse text generation strategies and LLM architectures, underscoring the power of stylometric analysis even in its more parsimonious forms (Salnikov & Bonch-Osmolovskaya, 2025).

In summary, stylometric analysis provides a multi-faceted and powerful toolkit for detecting machine-generated text. From foundational features like lexical and syntactic variation to advanced integrations with psycholinguistics and efficient metrics like the trigram-cosine delta, these methods exploit the fundamental differences between human cognition and statistical language modeling. The continued development and refinement of these stylometric techniques are crucial for building effective defenses against the erosion of academic integrity.

## 2.2 Machine Learning Models

While stylometric analysis provides the essential features for distinguishing human writing from machine-generated text, machine learning models serve as the computational engines that learn to recognize the complex patterns within those features. These models have become increasingly sophisticated, evolving from traditional supervised classifiers to advanced ensemble methods and neural architectures. The integration of machine learning with stylometric features has consistently demonstrated high accuracy in authorship verification tasks, making it a cornerstone of modern detection systems (Maaloul, 2024; Agrahari et al., 2025).

### 2.2.1 Supervised Classification

Supervised machine learning forms the bedrock of most text detection systems. In this paradigm, models are trained on labeled datasets containing both human-written and AI-generated texts, learning to map the extracted stylometric features to their respective classes. The effectiveness of this approach hinges on the quality and discriminative power of the features provided.

Research by Maaloul (2024) demonstrates the efficacy of supervised classification using a comprehensive set of stylometric features. The study employed features such as lexical diversity (including type-token ratio and vocabulary richness), syntactic complexity (measured through parse tree depth and clause density), and readability scores. These features were used to train several classical classifiers, including Support Vector Machines (SVM), Random Forests, and Logistic Regression. The results indicated that models trained on this rich feature set could effectively differentiate AI-generated academic content from human writing, achieving robust performance across various test conditions. The study underscored the importance of feature engineering in the supervised learning pipeline, showing that well-chosen stylometric attributes could compensate for the relative simplicity of the underlying classifier (Maaloul, 2024).

Similarly, Agrahari et al. (2025) explored supervised classification in the context of the GenAI Detection Task 2, which focused on multilingual detection of AI-generated essays. Their approach involved training classifiers on a diverse set of linguistic features extracted from essays written in multiple languages. The researchers found that while language-specific features were valuable, a core set of cross-linguistic stylistic indicators—such as sentence length variability, function word frequency, and part-of-speech distribution—enabled their models to generalize effectively across English, German, and Spanish texts. This work highlights the potential of supervised methods to operate in multilingual academic environments, a critical capability given the global nature of higher education and research (Agrahari et al., 2025).

### 2.2.2 Ensemble Learning

Ensemble learning methods, which combine multiple base classifiers to produce a single, more robust predictor, have shown particular promise in the domain of AI text detection. By aggregating the decisions of diverse models, ensemble techniques can mitigate the biases and weaknesses of individual classifiers, leading to improved accuracy and generalization.

Berriche and Larabi-Marie-Sainte (2024) conducted a comprehensive study on detecting ChatGPT-generated text through writing style analysis, with a specific focus on ensemble methods. Their approach involved extracting a wide range of stylistic features from a corpus of human and ChatGPT-generated texts on identical prompts. They then applied several ensemble techniques, including:

**XGBoost:** An optimized gradient boosting framework that builds an ensemble of decision trees sequentially, with each new tree correcting errors made by the previous ones. XGBoost demonstrated exceptional performance, achieving high accuracy in distinguishing between human and machine authors.

**Stacking (Stacked Generalization):** This meta-learning approach involves training multiple base classifiers and then using a meta-classifier to learn how to best combine their predictions. The stacking ensemble in this study leveraged the complementary strengths of different algorithms, resulting in a detection system that was more robust than any of its individual components.

The study reported that these ensemble methods achieved accuracy rates of up to 98%, demonstrating the power of combining multiple learning algorithms for this challenging task. The authors attributed this success to the ensemble's ability to capture different facets of stylistic variation, with each base classifier attending to distinct patterns in the feature space (Berriche & Larabi-Marie-Sainte, 2024). This high level of accuracy, achieved on diverse prompts and text types, underscores the viability of ensemble learning for practical deployment in academic integrity monitoring.



### 2.2.3 Graph Neural Networks (GNNs)

The most recent advances in machine learning for text detection have seen the incorporation of neural network architectures capable of capturing structural and relational information within texts. Graph Neural Networks (GNNs) represent a particularly innovative direction, as they can model text not merely as a sequence of words or a bag of features, but as a complex network of interconnected linguistic elements.

