Human Versus Machine Intelligence: Assessing Natural Language Generation Models Through Complex Systems Theory

Enrico De Santis, *Member, IEEE*, Alessio Martino, *Member, IEEE*, and Antonello Rizzi, *Senior Member, IEEE*

***Abstract*—The introduction of Transformer architectures – with the self-attention mechanism – in automatic Natural Language Generation (NLG) is a breakthrough in solving general task-oriented problems, such as the simple production of long text excerpts that resemble ones written by humans. While the performance of GPT-X architectures is there for all to see, many efforts are underway to penetrate the secrets of these black-boxes in terms of intelligent information processing whose output statistical distributions resemble that of natural language. In this work, through the complexity science framework, a comparative study of the stochastic processes underlying the texts produced by the English version of GPT-2 with respect to texts produced by human beings, notably novels in English and programming codes, is offered. The investigation, of a methodological nature, consists first of all of an analysis phase in which the Multifractal Detrended Fluctuation Analysis and the Recurrence Quantification Analysis – together with Zipf's law and approximate entropy – are adopted to characterize long-term correlations, regularities and recurrences in human and machine-produced texts. Results show several peculiarities and trends in terms of long-range correlations and recurrences in the last case. The synthesis phase, on the other hand, uses the complexity measures to build synthetic text descriptors – hence a suitable text embedding – which serve to constitute the features for feeding a machine learning system designed to operate feature selection through an evolutionary technique. Using multivariate analysis, it is then shown the grouping tendency of the three analyzed text types, allowing to place GTP-2 texts in between natural language texts and computer codes. Similarly, the classification task demonstrates that, given the high accuracy obtained in the automatic discrimination of text classes, the proposed set of complexity measures is highly informative. These interesting results allow us to add another piece to the theoretical understanding of the surprising results obtained by NLG systems based on deep learning and let us to improve the design of new informetrics or text mining systems for text classification, fake news detection, or even plagiarism detection.**

***Index Terms*—Natural language generation, GPT models, multifractal analysis, recurrence quantification analysis, Zipf's law, quantitative linguistics, complexity science, text classification.**

Manuscript received 27 March 2023; revised 24 October 2023; accepted 21 January 2024. Date of publication 24 January 2024; date of current version 5 June 2024. Recommended for acceptance by M. Choudhury. *(Corresponding author: Enrico De Santis.)*

Enrico De Santis and Antonello Rizzi are with the Department of Information Engineering, Electronics and Telecommunications, University of Rome "La Sapienza", 00184 Rome, Italy (e-mail: enrico.desantis@uniroma1.it; antonello.rizzi@uniroma1.it).

Alessio Martino is with the Department of Business and Management, LUISS University, 00197 Rome, Italy (e-mail: amartino@luiss.it).

Digital Object Identifier 10.1109/TPAMI.2024.3358168

I. Introduction

IN THE *Tractatus*, the Austrian philosopher Ludwig Wittgenstein (1889–1951) – who spent most of his life dealing language –, sought to demarcate *sense* from *nonsense* [1]. One might wonder what Wittgenstein would have thought about sense if he had been able to read a text excerpt produced with a state-of-the-art (pre-trained) generative language model, such as those belonging to the GPT-X neural architectures. In fact, beyond the easy "exceptionalism" in the field of natural language generation (NLG), state-of-the-art generative pre-trained models are performing astonishingly in letting machines to synthesize text in a way that resembles spoken or written language, as typically employed by humans. Perhaps Wittgenstein would have been one of the few to understand the potential of these generative models in what he already claimed: "the meaning of a word is its use in the language". A claim that is strictly related to the so-called "Distributional Semantics"[1] [2], which is related to the thinking of the American linguist Z. S. Harris who in the Fifties affirmed that [3]: "words that are used and occur in the same contexts tend to purport similar meanings". It can be said, in general terms, that modern generative language models based on deep learning are in the same way grounded on the distributional hypothesis. Still, they manage to capture the intimate hierarchical and syntactic-grammatical structure in a truly surprising manner by deeply mimicking the peculiarities of human language. The leap in quality is achieved thanks to the Transformer [4], which is a relatively simple modular architecture. Specifically, it is a well-suited deep learning model that adopts the mechanism of self-attention by differently weighting the significance of each part of the input data, solving to some extent the well-known co-reference problem in a hierarchical fashion [5].

On the other hand, natural language is a system functioning at the interface between biology and social interactions [6]. From the perspective of the science of complexity (and also for linguistics), it is a "discrete combinatorial system" [7] produced by the brain and organized as a complex system [8], [9], [10], [11] structured, in turn, in a hierarchical fashion (i.e., characters, morphemes, words, sentences, etc.). If, on one hand, natural language consists of "making infinite use of finite means" [12],

[1] Distributional Semantics is grounded on the so-called "distributional hypothesis", that is: similarity of meaning correlates with similarity of distribution of words in a text.

many linguists believe that part of the solution to the meaning (or sense) problem is hidden in the so-called *long-distance dependencies* [13], for which the appearance of power-laws are a hallmark of complexity [14]. It is worth noting that, in light of some evolutionary scenarios related to languages, for example discussed in [15] or in [16], long-distance dependencies are not only a feature of human sequences but also of animal ones. Moreover, a given text can be conceived as a symbolic stochastic dynamic system [17]. At the same time, modern neural network architectures during training can incorporate rich dynamical features. In natural language modeling, starting from 2017, researchers showed that self-supervised learning for solving some linguistic tasks (word masking, next sentence prediction, predicting the words likely to occur around a given word, etc.) may allow the construction of rich word-token specific deep contextual representations of human language. Hence, incorporating both the dynamics of language and its hierarchical organization (including word classes, the syntactic structure, such as grammatical relations or dependencies) [18], [19] these models can mimic some cognitive abilities reserved for human learning, approaching at the same time, and for some extent, the General Artificial Intelligence systems [20].

At the time of this writing, OpenAI released ChatGPT,[2] an instruction-based multi-task architecture belonging to the family of Large Language Models (LLMs) – accessible through a web browser – based on GPT-3.5 that is going viral for its formidable performances, not only in chatting but also in generating summaries, translations, writing songs, poetry or stories, demonstrating real semantic capabilities and good knowledge (not perfect) of the world. The system can also generate meaningful programming language codes (i.e., Python, C, Javascript, etc.) starting from a natural language explanation or request (*prompt*). We can consider the present time the "Year 0" of these technologies and we can expect still important improvement in the near future.

However, while the Australian philosopher and cognitive scientist David Chalmers described GPT-3 as "one of the most interesting and important AI systems ever produced" [20], one may try to quantify – within the framework of complexity science – similarities and differences between human and machine-generated texts. This research can be useful both as a purely scientific investigation at the cross-fields between AI and complex systems and in the application domain itself (i.e., comparative studies, content validation, plagiarism detection, text mining, fake news detection, etc.).

The following research – within a methodological perspective – aims mainly to provide the reader with an x-ray of the deep generative models, able to generate quality texts, through a series of heterogeneous complexity measures belonging to the complex system analysis framework. Complexity measures, obtained through suitable complexity indices, constitute a precise digital footprint of a text excerpt (which, for the sake of comparison, will be provided also for texts generated by humans), allowing to highlight the hallmark of complexity and, to some extent, to characterize the underlying random or chaotic behavior. Methodologically, the study proceeds via two main (consequential) phases instantiated through two different frameworks. The first one consists of an analytical phase where the measured complexity indices – together with their underlying theoretical framework – allow us to give a near-complete and heterogeneous picture of the complexity and chaotic/random behavior of texts (in English) generated by the machine compared to those generated by humans. The second is, as instead, a phase of synthesis, in which the complexity measures together with the entire experience obtained in the analysis phase, are used for instantiating a classification problem addressed through a well-suited machine learning technique. In other words, the complexity indices will participate in an embedding procedure towards a Euclidean space, hence forming a feature vector, able to synthetically describe a given text excerpt. For the analytical phase, after a suitable text transformation in a numeric time series (where it applies), several complexity measures are collected. The first one is the Zipf's exponent underlying the Zipf's law on word base [21], [22]. The second one consists of a set of indices estimated through the Multifractal Detrended Fluctuation Analysis (MFDFA) framework [23] allowing to deeply characterize long-range correlations [24] and, in general, to investigate the richness of the correlation structure underlying a given time series [25], [26]. The third set of indices is obtained by means of the Recurrence Quantification Analysis (RQA) [27], a consolidated methodology in the analysis of complex systems that allows characterizing the recurrence structure of a time series in a simple and direct way. The last index is the Approximate Entropy (ApEn) [28], [29], which measures synthetically the amount of regularity and the unpredictability of fluctuations over time series data. Part of the proposed methodology (specifically the MFDFA framework) is mediated by our previous research work with which we studied the morphological characteristics of a set of ancient and modern texts belonging to different linguistic strains [30].

It is worth noting that each adopted analysis framework is language-agnostic, as no prior assumptions are made with respect to the input text language. They allow depicting the complexity of a given text from a specific and differentiated perspective, that is not only by the correlation or recurrence structures but also by wondering to which extent the behavior of a synthetically generated text can be considered random (i.e., random walk-like noise) and at the same time chaotic. Through a series of statistical analyses a deep characterization of differences and similarities of texts produced by a machine compared to ones generated by humans will be provided. Instead, the embedding of text useful to generate feature vectors feeding a machine learning procedure is here adopted not to find the world's best classifier able in discriminating human-generated with machine-generated text excerpts, but for providing a better characterization of similarities/differences between them. This task is carried out by means of an ad-hoc classifier system boosted with a features selection procedure performed with a wrapper-based technique, in turn, realized with an evolutive meta-heuristic (i.e., a genetic algorithm).

As concerns the experimental part, the comparison will be done between three text corpora. The first one is obtained

[2]https://openai.com/blog/chatgpt/

through the GPT-2 architecture by varying the temperature meta-parameter. The second one consists of 80 literary classic novels – written by human writers – retrieved from the Gutenberg Project website. The last one is constituted by several versions of the Linux kernel source code in C language. In the latter case, what is interesting is that i) a source code is a syntactically stable language whose syntax is close to Chomsky Generative Grammars, ii) the rigid syntax of a structured language constrains humans during their creative process in programming a computer in a rigid syntactic scheme compared to writing novels. Hence, we may question how a deductive system like GPT-X (trained by induction) behaves compared to humans in writing literary text and in writing programming codes.

The current paper is organized as follows. Section II provides an overview of the main scientific papers in which the following study is located. Section III provides a brief description of the systems and, in particular, of the Transformer architecture. Section IV describes the complexity measures adopted to describe the analyzed texts. Section V deals with the corpora adopted for the investigation. In Section VI we report the experiments carried out detailing both the analysis phase (performed through the complexity framework) and the synthesis phase (performed via a machine learning approach). Conclusions and final comments are offered in Section VII.

