Fingerprinting Large Language Models Using Signal Processing

Daniel McGonigle
Advisor: Dr. Joris Roos

University of Massachusetts Lowell, Mathematics Department
Lowell, MA

November 30, 2025

Acknowledgments. I would like to express my deepest gratitude to Dr. Joris Roos for his guidance, encouragement, and invaluable feedback throughout this project.

Contents

1. Introduction

Since the release of GPT-3 [2], large language models (LLMs) based on transformer architecture have revolutionized natural language processing by enabling fluent text generation that can closely mimic human style and reasoning. As the distinction between human- and machine-authored text becomes increasingly subtle, reliable methods for text that was produced by LLMs has become a significant challenge. Furthermore, there is value in attributing text to specific models, or identifying "fingerprints" imparted by particular models that aid in attribution. Applications range from academic integrity and misinformation tracking to model auditing and authenticity verification.

In this project, we explore an alternative perspective: that each model's generation process may leave a measurable spectral signature when its token probability sequence is treated as a signal. Specifically, we hypothesize that human and model text differ in the temporal and frequency-domain characteristics of these probability signals due to differences in attention dynamics and sampling noise.

To investigate this hypothesis, we conducted a systematic comparison between human-written and LLM-generated text using Fourier analysis and wavelet transforms applied to token log-probability sequences. The analysis focused on identifying characteristic frequency bands, entropy levels, and power-spectrum shapes that could distinguish machine- from human-generated text. Complementary statistical and distributional metrics aimed at probing frequency characteristics were also used to evaluate the separability of these groups in a non-spectral space.

There are two goals in this work: The first goal is to assess whether frequency-domain analysis provides meaningful discriminatory power between human and model text. The second is to establish a foundation for spectral fingerprinting, with the hope that this can be done in a model-agnostic manner. The findings of this study demonstrate some promising signal-level regularities that suggest LLMs possess spectral patterns across generations, motivating future research into cross-model generalization, temporal dynamics of attention mechanisms, and the integration of spectral features into broader model-audit frameworks.

2. Background and Related Work

This body of works covers a lot of ground from LLMs and detecting generative content, to tools for analyzing signals in the frequency domain, including computer science methodologies for classification. This chapter serves as a brief introduction to these topics and discusses some of the relevant work.

2.1. Large Language Models.

LLMs are autoregressive neural networks trained to predict the next token in a sequence given all preceding context. During generation, each output token is sampled from a probability distribution $P(t_i \mid t_{<i})$, representing the model's estimated likelihood of possible continuations at position i. These token-level probabilities capture a model's evolving internal state and confidence: when the model is highly certain, the distribution is sharply peaked; when uncertain, it is flatter. The logarithm of these probabilities, or *log-probabilities*, are particularly useful because they linearize multiplicative relationships, stabilize numerical variation, and directly reflect the additive structure of sequence likelihoods.

2.2. Detecting Machine-Generated Text.

Most existing detection techniques focus on lexical or syntactic cues, statistical irregularities such as word, phrase or punctuation probabilities, or leveraging neural networks as in [12]. Some approaches targeting specific data domains have showed limited success, as in DetectRL [11]. Zhang et al. [13] introduced

a zero-shot detection approach that operates directly on token probability distributions, showing that simple statistical measures such as likelihood variance and divergence between human and model token-prob histograms can achieve strong performance (AUROC ≈ 0.9) when distinguishing GPT-3 and ChatGPT text from human writing. Yet these methods often fail to generalize across models or fine-tuning conditions.

2.3. **Analyzing Signals in the Frequency Domain.** Signal analysis in the frequency domain provides a powerful framework for understanding the underlying structure, periodicity, and energy distribution of time-varying data. Rather than analyzing a signal $x(t)$ in the time domain—where the focus is on its instantaneous amplitude or value at each time point—frequency domain analysis decomposes the signal into constituent sinusoids of different frequencies, allowing researchers to study the spectral content and dominant oscillatory components.

2.3.1. *The Fourier Transform and Its Discrete Counterparts*

The *Fourier Transform (FT)* expresses a signal as a sum of complex exponentials, mapping it from the time domain to the frequency domain [8]. For a continuous signal $x(t)$, the Fourier Transform is defined as

$$X(f) = \int_{-\infty}^{\infty} x(t)\, e^{-j2\pi ft}\, dt, \tag{1}$$

where $X(f)$ represents the complex-valued frequency spectrum. In analyzing the token-level log-probabilities of LLM output, the "time" domain for this research consists of discrete token positions. This use of discrete samples requires us to utilize the *Discrete Fourier Transform (DFT)*, which is defined:

$$X_k = \sum_{n=0}^{N-1} x_n\, e^{-j2\pi kn/N}, \quad k = 0, 1, \ldots, N-1. \tag{2}$$

The DFT converts a sequence of N samples into N complex coefficients corresponding to frequency bins. Efficient computation of the DFT is achieved using the *Fast Fourier Transform (FFT)* algorithm [3], which reduces computational complexity from $O(N^2)$ to $O(N \log N)$. The FFT underpins much of spectral analysis, providing useful tools for power spectral density estimation, filtering, and feature extraction in both scientific and engineering applications.

2.3.2. *Limitations of Purely Frequency-Domain Analysis*

While the Fourier Transform is effective for stationary signals—those whose frequency composition does not change over time—it is less suited for *nonstationary* or *transient* data, where the frequency content varies. The Fourier spectrum captures signal patterns by determining global frequency coefficients, but loses features that are temporally localized. This limitation motivated the development of *time-frequency representations*, which aim to capture how frequency content evolves over time.

