Repeated Sequences Reveal Gaps between Large Language Models and Natural Language

Kumiko Tanaka-Ishii
Waseda University, Japan.
kumiko@waseda.jp

Abstract

Evaluating whether large language models (LLMs) capture the structure of natural language beyond local fluency remains an open challenge. Existing evaluation methods, largely based on task performance or short-context behavior, provide limited insight into the long-range statistical organization of generated text.

We propose a complementary evaluation framework based on repeated subsequences. By analyzing their distribution across scales and relating it to higher-order Rényi entropies, we probe how texts reuse previously established structure under finite-length conditions. Experiments on human-written texts and length-matched GPT-generated texts show that, while power-law models can describe restricted ranges of block length, the observed entropy growth is often equally or better characterized by logarithmic–power forms.

Across datasets, natural language exhibits stable entropy-growth patterns over accessible ranges, with consistent average behavior despite variability across individual texts. In contrast, GPT-generated texts show systematic and statistically significant shifts in estimated exponents with model size. These results demonstrate that repeated-subsequence entropy provides a quantitative structural diagnostic that reveals systematic differences in long-range organization, distinguishing natural language from state-of-the-art LLM outputs beyond surface-level fluency.

1 Introduction

Recent large language models (LLMs) generate highly fluent and coherent text, achieving strong performance across a wide range of language tasks. However, it remains unclear whether such models capture the long-range organization of natural language beyond local consistency. Improvements in next-token prediction do not necessarily imply that generated texts exhibit human-like structure at larger scales.

Most evaluations of language models rely on task-based benchmarks or short-context analyses, which primarily measure task performance (Brown et al., 2020; Bubeck et al., 2023). While effective for downstream performance, these approaches provide only indirect evidence about the global statistical structure of generated text. Prior work has identified systematic issues in generated text, including excessive repetition and reduced diversity (Holtzman et al., 2020; Welleck et al., 2020), suggesting that high task performance does not necessarily imply human-like long-range organization.

These limitations suggest that, while LLMs are effective at producing locally coherent outputs, they may struggle to sustain globally consistent organization over long spans. In natural language, expressions are not used in isolation but are repeatedly referred to, reused, and recombined in different contexts. Such organization can be understood as a form of *reference structure*, in which previously established elements are revisited and integrated over long distances.

Since reference structure is primarily realized through reuse, it should be observable as repetition in the sequence. We therefore analyze repetition as a distributional property across scales, capturing how subsequences of different lengths are reused throughout a text.

Repetition has long been recognized as a fundamental property of symbolic sequences in information theory (Ziv and Lempel, 1978; Ornstein and Weiss, 1993). Building on this line of work, we relate repetition statistics to higher-order Rényi entropies, providing a finite-length characterization of entropy growth. As discussed in Section 3, different growth regimes correspond to qualitatively distinct types of structural organization.

Using this framework, we compare natural language and LLM-generated texts. We find that natural language exhibits stable entropy growth, whereas LLM-generated text shows systematic

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8367–8382

shifts in exponents with model size, indicating differences in how structure is reused over long ranges.

Our contributions are as follows: (1) we propose a distributional formulation of repetition that enables a finite-length characterization of entropy growth via higher-order Rényi entropies; and (2) we demonstrate systematic differences between natural language and LLM-generated text in entropy growth behavior, revealing distinct patterns of structural reuse.

2 Related Work

2.1 Evaluating long-range structure in language models

Most evaluations of language models rely on task-based benchmarks or short-context analyses, which primarily measure task performance (Brown et al., 2020; Bubeck et al., 2023). While effective for assessing downstream performance, these approaches provide only indirect evidence about the global statistical organization of generated text.

A complementary line of work has examined language models from an information-theoretic perspective, analyzing properties such as entropy rates, calibration, and memory usage (Braverman et al., 2020), as well as inductive biases toward long-range dependencies (Hahn, 2020; Merrill et al., 2021). Recent studies have also tackled limitations in effectively utilizing long contexts and maintaining coherent structure across extended sequences (Press et al., 2022).

While these approaches provide valuable insights into model behavior, they do not directly address how generated texts reuse structure across scales. Our work addresses this gap by focusing on repeated subsequences as a distributional signal of structural reuse in long texts.

2.2 Repetition in natural and generated texts

Repetition has long been studied as a signature of long-range structure in symbolic sequences. Related ideas arise in universal compression schemes such as Lempel–Ziv coding, where repeated subsequences determine compression performance (Ziv and Lempel, 1978).

Analyses based on maximal repeated subsequences capture extreme instances of such repetition. For i.i.d. sources over a finite alphabet, maximal repetition grows logarithmically with sequence length (Ornstein and Weiss, 1993; Wyner and Ziv, 1989), whereas natural language exhibits stronger maximal repetition growth (Dębowski, 2015). However, such statistics are numerically unstable in finite texts, which limits their use as a primary diagnostic.

In the context of neural text generation, repetition has also been studied as a generation artifact, including degeneration and excessive repetition (Holtzman et al., 2020; Welleck et al., 2020). Subsequent work has further shown that such behaviors are strongly influenced by the choice of decoding strategy, which can lead to qualitatively different trade-offs between diversity, coherence, and repetition across tasks (Wiher et al., 2022). These approaches focus on mitigating or regulating repetition, rather than analyzing it as a structural property of the sequence.

Overall, while repetition has been examined through extreme statistics, compression-based methods, and generation behavior, systematic distributional analyses of repeated subsequences across scales remain limited. This gap motivates the present work.

2.3 Entropy scaling in natural language

A complementary perspective on language structure is to examine how information grows as the context length increases.

A classical approach to long-range structure analyzes the scaling of block entropy, typically using Shannon entropy (Shannon, 1948, 1951). Empirical studies report sublinear entropy growth in natural language, often discussed in the context of power-law-like behavior (Hilberg, 1990; Dębowski, 2011). Entropy estimates converge extremely slowly (Dębowski, 2011; Takahira et al., 2016), leaving open whether observed scaling reflects asymptotic behavior or finite-size effects.

Higher-order Rényi entropies have also been considered in the theoretical analysis of entropy rates (Dębowski, 2011), and in empirical studies of linguistic statistics. In particular, Tanaka-Ishii and Aihara (2015) show that certain Rényi-based measures, such as Yule's K, exhibit approximate constancy with respect to text length, highlighting scale-invariant properties of lexical distributions.

However, prior work has primarily used Rényi entropies either in theoretical settings or as static measures of distributions, rather than to characterize how entropy varies with context length. In contrast, the present work uses Rényi entropies to analyze entropy growth as a function of context

length, linking them to the distribution of repeated subsequences and enabling a finite-length characterization of scaling behavior.

3 Three Entropy Growth Regimes and Structural Reuse

This paper studies how information grows as longer contexts are considered. Let $\mathcal{A}$ be a finite alphabet, and let $X = x_1, x_2, \ldots, x_n$ be a sequence of length n over $\mathcal{A}$. We call a contiguous subsequence of length m a *block*. Let $H_1(m)$ denote the Shannon entropy of blocks of length m. The growth rate of $H_1(m)$ characterizes how much new information is introduced as the block length increases.

Entropy growth is commonly decomposed into an extensive linear term and a subextensive correction (Hilberg, 1990; Dębowski, 2020)

$$H_1(m) = h_1 m + G(m),$$

where h_1 is the entropy rate and $G(m)$ captures deviations from linear growth. If the subextensive term satisfies $G(m) = o(m)$, then $H_1(m)/m \to h_1$ as $m \to \infty$. It remains an open question whether the entropy rate h_1 of natural language is strictly positive; in particular, it has been conjectured that it may vanish under Hilberg-type scaling (Hilberg, 1990; Dębowski, 2020). For finite and empirically accessible ranges of m, the subextensive term $G(m)$ may dominate the observed behavior and thus captures the effective structure of entropy growth.

Three qualitatively distinct regimes have been considered for $G(m)$. First, for i.i.d. or finite-order Markov systems, $G(m) = O(1)$. Second, systems with long-range dependencies and expanding structural degrees of freedom exhibit sublinear growth, often approximated by a power law (Hilberg, 1990):

$$G(m) \propto m^{\beta}. \tag{1}$$

This regime has been linked to grammar-based structure in natural language, where hierarchical rules and expanding sets of patterns give rise to power-law growth (Dębowski, 2020).