Valdez-Valenzuela and Gómez-Adorno (2024) explored this frontier in their work presented at the PAN lab on Generative AI Authorship Verification. Their approach, developed by team iimasnlp, represents a significant methodological advancement by integrating three complementary components:

**Pre-trained Language Models (PLMs):** The researchers leveraged large, pre-trained transformer models such as RoBERTa to obtain rich, contextualized representations of the input texts. These models capture deep semantic and syntactic information that traditional feature engineering might miss.

**Stylometric Features:** In parallel with the neural representations, the team extracted a comprehensive set of traditional stylometric features, including character n-grams, word n-grams, and various readability and complexity indices. This combination ensured that both surface-level stylistic patterns and deep linguistic structures were considered.

**Graph Neural Networks:** The core innovation was the use of GNNs to model the relational structure within and between texts. By representing documents as graphs—where nodes might represent sentences, paragraphs, or linguistic units, and edges represent semantic or sequential relationships—the GNN could learn to capture holistic document structure. This is particularly relevant for authorship verification, as the way an author organizes and connects ideas across a document can be a powerful stylistic signature.

The integration of these components within a GNN framework enhanced both classification accuracy and robustness. The graph-based approach allowed the model to consider not just isolated features but the overall architecture of the text, making it more resilient to superficial manipulations that might deceive simpler classifiers. The authors demonstrated that this multi-modal, graph-based approach outperformed methods relying solely on either PLMs or stylometric features alone, highlighting the synergistic potential of combining neural and symbolic representations (Valdez-Valenzuela & Gómez-Adorno, 2024).

In summary, machine learning models for detecting machine-generated text have evolved along a trajectory of increasing sophistication. From well-established supervised classifiers leveraging engineered stylometric features (Maaloul, 2024; Agrahari et al., 2025), to powerful ensemble methods like XGBoost and Stacking that achieve near-perfect accuracy (Berriche & Larabi-Marie-Sainte, 2024),

and finally to cutting-edge neural architectures like GNNs that model the relational structure of text (Valdez-Valenzuela & Gómez-Adorno, 2024), these techniques form the computational backbone of modern authorship verification systems. The continued advancement of these models, particularly in their ability to generalize across domains and resist adversarial attacks, remains a central focus of research in this field.

## 2.3 Multilingual and Cross-Domain Approaches

As the use of Large Language Models becomes a global phenomenon, the challenge of detecting machine-generated text extends far beyond English-language content. Academic writing occurs in numerous languages, and LLMs are increasingly proficient in generating high-quality text in many of them. Furthermore, the stylistic characteristics of writing vary significantly across different domains—from formal academic essays and research articles to informal social media posts and creative writing. Effective detection systems must therefore be capable of operating across both linguistic and domain boundaries, adapting to diverse stylistic norms while maintaining accuracy (Agrahari et al., 2025; Chhatwal & Zhao, 2024; Al-Shaibani & Ahmed, 2026).

### 2.3.1 Multilingual Detection

The development of multilingual detection systems presents unique challenges. Languages differ fundamentally in their morphology, syntax, and orthography, meaning that features discriminative in one language may be irrelevant or misleading in another. Researchers have addressed this challenge through two primary approaches: leveraging multilingual pre-trained language models and identifying language-independent stylometric features.

Agrahari et al. (2025) tackled the multilingual detection problem in the context of the GenAI Detection Task 2, which specifically focused on identifying AI-generated essays across multiple languages. Their methodology, named EssayDetect, combined several innovative strategies to achieve robust cross-lingual performance:

**PLM-Based Methods:** The researchers leveraged multilingual pre-trained language models (mPLMs) such as XLM-RoBERTa and mBERT, which are trained on vast corpora covering dozens of languages. These models develop cross-lingual representations that capture linguistic regularities shared across languages, enabling them to transfer knowledge from high-resource languages to low-resource ones. By fine-tuning these mPLMs on a mixture of human and AI-generated essays from multiple languages, the team developed detectors that could effectively identify machine-generated text regardless of the language of composition.



**Stylometric Feature-Based Techniques:** Complementing the neural approaches, the study also explored language-independent stylometric features. These included universal metrics such as sentence length distribution, punctuation density, type-token ratio, and the frequency of function words. By combining these language-agnostic features with mPLM representations, the detection system achieved improved performance and generalizability across English, German, Spanish, and other languages represented in the task (Agrahari et al., 2025).

