

Identifying the Linguistic Fingerprints of Generative AI: A Comparative Analysis of Gemma-2B and Mistral-7B Textual Outputs

Dimitris Bilianos
National and Kapodistrian University of Athens
dbilianos@ill.uoa.gr

Abstract

The rapid proliferation of large language models (LLMs) necessitates new methods for distinguishing between texts generated by different AI systems. This study investigates the existence of unique linguistic fingerprints in the outputs of two prominent open-source LLMs, Gemma-2B and Mistral-7B. A controlled corpus of 400 texts was generated, consisting of 200 responses from each model across a variety of prompts. A feature engineering approach was then employed to quantify key stylistic attributes, including readability scores, lexical diversity, and structural metrics. An SVM classifier was subsequently trained on this feature-rich dataset to perform model attribution. The results demonstrate a high degree of distinguishability between the models, with the classifier achieving an overall accuracy of 89% and a macro F1-score of 87%. These findings provide strong empirical evidence that LLMs may possess measurably distinct linguistic signatures, a crucial insight for the fields of AI text detection, digital forensics, and understanding the stylistic nuances of modern generative models.

1. Introduction

In the rapidly evolving landscape of artificial intelligence, large language models (LLMs) have become ubiquitous, generating content that ranges from creative prose and scientific summaries to code and conversational responses. As these models become more sophisticated and their outputs more stylistically nuanced, a critical new frontier in computational linguistics is emerging: the fine-grained attribution of AI-generated text. While the broad distinction between human and machine-authored content has been a subject of a lot of research, the ability to differentiate between texts produced by various generative models, or even specific versions of the same model, remains more challenging (see among others Mikros et al. 2023). This task is not merely an academic exercise, as it carries profound implications for content verification, intellectual property rights, and the detection of misinformation.

This research tries to address this gap by developing a purely computational framework to identify the unique linguistic fingerprints left by different LLMs. It can thus be hypothesized that despite their shared foundational architectures (e.g., the transformer), distinct training methodologies, data biases, and architectural modifications imbue each model with a unique stylistic and structural signature. These signatures are not readily apparent through simple text

analysis but are embedded in subtle, high-dimensional patterns that can be identified and classified using machine learning techniques; thus by systematically generating a corpus from a set of LLMs and then subjecting this data to a rigorous, feature-rich analysis, there could be demonstrated the feasibility of automated authorship attribution within the generative AI domain. The ultimate goal should be to move beyond mere classification to provide a deeper understanding of the computational linguistic features that define each model's distinct generative style.

This work specifically investigates the linguistic fingerprints of two prominent open-source LLMs, Gemma-2B and Mistral-7B. The experiment is designed to navigate real-world computational limitations in GPU memory and processing time by creating a carefully curated corpus of 400 generated texts (200 from each model).

The remainder of this paper is structured as follows: Section 2 provides a comprehensive literature review on authorship attribution, with a focus on both traditional human-authored text and the nascent field of AI-generated content. Section 3 details the methodology, including the design of the LLM corpus, the extensive set of computational linguistic features extracted, and the machine learning models employed for attribution. Section 4 presents the experimental results. Finally, Section 5 discusses the broader implications of the findings for content authenticity, AI safety, and the future of computational linguistics research, concluding with a summary of this contribution and directions for future work.

2. Literature Review

The field of authorship attribution (AA), a core subfield of computational linguistics, has a rich history spanning over a century. Traditionally, AA has focused on identifying the human author of an anonymous text by analyzing their unique writing style, or "stylometry" (Stamatatos, 2009; Juola, 2008). Early work relied on simple statistical measures, such as word length distribution and sentence length (Mendenhall, 1887), while later methods incorporated a wider array of linguistic features, including part-of-speech tags, n-grams, and function word frequencies (Stamatatos, 2009). The advent of machine learning enabled more sophisticated approaches, treating AA as a classification problem where a model learns to map a text's features to one of a set of known authors. These methods have been highly successful in forensic linguistics, literary scholarship, and intellectual property disputes (Abbasi & Chen, 2008).

The recent proliferation of powerful LLMs has fundamentally challenged this traditional paradigm. The core problem has shifted from distinguishing among human authors to differentiating between human and machine authorship, and more recently, among different generative AI models themselves. Initial research on detecting AI-generated text primarily focused on binary classification (human vs. AI) (e.g., Mikros et al., 2023, first subtask). Newer models produce outputs that are increasingly difficult to distinguish from human-written text (Wu et al., 2025; Kumarage et al., 2024). This evolution underscores the need for more nuanced, sophisticated attribution methods that can handle the growing subtlety of AI-generated content.

