Letter

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, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

[2] Department of Computer Science and Engineering, School of Fundamental Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

(Received 25 July 2023; revised 24 January 2024; accepted 9 April 2024; published 2 May 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.

DOI: 10.1103/PhysRevResearch.6.L022028

Introduction. The correlation dimension of Grassberger and Procaccia [1] 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 [1], random processes [2], and sequences sampled from complex networks [3].

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 [4–6], 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.* [7], 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 [8–10]. 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) [11–14] 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 [12], Llama-2 [13], and "Yi" [14]) 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 an approach to examine the underlying structures of pretrained neural networks, thus shedding light on the intricate ways they mirror human intelligence.

These 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 [15,16] 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.

Method. Let (S, d) be a metric space and $[x_1, x_2, \ldots, x_N]$ be a point sequence, where $x_t \in S$ for $t = 1, \ldots, N$. The Grassberger-Procaccia algorithm [1] (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

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

*duxin@aoni.waseda.jp
†kumiko@waseda.jp

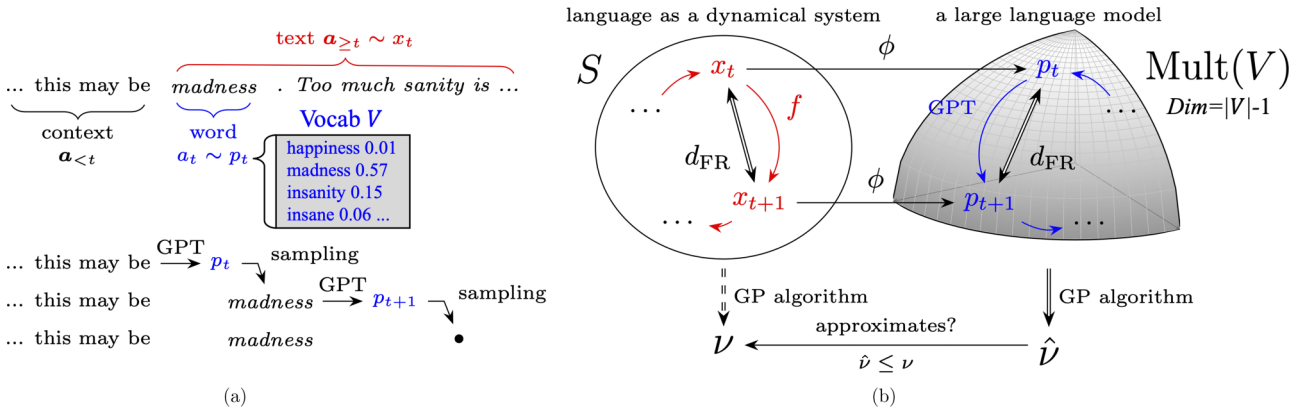

FIG. 1. Our model of language as a stochastic dynamical system. (a) The difference between the system state x_t and the next-word probability distribution p_t. (b) $\{p_t\}$ [where $p_t \in \mathrm{Mult}(V)$] as the image of $\{x_t\}$ (where $x_t \in S$) through the marginalization mapping ϕ in formula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

# 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 [17]. For example, the Hénon map has $\nu = 1.21 \pm 0.01$ [1], which is close to its Hausdorff dimension of 1.261 ± 0.003 [18]. GP can be generalized to apply to a sequence in a more general smooth manifold [17].

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, $\boldsymbol{a} = [a_1, a_2, \ldots, a_t, \ldots,]$, where $a_t \in V$, is associated with a sequence of system states, $[x_1, x_2, \ldots, x_t, \ldots,]$. As demonstrated in Fig. 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 probability of any text to occur as $\boldsymbol{a}_{\geqslant t} = [a_t, a_{t+1}, \ldots,]$, following a context $\boldsymbol{a}_{<t} = [a_1, \ldots, a_{t-1}]$. Furthermore, we consider the next-word probability distribution p_t over the vocabulary V. x_t and p_t are formally defined as follows:

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

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

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 the Supplemental Material [19].