II. Related Works

The interpretation of human writing as a complex system where long-range correlations have a prominent role has a long tradition [8], [31]. In [32], the authors try to model certain features of human language complexity by means of advanced concepts borrowed from statistical mechanics, using a suitable encoding procedure for transforming text in time series. In [33], the authors use Detrended Fluctuation Analysis (DFA) and Grassberger-Proccacia analysis (GP) methods in order to study language characteristics adopting both the word-frequency and the sentence-length mapping. While the GP analysis indicates that linguistic signals may be considered as the manifestation of a complex system of high dimensionality, differently from random signals, the DFA method is found additionally able to distinguish a natural language signal from a computer code signal. A study through the MFDFA technique applied to sentence length – measured by the number of words between two full stops – in a large corpus of world-famous literary texts is provided in [34]. The authors show that an appealing and aesthetic optimum appears somewhere in between and involves self-similar, cascade-like alternation of various lengths of sentences together with a $1/f^{\beta}$ scaling of the spectrum. Furthermore, the authors indicate how the recurrence statistics of full stops can be descriptive of the writing style. An analogous study based on the same mapping on a smaller corpus of literary English text is offered in [35]. The fractal structure of long human-language records by mapping large samples of texts onto time series is faced in [36], through the original Rescaled Range Analysis proposed by Harold E. Hurst. Interestingly, the authors transform a text into a discrete time series mapping the words of the text using Zipf's analysis, i.e., constructing the time series as the sequence of indices of the list ordered according to the frequency of appearance of the words – the same encoding scheme that will be used in our research. In line with Zipf's research, the authors claim that the specific numerical assignment can be considered meaningful because it minimizes the effort in lexical access in the rank-ordered list of words when writing the whole text. In [37], [38], the author measured the generalized Hurst exponent and other multifractal characteristics of original and translated texts, through the partition function method. The author, in both works, through a different level of investigation, shows the multifractal behavior of shuffled and normal text for a number of literary works, some of which were translated into the synthetic language Esperanto. The proposed framework is also adopted for linking complexity to quality in texts. In [39], the authors compare computer programs and natural language texts in terms of complexity and long-range correlations, trying to find similarities and differences. In [40], the authors propose an interesting and recent research – in line with our study – aiming to investigate to which extent artificial texts generated by Long Short-Term Memory (LSTM) networks resemble those generated by humans. The authors measured several complexity indices, such as word-frequency statistics, long-range correlations, and entropy measures, comparing RNN-generated texts with Markov models of various orders and human-generated literary texts, showing that LSTM-generated texts are able to reproduce long-range correlations at scales comparable to those found in natural language. In the specific context of comparing linguistic patterns in human and LLM-generated text, it is worth signaling [41], where the authors investigate – from a more linguistic perspective – several measurable linguistic dimensions, including morphological, syntactic, psychometric and sociolinguistic aspects, starting from contemporary articles from the New York Times. By feeding the articles' headlines to a set of specific LLMs (LLaMa [42]), and comparing the original articles with the ones generated by the machine, the authors find intriguing similarities and differences, such as the restricted vocabulary used by the machine, the use of a more objective language (through the adoption of symbols or numbers) in comparison to the intensive adoption of adjectives in texts written by humans. They also observed variations in terms of syntactic structures, both for dependency and constituent representations, specifically in the use of dependency and constituent types, as well as the length of spans across both types of texts. A study of structural characteristics of modern deep contextual language models and how they learn major aspects of language structure without any explicit supervision is provided in [18]. The authors remark that these models (the ones based on contextual embedding, attention mechanisms and grounded on self-supervised learning tasks) mimic stunningly parse tree distances to a very high degree, allowing approximate reconstruction of the sentence tree structures normally assumed by linguists. Finally, another intriguing study about compositionality properties[3] of language generated by deep neural models is provided in [44].

[3]That is, "the meaning of a whole is a function of the meanings of the parts and of the way they are syntactically combined" [43].

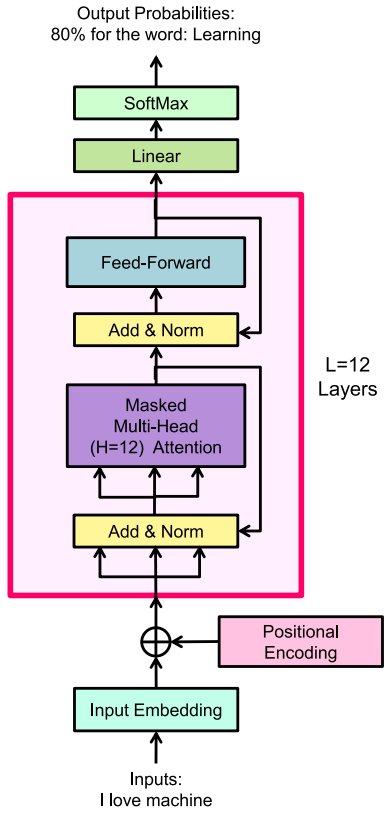


Fig. 1. GPT-2 detailed architecture.

III. Generative Pre-Trained Transformer

GPT models by OpenAI are robust LLM based on the Transformer (deep) neural network architecture, excelling in various NLP tasks like question answering, text summarization and other interesting linguistic tasks. The first version (GPT-1) [45] was released in 2018, the second (GPT-2) [46] in 2019 and the third (GPT-3) [47] in May 2020. At the time of writing, the latest versions underlying the powerful ChatGPT chatbot are not open source. After an unsupervised training stage on a huge corpus of documents, the GPT architecture can be used for various downstream tasks. In particular, GPT can be applied with either a small supervised training of a few examples or without, as in Zero-Shot setting. In both cases, GPT outperforms very often state-of-the-art models trained in a supervised way [47].

Built on the Transformer's Attention mechanism, GPT can manage *long-term dependencies* in text. Despite limitations like unidirectionality, its transfer learning capabilities make it a breakthrough in NLG. GPT is a task-agnostic model that, with a simple set-up and minimal changes to the model architecture, can perform non-trivial tasks via transfer learning (GPT-3 performs very well also without any fine-tuning). In general, transfer learning is so effective that in some problems, such as commonsense reasoning, question answering or textual entailment, GPT manages to obtain results far superior to models with task-specific architectures.

GPT-2, shown in Fig. 1, comprises $N = 12$ identical Decoder layers. It employs an auto-regressive scheme and lacks a second layer of Self-Attention, making it unidirectional.

The pre-training objective is

$$L_1(U) = \sum_i \log(P(u_i|u_{i-k}, \ldots, u_{i-1}; \Theta)), \qquad (1)$$

where U is the corpus of tokens, k is the context window size, and Θ are the model parameters.

Specifically, the input of GPT-2 consists of a set of vectors – within a suitable context window – obtained by a word-embedding transformation of tokens to which is applied a positional encoding scheme – a signal that indicates the order of the words in the sequence to the Transformer blocks. The embedding size can vary depending on the largeness of the model (768 - small, 1,024 - medium, 1,280 - large, 1,600 - extra large). The context window is generally 1,024, as for GPT-2. Hence, part of the trained model is a matrix containing a positional encoding vector for each of the 1,024 positions in the (sentence) input. Then, input tokens are processed by first passing them through the self-attention process and then passing them through its neural network layer (self-attention and neural network layer constitute the Transformer block – see next section). The output is then passed through the next Transformer blocks, each of which has its weights in both self-attention and the neural network sublayers. The output of the stack of Transformers is processed by a linear transformation followed by a softmax layer. After the pre-training on the objective expressed in (1), supervised training can be carried out on a labeled data set, to adapt the parameters of the model to the application task.

GPT-2 uses a data set called *WebText* [46] collected by the OpenAI researchers. WebText consists of texts belonging to about 45 M websites which have been carefully selected based on the content. The result is a data set containing more than 8 M documents, i.e., 40 GB of text (with a learned vocabulary of more than 50,257 words). In fact, if GPT-1 was trained on a corpus of about 7,000 books, GPT-3 is trained on 570 GB of heterogeneous text. The various GPT versions also differ in the number of trainable parameters. GPT-1 has 17 M parameters, the biggest GPT-2 model has a total of 1.5B parameters while GPT-3 has 175B parameters. Researchers have shown that by increasing the number of parameters by about an order of magnitude (without substantially changing the architecture as in GPT-2 and GPT-3) the linguistic capabilities of the language model improve not only quantitatively but also qualitatively, especially in the Zero-Shot case.

A. The Transformer and the Self-Attention Mechanism

The Transformer is a neural network architecture that became famous in 2017 [4] with Bidirectional Encoder Representations from Transformers (BERT). The transformer outperforms RNNs in NLP and multi-modal tasks. Multi-headed attention for contextualization is a computationally intensive task, but the computational burden is mitigated by the support for parallel training by employing GPUs or TPUs. The multi-headed attention mechanism dynamically assigns a weight to every pair of words in the sequence. The weight indicates how much the model should "pay attention to" the first word when computing the representation of the second one. Transformers can adopt

multiple attention heads (in parallel) and each one can potentially capture a completely different word–word relation [18]. The self-attention mechanism is mainly composed of a suitable dot product strategy, consisting of the following quantities:

- $\vec{q}$ and $\vec{k}$, denoting vectors of dimension d_k, containing the queries and keys, respectively;
- $\vec{v}$ denoting a vector of dimension d_v, containing the values;
- $\vec{Q}$, $\vec{K}$, and $\vec{V}$ denoting matrices collecting together sets of queries, keys, and values, respectively;
- $\vec{W}^Q$, $\vec{W}^K$, and $\vec{W}^W$ are projection matrices used in generating different subspace representations of the query, key, and value matrices;
- $\vec{W}^0$, denoting a projection matrix for the multi-head output.

The attention can be considered a mapping between a query and a set of key-value pairs to an output. Within the general attention setting (e.g., adopted in machine translation) the self-attention captures the relationships between the different elements (in this case, the words) of the same sentence.

The Transformer implements a scaled dot-product attention [4] involving the dot product of each query $\vec{q}$ with all of the keys $\vec{k}$. Subsequently, each result is divided by a constant factor and passed to a softmax function. In matrix form, the attention mechanism is computed as

$$\text{attention}(\vec{Q}, \vec{K}, \vec{V}) = \text{softmax}\left(\frac{\vec{Q}\vec{K}^T}{\sqrt{d_k}}\right)\vec{V}. \tag{2}$$

The scaling factor $\sqrt{d_k}^{-1}$ counteracts the softmax function alleviating the well-known vanishing gradients problem [4].

The single-head attention scheme can be replicated linearly projecting the queries, keys and values h times, using a different learned projection each time. The rationale behind multi-head attention – in spite of single-head – is to allow the attention function to extract information from different representation subspaces. It is worth noting that the attention mechanism in Language Models (LLMs) is generally masked, meaning that it prevents the model from peeking into future tokens in the sequence, thereby ensuring a causal or autoregressive generation of text.

Multi-head attention is a powerful tool for incorporating highly hierarchical features of natural language (syntactic-grammar rules, long-range correlations, etc.), as deeply shown in [18], [44] from a linguistic perspective. Moreover, in the current research, we ground on the hypothesis that the self-attention mechanism herein briefly illustrated is the cause of the complex structures that can be found in machine-generated texts when using the GPT-X model. This hypothesis is strengthened by some studies, such as in [19], where the structure of the self-attention mechanism is investigated on three levels of granularity (the attention-head level, the model level, and the neuron level) through a suitable visualization methodology. The study wonders how attention in GPT-2 captures long-distance relationships versus short-distance ones, attributing this phenomenon to the deep layers within the hierarchical organization. Our investigation will focus on GPT-2 language model.

IV. Complexity Indices for Text Modelling

As regards investigations that directly concern the processing of a time series (MFDFA, RQA and Approximate Entropy), the texts were coded following the methodology illustrated in [36] (or in [30] where, instead of the sequence of words, the authors adopted the sequence of POS-tags), i.e., mapping each word with the index (rank) relating to the list containing the words ordered according to their frequency of appearance. Hence, the most frequent word has index $r = 1$, the second in the list is given $r = 2$, and so on. In other words, by means of Zipf's analysis, each word in the original text can be replaced by its corresponding index r. Then, at position t starting from the beginning of the text, we have the corresponding index $r(t)$ [36].