2.3.3. *Wavelet Transforms and Time-Frequency Analysis*

The *Wavelet Transform (WT)* addresses the shortcomings of the Fourier framework by providing a multi-resolution view of a signal. Instead of fixed sinusoidal bases, wavelet analysis uses localized "wavelets"—functions that are dilated and translated versions of a

mother wavelet—to analyze both short, high-frequency events and long, low-frequency trends [6], [8].

For a continuous-time signal $x(t)$, the *Continuous Wavelet Transform (CWT)* is defined as

$$W(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t)\, \psi^*\left(\frac{t-b}{a}\right) dt, \tag{3}$$

where a and b denote the scale and translation parameters, respectively, and $\psi(t)$ is the mother wavelet. The *Discrete Wavelet Transform (DWT)* discretizes these parameters, yielding efficient computational schemes with compact representations of the signal across scales.

Wavelet analysis has proven useful for analyzing signals with nonstationary or multi-scale structure—ranging from geophysical and biomedical signals to linguistic data [1].

3. Methods

In this research, these per-token log-probabilities are treated not merely as statistical outcomes but as a *temporal signal* that evolves as the model generates text. Each step in generation corresponds to a new sample in a discrete-time series whose fluctuations encode the model's shifting certainty, stylistic rhythm, and contextual transitions. This framing allows the use of signal-processing tools—such as Fourier and wavelet transforms—to examine structure in the *frequency domain* rather than only the token-distribution domain. If model outputs differ systematically from human writing in the smoothness, periodicity, or spectral composition of their log-probability sequences, these differences can be interpreted as latent *fingerprints* of the model's internal generative dynamics. By analyzing log-probabilities as signals, we aim to uncover whether LLMs exhibit distinctive frequency-domain patterns that remain stable across text samples and model families, providing a potential foundation for model attribution and authenticity detection.

3.1. **Dataset Curation.**

A rigorous and transparent dataset curation process is essential for any research endeavor involving linguistic or generative text analysis. This process encompasses corpus selection, data acquisition, cleaning, annotation, and metadata management. For the present study, the goal was to construct a corpus that balances diversity of authorship, disciplinary representation, and writing quality, while maintaining consistency in format and linguistic features suitable for quantitative and spectral analysis. We sought several key characteristics to amplify signal differences related to LLM patterns and minimize signal differences based on domain-specific text, formatting errors or technical jargon:

1. To prevent "leakage" of LLM-generated text into the human-generated corpus, we select datasets that were curated before 2020.
2. We avoid signal differences based on poor writing style by selecting documents that are professionally or semi-professionally written.
3. Care is taken to avoid patterns that are irrelevant, such as structured elements (bibliography, table of contents, etc).

3.1.1. ***Source Corpus: Reuters 50/50 (Subset of RCV1)***

The primary dataset used in this study is the *Reuters 50/50* corpus, a balanced subset derived from the *Reuters Corpus Volume 1 (RCV1)* [4]. The full RCV1 dataset contains over 800,000 newswire stories published by Reuters Ltd. between August 20, 1996 and August 19, 1997. Each article was manually and algorithmically categorized into hierarchical

topic codes, industry sectors, and regions, forming one of the most influential benchmark corpora for text classification, information retrieval, and distributional semantics.
Dataset composition. The Reuters 50/50 subset is a widely used balanced sampling of the RCV1 corpus designed for binary classification and stylistic comparison tasks. It consists of 50 selected topic categories drawn from the full taxonomy, with approximately equal representation of documents per category, resulting in a roughly uniform distribution across thematic domains such as politics, economics, international relations, science, and culture. This balance reduces topical bias and supports fair evaluation of linguistic or statistical features independent of subject matter.
Preprocessing and normalization. All Reuters documents were processed in a way to minimize patterns that might manifest as signal differences between human- and machine-generated text. Documents were normalized to UTF–8, spelling was corrected, and dates were standardized to a consistent format.
Generation. The LLM-generated corpus was produced using Meta's *Llama 3.1* model, a 70 billion-parameter transformer architecture deployed in quantized form for efficient inference. Through iterative prompt engineering, a standardized procedure was developed to generate machine-authored documents that mirror the stylistic and structural characteristics of human writing while avoiding confounding artifacts such as domain-specific jargon, formatting inconsistencies, or explicit references to named entities. The objective was to construct a corpus in which any measurable signal differences between human- and model-generated text could be attributed primarily to stylistic and probabilistic properties of the language model itself rather than to topical or contextual factors.

The generation methodology for each human-written document proceeded as follows:

1. The LLM was prompted to generate a document within the same general topical domain as the source text (e.g., business, finance, politics).
2. The model was instructed to approximate the *length*, *style*, and *organizational structure* of the corresponding human-written sample, including paragraph count and rhetorical tone.
3. Where applicable, the LLM was encouraged to reuse proper names, recurring entities, or unique subjects from the original to maintain thematic alignment while still producing novel phrasing.

This approach yielded a synthetic corpus that is topically aligned and stylistically comparable to the human-authored datasets (MICUSP and Reuters 50/50), enabling controlled comparative analysis of statistical and spectral features.

3.1.2. ***Source Corpus: The Michigan Corpus of Upper-Level Student Papers (MICUSP)***

A secondary dataset used in preliminary analysis for this study is the *Michigan Corpus of Upper-level Student Papers (MICUSP)*, a well-documented collection of academic writing produced by proficient undergraduate and graduate students at the University of Michigan [7], [9]. The corpus was developed to represent high-quality student writing across a wide range of academic disciplines, capturing authentic examples of advanced learner language in academic contexts.

Römer and O'Donnell [9] describe the design and compilation of MICUSP (Part 1), emphasizing the corpus's multi-disciplinary structure and genre taxonomy. The resulting collection comprises approximately 830 student papers totaling 2.6 million words, with metadata including discipline, genre, grade level, native language, and gender of the author.

