Correlation Dimension of Natural Language in a Statistical Manifold

Xin Du[1, *] and Kumiko Tanaka-Ishii[2, †]
[1] *Waseda Research Institute for Science and Engineering,*
Waseda University.
[2] *Department of Computer Science and Engineering,*
School of Fundamental Science and Engineering,
Waseda University.

(Dated: May 16, 2024)

The correlation dimension of natural language is measured by applying the Grassberger-Procaccia algorithm to high-dimensional sequences produced by a large-scale language model. This method, previously studied only in a Euclidean space, is reformulated in a statistical manifold via the Fisher-Rao distance. Language exhibits a multifractal, with global self-similarity and a universal dimension around 6.5, which is smaller than those of simple discrete random sequences and larger than that of a Barabási-Albert process. Long memory is the key to producing self-similarity. Our method is applicable to any probabilistic model of real-world discrete sequences, and we show an application to music data.

CONTENTS

[*] duxin.ac@gmail.com

[†] kumiko@waseda.jp

I. INTRODUCTION

The correlation dimension of Grassberger and Procaccia (1983) quantifies the degree of recurrence in a system's evolution and has been applied to examine the characteristics of sequential data, such as the trajectories of strange attractors (Grassberger and Procaccia, 1983), random processes (Osborne and Provenzale, 1989), and sequences sampled from complex networks (Lacasa and Gómez-Gardenes, 2013).

In this letter, we report the correlation dimension of natural language by regarding texts as the trajectories of a language dynamical system. In contrast to the long-memory quality of natural language as reported in (Altmann *et al.*, 2012; Li, 1989; Tanaka-Ishii and Bunde, 2016), the correlation dimension of natural language has barely been studied because of its high dimensionality and discrete nature. An exceptional previous work, to the best of our knowledge, was that of Doxas *et al.* (2010), who measured the correlation dimension of language in terms of a set of paragraphs. Every paragraph was represented as a vector, with each dimension being the logarithm of a word's frequency. The distance between two paragraphs was measured as the Euclidean distance. Such a representation has also been used for measuring other scaling factors of language (Ausloos, 2012; Kobayashi and Tanaka-Ishii, 2018; Tanaka-Ishii and Kobayashi, 2018). However, without a rigorous definition of language as a dynamical system, the correlation dimension is difficult to interpret, and its value may easily depend on the setting. For example, the dimension would vary greatly between handling word frequencies logarithmically and nonlogarithmically.

Today, language representation has become elaborate by incorporating semantic ambiguity and long context. *Large language models* (LLMs) (OpenAI, 2023; Radford *et al.*, 2019; Touvron *et al.*, 2023; Yi, 2024) such as ChatGPT generate texts that are hardly distinguishable from human-generated texts. The generation process is autoregressive, which naturally associates a dynamical system. Such state-of-the-art (SOTA) models (i.e., the GPT series, including GPT-4 (OpenAI, 2023), Llama-2 (Touvron *et al.*, 2023), and "Yi" (Yi, 2024)) have opened a new possibility of studying the physical nature of language as a complex dynamical system. Furthermore, exploration of the fractal dimension of language offers a novel approach to examine the underlying structures of pretrained neural networks, thus shedding light on the intricate ways they mirror human intelligence.

These new systems, however, are not defined in a Euclidean space and thus require reformulation of the state space and the metric between states. Because a neural model assumes a probability space, the analysis method that was originally defined in a Euclidean space must be accommodated in a space of probability distributions, and the distance metric must be statistical. Specifically, we consider a statistical manifold (Amari, 2012; Rao, 1992) whose metric is the Fisher information metric. Hence, this letter proposes a rigorous formalization to analyze the universal properties of these GPT models, thus representing language as an original dynamical system. Although we report results mainly for language, given the impact of ChatGPT, our formalization applies to any other GPT neural models for real-world sequences, such as DNA, music, programming sources, and finance data. To demonstrate this possibility, we show an application to music.

II. METHOD

Let (S, d) be a metric space and $[x_1, x_2, \cdots, x_N]$ be a point sequence, where $x_t \in S$ for $t = 1, \cdots, N$. The Grassberger-Procaccia algorithm (Grassberger and Procaccia, 1983) (GP in the following) defines the correlation dimension of this point sequence in terms of an exponent ν via the growth of the correlation integral $C(\varepsilon)$, as follows:

$$C(\varepsilon) \sim \varepsilon^{\nu} \quad \text{as } \varepsilon \to 0, \tag{1}$$

where

$$C(\varepsilon) = \lim_{N\to\infty} \frac{1}{N^2} \sum_{1\le t,s\le N} \#\Big\{(t,s) : d(x_t, x_s) < \varepsilon\Big\}, \tag{2}$$

$\#$ denotes a set's size, and d is the distance metric. In the original GP, the sequence lies in a Euclidean space and d is the Euclidean distance. For an ergodic sequence, the correlation dimension suggests the values of other fractal dimensions such as the Haussdorf dimension (Pesin, 1993). For example, the Hénon map has $\nu = 1.21 \pm 0.01$ (Grassberger and Procaccia, 1983), which is close to its Hausdorff dimension of 1.261 ± 0.003 (Russell *et al.*, 1980). GP can be generalized to apply to a sequence in a more general smooth manifold (Pesin, 1993).

In our study, we examine natural language through this correlation dimension. Thus far, language texts have typically been considered in a Euclidean space. However, recent large language models have shown unprecedented performance in the form of an autoregressive system, which is defined in a probability space. Hence, we are motivated to measure the correlation dimension in a statistical manifold.

We consider a language dynamical system $\{x_t\}$ that develops word by word: $f : x_t \mapsto x_{t+1}$. Let V represent a vocabulary that comprises all unique words. A sequence of words, $\boldsymbol{a} = [a_1, a_2, \cdots, a_t, \cdots]$, where $a_t \in V$, is associated with a sequence of system states, $[x_1, x_2, \cdots, x_t, \cdots]$. As demonstrated in Figure 1(a) at the top, we define each state x_t as a probability distribution over the set Γ of all word sequences. x_t measures the

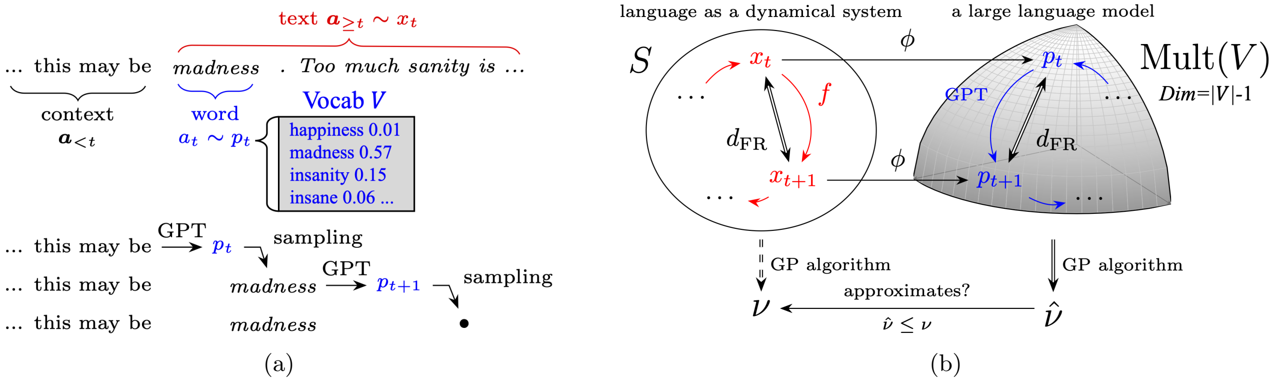


FIG. 1 Our model of language as a stochastic dynamical system. (a) The difference between the system state x_t and the next-word probability distribution p_t. (b) $\{p_t\}$ (where $p_t \in$ Mult(V)) as the image of $\{x_t\}$ (where $x_t \in S$) through the marginalization mapping ϕ in Formula (5). In this study, we use $\hat{\nu}$ to approximate ν.

probability of any text to occur as $\boldsymbol{a}_{\geq t} = [a_t, a_{t+1}, \cdots]$, following a *context* $\boldsymbol{a}_{<t} = [a_1, \cdots, a_{t-1}]$. Furthermore, we consider the next-word probability distribution p_t over the vocabulary V. x_t and p_t are formally defined as follows:

$$x_t(\boldsymbol{a}_{\geq t}) = \mathrm{P}(\boldsymbol{a}_{\geq t} \mid \boldsymbol{a}_{<t}) \qquad \forall \boldsymbol{a}_{\geq t} \in \Gamma, \tag{3}$$

$$p_t(w) = \mathrm{P}(a_t = w \mid \boldsymbol{a}_{<t}) \qquad \forall w \in V. \tag{4}$$

p_t can be represented as the image of x_t by a mapping ϕ:

$$p_t = \phi(x_t). \tag{5}$$

Here, ϕ is the marginalization across Γ and is linear with respect to a mixture of distributions, as explained in Supp. A.2.

Hence, a language state x_t is represented as a probability function instead of a point in a Euclidean space. The correlation dimension ν can be defined for the sequence $\{x_t\}$ as long as the distance metric d in Formula (2) is specified between any pair of states x_t and x_s. However, direct acquisition of $d(x_t, x_s)$ is nontrivial because $\{x_t\}$ as a language is unobservable. One new alternative path today is to represent x_t via p_t, where p_t is produced by a large language (especially a GPT-like) model (LLM). We denote the correlation dimension of the sequence $\{p_t\}$ as $\hat{\nu}$. Our approach is summarized in Figure 1(b) at the bottom. Supp. B provides a brief introduction to GPT-like LLMs.

Theoretically, $\hat{\nu} = \nu$ when the sequence of words is generated by a Markov process. We prove this in Supp. A.4. Natural language exhibits the Markov property to a certain extent, but strictly speaking, it violates the property. This phenomenon has been studied in terms of long memory (Altmann *et al.*, 2012, 2009; Li, 1989; Tanaka-Ishii and Bunde, 2016), as mentioned in the Introduction. Therefore, the $\hat{\nu}$ acquired from p_t will remain an approximation of ν. In general, $\hat{\nu} \leq \nu$ holds (Peitgen *et al.*, 1992) and $\hat{\nu}$ thus constitutes a lower bound of ν.

The distance metric d in Formula (2) is chosen as the Fisher-Rao distance, defined as the geodesic distance on a statistical manifold generated by Fisher information (Amari, 2012). When $\{p_t\}$ is presumed to follow a multinoulli distribution (over the vocabulary V), the statistical manifold is the space of all multinoulli distributions over V, denoted as Mult(V), as shown at the top right in Figure 1(b). Mult(V) has a (topological) dimension of $|V| - 1$ and is isometric to the positive orthant of a hypersphere. The Fisher-Rao distance is analytically equal to twice the Bhattacharyya angle, as follows:

$$d_{\mathrm{FR}}(p_t, p_s) = 2 \arccos \left(\sum_{w \in V} \sqrt{p_t(w) p_s(w)} \right) \qquad \text{for } t, s = 1, 2, \cdots, N. \tag{6}$$

This statistical manifold is a Riemannian manifold of constant curvature (as it constitutes a part of a hypersphere), sharing many favorable topological properties with Euclidean spaces. Particularly, the Marstrand projection theorems (Falconer, 2004; Marstrand, 1954) for Euclidean spaces, which state that linear mappings almost surely preserve a set's Hausdorff dimension, can be generalized to such Riemannian manifolds. Recently, Balogh and Iseli (2016) proved Marstrand-like theorems for sets on a 2-sphere. Because the mapping $\phi : x_t \mapsto p_t$ is linear, as mentioned before and proved in Supp. A.2, these theorems could be generalized to suggest the equality $\nu = \hat{\nu}$. This possible generalization goes beyond this letter's scope; even if it were true, Marstrand-like theorems do not guarantee a specific linear mapping (i.e., ϕ) to be dimension-preserving. Nevertheless, these theorems motivate our proposal to analyze ν via its lower bound $\hat{\nu}$.

The calculation of distances over N timesteps takes $O(|V| \cdot N^2)$ time, with a vocabulary size $|V|$ around 10^4. This computational cost can be reduced to $O(M \cdot N^2)$ through dimension reduction from $\{p_t\}$ to $\{q_t\}$, without altering the estimated correlation dimension $\hat{\nu}$, where

$M \ll |V|$ is the new, smaller dimensionality. For $t = 1, \cdots, N$, the dimension-reduction projection transforms p_t to q_t as follows:

$$q_t(m) = \sum_{w \in \Phi^{-1}(\{m\})} p_t(w), \quad \forall m = 1, \cdots, M. \tag{7}$$

Here, Φ is determined via the modulo function: $\Phi(w) = \mathrm{index}(w) \bmod M$, where $\mathrm{index}(w)$ indicates a word's index in the vocabulary. Essentially, we "randomly" group words from the extensive vocabulary V in a smaller set $\{1, \cdots, M\}$ and estimate $\hat{\nu}$ according to this condensed vocabulary. We empirically validated this method, which is rooted in Marstand's projection theorem, as detailed in Supp. C. Specifically, dimensionality reduction from approximately 50,000 to 1,000 retained the consistency of estimating $\hat{\nu}$ and achieved up to 50X faster computation.

III. RESULTS

Before showing the correlation dimension, we examine language's inherent self-similarity. Figure 2 includes a plot showing the probability p_t of encountering "," (commas) and ";" semicolons over $t = 1, 2, \cdots, N$ in an English translation of *Don Quixote* by Miguel de Cervantes from Project Gutenberg [1]. These punctuation marks, chosen for their high frequency, illustrate the role of semantic ambiguity. Each p_t represents a point in $\mathrm{Mult}(V)$, a probability vector of the next-word occurrence, estimated using GPT2-xl (Radford *et al.*, 2019). The figure maps these points, varying with input context $\boldsymbol{a}_{<t}$, and classifies them by Shannon entropy $H(p_t)$, revealing self-similarity in both low- and high-entropy regions through magnified views at different scales. Nevertheless, a thorough assessment of this self-similarity necessitates examining the high-dimensional space of $\mathrm{Mult}(V)$, beyond the limits of a two-dimensional display that cannot represent correlation dimensions above 2.