Third, studies of predictive information and complexity have identified logarithmic or log-power growth as a distinct regime,

$$G(m) \propto (\log m)^{\gamma}, \tag{2}$$

associated with strong structural reuse (Bialek et al., 2001). Such growth is consistent with systems whose effective description length increases much more slowly than sequence length (Kolmogorov, 1965; Li and Vitányi, 2008). While power-law behavior has been discussed for natural language, as mentioned previously, logarithmic or log-power scaling has not been systematically explored in empirical studies.

These regimes are formulated for Shannon entropy but are speculated to extend to higher-order entropy measures, which are derived from the same block distribution. Differences between entropy orders primarily reflect different weighting of frequent versus rare patterns, while the overall regime is governed by the growth of distinct blocks with length.

In this work, we show that higher-order entropy measures are compatible with a logarithmic-power form, suggesting that entropy growth in natural language lies near this regime at accessible sequence lengths. As discussed in Section 6, this behavior admits an interpretation in terms of strong structural reuse, where texts are generated through recombination and re-indexing of shared linguistic resources.

4 Proposed Method

4.1 Counting Repeated Subsequences

We formally define the number of repeated subsequences and propose a method to characterize their behavior, which reveals systematic differences between natural language texts and state-of-the-art GPT-generated texts.

As mentioned previously, let $X = x_1, x_2, \ldots, x_n$ be a sequence over $\mathcal{A}$, a finite set of alphabet. For $1 \le i < j \le n$, let x_i^j denote the contiguous subsequence from position i to $j-1$. A subsequence is said to be repeated if two subsequences of length m starting at distinct positions i and j coincide, i.e., $x_i^{i+m} = x_j^{j+m}$ for some $i \neq j$.

Let m denote the length of a repeated subsequence, and call any consecutive subsequence a *block*. In a sequence of total length n, there are $T_m = n - m + 1$ blocks of length m. Let K_m denote the number of *distinct* block types of length m. The number of repetitions of length-m blocks is then defined as

$$D_m = T_m - K_m. \tag{3}$$

For example, for a sequence $X =$ "banana" of length $n = 6$, when $m = 2$ the blocks are

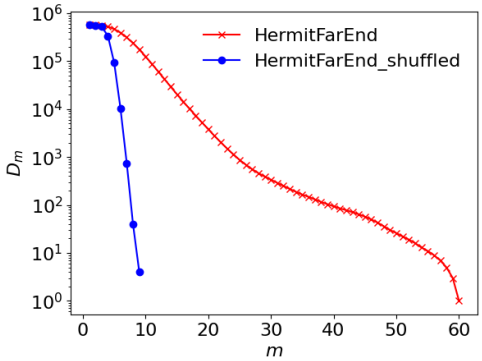

Figure 1: Number of repeated blocks D_m as a function of block length m for *The Hermit of Far End* (red) and its randomly shuffled counterpart (blue).

ba, an, na, an, na, yielding $T_2 = 5 = 6 - 2 + 1$. There are three distinct blocks, ba, an, na, so that $K_2 = 3$ and $D_2 = 2$. Indeed, the blocks "an" and "na" are each repeated once; their total number of repetitions is D_2 for the sequence "banana". Since $1 \leq K_m \leq T_m$, where $K_m = 1$ corresponds to the case in which all blocks are identical and $K_m = T_m$ corresponds to the case in which all blocks are distinct, the range of D_m is $0 \leq D_m \leq T_m - 1$.

Figure 1 shows the observed values of D_m (vertical axis, logarithmic scale) as a function of m (horizontal axis) for *The Hermit of Far End* (by Margaret Pedler, $n = 586{,}533$, red) and its shuffled version (blue). The shuffled text exhibits behavior similar to that of a Bernoulli process, as shown in Supplementary A. For small values of m, essentially all $|\mathcal{A}|^m$ possible blocks appear in the sequence. As m increases, the number of repetitions decreases sharply.

In contrast, natural language displays a markedly different and strongly non-linear pattern. In particular, repetitions persist up to block lengths close to 60, indicating that while the maximally repeated subsequence can be very long, the distribution of shorter repeated blocks already reflects characteristic structural properties of the text.

Nevertheless, the absolute vertical position of D_m is dominated by the document length n, and its non-linear shape is difficult to capture with a simple functional form. Since the vertical axis is plotted on a logarithmic scale, this observation naturally motivates an information-theoretic characterization of repetition, which we develop in the next subsection.

4.2 Higher-order Rényi Entropy

Let S_m denote the total number of possible block types of length m over the alphabet $\mathcal{A}$. For a completely random sequence, $S_m = |\mathcal{A}|^m$, whereas for structured systems such as natural language, typically $S_m \ll |\mathcal{A}|^m$. Among these S_m possible blocks, only a subset actually appears in a given sequence; this observed subset is counted by K_m. Let p_w denote the probability of occurrence of a particular block w. For the moment, we restrict attention to blocks of fixed length $|w| = m$. The probability that w does not appear in any of the T_m positions is $(1 - p_w)^{T_m}$, and hence the probability that it appears at least once is $1-(1-p_w)^{T_m}$, whose sum over $|w| = m$ is $E[K_m]$, the expected number of distinct observed blocks. Therefore, from formula (3),

$$E[D_m] = T_m - \sum_{|w|=m} \left(1 - (1 - p_w)^{T_m}\right). \quad (4)$$

For a sufficiently long sequence, T_m is large, and thus $(1 - p_w)^{T_m} \approx e^{-T_m p_w}$, which can be expanded as

$$e^{-T_m p_w} = 1 - T_m p_w + \frac{T_m^2 p_w^2}{2!} - \frac{T_m^3 p_w^3}{3!} + \cdots.$$

Substituting this expansion into (4), the first-order terms cancel, yielding

$$E[D_m] \approx \sum_{|w|=m} \left(\frac{T_m^2 p_w^2}{2!} - \frac{T_m^3 p_w^3}{3!} + \cdots\right). \quad (5)$$

Therefore, $E[D_m]$ is characterized by the spectrum $\sum_{|w|=m} p_w^\alpha$ with $\alpha \geq 2$.

As noted above, we wanted to analyze $E[D_m]$ on a logarithmic scale. This spectrum is naturally captured by the Rényi entropy of order α (Rényi, 1961), defined as

$$H_\alpha(m) = \frac{1}{1-\alpha} \log_2 \sum_{|w|=m} p_w^\alpha. \quad (6)$$

Larger values of α place greater weight on frequently occurring blocks, whereas smaller values emphasize overall diversity. As is well known, $H_\alpha(m) \to H_1(m)$ as $\alpha \to 1$. Furthermore, if the underlying stochastic process is stationary and ergodic, and if the Rényi entropy rate h_α exists, then $H_\alpha(m)/m \to h_\alpha$ as $m \to \infty$ (Cover and Thomas, 2006). For a Bernoulli process with $p = 0.5$, we have $h_\alpha = 1$ for all α (see Supplementary B).

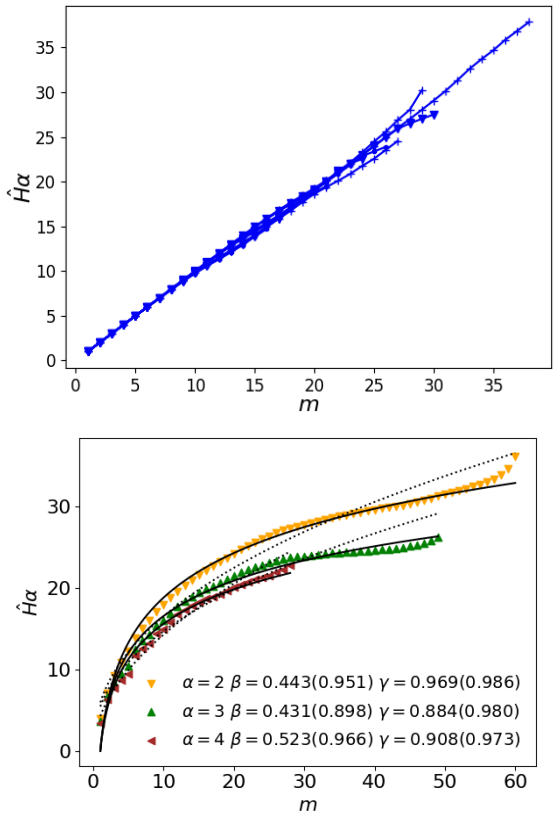

Figure 2: Empirical higher-order Rényi entropy $\hat{H}_\alpha(m)$ for $\alpha = 2, 3, 4$. Top: Bernoulli process with $p = 0.5$ for sequence lengths 10k, 100k, and 1M (all curves collapse). Bottom: *The Hermit of Far End*.