Chhatwal and Zhao (2024) further advanced this line of research in their comprehensive study on differentiating human and machine-generated texts in a multilingual setting. Their work systematically compared the effectiveness of various detection approaches across languages with different structural properties. Key findings from their research include:

**Cross-Lingual Transfer:** Detection models trained predominantly on English data could be successfully adapted to other languages through transfer learning techniques, though performance degradation varied depending on the linguistic similarity between the source and target languages.

**Language-Specific Calibration:** The researchers found that optimal feature sets and model architectures varied across languages. For instance, morphologically rich languages like German and Arabic benefited more from character-level features than word-level features, while English showed strong results with both.

**Ensemble Strategies for Multilinguality:** The most robust performance across all tested languages was achieved by ensemble models that combined language-specific detectors with a shared cross-lingual backbone, allowing the system to leverage both universal patterns and language-specific stylistic cues (Chhatwal & Zhao, 2024).

### 2.3.2 Domain-Specific Analysis

Beyond multilingual considerations, the domain or genre of writing significantly influences stylistic expression. Academic writing follows different conventions than social media posts, news articles, or creative fiction. These domain-specific norms affect everything from vocabulary choice and sentence structure to the use of rhetorical devices and citation practices. Developing detection models that can generalize across domains—or adapt to domain-specific patterns—is therefore essential for practical deployment.

Al-Shaibani and Ahmed (2026) addressed this challenge in the context of Arabic machine-generated text detection, with a particular focus on cross-domain evaluation. Their research, which examined texts from academic, social media, and news domains, yielded several important insights:

**Domain-Specific Linguistic Patterns:** The study identified distinct linguistic fingerprints associated with different domains. Academic texts, whether human or AI-generated, exhibited greater syntactic complexity, higher lexical density, and more frequent use of nominalizations and passive constructions. Social media texts, in contrast, showed shorter sentences, more informal vocabulary, and greater use of emotive language and first-person pronouns. These domain-specific patterns meant that a detector trained exclusively on academic texts might misinterpret stylistic features in social media texts, leading to misclassification.

**Cross-Domain Robustness:** The researchers systematically evaluated how well detection models trained on one domain performed when tested on others. They found that models trained on academic texts showed moderate degradation (10-15% reduction in accuracy) when applied to news texts, but more severe degradation (20-25%) when applied to social media texts. This asymmetry highlighted the greater stylistic distance between formal academic writing and informal social media discourse.

**Domain-Adaptation Strategies:** To address these challenges, the study explored several domain-adaptation techniques. Fine-tuning pre-trained models on small amounts of target-domain data proved effective, recovering much of the lost performance. Additionally, the inclusion of domain-invariant stylometric features—such as basic punctuation patterns and function word frequencies that remain relatively stable across domains—helped improve cross-domain generalization. The most successful approach involved training on a mixture of domains, which exposed the model to diverse stylistic conventions and reduced overfitting to any single domain's patterns (Al-Shaibani & Ahmed, 2026).

The implications of this research for academic integrity monitoring are significant. Educational institutions must contend with AI-generated text in various contexts—not only formal essays and theses but also discussion forum posts, reflective journals, and even email communications. Detection systems deployed in these settings must be robust to this diversity, or else risk systematic biases that could unfairly target certain types of writing or, conversely, fail to detect violations in less familiar domains.

Furthermore, the intersection of multilingual and cross-domain challenges presents an even more complex frontier. A detector must be able to recognize, for example, that an AI-generated discussion post in Arabic exhibits different stylistic patterns than an AI-generated research article in English, while still accurately distinguishing both from their human-written counterparts. Addressing this multidimensional variation requires continued research into transfer learning, domain adaptation, and the fundamental linguistic properties that distinguish human cognition from machine generation across all languages and contexts (Agrahari et al., 2025; Chhatwal & Zhao, 2024; Al-Shaibani & Ahmed, 2026).

In summary, multilingual and cross-domain approaches represent essential frontiers in the development of robust text detection systems. By leveraging multilingual pre-trained language models, identifying language-independent stylometric features, and developing sophisticated domain-adaptation strategies, researchers are building detectors capable of operating in the diverse linguistic and contextual

environments where academic writing actually occurs. These efforts are crucial for ensuring that the benefits of AI detection technology can be realized globally and equitably, without being limited to specific languages or genres.

## 2.4 Benchmarking and Evaluation

The development of robust detection methods for machine-generated text necessitates rigorous evaluation frameworks that allow for fair comparison between different approaches and reliable assessment of their real-world performance. Benchmarking and evaluation encompass two critical components: the creation of high-quality benchmark datasets that represent the diversity of human and machine-generated texts, and the application of appropriate performance metrics that capture different dimensions of detection effectiveness (Przystalski et al., 2026; Agrahari et al., 2025; Berriche & Larabi-Marie-Sainte, 2024; Tatavarthi, Abri, & Attar, 2025).