We conjecture that the trajectory has two kinds of fractals: local and global. The local fractals, potentially arising from simple word distributions across contexts akin to those in topic models like LDA (Blei *et al.*, 2003), are evident in low-entropy areas where single words predominate. In Supp. D.1, we show that even i.i.d. samples from a Dirichlet distribution (a commonly assumed prior for multinoulli distributions) can reproduce the local fractal seen in Figure 2. The local kind's occurrence could be related to the finding in Doxas *et al.* (2010) that topic models can reproduce self-similar patterns. However, the local kind is not especially concerned in this letter because it characterizes single words and hence does not reveal the nature of the original system $\{x_t\}$.

In this letter, we are mainly interested in the correlation dimension of the global phenomenon. Unlike the local kind, the global fractals represent high-entropy regions that are governed by the trajectory's global development. Hence, we consider points in the higher-entropy region, as filtered by a parameter η:

$$\max_{w \in V} p_t(w) < \eta. \tag{8}$$

Figure 3(a) shows the correlation integral from Formula (2) with respect to ε for *Don Quixote* in terms of different probability thresholds η in Formula (8). As η decreases to 0.5 (red curve), the linear region becomes visible across all scales, and the correlation dimension (given by the slope) converges to $\hat{\nu} = 6.42$. In contrast, the curve for $\eta = 1.0$ (i.e., when no timesteps are excluded) shows great deviation from the other curves, especially at smaller ε values, producing a local correlation dimension that drops below 2.0. Hence, unless mentioned otherwise, $\eta = 0.5$ in this letter. For $\eta = 0.5$, Figure 3(a) shows a long span across more than six orders of magnitude, from 10^{-1} to 10^{-8} on the vertical axis.

Figure 3(b) characterizes the effect of N, the length of the text used to estimate the correlation dimension. The longest text fragment had 150,000 words and is indicated by the red curve. Convergence is visible for all N, starting from $N = 500$. Unless mentioned otherwise, $N = 150,000$ here.

We also investigated the effect of the context length, denoted as c. Ideally, an LLM estimates the distribution p_t by using the whole text $[a_1, \cdots, a_{t-1}]$ before timestep t as the context, but in practice, a maximum context length c is often set; that is,

$$\begin{aligned} p_t^{(c)}(w) &= \mathrm{P}(a_t = w \mid a_{t-c}, a_{t-c+1}, \cdots, a_{t-1}) \\ &\approx p_t(w) \quad \forall w \in V. \end{aligned} \tag{9}$$

Unless mentioned otherwise, all results in this letter were obtained with $c = 512$.

Figure 3(c) shows the correlation dimension with values of c as small as 1 (i.e., a Markov model). For context lengths above 32, a clear linear scaling phenomenon is observed across all scales, which resembles the case of $c = 512$. As c decreases, the linear-scaling region becomes narrower and the self-similarity becomes less evident. Dependency of the correlation dimension on c is seen only for the global fractal, whereas the dimension is consistent across c values for the local fractals, as detailed in Supp. D.2.

This difference in the behavior of local and global fractals suggests a fundamental difference between these two kinds. The local fractal does not depend on c, whereas the global fractal requires large c to appear. While the local fractal may stem from mixed word-frequency distributions in topic models, as observed by Doxas *et al.* (2010) and mentioned above, the global fractal is due to long

[1] https://www.gutenberg.org/ebooks/996

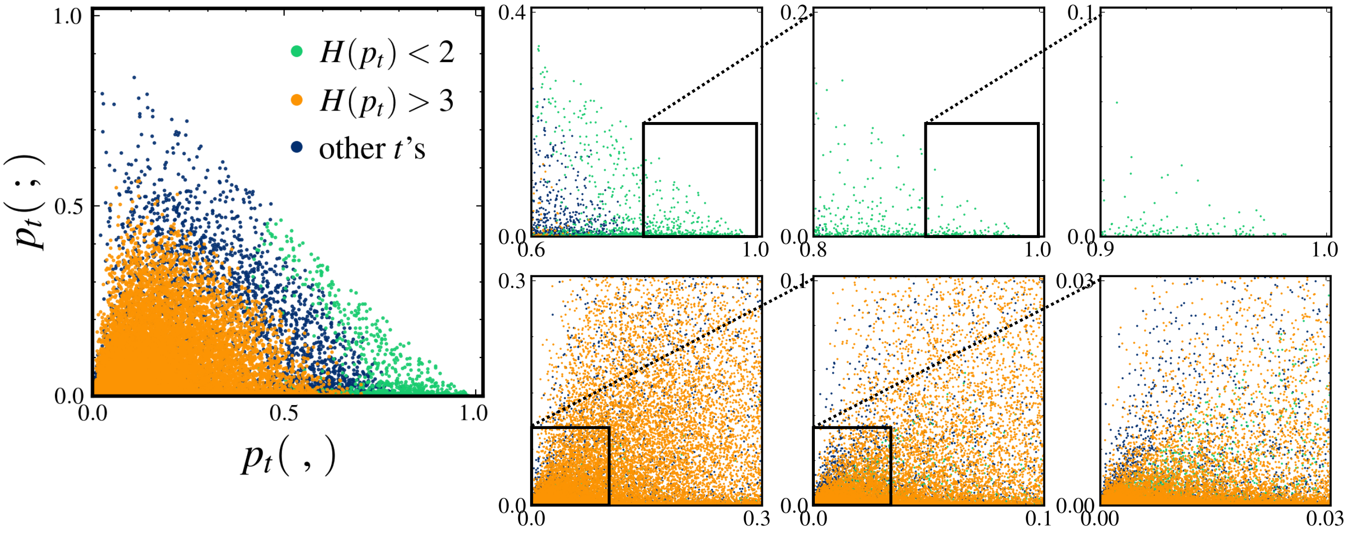


FIG. 2 Sequence of distributions p_t underlying the words in *Don Quixote*, as visualized for words "," (comma) and ";" (semicolon). Each point represents one timestep. The green points represents timesteps at which p_t(",") dominates and the Shannon entropy $H(p_t) < 2.0$, whereas the orange points correspond to high-entropy states with $H(p_t) > 3.0$. Self-similar patterns are observed in both the green and orange regions.

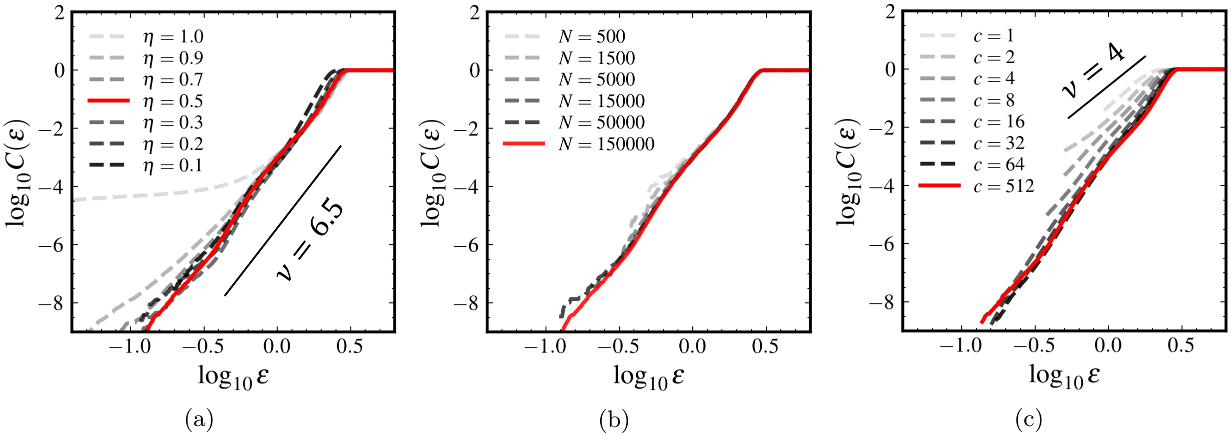
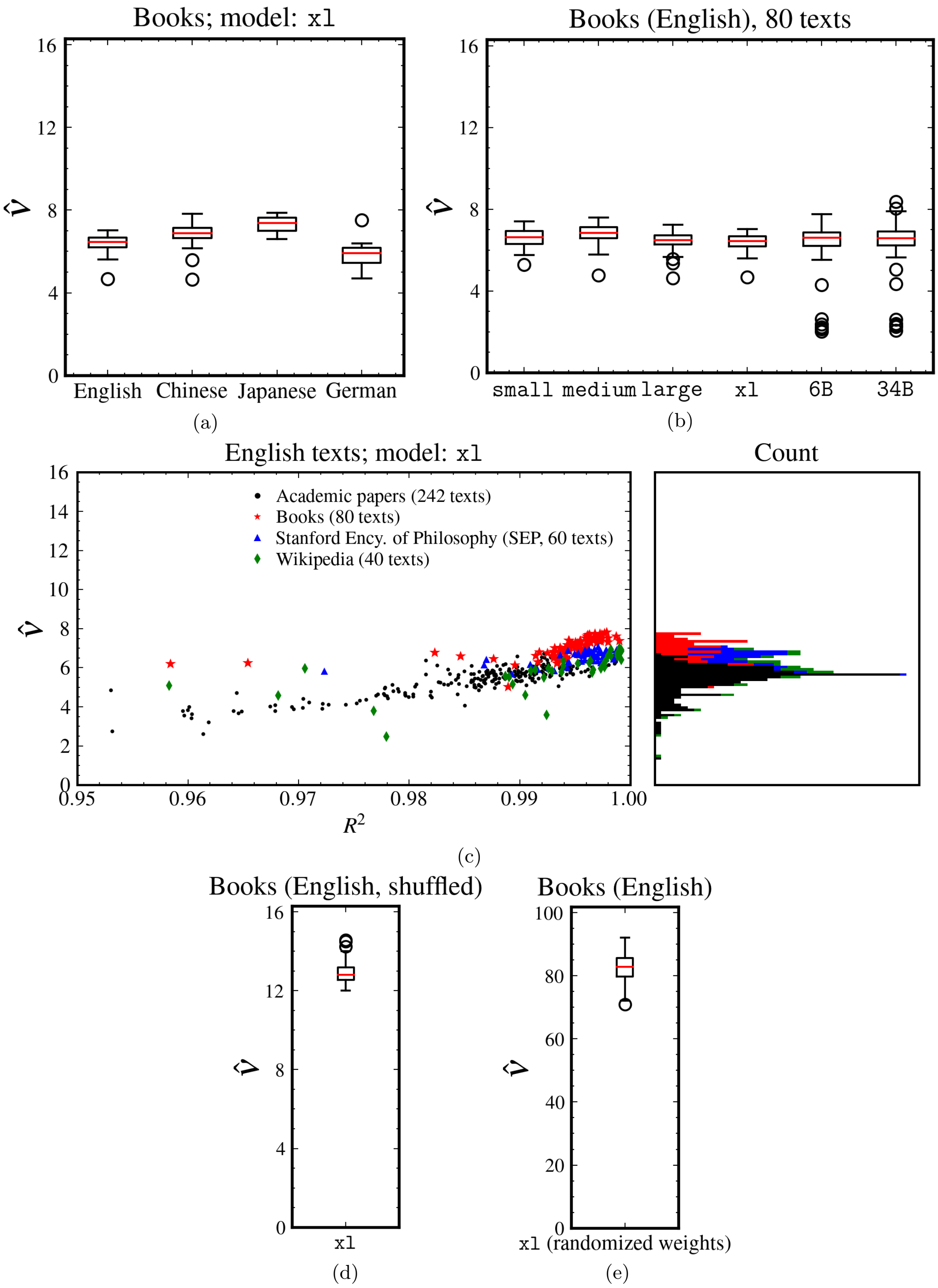


FIG. 3 Correlation integral curves as defined by Formula (2) and estimated with GPT2-`xl` with respect to (a) the maximum-probability threshold η in Formula (8), (b) the sequence length N, and (c) the context length c in Formula (9).

memory and was anticipated in the literature (Altmann *et al.*, 2012; Li, 1989; Tanaka-Ishii and Bunde, 2016). Although self-similarity and long memory have often been studied separately and were even conjectured as different aspects of a scale-invariant process (Abry *et al.*, 2003), they show interesting coordination for natural language. More results on a larger dataset are provided in Supp. F.2.

To further investigate the properties of natural language, we conducted a larger-scale analysis of long texts, which were divided into two groups: books in multiple languages and English articles in multiple genres, as detailed in Supp. E. The first group included 144 single-author books from Project Gutenberg and Aozora Bunko, comprising 80 in English, 32 in Chinese, 16 in German, and 16 in Japanese. The second group included 342 long English texts from different sources. We obtained all the results in this large-scale analysis by applying the dimension-reduction method given in Formula (7).

Figures 4 (a) and (b) show the large-scale results on the books for the correlation dimension $\hat{\nu}$ with respect to (a) different languages and (b) various model sizes. The former results (a) were produced using the GPT2 model of size `xl` (denoting "extra-large"), with $\approx 10^9$ parameters. For the latter results (b), we tested models of different sizes from `small` ($\approx 10^6$ parameters) to `34B` (3.4×10^{10}). For the sizes up to `xl`, we used the GPT2 model; for `6B` and `34B`, we used the Yi model (Yi, 2024), which offers the SOTA capability in English among all



(a)

(b)

(c)

(d)

(e)

FIG. 4 Correlation dimensions of (a) all books grouped by language, as estimated using GPT2-`xl`; (b) English books as estimated using GPT with different model sizes (GPT2 from `small` to `xl` and the Yi model for `6b` and `34b`); (c) English texts from various sources with the R^2 scores (horizontal axis) of their linear fits to the correlation integral curves; (d) shuffled English books evaluated with GPT2-`xl`; and (e) English books evaluated with weight-randomized GPT2-`xl`.

publicly available LLMs. For all tested model sizes, the average correlation dimension remains constant. Outliers occur more frequently for the two Yi models (`6B` and `34B`), which was possibly due to those models' use of a lower numerical precision (16-bit floating-point numbers).