Figure 2 shows the empirical spectra $\hat{H}_\alpha(m) = \frac{1}{1-\alpha} \log_2 \sum_{i=1}^{K_m} \hat{p}_{w_i}^\alpha$, where $\hat{p}_w \equiv \#w/T_m$ and $\#w$ denotes the frequency of block type w in the sequence, restricted to $\#w \geq \alpha$. The rationale for approximating p_w by this empirical distribution is deferred to Supplementary C, and we proceed with the main text.

The top panel shows higher-order Rényi entropies for three Bernoulli processes of different lengths, with $p = 0.5$ and $\alpha = 2, 3, 4$, yielding nine curves in total. All curves collapse onto a single line with slope 1. In contrast, the bottom panel shows $\hat{H}_\alpha(m)$ for *The Hermit of Far End*, where $\alpha = 2, 3, 4$ are shown in yellow, green, and brown, respectively. The shuffled versions of *The Hermit of Far End* exhibit linear collapse similar to the Bernoulli case and are omitted for clarity. Larger values of α yield shorter curves due to the restriction $\#w \geq \alpha$.

4.3 Functional Characterization of Spectrum

We now seek a functional description of the empirical spectrum $\hat{H}_\alpha(m)$. The empirical $\hat{H}_\alpha(m)$ increases smoothly over a wide range of m, but shows a sharp rise as it approaches the upper bound imposed by the total number of blocks T_m.

Following the derivation in Supplementary D, it can be shown that

$$H_\alpha(m) \approx \log_2 S_m - \Delta_\alpha, \tag{7}$$

where Δ_α is expressed as a weighted logarithm of Touchard polynomials. For example, for $\alpha = 2$ we have $\Delta_2 = \log_2 \left(1 + \frac{1}{\lambda_m}\right)$ with $\lambda_m \equiv T_m/S_m$, while Δ_3 and Δ_4 are given in the same appendix. The term Δ_α therefore represents a finite-size correction arising from the upper bound imposed by T_m. Consequently, we focus on modeling the leading term $\log_2 S_m$, which captures the effective growth of the the number of distinguishable blocks underlying entropy scaling.

The classification in Section 3, given by (1) and (2), suggests two distinct forms of sublinear entropy growth. The first model is the power-law ansatz:

$$\log_2 S_m \propto m^\beta, \tag{8}$$

where β is the exponent. This form corresponds to systems in which the effective number of distinguishable blocks continues to expand with block length, reflecting increasing structural degrees of freedom. The second model is the logarithmic–power ansatz:

$$\log_2 S_m \propto (\log_2 m)^\gamma, \tag{9}$$

where γ is the exponent. This form corresponds to entropy growth dominated by the reuse and recombination of previously established structure.

We do not include an explicit linear term hm in the regression (see Section 3), because over the finite range of m considered, it is not reliably separable from the subextensive component. As our focus is on entropy growth behavior, we model $\log S_m$ directly using sublinear forms.

Direct estimation of β or γ from K_m alone is not feasible, since $K_m \ll S_m$ for large m. Alternatively, estimating these exponents directly from (7) by substituting the ansatz leads to unstable results, because the logarithmic correction term Δ_α can be non-negligible at finite lengths. We therefore adopt a two-stage estimation procedure:

1. Estimate λ_m from its functional relation with D_m/T_m, as derived in Supplementary E, and

2. Estimate β and γ by fitting $\log_2 S_m$ in $\hat{H}_\alpha(m) = \log_2 S_m - \Delta_\alpha$, where Δ_α depends

only on λ_m and $\hat{H}_\alpha(m)$ is empirically estimated.

Figure 2 (bottom) shows the resulting fits to $\hat{H}_\alpha(m)$ for $\alpha = 2, 3, 4$ for *The Hermit of Far End*. Power-law fits are shown as dotted lines, while log-power fits are shown as solid lines. The power-law model systematically overestimates the growth rate, particularly for $\alpha = 2$ and 3, where it increases too rapidly at large m. In contrast, the log-power model captures the overall trend substantially better across all values of α. The legend reports the estimated values of β and γ, together with the coefficient of determination, as follows:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

where, y_i are observed values ($\hat{H}_\alpha(m) + \Delta_\alpha$), $\hat{y}_i$ are fitted values to the model (power or log-power of m), and $\bar{y}$ is the mean of the observed values. Values closer to 1 indicate a better fit. For *The Hermit of Far End*, although the power-law model (dotted lines) yields R^2 values in the range 0.90–0.96, its deviations are visually larger than those of the log-power model (solid lines), for which $R^2 > 0.97$. Hence, the log-power model is preferred over the power-law model for all tested values of α.

As we will show below, empirical estimates of $\hat{H}_\alpha(m)$ for natural language often lie near the boundary between these two regimes, making it difficult to distinguish between a pure power-law and a log-power form at accessible sequence lengths.

5 Experiments

We evaluate long-range organization in text by analyzing repeated-subsequence statistics. Differences between natural language and LLM outputs are assessed via fitted entropy-growth parameters and statistical tests. All comparisons are conducted on collections of long texts with matched length distributions.

5.1 Dataset

Our analysis targets entropy growth over a wide range of block lengths, which requires long and coherent sequences. This substantially restricts the choice of data: among naturally occurring texts, extended narratives such as novels provide one of the few sources of sufficiently long sequences. For the same reason, previous studies of maximal repetition have also focused on novels (Dębowski, 2015),

Table 1: Datasets with the mean and standard deviation of text length. Natural language texts are sampled from Project Gutenberg to match the length distribution of the corresponding GPT datasets.

dataset	number	length (chars)
GPT-generated text		
gpt-3.5turbo	100	35044.91±2287.31
gpt-4o-mini	100	110888.61±23378.64
gpt-5-mini	100	347044.85±19793.48
gpt-5	100	601187.13±24972.87
Natural language text		
nl-3.5turbo	100	34741.86±1997.81
nl-4o-mini	100	108762.77±24223.76
nl-5-mini	100	346913.54±18034.12
nl-5	100	593630.96±27223.67

which provides an additional motivation for our choice of data.

Based on these considerations, we analyze both natural language texts and texts generated by GPT models. Specifically, we consider outputs from gpt-3.5turbo, gpt-4o-mini, gpt-5-mini, and gpt-5, as well as human-written natural language texts. Earlier models such as gpt-1 and gpt-2 are not included, as they are unable to reliably generate texts of sufficient length.

For non-stationary data, the behavior of $H_\alpha(m)$ may still depend on the overall sequence length n. However, as we will show below, the estimated exponents for natural language exhibit a degree of universality. To control for length effects, we first sampled long stories from each GPT model under the generation settings described in Supplementary F. For comparison, we sampled natural language texts from the Project Gutenberg corpus, so that length distributions are matched to those of corresponding gpt-X.

Before sampling, we preprocessed all texts from the Standardized Project Gutenberg Corpus (Gerlach and Font-Clos, 2018) by removing metadata and layout-related whitespace. We restrict our analysis to languages that use alphabetic writing systems, leaving extensions to non-alphabetic scripts for future work. For each data category, we sampled 100 texts.

Finally, we note a potential domain-related consideration arising from the use of narrative texts. As mentioned, long coherent texts are not abundantly available outside of narrative genres, which is the main reason for our choice. To assess whether our findings are specific to narratives, we conducted additional experiments using GPT-generated texts under alternative prompts, including essays and scientific-style writing. These results are reported

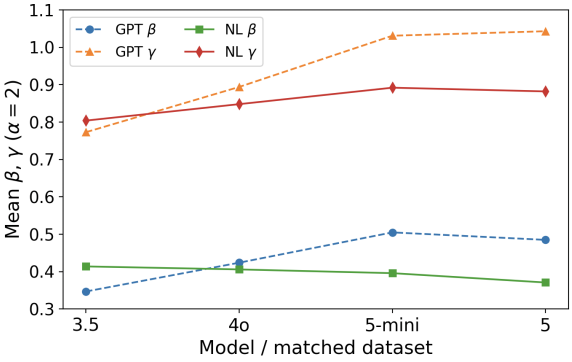

Figure 3: Mean exponents β and γ for $\alpha = 2$. Natural language remains approximately stable across datasets, whereas GPT-generated text exhibits a monotonic increase with model size. This figure summarizes the central tendency of the distributions shown in Figures 4 and 5.

