The Aesthetics of Generation

Andrei Skobtsov

Corso di Informatica Magistrale, Università' degli Studi di Milano.

Abstract

The rapid advancement of Large Language Models (LLMs) prompts a critical question: do these models possess a recognizable, persistent stylistic "voice"? This study investigates the aesthetics of machine generation by quantitatively profiling the stylistic identities of four major LLMs (GPT, Claude, LLaMA, and Mistral) across four distinct genres (narration, argumentation, description, and dialogue). Using the StyloMetrix tool, I extracted syntactic and grammatical features from a corpus of generated texts. A Random Forest classifier was employed to evaluate stylistic identifiability. SHAP (SHapley Additive exPlanations) analysis revealed that these signatures are deeply rooted in subconscious syntactic structures, such as Part-of-Speech distributions, rather than overt vocabulary. However, a "Leave-One-Genre-Out" robustness evaluation demonstrated that these stylistic fingerprints are fragile. While a faint stylistic baseline persists across structurally similar prose, the degree of stylistic retention proves to be highly genre-dependent. The models' stylometric identifiability experiences a severe degradation when subjected to the distinct structural constraints of out-of-distribution formats like dialogue, whereas this classification drop is notably less pronounced among continuous prose genres. Visualized through t-SNE topological landscapes, these findings mathematically suggest that an LLM's "style" is not an immutable trait, but is highly dependent on the structural parameters of the prompt.

Keywords: Large Language Models, Stylometry, Authorship Attribution, Feature Interpretability, Generative AI

1 Introduction

The rapid evolution of Large Language Models (LLMs) has fundamentally blurred the boundaries between computational generation and human creativity. As models such as GPT, Claude, LLaMA, and Mistral achieve near-perfect syntactic fluency,

research is increasingly shifting from evaluating basic text coherence to investigating deeper, more nuanced expressive capabilities. This project explores the "Aesthetics of Generation": the hypothesis that LLMs exhibit distinctive stylistic identities in their generated outputs. Beyond surface-level metrics of quality, I ask whether models possess a recognizable "voice" that persists across tasks, topics, and prompts, and reflect on the philosophical dimension of style when it emerges from probabilistic generation rather than human intention.

Stylometry, or computational stylistics, provides the methodological foundation for this investigation. Traditionally, stylometry relies on the statistical analysis of linguistic features, extracted from a collection of texts, to characterize the unique style of an author or document. These methods draw out subtle differences and similarities invisible to the naked eye, relying on lexical, grammatical, and syntactic metrics to group texts based on their linguistic affinity. Advanced tools like StyloMetrix have recently been developed to extract normalized stylometric text representations covering grammar, syntax, and lexicon. These vectors have proven highly effective as inputs for machine learning classifiers, achieving robust results in content and genre classification tasks.

In recent years, stylometric methodologies have been adapted to address the proliferation of artificial text. Foundational work by Zellers et al. (2019) on defending against neural fake news established the critical need for systems capable of identifying machine-generated prose. This has evolved into specialized authorship attribution systems, such as the StyloAI framework proposed by Opara (2024), which distinguishes AI-generated content through targeted stylometric analysis. Similarly, Kumarage et al. (2023) demonstrated the efficacy of stylometric detection for identifying AI-generated text within the constrained, highly specific context of social media timelines.

However, while existing literature predominantly focuses on the binary classification of human versus machine text, there is a gap in profiling the idiosyncratic stylistic "fingerprints" of individual LLMs against one another. This project aims to bridge that gap by characterizing the specific stylistic identities of four major models: GPT, LLaMA, Claude, and Mistral.

To accomplish this, I generated a balanced corpus of model outputs across four distinct genres: narration, argumentation, dialogue, and description. By extracting quantitative stylometric features, ranging from part-of-speech ratios and lexical diversity to syntactic depth, I perform a comparative evaluation using dimensionality reduction and Random Forest classifiers. Ultimately, this study aims to determine not only if models have a core stylistic baseline, but also how robust these signatures remain when subjected to the out-of-distribution structural constraints imposed by different textual genres.

2 Research Question and Methodology

2.1 Research Question and Objectives

The primary objective of this project is to quantitatively investigate whether Large Language Models (LLMs) possess an inherent and distinct stylistic identity that persists independent of the semantic content of their output.

This leads to the formulation of two core research questions: This leads to the formulation of two core research questions:

1. **Stylistic Identifiability:** Can quantitative stylometric features (e.g., lexical diversity, syntactic structures, part-of-speech distributions) reliably distinguish the outputs of different LLMs to establish a baseline stylistic fingerprint?
2. **Stylistic Robustness:** Are these stylistic signatures immutable characteristics of the models, or do they degrade when subjected to the structural constraints and formatting shifts imposed by different textual genres?