Hence, for all languages, an average correlation dimension of around $\hat{\nu} = 6.5$ is observed: 6.39 ± 0.40 for English, 6.81 ± 0.58 for Chinese, 7.30 ± 0.41 for Japanese, and 5.84 ± 0.70 for German (± indicates the standard deviation). These results suggest the possible existence of a common dimension for natural language, with a lower bound of 6.5 under our settings.

Figure 4(c) shows the correlation dimension (vertical axis) for English texts in four genres: books, academic papers (Kershaw and Koeling, 2020), the Stanford Encyclopedia of Philosophy (SEP) [2], and Wikipedia webpages. For each text, the horizontal axis indicates the coefficient of determination, R^2, for the correlation integral curve's linear fit. A larger R^2 value (maximum 1) implies more significant self-similarity in a text. The right side of (c) shows the distribution of the dimension values grouped by genre.

As seen in the figure, most texts have a correlation dimension around 6, especially those estimated with high R^2 scores. The SEP texts (blue) have the most concentrated range of dimensions, at 6.57 ± 0.32 with $R^2 > 0.99$ for over 90% of the texts. In contrast, the academic papers (black) show the most scattered distribution of the correlation dimension. This is deemed natural, as the SEP texts have the highest quality, whereas the academic papers include irregular notations such as chemical and mathematical formulas, which obscure a text's self-similarity.

The universal correlation dimension value, $\nu \approx 6.5$, can be understood through the lens of the "information dimension" (Farmer, 1982), which coincides with ν under ergodic conditions (Pesin, 1993). The information dimension reflects how information, or the log count of unique contexts, scales with the statistical manifold's resolution. Contexts are deemed the same if their p_t values are indistinguishably close within a certain threshold. Essentially, doubling the resolution would reveal about $2^{6.5} \approx 90$ times more distinct contexts that were previously considered identical. Therefore, ν quantifies the average "redundancy" in the diversity of texts conveying similar messages.

We also compared several theoretical random processes. As analyzed using a GPT2-`xl` model and shown in Figure 4(d), shuffled word sequences exhibited an average correlation dimension of 13.0, indicating inherent self-similarity despite the shuffling. As seen in Figure 4(e), randomization of the GPT2-`xl` model's weights significantly increased the correlation dimensions to an average of 80. This result suggests purely random outputs, unlike text shuffling, which retains some linguistic structures, like a bag-of-words approach.

Analyses of additional random processes, as detailed in Supp. G, showed that a uniform white-noise process on the statistical manifold S yielded a correlation dimension over 100. Symmetric Dirichlet distributions in high-entropy regions consistently produced dimensions above 10. Conversely, Barabási-Albert (BA) networks (Barabási and Albert, 1999), which are special cases of a Simon process, demonstrated a correlation dimension of 2.00 ± 0.003, and a fractal variant (Rak and Rak, 2020) produced $2 \sim 3.5$. In terms of complexity via the correlation dimension, this places natural language above BA networks but below white noise.

In Supp. H, we further investigate the relationship between the statistical manifold and conventional Euclidean spaces with respect to the correlation dimension. For BA models, the dimension remains the same whether measured in a Euclidean space or the manifold, thus emphasizing the comparability. However, language data reveals a different story: Euclidean metrics yield compromised linearity in comparison to Fisher-Rao metrics, thus underscoring that the Fisher-Rao distance more accurately captures language's inherent self-similarity.

Recently, LLMs have also been developed for processing data beyond natural language, and one successful example is for acoustic waves compressed into discrete sequences (Copet *et al.*, 2023). To demonstrate the applicability of our analysis, we used the `GTZAN` dataset (Tzanetakis and Cook, 2002), which comprises 1000 recorded music pieces categorized in 10 genres. Briefly, we observed clear self-similarity in the compressed music data. The correlation dimension was found to depend on the genre: classical music showed the smallest dimension at 5.44 ± 1.13, much smaller than the dimensions for metal music at 7.27 ± 0.96 and rock music at 7.42 ± 0.87. None of the music genres showed a correlation dimension as large as that of white noise, as mentioned previously, even though the analysis was based on recorded data. The details of this analysis are given in Supp. I.

In closing, we recognize this study's limitation of viewing text as a dynamical system akin to the GPT model, which overlooks the potential of representing words as leaf nodes in a syntactic tree, as suggested by generative and context-free grammars (CFGs) (Chomsky, 2014). Although promising, that complex linguistic framework exceeds our current scope, and we expect to explore it in the future.

[2] `https://plato.stanford.edu/`

ACKNOWLEDGMENTS

This work was supported by JST CREST Grant Number JPMJCR2114 and JSPS KAKENHI Grant Number JP20K20492.

REFERENCES

Abry, Patrice, Patrick Flandrin, Murad S Taqqu, *et al.* (2003), "Self-similarity and long-range dependence through the wavelet lens," Theory and applications of long-range dependence **1**, 527–556.

Altmann, Edouard G, Giampaolo Cristadoro, and Mirko D. Esposti (2012), "On the origin of long-range correlations in texts," Proceedings of the National Academy of Sciences **109** (29), 11582–11587.

Altmann, Eduardo G, Janet B. Pierrehumbert, and Adilson E. Motter (2009), "Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words," PLoS One **4** (e7678).

Amari, Shun-ichi (2012), *Differential-geometrical methods in statistics*, Vol. 28 (Springer Science & Business Media).

Ausloos, Marcel (2012), "Measuring complexity with multifractals in texts. translation effects," Chaos, Solitons & Fractals **45** (11), 1349–1357.

Balogh, Zoltán, and Annina Iseli (2016), "Dimensions of projections of sets on riemannian surfaces of constant curvature," Proceedings of the American Mathematical Society **144** (7), 2939–2951.

Barabási, Albert-László, and Réka Albert (1999), "Emergence of scaling in random networks," science **286** (5439), 509–512.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003), "Latent dirichlet allocation," Journal of machine Learning research **3** (Jan), 993–1022.

Brown, Tom, Benjamin Mann, Nick Ryder, *et al.* (2020), "Language models are few-shot learners," Advances in neural information processing systems **33**, 1877–1901.

Chomsky, Noam (2014), *Aspects of the Theory of Syntax*, 11 (MIT press).

Copet, Jade, Felix Kreuk, Itai Gat, *et al.* (2023), "Simple and controllable music generation," in *Thirty-seventh Conference on Neural Information Processing Systems*.

Doxas, Isidoros, Simon Dennis, and William L Oliver (2010), "The dimensionality of discourse," Proceedings of the National Academy of Sciences **107** (11), 4866–4871.

Falconer, Kenneth (2004), *Fractal geometry: mathematical foundations and applications* (John Wiley & Sons).

Farmer, J Doyne (1982), "Information dimension and the probabilistic structure of chaos," Zeitschrift für Naturforschung A **37** (11), 1304–1326.

Grassberger, Peter, and Itamar Procaccia (1983), "Characterization of strange attractors," Physical review letters **50** (5), 346.

Kershaw, Daniel James, and R. Koeling (2020), "Elsevier oa cc-by corpus," ArXiv **abs/2008.00774**.

Kobayashi, Tatsuru, and Kumiko Tanaka-Ishii (2018), "Taylor's law for human linguistic sequences," Proceedings of the 56th Annual Meeting of the Association for Computational Lingusitics , 1138–1148.

Lacasa, Lucas, and Jesús Gómez-Gardenes (2013), "Correlation dimension of complex networks," Physical review letters **110** (16), 168703.

Li, Wentian (1989), "Mutual information functions of natural language texts," (Citeseer).

Marstrand, John M (1954), "Some fundamental geometrical properties of plane sets of fractional dimensions," Proceedings of the London Mathematical Society **3** (1), 257–302.

OpenAI, (2023), "Gpt-4 technical report," arXiv:2303.08774 [cs.CL].

Osborne, A Ro, and A Provenzale (1989), "Finite correlation dimension for stochastic systems with power-law spectra," Physica D: Nonlinear Phenomena **35** (3), 357–381.

Peitgen, Heinz-Otto, Hartmut Jürgens, Dietmar Saupe, and Mitchell J Feigenbaum (1992), *Chaos and fractals: new frontiers of science*, Vol. 7 (Springer).

Pesin, Ya B (1993), "On rigorous mathematical definitions of correlation dimension and generalized spectrum for dimensions," Journal of statistical physics **71**, 529–547.

Radford, Alec, Jeffrey Wu, Rewon Child, *et al.* (2019), "Language models are unsupervised multitask learners," OpenAI blog **1** (8), 9.

Rak, Rafał, and Ewa Rak (2020), "The fractional preferential attachment scale-free network model," Entropy **22** (5), 509.

Rao, C Radhakrishna (1992), "Information and the accuracy attainable in the estimation of statistical parameters," in *Breakthroughs in Statistics: Foundations and basic theory* (Springer) pp. 235–247.

Russell, David A, James D Hanson, and Edward Ott (1980), "Dimension of strange attractors," Physical Review Letters **45** (14), 1175.

Simon, Herbert A (1955), "On a class of skew distribution functions," Biometrika **42** (3/4), 425–440.

Tanaka-Ishii, Kumiko, and Armin Bunde (2016), "Long-range memory in literary texts: On the universal clustering of the rare words," PLoS One **11** (11), e0164658.

Tanaka-Ishii, Kumiko, and Tatsuru Kobayashi (2018), "Taylor's law for linguistic sequences and random walk models," Journal of Physics Communications **2** (11), 089401.

Touvron, Hugo, Louis Martin, Kevin Stone, *et al.* (2023), "Llama 2: Open foundation and fine-tuned chat models," arXiv preprint arXiv:2307.09288.

Tzanetakis, George, and Perry Cook (2002), "Musical genre classification of audio signals," IEEE Transactions on speech and audio processing **10** (5), 293–302.

Yi, (2024), "The yi model," `https://huggingface.co/01-ai/Yi-34B`, visited in January 2024.

Appendix A: Properties of the Mapping $\phi : x_t \mapsto p_t$

1. Formulation

In the main text, we consider two sequences of probability distributions, i.e., $\{x_t\}$ and $\{p_t\}$, which are related by the mapping $\phi : x_t \mapsto p_t$ in 5. We explain the formulation of ϕ here.

Recall that x_t denotes the language dynamical system's state at timestep t and is defined as a distribution over the set Γ of all sequences of words, as in 3. Then, p_t is defined over the vocabulary V and characterizes the probability of a word w to occur as the next word following a given context. The probability $p_t(w)$ for any $w \in V$ is defined as the probability that an arbitrary closed text has w as its first word, thus giving the following formulation of ϕ:

$$p_t(w) = \phi(x_t)(w) := \sum_{\substack{\boldsymbol{a}_{\geq t} \in \Gamma \\ a_t = w}} x_t(\boldsymbol{a}_{\geq t}) \qquad \forall w \in V. \tag{A1}$$

x_t and p_t are defined over different sets but are essentially consistent with respect to the same probability measure μ on a probability space $(\Gamma, \mathcal{F}, \mu)$, where $\mathcal{F} = \{\Lambda : \Lambda \subset \Gamma\}$ denotes the power set of Γ. Here, $\mu : \mathcal{F} \to [0,1]$ is defined as follows:

$$\mu(\Lambda) = \sum_{\boldsymbol{a}_{\geq t} \in \Lambda} x_t(\boldsymbol{a}_{\geq t}), \qquad \forall \Lambda \subset \Gamma. \tag{A2}$$

Hence, x_t are p_t are both consistent with respect to μ:

$$x_t(\boldsymbol{a}_{\geq t}) = \mu(\{\boldsymbol{a}_{\geq t}\}) \qquad \forall \boldsymbol{a}_{\geq t} \in \Gamma, \tag{A3}$$

$$p_t(w) = \mu\left(\coprod_{\substack{\boldsymbol{a}_{\geq t} \in \Gamma \\ a_t = w}} \{\boldsymbol{a}_{\geq t}\} \right) \qquad \forall w \in V. \tag{A4}$$

2. Linearity

The mapping ϕ is linear with respect to the mixture of probability distributions within $\{x_t\}$. That is, for any two distributions x_t, x_s, and any mixture weight $\alpha \in [0,1]$, ϕ satisfies the following:

$$\phi\big(\alpha x_t + (1-\alpha) x_s\big) = \alpha \phi(x_t) + (1-\alpha)\phi(x_s). \tag{A5}$$

This equality can be obtained directly from the definition of ϕ in Formula (A1). The left side of Formula (A1) is calculated as follows for any $w \in V$:

$$\begin{aligned}
\phi\big(\alpha x_t + (1-\alpha) x_s\big)(w) &= \sum_{\substack{\boldsymbol{a}_{\geq t} \in \Gamma \\ a_t = w}} \big(\alpha x_t + (1-\alpha) x_s\big)(\boldsymbol{a}_{\geq t}) && \text{(A6)} \\
&= \sum_{\substack{\boldsymbol{a}_{\geq t} \in \Gamma \\ a_t = w}} \Big(\alpha x_t(\boldsymbol{a}_{\geq t}) + (1-\alpha) x_s(\boldsymbol{a}_{\geq t})\Big) && \text{(A7)} \\
&= \alpha \sum_{\substack{\boldsymbol{a}_{\geq t} \in \Gamma \\ a_t = w}} x_t(\boldsymbol{a}_{\geq t}) + (1-\alpha) \sum_{\substack{\boldsymbol{a}_{\geq s} \in \Gamma \\ a_s = w}} x_s(\boldsymbol{a}_{\geq s}) && \text{(A8)} \\
&= \alpha p_t(w) + (1-\alpha) p_s(w). && \text{(A9)}
\end{aligned}$$

3. Distance Distortion Rate

In this work, we are especially interested in how the mapping $\phi : x_t \mapsto p_t$ would distort the Fisher-Rao distance between any two states x_t and x_s. The distortion is measured by the following rate:

$$r(x_t, x_s) \equiv d_{\mathrm{FR}}(x_t, x_s)/d_{\mathrm{FR}}(p_t, p_s), \tag{A10}$$

where $p_t = \phi(x_t)$ and $p_s = \phi(x_s)$ as defined in Formula (A1).