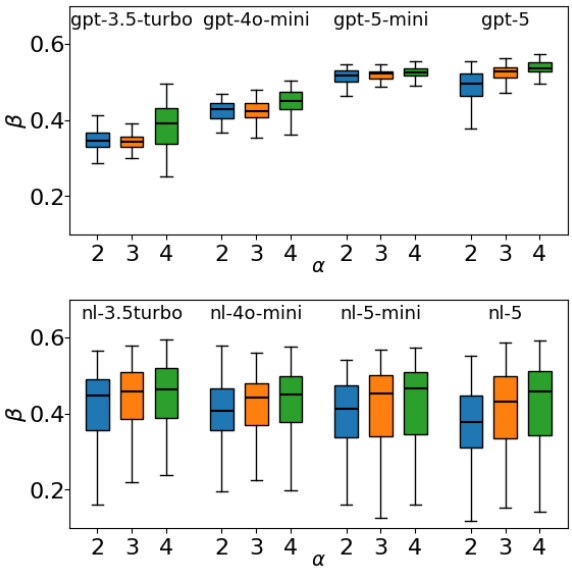

Figure 4: Boxplots of β for each dataset.

in Supplementary G. Although such texts are typically much shorter, the overall trends remain almost consistent with those observed for narratives.

5.2 β and γ

Figure 3 provides a consolidated summary view of the mean exponents β and γ for $\alpha = 2$. Natural language remains approximately stable across datasets despite large differences in text length, whereas GPT-generated text exhibits a clear monotonic increase in both β and γ with model size. For gpt-5 and gpt-5-mini, both β and γ are substantially larger than for natural language.

This contrast is further supported by the full distributions shown in Figure 4 and Figure 5 (See Supplementary H for numerical values). Natural language displays substantial variability across individual texts but maintains stable mean values,

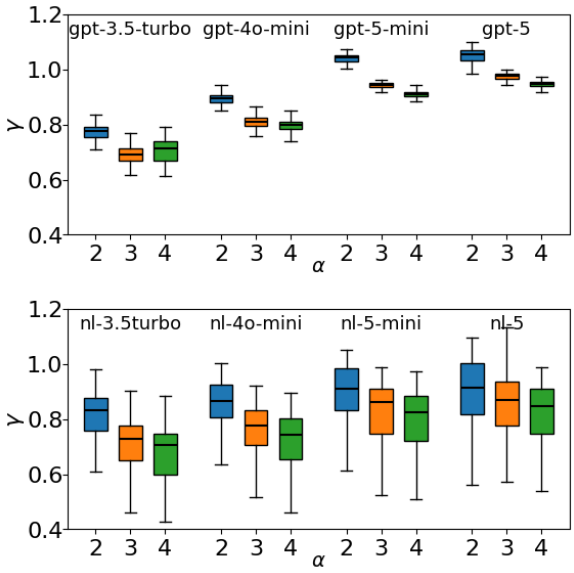

Figure 5: Boxplots of γ for each dataset.

suggesting a form of weak universality in which growth behavior emerges at the population level. In contrast, GPT-generated texts are more homogeneous and exhibit a systematic shift toward larger exponent values as model size increases from gpt-3.5 to gpt-5. As a reference, representative examples of $\hat{H}_\alpha(m)$ for individual GPT-generated text samples are shown in Supplementary I. These examples provide qualitative context for interpreting the large-scale statistical analyses in this section.

Welch two-sample t-tests indicate clear statistical separation between GPT-generated and natural language texts (gpt-5 vs. nl-5 and gpt-5-mini vs. nl-5-mini, with $p \approx 0$ for both β and γ). Within natural language, comparisons between nl-5 vs. nl-5-mini do not yield statistically significant differences ($p = 0.12$ for β, $p = 0.94$ for γ). This is consistent with higher variability but no systematic shift in exponent values, in contrast to the monotonic increase observed for GPT models.

Turning to the dependence on the Rényi order α, we observe a systematic increase of β with α. Since larger α places greater weight on high-probability events, this indicates that frequent blocks exhibit stronger effective growth in structural degrees of freedom. In contrast, the parameter γ decreases with α, implying that lower-order entropies, which remain sensitive to rare events, exhibit more pronounced convexity. This suggests that rare patterns are more strongly governed by structural reuse, leading to slower effective growth. Taken together, these trends indicate that entropy growth is not uniform across the distribution of patterns: frequent and rare structures follow qualitatively different

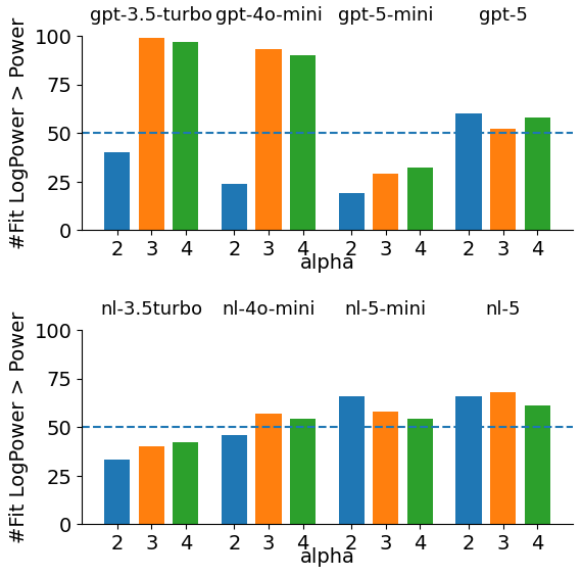

Figure 6: Number of texts for which the log-power model achieves a higher coefficient of determination than the power-law model.

behaviors. This heterogeneity is observed in both natural language and GPT-generated text.

Across all datasets, the average coefficient of determination R^2 exceeds 0.87 and is typically close to 0.95. Figure 6 summarizes model preference by counting the number of texts for which the log-power fit achieves a higher R^2 than the power-law fit (see Supplementary H for numerical values).

For GPT-generated texts (Figure 6 top), smaller models show a strong preference for the log-power model at $\alpha = 3$ and 4, suggesting the presence of many repeated but relatively uninformative sequences. In gpt-5-mini, the power-law tendency becomes stronger, while gpt-5 approaches the behavior observed in natural language. These results indicate that the relative fit of power-law versus log-power models depends strongly on the model.

In contrast, for natural language texts, model preference varies systematically with text length. Shorter texts tend to favor the power-law model, suggesting continued introduction of new information, whereas longer texts increasingly favor the log-power model. This transition indicates that, as text length increases, structural reuse through reference mechanisms becomes more prominent, leading to more convex entropy growth and improved fit by the log-power model. An extreme example is shown in Figure 12 (of Supplementary J), which plots $\hat{H}_\alpha(m)$ for the complete works of Shakespeare ($n = 5{,}442{,}126$). Although this text contains a very long maximally repeated sequence, the effective growth of information saturates at

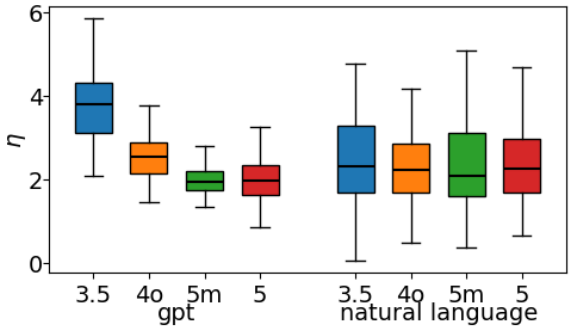

Figure 7: Boxplots of the exponent η governing the growth of maximally repeated subsequences (baseline comparison).

moderate block lengths. This behavior strongly favors a log-power function, and may even suggest an asymptotic form closer to a log–log power law.

5.3 Comparison with maximal repetition

For comparison with prior work, we analyze the growth of maximally repeated subsequences. A maximally repeated subsequence is defined as a subsequence $x_i^{i+m_{\max}}$ such that

$$m_{\max} = \max\left\{\ell \;\middle|\; \exists\, i \neq j \text{ with } x_i^{i+\ell} = x_j^{j+\ell}\right\}. \tag{10}$$

It is known that

$$m_{\max}(n) \propto (\log_2 n)^{\eta}, \tag{11}$$

where $\eta \to 1$ for i.i.d. sequences (Ornstein and Weiss, 1993), whereas empirical studies of natural language report $\eta > 1$ (Dębowski, 2015).