A. Zipf's Laws for Words

Zipf's law for word frequencies is one of the best-known statistical regularities of language [48]. The law states that, in a statistically significant long text, if one calculates the frequency of each word and then sorts such words according to their respective frequency, there is a power-law relationship between the frequency of the word and its rank. The law is often associated with the principle of least effort, that is, language evolves in a way that minimizes the overall work spent in communication, balancing the effort between speakers and listeners [49]. Formally, let $n(r)$ be the frequency of the rth most frequent word, the Zipf's law reads as

$$n(r) \propto \frac{1}{r^\beta}, \tag{3}$$

with β being a constant, which has been empirically demonstrated to be close to 1 for many human languages, although there can be variations [50]. It is worth noting that also thanks to the availability of (long) texts in electronic format in recent years, the Zipf's law has been studied in specific contexts by connecting cognitive and developmental aspects of children to the variation of its exponent – see [51] for more details. In the current study, the β parameter is estimated with a linear regression after a log-log scale transformation.

B. Multifractal Analysis and Long-Range Correlations

MFDFA [23] is a powerful tool for assessing the complexity of a non-stationary time series from the perspective of fractal behavior. Given a time series $X = \{x_k\}_{k=1}^N$ consisting in N time-samples fulfilling suitable properties [23], the objective is to obtain a reliable qth order fluctuation function $F_q(s) \sim s^{h(q)}$, that is, if X is long-range power-law correlated it will increase, for large values of the time scales s, as power-law. The first step of the procedure consists of obtaining the "profile"

$$Y(i) \equiv \sum_{k=1}^{i}(x_k - \langle x \rangle), i = 1, \ldots, N, \tag{4}$$

where $\langle x \rangle$ is the mean of X. Then the profile is divided into a series of non-overlapping segments where for each segment a polynomial trend is estimated and then it is subtracted from the

profile portion. Linear, quadratic, cubic, or higher-order polynomials can be used, yielding to high-order polynomial detrending procedures (MFDFA1, MFDFA2, etc.). For each scale s and for each segment, ν the variance $F^2(s,\nu)$ is computed. Finally the qth order fluctuation function $F_q(s)$ is obtained through the Hölder mean of parameter q, computed by averaging all detrended segments, that is

$$F_q(s) \equiv \left\{ \frac{1}{2N_s} \sum_{\nu=1}^{2N_s} \left[F^2(s,v)\right]^{q/2} \right\}^{1/q}. \tag{5}$$

For $q = 2$ the standard DFA is recovered. The interest is in how the generalized q-dependent fluctuation $F_q(s)$ depends on the time scale s for different values of q. The fluctuation function $F_q(s)$, which depends on the DFA order m, will increase with increasing s. Finally, the scaling behavior of the fluctuation functions is assessed by analyzing the log-log plots $F_q(s) \sim s^{h(q)}$ versus s. The exponent $h(q)$ may depend on q. The quantity $h(q)$ is known as the generalized Hurst exponent and for stationary time series $h(2)$ is identical to the Hurst exponent H [23]. If the generalized Hurst exponent $h(q)$ is found independent of q, the time series is monofractal, i.e., it shows a uniform scaling over all magnitude scales of the fluctuations. Conversely, it is multifractal when $h(q)$ depends appreciably on q, so that small fluctuations scale differently from large ones. For stationary processes, long-term memory properties can be safely described by the power-law like decreasing of the autocorrelation summarized by the value of the so-called Hurst exponent $\mathrm{H} = h(2)$. Depending on the value of H, a signal can be classified as correlated or persistent, i.e., it has long memory if $0.5 < \mathrm{H} \leq 1$, while it is considered anticorrelated or antipersistent, i.e., it has short memory, if $0 < \mathrm{H} < 0.5$. The case with $\mathrm{H} = 0.5$ denotes uncorrelated white noise. Monofractal signals are homogeneous because they have the same scaling properties, while multifractal ones require an infinite number of indices to characterize their scaling behavior.

The generalized Hurst exponent is related to some classical multifractal indices [52]. Specifically, $h(q)$ is directly related to the classical multifractal scaling exponents, also called Rényi scaling exponent $\tau(q)$, by the relation $\tau(q) = qh(q) - 1$, where $\tau(2)$ is the correlation dimension. A compact and useful way to express the multifractal characteristic of a time series is the multifractal spectrum $f(\widetilde{\alpha})$ computed by the Legendre transform of $\tau(q)$, that is $f(\widetilde{\alpha}) = q\widetilde{\alpha} - \tau(q)$, where $\widetilde{\alpha}$, called singularity strength or Hölder exponent, is equal to $\widetilde{\alpha} = \frac{d\tau(q)}{dq}$. Finally, through the last formula, we find the following relations for the multifractal spectrum:

$$\widetilde{\alpha} = h(q) + q\frac{d\tau(q)}{dq} \text{ and } f(\widetilde{\alpha}) = q[\widetilde{\alpha} - h(q)] + 1. \tag{6}$$

A concise way of describing the multifractal signature of a complex system consists of analyzing the shape characteristics of some multifractal function, such as the Rényi scaling exponent $\tau(q)$, the generalized Hurst exponent $h(q)$, as functions of the qth order fluctuation and the multifractal spectrum $f(\widetilde{\alpha})$, as function of the Hölder exponent $\widetilde{\alpha}$. If for multifractal time series obtained from a pure analytical model, such as the well-known binomial cascade, the spectrum is a near-perfect symmetric parabola, for real-world time series the spectrum can be asymmetric and distorted by some degree.

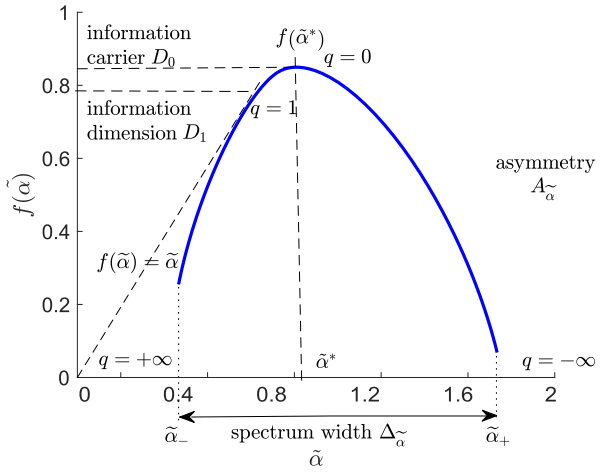


Fig. 2. Schematic presentation of the main parameters of a multifractal spectrum.

The richness of the multifractal properties is obtained through the so-called multifractal strength of the spectrum $\Delta_{\widetilde{\alpha}} = \widetilde{\alpha}_+ - \widetilde{\alpha}_-$, where $\widetilde{\alpha}_+$ and $\widetilde{\alpha}_-$ are the two extreme values at two ends of the multifractal spectrum support, respectively. The more $\Delta_{\widetilde{\alpha}}$ is elevated, the more the multifractality characteristic is prominent. Monofractal signals see a near-linear $h(q)$, while the multifractal ones are characterized by a typical s-shaped form, hence a nonlinear $h(q)$; theoretically, this fact is translated in a wide multifractal spectrum in the first case, while in the latter one it reduces to a single point. Furthermore, finite size effects and other joined phenomena can corrupt the spectrum. It is worth noting that the multifractal strength $\Delta_{\widetilde{\alpha}}$ is related to the degree of difference between $h(+\infty)$ and $h(-\infty)$. Nevertheless, the asymmetry in the shape of the multifractal spectrum can indicate the presence of a particular system [53] – such as in the case of asymmetric binomial cascade, with a peculiar dynamical behavior – or noise corruption on smaller scales (filtered out by $q > 0$ – left asymmetry) or higher scales (filtered out by $q < 0$ – right asymmetry). The asymmetry can be evaluated through a suitable asymmetry index $A_{\widetilde{\alpha}}$ [54], that reads as

$$A_{\widetilde{\alpha}} = \left(\Delta\widetilde{\alpha}_L - \Delta\widetilde{\alpha}_R\right) / \left(\Delta\widetilde{\alpha}_L + \Delta\widetilde{\alpha}_R\right), \tag{7}$$

where $\Delta\widetilde{\alpha}_L = \widetilde{\alpha}^* - \widetilde{\alpha}_-$ and $\Delta\widetilde{\alpha}_R = \widetilde{\alpha}_+ - \widetilde{\alpha}^*$, while $\widetilde{\alpha}^*$ is the $\widetilde{\alpha}$ value at maximum of $f(\widetilde{\alpha})$ (which corresponds to $q = 0$), i.e., the box counting dimension, and $\widetilde{\alpha}_-$ and $\widetilde{\alpha}_+$ denote the beginning and the end of $f(\widetilde{\alpha})$ support – as depicted in Fig. 2.

One way to delve into the analysis of multifractal systems is to distinguish from multifractality due to the broadness of the Probability Density Function (PDF) and multifractality due to the different correlations in small and large-scale fluctuations. This is done by evaluating two time series derived from the original one, namely the shuffled time series and the surrogate one [55]. The former is computed simply by shuffling at random the time indices, while the second is obtained by changing the

phases, computed through the Discrete Fourier Transform (DFT) of the original signal, drawing from a uniform distribution in $(-\pi, \pi)$. In this case, it can be demonstrated that the PDF tends to be normally distributed but correlations do not change. The shuffling procedure will destroy all long-range correlations and the corresponding shuffled time series will exhibit monofractal scaling. Conversely, the multifractality due to the fatness of the PDF signals is not affected by the shuffling procedure. Thus, since the shuffling of time series destroys the long-range correlation, that is if the multifractality belongs only to the long-range correlation, a constant value $h_{\text{shuffle}} = 0.5$ should be found. If both types of multifractality are present, the shuffled and the surrogate series will show weaker multifractality than the original series and this can be assessed by examining, for example, the multifractal spectrum $f(\widetilde{\alpha})$.

C. Recurrence Quantification Analysis

Recurrence Plots (RPs), introduced by Eckmann et al. in [56], are useful tools for describing the recurrence property of a deterministic dynamical system. RPs allow visualizing the time-dependent behavior of orbits x_i in phase space. RPs are the key elements of RQA, which allows to broadly characterize dynamical systems. The main steps of the RQA are: i) the reconstruction of the entire phase space of a time series, ii) the generation of the RP matrix, iii) the estimation of suitable descriptive indices from this matrix.

The principal step for obtaining a RP is to calculate the following $N \times N$ recurrence matrix [57]:

$$R_{i,j}(\epsilon) = \Theta(\epsilon - \|x_i - x_j\|_2), i, j, = 1, \ldots, N, \tag{8}$$

where $N = L - (m-1)\tau$, ϵ is a predefined cutoff distance, $\| \cdot \|_2$ is the Euclidean norm and $\Theta(x)$ is the Heaviside function. The phase space vector x_i can be reconstructed using Takens' time delay method, $x_i = (u_i, u_{i+\tau}, \ldots, u_{i+(m-1)\tau})$ [58], grounded on the observations u_i. The threshold ϵ defines a sphere centered at x_j: if x_i falls within this sphere, that is the state is close to x_j, then $R_{i,j} = 1$, otherwise $R_{i,j} = 0$. The RP consists in visualizing the binary matrix R (in black and white), while the non-thresholded matrix can be visualized as a colored heatmap. In this way, the RP is an instrument for the inspection of a high-dimensional phase space trajectory, that is its time evolution. On the other hand, RPs can describe the characteristics of large-scale and small-scale patterns of a dynamical system, starting from short and non-stationary data.