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 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 Fig. 1(b) at the bottom. The Supplemental Material [19] 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 the Supplemental Material [19]. 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 [4–6,20], as mentioned in the Introduction. Therefore, the $\hat{\nu}$ acquired from p_t will remain an approximation of ν. In general, $\hat{\nu} \leqslant \nu$ holds [21] 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 [16]. 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 $\mathrm{Mult}(V)$, as shown at the top right in Fig. 1(b). $\mathrm{Mult}(V)$ 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:

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

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 [22,23] 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 [24] proved Marstrand-like theorems for sets on a

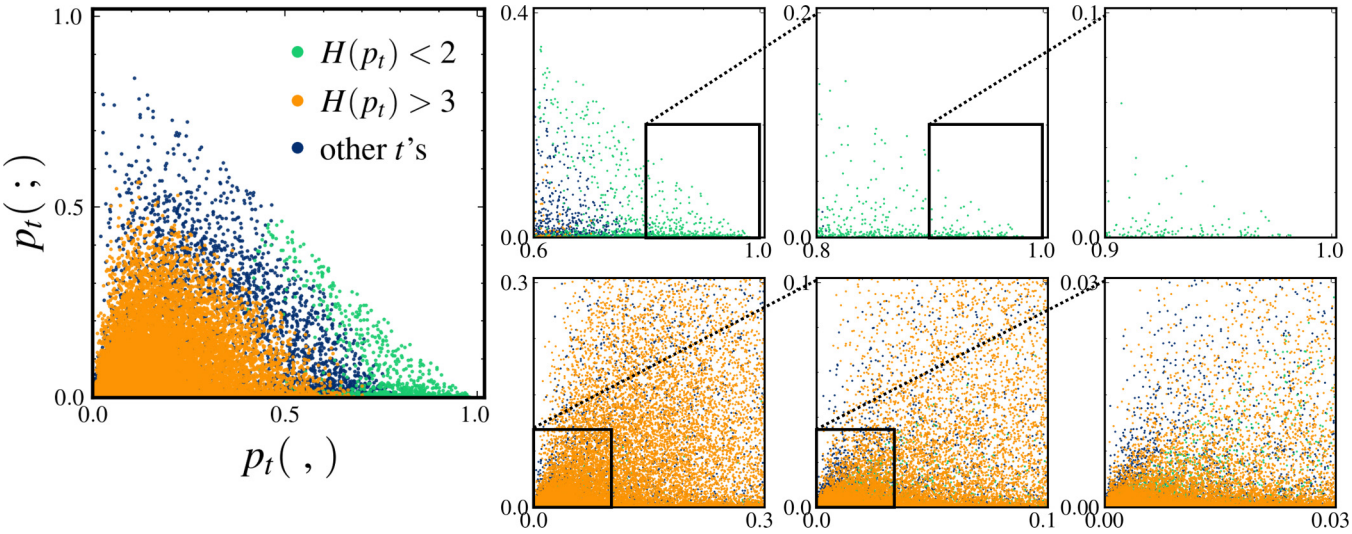

FIG. 2. Sequence of distributions p_t underlying the words in *Don Quixote*, as visualized for words "," (comma) and ";" (semicolon). Each point represents one timestep. The green points represent timesteps at which p_t(",") dominates and the Shannon entropy $H(p_t) < 2.0$, whereas the orange points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

2-sphere. Because the mapping $\phi : x_t \mapsto p_t$ is linear, as mentioned before and proved in the Supplemental Material [19], 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 $t = 1, \ldots, N$, the dimension-reduction projection transforms p_t to q_t as follows:

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

Here, Φ is determined via the modulo function $\Phi(w) = \text{index}(w) \bmod M$, where $\text{index}(w)$ indicates a word's index in the vocabulary. Essentially, we "randomly" group words from the extensive vocabulary V in a smaller set $\{1, \ldots, M\}$ 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 the Supplemental Material [19]. Specifically, dimensionality reduction from approximately 50 000 to 1000 retained the consistency of estimating $\hat{\nu}$ and achieved up to 50x faster computation.

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 $t = 1, 2, \ldots, N$ in an English translation of *Don Quixote* by Miguel de Cervantes from Project Gutenberg [25]. These punctuation marks, chosen for their high frequency, illustrate the role of semantic ambiguity. Each p_t represents a point in $\text{Mult}(V)$, a probability vector of the next-word occurrence, estimated using GPT2-`xl` [11]. 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 $\text{Mult}(V)$, beyond the limits of a two-dimensional display that cannot represent correlation dimensions above two.