Figure 7 shows boxplots of the exponent η governing maximal repetition growth. Overall, the qualitative trends observed for η are consistent with those obtained for the exponents β and γ. Earlier GPT models exhibit larger values of η, indicating the presence of many long but relatively uninformative repetitions. This tendency largely disappears in gpt-5, whose mean η values become comparable to those of natural language. In contrast, natural language consistently shows a larger variance in η, while its mean values remain stable across different text lengths, in line with the observations for β and γ.

Under maximal repetition, gpt-5 behaves similarly to natural language apart from the difference in variance, whereas the proposed approach reveals systematically larger exponents for GPT-generated texts, yielding a more pronounced distinction. Moreover, estimates based on maximal

repetition are numerically unstable due to their reliance on extreme statistics, as discussed in Supplementary K, which further limits their effectiveness as a diagnostic compared to the distribution-based method proposed here.

Finally, our results also refine previous estimates of η for natural language. Prior work reported values around $\eta \approx 3$ (Dębowski, 2015), based in part on aggregated corpora such as the complete works of Shakespeare, which exhibit substantial redundancy due to corpus-level aggregation (see Supplementary J). In contrast, we find that for individual texts, η typically converges to values slightly above 2, as shown in Figure 7, suggesting that earlier estimates may reflect corpus composition.

6 Discussion

Our results refine the interpretation of entropy growth in natural language by distinguishing between power-law and logarithmic–power scaling. Although both appear sublinear over finite ranges, they correspond to qualitatively different patterns of information accumulation: power-law growth is consistent with a continual increase in irreducible content, whereas logarithmic–power growth reflects a slower expansion of distinguishable structure.

This distinction can be interpreted through the lens of algorithmic information theory. Sequences generated by simple mechanisms, such as short programs that repeatedly reuse the same rules, admit descriptions whose Kolmogorov complexity grows much more slowly than sequence length (Kolmogorov, 1965; Li and Vitányi, 2008). By analogy, increases in description length may be dominated by specifying references to previously defined structures rather than encoding new content.

Applied to language, a log-power fit is consistent with strong structural reuse, where texts are generated through recombination and re-indexing of shared linguistic resources, yielding sublinear growth in informational complexity. This view remains compatible with Zipfian frequency distributions and unbounded vocabulary growth (Zipf, 1949; Herdan, 1964; Heaps, 1978), since lexical innovation can coexist with slower growth in syntactic and discourse-level structure.

Within this framework, the differences observed between natural language and GPT-generated text suggest a weaker manifestation of this structure-as-reference organization in current models. Although model outputs exhibit repetition and long-range dependencies, the estimated exponents and their dependence on Rényi order indicate a greater reliance on locally generated structure, possibly due to the next-token prediction objective and finite context utilization.

These results provide empirical support for viewing natural language as operating near a regime dominated by structural reuse. Further work is needed to clarify how training objectives and architectures shape repetition-based entropy growth.

7 Conclusion

We introduced a repetition-based framework for analyzing long-range structure in natural language and large language model outputs. By relating repeated subsequences to higher-order Rényi entropies and accounting for finite-size effects, we showed that entropy growth is often better characterized by logarithmic–power than by power-law models.

Across datasets, natural language exhibits stable and strongly sublinear growth patterns, whereas GPT-generated texts show systematic and statistically significant shifts in exponents with model size. These results reveal persistent differences in long-range statistical organization that are not captured by standard task-based evaluations.

Overall, our findings establish repeated-subsequence entropy as a quantitative diagnostic for long-range structure, providing a simple way in finite-length data to assess how closely model-generated text matches the statistical patterns of natural language.

8 Limitations

The limitations discussed below fall into three categories: scope, methodology, and interpretation.

Scope limitations. A first limitation concerns the scope of the data. The present analysis focuses on structural reuse within individual texts, whereas classical studies of Shannon entropy (H_1) typically analyze large heterogeneous corpora aggregated across multiple authors, topics, and styles, rather than single coherent documents. As a result, our approach captures intra-text organization but does not directly address cross-text variability or population-level statistics. Extending the framework to larger-scale datasets that enable analysis of both within-

text and across-text structure remains an important direction for future work.

A second limitation is that our experiments focus on GPT-family models as representative large language models. Although these models span multiple generations and sizes, the results do not necessarily generalize to all architectures or training paradigms. Models with different objectives, architectures, or external memory components may exhibit different entropy-growth behavior.

A third limitation is that our analysis is conducted at the alphabetic character level. This choice avoids tokenization-specific artifacts and enables a clean information-theoretic treatment, but does not directly operate on higher-level linguistic units such as words, morphemes, or syntactic constructions.

Methodological limitations. A fourth limitation is that our analysis does not directly resolve the presence of an extensive linear component in entropy growth. Because we focus on the growth behavior of $H_\alpha(m)$ rather than on the ratio $H_\alpha(m)/m$, we do not obtain a direct estimate of the entropy rate h_α or determine whether a linear term emerges at larger scales.

A fifth limitation concerns the range of block lengths considered. Our conclusions are based on entropy-growth behavior over empirically accessible values of m. Although we account for finite-size effects, longer texts or alternative estimation methods may reveal additional structure beyond the ranges considered here.

Interpretative limitations. A sixth limitation is that the proposed approach is not designed to measure task performance or downstream capabilities. Instead, it provides a complementary structural signal. Understanding how these measures relate to functional performance remains an open question.

Finally, while we observe systematic and statistically significant differences between natural language and LLM-generated text, our analysis is descriptive and does not identify the mechanisms responsible for these differences. Establishing causal links between entropy-growth patterns and model properties remains an important direction for future work.

Acknowledgements

This work was supported by JST CREST, Japan, Grant Number JPMJCR2114.

AI assistants were used during the development and writing of this work. All technical content, analyses, and conclusions are the sole responsibility of the author.

References

William Bialek, Ilya Nemenman, and Naftali Tishby. 2001. Predictability, complexity, and learning. *Neural Computation*, 13(11):2409–2463.

Mark Braverman, Xinyi Chen, Sham Kakade, Karthik Narasimhan, Cyril Zhang, and Yi Zhang. 2020. Calibration, entropy rates, and memory in language models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1089–1099.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2nd edition. Wiley-Interscience.

Łukasz Dębowski. 2011. Mixing, ergodic, and nonergodic processes with rapidly growing information between blocks. *IEEE Transactions on Information Theory*, 58(6):3392–3401.

Łukasz Dębowski. 2015. Maximal repetitions in written texts: Finite energy hypothesis vs. strong hilberg conjecture. *Entropy*, 17(8):5903–5919.

Łukasz Dębowski. 2020. *Information Theory Meets Power Laws : Stochastic processes and Language Models*. Wiley.

Martin Gerlach and Francesc Font-Clos. 2018. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *arXiv preprint*.

Michael Hahn. 2020. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171.

Harold S. Heaps. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press.

Gustav Herdan. 1964. *Quantitative Linguistics*. Butterworths.

Wolfgang Hilberg. 1990. Der bekannte grenzwert der redundanzfreien information in texten—eine fehlinterpretation der shannonschen experimente? *Frequenz*, 44:243–248.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference for Learning Representation (ICLR)*.

Donald E. Knuth. 1997. *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*, 3rd edition. Addison Wesley Longman, Reading, MA.

Andrei N. Kolmogorov. 1965. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7.

Ming Li and Paul M. B. Vitányi. 2008. *An Introduction to Kolmogorov Complexity and Its Applications*, 3rd edition. Springer.

William Merrill, Vivek Ramanujan, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Effects of parameter norm growth during transformer training: Inductive bias from gradient descent. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1766–1781, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rajeev Motwani and Prabhakar Raghavan. 1995. *Randomized Algorithms*. Cambridge University Press, Cambridge.

Donald S. Ornstein and Benjamin Weiss. 1993. Entropy and data compression schemes. *IEEE Transactions on Information Theory*, 39(1):78–83.

Ofir Press, Noah A. Smith, and Omer Levy. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR)*.

Alfréd Rényi. 1961. On measures of entropy and information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:547–561.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Claude E. Shannon. 1951. Prediction and entropy of printed english. *Bell System Technical Journal*, 30(1):50–64.

Ryosuke Takahira, Kumiko Tanaka-Ishii, and Łukasz Dębowski. 2016. Entropy rate estimates for natural language—a new extrapolation of compressed large-scale corpora. *Entropy*, 18(10):364.

Kumiko Tanaka-Ishii and Shunsuke Aihara. 2015. Computational constancy measures of Texts—Yule's k and rényi's entropy. *Computational Linguistics*, 41(3):481–502.

