https://doi.org/10.17323/jle.2025.22211

Detecting LLM-Generated Text with Trigram–Cosine Stylometric Delta: An Unsupervised and Interpretable Approach

Egor Salnikov [1], **Anastasiya Bonch-Osmolovskaya** [2]

[1] HSE University, Moscow, Russia

[2] Vinogradov Institute of Russian Language, Moscow, Russia

ABSTRACT

Background: Contemporary methods for detecting synthetic text, including model-specific detectors and transformer-based classifiers, often rely on intensive training or on features tied to particular language models, which restricts their generalizability to unfamiliar LLMs and diverse domains.

Purpose: To advance text attribution research by introducing a stylometry-based approach that utilizes trigram-based cosine delta as a lightweight and interpretable metric for distinguishing LLM-generated texts from human-written texts, irrespective of the underlying generation strategy.

Method: A corpus of Russian diary entries was compiled, encompassing both authentic human-written texts and synthetic counterparts generated through few-shot prompting and finetuned LoRA models. To evaluate the effectiveness of the proposed approach, multiple stylometric-delta variations were examined, integrating uni-, bi-, and trigram features with Manhattan and cosine distance metrics.

Results: The evaluation demonstrated that the trigram–cosine delta consistently achieved the highest performance across experimental conditions, reaching an Adjusted Rand Index of approximately 0.70. This markedly surpassed both the finetuned RuModernBERT baseline (ARI ≈ 0.28) and the classic unigram-based delta (ARI ≈ 0.53). Importantly, the method proved effective not only within the Russian diary corpus but also when applied to the RuATD benchmark, where it successfully separated human-authored and machine-generated texts and produced coherent clustering of related model families.

Conclusion: The findings confirm that trigram–cosine stylometric delta offers a robust, interpretable, and computationally efficient strategy for detecting LLM-generated texts across diverse generation strategies, including few-shot prompting and finetuning. By capturing discourse-level stylistic cohesion, the method advances beyond surface fluency and provides a scalable, unsupervised alternative to classifier-based detectors. While current validation is limited to Russian diaries and selected generation models, the approach demonstrates clear potential for broader application across domains, languages, and emerging state-of-the-art LLMs.

KEYWORDS

LLM; synthetic text detection; stylometry; Burrows delta; cosine delta

Citation: 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, 11*(3), 138-151. https://doi.org/10.17323/jle.2025.22211

Correspondence:
Egor Salnikova,
esalnikov@hse.ru

Received: August 13, 2024
Accepted: September 18, 2025
Published: September 30, 2025

INTRODUCTION

The rapid development of large language models (LLMs) has significantly advanced natural language generation, but it has also intensified concerns regarding the authenticity and reliability of textual content across multiple domains such as education, journalism, and digital archiving (Aich et al., 2022; Bender et al., 2021; Huang et al., 2023; Sahoo et al., 2024; Gurioli et al., 2025). Despite ongoing progress (Fraser et al., 2025), the ability to reliably differentiate between human-authored and machine-generated texts remains limited (Wang, S. et al., 2024; Wang, Y. et al., 2023). This challenge has direct implications for academic integrity, information credibility, and the security of communication systems (Gressel et al., 2024; Roy et al., 2024).

A wide range of detection methods has been proposed, yet none of them offers a universal and sufficiently robust solution (Wu et al., 2025; Tang et al., 2024, Sadasivan et al., 2023). Model-dependent strategies, such as watermarking, introduce identifiable patterns during text generation (Kirchenbauer et al., 2023; Zhao et al., 2022), but these techniques cannot be applied retroactively to external or proprietary models. Distribution-matching approaches achieve precision when logit information is accessible (Gehrmann et al., 2019; Mitchell et al., 2023), but they fail when applied to black-box or previously unseen models. Supervised classifiers, including BERT-based detectors and systems like DetectGPT, demonstrate strong performance on in-domain data (Antoun et al., 2023; Bethany et al., 2024), although they require large volumes of labeled data, act as black boxes with limited interpretability, and generalize poorly to out-of-domain or finetuned outputs (Bakhtin et al., 2019; Li et al., 2023; Shamardina et al., 2022). Feature-based unsupervised methods avoid reliance on labeled datasets and provide higher interpretability, but their effectiveness declines when the text domain or stylistic register shifts significantly (Ma et al., 2023; Muñoz-Ortiz et al., 2023; Guo et al., 2023).

Stylometric techniques offer an alternative pathway grounded in authorship attribution studies. Burrows' delta and its variations rely on distance-based comparisons of frequent words and other stylistic features (Burrows, 2002; Hoover, 2004; Craig & Kinney, 2009). These methods have been successfully applied across languages and registers (Rybicki & Eder, 2011; Eder et al., 2016) and remain valued for their simplicity, interpretability, and low computational requirements. Recent research has suggested that even basic stylometric measures can differentiate human texts from LLM outputs in certain contexts (Rebora, 2023; Salnikov & Bonch-Osmolovskaya, 2023; Wang, S. et al., 2024). However, such studies often focus on zero-shot prompting and fail to address more advanced generation strategies such as finetuning, which can more effectively mimic authorial style and therefore complicate detection (Schuster et al., 2020; Zhang et al., 2024; Przystalski et al., 2025).

This gap highlights the need for an unsupervised detection approach that preserves interpretability and scalability while demonstrating robustness against different generation strategies. The present study therefore sets out to systematically assess the potential of stylometric delta methods for distinguishing between human-written and LLM-generated texts. Specifically, the study aims to determine whether trigram-based cosine delta, in comparison with alternative n-gram and distance metric configurations, provides a reliable, interpretable, and computationally efficient solution for text attribution across both few-shot prompting and finetuning scenarios.

LITERATURE REVIEW

Model-Dependent Approaches

Model-dependent approaches include methods such as watermarking, which introduce traceable patterns into generated outputs by manipulating token selection or probability distributions (Kirchenbauer et al., 2023; Zhao et al., 2023). These techniques demonstrate high effectiveness when applied to text produced by models under the researcher's control. Nevertheless, their reliance on pre-embedded signals makes them unsuitable for detecting content generated by external or proprietary systems, since signals cannot be retroactively incorporated into outputs created by third-party models. A related group of strategies is distribution matching, exemplified by the methods proposed by Gehrmann et al. (2019) and Mitchell et al. (2023). These approaches compare statistical regularities between known model outputs and test samples. Although they can achieve high precision when full access to the model is available, their performance decreases substantially when applied to black-box systems or previously unseen models, which limits their applicability in more general detection settings.

Supervised Classifier-Based Detection

A popular strategy involves training classifiers like BERT, RoBERTa, or DetectGPT using labeled samples (Shamardina et al., 2022; Li et al., 2023; Antoun et al., 2023; Bethany et al., 2024; Emi & Spero, 2024). For example, the Pangram Text classifier outperforms DetectGPT and many commercial tools in accuracy and generalization tasks. Though effective in-domain, these models carry significant drawbacks. Firstly, they require large quantities of labeled synthetic and human-written text, which is expensive and time-consuming to generate. Secondly, they often act as black boxes, offering little interpretability for their decisions. And lastly, they struggle to generalize to new domains or adaptation techniques like finetuning or fewshot prompting.