3.2. **Additional Datasets.** In addition to the Reuters-based comparisons, we evaluate two further datasets that have become standard testbeds for distinguishing human- and machine-generated text: **HC3** and **MAGE**. These datasets differ substantially in authorship sources, domains, and LLM families, providing complementary perspectives on how the log-probability signal varies across models and content types.

HC3 (Human ChatGPT Comparison Corpus). The HC3 dataset [10] consists of paired human-written and ChatGPT-generated answers to the same questions across multiple domains, including medicine, finance, and general knowledge. Because queries are shared across modalities, differences arise purely from stylistic and distributional properties of the text rather than topical content. The dataset is widely used for evaluating LLM-generated text detection and has been updated to include GPT-3.5 and GPT-4 outputs. In our experiments, HC3 provides a perspective on older-generation models (GPT-3.5), revealing consistent separability via statistics-, Fourier-, and wavelet-based features.

MAGE (Model-Agnostic Generation Evaluation). The MAGE dataset [5] introduces a large-scale, heterogeneous collection of human- and machine-authored responses spanning multiple tasks, domains, and modern LLM families. Unlike HC3, MAGE includes outputs from newer-generation models—such as ChatGPT variants, PaLM, and GPT-4—which produce more human-like distributions and exhibit reduced stylistic artifacts. As a result, MAGE constitutes a significantly more challenging classification environment, with many traditional detection signals reduced or absent. In our results, MAGE shows lower separability across feature families, highlighting the increasing difficulty of detecting machine-generated text as models continue to improve.

3.3. **Token-Level Grading of Text Samples.** To enable a consistent comparison between human- and model-generated writing, we represent each text as a sequence of token-level probabilities as assigned by a reference large language model (LLM). We use the term *grading* to refer to the generation of token-level log probabilities for a given text. We accomplish this by passing a text sample through an LLM in a non-generative (evaluation-only) mode and recording the model's conditional log-probability for each observed token, given all preceding context. Formally, for a text sequence $x = (x_1, x_2, \ldots, x_T)$, we obtain

$$\ell_t = \log P_\theta(x_t \mid x_{<t}),$$

where P_θ denotes the probability distribution defined by the model with parameters θ. The resulting series $\{\ell_t\}_{t=1}^{T}$ represents the model's internal assessment of how *expected* or *surprising* each token is within its surrounding context.

Motivation and Terminology

We adopt the term *grading* to describe this process because it parallels how a human evaluator might assign a score to a written text, albeit at a much finer (token-level) granularity. The grading model effectively produces a per-token "confidence profile" that reflects its own linguistic expectations and calibration. When applied to human-written text, these scores reflect how well the human sequence aligns with the model's learned distribution of language; when applied to model-generated text, they measure the model's self-assessment or, in cross-model cases, one model's evaluation of another's output.

Procedure

For each document in our dataset, we performed the following steps:

(1) Tokenize the text using the same tokenizer as the evaluating model (e.g., LLaMA or Mixtral tokenizer).

(2) Feed the full token sequence to the model in evaluation mode (`generate=False`) to obtain conditional log-probabilities for all tokens.
(3) Record and store the sequence of token log-probabilities, normalized probabilities, and optional derived features such as per-token entropy and perplexity.

This grading process was performed for both human-authored and LLM-generated documents, enabling direct statistical comparison between groups. Importantly, the same evaluation model was used consistently within each experimental condition to ensure comparability of scores.

Experimental Groupings

Using this method, we generated three distinct graded datasets:

(1) **Human-generated, LLaMA-graded (HG–LG):** Human-written documents graded by LLaMA.
(2) **LLaMA-generated, LLaMA-graded (LG–LG):** Model-generated documents graded by LLaMA.
(3) **LLaMA-generated, Mixtral-graded (LG–MG):** Model-generated documents graded by Mixtral, to study cross-model calibration and evaluator bias.

Each group thus consists of the same fundamental signal type—a sequence of token-level log-probabilities—but derived under different combinations of generator and evaluator. Throughout the remainder of this report, we refer to these sequences as *graded signals*, and their statistical and spectral properties form the basis of the analyses that follow.

3.4. **Analysis of Distributions.** A central question in this study is whether statistical properties of token-level signals differ systematically between human- and model-generated text, and whether those differences remain detectable when using models other than the text-generating model. The preliminary analysis was to first examine the marginal and joint distributions of token log-probabilities and related features. These analyses serve two primary purposes: (1) to characterize surface-level statistical differences that may underlie or interact with deeper temporal patterns, and (2) to assess how such differences vary depending on whether the text was generated or graded by different large language models (LLMs).

Specifically, we compare three experimental groupings:

(1) **Human-generated, LLaMA-graded (HG–LG):** Human-written documents graded by the LLaMA model.
(2) **LLaMA-generated, LLaMA-graded (LG–LG):** Model-generated documents graded by the same model that produced them.
(3) **LLaMA-generated, Mixtral-graded (LG–MG):** Model-generated documents graded by a distinct LLM (Mixtral) to test grading-model dependence.

For each group, we extract per-token log-probabilities, normalized probabilities, and derived quantities such as entropy. Distributions are then analyzed both marginally and through pairwise comparisons. Empirical cumulative distribution functions (ECDFs), kernel density estimates (KDEs), and Zipf plots are used to visualize deviations between groups. Statistical divergence metrics—including Kullback-Leibler (KL) divergence, Jensen-Shannon (JS) divergence, and the Wasserstein distance—quantify how strongly each pair of distributions differs.

