Zipfian Patterns in LLM-Generated Text: A Quantitative Analysis

Elisa Negrini - https://github.com/elisa-negrini/CognitiveDataScience-project.git

May 30, 2025

Abstract :

Zipf's law describes a power-law relationship between word frequency and rank, widely regarded as a statistical hallmark of natural language. This study investigates whether this scaling behavior persists in texts generated by a large language model (LLM) under diverse prompts and temperature settings. For each corpus, the Zipf exponent (α) is estimated after trimming low-frequency noise. Results reveal that higher-temperature outputs produce steeper Zipfian slopes, while lower temperatures yield more uniform distributions—yet all preserve the power-law structure. Corpus characteristics like vocabulary size and trimming threshold ($x_{\min}$) significantly impact α, suggesting that Zipfian structure is a robust emergent property of LLM-generated text.

1 Introduction

A persistent pattern in language is described by Zipf's law, which states that word frequency is inversely proportional to its rank. This power-law behavior extends beyond natural language into domains like user-generated passwords. Wang et al. (2017) showed that passwords follow Zipfian distributions, revealing deep regularities even in seemingly arbitrary and spontaneous human choices.

Inspired by this, the aim of this study is to investigate whether similar patterns emerge in texts generated by large language models (LLMs). Using corpora generated across different prompts and temperature settings, we analyze word frequency distributions to assess the presence and robustness of Zipfian structure in synthetic language.

2 Theoretical background

Zipf's law describes a consistent pattern in language where the frequency of a word is inversely proportional to its rank in the frequency list. This means that the most frequent word in a corpus typically occurs about twice as often as the second most frequent word, three times as often as the third, and so on.

More formally, if $f(r)$ is the frequency of the word with rank r, Zipf's law can be expressed as:

$$f(r) \propto \frac{1}{r^{\alpha}}$$

where α is the exponent of the distribution, typically close to 1 in natural language (Diamond J., 2023).

2.1 Cognitive and statistical explanations

Across natural languages and domains, word frequencies follow a robust power-law distribution, revealing deep statistical regularities in language. Both cognitive and statistical explanations have been proposed. Some argue that even random text (e.g., "monkey typing") can exhibit Zipf-like distributions, suggesting such patterns may arise without cognition. This idea dates back to Miller (1957), who showed that a monkey hitting typewriter keys at random—under minimal probabilistic constraints—can generate a rank-frequency distribution resembling Zipf's law. However, this view has been strongly challenged by Ferrer-i-Cancho and Elvevåg (2010), who demonstrated through rigorous statistical tests that random texts systematically fail to reproduce the statistical structure of real language. Their findings demonstrate that Zipfian distributions in natural language are a reflection of underlying cognitive, communicative, and structural constraints rather than being solely attributable to randomness.

Interestingly, Zipf's law has also been observed in non-human communication systems, such as dolphins and gorillas, suggesting that it may be a universal feature of complex systems.

Cognitive accounts of Zipf's law often invoke the Principle of Least Effort: speakers tend to prefer short, frequent words to minimize production effort, while listeners benefit from less ambiguous, though rarer, words. This trade-off leads to an efficient distribution of word use. Zipfian patterns are thought to emerge through language evolution to support communicative efficiency. Frequent, shorter words are easier to learn and remember, aiding comprehension and speech segmentation. These patterns align with

broader cognitive scaling laws, suggesting the brain may operate near critical states that balance stability and adaptability (Linders et al., 2020).

In recent work, Linders and collegues examined whether Zipf's law also applies to human-machine dialogues. They found that both human and agent-generated speech generally follow Zipfian distributions across features such as word frequency, word position, and turn length. However, agent speech tends to be more repetitive and less variable—particularly in task-based interactions. Notably, differences in turn length distributions emerged as a potential cue for distinguishing machine-generated dialogue from human speech.