Unsupervised, Feature-Based Methods

These methods rely on linguistically interpretable features such as function-word frequencies, syntactic complexity, lexical richness, and other stylistic indicators (Ma et al., 2023; Muñoz-Ortiz et al., 2023; Zaitsu & Jin, 2023; Guo et al., 2023; Fröhling & Zubiaga, 2021, Chhatwal & Zhao, 2025). They do not require training data, and their outputs can be directly traced back to specific linguistic features, which ensures transparency of interpretation (Kumarage & Liu, 2023; Opara, 2024; Weerasinghe et al., 2025). However, when applied across domains that differ in genre, topic, or authorial style, the relevance and effectiveness of such features may decline substantially, which reduces the reliability of these methods in cross-genre applications and in contexts involving evolving LLM outputs.

Stylometry-Based Techniques

Stylometry-based techniques have their origins in authorship attribution, where stylometric delta was introduced as a method for comparing ranked-frequency profiles, most often based on function words, across different corpora (Burrows, 2002). Subsequent studies have demonstrated its applicability to synthetic text detection. For example, Rebora (2023) employed Burrows' delta to distinguish ChatGPT-generated texts from Dickensian prose in a zero-shot setting. Although the study provided valuable insights, its scope was narrow, since it focused on a single stylistic register and did not consider texts produced through finetuning or domain-adapted LLMs. The principal strength of this approach lies in its simplicity and interpretability, which makes it appealing as an unsupervised method. Nevertheless, its broader applicability has not been thoroughly examined, particularly in contexts involving more sophisticated generation strategies such as finetuning or few-shot prompting. Moreover, Schuster et al. (2020) demonstrated that stylometry can be vulnerable when LLMs produce stylistically homogeneous content, such as uniform misinformation, in which case the method fails to discriminate between sources. This limitation underscores the necessity of further evaluating and refining stylometric techniques for contemporary detection tasks.

Comparative Analysis and Research Gap

In order to contextualize the present study, it is necessary to compare the main categories of existing detection methods along key dimensions that define their practical relevance. These dimensions include the degree of access required to underlying language models, the dependence on labeled training data, the interpretability of detection outcomes, the adaptability across domains and languages, and the robustness under advanced generation strategies such as finetuning or few-shot prompting. Table 1 provides a structured overview of these categories, enabling a systematic assessment of their relative strengths and limitations.

The comparison reveals critical limitations across existing approaches. Model-dependent methods such as watermarking and distribution matching presuppose either direct control over text generation or access to internal model parameters. While effective in controlled conditions, these techniques are inapplicable when dealing with outputs from external or proprietary systems (Kirchenbauer et al., 2023; Mitchell et al., 2023). Supervised classifiers demonstrate strong performance in in-domain settings (Antoun et al., 2023; Bethany et al., 2024), yet they are constrained by their reliance on large volumes of labeled data, lack of transparency, and limited generalizability to unseen domains or finetuned models (Bakhtin et al., 2019; Li et al., 2023). Feature-based unsupervised methods rely on linguistically interpretable indicators such as lexical richness or syntactic complexity, but they tend to degrade under shifts in genre or authorial style (Ma et al., 2023; Muñoz-Ortiz et al., 2023; Guo et al., 2023).

Stylometric approaches, particularly Burrows' delta and its modifications, combine methodological simplicity with interpretability and have a long tradition in authorship attribution (Burrows, 2002; Hoover, 2004; Rybicki & Eder, 2011; Eder et al., 2016). Initial attempts to adapt them for synthetic text detection indicate that they can differentiate between human-authored and LLM-generated texts in restricted settings (Rebora, 2023; Salnikov & Bonch-Osmolovskaya, 2023; Wang, S. et al., 2024). However, their performance has not been systematically tested in scenarios involving advanced generation techniques such as finetuning or few-shot prompting, and prior work has shown their vulnerability when models produce stylistically uniform outputs, for example in the case of misinformation (Schuster et al., 2020).

Table 1

Comparative Analysis of Detection Methods

Method Category	Model Access Needed	Labeled Data Required	Interpretability	Domain/Language Adaptability	Robustness to Finetuning/Few-Shot Prompting
Watermarking	Yes (owner-controlled)	No	Low	Low (requires pretraining integration)	Poor (cannot be applied post hoc)
Distribution Matching	Yes (requires logits)	No	Moderate	Limited to known models	Poor (fails with unseen or finetuned models)
Supervised Classifiers	No	Yes	Low (black-box)	Moderate (performance drops out-of-domain)	Poor (overfits and degrades on new models)
Unsupervised Feature-Based	No	No	High	Moderate (fragile across genres and domains)	Low (weak cross-domain performance)
Stylometry (e.g., Delta)	No	No	High	Historically applied across registers	Underexplored (particularly with finetuned and few-shot LLMs)

Taken together, these observations underscore a clear methodological gap. There remains a need for an unsupervised detection technique that is both interpretable and computationally efficient, yet at the same time robust to stylistic variation induced by different LLM generation strategies. To address this gap, the present study systematically evaluates stylometric delta with varying feature sets (unigrams, bigrams, trigrams) and distance metrics (Manhattan and cosine). The analysis is applied to human-written and synthetic diary entries generated through few-shot prompting and finetuning. This design makes it possible to assess the interpretability, scalability, and resilience of stylometry-based detection, thereby positioning it as a practical and transparent alternative to model-dependent or classifier-centric approaches.

METHOD

Research Design

The central hypothesis of the current study is that stylometric delta, when applied to the clustering of natural and synthetic texts, can serve as an effective basis for unsupervised detection of machine-generated content. To test this assumption, the experimental design was structured around systematic comparisons between human-authored diary entries and texts generated by large language models through different strategies, including few-shot prompting, finetuning with LoRA adapters, and Direct Preference Optimization (DPO). In addition, a finetuned transformer classifier, RuModernBERT, was employed as a supervised baseline in order to benchmark the performance of the proposed unsupervised approach.

The overall aim of the design was to assess whether stylometric delta can provide both efficiency and robustness in unsupervised detection tasks. By comparing across generation strategies and benchmarking against a supervised baseline, the study sought not only to evaluate the precision of the method but also to test its potential for scalability and generalization in broader applications of synthetic text detection.

Corpus and Data Preparation

The data for the study were extracted from the Prozhito corpus**,** a large archive of Russian diaries and ego-documents. To ensure diversity and representativeness, samples were constructed with consideration of authorial identity and subdomain characteristics. Randomized subsets of the corpus were divided into training and test partitions. The test sets were reserved as authentic examples of natural writing, while the training sets were used both for prompting and for the finetuning of generative models. Several Mistral-based LLMs were then trained on these subsets to approximate the stylistic properties of the original diaries. Using both pretrained LLMs and the finetuned Mistral variants, we generated synthetic corpora designed to mimic natural diaries across multiple stylistic domains.