### 2.4.1 Benchmark Datasets

The foundation of any empirical evaluation is the dataset upon which models are trained and tested. In the domain of AI text detection, the creation of benchmark datasets presents unique challenges. These datasets must adequately represent the stylistic diversity of human writing across different genres, disciplines, and demographic groups, while also capturing the output of various LLMs under different generation conditions.

Przystalski et al. (2026) made significant contributions to this area through their development of benchmark datasets specifically designed for stylometric analysis of short text samples. Their methodology involved:

**Wikipedia-Based Corpus:** The researchers constructed a dataset comprising human-written Wikipedia articles across multiple domains, including science, humanities, and current events. Wikipedia was chosen as a source of human text due to its collaborative editing process, which results in a relatively standardized yet still human-authored prose style. For each human-written article, they generated machine-written counterparts using various LLMs, including GPT-3.5 and GPT-4, prompted to produce text on the same topic with similar length constraints.

**Academic Essay Collection:** Complementing the Wikipedia corpus, the dataset also included a collection of authentic student essays from various disciplines and educational levels. These essays represented natural human writing in academic contexts, complete with the idiosyncrasies, occasional errors, and stylistic variations characteristic of student authors. Corresponding AI-generated essays were produced by prompting LLMs with the same essay prompts provided to students.



**Short Sample Focus:** A distinctive feature of this dataset was its emphasis on relatively short text samples (ranging from 100 to 500 words). This design choice reflected the practical reality that many academic integrity scenarios involve detecting AI-generated content in paragraphs or short essays rather than full-length documents. The researchers demonstrated that even with these shorter samples, stylometric methods could achieve impressive accuracy, though performance degraded as sample length decreased below certain thresholds (Przystalski et al., 2026).

Agrahari et al. (2025) extended this benchmarking effort to the multilingual domain through their work on the GenAI Detection Task 2. Their dataset construction involved:

**Multilingual Essays:** The researchers collected human-written essays in multiple languages, including English, German, Spanish, French, and Chinese. These essays covered a range of academic topics and were sourced from educational institutions and public writing repositories. For each language, they generated corresponding AI-written essays using state-of-the-art LLMs with multilingual capabilities.

**Controlled Generation:** To ensure fair comparison, the generation process was carefully controlled. The same prompts used for human writers were provided to the LLMs, and multiple generation temperatures were employed to capture the range of machine outputs from deterministic to highly creative. This controlled approach allowed for apples-to-apples comparisons between human and machine writing on identical tasks.

**Public Availability:** In the spirit of advancing the field, the benchmark datasets were made publicly available for research purposes, enabling other teams to evaluate their methods on the same standardized corpus and facilitating direct comparison between different detection approaches (Agrahari et al., 2025).

### 2.4.2 Performance Metrics

The evaluation of detection models requires a suite of metrics that capture different aspects of performance. No single metric tells the complete story, as trade-offs often exist between different performance dimensions—for example, between sensitivity and specificity.

Berriche and Larabi-Marie-Sainte (2024) employed a comprehensive set of metrics in their evaluation of ensemble methods for ChatGPT text detection:

**Accuracy:** The most intuitive metric, accuracy represents the proportion of all predictions that were correct. In their study, ensemble methods achieved accuracy rates up to 98%, indicating that the models correctly classified nearly all texts in their evaluation corpus. While impressive, the authors cautioned that accuracy alone can be misleading in imbalanced datasets, where a model might achieve high accuracy simply by predicting the majority class (Berriche & Larabi-Marie-Sainte, 2024).

**F1 Score:** The F1 score, calculated as the harmonic mean of precision and recall, provides a more balanced view of model performance, particularly when classes are imbalanced. Precision measures the proportion of texts classified as AI-generated that truly were AI-generated, while recall measures the proportion of actual AI-generated texts that were correctly identified. The F1 score combines these into a single metric that penalizes extreme imbalances. The study reported F1 scores consistently above 0.95, indicating that the ensemble models maintained both high precision and high recall (Berriche & Larabi-Marie-Sainte, 2024).

**Matthews Correlation Coefficient (MCC):** The MCC, which produces a value between -1 and +1, is widely regarded as a more informative metric for binary classification than either accuracy or F1 score. It takes into account all four categories of the confusion matrix and provides a balanced measure even when classes are of very different sizes. The researchers reported MCC values exceeding 0.95 for their best-performing ensemble models, confirming the robustness of their approach (Berriche & Larabi-Marie-Sainte, 2024).