Concerning the visual inspection of RPs, series that are deterministic show the existence of short line segments parallel to the main diagonal. The diagonal lines represent segments of the phase space trajectory that run parallel for some time, instead, the vertical lines represent segments that remain in the same phase space region for some time. In general, a graph showing small random spots (even dot-sized) is related to random noise, while a random-walk-like noise presents larger randomly placed spots. A RP with a regular texture can come from near-periodic and more deterministic series – more details in [59].

The quantitative analysis of RP, that is the RQA, is grounded on several measure variables. In this work, we adopt five indices: recurrence rate (RR), determinism (DET), entropy (ENTR), the averaged diagonal line length (LEN), laminarity (LAM) and trapping time (TT).

The recurrence rate is defined as

$$\text{RR}(\epsilon) = \frac{1}{N^2} \sum_{i,j}^{N} R_{i,j}(\epsilon), \tag{9}$$

that is, it simply counts the black dots in the RP.

The frequency distribution (i.e., the histogram) of the lengths l of the diagonal structures in the RP reads as $P^{\epsilon}(l)\{l_i; i = 1, 2, \ldots, N\}$. The determinism (or predictability) measure of the system (DET) is the ratio of recurrence points on the diagonal structures (of at least length l_{min}) to all recurrence points, that is

$$\text{DET} = \frac{\sum_{l=l_{\text{min}}}^{N} P^{\epsilon}(l)}{\sum_{l=1}^{N} P^{\epsilon}(l)}, \tag{10}$$

where l_{min} is the threshold parameter, which excludes the diagonal lines formed by the tangential motion of a phase space trajectory.

ENTR refers to the Shannon entropy of the frequency distribution of the diagonal line lengths

$$\text{ENTR} = -\sum_{l=l_{\text{min}}}^{N} p(l) \log(p(l)), \tag{11}$$

where $p(l) = \frac{P(l)}{\sum_{l=l_{\text{min}}}^{N} P(l)}$. The ENTR index is a complexity measure of the deterministic structure in a dynamical system. The more complex the deterministic structure, the larger the ENTR value.

The average length of diagonal lines is computed as

$$\text{LEN} = \frac{\sum_{l=l_{\text{min}}}^{N} lP(l)}{\sum_{l=l_{\text{min}}}^{N} P(l)}. \tag{12}$$

Moreover, the laminarity – i.e., the percentage of recurrence points that form vertical lines – reads as

$$\text{LAM} = \frac{\sum_{\nu=\nu_{\text{min}}}^{N} \nu P(\nu)}{\sum_{\nu=1}^{N} \nu P(\nu)}, \tag{13}$$

where $P(\nu)$ is the frequency distribution of the lengths ν of the vertical lines.

Finally, the average length of vertical lines, i.e., the trapping time, is

$$\text{TT} = \frac{\sum_{\nu=\nu_{\text{min}}}^{N} \nu P(\nu)}{\sum_{\nu=\nu_{\text{min}}} P(\nu)}. \tag{14}$$

The TT variable measures the mean time the system will abide at a specific state [59]. The important quantities to estimate are the embedding dimension and the time delay. In this work, the former is computed through the well-known False Nearest Neighbor method [60], while the latter is obtained by inspecting the first local minimum of the Mutual Information diagram of the time series. Another important parameter specific to the RP is the cutoff distance ϵ. Among the numerous methods to estimate this parameter, in this work the one described in [61] is used, which is based on the inspection of the graph RR varying ϵ.

The diagram exhibits a sigmoid curve (the density of recurrence points increases till a saturation zone) and ϵ must be kept i) low and ii) in the linear scaling region of the RR-ϵ diagram.

D. Approximate Entropy

ApEn is a statistical index used to quantify the predictability, hence the regularity, of a time series. Let $\mathbf{u} = \{u_1, u_2, \ldots, u_N\}$ be an N-length time series, let r be a positive real number and let $m \leq N$ be a non-negative integer. Further, let us define two slices $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ as two m-length windows of $\mathbf{u}$, namely $\mathbf{x}^{(i)} = \{u_i, u_{i+1}, \ldots, u_{i+m-1}\}$ and $\mathbf{x}^{(j)} = \{u_j, u_{j+1}, \ldots, u_{j+m-1}\}$. Given any two slices $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$, let $d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ be a distance measure defined as

$$d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \max_{k=1,\ldots,m} \left(\|\mathbf{x}_k^{(i)} - \mathbf{x}_k^{(j)}\|_1 \right), \tag{15}$$

namely, the maximum absolute difference between any two homologous entries in the two slices. Then, we calculate the value $C_i^m(r)$ as

$$C_i^m(r) = \frac{|\{j : d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \leq r\}|}{N - m + 1}, \quad 1 \leq i, j \leq N - m + 1, \tag{16}$$

namely, the numerator of $C_i^m(r)$ counts the number of slices of consecutive values of length m which are similar to a given slice, within a given resolution r. Finally, the ApEn index reads as

$$\text{ApEn}(m, r, N, \mathbf{u}) = \phi^m(r) - \phi^{m+1}(r), \tag{17}$$

where

$$\phi^m(r) = \frac{1}{n} \sum_{i=1}^{N-m+1} \log\left(C_i^m(r)\right). \tag{18}$$

It can be shown that ApEn $\to 0$ corresponds to a regular and predictable sequence.

V. Corpora

Three main corpora have been used for the experimental stage. For humans-generated texts, 80 classic novels written in English have been retrieved from the Gutenberg Project website,[4] with an average of 123473.3 tokens per novel. Instead, for English language texts generated by the machine, the original GPT-2 architecture[5] is adopted [46], allowing to collect 80 text excerpts with an average of 17028.3 tokens per text. In particular, the Tensorflow GPT-2-124 M model is wrapped within a filtering procedure designed suitably for filtering out too noisy, repetitive and very low-quality text excerpts. Especially with low values of the temperature parameter, GPT-2 can get trapped in a recurring state, repeating the same word indefinitely. During the text generation, the procedure performs a series of quality checks on the text excerpts, rejecting too noisy ones (e.g., with a higher density of non-common characters) or too repetitive ones. For GPT-2 the dimension of the vocabulary is 50,257 words, the context window is 1,024 tokens, the embedding dimension is 768, and the number of multi-head self-attention blocks is 12, like the number of decoder stacks. Particularly, for the current experiments, 20 text excerpts (of at least 15 k tokens) for each temperature parameter in the set $\mathrm{T} = \{0.7, 0.8, 0.9, 1.0\}$ are generated. In this study, we maintained fixed the output sampling parameters Top-P $= 1$[6] and Top-K $= 40$[7] [62] - an example of text samples generated by the machine is reported in Fig. 3. Comparing the two texts in the figure it can be seen that, concerning the temperature parameter $\mathrm{T} = 1.0$, the text excerpts obtained with a temperature $\mathrm{T} = 0.7$ are more repetitive with consequent degradation of the carried meaning. For the computer programs – i.e., texts written by humans but following a strict syntax – we collected 52 `C`-files pertaining to the Linux kernel[8] spanning from all the available versions, with an average of 3265.2 tokens per file. Specifically, the Linux corpus consists of several versions of the `fork.c`, `time.c`, `ptrace.c`, `sys.c` and `exit.c` files. As regards text preprocessing, for homogeneity between the different corpora, we have chosen to minimize these operations by retaining capital letters and numbers whilst removing punctuation and any non-standard character. We also chose a simple tokenizer that divides the text into tokens according to white spaces.

VI. Experiments

As mentioned in the Introduction, the experimental setting consists of an analysis phase – offered in next Section VI-A – and a synthesis phase – reported in Section VI-B. In the analysis phase, the time series – obtained by encoding the words of the text on the basis of the ranking underlying the Zipf's law as explained in Section IV – are subjected to the measurement of the complexity indices through the i) MFDFA, ii) the RQA, iii) the estimation of the coefficient β of the Zipf's law, iv) and the approximate entropy (ApEn) – see Section IV. Each framework will describe the underlying stochastic process from a different perspective. Therefore, both a detail on the estimated values through group analysis and by means of the Multivariate Analysis of Variance (MANOVA) statistical framework will be provided. We remark that the main aim, in this preliminary stage, is to x-ray some aspects of the complex behavior of a text generated by the GPT-2 architecture (e.g., the long-range correlations, the recurrence, the predictability, etc.) in comparison with texts written by human beings (English language novels and programming codes). The ultimate goal of the comparative analysis is to underline similarities and differences in the various cases examined. The second phase, which we call synthesis, instead proposes to investigate the informativeness of the complexity indices in solving a three-class classification problem with a suitable learning algorithm based on an evolutionary heuristic.

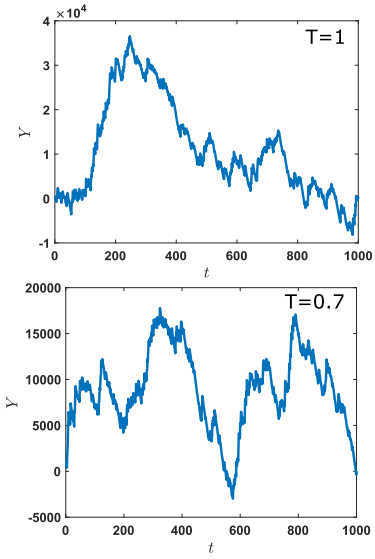
[4] https://www.gutenberg.org/

[5] https://github.com/openai/gpt-2

[6] Top-P, also called *nucleus sampling:* the next word is selected randomly based on the probability distribution conditioned by the previous word among the set of words that add a probability greater than or equal to P.

[7] In the output text, the next word is selected randomly based on the probability distribution conditioned by the preceding K words with the highest probability.

[8] Available at http://ftp.riken.jp/Linux/kernel.org/linux/kernel/.



GPT-2 generated text sample - Temperature 1

The origin of the modern world Heller believed in the supernatural. As an atheist he believed in supernatural explanations like that which have never been offered to On top of that he believed with great respect in the idea of a supernatural God (called Theism) claiming to have created it. This view of creationism in general and Theism in particular was supported by the writings of the German scientist Rudolf Hegel As 1785, Hegel considered creationism to be the main cause of the Protestant Reformation.

GPT-2 generated text sample - Temperature 0.7

The Nobel Prizes for Physics are awarded to those who have completed the most research or have led the most successful research projects in the last 12 years. They are given to the most distinguished scientists in the world, and to those who have made the most progress in solving problems in the last 12 years. The Nobel Prizes are awarded to those who have made the most progress in solving problems in the last 12 years. They are given to the most distinguished scientists in the world, and to those who have made the most progress in solving problems in the last 12 years. The Nobel Prize has been awarded to a Nobel laureate every three years since 1905, and every four years since 1910.

Fig. 3. Example of profile Y and text samples generated with GPT-2. Temperature T = 1 (upper panel), temperature T = 0.7 (lower panel).