3 Methods

3.1 Data Generation and Prompt Design

Textual data were collected from a large language model (LLM) through programmatic querying using the Mistral API. The goal was to generate diverse textual datasets for testing the applicability of Zipf's law under different generative conditions. The LLM was prompted to generate 500 individual texts per configuration, which were then merged into a single corpus for analysis.

Three distinct topics were selected to ensure lexical diversity and variation in linguistic structure: (1) horror stories, (2) scientific papers, and (3) surreal narratives. The third category was specifically designed to deviate from natural language distributions, with prompts encouraging the model to avoid adhering to Zipfian patterns.

For each topic, two temperature settings were used during generation: a low temperature (e.g., 0.2) to promote deterministic, repetitive output, and a high temperature (e.g., 0.9) to increase randomness and lexical variability. This setup was intended to explore how generation temperature affects word frequency distributions and the emergence or breakdown of Zipfian behavior.

Text generation was fully automated through Python scripts issuing structured prompts via POST requests to the Mistral API.

3.2 Data Analysis

For each text, the corpus was tokenized, and the rank and frequency of all word types were computed and stored in a dataframe. An example of the resulting frequency table is shown at page 4, which displays top-ranked words as well as a sample of entries from the long tail, where most tokens are *hapax legomena*—words that occur only once. All the texts are roughly 40 to 50 tokens long.

Before fitting Zipf's law to the rank-frequency data, the tail of the distribution was excluded. This decision is motivated by the fact that the tail contains a large number of low-frequency words, particularly *hapax legomena*, which introduce noise and distort the log-log plot. These infrequent tokens tend to flatten the slope of the curve, inflate the estimated Zipf exponent (α), and reduce the overall goodness of fit (R^2). Empirical studies have shown that Zipf's law holds most reliably in the head of the distribution, where the power-law behavior is more consistent (A Clauset et al., 2009). The lower bound for fitting, denoted as $x_{\min}$, was automatically determined using the powerlaw package, which selects the value that minimizes the Kolmogorov–Smirnov (KS) distance between the empirical distribution and the fitted power-law model. This trimming ensures a more accurate and meaningful estimation of the Zipfian relationship by excluding the noisy long tail.

Zipf's exponent α (the negative of the fitted slope) and the coefficient of determination R^2 were computed for each corpus using linear regression in log-log space. These metrics quantify how well the data follow Zipf's law: α reflects the steepness of the rank-frequency curve, while R^2 indicates the goodness of fit. The results are visualized in Figure 1, where each subplot displays the fitted curve, the theoretical Zipf line ($\alpha = 1$), and the top words annotated on the empirical distribution. Additional metadata, such as the number of unique words before and after trimming, the corresponding $x_{\min}$ and the length of the original text, are also reported in each subplot.

To assess the stability of our Zipf exponent estimates, a bootstrap resampling procedure was applied. For each corpus, we repeatedly resampled the rank-frequency dataset with replacement and computed Zipf's α for each resample. This allowed to estimate both the mean α and its associated 95% confidence interval, providing a robust measure of the uncertainty surrounding each exponent.

Interesting was also to try to understand whether temperature affects the Zipf exponent α. For each topic the two-sample Welch's t-test was performed on the bootstrap distributions of α between the high- and low-temperature versions.

Finally, to investigate the effect of temperature on Zipfian structure, two-sample Welch's t-tests on the bootstrap distributions of α is performed between the high- and low-temperature versions for each topic. This statistical comparison enabled to test whether the observed differences in α across temperature settings were significant.

Lastly pearson correlation coefficient is computed between the values of alpha, $x_{\min}$, the number of unique words before and after trimming.

4 Results

The estimated Zipf exponents, computed after trimming the low-frequency tail and applying bootstrap resampling, reveal some differences across both topics and temperature settings. In the Horror topic, the high-temperature text yielded a higher exponent (mean $\alpha_{\text{high\_t}} = 1.08$, 95% CI [1.06, 1.11]) than its low-temperature counterpart ($\alpha_{\text{low\_t}} = 1.01$, CI [0.99, 1.04]). A similar pattern is observed for Scientific texts ($\alpha_{\text{high\_t}} = 1.01$, CI [1.00, 1.03] vs. $\alpha_{\text{low\_t}} = 0.95$, CI [0.93, 0.96]) and Surreal Narrative ($\alpha_{\text{high\_t}} = 1.18$, CI [1.14, 1.23] vs. $\alpha_{\text{low\_t}} = 1.06$, CI [1.04, 1.07]).