Tatavarthi, Abri, and Attar (2025) provided additional insights into performance evaluation through their work on AI-generated text detection and source identification:

**Source Identification Accuracy:** Beyond simply distinguishing human from machine text, their research addressed the more granular task of identifying which specific LLM generated a given text. This required evaluation metrics that could capture performance across multiple classes. They employed macro-averaged and micro-averaged F1 scores to assess multi-class classification performance, finding that models could not only detect AI-generated text but also identify its source with reasonable accuracy (Tatavarthi et al., 2025).

**Confidence Calibration:** An important contribution of this research was the emphasis on confidence calibration—how well a model's predicted probabilities align with its actual accuracy. A well-calibrated model that predicts 90% probability of AI generation should be correct approximately 90% of the time. Poor calibration can lead to overconfidence in predictions, which is particularly problematic in high-stakes academic integrity contexts where false accusations can have serious consequences. The authors introduced calibration metrics and demonstrated that ensemble methods generally produced better-calibrated predictions than individual classifiers (Tatavarthi et al., 2025).

**Robustness Across Generation Parameters:** The study also evaluated how detection performance varied with different LLM generation parameters, such as temperature settings and prompt engineering. They found that while detection accuracy remained high across most conditions, texts generated with very high temperature settings were somewhat harder to detect, highlighting the importance of evaluating models across diverse generation conditions (Tatavarthi et al., 2025).

The collective findings from these benchmarking and evaluation studies reveal several important patterns. First, state-of-the-art detection models can achieve remarkable accuracy—up to 98% or higher—on well-constructed benchmark datasets (Przystalski et al., 2026; Berriche & Larabi-Marie-Sainte, 2024). Second, the choice of evaluation metrics matters: relying solely on accuracy can mask important performance dimensions, and comprehensive evaluation should include F1 scores, MCC, and calibration metrics (Berriche & Larabi-Marie-Sainte, 2024; Tatavarthi et al., 2025). Third, benchmark datasets must evolve alongside LLMs; as models improve and new generation techniques emerge, evaluation corpora must be updated to reflect the current landscape (Przystalski et al., 2026; Aghahari et al., 2025).

Despite these advances, important challenges remain in benchmarking and evaluation. The rapid pace of LLM development means that datasets can quickly become outdated, as newer models may exhibit different stylistic patterns than those on which detectors were trained. Additionally, most benchmarks currently focus on English and a few high-resource languages, leaving questions about performance in the world's many other languages largely unexplored (Aghahari et al., 2025). Finally, the ecological validity of benchmarks—how well performance on curated datasets predicts real-world performance in diverse educational and publishing contexts—remains an open question requiring ongoing investigation.

### 3. Challenges and Future Directions

Despite significant advances in the detection of machine-generated text, the field faces several persistent challenges that must be addressed to develop robust, practical, and ethically sound systems. The rapid evolution of Large Language Models, the diversity of textual domains, and the high-stakes nature of academic integrity applications all contribute to the complexity of this research area. This section critically examines the primary challenges facing current detection methods and outlines promising directions for future research (Salnikov & Bonch-Osmolovskaya, 2025; Al-Shaibani & Ahmed, 2026; Yan et al., 2025; Patargianakis & Limniotis, 2022).

#### 3.1 Generalizability

Perhaps the most fundamental challenge in AI text detection is ensuring that methods can generalize across different LLMs, text domains, and generation conditions. A detection system that performs well on texts from one specific model may fail dramatically when confronted with outputs from a newer or differently architected model (Salnikov & Bonch-Osmolovskaya, 2025; Al-Shaibani & Ahmed, 2026).



Salnikov and Bonch-Osmolovskaya (2025) systematically investigated this generalization challenge in their development of the trigram-cosine stylometric delta. Their research revealed several important dimensions of the generalizability problem:

**Cross-Model Generalization:** The researchers evaluated their detection method on texts generated by multiple LLMs, including GPT-3.5, GPT-4, and several open-source models. They found that detectors trained on outputs from a single model often exhibited significant performance degradation when tested on texts from different models. For instance, a detector optimized for GPT-3.5 might achieve 95% accuracy on held-out GPT-3.5 samples but drop to 75-80% accuracy on GPT-4 samples. This degradation occurred because different LLMs, even those from the same family, develop distinct stylistic fingerprints based on their training data, architecture, and alignment procedures.