Beyond the human-AI dichotomy, the challenge of attributing texts to specific generative models represents a new and complex problem. This task extends traditional authorship attribution to a non-human author set. Early work in this space has explored the use of model-generated watermarks (Kirchenbauer et al., 2023). However, these methods are not robust, as they can be circumvented through paraphrasing, text editing, or the absence of watermarking systems in open-source models. A more promising and generalizable approach involves analyzing the intrinsic computational linguistic properties (e.g. see Mikros, 2025 among others) of the generated text itself, to identify stylistic and structural differences between LLMs, finding variations in syntactic complexity, lexical richness, and the use of discourse markers.

3. Methodology

3.1 Dataset Creation

The foundation of this research is a meticulously constructed, two-class dataset of AI-generated texts¹, tailored to navigate real-world computational constraints. To ensure a robust and scientifically sound analysis, the corpus was generated from two leading open-source large language models, each representing a distinct "author." This selection focused on models that offer a strong balance between generative performance and feasibility within a standard cloud computing environment. The models selected are:

- **Gemma 2B-it:** A compact yet powerful model developed by Google, chosen for its efficiency and strong performance for its size.
- **Mistral 7B-Instruct-v0.1:** A highly competitive model from Mistral AI, representing a step up in scale and complexity from Gemma while remaining accessible for this study.

To create a balanced and representative corpus, a standardized prompting strategy was devised. A set of 200 varied prompts were created, encompassing different text types such as creative writing, technical explanations, and persuasive arguments. To ensure a broad and representative evaluation of the models, the 200 prompts were thus carefully designed to cover a wide range of linguistic styles and cognitive tasks. The set was divided into several key categories to elicit specific types of output, including: creative prompts (e.g., short stories, poetry, and dialogues), analytical prompts (e.g., explaining a complex scientific concept or summarizing a historical event), and persuasive prompts (e.g., arguments for or against a particular viewpoint). This structured approach minimizes the risk of the classifier learning to detect patterns from a single, narrow task, instead forcing it to identify more fundamental and generalizable stylistic fingerprints. Furthermore, the prompts were intentionally crafted to be

¹ In the interest of open science and reproducibility, the complete dataset used for this study will be made publicly available. This includes the original set of 200 prompts, the 400 corresponding texts generated by Gemma-2B and Mistral-7B, and the metadata detailing the generating model for each text. The dataset will be hosted on a public repository, ensuring that future researchers can replicate the findings, build upon this work, and use the corpus for their own analyses.

sufficiently complex to avoid generic, low-effort responses, encouraging the models to produce more nuanced and characteristic outputs.

For each prompt, a text sample was sequentially generated from both of the selected models. This resulted in a total corpus of 400 text documents (200 prompts \times 2 models). All generation parameters, such as temperature and top_p, were kept consistent across models for each prompt to minimize extraneous variables and isolate the inherent stylistic fingerprints. The generated texts were then stored in a structured format, with metadata including the generating model and the original prompt.

This controlled generation process was specifically adapted to overcome the significant computational limitations, such as limited GPU memory and disk space. The texts were sequentially generated and focusing on these two models, it was ensured that any discovered stylistic differences are a true reflection of the models' unique generative patterns and not artifacts of inconsistent data collection or resource-induced errors.

3.2 Feature Engineering and Classification Model

Having the curated dataset of LLM-generated texts, the next critical step was to transform the raw text into a set of quantifiable features that could serve as discriminators for the classification task. The approach in this study focused on a targeted set of computational linguistic features designed to capture the distinct stylistic and structural fingerprints of each generative model.

Feature Engineering

The feature set was derived from well-established metrics in the field of computational linguistics. Rather than a purely lexical approach, a multi-faceted set of features was extracted to capture the nuances of text style. For each text in the corpus, the following features were calculated:

- **Structural Features:** Word count and sentence count were used to capture the basic length and complexity of the generated texts.
- **Readability:** The Flesch-Reading Ease score was computed to measure the general readability of the text, a metric that reflects sentence length and word complexity.
- **Lexical Diversity:** The vocabulary richness of each text was measured by calculating its Type-Token Ratio (TTR), which quantifies the proportion of unique words to total words.
- **Named Entity Count:** The number of unique named entities (e.g., persons, organizations, locations) was counted to serve as a proxy for the factual density or stylistic preference for concrete references.

Classification Model

The task of attributing a text to either Gemma-2B or Mistral-7B was framed as a binary classification problem. Given the nature of the feature set and the size of the dataset, a Support

Vector Machine (SVM)-based algorithm with a linear kernel was chosen. The SVM is a robust and efficient classifier that has proven effective in high-dimensional spaces, making it a suitable choice for this stylometric feature set.

Prior to training the model, the feature data was normalized using StandardScaler to ensure that no single feature with a larger magnitude would disproportionately influence the classifier. The dataset was then partitioned into a training set (80%) and a testing set (20%) to facilitate rigorous model validation and performance evaluation on unseen data.

4. Experimental Results

The experimental results validate the core hypothesis that Gemma-2B and Mistral-7B possess distinct and measurable linguistic fingerprints. The trained SVM classifier demonstrated a high degree of effectiveness in attributing text to the correct generative model based on the engineered features.