Jacques Touchard. 1956. Nombres exponentiels et nombres de Bernoulli. *Canadian Journal of Mathematics*, 8:305–320.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations (ICLR)*.

Gian Wiher, Clara Meister, and Ryan Cotterell. 2022. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012.

Aaron D. Wyner and Jacob Ziv. 1989. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Transactions on Information Theory*, 35(6):1250–1258.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

Jacob Ziv and Abraham Lempel. 1978. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536.

Supplementary Material

A Repeated Sequences for a Bernoulli Process ($p = 0.5$)

Figure 8 shows the counts of repeated subsequences D_m (vertical axis, logarithmic scale) as a function of block length m (horizontal axis) for Bernoulli processes of length $n = 10\text{k}$, 100k, and 1m. Longer sequences appear higher in the plot, reflecting the larger number of available blocks.

As discussed in the main text (Section 4.1), shuffled natural language texts exhibit the same qualitative behavior as the Bernoulli process. For small m, essentially all $|\mathcal{A}|^m$ possible blocks appear in the sequence. As m increases, the number of repeated blocks decreases sharply. This transition corresponds to the well-known *birthday limit* (Motwani and Raghavan, 1995; Knuth, 1997).

B $h_\alpha = 1$ for a Bernoulli Process with $p = 0.5$

For a Bernoulli process, the following derivation applies. Let the probability of symbol 1 be p. Then the probability of a word w of length m containing k ones is given by

$$p_w = p^k(1-p)^{m-k}.$$

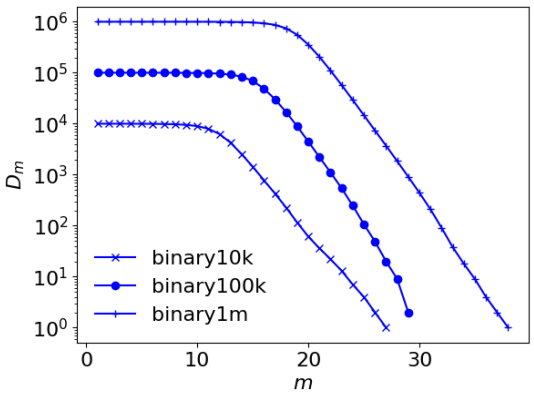

Figure 8: Repeated subsequences D_m for a Bernoulli process with $p = 0.5$ ($n = 10$k, 100k, and 1m).

Therefore,

$$\begin{aligned}\sum_{|w|=m} p_w^\alpha &= \sum_{k=0}^{m} \binom{m}{k} (p^\alpha)^k \left((1-p)^\alpha\right)^{m-k} \\ &= (p^\alpha + (1-p)^\alpha)^m .\end{aligned}$$

Consequently,

$$\begin{aligned}H_\alpha(m) &= \frac{1}{1-\alpha} \log_2 \left(\sum_{|w|=m} p_w^\alpha \right) \\ &= \frac{1}{1-\alpha} \log_2 \left(p^\alpha + (1-p)^\alpha\right)^m .\end{aligned}$$

In particular, when $p = 0.5$, the Rényi entropy grows linearly with m with slope 1, independently of the value of α.

C Rationale for Probability Estimation of p_w

Throughout this paper, we estimate the block probability p_w by the empirical frequency

$$\hat{p}_w = \frac{\#w}{T_m},$$

where $\#w$ denotes the number of occurrences of block w among the $T_m = n - m + 1$ length-m blocks extracted from a sequence of length n. This choice corresponds to defining p_w as the probability that block w is observed when a start position is sampled uniformly at random from the T_m available positions.

Under stationarity and ergodicity, this estimator is consistent for the true block probability $\Pr(X_1^m = w)$. More generally, for non-stationary texts, $\hat{p}_w$ estimates the position-averaged block distribution, which reflects the empirical statistics of the finite sequence under uniform sampling of block positions. This operational definition is well aligned with our finite-length analysis of repeated subsequences.

In contrast, normalizing by the number of possible block types S_m would not yield a meaningful probability distribution over observed blocks, since S_m counts types rather than sampling trials. Moreover, our finite-size corrections and occupancy-based derivations, including the definition of $\lambda_m = T_m/S_m$, are naturally formulated in terms of T_m as the number of block occurrences.

Furthermore, in this work, $\hat{H}_\alpha(m)$ is computed using only blocks with $\#w \geq \alpha$. The first reason is conceptual: the Rényi order α directly corresponds to the degree of repetition being emphasized, and blocks occurring fewer than α times do not meaningfully contribute to this regime. The second reason is empirical: when rarer blocks with $\#w < \alpha$ are included, the empirical estimates for a Bernoulli process systematically deviate from the theoretical predictions and fail to converge to the expected behavior.

D Derivation of Δ_2

To derive formula (7) in the main text, we consider the case $\alpha = 2$. Assume that words of length m occur independently according to a Poisson distribution with mean $\lambda_m = T_m/S_m$. Among the S_m possible word types, K_m types are observed in the sequence. Let C_i denote the number of occurrences of the i-th word type. For $\alpha = 2$, we have

$$\sum_{i=1}^{K_m} p_i^2 = \sum_{i=1}^{S_m} \left(\frac{C_i}{T_m}\right)^2 = \frac{1}{T_m^2} \sum_{i=1}^{S_m} C_i^2.$$

Since unobserved word types satisfy $C_i = 0$, the summation may be taken over all S_m types. Because $C_i \sim \text{Poisson}(\lambda_m)$, we obtain

$$E[C_i^2] = \text{Var}(C_i) + (E[C_i])^2 = \lambda_m + \lambda_m^2.$$

Assuming independence, it follows that

$$E\left[\sum_{i=1}^{S_m} C_i^2\right] = \sum_{i=1}^{S_m} E[C_i^2] = S_m(\lambda_m^2 + \lambda_m).$$

Therefore,

$$E\left[\sum_{i=1}^{K_m} p_i^2\right] = \frac{S_m(\lambda_m^2 + \lambda_m)}{T_m^2}.$$

Substituting $T_m = S_m \lambda_m$, we obtain

$$E\left[\sum_{i=1}^{K_m} p_i^2\right] = \frac{1}{S_m}\left(1 + \frac{1}{\lambda_m}\right).$$

To take logarithms, we use the approximation

$$E[-\log X] \approx -\log E[X],$$

which yields

$$\begin{aligned} H_2(m) &= -\log\left[\frac{1}{S_m}\left(1+\frac{1}{\lambda_m}\right)\right] \\ &= \log S_m - \log\left(1+\frac{1}{\lambda_m}\right). \end{aligned}$$

This yields the desired form, with correction term $\Delta_2 = \log\left(1+\frac{1}{\lambda_m}\right)$. As noted in the main text, these correction terms can be expressed as weighted logarithms of Touchard polynomials (Touchard, 1956). Similarly, for $\alpha = 3$ and $\alpha = 4$, the corresponding correction terms Δ_3 and Δ_4 are given by as

$$\begin{aligned} \Delta_3 &= \frac{1}{2}\log\left(1+\frac{3}{\lambda_m}+\frac{1}{\lambda_m^2}\right), \\ \Delta_4 &= \frac{1}{3}\log\left(1+\frac{6}{\lambda_m}+\frac{7}{\lambda_m^2}+\frac{1}{\lambda_m^3}\right). \end{aligned}$$

E Relation between λ_m and D_m

It can be shown that the relationship between $\lambda_m = T_m/S_m$ and D_m is given by

$$E[D_m] = T_m f(\lambda_m), \tag{12}$$

where

$$f(\lambda) \equiv \left(1 - \frac{1-e^{-\lambda}}{\lambda}\right). \tag{13}$$

When p_w is uniformly distributed (i.e., $p_w = 1/S_m$), we have

$$\sum_{|w|=m} e^{-T_m p_w} = S_m e^{-T_m/S_m}, \tag{14}$$

and hence, from formula (4) in the main text,

$$\begin{aligned} E[D_m] &\approx T_m - S_m(1-e^{-T_m/S_m}) \\ &= T_m\left(1 - \frac{1-e^{-T_m/S_m}}{T_m/S_m}\right) \\ &= T_m\left(1-\frac{1-e^{-\lambda_m}}{\lambda_m}\right). \end{aligned}$$

For a general distribution p_w, we use the expansion of formula (5) in the main text, whose first-order term cancels out. This implies that the probability of repetition is determined by the second moment. Following the standard technique of *effective uniformization*, we introduce a uniform distribution with the same second moment $q_m = \sum_{|w|=m} p_w^2$. Let $S_{\text{eff}} = 1/q_m$, and define

$$\tilde{p}_w \equiv \frac{1}{S_{\text{eff}}}. \tag{15}$$

Then,

$$\begin{aligned} \sum_{|w|=m} \tilde{p}_w^2 &= S_{\text{eff}}\left(\frac{1}{S_{\text{eff}}}\right)^2 && (16) \\ &= \frac{1}{S_{\text{eff}}} && (17) \\ &= q_m. && (18) \end{aligned}$$

Letting $\lambda_m \equiv T_m q_m$, where q_m is the second moment of the original distribution, we can apply the same argument as in the uniform case to obtain

$$\begin{aligned} E[D_m] &\approx T_m - S_{\text{eff}}(1-e^{-T_m/S_{\text{eff}}}) && (19) \\ &= T_m f(\lambda_m), && (20) \end{aligned}$$

which is independent of the specific form of $\tilde{p}_w$.