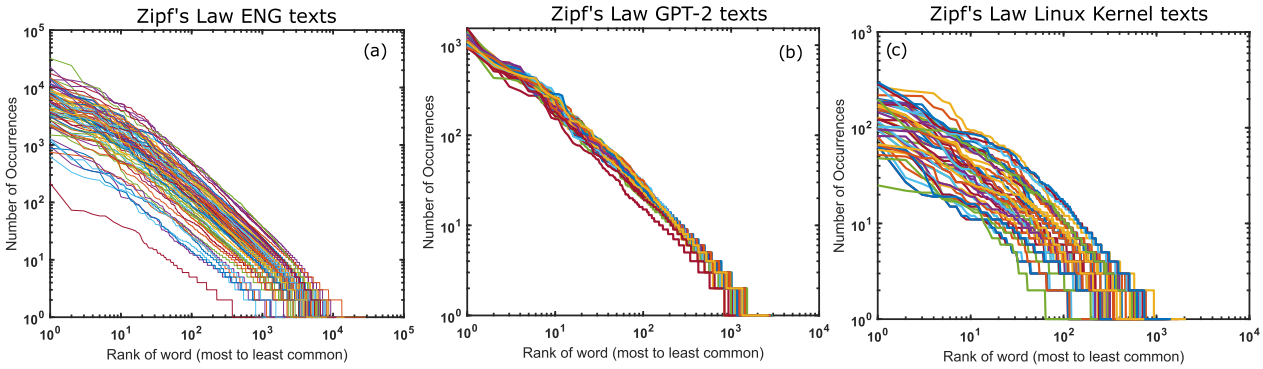


Fig. 4. Zipf's Law for a) English novels, b) GPT-2 texts, c) Linux source files.

A. Analysis

Zipf's Law: As well known, the Zipf's law has been extensively studied in quantitative linguistics in a large and typologically diverse ensemble of languages. In our study we compare this law considering the texts produced by human brains (English novels and programming codes – labelled ENG and LINUX, respectively, in the following) and texts produced by GPT-2 – labelled simply GPT-2 hereinafter. In Fig. 4 it is possible to note the trend of the law in the classic log-log plot (logarithm of rank order versus logarithm of frequency) for the three considered textual classes. We can see more variability in the ENG and LINUX cases than in GPT-2, an expected result since although there is variability in the temperature parameter for GPT-2, the latter can be considered as a single "author" compared to the other cases. In other words, at least in these experiments, the texts produced by GPT-2 appear more homogeneous even when the temperature parameter is varied. It should be noted that the estimation of the distribution parameters (intercept b and slope β – see Section IV-A) is carried out in the rectilinear area, i.e., in the range $[10^1, 10^3]$ for the rank variable. Evaluating the box-and-whisker diagram in the left panel of Fig. 14, we note an increasing trend of the (negative) value of β and the texts produced by GPT-2 are placed in-between the other two classes, with very compact values, slightly higher than the ENG case.

Multifractal Detrended Fluctuation Analysis: For these experiments, detrending is performed by a third-order polynomial (MFDFA3) while the qth order exponent is chosen in a wide range of values, that is ± 10. In Figs. 5, 6, and 7 are reported the singularity spectra (panels c)) and the qth order fluctuation function $F_q(s)$, $q = 2$, (panels d)) – see Section IV-B – for a single sample of the three families of text analyzed, that are the novel "Mrs Dalloway in Bond Street" by Virginia Woolf, a long text generated by GPT-2 setting the temperature to 0.7, a `C`-file (`sys.c`) extracted from the Linux kernel (ver. 2.6.0), respectively. The singularity spectrum and the main fractal indices $(\alpha_+, \alpha_-, \alpha^*, \Delta_\alpha, A_\alpha, \mathrm{H})$ are computed for the original series, the shuffled and surrogate ones. For the sake of brevity, these results are not reported for all 212 analyzed texts. The inserts c') and d') in the three above-mentioned figures depict - - for completeness – the singularity spectra (of the original series) and the qth order fluctuation functions $F_q(s)$, $q = 2$, for all samples.

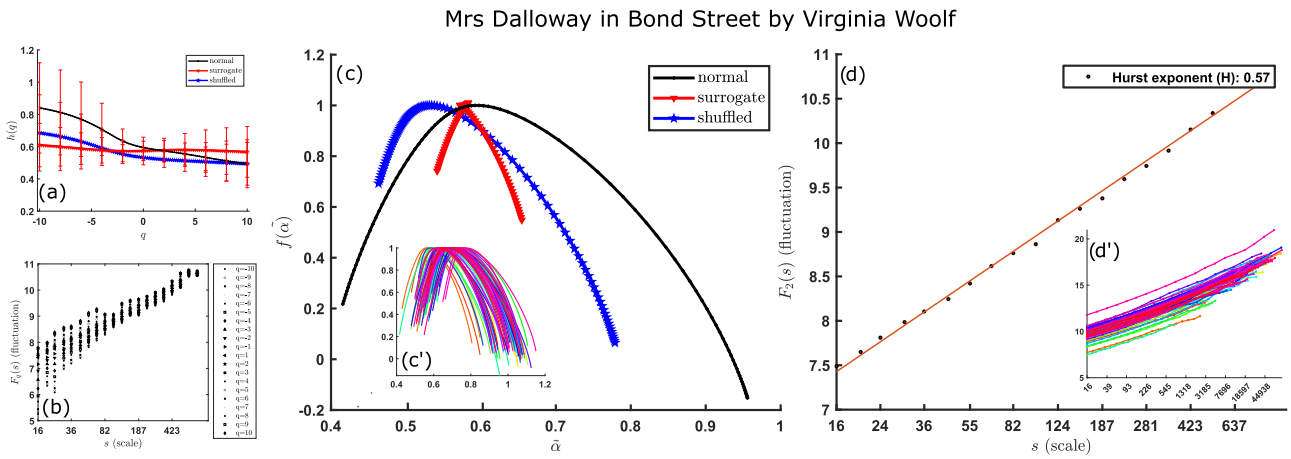


Fig. 5. MFDFA for "Mrs Dalloway in Bond Street" by Virginia Woolf. a) The generalized Hurst exponent $h(q)$ of the original, shuffled and surrogate time series. b) The fluctuation function $F_q(s)$ parameterized by the qth order exponent. c) Singularity spectrum $f(\tilde{\alpha})$ of the original, shuffled and surrogate time series. c') Singularity spectra for overall English novels. d) Fluctuation plot $F_2(s)$ and Hurts exponent H. d') Fluctuation plot for overall English novels.

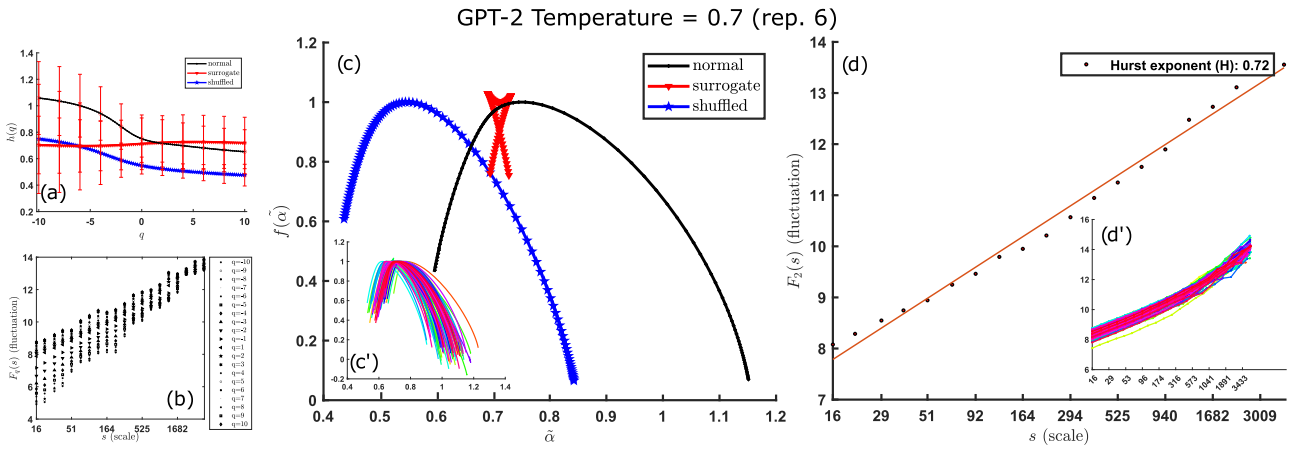


Fig. 6. MFDFA for GPT-2 generated text (temperature 0.7, Top-P=1, Top-K=40). a) The generalized Hurst exponent $h(q)$ of the original, shuffled and surrogate time series. b) The fluctuation function $F_q(s)$ parameterized by the qth order exponent. c) Singularity spectrum $f(\tilde{\alpha})$ of the original, shuffled and surrogate time series. c') Singularity spectra for overall GPT-2 texts. d) Fluctuation plot $F_2(s)$ and Hurts exponent H. d') Fluctuation plot for overall GPT-2 texts.

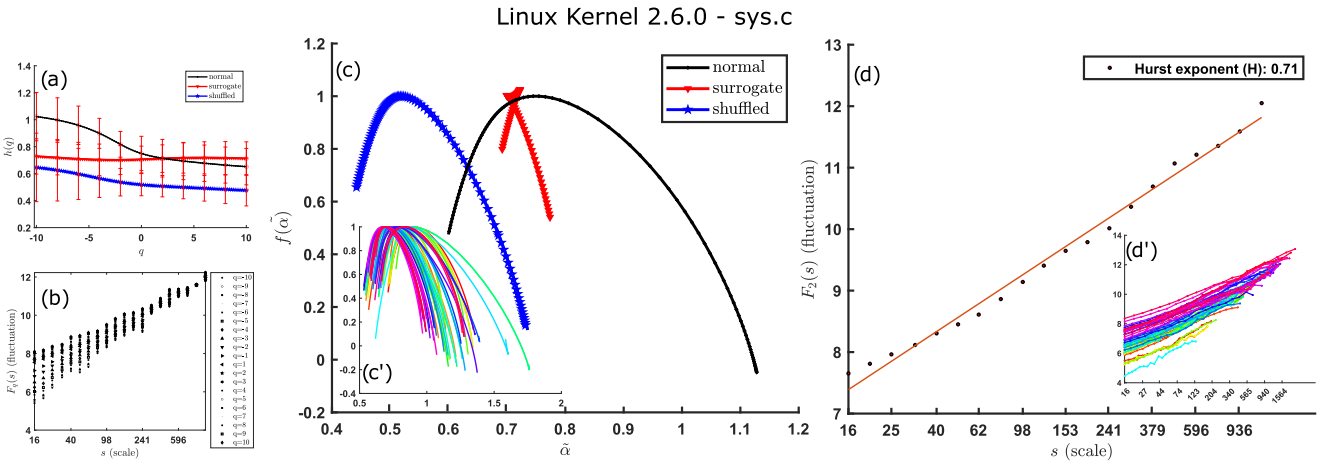


Fig. 7. MFDFA for the Linux kernel file `sys.c`. a) The generalized Hurst exponent $h(q)$ of the original, shuffled and surrogate time series. b) The fluctuation function $F_q(s)$ parameterized by the qth order exponent. c) Singularity spectrum $f(\tilde{\alpha})$ of the original, shuffled and surrogate time series. c') Singularity spectra for overall selected Linux kernel files. d) Fluctuation plot $F_2(s)$ and Hurts exponent H. d') Fluctuation plot for overall selected Linux kernel files.