These results, visualized in Figure 1, show that high-temperature texts tend to produce steeper Zipfian distributions, reflecting greater dominance of top-ranked tokens. At the same time, low-temperature outputs exhibit flatter distributions and reduced α values, suggesting more uniform frequency use. Importantly, for each topic, a two-sample Welch's t-test was performed on the bootstrap distributions of α, and in all cases, the differences between high- and low-temperature texts were statistically significant at the 5% level. This highlights how temperature systematically affects word frequency distribution.

It was also observed how temperature settings influence other corpus characteristics. In particular, higher temperatures -and the associated increase in α levels- are related to a larger number of unique words ($r = 0.837$, $p = 0.0376$) and to lower values of $x_{\min}$ ($r = -0.893$, $p = 0.0164$), reflecting greater lexical diversity. This, in turn, affects the number of tokens retained after trimming, which also showed a strong positive correlation with α ($r = 0.898$, $p = 0.015$). These findings suggest that preprocessing thresholds and corpus richness should be taken into account when interpreting differences in Zipfian scaling across conditions.

Finally, it is worth noting that even when explicitly prompted to generate a non-Zipfian text—as in the surreal narrative condition—the model still produced outputs that exhibited clear Zipfian structure. This suggests that Zipf's law is not easily disrupted by prompt-level interventions, and may instead emerge as an inherent property of the language modeling objective itself.

5 Conclusion

This study confirms that Zipf's law persists across LLM-generated texts, regardless of topic or sampling temperature. High-temperature outputs tend to produce steeper Zipfian slopes, while low-temperature texts exhibit more uniform word usage. Nevertheless, all conditions retain the characteristic power-law structure of word frequencies. We found that both the number of unique words and the trimming extent ($x_{\min}$) systematically affect Zipf exponent estimates: larger vocabularies yield higher α, while more aggressive trimming lowers it. This highlights the importance of corpus richness and preprocessing choices in analyzing Zipfian patterns in LLM-generated text.

Our findings align with recent work by Diamond (2023), who showed that even fully artificial languages invented by ChatGPT exhibit Zipfian structure comparable to natural English. Taken together, these results suggest that Zipf's law is not solely a feature of natural communication, but also an emergent property of language models trained on human data. Despite variations in prompt type or sampling parameters, Zipfian regularity appears to be deeply embedded in the generative behavior of large-scale language models.

References

Linders, G. M., & Louwerse, M. M. (2020, October). Zipf's Law in Human-Machine Dialog. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (pp. 1–8).

Diamond, J. (2023). "Genlangs" and Zipf's Law: Do languages generated by ChatGPT statistically look human?. *arXiv preprint arXiv:2304.12191.*

Wang, D., Cheng, H., Wang, P., Huang, X., & Jian, G. (2017). Zipf's law in passwords. *IEEE Transactions on Information Forensics and Security, 12*(11), 2776–2791.

Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology, 70*(2), 311–314.

Ferrer-i-Cancho, R., & Elvevåg, B. (2010). Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS One, 5*(3), e9411.

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review, 51*(4), 661-703.

Appendix

nr.	rank	word	frequency	rel_freq
0	1	the	38666	0.079808
1	2	and	15638	0.032278
2	3	a	13178	0.027200
3	4	of	12249	0.025283
...	...	...	...	...
7987	462	lain	1	0.000002
7988	462	colliding	1	0.000002
7989	462	conclusion	1	0.000002

Table 1
Frequencies of words in the `horror_hightemp` corpus.