F GPT Text Generation

GPT models, particularly earlier versions, impose practical limits on the maximum length of a single response. To generate sufficiently long texts for our analysis, we adopt a strategy in which each model is instructed to produce a single coherent story in multiple parts.

Specifically, for each GPT model, we provide the following prompt:

> You are a genius storyteller. I want you to generate a story longer than $numWords$ tokens in num parts. Please tell the story part by part.

The target length $numWords$ is set to 200,000 tokens, reflecting the typical scale of long human-written narratives (on the order of 100,000 tokens). The number of parts num is set to 20, so that each generated text is produced incrementally while maintaining global narrative consistency.

Table 2: Datasets generated for Essay and Scientific Articles

dataset	number	length (chars)
Essay		
gpt-3.5turbo	30	36489.23±2368.41
gpt-4o-mini	30	28196.33±7772.60
gpt-5-mini	30	157692.33±20627.09
gpt-5	30	313300.67±11774.73
Scientific Article		
gpt-3.5turbo	30	39002.53±5422.58
gpt-4o-mini	30	60039.40±8858.72
gpt-5-mini	30	193521.70±25617.89
gpt-5	30	370674.73±12729.61

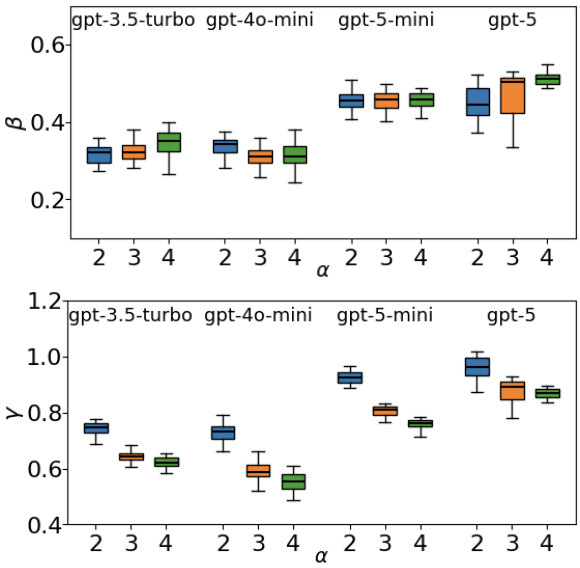

Figure 9: Boxplots of β and γ for GPT generated Essay dataset.

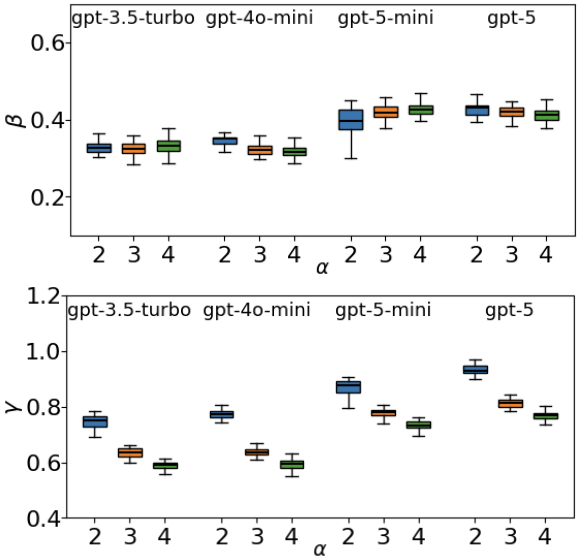

Figure 10: Boxplots of β and γ for GPT generated Scientific Article dataset.

G Domain Effects and Robustness

To examine domain dependence, we compare results across essays and scientific-style texts.

For natural language, obtaining texts of comparable length and coherence to novels in these domains is challenging. As noted in the main text, it is difficult to collect sufficiently long and structurally consistent essays or scientific articles, and the wide variation in topics further limits assembling well-matched data. Extending the analysis to large-scale corpora spanning multiple authors and domains therefore remains an important direction for future work.

In contrast, GPT-generated texts allow controlled generation of alternative domains. Using the same prompts as in the previous section, we generated 30 essays and 30 scientific-style texts by replacing "story". The resulting datasets are summarized in Table 2. These texts are approximately half as long as the generated stories, consistent with the tendency of human-written essays and scientific texts to be shorter than novels.

For these datasets, we constructed boxplots of the estimated exponents, analogous to those in Section 5.2. The results for essays are broadly consistent with those for novels in terms of distributions and model-dependent trends, although model preference favors power-law in a large majority of cases.

In contrast, scientific-style texts exhibit a quantitatively different pattern while preserving the same qualitative trends. The estimated values of β and γ are systematically smaller, and model preference favors power-law model over log-power model in most cases. At the same time, the dependence on model size remains consistent with the main analysis: both β and γ increase with model size, and their dependence on α follows the same pattern, with β increasing and γ decreasing as α grows.

These observations indicate that the overall structural behavior is shared across domains, but that its strength differs. In particular, the weaker log-power preference and smaller exponent values suggest that referential reuse is more limited in GPT-generated scientific-style texts. Overall, the results highlight that domain primarily affects the strength of entropy growth, while the overall trends remain similar.

H Actual Values Used to Produce Figures 4–7

Tables 3–6 report the numerical values used to generate Figures 4–7, respectively. Interpretation and

Table 3: Mean and standard deviation of the estimated exponent β (mean $\pm$ std), together with the coefficient of determination R^2, for $\alpha = 2, 3, 4$ across all datasets.

	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
GPT-generated text			
gpt-3.5	0.347 ± 0.032 (0.956 ± 0.014)	0.344 ± 0.032 (0.890 ± 0.027)	0.389 ± 0.053 (0.878 ± 0.051)
gpt-4o-mini	0.424 ± 0.026 (0.976 ± 0.010)	0.424 ± 0.030 (0.942 ± 0.022)	0.446 ± 0.034 (0.932 ± 0.029)
gpt-5-mini	0.505 ± 0.045 (0.977 ± 0.037)	0.518 ± 0.017 (0.976 ± 0.010)	0.524 ± 0.014 (0.970 ± 0.008)
gpt-5	0.485 ± 0.051 (0.960 ± 0.042)	0.520 ± 0.031 (0.967 ± 0.027)	0.533 ± 0.030 (0.964 ± 0.027)
Natural language text			
nl-3.5	0.414 ± 0.103 (0.952 ± 0.054)	0.427 ± 0.106 (0.933 ± 0.068)	0.439 ± 0.108 (0.920 ± 0.078)
nl-4o	0.406 ± 0.088 (0.941 ± 0.063)	0.418 ± 0.091 (0.920 ± 0.075)	0.425 ± 0.101 (0.910 ± 0.079)
nl-5-mini	0.396 ± 0.094 (0.933 ± 0.056)	0.419 ± 0.103 (0.923 ± 0.078)	0.431 ± 0.105 (0.921 ± 0.073)
nl-5	0.371 ± 0.108 (0.913 ± 0.072)	0.415 ± 0.096 (0.919 ± 0.072)	0.429 ± 0.100 (0.917 ± 0.071)

Table 4: Mean and standard deviation of the estimated exponent γ (mean $\pm$ std), together with the coefficient of determination R^2, for $\alpha = 2, 3, 4$ across all datasets.

