Fingerprinting Large Language Models with Signal Processing

Daniel McGonigle
Advisor: Dr. Joris Roos

University of Massachusetts Lowell, Mathematics Department
Lowell, MA

November 5, 2025

Abstract

This project investigates whether large language models (LLMs) exhibit measurable statistical and spectral differences from human-generated text. Using token-level log-probability sequences for both human- and machine-generated text, we analyze distributional and signal-based features to identify potential "fingerprints" of model generation. Power-spectral densities, wavelet decompositions, and entropy measures were computed for human- and machine-authored corpora, revealing some measurable deviations in frequency dynamics between human and machine-generated sources. A classification model trained on these spectral representations achieved limited success in discriminating between human and model text, particularly when the same model used to generate the text provided the log-probability signal. These findings suggest that there may be merit in frequency-domain analysis as a means to detect text generated by advanced LLMs.

Acknowledgments

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

Contents

Chapter 1

Introduction

Since the release of GPT-3 [3], 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.

Chapter 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 [13]. Some approaches targeting specific data domains have showed limited success, as in DetectRL [12]. Zhang et al. [14] 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 [2]. For a continuous signal $x(t)$, the Fourier Transform is defined as

$$X(f) = \int_{-\infty}^{\infty} x(t)\, e^{-j2\pi ft}\, dt, \tag{2.1}$$

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:

$$X_k = \sum_{n=0}^{N-1} x_n\, e^{-j2\pi kn/N}, \quad k = 0, 1, \ldots, N-1. \tag{2.2}$$

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 [4], 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 [5], [8].

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

$$W(a, b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t)\, \psi^*\left(\frac{t-b}{a}\right)\, dt, \tag{2.3}$$

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 and generative model outputs [1], [6]. Recent work has also explored wavelet-based representations in deep learning contexts, integrating spectral decomposition into neural architectures [7], [11].

Chapter 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.

3.1.1 Source Corpus: The Michigan Corpus of Upper-Level Student Papers (MICUSP)

A key dataset used in 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 [9], [10]. 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. The compilation of MICUSP followed a carefully designed methodology involving institutional collaboration, participant consent, and detailed genre classification.

Römer and O'Donnell [10] describe the design and compilation of MICUSP (Part 1), emphasizing the corpus's multi-disciplinary structure and genre taxonomy. Each text was assigned to one of 16 academic disciplines (e.g., Linguistics, Psychology, Mechanical Engineering) and classified into one of four genre families—*argumentative*, *analytical*,

report, or *narrative*—based on communicative purpose and rhetorical structure. 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.

O'Donnell and Römer [9] detail the annotation and online distribution of MICUSP (Part 2), which introduced a standardized XML format and consistent metadata schema. Annotation layers include paragraph and sentence boundaries, structural divisions (e.g., introduction, discussion, conclusion), and genre-level metadata. This structured representation facilitates computational approaches to text analysis, enabling tokenization, linguistic feature extraction, and model-based inference using reproducible workflows.

3.1.2 Preprocessing and Format Standardization

To ensure compatibility with machine learning and signal-processing pipelines, each document was normalized to a consistent UTF-8 text format. Headers, page numbers, and non-linguistic artifacts were removed. Tokenization, sentence segmentation, and part-of-speech tagging were performed using standard NLP preprocessing tools. For quantitative analysis, token-level sequences were encoded into numerical arrays representing log-probabilities, entropy, or spectral characteristics (e.g., frequency-domain features derived via Fourier or wavelet transforms). All processing steps were implemented in reproducible scripts to maintain consistency across datasets.

3.1.3 Ethical and Licensing Considerations

MICUSP is distributed under a research license for non-commercial academic use. Only publicly available materials or those distributed with appropriate institutional permission were included. No personally identifiable information was retained, and all data handling followed established corpus ethics and reproducibility standards.

3.2 Analysis of Distributions

3.3 Analysis of Signals

Chapter 4

Results

Chapter 5

Discussion

Chapter 6

Conclusion and Future Work

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] R. N. Bracewell, *The Fourier Transform and Its Applications*, 3rd. McGraw-Hill, 2000.

[3] T. B. Brown et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[4] 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.

[5] I. Daubechies, *Ten Lectures on Wavelets*. SIAM, 1992.

[6] P. Flandrin, *Time–Frequency/Time–Scale Analysis*. Academic Press, 1999.

[7] X. Liu, Y. Zhang, and Q. Li, "Wavelet neural networks for nonstationary time series forecasting," *Neural Networks*, vol. 165, pp. 158–171, 2023.

[8] 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.

[9] 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.

[10] 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.

[11] C. K. I. Williams, M. Seeger, and N. D. Lawrence, "Using wavelets in deep generative models," *Journal of Machine Learning Research*, vol. 21, no. 225, pp. 1–32, 2020.

[12] 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, Accessed November 2025. [Online]. Available: https://arxiv.org/abs/2410.23746.

[13] 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.

[14] 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