Domain Choice

Russian diaries were chosen as the primary domain of analysis because of their distinctive linguistic and stylistic properties. Previous research has emphasized the heterogeneity of diary writing, which resists formal unification and encompasses considerable variation in length, narrative style, communicative function, and the cultural and social backgrounds of authors (Bogdanova, 2008). This inherent diversity renders the diary genre an especially challenging and therefore informative target for evaluating text detection methods. Restricting the study to diaries provided a clearly delimited genre while simultaneously allowing the examination of stylistic heterogeneity within that boundary.

Dataset Construction

In total, eleven datasets were constructed, each consisting of 500 texts and divided into three major categories: one category representing authentic human-written diaries and two categories representing synthetic texts. Within the natural category, three datasets were selected and further partitioned into training and test subsets. Training sets were used both for model training and for few-shot generation, whereas test sets were reserved exclusively for stylometric delta experiments as gold-standard examples of natural texts.

The three natural datasets included:

(1) PRISHVIN, consisting of diary entries authored by Mikhail Mikhailovich Prishvin;
(2) RANDOM, composed of entries from twenty randomly selected authors spanning a period of three centuries;
(3) WORK, derived from randomly chosen “work diaries,” which describe professional activities and are characterized by reduced emotional content (Vorobeva et al., 2024).

The WORK dataset was assembled using a classifier trained on materials available at GitHub repository[1]. This dataset includes contributions from more than one hundred authors, which substantially diminishes the strength of individual authorial signals and increases stylistic noise. Consequently, this domain was considered the most difficult for stylometric delta to discriminate. For the classification task, the rubert-tiny2 model was finetuned on one thousand texts randomly sampled across historical periods, achieving a classification precision of approximately 0.94.

[1] https://github.com/laidhimonthegreen/prozhito_thesis?utm_source=chatgpt.com

Synthetic Data Generation

Few-Shot Prompting

To collect synthetic datasets for the second category, we prompted four LLMs to write diaries as if they were Prishvin. We used a few-shot learning method, providing the model with two original Prishvin's diaries for each prompt (Table 2). This approach has already demonstrated better results than regular prompting, making it easier for models to copy the particular style. LLMs used for pseudo-Prishvin diaries generation are: ChatGPT (GPT-3.5), GigaChat (GigaChat-Pro), Vicuna-1.5, Saiga/Mistral.

Finetuning with LoRA and DPO

The last category of our data was composed of synthetic diaries generated by models trained on natural diaries of the first category. As a base model we used the original Mistral-7B (Jiang et al., 2023). First, we trained two models on Prishvin's diaries using two different approaches: LoRA-adapter (Hu et al., 2021) and Direct Preference Optimization (DPO) (Rafailov, 2024).

Every model was trained to write diary entries based on an author specified in the system prompt, an instruction to write a diary entry, and a date (Table 3).

For all generations the fixed set of parameters was used:

(1) maximum length = 1024 tokens,
(2) temperature = 0.7,
(3) top_p = 0.9,
(4) seed = 42.

Training LoRA-adapters proved to be a sufficient method, which helps to shift the model's distribution closer to the target domain, while requiring only some model's parameters (<1%) to be updated.

On the other hand, DPO is often used to explicitly set a model's preferences. This approach is similar to Reinforcement Learning from Human Feedback (RLHF) (Knox & Stone, 2011), used by OpenAI to train GPT, and thus requires at least a pair of inputs during training. In our case every training sample contained two inputs: "chosen" (an original Prishvin diary entry) and "rejected" (a randomly chosen

Table 2

Few-Shot Generation Example

Prompt	Напиши дневниковую запись за 10 января 1921 от лица Михаила Михайловича Пришвина:
	Святочный сон: в лесу на сосне привесил я портрет своей Козочки, волки стали собираться к портрету, а я их стрелять. Разрушитель не дикий человек, а недоучка, тот, кто, выкрав, напр., из физического кабинета спираль Румкорфа, знает, что ее можно приспособить для закуривания папирос. – 1 Р. Снег мелкий, но частый и сильный ветер. Ночью на 29-е – буря и + 5 Р.
	Напиши дневниковую запись за 16 января 1921 от лица Михаила Михайловича Пришвина:
	Искусство есть способность человека изображать предмет своей веры и любви (Христ.). (Возрождение – предмет наслаждений.) Вера без дел мертва, а вера без любви – зла и есть, кажется (надо подумать), основа величайших злодейств. Зло существует на кредит любви. Сатира, пародия – должники искусства.
	Напиши дневниковую запись за 26 января 1921 от лица Михаила Михайловича Пришвина:
Generation result (GPT 3.5)	Сегодня я прочитал долгое и принципиальное письмо от одного из моих адвокатов. Мне предстоит сразу же начать подготовку к защите своей позиции, поэтому я встал рано и занялся прочтением литературы по юридической теории. Я размышляю над тем, как
	использовать закон в мою пользу, чтобы достичь моей цели. Я верю, что правая юридическая аргументация поможет мне достичь моей цели. Да будет мне помощью мудрость и да будет мне поддержкой судьба!

Table 3

Example of Prompt Used for Finetuning

System Prompt	Ты - Михаил Михайлович Пришвин. Ты ведешь дневник.
Prompt	Напиши дневниковую запись за 4 марта 1942 года от лица Михаила Михайловича Пришвина:
Generation result	Продолжение «Весны» в духе Гёльцнера: «после революции» и так далее.
	Как надо заплатить налог: ...

diary entry from the RANDOM dataset). DPO was expected to change the model's preferences and allow it to better capture Prishvin's style by gradually distinguishing it from other authors. However, later experiments showed that the DPO model struggled to learn Prishvin's style, being easily detectable even by the simplest delta.

In contrast, the model trained with LoRA-adapter showed peculiar results. Despite the low quality of generated texts, classic Burrows delta was unable to distinguish them from their natural counterparts. Seeing such promising results, we decided to train another two models using the data from RANDOM and WORK datasets, respectively.

Stylometric Delta Methodology

As a primary method for generated text detection we chose stylometric delta. Stylometry is the application of linguistics designed to evaluate the individual style of an author. Stylometry methods proved to be efficient in authorship attribution of texts (Hoover, 2004; Craig & Kinney, 2009). As author attribution is similar to text classification as "natural" or "generated," stylometry methods could also be productive for our study.

Since the emergence of Burrows' delta in 2002, this method (and its variations) is frequently used for authorship attribution (Stamatatos, 2009). It uses z-scores of normalized word frequencies to calculate distances between texts. Variants of delta are usually made by altering distance measures or normalization procedures (Eder et al., 2016; Argamon, 2008). Delta is now the most established measure in authorship attribution (Rybicki & Eder, 2011).

Previous works (Rebora, 2023; Salnikov & Bonch-Osmolovskaya, 2023) showed that even classic Burrows' delta often demonstrates positive results in distinguishing generated texts. However, these studies are limited by reliance on zero-shot prompting, without LLM finetuning. We aimed to overcome this by employing few-shot prompting, LoRA-adapters, and DPO.

Each dataset was truncated to 6000 tokens, which is sufficient for delta to perform as expected. For each configuration, the top 1000 most frequent n-grams were extracted from the union of all datasets, following best practices (Kestemont, 2014; Evert et al., 2017). Burrows' delta was then calculated as the mean absolute (Manhattan) or cosine distance between vectors. Experiments were repeated for unigrams, bigrams, and trigrams.