**Evolutionary Pressure:** As LLMs continue to evolve at a rapid pace, the challenge of cross-model generalization becomes increasingly acute. A detector developed today may be obsolete within months as new models with different stylistic characteristics are released. This creates an "arms race" dynamic in which detection methods must constantly adapt to keep pace with generation capabilities.

**Generation Strategy Variation:** The study also examined how different generation strategies—such as varying temperature settings, top-k sampling, and prompt engineering—affected detectability. They found that texts generated with higher temperature settings were generally harder to detect than those generated with lower temperatures. This variation within a single model further complicates the generalization challenge (Salnikov & Bonch-Osmolovskaya, 2025).

Al-Shaibani and Ahmed (2026) extended this analysis to the domain level in their work on Arabic machine-generated text detection. Their findings highlighted the importance of cross-domain generalization:

**Domain Shift:** The researchers systematically evaluated how detection models trained on one domain performed when tested on others. They observed significant performance degradation, with accuracy drops of 10-25% depending on the stylistic distance between domains. This domain shift problem arises because different genres have different stylistic norms, and models may inadvertently learn to associate domain-specific features with human or machine authorship rather than learning truly generalizable discriminative patterns.

**Stylistic Overlap:** Some domains showed greater stylistic overlap between human and machine writing, making detection more challenging. For example, machine-generated news articles were harder to distinguish from human-written ones than machine-generated social media posts, possibly because news writing follows more standardized conventions that LLMs can easily replicate.

**Domain-Adaptation Strategies:** To address these challenges, the study explored various domain-adaptation techniques, including fine-tuning on target-domain data, domain-adversarial training, and the use of domain-invariant features. While these approaches showed promise, none completely eliminated the performance gap, indicating that cross-domain generalization remains an open research problem (Al-Shaibani & Ahmed, 2026).

Future research directions for addressing generalizability challenges include:

**Adversarial Training:** Developing detectors that are explicitly trained to be robust to variations in model architecture, domain, and generation parameters through exposure to diverse examples during training.

**Continual Learning:** Creating detection systems that can continuously update themselves as new LLMs emerge, without catastrophic forgetting of previously learned patterns.

**Fundamental Feature Discovery:** Identifying stylometric features that are truly invariant across models and domains—properties of human cognition that machines cannot easily replicate regardless of their architecture or training data.

### 3.2 Computational Efficiency

For detection methods to be practically useful in real-world academic settings, they must be not only accurate but also computationally efficient. Educational institutions may need to screen thousands of student submissions, and publishing houses may need to evaluate numerous manuscript submissions. Lightweight, scalable detection methods are therefore essential (Yan et al., 2025).

Yan, Zhao, and Guo (2025) addressed this challenge directly in their work on lightweight detection systems. Their research introduced several innovations aimed at reducing computational overhead while maintaining detection accuracy:

**Zero-Shot Detection:** The researchers proposed a novel approach called "Once Call" that enables zero-shot detection of machine-generated text without requiring training on labeled examples. This method leverages the intrinsic properties of LLMs—specifically, their tendency to assign higher probabilities to certain token sequences—to distinguish machine outputs from human writing. By eliminating the need for training data and model fine-tuning, this approach achieves dramatic reductions in computational requirements.

**Efficiency Metrics:** The study introduced systematic efficiency metrics for detection systems, including inference time per text, memory footprint, and FLOPs required per classification. These metrics allow for

fair comparison between different methods along computational dimensions, complementing traditional accuracy-based evaluations.

**Trade-Off Analysis:** The researchers systematically explored the trade-off between computational efficiency and detection accuracy. They found that while the most accurate detectors tended to be computationally intensive, lightweight approaches could achieve surprisingly strong performance—often within 5-10% of state-of-the-art accuracy—while requiring orders of magnitude less computation.

**Practical Deployment Scenarios:** The study also considered practical deployment constraints in educational settings, such as the need for real-time processing, the limitations of institutional computing infrastructure, and the importance of energy efficiency for sustainability. These considerations led to recommendations for matching detection methods to specific use cases based on computational requirements (Yan et al., 2025).

Future research directions for improving computational efficiency include:

**Model Distillation:** Developing smaller, faster student models that approximate the performance of larger teacher models through knowledge distillation techniques.

**Feature Selection:** Identifying minimal feature sets that capture most of the discriminative information, enabling simpler and faster classifiers without significant accuracy loss.

**Hardware Acceleration:** Exploring the use of specialized hardware for efficient deployment of detection systems in educational contexts.

**Progressive Screening:** Developing multi-stage detection pipelines that use lightweight methods for initial screening and reserve more computationally intensive analyses for borderline cases.