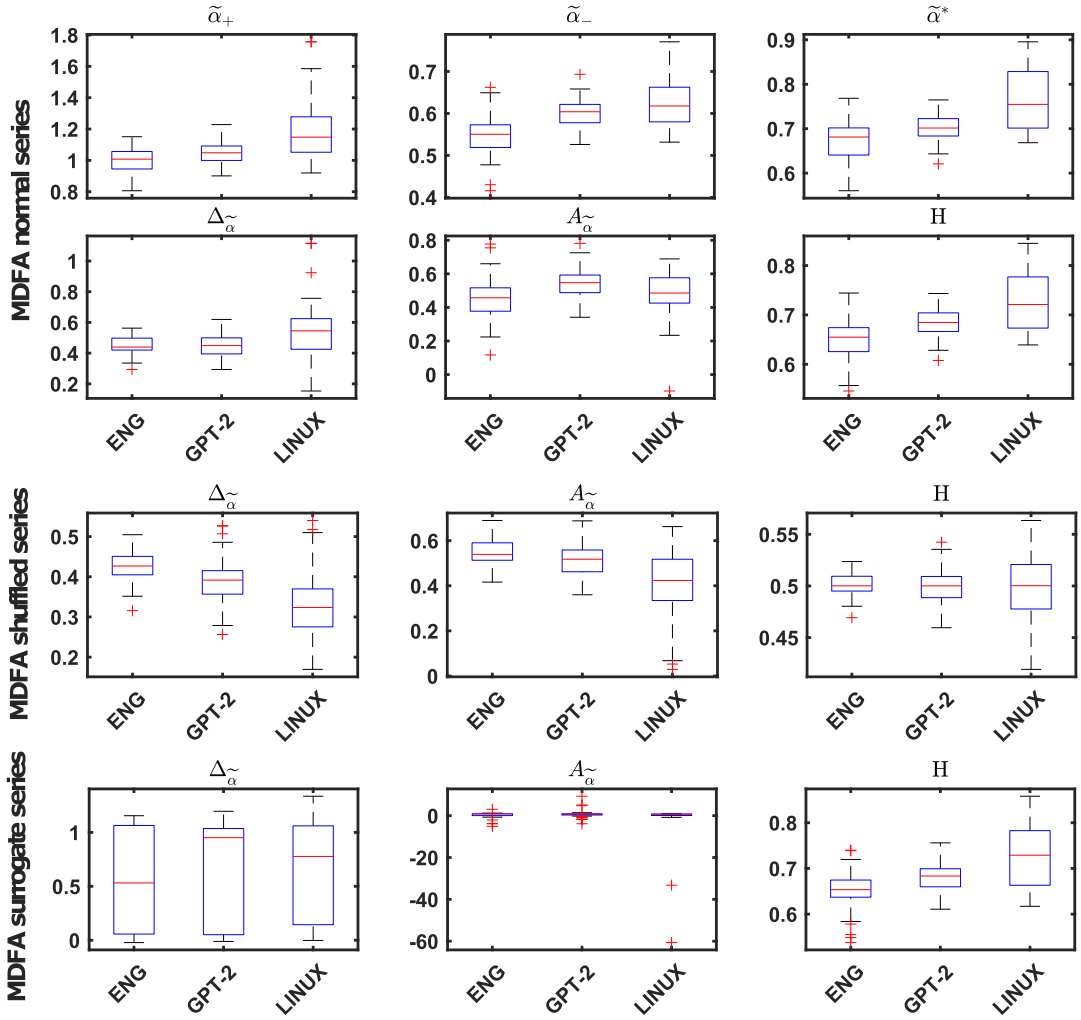


Fig. 8. Box-and-whisker plots of the main MFDFA indices for the original (upper panel), shuffled (middle panel), surrogate (lower panel).

Finally, panels a) and b) of the three figures show the generalized Hurst exponent $h(q)$ of the original, shuffled and surrogate time series (where also here it can be appreciated the multifractality degree) and the fluctuation function $F_q(s)$ parameterized by the qth order exponent. In Fig. 8 is reported the box-and-whisker diagram for each group computed over the main fractal indices mentioned above for the original time series, the shuffled and the surrogate ones (in the last two cases, for brevity, they are considered only the multifractal signature Δ_α, the asymmetry index A_α, and Hurst exponent H).

By comparing the three Figs. 5, 6, and 7 and the respective inserts, it can be seen that all three families of analyzed texts present a form of long-term memory, i.e., long-term correlations (see the singularity spectrum $f(\widetilde{\alpha})$ and the slope of the $F_q(s)$ ($q = 2$) plot). For English texts produced by human beings, this specific result is in line with what is reported in the literature. In our study, it can be appreciated that even for texts generated by a machine such as GPT-2 there are long-term correlations. Specifically, looking at Hurst's monofractal exponent H (Fig. 8) we note that there is an increasing trend among the three text typologies. In particular, novels (ENG) have a lower value (but still higher than the case $H = 0.5$ related to uncorrelated noise) while programming codes (LINUX) show a high degree of long-term correlation. The texts generated by GPT-2 (GPT-2) are placed in an intermediate zone (with a median H slightly higher than 0.7). As artificially-created series are concerned, it can be seen that the correlations are not destroyed by the transformation of the frequency spectrum (surrogate series) while, as expected, they are completely canceled by the random shuffling procedure. This is true for all three analyzed cases. By reviewing the main multifractal indices, also in this case we note a growing trend for the multifractal signature Δ_α, but the difference between ENG and GPT-2 is less marked than in the LINUX case. In any case, these values are much higher than the (theoretical) null value, showing a good degree of multifractality. Random shuffling inverts the trend for the singularity strength, so this operation affects more GPT-2 and LINUX than ENG. This is attributable to the presence of a higher degree of intermittence in the process underlying machine-generated texts or related to programming codes than those produced by human brains. As regards the asymmetry index $A_{\widetilde{\alpha}}$, for all texts, the obtained value shows a pronounced right-side asymmetry (generally a

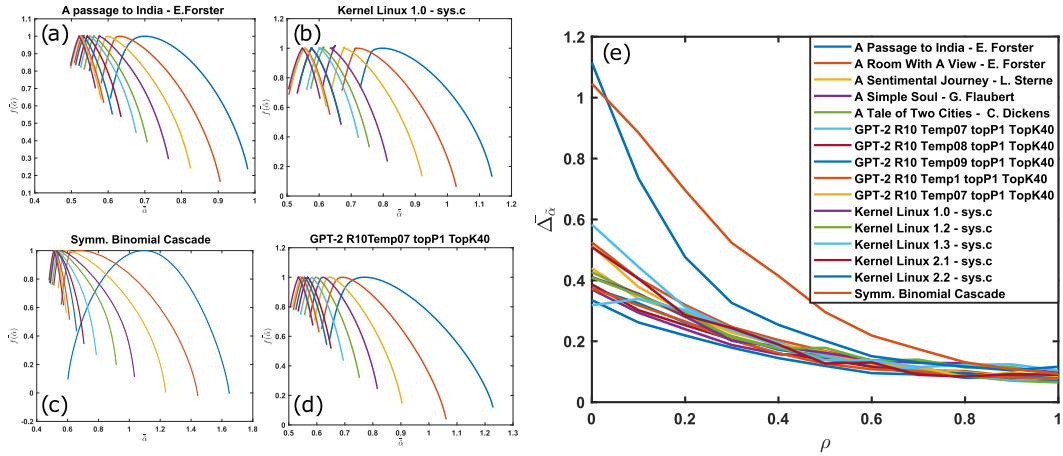


Fig. 9. Robustness analysis by increasing the degradation with incremental (shuffling) noise. a) ENG, b) LINUX, c) Symmetric binomial cascade series, d) GPT-2, e) multifractal signatures for 5 samples per class, including the Symmetric binomial cascade.

rare case [54]), hence a more uniform hierarchical organization on the level of smaller fluctuations and more noise-like behavior of the large fluctuations [54]. The index $A_{\widetilde{\alpha}}$ is zeroed in the case of a surrogate series. This means that the information difference given by the behavior on scales with higher values and scales with lower values cancels out.

As a last experiment, we propose a robustness analysis, specifically on the multifractal signature index Δ_α which is known to measure the richness of the multifractal spectrum $f(\widetilde{\alpha})$ (through the measure of the amplitude of the singularity spectrum). The test is generated through an incremental random shuffling procedure measured by a (normalized) ρ parameter. The first step is to randomly select x indices and shuffle them. Consequently, this number x is increased until obtaining (in case of $\rho = 1$) the completely randomized (shuffled) series. Fig. 9 shows the singularity spectra for the texts belonging to the three families (panels, a), b), d)) and for the typical case of the well-known symmetrical Binomial Cascade (panel c)). Panel e) instead reports the trend of the multifractal signature Δ_α (averaged over 30 repetitions for each ρ value in $[0, 1]$) for 5 samples among each text class including the Binomial Cascade. As the value of ρ increases ($\rho = 1$ means complete shuffling) Δ_α decreases (starting from a moderately high value) up to a minimum value, which is not null due to the finiteness effects of the series under analysis. This proves that the Δ_α values obtained are noteworthy for all analyzed series.

From this first complexity analysis it can be observed that, although GPT-2 is qualitatively capable of producing texts in some cases comparable to those produced by a human being (and this is well known above all for high temperature values), texts produced by human beings are less predictable (at least in terms of Hurst's analysis) and have a less intermittent behavior (as a stochastic process). This is a result that can be expected if we imagine GPT-2 as a semi-deterministic symbol generator system with far fewer degrees of freedom of a brain (if they were comparable), albeit exceptionally large. The following analysis relating to RQA will enrich the picture that is emerging.

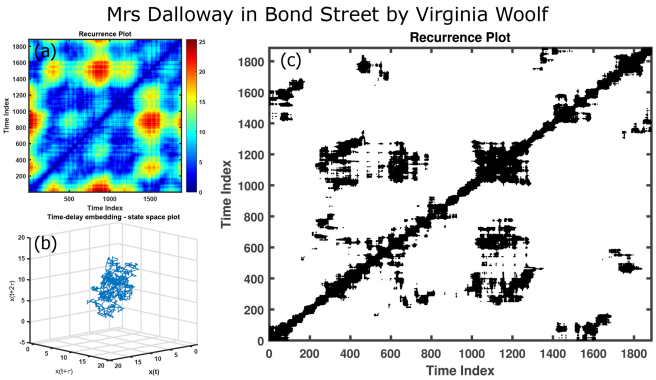


Fig. 10. Recurrence Quantification Analysis for "Mrs Dalloway in Bond Street" by Virginia Woolf. a) Non-thresholded RP, b) state space reconstruction ($\tau = 57, m = 3$), c) RP.