To address these questions, the project proposes a modular, object-oriented pipeline encompassing corpus generation, feature extraction, and comparative evaluation using machine learning classification and dimensionality reduction.

2.2 Corpus Generation

To ensure stylistic comparability, a balanced corpus of machine-generated text was created under strictly controlled, uniform prompts. I targeted four distinct models representing both open-source and proprietary architectures: GPT-4o-mini (OpenAI), Meta-Llama-3.1-8B-Instruct (Together AI), Claude-Sonnet-4-6 (Anthropic), and Mistral-small-latest (Mistral).

For each model, texts were generated across four distinct genres, using one constant prompt per genre:

- *Narration:* "Write a short story about a robot learning to paint."
- *Argumentation:* "Argue for or against the use of AI in creative writing."
- *Description:* "Describe a sunset over the ocean in vivid detail."
- *Dialogue:* "Write a dialogue between two friends discussing climate change."

The generation logic was encapsulated within a custom Python package. Specific Object-Oriented API wrappers were built for OpenAI, Anthropic, Mistral, and Together AI. A temperature of 0.7 was maintained across all API calls to balance deterministic adherence to the prompt with natural linguistic variance. The resulting corpus consists of 320 documents (20 texts per genre, per model).

2.3 Feature Extraction

To quantify stylistic variation, I utilized StyloMetrix, an open-source multilingual tool designed for deep stylometric analysis. While neural embeddings (like Word2Vec or BERT) capture semantic meaning, they operate as "black boxes" that obscure specific grammatical choices. StyloMetrix resolves this by generating interpretable, normalized vectors based on over 100 grammatical, lexical, and syntactic rules.

A custom preprocessing module was built to pass the raw text corpus through the StyloMetrix pipeline. The resulting feature matrix captures nuanced linguistic behaviors, including part-of-speech ratios (e.g., noun-to-verb ratios), syntactic depth, punctuation frequency, specific pronoun usage, and sentence length distribution. These features were subsequently merged with the corpus metadata to form a unified dataset for spatial and predictive analysis.

2.4 Comparative Evaluation and Robustness Testing

The analytical methodology is divided into three consecutive phases:

1. **Dimensionality Reduction and Visualization:** Because the extracted stylometric fingerprints exist in a 196-dimensional space (representing 196 distinct grammatical features), direct visual interpretation is mathematically impossible. To resolve this, I employ Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE). These algorithms compress the high-dimensional data into a two-dimensional plane while preserving the underlying linguistic relationships, ensuring that texts with similar grammatical structures remain plotted in close proximity. This allows to visually map the "aesthetic landscapes" of the models and observe their natural stylistic territories.
2. **Base Classification (In-Distribution):** A Random Forest Classifier is trained on a stratified split of standard prose texts (Narration, Argumentation, Description). Performance metrics (Accuracy, Precision, Recall, F1-Score) establish the baseline identifiability of the models' core stylistic fingerprints.
3. **Cross-Genre Robustness Testing (Out-of-Distribution):** To evaluate the fragility of the stylistic signatures, I employed a "Leave-One-Genre-Out" cross-validation strategy. The classifier is trained on three genres and subsequently evaluated on the completely unseen fourth genre. This rigorously tests whether the stylistic features learned from one structural format (e.g., standard prose) can successfully generalize to a structurally alien format (e.g., dialogue).

Following the methodological benchmarks established by Okulska et al. (2023) in the StyloMetrix documentation, I decided to use the Random Forest Classifier with 200 estimators, as decision trees are particularly well-suited for interpreting the non-linear grammatical ratios present in stylometric vectors. Additionally, SHAP (SHapley Additive exPlanations) values and Gini feature importances are extracted from the Random Forest model to determine the precise lexical and syntactic features driving the classification boundaries.

3 Experimental Results

The experimental evaluation was conducted around 3 objectives: an analysis of baseline stylistic identifiability across four distinct training permutations, a critical interpretation of feature variance, and a rigorous robustness evaluation across out-of-distribution textual genres.

3.1 Baseline Stylistic Identifiability

To establish whether LLMs possess a baseline stylistic identity, I executed a holistic "Leave-One-Genre-Out" methodology. Rather than training a single classifier, four separate Random Forest classifiers were trained. In each iteration, the model was trained on a stratified 80/20 split of three genres, holding the fourth genre completely unseen for later robustness testing.