Supervised Baseline

As a supervised baseline, we used RuModernBERT**,** a finetuned transformer classifier trained on one thousand randomly sampled texts. This allowed benchmarking of the stylometric delta approach against a state-of-the-art supervised model, ensuring robust evaluation of efficiency, scalability, and generalization potential.

RESULTS

Baseline Classifier Performance

The first stage of the evaluation focused on the performance of the finetuned RuModernBERT model in distinguishing natural diary entries from synthetic texts generated through few-shot prompting, LoRA finetuning, and Direct Preference Optimization (DPO). The classifier was trained in a supervised manner and was designed to assign each text to one of four categories: natural, few-shot, LoRA, or DPO. The quality of clustering, measured by the Adjusted Rand Index (ARI), reached only 0.28, which reflects a relatively weak correspondence between the predicted clusters and the true labels. A closer inspection of the results revealed frequent misclassification of Vicuna and GigaChat outputs as LoRA-generated texts, which illustrates the limited capacity of the model to capture structural distinctions among generation strategies.

Classic Delta (Manhattan and Unigrams)

In the next step, we applied an unsupervised approach based on the classic Burrows' delta. Hierarchical clustering was performed using Manhattan distances calculated over the one thousand most frequent unigrams, following z-score normalization. This configuration produced substantially more coherent clusters and was particularly successful in isolating few-shot outputs irrespective of their generative origin (see Dendrogram 1). Under this setting, the ARI increased to 0.53, which represents a notable improvement compared with the supervised baseline.

Cosine Delta with Unigrams

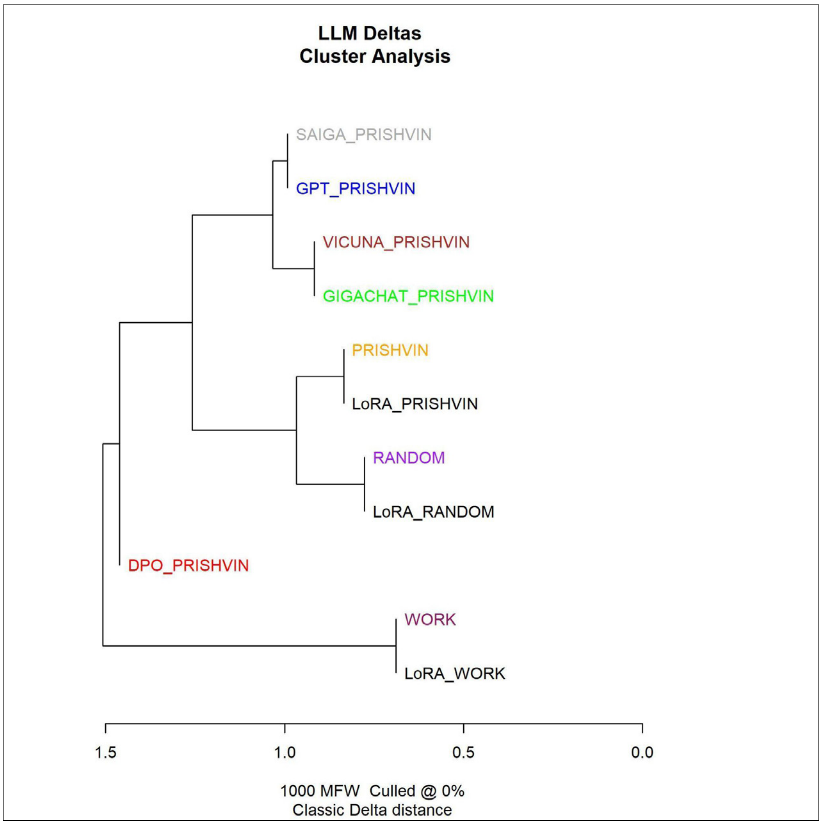
When Manhattan distance was replaced with cosine distance on the same unigram feature set, the overall results deteriorated. The Adjusted Rand Index dropped to approximately 0.12, indicating poor clustering performance. Although certain distinctions were visible, such as the partial separation between the PRISHVIN and LoRA_PRISHVIN datasets, the method failed to provide reliable clustering in more challenging domains. In particular, in the WORK dataset, LoRA-generated and natural texts were grouped together in an ambiguous manner (see Dendrogram 2).

Cosine Delta with Trigrams

To better capture stylistic nuances, we then employed cosine distance on the top 1000 trigrams. This configuration delivered the highest performance: the dendrogram clearly separated synthetic and human-authored texts across all domains (Dendrogram 3). Only a minor misclassification

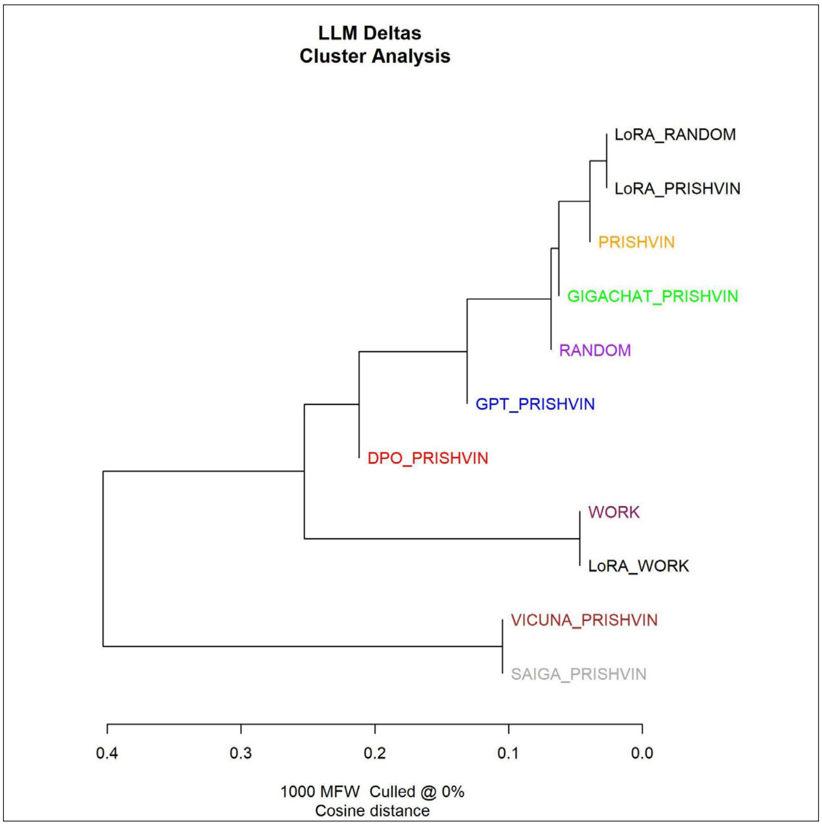
Dendrogram 1

Classic Delta (Manhattan Distance and Unigrams)



Dendrogram 2

Cosine Delta and Unigrams



occurred where LoRA_PRISHVIN clustered with fewshot outputs. Overall, ARI reached approximately 0.70, the highest of all methods.