Overall Classifier Performance

The model's overall performance was evaluated on a held-out test set comprising 80 texts (44 from Gemma and 36 from Mistral). The classifier achieved a high level of accuracy and a strong F1-score, indicating robust performance across both classes. The primary metrics are summarized in Table 1.

Table 1: Primary metrics	
Metric	Value
Accuracy	0.89
F1-Score	0.87

The accuracy of 89% signifies that the model correctly identified the originating model for nearly nine out of every ten texts it was presented with. This is a highly encouraging result for a classification task of this nature. The F1-Score of 0.87 confirms that this high accuracy is not a

result of a bias towards one class but reflects a balanced performance between precision and recall.

Detailed Per-Class Analysis

A more granular view of the classifier’s performance is provided by the classification report, which breaks down the metrics for each individual model.

Table 2: Classification report		
Metric	Gemma (Class 0)	Mistral (Class 1)
Precision	0.87	0.91
Recall	0.93	0.83
F1-Score	0.90	0.87
Support	44	36

The results indicate that while the model performed well for both classes, there were subtle differences in its attribution performance. The classifier was particularly effective at correctly identifying texts generated by Gemma, as shown by its high recall of 0.93. This means that 93% of the actual Gemma texts in the test set were correctly classified. Conversely, the model was slightly more precise in its predictions for Mistral, achieving a precision of 0.91. This suggests that when the classifier labeled a text as being from Mistral, it was correct 91% of the time.

In summary, the experimental results provide compelling quantitative evidence that the linguistic features extracted are sufficient to train a machine learning model to reliably distinguish between the textual outputs of Gemma-2B and Mistral-7B.

5. Discussion and Conclusion

The findings in this study provide compelling evidence that generative language models, even those of a similar scale, possess unique and measurable linguistic fingerprints. The success of the SVM classifier in distinguishing between Gemma-2B and Mistral-7B with high accuracy has significant implications for several key areas of research and practical application.

5.1 Broader Implications

The broader implications of this study can be divided into three parts; content authenticity, AI safety and computational linguistics. These implications are discussed in detail below.

Content authenticity: The ability to reliably attribute a text to its generative model presents a powerful new tool in the fight against misinformation and the rise of inauthentic content. As AI-generated text becomes more pervasive, methods for automated detection are essential. This research demonstrates that even without access to internal model parameters, a simple set of stylistic features can serve as an effective means of content provenance, helping to authenticate information and identify the source of synthetic media.

AI safety: The existence of these linguistic fingerprints opens up a new avenue for auditing and evaluating LLMs. Distinctive stylistic patterns could be a manifestation of underlying biases in a model's training data or architectural design. The methodology of this study could be adapted to serve as a diagnostic tool for AI developers to detect and mitigate undesirable stylistic biases, such as overly simplistic, emotionally manipulative, or dogmatic writing patterns, before a model is deployed to the public.

Computational linguistics: This work demonstrates that the problem of distinguishing between non-human authors can be effectively solved using a data-driven, machine learning approach with a minimal set of interpretable features.

5.2 Contributions and Future Work

This study introduces a reproducible research pipeline for generating and analyzing LLM outputs, demonstrating a new paradigm for scientific inquiry. Strong quantitative evidence is provided that two leading open-source LLMs, Gemma-2B and Mistral-7B, have demonstrably distinct linguistic fingerprints that are classifiable with high accuracy. Also, in the spirit of open science, the complete corpus of prompts and generated texts is made publicly available to serve as a resource for the broader research community.

Looking ahead, several directions for future work could build upon these findings. The methodology could be scaled to include a larger number of LLMs to explore the complexities of multi-class attribution. Furthermore, future research could investigate the causal links between a model's training data or architectural choices and the specific linguistic fingerprints that it develops. Finally, extending this methodology to detect hybrid texts, that is, mixtures of human- and AI-generated content, would be a crucial next step in applying this research to real-world content authentication challenges.

References

- Abbasi, A., & Chen, H. (2008). Writeprints: A Stylometric Approach to Author Identification. *IEEE Intelligent Systems*, 23(1), 38–47.
- Juola, P. (2008). Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1(3), 233–338.
- Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. *arXiv preprint arXiv:2301.10237*.
- Kumarage, T., Agrawal, G., Sheth, P., Moraffah, R., Chadha, A., Garland, J., & Liu, H. (2024). A survey of ai-generated text forensic systems: Detection, attribution, and characterization. *arXiv preprint arXiv:2403.01152*.
- Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, 9(220), 237-249.
- Mikros, G. (2025). Beyond the surface: stylometric analysis of GPT-4o's capacity for literary style imitation. *Digital Scholarship in the Humanities*, 40(2), 587-600.
- Mikros, G. K., Koursaris, A., Bilianos, D., & Markopoulos, G. (2023). AI-Writing Detection Using an Ensemble of Transformers and Stylometric Features. In *IberLEF@ SEPLN*.
- Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538–556.
- 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.