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 [26], are evident in low-entropy areas where single words predominate. In the Supplemental Material [19], 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 Fig. 2. The local kind's occurrence could be related to the finding in Doxas *et al.* [7] 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

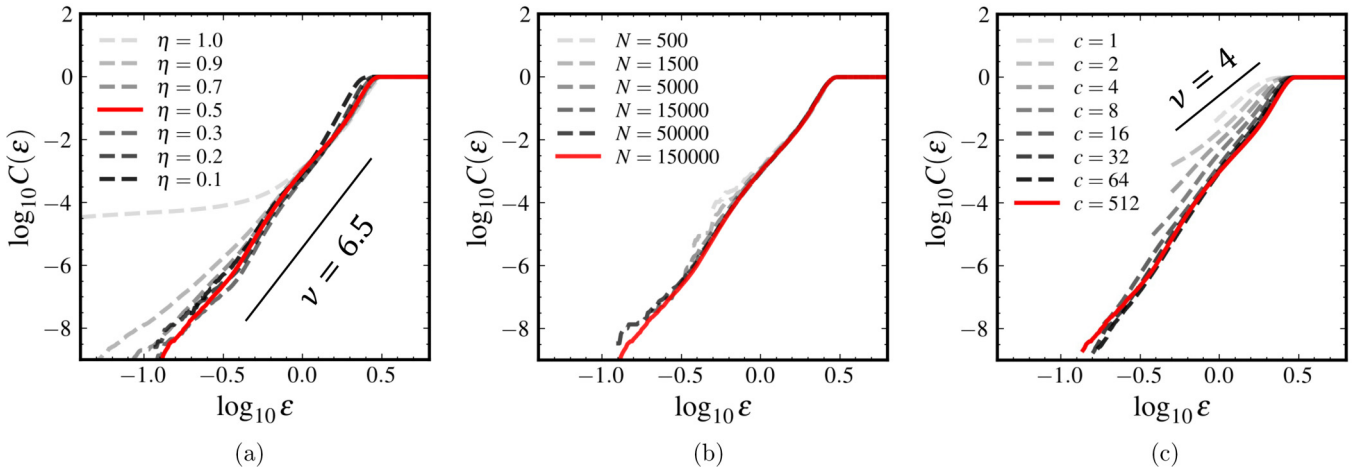

FIG. 3. Correlation integral curves as defined by formula (2) and estimated with GPT2-xl with respect to (a) the maximum-probability threshold η in formula (8), (b) the sequence length N, and (c) the context length c in Formula (9).

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 $\eta = 0.5$, Fig. 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 $[a_1, \ldots, a_{t-1}]$ before timestep t as the context, but in practice, a maximum context length c is often set; that is,

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

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 the Supplemental Material [19].

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.* [7] and mentioned above, the global fractal is due to long memory and was anticipated in the literature [4–6]. Although self-similarity and long memory have often been studied separately and were even conjectured as different aspects of a scale-invariant process [27], they show interesting coordination for natural language. More results on a larger dataset are provided in the Supplemental Material [19].

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 the Supplemental Material [19]. 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 4(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 [14], which offers the SOTA capability in English among all 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 ($\pm$ 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 [28], the Stanford Encyclopedia of Philosophy (SEP) [29], 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 one) implies more significant self-similarity in a text. The