### 3.3 Ethical Considerations

The deployment of stylometric techniques for authorship verification raises profound ethical questions that must be carefully addressed. These concerns span privacy, data protection, algorithmic fairness, and the potential for harmful misclassifications (Patergianakis & Limniotis, 2022).

Patergianakis and Limniotis (2022) provided a comprehensive examination of privacy issues in stylometric methods, identifying several critical ethical challenges:

**Privacy and Data Protection:** Stylometric analysis operates on the premise that writing style is a unique identifier—a kind of behavioral biometric. Just as fingerprints or retinal scans can identify individuals, stylometric profiles can potentially de-anonymize authors even when they write under pseudonyms or



attempt to conceal their identity. This raises significant privacy concerns, particularly when detection systems are applied to sensitive domains such as student writing, where individuals may have reasonable expectations of privacy. The researchers noted that stylometric profiles could potentially be used for purposes beyond academic integrity monitoring, such as surveillance or behavioral profiling, without adequate consent or oversight.

**Informed Consent:** In educational settings, students may not be fully informed about how their writing samples are being used to train and evaluate detection systems. The collection and storage of student texts for stylometric analysis raises questions about informed consent, data ownership, and the long-term retention of personal writing samples.

**Algorithmic Fairness and Bias:** Stylometric methods may exhibit systematic biases against certain groups of writers. For example, non-native speakers, writers from different cultural backgrounds, or individuals with certain learning differences may exhibit stylistic patterns that differ from the "typical" human writing represented in training data. If detection models are trained primarily on texts from native-speaking, academically successful writers, they may disproportionately flag texts from linguistic minorities as machine-generated, leading to false accusations and reinforcing educational inequities.

**Risk of Misclassification:** Perhaps the most immediate ethical concern is the potential for false positives—incorrectly classifying human-written text as machine-generated. In academic contexts, such misclassifications can have severe consequences, including accusations of academic dishonesty, damage to reputation, and disciplinary actions. The researchers emphasized that detection systems, no matter how accurate, will never achieve perfect performance, and institutions must have robust procedures for handling cases where algorithmic predictions conflict with human judgment.

**Transparency and Explainability:** Many state-of-the-art detection methods, particularly those based on deep learning, function as "black boxes" that provide predictions without explanations. This lack of transparency makes it difficult for individuals to understand why their writing was flagged, to contest automated decisions, or for institutions to audit detection systems for fairness and bias. The researchers argued for the development of more interpretable detection methods and for procedural safeguards that ensure human oversight of automated decisions (Patergianakis & Limniotis, 2022).

Future research directions for addressing ethical considerations include:

**Fairness-Aware Learning:** Developing detection methods that explicitly incorporate fairness constraints, ensuring that performance is equitable across demographic groups and writing styles.

**Privacy-Preserving Techniques:** Exploring techniques such as differential privacy, federated learning, and secure multi-party computation that enable detection without compromising individual privacy.

Explainable AI (XAI): Advancing research into interpretable detection methods that can provide explanations for their predictions, enabling meaningful human review and contestation.

Ethical Guidelines and Governance: Developing professional and institutional guidelines for the ethical deployment of detection systems, including standards for transparency, consent, data retention, and appeal processes.

Human-in-the-Loop Systems: Designing detection workflows that combine algorithmic predictions with human judgment, ensuring that automated systems support rather than supplant human decision-making in high-stakes contexts.

#### 4. Conclusion

This paper has provided a comprehensive review of contemporary techniques for detecting machine-generated text in academic writing, with a particular focus on stylometric fingerprinting and machine learning approaches. The proliferation of Large Language Models such as GPT-4 and ChatGPT has created an urgent need for robust detection methods to preserve academic integrity, and the research community has responded with a diverse array of innovative solutions.

The review has demonstrated that stylometric analysis remains a foundational technique, leveraging features such as lexical diversity, syntactic complexity, and punctuation patterns to capture the unique stylistic fingerprints of human authors. Recent advancements, including psycholinguistic mapping and the trigram-cosine delta metric, have further enhanced the power and interpretability of stylometric approaches. Machine learning models have evolved in parallel, progressing from supervised classification methods to sophisticated ensemble techniques like XGBoost and Stacking that achieve accuracy rates up to 98%, and finally to cutting-edge Graph Neural Networks that model the relational structure of texts in conjunction with pre-trained language models.