Evaluated on their respective in-distribution test sets, all four classifiers achieved exceptional macro-accuracies, consistently ranging between 96% and 98%. This near-perfect identifiability across all structural permutations strongly supports the hypothesis that LLMs do not write "neutrally." Regardless of which specific prose genres are combined, the models embed distinct, quantifiable stylistic fingerprints into their outputs that are highly recognizable in-distribution.

3.2 Feature Interpretability and Model Variance

To interpret the mechanics of these stylistic fingerprints, I extracted both Gini feature importances and SHAP (SHapley Additive exPlanations) values across the four iterations. A comparative analysis of these metrics reveals two critical insights regarding AI stylistics and machine learning interpretability.

First, the dominant stylometric features are fluid; they shift dynamically depending on which genres comprise the training data. There is no single universal variable for LLM identification. Instead, the models' stylistic identities exist as a contextual web of syntactic and lexical habits that shift based on the structural environment.

Second, a consistent mathematical discrepancy emerged between Gini importance and SHAP values. For example, in the classifier trained without Dialogue texts, Gini importance ranked `ST_TYPE_TOKEN_RATIO_LEMMAS` (the overall ratio of unique lemmas to total tokens) as the dominant feature. Conversely, SHAP analysis ranked `L_CONT_T` (the absolute count of unique lexical content words, such as nouns and verbs) as the primary driver.

This discrepancy exposes a well-documented bias in tree-based models: Gini importance measures the reduction of impurity during training and artificially inflates continuous variables with high cardinality (like decimal-based type-token ratios). SHAP, utilizing game theory, evaluates the model post-training to measure a feature's actual marginal contribution to the final prediction. Therefore, the SHAP values provide a more unbiased reflection of the model's logic, revealing that raw content vocabulary richness (`L_CONT_T`) exerts a stronger influence on identifying the LLM's true voice than broad diversity ratios.

3.3 Aesthetic Landscapes

To visually map these stylistic fingerprints, I applied t-Distributed Stochastic Neighbor Embedding (t-SNE) coupled with Kernel Density Estimation (KDE) to the 196-dimensional StyloMetrix vectors. This generated topographical "Aesthetic Landscapes."

When mapped by model, the landscape displays distinct "territories" for GPT, Claude, LLaMA, and Mistral, with dense, isolated topographical peaks indicating highly concentrated stylistic habits. However, when the exact same vector space is mapped by *genre*, a critical structural divide emerges. The standard prose genres (Narration, Description, Argumentation) form a massive, overlapping continent, indicating shared structural DNA. Conversely, the Dialogue texts form a completely isolated topological island, visually foreshadowing the limitations of the models' stylistic permanence.

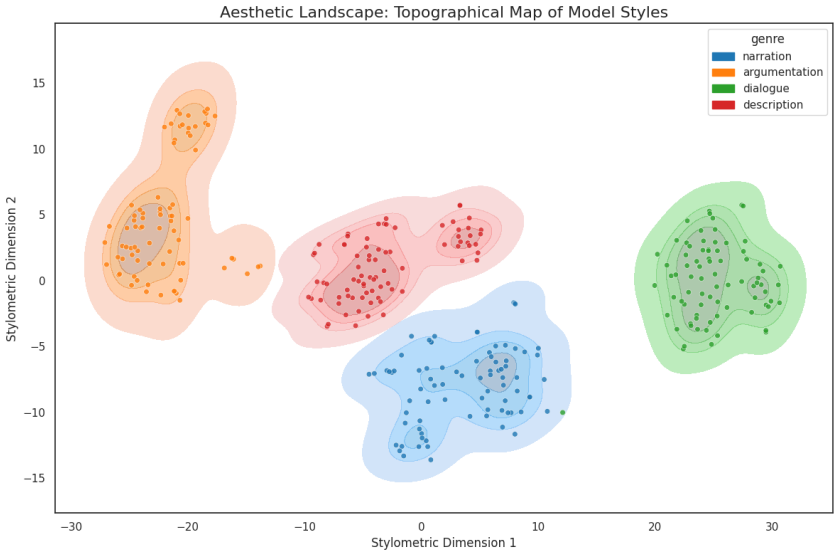

Fig. 1 Aesthetic Landscape mapped by Genre (t-SNE and KDE). Standard prose formats form an overlapping structural continuum, while Dialogue forms an isolated topographical island.

3.4 Stylistic Robustness (Cross-Genre Evaluation)

To determine if these stylistic signatures are immutable traits or highly fragile masks, the four classifiers were evaluated on their respective unseen, out-of-distribution genres.