The goal of this analysis is not merely to detect separability, but to characterize *what form* these separations take. For example, a consistent rightward shift in the LLaMA-generated distributions would suggest that the model assigns higher confidence to its own outputs compared to human-written text, whereas discrepancies between LLaMA-graded

and Mixtral-graded distributions may indicate calibration or alignment differences between evaluators. Together, these distributional experiments provide a statistical foundation for subsequent analyses in the frequency domain, where we explore how such differences manifest as structured temporal or spectral patterns across token sequences.

3.5. **Analysis of Signals.** While distributional comparisons reveal static statistical differences between groups, they do not capture how those differences evolve or manifest over the sequence of tokens. To investigate temporal structure, we treat each graded text as a discrete one-dimensional signal composed of token-level log-probabilities, $\{\ell_t\}_{t=1}^{T}$. We then analyze these sequences using tools from signal processing, with the aim of identifying rhythmic or self-similar patterns characteristic of human versus model generation.

The methods described in this section include frequency-domain analysis using the Fourier transform, time-frequency analysis using wavelet transforms, and additional time-series techniques such as variance, autocorrelation, and entropy-based metrics.

3.5.1. *Fourier Transform Analysis*

Having introduced the theoretical basis of the Fourier transform earlier, here we focus on the specific spectral features derived from the token-level log-probability signals. Each sequence $\{\ell_t\}$ was treated as a discrete signal and transformed using the real-valued fast Fourier transform (FFT) to obtain its power spectrum, $P_k = |\hat{\ell}_k|^2$. The power spectrum captures how the total signal variance is distributed across frequency components, reflecting the degree of smoothness or volatility in the underlying confidence dynamics.

From each normalized power spectrum, we computed the following summary features:

- **High-frequency energy ratios** (`fft_high_energy_ratio_cut{c}`), with $c \in \{0.5, 0.75, 0.9\}$:

  The fraction of total spectral power in frequencies $\geq c$ times the Nyquist frequency. Formally,

  $$R_c = \frac{\sum_{k \geq \lfloor c \cdot K \rfloor} P_k}{\sum_{k=0}^{K-1} P_k},$$

  where K is the number of FFT bins. Larger ratios indicate greater concentration of energy at higher frequencies, corresponding to more rapid, localized fluctuations in model-assigned confidence. Lower ratios reflect smoother, more predictable token-probability sequences.
- **Spectral centroid:** The power-weighted mean frequency, defined as

  $$f_{\text{centroid}} = \frac{\sum_k k\, P_k}{\sum_k P_k}.$$

  This metric indicates where the "center of mass" of spectral energy lies. Human-written text typically exhibits lower centroids (energy concentrated at lower frequencies), whereas model-generated text may have higher centroids reflecting finer-scale oscillations in confidence.
- **Spectral slope:** The slope of a least-squares linear fit of $\log(P_k + \varepsilon)$ versus frequency index k, where ε prevents numerical instability for near-zero powers. This measure captures the overall decay rate of spectral energy with frequency. Steeper negative slopes indicate dominance of low frequencies (smooth, structured confidence evolution), while flatter or positive slopes suggest relatively more high-frequency noise or abrupt variability.

Together, these spectral metrics provide compact, interpretable summaries of signal smoothness and temporal coherence. They complement the distributional and autocorrelation analyses by quantifying the relative balance of slow versus fast variations in

token-level confidence, thereby revealing characteristic frequency signatures that may differentiate human from model text.

3.5.2. *Wavelet Transform Analysis*

While Fourier transforms reveal global frequency content, they are less suited to identifying how those frequencies vary locally in time. To capture both time and frequency information simultaneously, we apply continuous and discrete wavelet transforms (CWT and DWT). The wavelet transform of ℓ_t with respect to a mother wavelet ψ is defined as

$$W(a,b) = \frac{1}{\sqrt{a}} \sum_t \ell_t \, \psi^* \left(\frac{t-b}{a} \right),$$

where a and b represent scale and translation parameters, respectively.

We explored several wavelet families with different time-frequency trade-offs:

- **Daubechies (db2, db4, db8):** Provide compact support and varying smoothness. Lower-order Daubechies capture sharper transitions, while higher-order ones capture smoother, more global patterns in the signal.
- **Symlets (sym4):** Nearly symmetrical variants of Daubechies wavelets that reduce phase distortion, allowing more interpretable localization of transient features.
- **Coiflets (coif1):** Offer improved symmetry and higher vanishing moments for both the wavelet and scaling functions, making them well-suited for capturing slow-varying confidence trends.
- **Biorthogonal (bior3.3):** Enable separate decomposition and reconstruction filters, allowing perfect signal reconstruction with increased symmetry and linear phase—useful for comparing wavelet energies across documents of different lengths.

Wavelet coefficients were used to compute multi-scale energy spectra and entropy measures, which quantify how signal complexity varies across scales. By averaging over documents within each group, we obtained characteristic signatures describing the temporal organization of human- versus model-generated token probabilities.

3.5.3. *Statistical Signal Features*

In addition to spectral and wavelet analyses, we examined several time-series properties of the graded signals to capture different aspects of variability and temporal dependence:

- **Coefficient of Variation:** Amplitude fluctuations in log-probability over a rolling window; lower variance suggests smoother, more confident generation.
- **Autocorrelation:** Measures how strongly a token's log-probability depends on preceding tokens. We compute the autocorrelation function $r(\tau)$ for lags $\tau = 1, 2, \ldots$, as well as the integrated autocorrelation time to summarize persistence of confidence dynamics.
- **Token-Token Difference Variance:** We compute the variance of the difference between each adjacent pair of tokens, which captures the local smoothness of the probability signal.

These methods complement the spectral analyses by providing interpretable, scale-agnostic features describing signal stability, smoothness, and self-similarity.