Comparative Performance

Comparing ARI scores across methods shows a clear progression: the RuModernBERT classifier scored 0.28, Classic Delta with unigrams reached 0.53, Cosine Delta with unigrams dropped to 0.12, while Cosine Delta with trigrams achieved a peak score of 0.70. The Adjusted Rand Index values for all methods are presented in Table 4.

Out of Domain Evaluation

To further examine the general applicability of the trigram cosine delta, we conducted an evaluation using data from the RuATD 2022 generated text detection competition. This dataset included outputs from thirteen different language models, which allowed us to test the method in a more heterogeneous setting. The experiment was carried out in a multiclass configuration using only the validation subset. Since delta methods tend to perform more reliably on longer inputs, we aggregated all texts produced by each model, as well as the human-authored texts, into single datapoints. These datapoints were then compared using the one thousand most frequent trigrams extracted from the combined validation dataset.

The results are visualized in Dendrogram 4, which illustrates the clustering structure produced by the method. The dendrogram provides a clear separation between human-written and machine-generated texts and further demonstrates coherent subgrouping of language models according to their architectural families, such as the ruGPT3 and mT5 clusters. This outcome supports the potential of the proposed method to generalize across datasets with varying domains and stylistic characteristics.

DISCUSSION

Previous stylometry-based research has primarily concentrated on zero-shot detection of prompt-generated synthetic texts. Such studies demonstrated that basic stylometric measures, including Burrows' delta, are capable of distinguishing between human-authored and machine-generated texts, although often only at a superficial level (Rebora, 2023; Salnikov & Bonch-Osmolovskaya, 2023). However, these works did not explore the resilience of stylometry when applied to more advanced generation strategies, in particular finetuning, which is designed to approximate individual authorial style more closely. The results of the present study extend this line of research by demonstrating that stylometric detection remains effective even when LLMs are finetuned, provided that the method incorporates sufficiently complex features and an appropriate distance metric. Specifically, the use of trigram features in combination with cosine distance produced consistently enhanced discrimination.

At the same time, the evaluation carried out in this study has clear limitations. The analysis was restricted to Russian diary texts and to a relatively narrow set of generation techniques. Consequently, the extent to which the trigram cosine delta can be generalized to other genres, such as news reporting or academic writing, or to languages with markedly different syntactic structures, remains uncertain. Its performance against newer LLMs that are capable of producing highly polished and stylistically nuanced outputs also requires further validation. Earlier research has shown that stylometry may fail in contexts where machine-generated texts are deliberately homogenized, for example in the case of misinformation that is stylistically uniform (Schuster et al., 2019). For this reason, the findings presented here should be regarded as encouraging but preliminary, and they must be interpreted within the methodological and domain-specific constraints of the study.

A comparison with alternative methods further illustrates the advantages of the proposed approach. The BERT-based classifier struggled to capture subtle stylistic distinctions, whereas the classical delta method achieved moderate success, particularly in isolating few-shot text clusters. In contrast, the trigram-based cosine delta consistently produced robust separation between natural and synthetic texts, thereby confirming its utility and supporting the central hypothesis of the research. Qualitative inspection of the outputs further corroborates these results. Few-shot generations tend to display overly simplified phrasing and

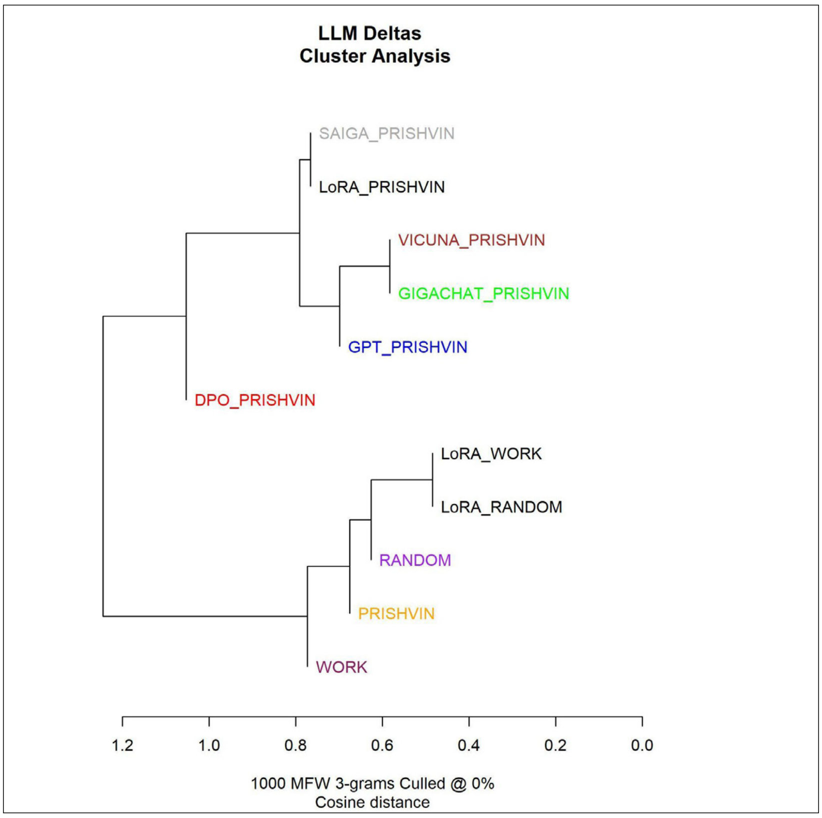
Table 4

Adjusted Rand Scores for Different Methods

Method	Adjusted Rand Score
RuModernBERT (Baseline)	0.2775
Classic Delta	0.5343
Cosine Delta and Unigrams	0.1204
Cosine Delta and Trigrams	0.6983