In this section, we show that $r(x_t, x_s) \geq 1$ in general, i.e., it has a lower bound of 1.

Lemma 1. *(Lower Bound of the Distance Distortion Rate) The distortion rate $r(x_t, x_s)$ is no smaller than 1 for any x_t and x_s.*

Proof. For any $\boldsymbol{a}_{\geq t} \in \Gamma$, x_t (and x_s) can be decomposed as follows:

$$\begin{aligned} x_t(\boldsymbol{a}_{\geq t}) &\equiv \mathrm{P}(\boldsymbol{a}_{\geq t} \mid \boldsymbol{a}_{<t}) && \text{(A11)} \\ &= \mathrm{P}(a_t \mid \boldsymbol{a}_{<t}) \cdot \mathrm{P}(\boldsymbol{a}_{\geq t+1} \mid \boldsymbol{a}_{<t}, a_t) && \text{(A12)} \\ &= p_t(a_t) \cdot \mathrm{P}(\boldsymbol{a}_{\geq t+1} \mid \boldsymbol{a}_{<t}, a_t). && \text{(A13)} \end{aligned}$$

For simplicity of notation, let $x_{t+1}(\boldsymbol{a}_{\geq t+1} \mid a_t) \equiv \mathrm{P}(\boldsymbol{a}_{\geq t+1} \mid \boldsymbol{a}_{<t}, a_t)$.

Hence, the distance $d_{\mathrm{FR}}(x_t, x_s)$ can be decomposed as follows:

$$\begin{aligned} d_{\mathrm{FR}}(x_t, x_s) &= 2 \arccos \sum_{\boldsymbol{b} \in \Gamma} \sqrt{x_t(\boldsymbol{b}) \cdot x_s(\boldsymbol{b})} && \text{(A14)} \\ &= 2 \arccos \sum_{\boldsymbol{b} \in \Gamma} \sqrt{p_t(b_1) x_{t+1}(\boldsymbol{b}_{\geq 2} \mid b_1) \cdot p_s(b_1) x_{s+1}(\boldsymbol{b}_{\geq 2} \mid b_1)} && \text{(A15)} \\ &= 2 \arccos \sum_{b_1 \in V} \sqrt{p_t(b_1) p_s(b_1)} \underbrace{\sum_{\boldsymbol{b}_{\geq 2} \in \Gamma} \sqrt{x_{t+1}(\boldsymbol{b}_{\geq 2} \mid b_1) x_{s+1}(\boldsymbol{b}_{\geq 2} \mid b_1)}}_{\cos\left[\frac{1}{2} d_{\mathrm{FR}}\left(x_{t+1}(\cdot|b_1), x_{s+1}(\cdot|b_2)\right)\right] \quad \leq 1} && \text{(A16)} \\ &\geq 2 \arccos \sum_{b_1 \in V} \sqrt{p_t(b_1) p_s(b_1)} && \text{(A17)} \\ &= d_{\mathrm{FR}}(p_t, p_s), && \text{(A18)} \end{aligned}$$

where $\boldsymbol{b} \in \Gamma$ denotes any closed text, and b_1 and $\boldsymbol{b}_{\geq 2} = [b_2, b_3 \cdots]$ represent the first word and the remainder of the text, respectively. This implies $r(x_t, x_s) \geq 1$. □

4. Dimension Preservation for Markov Processes

In this section, we analyze whether ϕ preserves the correlation dimension of a language system $\{x_t\}$ when the system follows certain Markov conditions. In other words, we examine whether $\nu = \hat{\nu}$ holds, where ν and $\hat{\nu}$ are the respective correlation dimensions of $\{x_t\}$ and $\{p_t\} = \{\phi(x_t)\}$.

We consider two kinds of Markov conditions, and we show that under either kind of condition, the distortion rate $r(x_t, x_s)$ defined in Formula (A10) is bounded above. That is, there exists a constant C such that

$$r(x_t, x_s) < C \tag{A19}$$

for any x_t and x_s. Because $r(x_t, x_s) \geq 1$ holds in general (see Lemma 1 in Section A.3), the boundedness of $r(x_t, x_s)$ in Formula (A19) implies the bi-Lipschitz characteristic of ϕ and thus the equality $\nu = \hat{\nu}$.

The first kind of Markov condition specifies the case when $\boldsymbol{a}_{\geq t}$ and $\boldsymbol{a}_{\geq s}$ are generated by the same Markov process with different initial states. For the second kind, $\boldsymbol{a}_{\geq t}$ and $\boldsymbol{a}_{\geq s}$ follow two different Markov processes, and we consider the case when their initial states get infinitesimally close to each other. The two kinds of conditions are examined in Sections A.4.a and A.4.b, respectively.

a. When $\boldsymbol{a}_{\geq t}$ and $\boldsymbol{a}_{\geq s}$ Follow the Same Markov Process

A Markov process can be represented by its transition matrix A. A text $\boldsymbol{a}_{\geq t}$ (or $\boldsymbol{a}_{\geq s}$) is said to follow a Markov process if $\forall \tau \geq t$ (or $\forall \tau \geq s$),

$$p_\tau \equiv \mathrm{P}(a_\tau \mid \boldsymbol{a}_{<\tau}) = \mathrm{P}(a_\tau \mid a_{\tau-1}) =: A_{a_{\tau-1}, a_\tau}, \tag{A20}$$

where $A_{a_{\tau-1}, a_\tau}$ represents the transition probability from word $a_{\tau-1}$ to a_τ, i.e., the probability that a_τ occurs immediately after $a_{\tau-1}$.

Theorem 2. *Consider two Markov processes that are defined over a vocabulary V and have the same transition matrix A but different initial states p_t and p_s. Then, the distance distortion rate $r(x_t, x_s) = 1$.*

Proof.

$$\cos\frac{d_{\mathrm{FR}}(x_t, x_s)}{2} = \sum_{\boldsymbol{b}\in\Gamma}\sqrt{x_t(\boldsymbol{b})x_s(\boldsymbol{b})} \tag{A21}$$

$$= \sum_{b_1\in V}\sqrt{p_t(b_1)p_s(b_1)}\sum_{\boldsymbol{b}_{\geq 2}\in\Gamma}\sqrt{x_{t+1}(\boldsymbol{b}_{\geq 2}\mid b_1)\ x_{s+1}(\boldsymbol{b}_{\geq 2}\mid b_1)}, \tag{A22}$$

where $x_{t+1}(\boldsymbol{b}_{\geq 2}\mid b_1) \equiv \mathrm{P}(\boldsymbol{a}_{\geq t+1} = \boldsymbol{b}_{\geq 2}\mid a_t = b_1, a_{<t} = \boldsymbol{a}_{<t})$, and $x_{s+1}(\boldsymbol{b}_{\geq 2}\mid b_1)$ is defined similarly. Owing to the Markov property of $\boldsymbol{a}$, $x_{t+1}(\boldsymbol{b}_{\geq 2}\mid b_1)$ is decomposed as follows:

$$x_{t+1}(\boldsymbol{b}_{\geq 2}\mid b_1) = \prod_{\tau=2}^{|\boldsymbol{b}|}A_{b_\tau, b_{\tau+1}} = x_{s+1}(\boldsymbol{b}_{\geq 2}\mid b_1). \tag{A23}$$

Hence, the second term of Formula (A22) simplifies to

$$\sum_{\boldsymbol{b}_{\geq 2}\in\Gamma}\sqrt{x_{t+1}(\boldsymbol{b}_{\geq 2}\mid b_1)\ x_{s+1}(\boldsymbol{b}_{\geq 2}\mid b_1)} = \sum_{\boldsymbol{b}_{\geq 2}\in\Gamma}x_{t+1}(\boldsymbol{b}_{\geq 2}\mid b_1) = 1. \tag{A24}$$

Therefore,

$$\cos\frac{d_{\mathrm{FR}}(x_t, x_s)}{2} = \sum_{b_1\in V}\sqrt{p_t(b_1)p_s(b_1)} = \cos\frac{d_{\mathrm{FR}}(p_t, p_s)}{2}, \tag{A25}$$

which implies $d_{\mathrm{FR}}(x_t, x_s) = d_{\mathrm{FR}}(p_t, p_s)$ and thus $r(x_t, x_s) = 1$ for any x_t and x_s. □

b. When $\boldsymbol{a}_{\geq t}$ and $\boldsymbol{a}_{\geq s}$ Follow Different Markov Processes

For the second kind of Markov condition, we assume the two word sequences $\boldsymbol{a}_{\geq t}$ and $\boldsymbol{a}_{\geq s}$ to be generated by two separate Markov processes. The transition matrices are denoted as A and B, respectively.

The difference between A and B is quantified by a row-wise metric Δ_w ($w \in V$) that is defined as the Fisher-Rao distance between the transition probabilities from state w to all states:

$$\Delta_w = 2\arccos\sum_{j\in V}\sqrt{A_{w,j}B_{w,j}}. \tag{A26}$$

The average difference across the columns j is reflected as the distance between the next-word probability distributions. For simplicity, we restrict our discussion to Markov processes that satisfy the following:

$$\Delta_w \leq d_{\mathrm{FR}}(p_t, p_s) \qquad w \in V. \tag{A27}$$

This is analogous to bounding the matrix norm $\|A - B\|$ by $d_{\mathrm{FR}}(p_t, p_s)$.

Here, the calculation of $d_{\mathrm{FR}}(x_t, x_s)$ is more difficult than in the case above, and we must consider texts of different lengths. In other words, the "end" of a text must be clearly defined. Γ was abstractly defined as "the set of all texts," but this definition must be made rigorous here to incorporate the text lengths.

Hence, we restrict Γ to the set of all *closed* texts, i.e., those end with a special closing token denoted as `<END>`. Note that the probabilities of these closed texts also determine the probability of any unclosed text (i.e., one without `<END>`) through aggregation of all closed texts that have the unclosed text as their prefix. Each word in the vocabulary besides `<END>` can be seen as an unclosed text of length one.

For the two Markov processes, `<END>` can be understood as an "absorbing state," because the transition probability from `<END>` to any other state is zero: a closed text will not return to being unclosed. In other words, $A_{\texttt{<END>},\texttt{<END>}} = 1$, and $A_{\texttt{<END>},w} = 0$ ($\forall w \in V\backslash$`<END>`). We consider the simplest case in which the transition probabilities to `<END>` are equal, with a value denoted as ρ:

$$A_{w,\texttt{<END>}} = B_{w,\texttt{<END>}} = \rho \qquad \forall w \in V, \tag{A28}$$

where $0 < \rho < 1$ is a constant.

Theorem 3. *For a pair of Markov processes A and B defined over V with an absorbing state* `<END>`*, if the two conditions in Formulas (A27) and (A28) are met, then*

$$\lim_{d_{FR}(p_t,p_s)\to 0} r(x_t,x_s) \leq \rho^{-1/2}. \tag{A29}$$

Proof. Recall that the Fisher-Rao distances between x_t and x_s and between p_t and p_s are defined as follows:

$$d_{\mathrm{FR}}(x_t,x_s) = 2\arccos \sum_{a\in\Gamma} \sqrt{x_t(\boldsymbol{a})x_s(\boldsymbol{a})}, \tag{A30}$$

$$d_{\mathrm{FR}}(p_t,p_s) = 2\arccos \sum_{w\in V} \sqrt{p_t(w)p_s(w)}. \tag{A31}$$

Rearranging the sequences within Γ by considering different sequence lengths n from 1 to infinity, we have the following:

$$d_{\mathrm{FR}}(x_t,x_s) = 2\arccos \sum_{\text{closed }\boldsymbol{a}} \sqrt{x_t(\boldsymbol{a})x_s(\boldsymbol{a})} = 2\arccos\sum_{n=1}^{\infty} H_n, \tag{A32}$$

where

$$H_n = \sum_{\substack{\text{unclosed }\boldsymbol{a}\\ |\boldsymbol{a}|=n-1}} \sqrt{x_t([\boldsymbol{a},\texttt{<END>}])\cdot x_s([\boldsymbol{a},\texttt{<END>}])}. \tag{A33}$$

Here, $|\boldsymbol{a}|$ notes the length of the sequence $\boldsymbol{a}$; $[\boldsymbol{a},\texttt{<END>}]$ represents a closed text formed by concatenating the unclosed text $\boldsymbol{a}$ with `<END>`. In particular, $H_1 = \rho$.