3.6. **Classification and Validation.** Having extracted a diverse set of statistical and spectral features from the graded signals, we next evaluated their ability to discriminate between human- and model-generated text. This section describes the procedures used to (1) identify effective thresholds for individual features, (2) combine multiple features

using a random forest classifier, and (3) validate classification performance through both cross-validation and independent train-test splits.

3.6.1. *Feature-Level Threshold Analysis*

For each scalar feature (e.g., spectral centroid, high-frequency energy ratio, autocorrelation time, entropy), we examined its separability between groups by fitting a simple one-dimensional threshold classifier. Thresholds were determined by maximizing accuracy on the training set or, equivalently, by selecting the cut-point that minimized total classification error across the two classes. Receiver Operating Characteristic (ROC) curves and Area Under the ROC Curve (AUROC) values were computed to quantify discriminative strength for each feature independently. The F1 harmonic score was used as a threshold-independent metric for accuracy, factoring in precision and recall. This step provided interpretable baselines and helped identify which features carried the most individual predictive signal prior to multivariate modeling.

3.6.2. *Random Forest Classification*

To jointly model multiple, potentially interacting features, we employed a random forest classifier (RFC) implemented with `scikit-learn`. Random forests are ensemble models composed of multiple decision trees trained on bootstrapped subsets of the data, each considering a random subset of available features at each split. The final prediction is obtained by majority vote over all trees. Feature vectors were standardized prior to training, and hyperparameters such as the number of estimators, tree depth, and minimum samples per leaf were tuned through grid search to balance bias and variance. The model outputs both class predictions and feature importance scores, allowing interpretation of which features most strongly contributed to discrimination.

3.6.3. *Cross-Validation and Train/Test Evaluation*

Model performance was assessed using both k-fold cross-validation and independent train/test splits to ensure robustness.

- **Cross-validation:** We used stratified k-fold cross-validation (typically $k = 5$) to estimate variability in model performance across folds. For each fold, metrics including accuracy, precision, recall, F_1, and AUROC were computed and averaged.
- **Train/test splits:** In addition to cross-validation, a held-out test set (commonly 20% of the data) was used for final model evaluation. The model was trained on the remaining 80% and tested once on unseen samples to approximate out-of-sample generalization performance.

This dual evaluation strategy ensured that both threshold-based and multivariate classifiers were validated against overfitting and provided reliable estimates of discriminative performance.

3.6.4. *Interpretation*

Threshold analysis allowed us to identify features with strong individual separation power, while the random forest classifier leveraged complementary information across heterogeneous features. Together, these approaches quantify the extent to which statistical and spectral properties of the graded signals encode detectable differences between human- and model-generated text, forming the methodological foundation for the performance results presented in the following chapter.

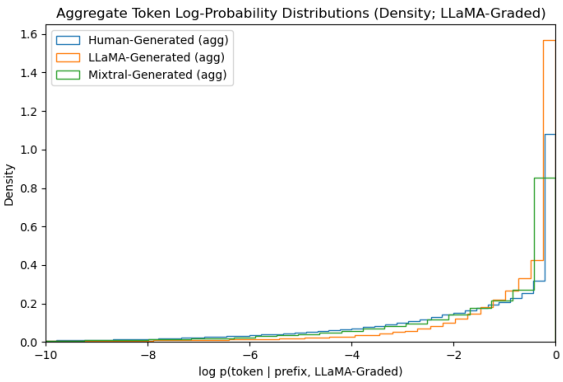
4. Results

4.1. **Statistical Analysis.** Statistical analysis uncovered distribution differences between the different populations of document token log probabilities. In particular, the ECDF and aggregate token log-probability distribution between human-authored and AI-generated Reuters documents showed some interesting differences.

4.1.1. *Empirical Cumulative Distribution Function*

Aggregate Token Log-Probability Distributions

This plot shows the empirical distribution of token log-probabilities (ECDF-style density) aggregated over 1,000 documents from each modality: Human-Generated, Llama-Graded (HGLG); Llama-Generated, Llama-Graded (LGLG); and Mixtral-Generated, Llama-Graded (MGLG). A clear pattern emerges: LGLG tokens cluster far more tightly near zero (high probability), indicating that LLaMA assigns consistently confident log-probabilities to its own generated text. In contrast, HGLG and MGLG exhibit broader distributions that extend further into lower-probability regions. MGLG displays a slightly sharper rise near 0 than HGLG, suggesting marginally more confident token predictions than in human-written text.



4.1.2. *Distribution Distance Measures*

Distribution distance measures indicate that the difference between log probabilities from human- and machine-generated text is an order of magnitude greater than the difference between machine-generated text that is "graded" by the same model vs. a different family of model:

Table 1. Divergence and distance metrics between aggregated distributions. 1,000 pairs of Human-Generated, Llama-Graded (HGLG), Llama-Generated, Llama-Graded (LGLG), and Mixtral-Generated, Llama-Graded documents. The difference between human and model-generated distributions are consistently greater than differences between two model-generated distributions.

Pair	**KL**$(a \mid\mid b)$	**KL**$(b \mid\mid a)$	**JSD**	$\sqrt{\textbf{JSD}}$	**KS**	**Wasserstein-1**
HGLG (agg) vs LGLG (agg)	0.151 732	0.131 382	0.034 476	0.185 676	0.204 903	0.849 863
HGLG (agg) vs MGLG (agg)	0.320 362	0.178 440	0.050 375	0.224 444	0.067 845	0.346 447
LGLG (agg) vs MGLG (agg)	0.058 363	0.030 558	0.008 051	0.089 725	0.137 712	0.583 223

4.1.3. *Token Log Probability Moving Averages*

Finally, one of the most compelling pieces of visual evidence from analyzing the log-probability signal characteristics using a moving average chart. You can see that the high frequency patterns for an individual document are far more greatly pronounced

in human-generated text than machine-generated text. There is additionally a small difference in the high frequency patterns between the LLM-generated text as graded by the same model (Llama) vs. another model (Mixtral). This is consistent with the aggregate token log-probability distribution chart.