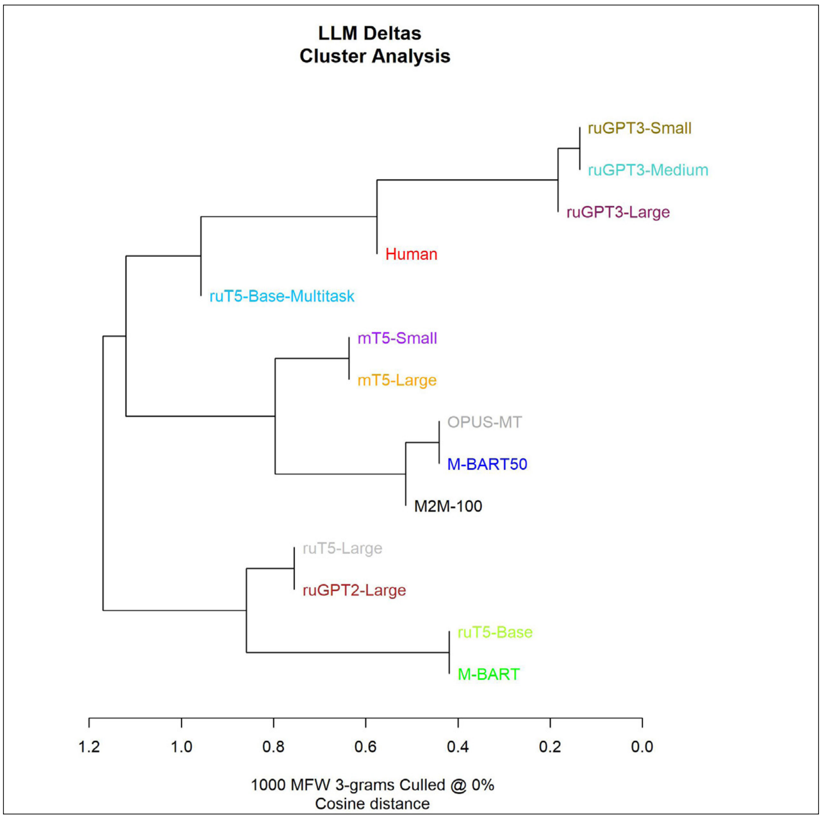
Dendrogram 3

Cosine Delta and Trigrams



Dendrogram 4

Cosine Trigram Delta and RuATD Data



weak stylistic alignment, while LoRA outputs more clearly attempt to replicate authorial style but often lack structural coherence (see Tables 5 and 6). The effectiveness of the trigram-based delta highlights the importance of capturing discourse-level cohesion, including connective structures and syntactic markers, as shown in Table 7. These elements play a critical role in accurate stylistic classification and appear to represent an area where LLMs still exhibit detectable shortcomings.

A central strength of the trigram-based approach lies in its capacity to capture discourse-level cohesion patterns that extend beyond surface fluency and reflect deeper syntactic complexity. In Russian, for example, repeated connective phrases such as «о том, что» function as markers of cohesion and are difficult for LLMs to reproduce consistently. This limitation is particularly evident in the case of finetuned models, which often generate texts that appear locally fluent but lack structural nuance at the global level (Muñoz-Ortiz et al., 2023). By incorporating trigrams, the method effectively encodes these cohesion cues, allowing the detection system to take advantage of weaknesses in the ability of LLMs to replicate authentic discourse-level style.

Equally important is the transparency of the proposed method. Because it relies on ranked and distance-based comparisons of n-grams, the approach aligns with broader research advocating for explainable linguistic techniques in the detection of synthetic text. This feature allows users to trace stylistic anomalies directly, rather than relying exclusively on opaque classification outcomes. Such interpretability enhances the method's practical applicability, especially in settings where auditability, adaptability to diverse domains, and methodological clarity are critical.

The findings of the present study confirm the core hypothesis that stylometric delta, when configured with trigram features and cosine distance, provides a reliable basis for the unsupervised detection of LLM-generated texts. The trigram–cosine configuration achieved an Adjusted Rand Index of approximately 0.70, which represents a substantial improvement over both the RuModernBERT baseline (ARI ≈ 0.28) and the classical unigram-based delta (ARI ≈ 0.53). The advantage was particularly pronounced in complex domains such as multi-author "WORK" diaries, where stylistic noise makes discrimination especially challenging.

Limitations

Nevertheless, our study has limitations. We tested only Russian diary texts, which limits generalization to other genres or languages. Similarly, emerging LLMs (GPT-4o/o1, LlaMA 3.1, Qwen 2.5) might better emulate individual style; our method's robustness against such models remains unproven.

Future work should evaluate wider text types – news, reviews, academic prose – in multiple languages, and measure sensitivity to variables like text length, authorial diversity, and intentional style copying. Testing against stronger foundation models will further assess the generality and lasting relevance of the trigram cosine delta approach.

Table 5

Examples of Texts Generated by Larger LLM Using Fewshot Learning (SAIGA_PRISHVIN) and Smaller LLM with fFnetuned LoRA-Adapter (LoRA_PRISHVIN).

SAIGA_PRISHVIN	LoRA_PRISHVIN
Сегодня я проснулся рано и решил посмотреть на восход солнца. Я вышел на улицу и увидел, что солнце еще не поднялось над горизонтом. Мне было интересно посмотреть, как оно будет взойти сегодня.	У нас доктор сословный, с его зелеными зубами, мертвым лицом, а смуты нет во мне. Я знал, что это он сам даст мне сестрам свое мнение и потом не позвонит по этому поводу: «Документов у нас нет».

Note. Despite producing mostly coherent text SAIGA_PRISHVIN fails to copy Prishvin individual style. On the other hand LoRA_PRISHVIN while also generating some nonsense, better captures Prishvin's individual style.

Table 6

Examples of original and synthetic texts of WORK domain

WORK	LoRA_WORK
Проходило оперативное совещание (по селектору). Энергосистемы повсеместно вышли на максимум энергонагрузок. К 1 января 1975 г. нагрузка будет (как обычно) снижаться, так как многие предприятия страны выполнили, а некоторые и перевыполнили свои планы.	Вышел в эфир первый выпуск телепредставления «Детские годы» о Горчакове. Длинное название ленты: «На смену литорине». Подобные темы очень не любят цензура и у внутрипартийной оппозиции.

Note. Due to domain's specifics (mostly simplicity) stylistic differences between original and generated texts are not very obvious.

Table 7

Randomly Picked Segments of Lemmatized Original and Generated Texts from WORK Domain

WORK	LoRA_WORK
нов заменить брусилов. володя вебер выслать соня 15 руб. и от (брат) колоть я сигара. * 15. * у я сильно заболеть правый бок. фрайнд приехать из крб. уговориться с он на случай надобность воспользоваться ссуда по 1200 акция по 250 р. и поручить он когда быть в берлин попытаться составить синдикат о покупка у я 10 т. акция. жалование i [половина] февра [аля] – 434. после урок бесконечный заседание – о архитектурный рисование – бродский, синайский, юон, я – против тырс. наконец получить согласие бродский на приглашение белкин. вечер телефон с он и с тата. (1) заканчивать прополка кукуруза. (2) гавриленко, тинавая, шаповалов – не работать . некрасово. событь весь день	прилепить полянка и клевер. панно делать и я поставить за она. пять колесо 16 колесо. сухой погода и добавлять зятка. яровый косово сносить. броса вода 12 бараболь по овес дать хуба 2 барабаль. выкорочный колесо 5. серени в прошед. гр. прох. < продолжение отменно стыдный. не оставить ничто никакой вывод ...> < первый сентябрь 1925 год > учить паренек овес спир 3 день. < один время перескакано, продолжиться следующий раз уже полный смысл ответ и возращение ... > петь вчера в поле попасть. пелед навеять. сирника в прошлый. гр. закончить 132 кубок. провотка и ножица. былед. гляз. одежда и рубашка. сено к 10 декабрь. быть весь день дома. много закурить. работать на пик

Note. Frequent trigrams are marked with color. The fragment «с он и с» is actually two frequent trigrams – «с он и» , «он и с». Trigrams of original text are much closer to connectors than those of generated text, thus embedding more syntactic information.

CONCLUSION

This study addressed the challenge of detecting synthetic text across outputs generated through finetuning and few-shot prompting by introducing a lightweight and interpretable detection method based on trigram cosine stylometric delta. The proposed approach achieved an Adjusted Rand Index of approximately 0.70, which substantially outperformed both the finetuned RuModernBERT classifier (ARI ≈ 0.28) and the classical unigram-based delta measure (ARI ≈ 0.53). These results provide strong confirmation of the central hypothesis regarding the effectiveness of higher-order n-gram features combined with cosine distance for unsupervised text detection.