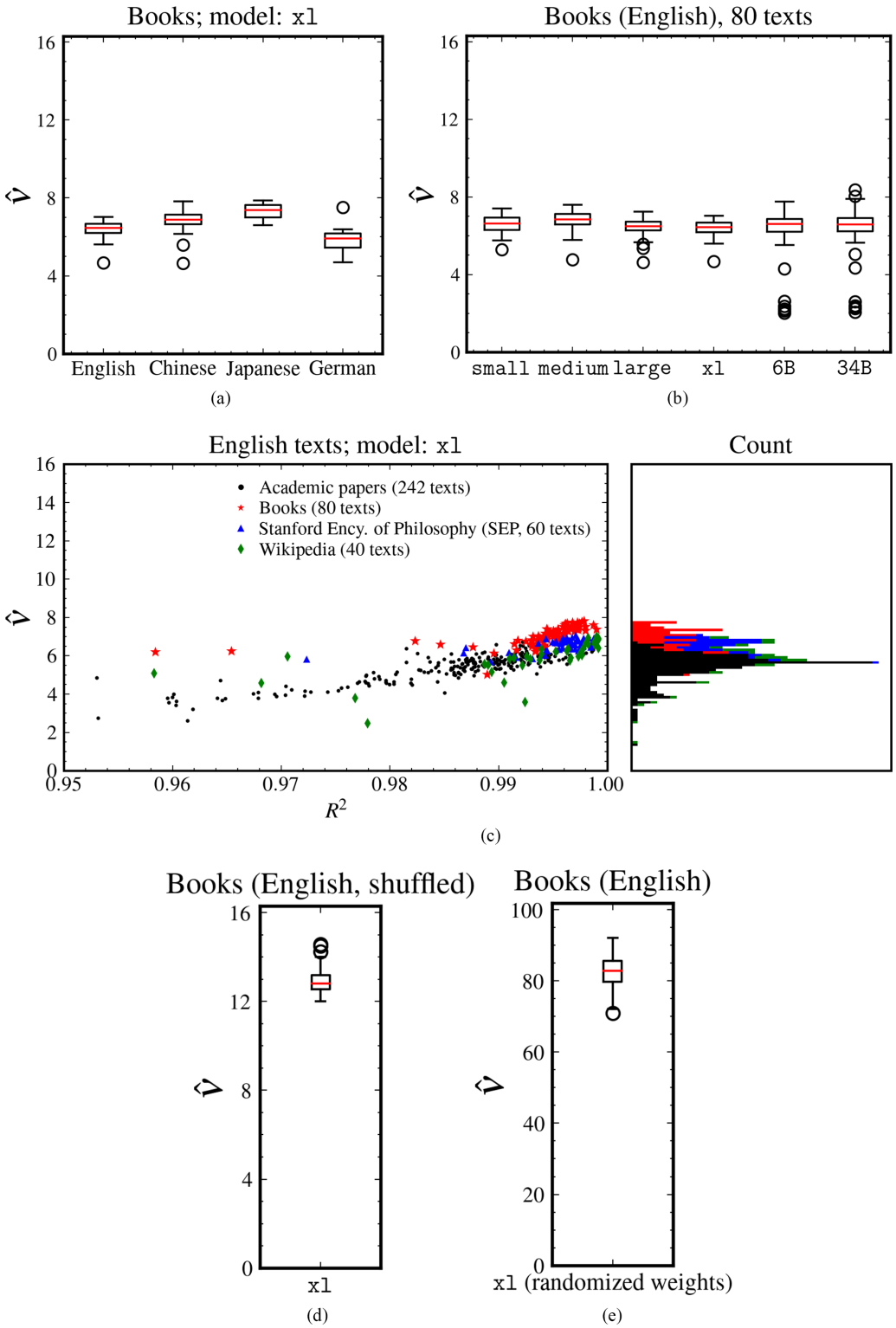

FIG. 4. Correlation dimensions of (a) all books grouped by language, as estimated using GPT2-xl; (b) English books as estimated using GPT with different model sizes (GPT2 from small to xl and the Yi model for 6b and 34b); (c) English texts from various sources with the R^2 scores (horizontal axis) of 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.

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 six, 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" [30], which coincides with ν under ergodic conditions [17]. 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 Fig. 4(d), shuffled word sequences exhibited an average correlation dimension of 13.0, indicating inherent self-similarity despite the shuffling. As seen in Fig. 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 the Supplemental Material [19], 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 ten. Conversely, Barabási-Albert (BA) networks [31], which are special cases of a Simon process, demonstrated a correlation dimension of 2.00 ± 0.003, and a fractal variant [32] 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 the Supplemental Material [19], 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 [33]. To demonstrate the applicability of our analysis, we used the `GTZAN` dataset [34], which comprises 1000 recorded music pieces categorized in ten 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 the Supplemental Material [19].

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) [35]. Although promising, that complex linguistic framework exceeds our current scope, and we expect to explore it in the future.

Acknowledgment. This work was supported by JST CREST Grant No. JPMJCR2114, and JSPS KAKENHI Grant No. JP20K20492.

[1] P. Grassberger and I. Procaccia, Characterization of strange attractors, Phys. Rev. Lett. **50**, 346 (1983).

[2] A. R. Osborne and A. Provenzale, Finite correlation dimension for stochastic systems with power-law spectra, Physica D **35**, 357 (1989).