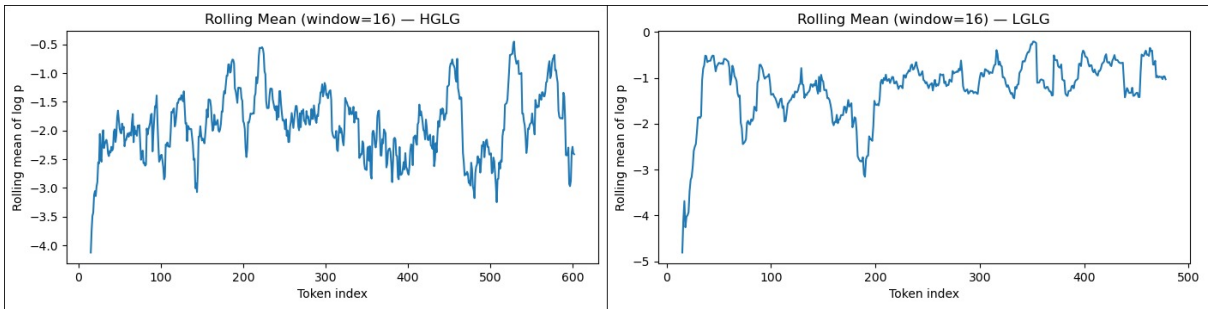


FIGURE 1. Moving average (window size 16) of token log-probabilities for a single example document from each modality: Human-Generated, Llama-Graded (HGLG); Llama-Generated, Llama-Graded (LGLG). The human-generated sequence shows much more pronounced peaks and troughs in the moving average, suggesting sharper changes in model confidence across the document, whereas both LGLG and MGLG (omitted from figure) exhibit smoother, more plateau-like behavior. No strong conclusions can be drawn from a single data point, but the contrast is nevertheless suggestive.

4.2. **Signal Analysis Results.** Classification based on signal-based features had a couple of clear patterns: 1. Older models were far more separable, meaning there is stronger difference between token probability patterns created by LLMs and humans, compared with more subtle differences for state of the art LLMs. 2. Token log probabilities produced by the same model that generated the text appears to be biased in favor of higher probabilities, and is therefore more clearly separable from the human text from a classification standpoint.

We use the macro F1-harmonic score because it balances precision and recall while weighting all classes equally, preventing performance on the dominant class from overwhelming the evaluation. The harmonic formulation further penalizes extreme imbalances between precision and recall, making it a sensitive and robust metric for detecting subtle differences in classification performance across modalities.

$$\text{Macro-F1} = \frac{1}{K}\sum_{k=1}^{K}\frac{2\,\text{Precision}_k\,\text{Recall}_k}{\text{Precision}_k + \text{Recall}_k},$$
$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k},$$
$$\text{Recall}_k = \frac{TP_k}{TP_k + FN_k}.$$

4.2.1. ***Statistical Feature Results***

Statistical features achieved consistently strong performance across all datasets, with macro F1 scores typically ranging from 0.80 to nearly 0.99. These features were especially effective in cases where the model needed to distinguish between human-generated and LLM-generated text (e.g., HC3 and Reuters 50/50), suggesting that simple descriptive

measures of the log-probability signal—such as coefficient of variation, local means, and variation of differences—capture robust and discriminative structure across modalities.

TABLE 2. Macro F1-harmonic classification performance using statistics-based features only.

Dataset / Classification Pair	**Macro F1-harmonic**
rcv1-llama-mixtral	0.800
rcv1-mixtral-llama	0.870
rcv1-llama-llama	0.970
hc3-llama	0.984
hc3-mixtral	0.989
mage-llama	0.876
mage-mixtral	0.806