By considering the last word ∂a of any sequence $\boldsymbol{a}$, we can see that

$$\begin{aligned}
H_{n+1} = &\sum_{\substack{\text{unclosed }\boldsymbol{a}\\ |\boldsymbol{a}|=n-1}} \sqrt{\frac{x_t([\boldsymbol{a},\texttt{<END>}])}{A_{\partial a,\texttt{<END>}}}\cdot\sum_{w\in V\setminus\texttt{<END>}} A_{\partial a,w}A_{w,\texttt{<END>}}} && \text{(A34)}\\
&\cdot\sqrt{\frac{x_s([\boldsymbol{a},\texttt{<END>}])}{B_{\partial a,\texttt{<END>}}}\cdot\sum_{w\in V\setminus\texttt{<END>}} B_{\partial a,w}B_{w,\texttt{<END>}}} && \text{(A35)}\\
= &\sum_{\substack{\text{unclosed } a\\ |a|=n-1\\ a_1=i}} \left\{\sqrt{x_t([\boldsymbol{a},\texttt{<END>}])\cdot x_s([\boldsymbol{a},\texttt{<END>}])}\cdot\sum_{w\in V\setminus\texttt{<END>}}\sqrt{A_{\partial a,w}B_{\partial a,w}}\right\} && \text{(A36)}\\
= &\sum_{\substack{\text{unclosed } a\\ |a|=n-1\\ a_1=i}} \left\{\sqrt{x_t([\boldsymbol{a},\texttt{<END>}])\cdot x_t([\boldsymbol{a},\texttt{<END>}])}\cdot\left(\cos\frac{\Delta_{\partial a}}{2}-\rho\right)\right\} && \text{(A37)}\\
\geq &\sum_{\substack{\text{unclosed } a\\ |a|=n-1\\ a_1=i}} \left\{\sqrt{x_t([\boldsymbol{a},\texttt{<END>}])\cdot x_t([\boldsymbol{a},\texttt{<END>}])}\cdot\left(\cos\frac{d_{\mathrm{FR}}(p_t,p_s)}{2}-\rho\right)\right\} && \text{(A38)}\\
= &\left(\cos\frac{d_{\mathrm{FR}}(p_t,p_s)}{2}-\rho\right)H_n, && \text{(A39)}
\end{aligned}$$

where the inequality is due to the condition in Formula (A27).

Next, by combining Formulas (A39) and (A32) and applying the formula for the sum of a geometric series, we have

$$d_{\mathrm{FR}}(x_t,x_s) \leq 2\arccos\frac{\rho}{1+\rho-\cos\frac{d_{\mathrm{FR}}(p_t,p_s)}{2}} =: F(d_{\mathrm{FR}}(p_t,p_s)). \tag{A40}$$

When $d_{\mathrm{FR}}(p_t,p_s)\to 0$, it follows that $F(d_{\mathrm{FR}}(p_t,p_s))\to 0$ as well. Furthermore,

$$\lim_{d_{\mathrm{FR}}(p_t,p_s)\to 0} r(x_t,x_s) \leq \lim_{d_{\mathrm{FR}}(p_t,p_s)\to 0}\frac{F(d_{\mathrm{FR}}(p_t,p_s))}{d_{\mathrm{FR}}(p_t,p_s)} = \rho^{-1/2}, \tag{A41}$$

which implies Theorem 3. □

Furthermore, as all i.i.d. processes meet the two conditions in Formulas (A27) and (A28), we immediately obtain the following corollary for i.i.d. processes.

Corollary 4. *Consider two i.i.d. processes that are defined over a vocabulary V and represented by probability vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. The first entries of $\boldsymbol{u}$ and $\boldsymbol{v}$, denoted respectively as u_1 and v_1, specify the occurrence probability of `<END>` at any timestep. Then, it follows that*

$$\lim_{d_{FR}(p_t,p_s)\to 0} r(x_t, x_s) \leq u_1^{-1/2}. \tag{A42}$$

Appendix B: GPT-Like Large-Scale Language Models

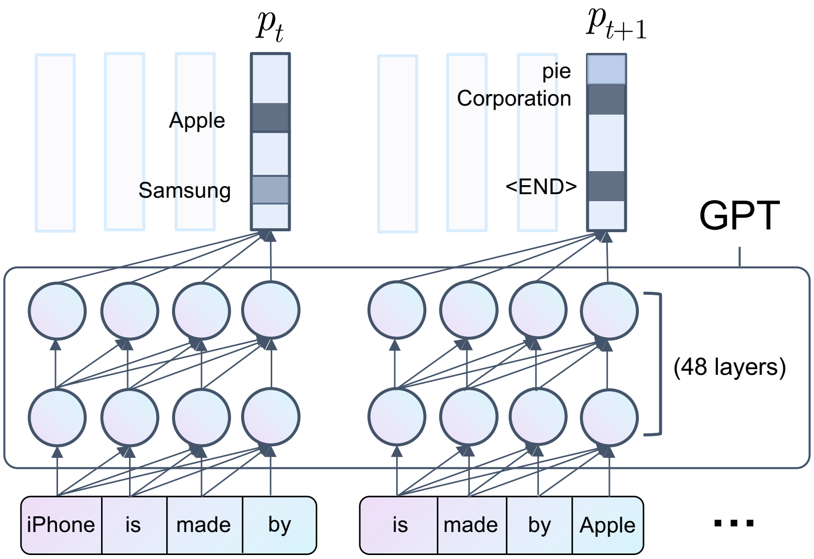


FIG. 5 Predicting the probability distribution p_t over a vocabulary with the GPT-`xl` model, which has 48 layers.

Generative pretrained transformers (GPTs) (Brown *et al.*, 2020; Radford *et al.*, 2019) comprise a class of statistical language models that are implemented with neural networks. Large-scale GPT models, including the representative *ChatGPT* by OpenAI, have shown quasi-human-level performance in language understanding and question answering.

Figure 5 illustrates how p_t was estimated in this work by using a GPT model, such as one with 48 neural-network layers (i.e., GPT2-`xl`). An input sequence of words is converted to vectors and then processed by multiple neural-network layers. To guarantee GPT's "autoregressive" nature, at every layer, the words visible during processing at timestep t are limited to those previous to t.

A GPT model can process a variable-length sequence of words, subject to a maximum length c as given in 9 in the main text. Figure 5 shows the case of $c = 4$, in which a context of four words is used for predicting p_t at any timestep t. For the work described in the main text, we used $c = 512$ unless specified otherwise.

LLMs are adept at estimating p_t, as they are trained with the objective of minimizing the statistical discrepancy between p_t for the actual data and the model's estimate. More advanced and larger language models enhance the precision of p_t estimation, bringing the measured correlation dimension closer to its true, universal value.

We used pretrained GPT-like models that are publicly available at `https://huggingface.co/models`. For the model size referred to as `xl` in the main text, we used `gpt2-xl` for English, `nlp-waseda/gpt2-xl-japanese` for Japanese, and `malteos/gpt2-xl-wechsel-german` for German. The tags `xl`, `large`, `medium`, and `small` correspond to GPT2 with 1.5 billion, 762 million, 345 million, and 117 million parameters, respectively.

For Chinese, we did not find a model that was consistent with the usual `xl` specification. However, we found a larger model for Chinese at `https://huggingface.co/IDEA-CCNL/Wenzhong2.0-GPT2-3.5B-chinese`, with around twice as many parameters as the usual `xl` specification. We also refer to this model as `xl`, as used in 4.

For even larger models, we considered the "Yi" family of GPT models provided at `https://huggingface.co/01-ai/Yi-34B`. Specifically, we used two models with 6 billion and 34 billion parameters. Among all publicly available models, the 34B Yi model is the state of the art for English, performing even better than several models with more parameters (e.g., Llama-2 70B).

Models that are larger than `xl` often adopt half-precision floating-point numbers for their parameters. Accordingly, the numerical precision in evaluating p_t is lower than with `xl` models. As a result, the dimension values may be more scattered, as was seen in 4.

Appendix C: Dimension Reduction

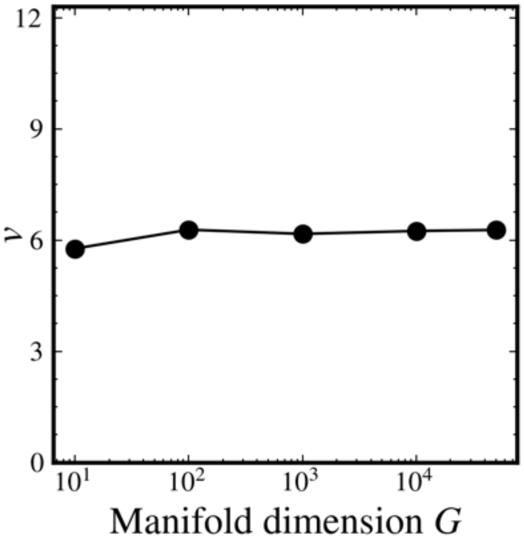


FIG. 6 Correlation dimension of *Don Quixote* as estimated using GPT2-`xl`, with respect to the statistical manifold's dimension G.

We proposed a method to efficiently estimate the correlation dimension $\hat{\nu}$. With this method, we have acquired dimension values that are indistinguishable from those calculated with the direct, naive method, which is 50X slower in our setting.

As detailed in 7, we mapped each distribution p_t into a new distribution q_t with reduced dimensionality. Consequently, with q_t, the correlation dimension can be estimated at a lower computational cost using the following metric:

$$\tilde{d}_{\mathrm{FR}}(p_t, p_s) \equiv 2 \arccos\left(\sum_{m=1}^{M} \sqrt{q_t(m) q_s(m)}\right). \tag{C1}$$

Empirically, varying the manifold dimension M does not alter the results significantly. Figure 6 shows how the estimated correlation dimension $\hat{\nu}$ evolves in relation to M. The rightmost point represents the case of $M = |V|$, i.e., no reduction in the manifold's (topological) dimension. As M decreases, $\hat{\nu}$ barely changes until M reaches its smallest value of 100. This suggests the reliability of the projection $p_t \mapsto q_t$ in preserving $\hat{\nu}$. In this letter, we employed this dimension reduction method in the extensive large-scale experiments, as shown in 4, where M was set to 1000.

In Section G, this dimension reduction method was also applied to acquire a low-dimensional visualization of a language system.

Appendix D: Local Fractality

1. Comparison with Dirichlet Distribution

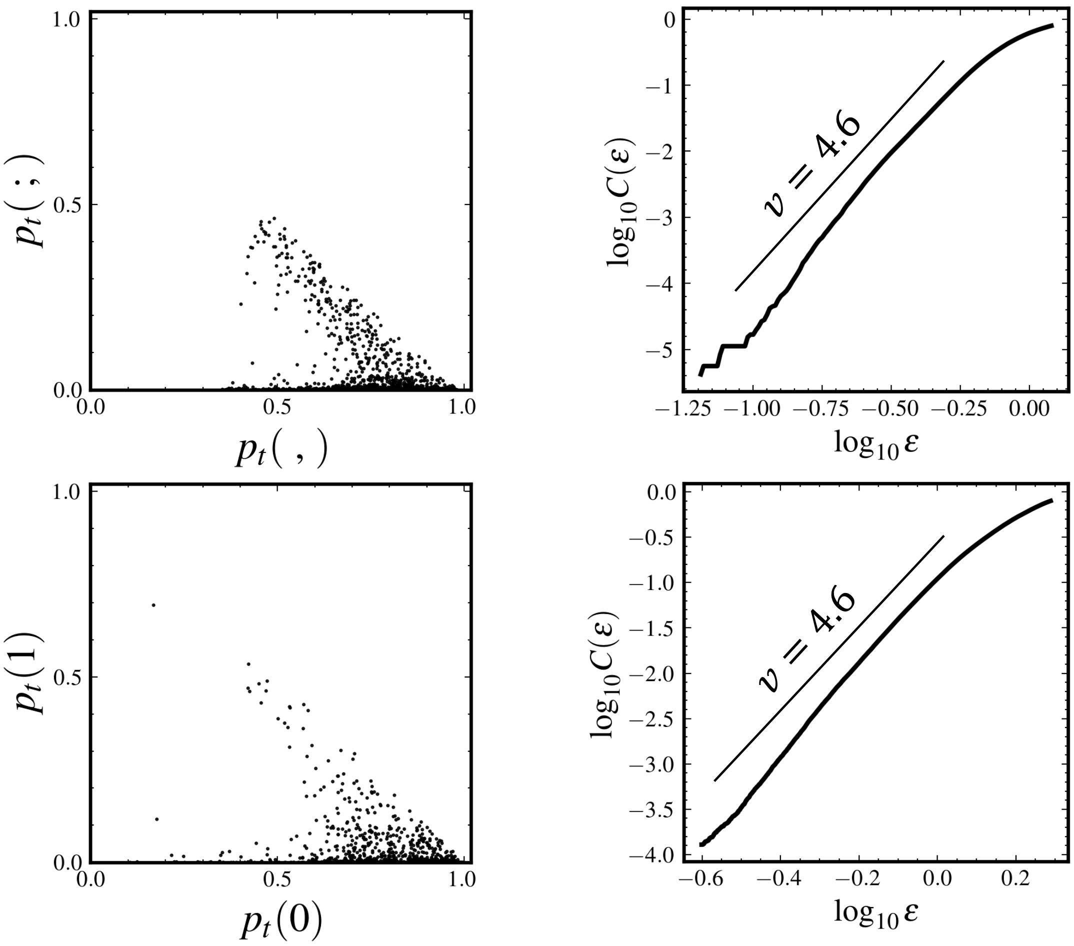


FIG. 7 Comparison between *Don Quixote* (top row) and a sequence of i.i.d. samples drawn from a Dirichlet distribution (bottom row). Left column: probabilities in two dimensions. Right column: correlation integral curves estimated using the points from the left column. It can be seen that the local fractal is reproduced well by the Dirichlet distribution.

As mentioned in the main text, local fractals are observed in low-entropy regions. These fractals show a relatively simpler self-similar pattern that also appears in a Dirichlet distribution.

A Dirichlet distribution is a continuous probability distribution defined on a K-dimensional probability simplex:

$$[z_1, \cdots, z_K] \subset [0,1]^K, \qquad \sum_{k=1}^{K} z_k = 1, \tag{D1}$$

as specified by a parameter vector denoted as $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_K] \in \mathbb{R}^K$. The probability density function of a Dirichlet distribution over $[z_1, \cdots, z_K]$ is defined as follows:

$$p(z_1, \cdots, z_K) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} z_k^{\alpha_k - 1}, \tag{D2}$$

where $B(\alpha)$ is the partition function.

We consider a Dirichlet distribution with $K = 50257$, which is the same as the vocabulary size used by the English GPT2 models. We set the distribution's parameter vector, $\boldsymbol{\alpha} \in \mathbb{R}^{50257}$, such that $\alpha_1 = 3$, $\alpha_2 = 0.2$, and $\alpha_k = 2.2\times10^{-5}$ for $k = 3, 4, \cdots, 50257$.