The results revealed two distinct phenomena:

1. **The Prose Bleed (Partial Robustness):** When holding out structurally similar prose formats, the classifiers maintained accuracies well above the random guessing baseline of 25%. Accuracy settled at 53% for Argumentation, 51% for Narration, and 44% for Description. While representing a significant degradation from the 96–98% baselines, these scores statistically indicate that a faint, underlying stylistic signature "bleeds" across different types of prose. However, robustness varied notably between the models: GPT and Mistral consistently exhibited lower identifiability across unseen genres compared to Claude and LLaMA, with their recall occasionally dropping to 0% in specific cross-genre evaluations.
2. **Severe Performance Degradation:** The most profound finding occurred when the classifier was tested on unseen Dialogue texts. Rather than maintaining the partial robustness seen in other genres, accuracy degraded significantly to 31%, characterized by a single-column confusion matrix where the classifier overwhelmingly defaulted to predicting Claude. Statistically, the model experienced severe covariate shift: the rigid structural constraints of dialogue completely overwrote the models' baseline prose fingerprints.

These results suggest a fundamental limitation in AI stylistics: an LLM's "style" is not a unified, persistent voice, but rather a set of flexible, probabilistic modes that are heavily subordinated to the structural constraints of the user's prompt.

Table 1 Cross-Genre Robustness Evaluation (Macro Averages)

Evaluation Set	Accuracy	Precision	Recall	F1-Score
Average Baseline (In-Distribution)	0.98	0.98	0.98	0.98
Argumentation (Out-of-Distribution)	0.53	0.63	0.52	0.46
Narration (Out-of-Distribution)	0.51	0.55	0.51	0.51
Description (Out-of-Distribution)	0.44	0.39	0.44	0.37
Dialogue (Out-of-Distribution)	0.31	0.16	0.25	0.12

Note: The Dialogue evaluation resulted in a severe performance degradation, heavily skewing precision and F1-scores.

4 Concluding Remarks

This project investigated the "Aesthetics of Generation" by asking whether Large Language Models possess a persistent, recognizable stylistic identity. Our experimental results confirm that LLMs do not simply generate neutral text; rather, they embed distinct, quantifiable stylistic fingerprints into their outputs. When evaluated within the bounds of standard prose, these signatures are highly identifiable, driven not just by overt vocabulary choices, but by underlying grammatical structures such as syntactic depth and Part-of-Speech distributions.

However, the critical finding of this study is that these stylistic identities are heavily genre-dependent. The robustness evaluation demonstrated that an LLM's "voice" is not a static, universal trait. While a partial stylistic baseline persists across structurally similar prose formats, the models' stylometric identifiability degrades significantly when subjected to out-of-distribution textual formats, such as dialogue. In these instances, the rigid structural constraints of the prompted genre override the model's baseline stylistic habits. Ultimately, this suggests that what we perceive as "style" in LLMs is highly flexible and dynamically adapts to the structural parameters of the user's prompt, rather than remaining constant across all types of text.

Future Work and Methodological Improvements

To build upon these findings and achieve more robust cross-genre classification results, future research should explore the following methodological improvements:

1. **Corpus Scaling and Dimensionality Reduction:** The current study relies on a constrained dataset (20 generations per prompt per model), which, when mapped against 196 stylometric features, introduces a risk of overfitting due to high dimensionality. Future iterations must significantly expand the corpus volume to ensure the statistical reliability of the classifier and prevent the decision trees from memorizing sample-specific noise.

2. **Stratified Multi-Domain Training:** Instead of conducting a strict out-of-distribution holdout test, future models could be trained on a proportionally stratified dataset containing all genres. This would allow the classifier to learn the holistic stylistic profile of an LLM across multiple modes (e.g., learning both GPT's descriptive patterns and its conversational structures simultaneously) rather than attempting to extrapolate one from the other.
3. **Prompt Diversity and Semantic Decoupling:** Because this study utilized a single, uniform prompt per genre, there is a risk that the classifier inadvertently learned prompt-specific semantic artifacts rather than pure stylistic syntax. Future studies must employ intra-genre prompt diversity, using dozens of vastly different prompts within the same genre category,to successfully decouple a model's underlying stylistic baseline from the semantic content of the prompt.

AI Acknowledgments. I would like the acknowledge the use of AI in this project. In particular Gemini and Deepseek. AI has been used to help with generating some of the code, especially in the visualisation aspect , the debugging, help with theoretical aspects and help with LaTex. All parts used with AI have been reviewed personally.