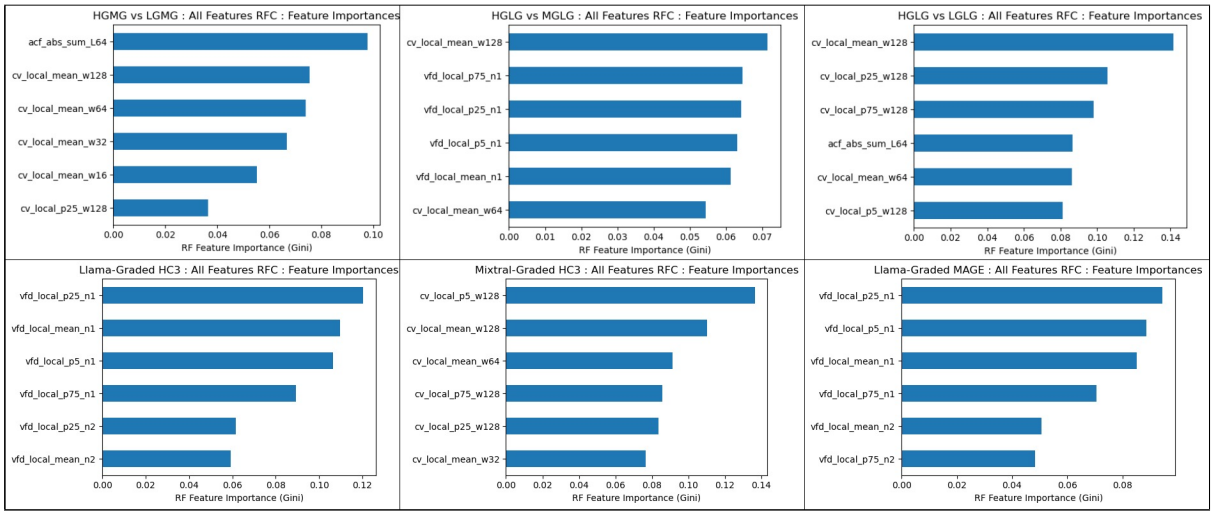


FIGURE 2. Random forest classifier feature importances for all statistics-based features across the six classification settings. Although the exact ranking varies by dataset and pair, coefficient-of-variation (cv_local_*) features are typically the most important, with variation-of-first-differences (vfd_local_*) and autocorrelation (acf_*) features also contributing. Here, `w#` denotes the window size in tokens, `n#` indicates whether the feature is computed on first (`n1`) or second (`n2`) differences, and `p#` is the percentile (e.g., p25, p75) used in the statistic.

4.2.2. *Fourier Feature Results*

Fourier-based features showed weak to moderate performance overall, with F1 scores generally between 0.60 and 0.88. While they sometimes provided meaningful signal, they were typically outperformed by both statistics-based and wavelet-based features. This pattern suggests that global frequency characteristics of the log-probability sequence do contain differences between modalities, but these differences may be too coarse or diffuse compared to the more localized features extracted by the other approaches.

TABLE 3. Macro F1-harmonic classification performance using Fourier-based features only.

Dataset / Classification Pair	Macro F1-harmonic
rcv1-llama-mixtral	0.675
rcv1-mixtral-llama	0.692
rcv1-llama-llama	0.878
hc3-llama	0.845
hc3-mixtral	0.974
mage-llama	0.602
mage-mixtral	0.648

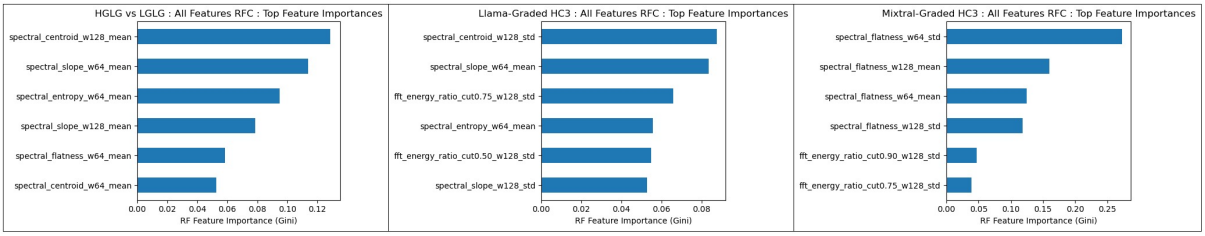


FIGURE 3. Random forest feature importances for Fourier-based features on the three settings where they provide reasonably strong classification performance: HGLG vs. LGLG (left), Llama-graded HC3 (center), and Mixtral-graded HC3 (right). Across these cases, spectral centroid and spectral slope features (especially with window sizes 64 and 128) are consistently important, with spectral entropy and high-frequency energy ratios (fft_energy_ratio_cut*) also contributing. For Mixtral-graded HC3, spectral flatness dominates, indicating that differences in how "peaky" or "noise-like" the spectra are play a key role in separating modalities.

4.2.3. *Wavelet Feature Results*

Wavelet-based features performed very strongly, in several cases matching or surpassing the statistics-based features, with F1 scores reaching as high as 0.99. Their strength was especially notable in fine-grained comparisons (e.g., rcv1-mixtral-llama), where multiscale structure in the log-probability signal appears highly discriminatory. This indicates that wavelets' ability to capture localized, scale-dependent fluctuations offers a rich representation that complements and sometimes exceeds the discriminative power of purely statistical or Fourier-based approaches.

TABLE 4. Macro F1-harmonic classification performance using wavelet-based features only.

Dataset / Classification Pair	Macro F1-harmonic
rcv1-llama-mixtral	0.729
rcv1-mixtral-llama	0.993
rcv1-llama-llama	0.965
hc3-llama	0.958
hc3-mixtral	0.978
mage-llama	0.677
mage-mixtral	0.703

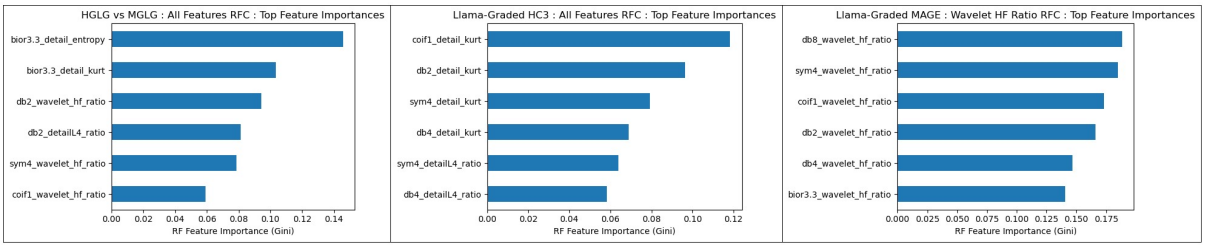


FIGURE 4. Random forest feature importances for wavelet-based features across selected classification settings. For the Reuters-style datasets (HGLG vs. MGLG, using Llama 3.1 and Mixtral) and for HC3 (GPT-3.5), wavelet detail *kurtosis* and *entropy* features consistently emerge as key discriminators, suggesting that higher-order statistics of the wavelet coefficients capture modality-specific structure in the log-probability signal. In contrast, for the MAGE dataset involving newer models, high-frequency wavelet energy ratios dominate while kurtosis and entropy play a lesser role, and the overall separability is noticeably weaker, indicating that these newer systems produce log-probability patterns that are harder to distinguish using the same wavelet feature family.