Unlike classifier-centric or transformer-based detectors that depend on extensive labeled data and often operate as black-box systems, the present method relies on transparent and linguistically interpretable features. This characteristic not only enhances its suitability for academic scrutiny but also facilitates practical deployment in real-world applications. The findings further extend earlier stylometric research by demonstrating that trigram-based features capture discourse-level cohesion and syntactic complexity, and that these properties remain discriminative even in the context of finetuned models.

At the same time, the study has several limitations. The evaluation was restricted to Russian diary texts and a specific set of generation strategies, which constrains the scope of generalization to other genres, languages, and more advanced large language models. Future work should therefore test the approach across a wider range of textual domains, stylistic registers, and contemporary LLMs in order to validate its broader applicability.

In conclusion, this research contributes a scalable, transparent, and effective unsupervised technique for the detection of LLM-generated content. By combining interpretability with computational efficiency, the trigram cosine stylometric delta represents a meaningful advancement in the field of synthetic text detection and provides a foundation for further methodological development.

AI DISCLOSURE STATEMENT

During the preparation of this manuscript, the authors used GPT-4o and GPT-5 exclusively for language editing and stylistic refinement. All AI-generated suggestions were reviewed, revised, and approved by the authors. No AI tools were used in the design, analysis, interpretation, or presentation of the research data; all scientific conclusions are based on the authors' independent work, for which they take full responsibility.

DECLARATION OF COMPETING INTEREST

None declared.

AUTHORS' CONTRIBUTIONS

Egor Salnikov: conceptualization; software; formal analysis; methodology; investigation; visualization; writing – original draft; writing – review & editing.

Anastasiya Bonch-Osmolovskaya: conceptualization; data curation ; methodology; project administration; supervision; resources; writing – review & editing.

REFERENCES

Aich, A., Bhattacharya, S., & Parde, N. (2022). Demystifying neural fake news via linguistic feature-based interpretation. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 6586-6599). International Committee on Computational Linguistics

Antoun, W., Mouilleron, V., Sagot, B., & Seddah, D. (2023). Towards a robust detection of Language Model generated text: Is ChatGPT that easy to detect? In *Actes de CORIA-TALN 2023* (vol. 1, pp. 14–27). ATALA.

Argamon, S. (2008). Interpreting Burrows's delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing, 23*(2), 131-147. https://doi.org/10.1093/llc/fqn003

Bakhtin, A., Gross, S., Ott, M., Deng, Y., Ranzato, M. A., & Szlam, A. (2019). *Real or fake? Learning to discriminate machine from human generated text.* arXiv:1906.03351. https://doi.org/10.48550/arXiv.1906.03351

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). ACM. https://doi.org/10.1145/3442188.344592

Bethany, M., Wherry, B., Bethany, E., Vishwamitra, N., & Najafirad, P. (2024*). Deciphering textual authenticity: A generalized strategy through the Lens of Large Language Semantics for detecting human vs. machine-generated text*. arXiv:2401.09407. https://doi.org/10.48550/arXiv.2401.09407

Bogdanova, E. (2008). Language features of diary genre. *Philological Sciences*, *1*(1), 28-33. https://doi.org/10.30853/filnauki

Vorobeva, V., Bonch-Osmolovskaya, A., Kryukov, A., & Podriadchikova, M. (2025). Distant reading of Soviet diaries. *Digital Scholarship in the Humanities, 40*(3), 956–968, https://doi.org/10.1093/llc/fqaf038

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*(3), 267-287. https://doi.org/10.1093/llc/17.3.267

Chhatwal, G. S., & Zhao, J. (2025). Multitask learning for authenticity and authorship detection. *Electronics, 14*(6), 1113. https://doi.org/10.3390/electronics14061113

Craig, H., & Kinney, A. F. (2009). *Shakespeare, computers, and the mystery of authorship.* Cambridge University Press.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Li, L., & Sui, Z. (2022). *A survey on in-context learning*. arXiv:2301.00234. https://doi.org/10.48550/arXiv.2301.00234

Eder M., Rybicki J., Kestemont M. (2016). Stylometry with R: A package for computational text analysis. *The R Journal, 8*(1), 107-121. https://doi.org/10.32614/rj-2016-007

Eder, M., Kestemont, M., & Rybicki, J. (2013). Stylometry with R: A suite of tools. *Digital Humanities 2013: Conference Abstracts* (pp. 487–489). University of Nebraska.

Emi, B., & Spero, M. (2024). *Technical report on the pangram ai-generated text classifier*. arXiv:2402.14873. https://doi.org/10.48550/arXiv.2402.14873

Crothers, E., Japkowicz, N., & Herna, L. (20023). Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access* (vol. 11, pp. 70977-71002). IEEE. https://doi.org/10.1109/ACCESS.2023.3294090

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., & Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities, 32*(suppl_2), ii4-ii16. https://doi.org/10.1093/llc/fqx023

Fraser, K. C., Dawkins, H., & Kiritchenko, S. (2025). Detecting AI-generated text: Factors influencing detectability with current methods. *Journal of Artificial Intelligence Research, 82*, 2233-2278. https://doi.org/10.1613/jair.1.16665

Fröhling, L., & Zubiaga, A. (2021). Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science, 7*, e443. https://doi.org/10.7717/peerj-cs.443

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). Gltr: Statistical detection and visualization of generated text. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 111–116). Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-3019

Gressel, G., Pankajakshan, R., & Mirsky, Y. (2024). Exploiting LLMs for scam automation: A looming threat. In *Proceedings of the 3rd ACM Workshop on the Security Implications of Deepfakes and Cheapfakes* (pp. 20-24). ACM. https://doi.org/10.1145/3660354.3660356

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). *How close is Chatgpt to human experts? Comparison corpus, evaluation, and detection*. arXiv:2301.07597.

Gurioli, A., Gabbrielli, M., & Zacchiroli, S. (2025). Is this you, LLM? Recognizing AI-written programs with multilingual code stylometry. In *2025 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)* (pp. 394-405). IEEE. https://doi.org/10.48550/arXiv.2412.14611

Hoover, D. L. (2004). Testing Burrows's delta. *Literary and Linguistic Computing, 19*(4), 453-475. https://doi.org/10.1093/llc/19.4.453

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions*. ACM Transactions on Information Systems*. https://doi.org/10.1145/3703155

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *Lora: Low-rank adaptation of large language models*. arXiv:2106.09685. https://doi.org/10.48550/arXiv.2106.09685

Hu, X., Ou, W., Acharya, S., Ding, S. H., D'Gama, R., & Yu, H. (2023). TDRLM: Stylometric learning for authorship verification by topic-debiasing. *Expert Systems with Applications, 233*, 120745. https://doi.org/10.1016/j.eswa.2023.1207

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B.* arXiv:2310.06825. https://doi.org/10.48550/arXiv.2310.06825

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. In *International Conference on Machine Learning* (PMLR) (pp. 17061-17084). ML Research Press.