Recurrence Quantification Analysis: In this study, RQA and the RP are obtained by reconstructing the phase space trajectory through the evaluation of Mutual Information (time delay τ) and the method known as False Nearest Neighbors (minimum embedding dimension m). The RQA is performed on the profile Y – (4) – obtained by transforming the time series in a random walk – see Section IV-B. We remark that RP and recurrence quantifications are strongly dependent on the sequential organization of the time series. By contrast, standard statistical measures such as mean and standard deviation are sequence independent [61]. In Figs. 10, 11, and 12 are reported the RP (panel c)), the non-thresholded RP, i.e., the heatmap of the recurrence matrix (panel a)) and the dynamics in the three-dimensional reconstructed phase space, for three text samples pertaining to ENG, GPT-2 and LINUX classes. Visual inspection of the RPs shows patchy areas with many irregularities. This is an expected behavior as we are analyzing a process that behaves like a random walk. Overall, the texts produced with GPT-2 appear to be less irregular than those written by humans. In order to give an all-encompassing look at the dynamic behavior of the series, in Fig. 13 the box-and-whisker diagrams of the quantities

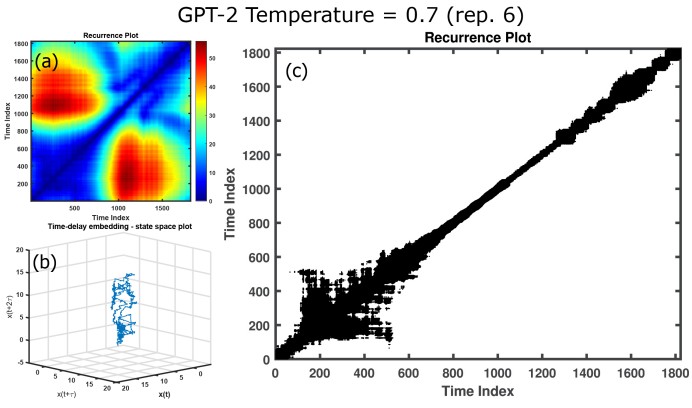


Fig. 11. Recurrence Quantification Analysis for GPT-2 generated text (temperature 0.7, topP $= 1$, topK $= 40$). a) Non-thresholded RP, b) state space reconstruction ($\tau = 54$, $m = 4$), c) RP.

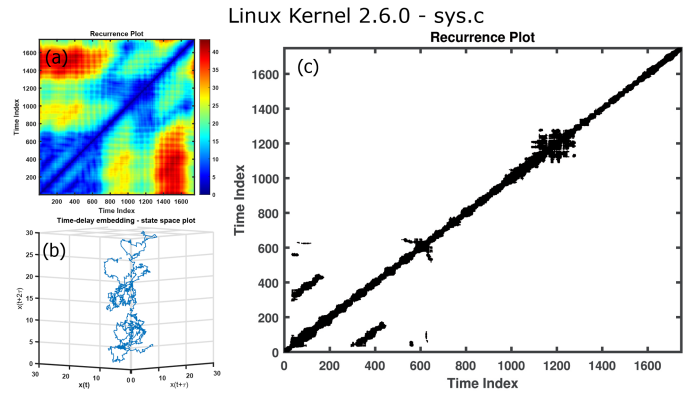


Fig. 12. Recurrence Quantification Analysis for the Linux kernel file `sys.c`. a) Non-thresholded RP, b) state space reconstruction ($\tau = 41$, $m = 5$), c) RP.

computed through the RQA are reported (see also Section IV-C). By analyzing the various RQA measures, it is noted that GPT-2 has a density of recurring points (RR) similar to novels (slightly higher on average) and certainly higher than programming codes. Unlike the average diagonal line length (LEN) – related to trapping time –, which is similar in all three cases, GPT-2 shows higher values than ENG and LINUX for the remaining measures; specifically for the measure of determinism or rule-obeying dynamics (DET), for entropy (ENTR), for laminarity (LAM) and for the laminarity time (TT) – mean time the system will abide at a specific state – of the underlying dynamical system. Hence, GPT-2 shows a more deterministic behavior (DET) with a more complex structure than the other cases (ENTR). Specifically, GPT-2 shows surprisingly higher aperiodicity. At the same time, looking at the LAM measurement, we find that the dynamic underlying GPT-2 has a greater degree of laminarity and therefore a higher degree of intermittence. Furthermore, GPT-2 tends to stay in a default state more often than novels or programming code. Looking specifically at DET and LEN, the underlying dynamical system for all cases can be conceived as a mixture of chaotic and random behavior as expected, even if GPT-2 seems to have a less random behavior. It should be noted that these considerations on the behavior of the three systems are to be considered tendentious since we are analyzing real systems, which can be corrupted by noise with non-smooth dynamics. However, it can be asserted that the texts tend to be similar to random walks with a certain superimposed chaotic dynamic. We hypothesize that the underlying chaotic dynamic is related to a slow change in the dynamic of the semantic content of a text, while the local variability is given by the noise-like behavior, which has a specific correlation structure, keeping in mind that semantically a concept can be expressed in an almost infinite number of ways.

Approximate Entropy: As far as the approximate entropy (described in Section IV-D) is concerned, we do not notice great differences between the three types of text – see the right panel of Fig. 14 –, especially if we evaluate the great variability that exists between the measures. The trend tends to see higher values for novels (low predictability or pattern recurrence), while lower values are collected for programming languages. GPT-2 is placed in a middle position. A joint multivariate analysis of variance (MANOVA) of all the variables analyzed separately will show that the clues obtained by investigating the single variables fit into a common and rational frame of reference.

Multivariate Analysis of Variance: In order to have a general comparative analysis, we propose a multivariate statistical test using the MANOVA technique, offering a multidimensional view where factors (i.e., complexity measures) are not only considered independently but also in their interplay. The analysis is carried out by organizing the indices – see Section IV – extracted for each text by column in a suitable data matrix. The purpose of MANOVA is to determine whether data from several groups (levels) of a factor have a common mean. It decomposes the total variation through the within- and between-groups variation but, in this case, these quantities are strictly related to the covariance matrix, hence the multivariate capability. MANOVA involves several matrix manipulations such as the Singular Value Decomposition where the derived eigenvalues are used for the test statistics [63]. Interestingly, MANOVA provides a set of canonical variables lying in a latent space, similar to Principal Component Analysis (PCA). Yet, while PCA computes the combination of the original variables that has the largest possible variation, MANOVA looks for the linear combination of the original variables that has the largest separation between groups. The size d of the group means obtained is equal to 2 (p-values $= 1.0 \times 10^{-82}, 0.2538 \times 10^{-82}$) indicating that our group means lies in a space of dimension equal to $d = 2$ leading to reject the null hypothesis that all group means are equal. In Fig. 15 is offered the scatter plot of the first two canonical variables (c_1, c_2), that is a projection (with information loss) of multidimensional data over a 2D plane. The figure shows, even at low dimensions, a tendency of the data to cluster. Insert b) of the figure shows the dendrogram which underlines the closeness of novels with texts generated by GPT-2, compared to programming codes. Ultimately, the measures of complexity turn out to be statistically descriptive, therefore informative in their heterogeneity, of the various text families within a multivariate setting. After the quantitative and detailed analysis of the complexity of the texts performed from

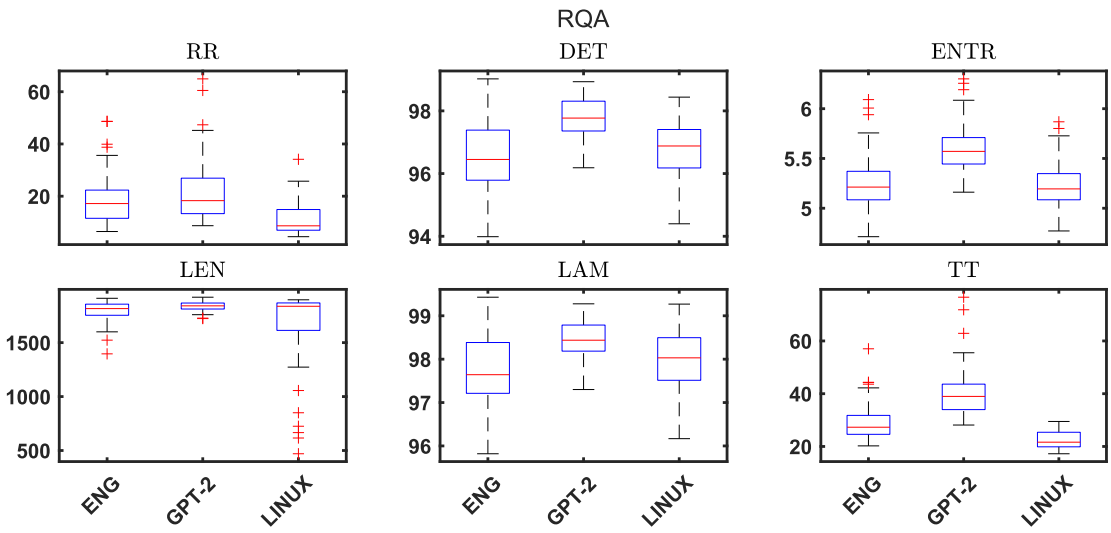


Fig. 13. Box-and-whisker plots of the main RQA indices.

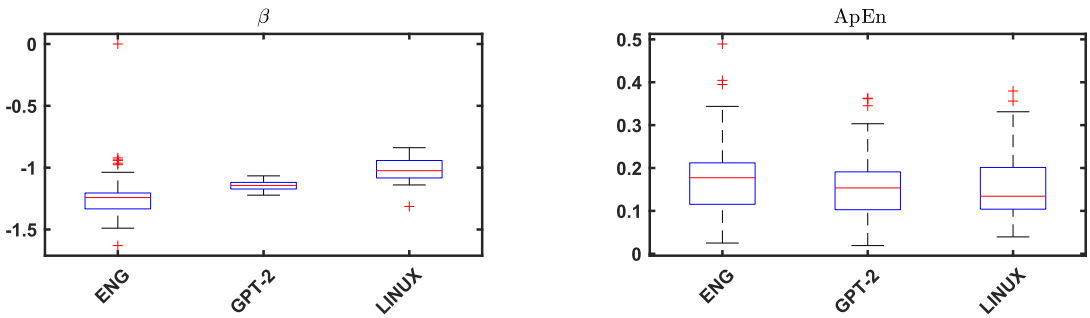


Fig. 14. Box-and-whisker plots of the β parameter estimated from the Zipf's law and the $ApEn$ parameter.

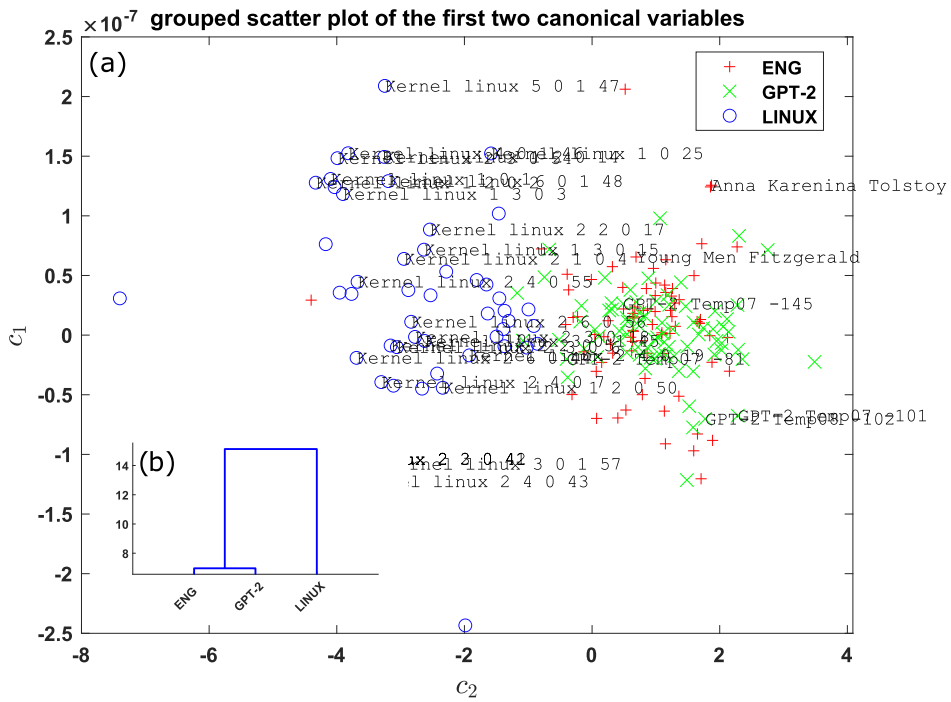


Fig. 15. MANOVA. scatter-plot of the the first two canonical components c_1, c_2 for the entire corpus consisting of the English novel texts, GPT-2 generated texts and for a selection of the Linux kernel C-files (the size of group symbols is proportional to the singularity strength $\Delta_{\tilde{\alpha}}$).