Figure 7 shows a comparison between *Don Quixote* (upper) and i.i.d. samples drawn from the Dirichlet distribution (lower), with restriction to the *low*-entropy region. The local fractal shown in Figure 7 (upper left) is specific to the word ",": a point p_t was selected to appear if $H(p_t) < 3$ and "," has the largest probability in p_t across the vocabulary. The left-hand plots show the probability for each pair of words, while the right-hand plots show the correlation integral curves estimated from the points on the left. As seen in the figure, the simple Dirichlet distribution approximates the real text well. Therefore, the local fractal, which appears when the subsequent word is almost determined, is reproducible with the Dirichlet distribution.

Note, however, that a Dirichlet distribution has completely different behavior with respect to global self-similarity. When the samples from a Dirichlet distribution are restricted to the *high*-entropy region, they do not exhibit the self-similar pattern of language. We demonstrate this in Section G.2.

2. Local Fractals under Small Context Length

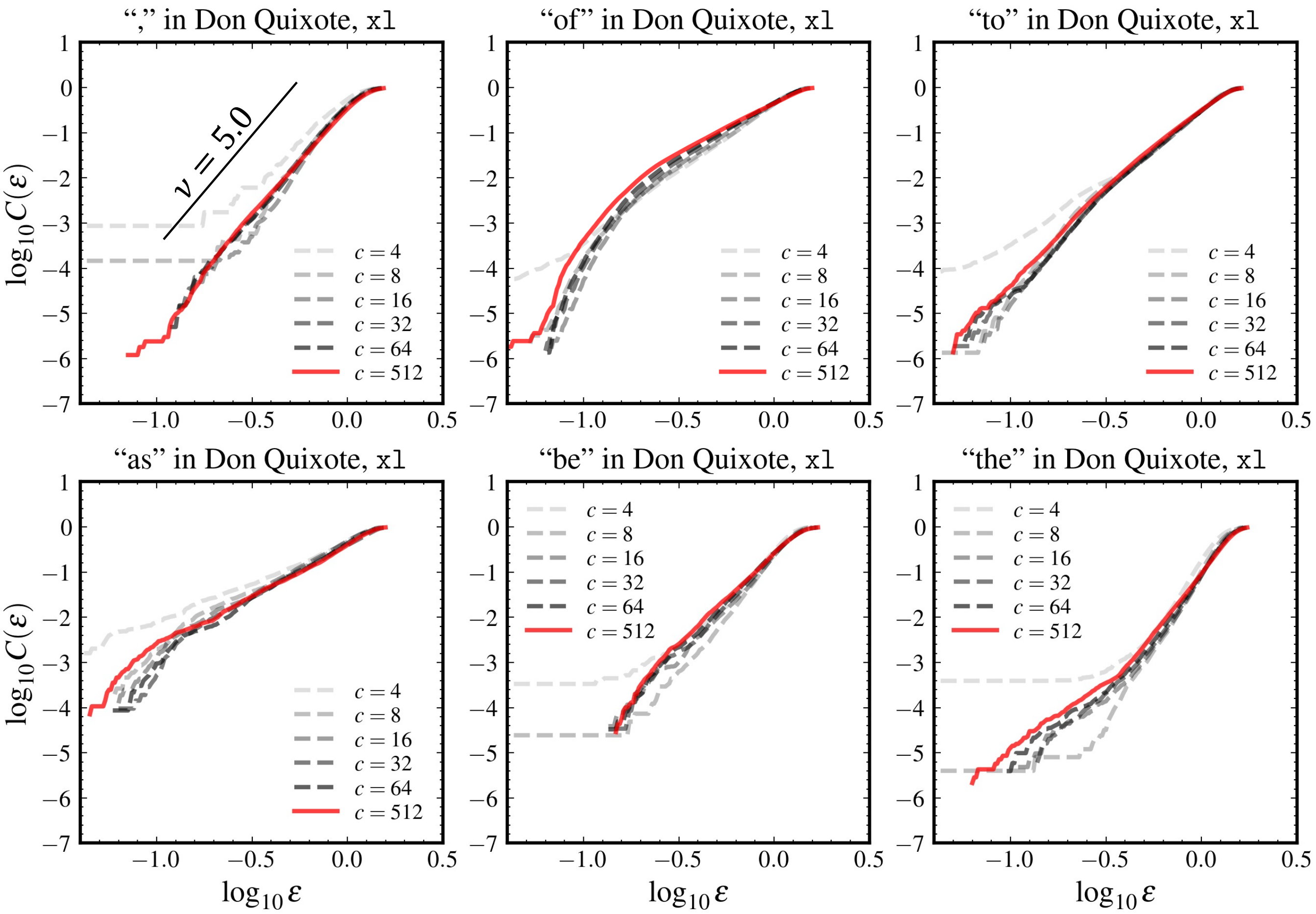


FIG. 8 Correlation integral curves for the local regions for six frequent words with respect to different context length c, estimated with GPT2-xl and a 100,000-word text segment in Don Quixote. The local region for a word was identified by selecting timesteps t among $\{p_t\}$ at which the word has the largest probability mass across the vocabulary and this probability mass exceeds $\eta = 0.5$. We were not able to examine $c < 4$ because the number of timesteps when p_t falls in a local region decreased to zero.

In our statistical manifold, each word corresponds to its own low-entropy region where a local fractal may appear. Hence, local fractals at different words may show varied dimensions. Here, we examined several frequent words ",", "of", "to", "as", "be", "the". Furthermore, we examined the effect of c on their local patterns.

Figure 8 displays the correlation integral curves for specific local regions, with each subfigure corresponding to a distinct word and each plot reflecting various c values, ranging from 512 to 4.

The local region associated with a word was identified by selecting timesteps t from the set $\{p_t\}$ where the word exhibits the highest probability mass within the vocabulary, and this mass surpasses a threshold of $\eta = 0.5$. We refrained from examining c values lower than 4 since, at $c < 4$, the frequency of timesteps where p_t aligns with the local area was too scant (fewer than 50 instances), making it challenging to derive accurate estimates.

Figure 8 reveals the *absence* of a consistent shift in the correlation integral curve observed as c decreases. The curves for smaller c values show fluctuations around the curve for $c = 512$. While occasional upward shifts at $c = 4$ can be noted, these do not maintain consistency across different words.

This observation contrasts with Figure 3(c) of the main text which displays a gradual shift of the global correlation dimension with respect to c is evident.

This contrast underlines a fundamental disparity between global and local fractal phenomena as discussed in the text. Unlike the global correlation dimension that characterizes the long-memory property of language, the local dimensions characterize certain word-specific characteristics that are invariant under a reduction of context diversity (i.e., decreasing c).

We present a possible explanation for this invariance of the local dimensions under varying c. Consider a specific word. A context is a word sequence, and the word's occurring frequency is measured at each location of the sequence, thus producing a mean μ and variance σ^2 of the frequency distribution for this context. Thus, a variation is expected across different contexts. Assuming this variation abides with a scaling law: $\sigma^2 \sim \mu^\gamma$, where γ is the scaling exponent and is somehow related to the local correlation dimension. Reducing the context length c is equivalent to averaging multiple contexts and forces the LLM to estimate the averaged occurring probability of the word.

The following theorem shows that γ is preserved under random pairing and merging of contexts.

Theorem 5. *(Invariance of γ under context merging) Consider a set of independent contexts $\{\boldsymbol{a}^1, \boldsymbol{a}^2, ..., \boldsymbol{a}^{2L}, ...\}$ and a word w. Define $p(w|\boldsymbol{a})$ as the occuring frequency of the word w in a context $\boldsymbol{a}$. If the variance and mean of the word's frequency $p(w|\boldsymbol{a})$ across these contexts follow the scaling relationship $Var_l[p(w|\boldsymbol{a}^l)] \propto \mathbb{E}_l[p(w|\boldsymbol{a}^l)]^\gamma$. Then, upon merging these contexts pairwise into a new set $\{\bar{\boldsymbol{a}}\}$, where $\bar{\boldsymbol{a}}^l = \{\boldsymbol{a}_{2l-1}; \boldsymbol{a}_{2l}\}$ and the frequency $p(w|\bar{\boldsymbol{a}}^l) = (p(w|\boldsymbol{a}^{2l-1}) + p(w|\boldsymbol{a}^{2l}))/2$, the resulting word frequency $p(w|\bar{\boldsymbol{a}}^l)$ still follows the same scaling relationship with the exponent γ remaining unchanged.*

Proof. Define $\mu := \frac{1}{2L}\sum_{l=1}^{2L} p(w|\boldsymbol{a}^l)$ and $\bar{\mu} := \frac{1}{L}\sum_{l=1}^{L} p(w|\bar{\boldsymbol{a}}^l)$. Also, let $\sigma^2 := \frac{1}{2L}\sum_{l=1}^{2L}(p(w|\boldsymbol{a}^l) - \mu)^2$ and $\bar{\sigma}^2 := \frac{1}{L}\sum_{l=1}^{L}(p(w|\bar{\boldsymbol{a}}^l) - \bar{\mu})^2$. It is straightforward to show that $\bar{\mu} = \mu$. Furthermore,

$$\begin{aligned}\bar{\sigma}^2 &= \frac{1}{4L}\sum_{l=1}^{L}\left[(p(w|\boldsymbol{a}^{2l-1}) - \mu) + (p(w|\boldsymbol{a}^{2l}) - \mu)\right]^2 \\ &= \frac{1}{2}\sigma^2 + \frac{1}{2L}\sum_{l=1}^{L}(p(w|\boldsymbol{a}^{2l-1}) - \mu)(p(w|\boldsymbol{a}^{2l}) - \mu),\end{aligned}$$

where the cross-term vanishes as $L \to \infty$ due to independence. Thus, $\bar{\sigma}^2 \to \sigma^2/2$.

Given $\sigma^2 = \beta\mu^\gamma$, it follows that $\bar{\sigma}^2 = \frac{\beta}{2}\bar{\mu}^\gamma$ at large L, indicating γ remains unchanged under context merging. □

Topic models (Blei *et al.*, 2003) have been successful in modeling this variation of word frequency distribution across contexts by considering the notion of topics. Interestingly, Doxas *et al.* (2010) also showed that their observed scaling structure in language sequences can be reproduced by topic models.

Appendix E: Our Data

TABLE I Summary of the datasets used in this letter.

Dataset	language	# sequences	sequence length
	Books		
Gutenberg	English	80	150,000
	Chinese	32	150,000
	German	16	150,000
Aozora-bunko	Japanese	16	150,000
Stanford Encyclopedia of Philosophy	English	60	25,000
Wikipedia webpages	English	40	30,000
Academic papers	English	242	15,000
	Music		
`gtzan`	-	1000	1,500

Table I summarizes the datasets used in this letter.

For books, we used texts from Project Gutenberg, except for the Japanese texts, which were taken from Aozora Bunko [3]. A minimum file size of 1 megabyte was used as a threshold to obtain the longest texts in each collection. For each book, we skipped the first 50,000 words, because this first section includes catalog information and any prologue to the main text. The next 150,000 words of every text were thus used to estimate the correlation dimension.

The texts were split into words by using the tokenizers released with the GPT models. The tokenizers split a text into "sub-word" units, which is a common technique to deal with rare words within a fixed-size vocabulary.

English texts from three other sources were also tested with our method: the Stanford Encyclopedia of Philosophy (SEP, `https://plato.stanford.edu/`), Wikipedia webpages, and academic papers (`https://huggingface.co/datasets/orieg/elsevier-oa-cc-by`). Texts beyond a certain length threshold were chosen: 25,000 words for SEP, 30,000 words for Wikipedia, and 15,000 words for the academic papers.

We also tested our method on the `gtzan` music dataset (Tzanetakis and Cook, 2002) of WAV files. This dataset contains 1000 pieces of music (30 s each) categorized into 10 genres: rock, country, metal, blues, disco, pop, hip-hop, jazz, reggae, and classical. Each genre has 100 music pieces.

The `encodec` library was used to compress each music piece into a sequence of discrete codes constituting a vocabulary of size 2048; i.e., $|V| = 2048$. Each 30-s piece was encoded as a sequence of 1500 timesteps.

In general, the `encodec` library compresses a music piece into four tracks (sequences) based on four different codebooks (vocabularies), of which one is the main track. We considered only the main track because it contains the most information in a music piece.

[3] `https://www.aozora.gr.jp/`

Appendix F: Supplementary Results on Book Texts

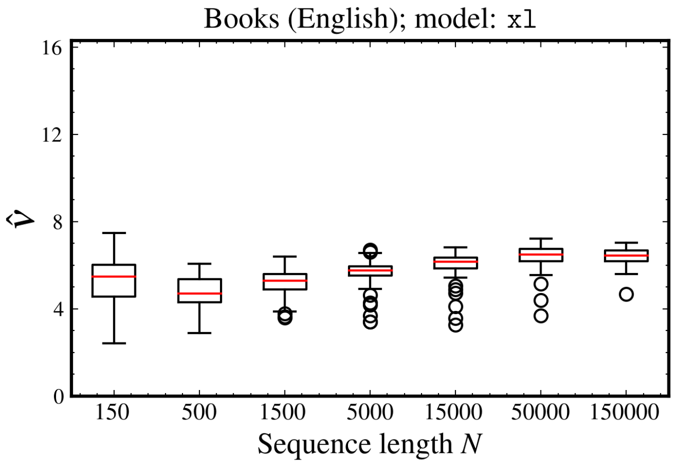


FIG. 9 Distribution of the correlation dimensions for the 80 English books in the Gutenberg Project, as measured using text fragments of different lengths.