The expansion of detection research into multilingual and cross-domain contexts represents a crucial development, recognizing that academic writing occurs in diverse linguistic and stylistic environments. Researchers have leveraged multilingual pre-trained language models and language-independent stylometric features to build detectors capable of operating across English, German, Spanish, Arabic, and other languages. Similarly, domain-adaptation strategies have been developed to address the stylistic differences between academic essays, news articles, and social media posts. Benchmarking efforts have produced valuable datasets and evaluation frameworks, with metrics such as accuracy, F1 score, Matthews Correlation Coefficient, and confidence calibration providing comprehensive assessments of detection performance.

Despite these advances, significant challenges persist. Generalizability across different LLMs and domains remains elusive, as detectors trained on specific models or genres often fail when confronted with unfamiliar styles. The rapid evolution of language models creates an ongoing "arms race" between generation and detection capabilities. Computational efficiency is another critical concern, as practical

deployment in educational settings requires lightweight, scalable solutions that can process thousands of submissions without excessive resource demands. Perhaps most importantly, ethical considerations surrounding privacy, algorithmic fairness, misclassification risk, and transparency demand careful attention. Stylometric profiles function as behavioral biometrics with profound privacy implications, and biased detectors could disproportionately harm linguistic minorities or non-native speakers.

The integration of stylometric analysis with advanced machine learning offers a promising pathway for safeguarding academic integrity, but realizing this potential requires continued research across multiple fronts. Future work should focus on developing detectors that are robust to model evolution and domain variation, computationally efficient enough for widespread deployment, and designed with ethical principles at their core. Fairness-aware learning, privacy-preserving techniques, explainable AI, and human-in-the-loop systems represent important directions for ensuring that detection technology serves educational values rather than undermining them.

In conclusion, the detection of machine-generated text in academic writing is a dynamic and rapidly evolving field at the intersection of computational linguistics, machine learning, and ethics. The techniques reviewed in this paper provide a strong foundation for current practice, but ongoing innovation is essential to keep pace with advancing language models and to address the complex challenges that lie ahead. Researchers, educators, and policymakers must work together to develop detection systems that are not only accurate but also fair, transparent, and respectful of individual privacy—tools that support academic integrity without compromising the educational mission they seek to protect.

## References

- Agrahari, S., Jayant, S., Kumar, S., & Singh, S.R. (2025). EssayDetect at GenAI Detection Task 2: Guardians of Academic Integrity: Multilingual Detection of AI-Generated Essays. Proceedings - International Conference on Computational Linguistics, COLING.
- Al-Shaibani, M.S., & Ahmed, M. (2026). Arabic machine-generated text detection: Stylometric analysis and cross-model evaluation. Expert Systems with Applications.
- Berriche, L., & Larabi-Marie-Sainte, S. (2024). Unveiling ChatGPT text using writing style. Heliyon.
- Chhatwal, G.S., & Zhao, J. (2024). Unveiling the Source: Differentiating Human and Machine-Generated Texts in a Multilingual Setting. Proceedings - 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2024.



Maaloul, K. (2024). Identifying AI-Written Text in Academia: A Machine Learning-Based Framework. Proceedings - ECTE-Tech 2024: International Conference on Electrical, Computer, Telecommunication, and Energy Technologies.

Opara, C. (2025). Distinguishing AI-Generated and Human-Written Text Through Psycholinguistic Analysis. Lecture Notes in Computer Science.

Patergianakis, A., & Limniotis, K. (2022). Privacy Issues in Stylometric Methods. Cryptography.

Przystalski, K., Argasiński, J.K., Grabska-Gradzińska, I., & Ochab, J.K. (2026). Stylometry recognizes human and LLM-generated texts in short samples. Expert Systems with Applications.

Ramnial, H., Panchoo, S., & Pudaruth, S. (2016). Authorship attribution using stylometry and machine learning techniques. Advances in Intelligent Systems and Computing.

Salnikov, E., & Bonch-Osmolovskaya, A. (2025). Detecting LLM-Generated Text with Trigram-Cosine Stylometric Delta: An Unsupervised and Interpretable Approach. Journal of Language and Education.

Tatavarthi, A.P., Abri, F., & Attar, N. (2025). AI-Generated Text Detection and Source Identification. Journal of Advances in Information Technology.

Valdez-Valenzuela, A., & Gómez-Adorno, H. (2024). Team iimasnlp at PAN: Leveraging Graph Neural Networks and Large Language Models for Generative AI Authorship Verification. CEUR Workshop Proceedings.

Yan, J., Zhao, W., & Guo, H. (2025). A Lightweight Detector: Zero-Shot Detection of Machine-Generated Text with Once Call. 2025 5th International Conference on Artificial Intelligence, Big Data and Algorithms, CAIBDA 2025.