Fingerprinting Large Language Models with Signal Processing

Daniel McGonigle
Advisor: Dr. Joris Roos

University of Massachusetts Lowell, Mathematics Department
Lowell, MA

November 4, 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 [1], 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

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.

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.

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 [3]. Some approaches targeting specific data domains have showed limited success, as in DetectRL [2]. Zhang et al. [4] 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.

Chapter 3

Methods

3.1 Dataset Curation

3.2 Analysis of Distributions

3.3 Analysis of Signals

Chapter 4

Results

Chapter 5

Discussion

Chapter 6

Conclusion and Future Work

References

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

[2] 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.

[3] 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.

[4] 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