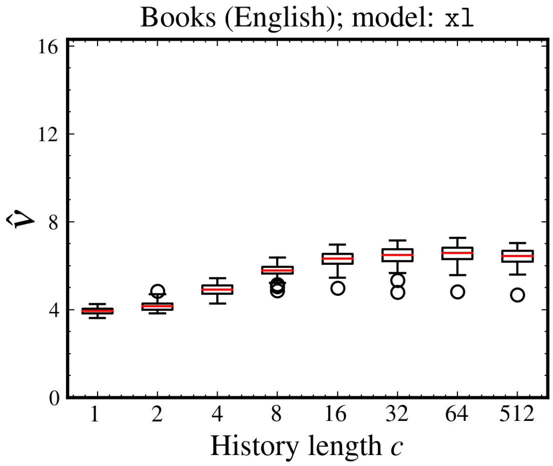


FIG. 10 Distribution of the correlation dimensions for the 80 English books in the Gutenberg Project with respect to c.

In 3 we reported the convergence of the correlation dimension with respect to the number of timesteps, N, and the context length c for the book Don Quixote. Here, we provide the results for the entire Gutenberg corpus of 80 English books. Briefly, we observed the same convergence with respect to increasing N or c as in the case of Don Quixote.

1. Effect of Sequence Length

Different values of N represent different lengths of the text fragment taken from a book to estimate the correlation dimension. As explained above, we discarded the first 50,000 words to avoid catalog information. Then, starting from word 50,001, we took text fragments with different numbers of words. The longest fragment had 150,000 words.

Figure 9 shows the distributions of the correlation dimensions (vertical axis) for the 80 books with N ranging from 150 to 150,000. A clear convergence to a dimension around 6.5 is visible as N increases. The correlation dimension's variance between books is large for small N but decreases gradually as N increases to 150,000. When $N = 150000$, the correlation dimension is 6.39 ± 0.40, as mentioned in the main text.

2. Effect of Context Length

The context length c defined in 9 determines how many previous words are visible to the language model. For LLMs, c cannot be infinitely large. We set $c = 512$ for all textual data in this letter.

It is of great interest how the context length c affects the correlation dimension of natural language. A small c value represents a short-memory approximation to natural language, which has been shown to have long-term memory. In particular, $c = 1$ corresponds to a Markov process (forced by a real text).

Figure 10 shows the distribution of correlation dimensions for the 80 books with respect to increasing c. When $c = 1$, the correlation dimensions are close to 4.0 with small variance. The dimension increases to around 6.5 when c exceeds 16. At $c = 512$, the correlation dimension is 6.39 ± 0.40, as mentioned in the main text. This difference between the results with $c = 1$ and $c = 512$ further validates our finding in this letter that the self-similarity of language is related to the existence of long memory in text.

Appendix G: Comparison with Random Processes

We tested three random processes in the statistical manifold S of multinoulli distributions over $V = \{1, 2, \cdots, K\}$, a vocabulary of K words.

1. Uniform White Noise

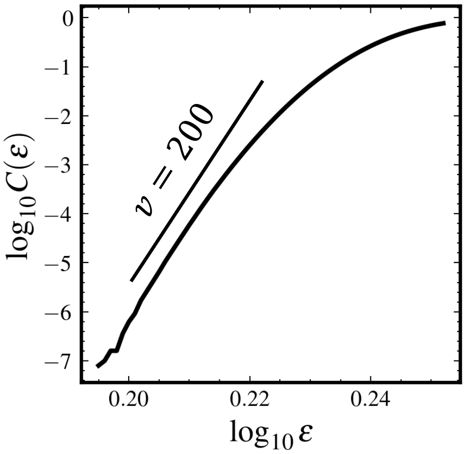


FIG. 11 Correlation integral curve of a uniform white-noise process in a statistical manifold. The process had $N = 10000$ timesteps. $C(\varepsilon)$ was calculated using 2 and the Fisher-Rao distance in 6.

The first random process was a uniform white-noise process. The manifold (S, d_{FR}) is isometric to the positive orthant of a hypersphere with coordinates $(\sqrt{\theta_1}, \sqrt{\theta_2}, \cdots, \sqrt{\theta_K})$. Therefore, a uniform white-noise process can be generated by uniformly drawing samples on this orthant, which can be done efficiently by normalizing Gaussian-distributed random vectors. We consider the uniform white-noise process in S to be an analog of a Gaussian white-noise process in a Euclidean space, because each one is the "entropy-maximizing" distribution in its corresponding space.

As shown in Figure 11, the correlation dimension of such a uniform white-noise process is large at over 100, which is identical to that of a Gaussian white-noise process in a Euclidean space.

2. Symmetric Dirichlet White Noise

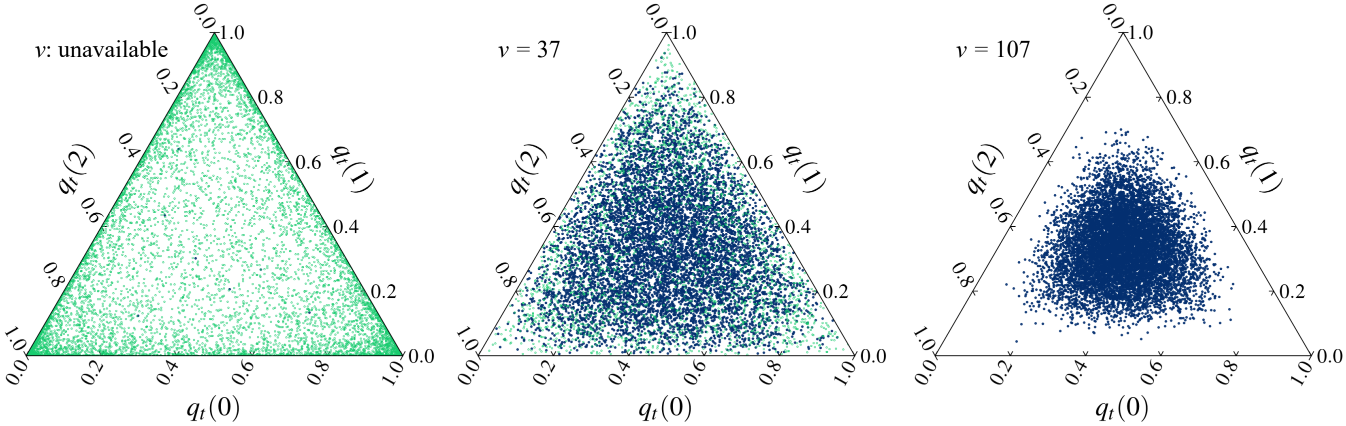


FIG. 12 Illustration of i.i.d. samples drawn from a symmetric Dirichlet distribution with increasing parameter values of $\alpha_k = 1/K$ (left), $5/K$ (middle), and $20/K$ ($k = 1, 2, \cdots, K$). The samples were probability vectors $p_t \in \mathbb{R}^{50257}$ and were projected to $q_t \in \mathbb{R}^3$ via 7, as detailed in Section C. The points are blue for $H(p_t) \geq 3.0$ and green otherwise. The estimated correlation dimension of the blue points is shown in each plot's upper-left corner. For the left plot, the correlation dimension is unavailable because there are almost no blue points.

The second random process was a symmetric Dirichlet white-noise process. In Section D, we showed that i.i.d. samples from a Dirichlet distribution reproduce the local self-similarity of language. A natural question is whether global self-similarity is also reproducible by a Dirichlet distribution. To show that this is not the case, we generated i.i.d. samples from several symmetric Dirichlet distributions. For a symmetric Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$ ($\boldsymbol{\alpha} \in \mathbb{R}^K_+$), the parameters α_k ($k = 1, \cdots, K$) were set identically through k. Three different symmetric Dirichlet distributions with parameters $\alpha_k = 1/K$, $5/K$, and $20/K$ ($k = 1, 2, \cdots, K$) were generated. For consistency with the English GPT2 models, K was set to 50257. Time series of these Dirichlet distributions were produced as $p_t \in \mathbb{R}^{50257}$ and then projected to $q_t \in \mathbb{R}^3$ via Formula (7). Figure 12 shows a plot map in these three dimensions. The points are colored blue for $H(p_t) \geq 3.0$ and green otherwise.

As seen in the figure, the correlation dimension depended on α_k. For a larger α_k, in the middle and right plots, the high-entropy blue points produced large correlation dimensions of 37 and 107, respectively, which indicates that these processes were barely self-similar. For a smaller α_k (left plot), all the points are green, and the correlation dimension of the high-entropy points is thus unavailable. These results show that simple i.i.d. samples from a Dirichlet distribution cannot reproduce the global fractals of natural language.

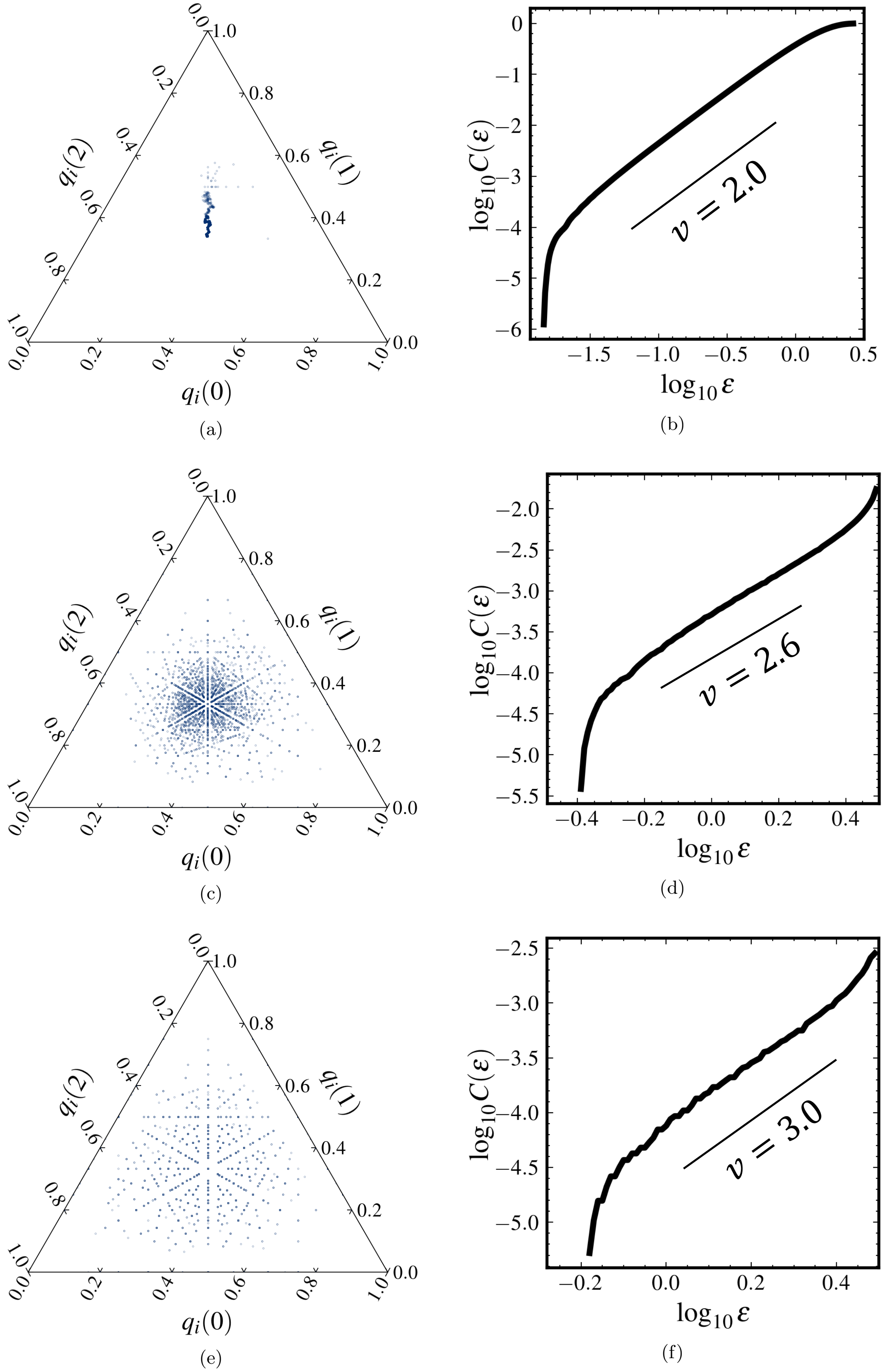
3. Barabási-Albert Network and A Fractional Variant

The third random process was the generation process of a Barabási-Albert (BA) network (Barabási and Albert, 1999). This process is driven by a property called "preferential attachment" and is a special case of the Simon model (Simon, 1955). A BA network starts as a small network of m_0 nodes, where each node is randomly connected to $m \geq m_0$ nodes. At each iteration, a new node is added to the network and randomly connected to an existing node in proportion to the nodes' (instantaneous) degrees. Consider the generation process of a BA network with K nodes, starting from m_0 nodes. For $K - m_0$ times, a new node is connected to an existing node, following a multinoulli distribution over all K nodes, including those that have not yet been appended to the network at time t.

We take the connection probability as the "next-word" probability p_t. Therefore, at time t ($t = 1, 2, \cdots, K - m_0$), the connection probability $p_t(k)$ is calculated as follows:

$$p_t(k) = \begin{cases} \dfrac{\deg_t(k)}{\sum_{k'=1}^{m_0+t-1} \deg_t(k')} & \text{if } k \leq m_0 + t - 1 \\ 0 & \text{otherwise,} \end{cases} \tag{G1}$$

where $\deg_t(k)$ is the degree of node k at time t.



(a)

(b)

(c)

(d)

(e)

(f)

FIG. 13 Trajectory (left) and correlation integral curve (right) of (a-b) a Barabási-Albert model, (c-d) a FAPA model with $\kappa = 0.005$, and (e-f) a FAPA model with $\kappa = 0.005$. The trajectories $\{q_t\}$ were acquired from $\{p_t\}$ by the dimension-reducing mapping in 7. $m = 1$ for all models.

Because the evolution of a BA network depends only on the nodes' current numbers of connections, the connection probability distribution p_t identifies the state of the network as a dynamical system.