	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
GPT-generated text			
gpt-3.5	0.773 ± 0.031 (0.953 ± 0.014)	0.690 ± 0.035 (0.965 ± 0.007)	0.703 ± 0.050 (0.950 ± 0.010)
gpt-4o-mini	0.894 ± 0.023 (0.963 ± 0.013)	0.807 ± 0.026 (0.975 ± 0.005)	0.795 ± 0.026 (0.966 ± 0.005)
gpt-5-mini	1.031 ± 0.041 (0.965 ± 0.013)	0.942 ± 0.011 (0.970 ± 0.007)	0.910 ± 0.011 (0.966 ± 0.006)
gpt-5	1.043 ± 0.048 (0.976 ± 0.012)	0.971 ± 0.023 (0.975 ± 0.006)	0.944 ± 0.021 (0.972 ± 0.005)
Natural language text			
nl-3.5	0.804 ± 0.120 (0.937 ± 0.045)	0.705 ± 0.108 (0.929 ± 0.055)	0.669 ± 0.111 (0.919 ± 0.051)
nl-4o	0.848 ± 0.108 (0.941 ± 0.068)	0.755 ± 0.107 (0.929 ± 0.096)	0.716 ± 0.117 (0.917 ± 0.096)
nl-5-mini	0.892 ± 0.112 (0.961 ± 0.031)	0.818 ± 0.119 (0.952 ± 0.038)	0.793 ± 0.114 (0.946 ± 0.038)
nl-5	0.882 ± 0.142 (0.944 ± 0.064)	0.844 ± 0.122 (0.945 ± 0.067)	0.818 ± 0.122 (0.937 ± 0.076)

discussion of these results are provided in the main text.

Table 5: Number of texts for which the log-power model provides a better fit than the power-law model (LP > P).

	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
GPT-generated text			
gpt-3.5	40	99	97
gpt-4o-mini	24	93	90
gpt-5-mini	19	29	32
gpt-5	60	52	58
Natural language text			
nl-3.5	33	40	42
nl-4o	46	57	54
nl-5-mini	66	58	54
nl-5	66	68	61

Table 6: Exponent η of maximal repetition growth. Values are reported as mean $\pm$ standard deviation together with the coefficient of determination R^2. The column "fail" indicates the number of texts for which the estimation did not converge. Each category contains 100 instances.

Dataset	η (mean $\pm$ std)	R^2	fail
GPT-generated text			
gpt-3.5-turbo	3.815 ± 0.929	0.812 ± 0.119	0
gpt-4o-mini	2.608 ± 0.603	0.862 ± 0.085	0
gpt-5-mini	2.091 ± 0.616	0.829 ± 0.113	0
gpt-5	2.085 ± 0.646	0.810 ± 0.138	0
Natural language text			
nl-3.5	2.855 ± 2.172	0.686 ± 0.235	1
nl-4o	2.732 ± 2.075	0.709 ± 0.225	6
nl-5-mini	2.598 ± 1.788	0.720 ± 0.226	1
nl-5	2.454 ± 1.263	0.715 ± 0.229	7

I $\hat{H}_\alpha(m)$ for Samples of GPT-Generated Texts

Figure 11 shows $\hat{H}_\alpha(m)$ for one representative sample from each of the GPT models considered (gpt-3.5-turbo, gpt-4o-mini, gpt-5-mini, and gpt-5).

For gpt-3.5-turbo and gpt-4o-mini, the empirical curves exhibit a clear preference for the log-power functional form at higher Rényi orders ($\alpha = 3, 4$), indicating increasingly convex entropy growth. In contrast, for $\alpha = 2$, GPT-generated texts typically favor the power-law model across all versions, suggesting that these models continue to introduce new block types even at large text lengths.

J $H_\alpha(m)$ of Shakespeare

Figure 12 shows $\hat{H}_\alpha(m)$ for the complete works of Shakespeare. Although repeated subsequences are observed up to block lengths of approximately $m \approx 350$, the effective growth of entropy continues only up to around $m \approx 50$, beyond which the increase in $\hat{H}_\alpha(m)$ becomes very slow. This figure therefore illustrates an extreme case of long-range repetition. In contrast, for typical individual texts, $\hat{H}_\alpha(m)$ more closely resembles the behavior shown in Figure 2 (bottom), where effective entropy growth decreases gradually and smoothly as m increases.

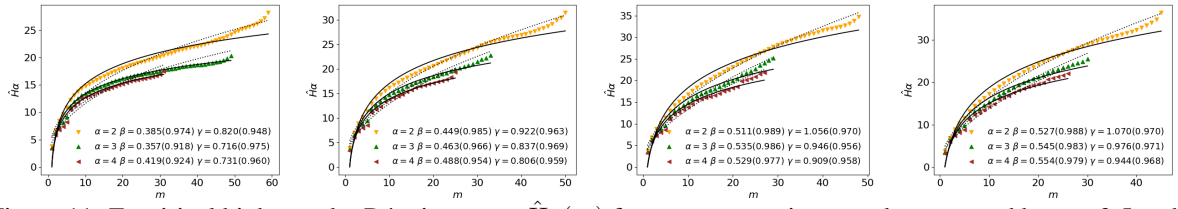

Figure 11: Empirical higher-order Rényi entropy $\hat{H}_\alpha(m)$ for a representative sample generated by gpt-3.5-turbo 4o-mini, 5-mini and 5.

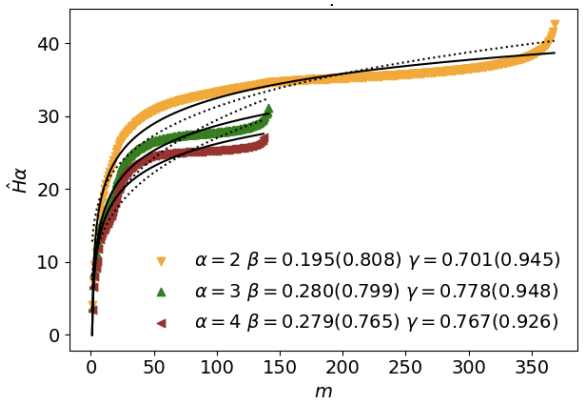

Figure 12: Empirical higher-order Rényi entropy $\hat{H}_\alpha(m)$ for $\alpha = 2, 3, 4$ computed on the complete works of Shakespeare.

K Maximal Repetition: Definition, Examples, and Limitations

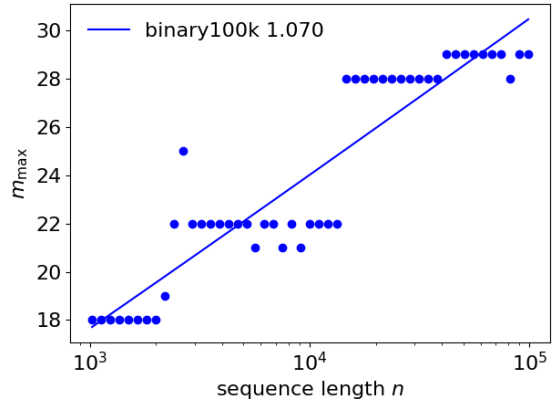

Figure 13: Maximal repetition length for a Bernoulli process.

Figure 13 shows the behavior for a Bernoulli process with $p = 0.5$, where $\eta \approx 1$, consistent with theory. Figure 14 shows results for *The Hermit of Far End*. The shuffled version yields $\eta \approx 1$, whereas the original text gives $\eta \approx 2.4$, indicating stronger repetition.

The growth curves display a stepwise structure, reflecting the fact that only a very small number

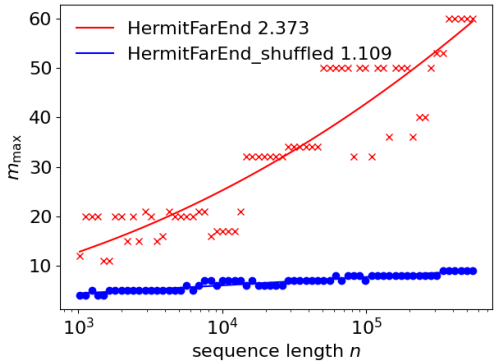

Figure 14: Growth of the maximally repeated subsequence length for a natural language text (red) and its shuffled counterpart (blue).

of subsequences attain the maximal length. As a result, the estimates are numerically unstable and converge slowly, and for a non-negligible fraction of texts, η cannot be reliably estimated.

To reduce noise, the curves in Figure 14 were computed using logarithmically spaced sampling points, where each point represents the median over k samples (with $k = 5$). Despite this smoothing, the estimates remain unstable, as illustrated by the shuffled text, for which $\eta = 1.1$.

These limitations motivate the use of distribution-based measures, which capture repetition across scales rather than relying on extreme statistics.