Knox, W. B., & Stone, P. (2011). Augmenting reinforcement learning with human feedback. In *ICML 2011 Workshop on New Developments in Imitation Learning* (vol. 855, No. 3). ACM. https://doi.org/10.1007/978-0-387-30164-8_714

Kumarage, T., Garland, J., Bhattacharjee, A., Trapeznikov, K., Ruston, S., & Liu, H. (2023). *Stylometric detection of AI-generated text in Twitter timelines*. arXiv:2303.03697. https://doi.org/10.48550/arXiv.2303.03697

Kumarage, T., & Liu, H. (2023). Neural authorship attribution: Stylometric analysis on large language models. *In 2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery* (pp. 51-54). IEEE. https://doi.org/10.1109/cyberc58899.2023.00019

Li, Y., Li, Q., Cui, L., Bi, W., Wang, L., Yang, L., Shi, S., & Zhang, Y. (2023). *Deepfake text detection in the wild*. arXiv:2305.13242. https://doi.org/10.48550/arXiv.2305.13242

Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., & Liu, X. (2023). *Is this abstract generated by ai? A research for the gap between ai-generated scientific text and human-written scientific text*. arXiv:2301. 10416, https://doi.org/10.48550/arXiv.2301.10416

Maisto, A. (2025). Collaborative storytelling and LLM: A linguistic analysis of automatically-generated role-playing game sessions. arXiv:2503.20623. https://doi.org/10.48550/arXiv.2503.20623

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning* (PMLR) (pp. 24950-24962). ML Research Press. https://doi.org/10.48550/arXiv.2301.11305

Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting linguistic patterns in human and llm-generated text. *Artificial Intelligence Review, 57*, article number 265. https://doi.org/10.1007/s10462-024-10903-2

Opara, C. (2024). StyloAI: Distinguishing AI-generated content with stylometric analysis. In A.M. Olney, IA. Chounta, Z. Liu, O.C. Santos, & I.I. Bittencourt (Eds.), *Communications in Computer and Information Science, 2151*. Springer. https://doi.org/10.1007/978-3-031-64312-5_13

Przystalski, K., Argasiński, J. K., Grabska-Gradzińska, I., & Ochab, J. (2025). Stylometry recognizes human and LLM-generated texts in short samples. *Expert Systems with Applications, 129001*. https://doi.org/10.1016/j.eswa.2025.129001

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems, 36,* 53728-53741, https://doi.org/10.48550/arXiv.2305.18290*.*

Rebora, S. (2023). GPT-3 vs. Delta. Applying Stylometry to Large Language Models. In *La memoria digitale: Forme del testo e organizzazione della conoscenza*. Atti del XII Convegno Annuale AIUCD (pp. 292-297). AlmaDL. https://doi.org/10.6092/unibo/amsacta/7721

Roy, S. S., Thota, P., Naragam, K. V., & Nilizadeh, S. (2024, May). From chatbots to Phishbots?: Phishing scam generation in commercial Large Language Models. In *2024 IEEE Symposium on Security and Privacy* (pp. 221-221). IEEE Computer Society. https://doi.org/10.1109/SP54263.2024.00182

Rybicki, J., & Eder, M. (2011). Deeper Delta across genres and languages: Do we really need the most frequent words? *Literary and Linguistic Computing, 26*(3), 315-321, https://doi.org/10.1093/llc/fqr031

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). *Can AI-generated text be reliably detected*? arXiv:2303.11156, https://doi.org/10.48550/arXiv.2303.11156

Sahoo, P., Meharia, P., Ghosh, A., Saha, S., Jain, V., & Chadha, A. (2024). A comprehensive survey of hallucination in large language, image, Video and audio foundation models. *Findings of the Association for Computational Linguistics* (EMNLP) (pp. 11709-11724). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.685

Salnikov, E., & Bonch-Osmolovskaya, A. (2023). Using stylometry to detect generated content. In *Proceedings of the DH Russia 2023 Conference* (pp. 176-182). Siberian Federal University.

Schuster, T., Schuster, R., Shah, D. J., & Barzilay, R. (2020). The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics, 46*(2), 499-510, https://doi.org/10.1162/coli_a_00380

Shamardina, T., Mikhailov, V., Chernianskii, D., Fenogenova, A., Saidov, M., Valeeva, A., Shavrina, T., Smurov, I., Tutubalina, E., & Artemova, E. (2022). *Findings of the the ruatd shared task 2022 on artificial text detection in Russian*. arXiv:2206.01583. https://doi.org/10.48550/arXiv.2206.01583

Smith, P. & Aldridge, W. (2011). Improving authorship attribution: Optimizing Burrows' delta method. *Journal of Quantitative Linguistics, 18*(1), 63–88, https://doi.org/10.1080/09296174.2011.533591

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538-556. https://doi.org/10.1002/asi.21001

Tang, R., Chuang, Y. N., & Hu, X. (2024). The science of detecting LLM-generated text. *Communications of the ACM, 67*(4), 50-59. https://doi.org/10.1145/3624725

Wang, S., Cristianini, N., & Hood, B. (2024). Stylometric comparison between ChatGPT and Human Essays. In *Workshop Proceedings of the 18th International AAAI Conference on Web and Social Media*, *Workshop: REAL-Info 2024: First Workshop on Reliable Evaluation of LLMs for Factual Information*. AAAI. https://doi.org/10.36190/2024.24

Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., Whitehouse, C., Afzal, O. M., Mahmoud, T., Sasaki, T., Arnold, T., Aji, A. F., Habash, N., Gurevych, I., & Nakov, P. (2023). M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Vol. 1: Long Papers*, pp. 1369–1407). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-long.83

Wang, Y., Pan, Y., Yan, M., Su, Z., & Luan, T. H. (2023). A survey on ChatGPT: AI-generated contents, challenges, and solutions. *IEEE Open Journal of the Computer Society, 4*, 1-20. IEEE Computer Society. https://doi.org/10.1109/OJCS.2023.3300321

Weerasinghe, J., Seepersaud, O., Smothers, G., Jose, J., & Greenstadt, R. (2025). Be sure to use the same writing style: Applying authorship verification on Large-Language-Model-generated texts. *Applied Sciences, 15*(5), 2467. https://doi.org/10.3390/app15052467

Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., & Wong, D. F. (2025). A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, *51*(1), 275-338. MIT Press. https://doi.org/10.1162/coli_a_00549

Zaitsu, W., & Jin, M. (2023). Distinguishing ChatGPT (-3.5,-4)-generated and human-written papers through Japanese stylometric analysis. *PLoS One, 18*(8), e0288453. https://doi.org/10.1371/journal.pone.0288453

Zhang, B., Liu, Z., Cherry, C., & Firat, O. (2024). *When scaling meets LLM finetuning: The effect of data, model and finetuning method.* arXiv:2402.17193. https://doi.org/10.48550/arXiv.2402.17193

Zhao, X., Li, L., & Wang, Y. X. (2022). *Distillation-resistant watermarking for model protection in NLP.* arXiv:2210.03312. https://doi.org/10.48550/arXiv.2210.03312