Wavelet detail coefficient kurtosis measures how often the log-probability sequence contains rare but sharp local deviations at specific scales. High kurtosis indicates a signal dominated by small coefficients with occasional large spikes, while low kurtosis reflects smoother, more uniform structure. Its usefulness as a feature suggests that the distinguishability between human and model text partly arises from differences in how abruptly the token probabilities change over short spans.

Wavelet detail coefficient entropy, by contrast, summarizes the overall disorder or unpredictability of the wavelet coefficients, indicating how evenly energy is spread across scales and positions. The strong performance as classification features suggests that human- and model-generated texts differ not just in average magnitude or frequency content, but in the higher-order structure and local complexity of their log-probability fluctuations.

5. Conclusion and Future Work

In this study, we compared aggregated distributions of LLM-derived token log-probabilities for human- and model-generated text using a range of techniques that included statistical measures and signal processing tools. We looked at 3 datasets in order to evaluate robustness as well as to identify some of the differences between sequences created by a variety of different LLMs. The results highlight measurable but nuanced differences between the underlying probability structures of the evaluated sources, suggesting that signal-based signatures of large language models can often be detected even when aggregated across samples. Local patterns tended to be far more pronounced than global patterns, with simple statistical measures, with the most predictive features being coefficient of variation and variation of first differences. Wavelet-based features also tended to fare well with separating human- from llm-generated work, especially when looking at higher frequency characteristics.

The findings from this research also appear to suggest that differences between human- and llm-generated text are far harder to discern when evaluating text written by many models, such as those utilized in the MAGE datset. It is possible also that this could be

a result of this technique having difficulty with newer models, but more research would have to be done with newer models in order to make that determination.

There are two main areas of future work we'd like to explore. The first would be to evaluate additional feature representations, adding more focus to lower frequency features and trying to produce more robust features than just the high frequency space that we probed in this work. We additionally want to expand on the findings from this paper by incorporating newer models and trying to learn characteristics and patterns that are specific to particular LLMs or LLM families. The ideal goal would be a multi-class classification capability that identifies the particular model (or model family) that a body of text was generated by.

References

[1] P. S. Addison, *The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance*, 2nd. CRC Press, 2017.

[2] T. B. Brown et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020. [Online]. Available: https://arxiv.org/pdf/2005.14165.

[3] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex fourier series," *Mathematics of Computation*, vol. 19, no. 90, pp. 297–301, 1965. [Online]. Available: https://www.cs.jhu.edu/~misha/ReadingSeminar/Papers/Cooley65.pdf.

[4] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," *Journal of Machine Learning Research*, vol. 5, pp. 361–397, 2004. [Online]. Available: https://jmlr.csail.mit.edu/papers/v5/lewis04a.html.

[5] Y. Li et al., "Mage: Machine-generated text detection," *arXiv preprint arXiv:2305.13242*, 2023. [Online]. Available: https://arxiv.org/pdf/2305.13242.

[6] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989. [Online]. Available: https://www.di.ens.fr/~mallat/papiers/MallatTheory89.pdf.

[7] M. B. O'Donnell and U. Römer, "From student hard drive to web corpus (part 2): The annotation and online distribution of the michigan corpus of upper-level student papers (micusp)," *Corpora*, vol. 7, no. 1, pp. 1–18, 2012. [Online]. Available: http://uteroemer.weebly.com/uploads/5/5/7/7/5577406/odonnell_and_roemer_corpora_article_2012.pdf.

[8] A. V. Oppenheim, *Signals and systems*, MIT OpenCourseWare, Massachusetts Institute of Technology, Available at https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-003-signals-and-systems-fall-2011/, 2011.

[9] U. Römer and M. B. O'Donnell, "From student hard drive to web corpus (part 1): The design, compilation and genre classification of the michigan corpus of upper-level student papers (micusp)," *Corpora*, vol. 6, no. 2, pp. 159–177, 2011. [Online]. Available: http://uteroemer.weebly.com/uploads/5/5/7/7/5577406/roemer_and_odonnell_corpora_article_2011.pdf.

[10] Z. Su, X. Wu, W. Zhou, G. Ma, and S. Hu, "Hc3 plus: A semantic-invariant human chatgpt comparison corpus," *arXiv preprint arXiv:2309.02731*, 2023. [Online]. Available: https://arxiv.org/pdf/2309.02731v4.

[11] Z.-X. Wu, C. Liu, Y. Wang, Y. Zhang, and X. Zeng, "Detectrl: Benchmarking llm-generated text detection in real-world scenarios," *arXiv preprint arXiv:2410.23746*, 2024. [Online]. Available: https://arxiv.org/abs/2410.23746.

[12] Z.-X. Wu, T. Wang, Y. Wang, Y. Huang, and X. Zeng, "A survey on llm-generated text detection: Necessity, methods, and future directions," *Computational Linguistics*, vol. 51, no. 1, pp. 275–316, 2023. DOI: 10.1162/coli_a_00483. [Online]. Available: https://direct.mit.edu/coli/article/51/1/275/127462/A-Survey-on-LLM-Generated-Text-Detection-Necessity.

[13] J. Zhang, Z. Yang, F. Li, Y. Wang, and X. Zhao, "Zero-shot detection of llm-generated text using token probability distributions," in *Proceedings of the 2024 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), 2024, pp. 15 362–15 378. [Online]. Available: https://aclanthology.org/2024.emnlp-main.971.

Appendix A. Additional Tables and Figures

Appendix B. Implementation Details