In addition to the standard BA model, we considered a fractional variant. This non-standard BA, called *fractional anti-preferential attachment* (FAPA), is based on Rak and Rak (2020). Unlike the standard BA, the probability distribution p_t over all existing node is truncated, and thus only a fraction of the nodes are allowed to be connected to new nodes. In a FAPA, the truncation is "anti-preferential," that is, low-degree nodes remain in this truncation while all high-degree nodes are assigned probability mass zero.

The truncation rate is controlled by a ratio κ, and Formula (G1) is modified as follows:

$$p_t(k) = \begin{cases} \dfrac{\deg_t(k)}{\sum_{k' \in L_t} \deg_t(k')} & \text{if } k \in L_t \\ 0 & \text{otherwise,} \end{cases} \tag{G2}$$

where L_t is the set of nodes which are among the $|L_t| = \kappa(m_0 + t - 1)$ nodes with the lowest degrees.

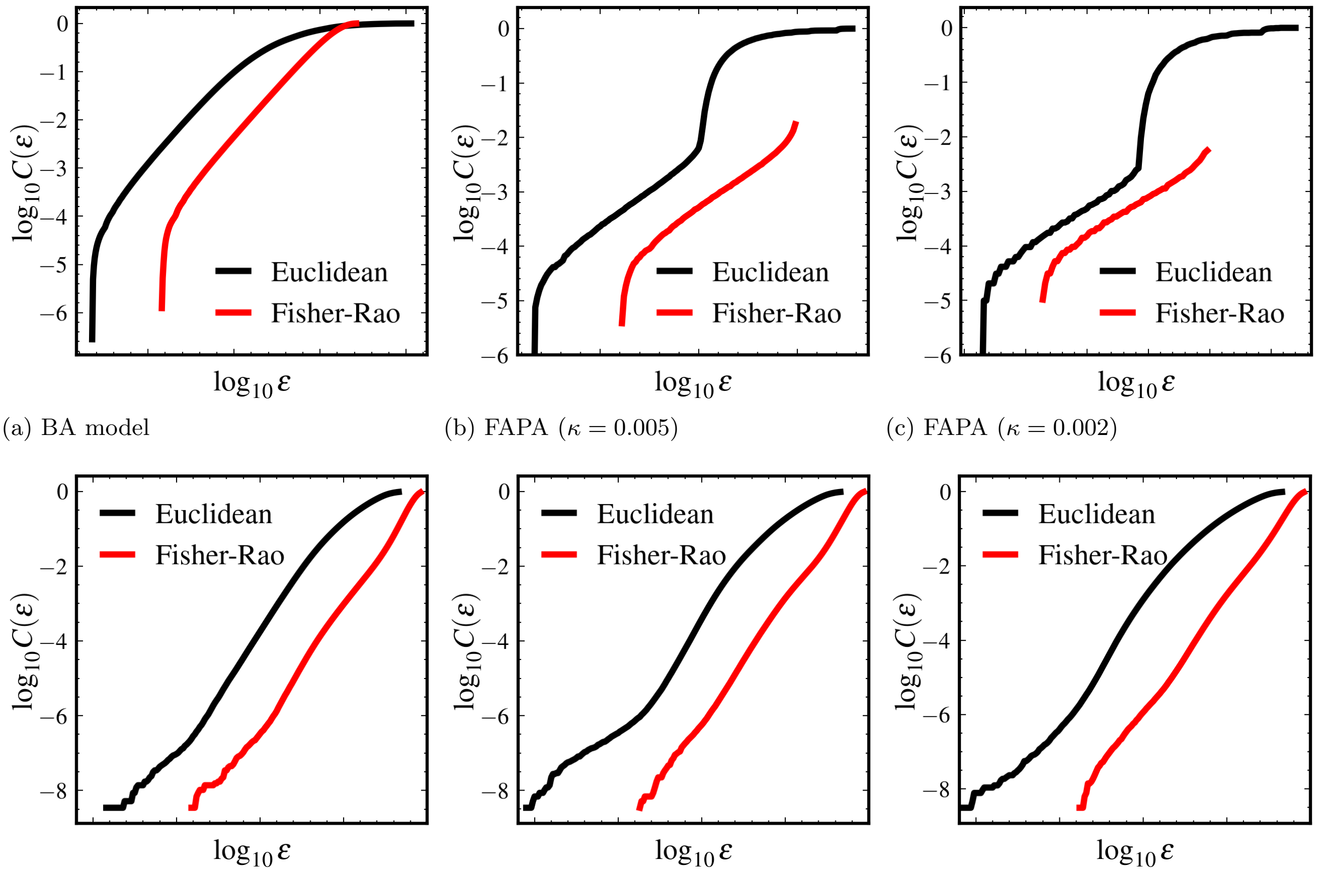
Figure 13 demonstrates the simulated processes of the standard BA model (top row) and two FAPA models with $\kappa = 0.005$ (middle row) and $\kappa = 0.002$ (bottom row), respectively. All the models started from a single node ($m_0 = 1$); at every iteration, one node is added to the network (i.e., $m = 1$); 10,000 iterations were simulated. Hence, p_t is a multinoulli distribution over 10,000 classes (nodes).

The left-hand side of Figure 13 displays the trajectories with the dimension reduced to 2D by using the dimension-reduction mapping in 7, and the right-hand side shows their corresponding correlation integral curves calculated with 2.

A scaling structure can be seen in all three trajectories. The correlation dimension was 2.0 for the standard BA (top row), 2.6 for the FAPA with $\kappa = 0.005$, and 3.0 for $\kappa = 0.002$. The FAPA model with the smaller κ value showed a larger correlation dimension.

The BA model is considered as a correspondent to the random walk model in Euclidean spaces, and the FAPA model to the fractional Brownian motion (fBm). The parameter κ of the FAPA model is analogous to the Hurst exponent of fBm models.

Appendix H: Using Euclidean Distance Metric



(a) BA model

(b) FAPA ($\kappa = 0.005$)

(c) FAPA ($\kappa = 0.002$)

(d) *Don Quixote*

(e) *History of Modern Philosophy*

(f) *"The Man Who Laughs"*

FIG. 14 Correlation integral curves depicted using either the Fisher-Rao distance metric (in red), or the conventional Euclidean distance metric (in black). Figures (a-c) present comparisons for three models outlined in Section G.3 and Figure 13: (a) a standard BA model, (b) a FAPA model with $\kappa = 0.005$, and (c) a FAPA model with $\kappa = 0.002$. Figures (d-f) present curves for three literary sequences: (d) *Don Quixote* by Miguel de Cervantes, (e) *History of Modern Philosophy* by Richard Falckenberg, and (f) *The Man Who Laughs* by Victor Hugo. The black curves have been shifted horizontally for a clearer comparative illustration alongside the red curves.

This section explores an alternative method for analyzing the language sequences $\{p_t\}$ by measuring the distances between p_t and p_s $(\forall t, s)$ using the conventional Euclidean distance metric, $\|p_t - p_s\|$. This approach contrasts with the Fisher-Rao distance method that is outlined in 6 and used throughout this paper.

Figure 14 illustrates the correlation integral curves for six distinct sequences $\{p_t\}$. Each graph depicts a single sequence, showing two curves: one generated by using the Euclidean distance metric (in black) or the other using the Fisher-Rao distance metric (in red). Figures 14(a-c) present simulated processes from the three BA models discussed in Section G.3: (a) the standard BA model, (b) a FAPA model with $\kappa = 0.005$, and (c) a FAPA model with $\kappa = 0.002$. Figures 14(d-f) feature three language sequences extracted from text segments of (d) *Don Quixote* by Miguel de Cervantes (Gutenberg NO. 996), (e) *History of Modern Philosophy* by Richard Falckenberg (Gutenberg NO. 11100), and (f) *The Man Who Laughs* by Victor Hugo (Gutenberg NO. 12587).

Observations reveal that for the simple models depicted in (a-c), the slopes obtained using Euclidean distances align with those from the Fisher-Rao distance. In (b) and (c), a non-linear region emerges at larger values of ε when employing the Euclidean distance, an effect not observed with the Fisher-Rao metric. However, for the actual language sequences displayed in (d-f), the Euclidean-based curves (in black) demonstrate compromised linearity compared with the Fisher-Rao-based curves (in red), making it challenging to pinpoint linear regions for curves in (e) and (f).

These findings highlight two critical insights. Firstly, there is a notable consistency in the correlation dimension values between statistical manifolds and traditional Euclidean spaces, especially for well structured simple processes like the BA models. Thus, the universally observed dimension value of 6.5 for natural language can be analogously compared with random processes defined in Euclidean spaces, such as fractional Brownian motion with a Hurst exponent of $H = 0.15$. This consistency, however, may be limited to $\{p_t\}$ sequences situated near the central region of the statistical manifold—where the global fractal structure of language is located, a condition met by the three BA models.

Secondly, the self-similarity property of language, as observed in the global fractal, becomes significantly more pronounced and apparent when analyzed within a statistical manifold utilizing the Fisher-Rao distance metric, compared with a traditional English space. This distinction is expected, given that the Euclidean distance between two probability vectors lacks mathematical relevance, a point previously discussed in the main text.

Appendix I: Correlation dimension of music data

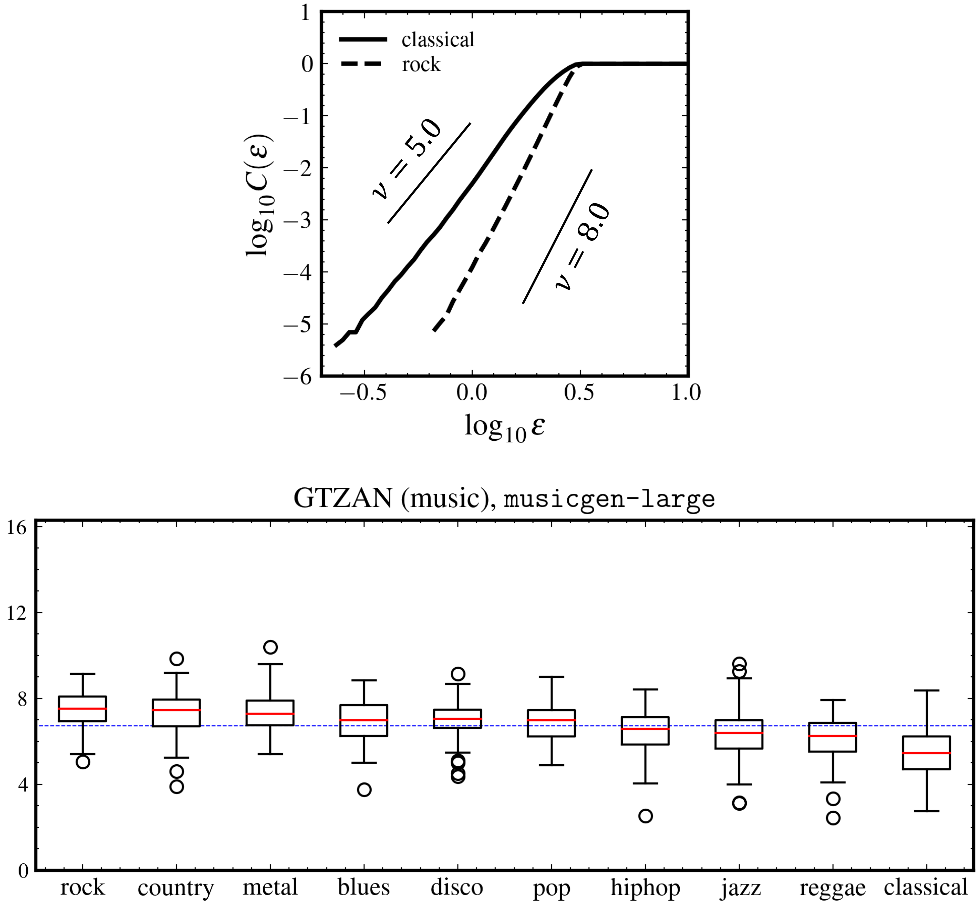


FIG. 15 (Upper) Correlation integral curves for two music pieces categorized as "classical" and "rock" in the `GTZAN` dataset. (Lower) Correlation dimensions of compressed music pieces categorized by genre. The dashed blue line indicates the average over all genres.

As mentioned in Section E, each music piece was encoded as a sequence of discrete codes from a "vocabulary" of 2048 "words." Then, we acquired the system states p_t for each timestep t ($t = 1, \cdots, N$) by using an LLM trained for music generation. In this letter, we used the `musicgen` model (Copet *et al.*, 2023) of size `large` (with 3×10^9 parameters). The `musicgen` model works in the same way as GPT2. Every music piece in the `GTZAN` dataset was compressed into a discrete sequence of 1,500 timesteps, i.e., $N = 1500$.

The context length c defined in 9 was not limited: we used the whole context without approximation. Unlike for texts, which have tens of thousands of timesteps, the music sequences only had 1,500 timesteps, which is within the `musicgen` model's limitation on the context length.

In estimating the correlation dimension $\hat{\nu}$ for the music data, we excluded the low-entropy timesteps from $\{p_t\}$. The maximum-probability threshold η was set to 0.5, the same value used for characterizing texts.

Figure 15 (upper) shows the correlation integral curves for two typical music pieces in the dataset. Linear scaling is clearly visible throughout the whole regions for both pieces. The music piece categorized as "classical" had a correlation dimension (i.e., the slope) around 5.0, while the piece categorized as "rock" had a dimension of 8.0. Next, Figure 15 (lower) shows the distribution of the correlation dimensions of all 1000 pieces, grouped by genre. Rock music had the highest correlation dimension on average, while classical music had the lowest.

This discrepancy in the correlation dimension between music genres aligns well with our perception that classical music is less random than rock or metal music. Nevertheless, as music pieces may show large variety in terms of many other factors (e.g., timbre), we are less likely to observe a universal value of the correlation dimension for music than for language.