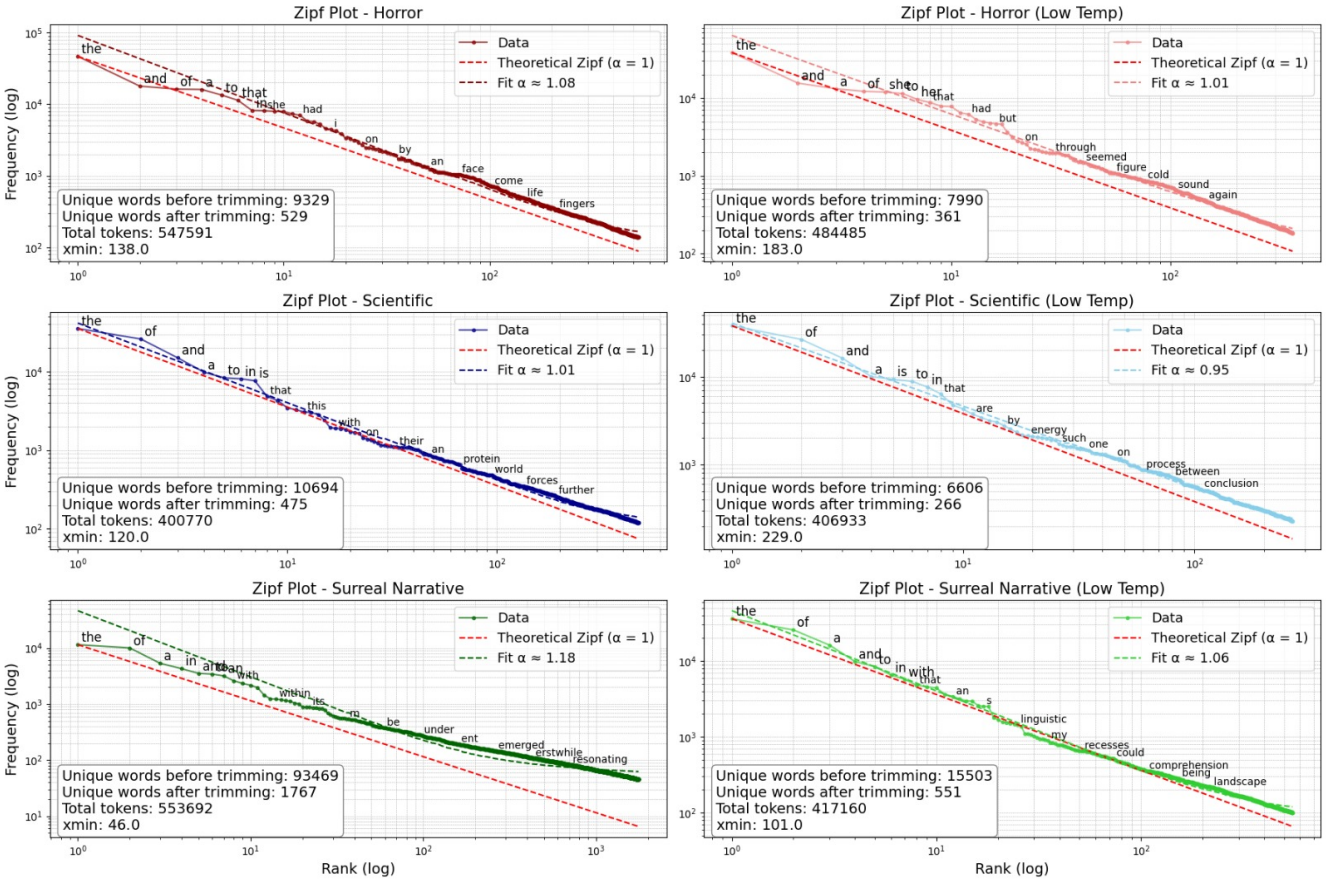

Figure 1. Zipf plots for all six corpora (3 topics × 2 temperatures). Frequencies vs. ranks (log-log) with Zipf reference $\alpha = 1$ and empirical fits.