[3] L. Lacasa and J. Gómez-Gardenes, Correlation dimension of complex networks, Phys. Rev. Lett. **110**, 168703 (2013).

[4] W. Li, Mutual information functions of natural language texts (Citeseer, 1989)

[5] E. G. Altmann, G. Cristadoro, and M. D. Esposti, On the origin of long-range correlations in texts, Proc. Natl. Acad. Sci. USA **109**, 11582 (2012).

[6] K. Tanaka-Ishii and A. Bunde, Long-range memory in literary texts: On the universal clustering of the rare words, PLoS ONE **11**, e0164658 (2016).

[7] I. Doxas, S. Dennis, and W. L. Oliver, The dimensionality of discourse, Proc. Natl. Acad. Sci. USA **107**, 4866 (2010).

[8] T. Kobayashi and K. Tanaka-Ishii, Taylor's law for human linguistic sequences, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistic* (2018), p. 1138.

[9] K. Tanaka-Ishii and T. Kobayashi, Taylor's law for linguistic sequences and random walk models, J. Phys. Commun. **2**, 115024 (2018).

[10] M. Ausloos, Measuring complexity with multifractals in texts. Translation effects, Chaos, Solitons Fractals **45**, 1349 (2012).

[11] A. Radford, J. Wu, R. Child *et al.*, Language models are unsupervised multitask learners, OpenAI blog **1**, 9 (2019).

[12] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman *et al.*, GPT-4 technical report, arXiv:2303.08774.

[13] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, Llama 2: Open foundation and fine-tuned chat models, arXiv:2307.09288.

[14] The Yi model, (2024), visited in January 2024, https://huggingface.co/01-ai/Yi-34B.

[15] C. R. Rao, Information and the accuracy attainable in the estimation of statistical parameters, in *Breakthroughs in Statistics: Foundations and Basic Theory* (Springer, New York, 1992), pp. 235–247.

[16] S.-I. Amari, *Differential-Geometrical Methods in Statistics*, Vol. 28 (Springer, New York, 2012).
[17] Y. B. Pesin, On rigorous mathematical definitions of correlation dimension and generalized spectrum for dimensions, J. Stat. Phys. **71**, 529 (1993).
[18] D. A. Russell, J. D. Hanson, and E. Ott, Dimension of strange attractors, Phys. Rev. Lett. **45**, 1175 (1980).
[19] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.6.L022028 for theoretical analysis of our method and extensive evaluation results.
[20] E. G. Altmann, J. B. Pierrehumbert, and A. E. Motter, Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words, PLoS ONE **4**, e7678 (2009).
[21] H.-O. Peitgen, H. Jürgens, D. Saupe, and M. J. Feigenbaum, *Chaos and Fractals: New Frontiers of Science*, Vol. 7 (Springer, New York, 1992).
[22] J. M. Marstrand, Some fundamental geometrical properties of plane sets of fractional dimensions, Proc. London Math. Soc. **s3-4**, 257 (1954).
[23] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications* (John Wiley & Sons, England, 2004).
[24] Z. Balogh and A. Iseli, Dimensions of projections of sets on riemannian surfaces of constant curvature, Proc. Am. Math. Soc. **144**, 2939 (2016).
[25] https://www.gutenberg.org/ebooks/996.
[26] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. **3**, 993 (2003).
[27] P. Abry, P. Flandrin, M. S. Taqqu, and D. Veitch, Self-similarity and long-range dependence through the wavelet lens, Theor. Applic. Long-Range Depend. **1**, 527 (2003).
[28] D. Kershaw and R. Koeling, Elsevier OA CC-BY corpus, arXiv:2008.00774.
[29] https://plato.stanford.edu/.
[30] J. D. Farmer, Information dimension and the probabilistic structure of chaos, Z. Naturforsch. A **37**, 1304 (1982).
[31] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).
[32] R. Rak and E. Rak, The fractional preferential attachment scale-free network model, Entropy **22**, 509 (2020).
[33] J. Copet *et al.*, Simple and controllable music generation, in *Advances in Neural Information Processing Systems* (New Orleans, Louisiana, 2023).
[34] G. Tzanetakis and P. Cook, Musical genre classification of audio signals, IEEE Trans. Speech Audio Process. **10**, 293 (2002).
[35] N. Chomsky, *Aspects of the Theory of Syntax*, Vol. 11 (MIT Press, Cambridge, MA, 2014).