TABLE I
ALL THE COMPLEXITY MEASURES ADOPTED AS FEATURES FOR THE THREE-CLASS CLASSIFICATION PROBLEM

Feature	Explanation
$\widetilde{\alpha}_+$	right extreme of the singularity spectrum $f(\widetilde{\alpha})$
$\widetilde{\alpha}_-$	left extreme of the singularity spectrum $f(\widetilde{\alpha})$
$\widetilde{\alpha}^*$	maximum α value of the singularity spectrum
$\Delta_{\widetilde{\alpha}}$	amplitude of the singularity spectrum
$A_{\widetilde{\alpha}}$	asymmetry index of the singularity spectrum
H	Hurst exponent
τ_q	correlation dimension
m	embedding dimension of the dynamical trajectory
τ	time-delay of the dynamical trajectory
RR	recurrence rate
DET	determinism
ENTR	entropy
LEN	averaged diagonal line length
LAM	laminarity
TT	trapping time
β	exponent of the Zif's law
b	intercepts of the Zipf's law
ApEn	Approximate Entropy

different perspectives, we are now going to synthetically evaluate the information content linked to the complexity indices, framing the problem in a predictive context using a feature selection technique through an evolutionary machine learning procedure.

B. Synthesis

In order to further address whether the indices described in Section IV are indeed peculiar to characterize texts generated by humans and machines, we considered the corpus of 212 texts (see Section V) divided in three classes: ENG (80 documents), GPT-2 (80 documents) and LINUX (52 documents). Each text is represented by a real-valued feature vector obtained by the concatenation of 18 complexity indices (see Section IV), summarized for the sake of simplicity in Table I.

The set of 212 observations has been split into training, validation and test set with a ratio of 50%-25%-25% and then normalized, in order to avoid implicit weighting phenomena due to different features spanning different ranges. A RBF-Support Vector Machine [64] has been selected as classification system in order to discriminate between the three classes. The SVM is wrapped by a genetic algorithm [65] which is in charge of performing hyperparameter optimization and, eventually, feature selection. In other words, for this analysis, we rely on a well-known statistical classifier serving as an 'external validator' on the soundness of the selected features for characterizing different sources of text. In detail, we tackled five different classification problems:

Problem 1: all of the 18 features are used to solve the classification problem;
Problem 2: the genetic algorithm acts as a feature selector, hence it will return a subset of the 18 features deemed most informative to solve the classification problem;
Problem 3: a subset of 7 MFDFA-related features is used to solve the classification problem;

TABLE II
TEST SET ACCURACY (AVERAGE $\pm$ STANDARD DEVIATION ACROSS 10 DIFFERENT DATASET PARTITIONS)

Experiment	Accuracy [%]
Problem 1	89.81 ± 5.20
Problem 2	93.58 ± 3.11
Problem 3	38.30 ± 18.51
Problem 4	66.98 ± 22.99
Problem 5	89.81 ± 4.64

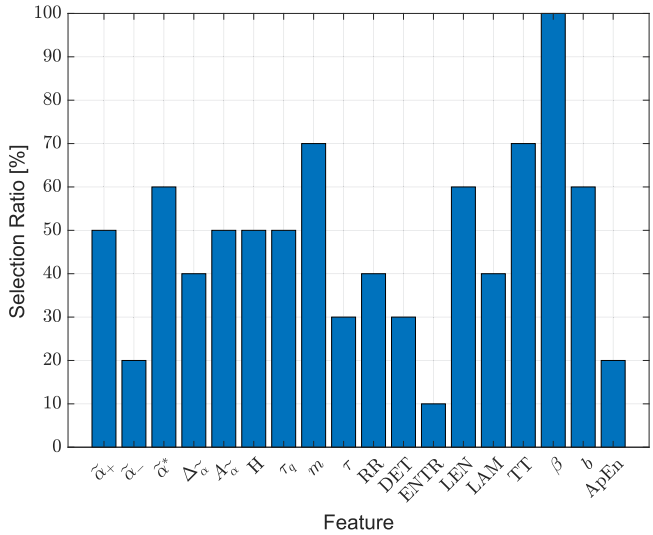


Fig. 16. Ratio of selection for each of the 18 features from Table I.

Problem 4: a subset of 5 RQA-related features is used to solve the classification problem;
Problem 5: a subset of 13 features corresponding to MFDFA, RQA, ApEn and Zipf's law is used to solve the classification problem.

The results for all of the 5 experiments are summarized in Table II. From Table II, it is possible to notice that MFDFA or RQA, if considered alone, do not provide a sound characterization of the texts ($p > 0.05$ with respect to using all features), whereas there are no statistically significant differences between Problems 1, 2 and 5 although, in absolute terms, the genetic algorithm-driven feature selection experiment yields a 4% accuracy improvement (on average). In Fig. 16 we show the results of the feature selection in terms of percentage of selection for each of the 18 features across the 10 different dataset partitions. It is possible to notice that, on average, entropy-based measures are the least important characteristics, whilst the Zipf's law exponent is considered as one of the most important descriptors. The lack of a neat polarization towards one (or a family of) complexity measure(s) confirms our claim that no individual indices are suitable for discriminating between human- and machine-generated texts.

VII. CONCLUSION

Heinz von Foerster, the Austrian American cognitive scientist who was the originator of Second-order cybernetics, in 1969

claimed [66]: "I am still baffled by the mystery that when Jim, a friend of Joe, hears the noises that are associated with reading aloud from the black marks that follow: ANN IS THE SISTER OF JOE – or just sees these marks – knows that indeed Ann is the sister of Joe, and, *de facto*, changes his whole attitude toward the world, commensurate with his new insight into a relational structure of elements in this world". We are all the more amazed that "changes toward the world" nowadays can be induced in our brains by a machine-generated text which, although according to its creator it has no knowledge of what it says [67], can *deceive* human beings in some cases. It can be asserted that the ability to mimic the production of natural language consists in learning the underlying statistic features (syntactic-grammatical structure, content distribution, etc.), which we know to be hierarchically organized. In this study, of methodological flavor, we have provided a quantitative external characterization of texts generated via the GPT-2 architecture in comparison to texts produced by human brains. This has been achieved through the framework of complexity science by analyzing regularities, patterns and recurrences (through specific measures) in order to investigate where the intelligent behavior of the machine is hidden. The analysis phase showed that the machine-generated text possesses long-term correlations, a peculiar multifractal distribution and specific recurrence patterns. Furthermore, machine-generated texts are placed somewhere between fluid texts produced by humans (English-language novels) and programming codes belonging to the Linux kernel. It has been experimentally demonstrated, through multivariate analysis, that (at least for the analyzed corpus) the heterogeneous measures of complexity are informative and allow to discriminate between the three analyzed text families. In addition, from a synthetic perspective, the analyzed measures have been used as descriptors of a text to build a feature vector in order to train a machine learning system operating feature selection. The experiments showed that the two families of complexity measures (related to the MFDFA and to the RQA), in their joint usage, are more predictive than their adoption alone. This claim has also been confirmed by the automatic feature selection procedure by means of a genetic algorithm, which does not show any bias towards any particular family of measures of complexity (RQA or MFDFA or ApEn or Zipf's) and, as instead, always looks for some compromises between features, with the exponent of the Zipf's law β emerging as the most significant feature (being it selected for all of the 10 different runs of the training procedure). This shows the potential of the complexity framework to analyze texts, both from a theoretical (e.g., in quantitative linguistics) and an applied perspective, for example in AI and related fields such as content quality control (e.g., to filter out machine-generated contents), fake news detection (e.g., spam or bot activities), plagiarism detection (e.g., in educational institutions or scientific research). We hypothesize that while multifractal analysis can be useful in synthetically understanding the statistical richness and hierarchical organization underlying a machine-generated text, recurrence analysis is related to the quality of the text. The rapid advancement of LLMs, specifically related to instruction fine-tuned models often further aligned with Reinforcement learning from human feedback (RLHF), opens the way to new challenges associated with modeling these systems according to complexity theory. We hypothesize that these new powerful models, with the possibility of modeling also author's style, have more degrees of freedom that translate into richer structures and patterns identifiable with more than one complexity index - as proposed in the current investigation - together with canonical stylometric measures. To enrich the analysis, we argue that might be necessary to model other characteristics of the text such as the length of the sentences or the sequence of POS-tags directly. As future works, in a companion paper, we have planned to go in depth on these interesting questions while maintaining the general claim that concerns the characterization of texts generated by machines with respect to some methodologies made available by the complexity sciences.

References

[1] K. B. Jensen and R. T. Craig, *The International Encyclopedia of Communication Theory and Philosophy, 4 Volume Set*. Hoboken, NJ, USA: Wiley, 2016.

[2] A. Lenci, "Distributional semantics in linguistic and cognitive research," *Italian J. Linguistics*, vol. 20, no. 1, pp. 1–31, 2008.

[3] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2/3, pp. 146–162, 1954.

[4] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[6] J. Kwapień and S. Drożdż, "Physical approach to complex systems," *Phys. Rep.*, vol. 515, no. 3, pp. 115–226, 2012.

[7] S. Pinker, *The Language Instinct: The New Science of Language and Mind*, vol. 7529. New York, NY, USA: Penguin, 1995.

[8] A. Baronchelli, V. Loreto, and F. Tria, "Language dynamics," *Adv. Complex Syst.*, vol. 15, no. 03n04, 2012, Art. no. 1203002, doi: 10.1109/TVT.2022.3148796.

[9] R. M. Roxas-Villanueva, M. K. Nambatac, and G. Tapang, "Characterizing English poetic style using complex networks," *Int. J. Modern Phys. C*, vol. 23, no. 02, 2012, Art. no. 1250009.

[10] T. Yasseri, A. Kornai, and J. Kertész, "A practical approach to language complexity: A Wikipedia case study," *PLoS One*, vol. 7, no. 11, 2012, Art. no. e48386.

[11] M. A. Montemurro and D. H. Zanette, "Towards the quantification of the semantic information encoded in written language," *Adv. Complex Syst.*, vol. 13, no. 02, pp. 135–153, 2010.

[12] W. F. von Humboldt and A. F. Pott, "Ueber Die Verschiedenheit des Menschlichen Sprachbaues und Ihren Einfluss Auf Die Geistige Entwickelung des Menschengeschlechts: Mit Erläuternden Anmerkungen und Excursen Sowie Als Einleitung: Wilhelm Von Humboldt und Die Sprachenwissenschaft Von A. F. Pott," Berlin: S. Calvary & Co, 1876.

[13] M. J. Traxler, M. Boudewyn, and J. Loudermilk, "What's special about human language? The contents of the "narrow language faculty" revisited," *Lang. Linguistics Compass*, vol. 6, no. 10, pp. 611–621, 2012.

[14] E. G. Altmann, G. Cristadoro, and M. D. Esposti, "On the origin of long-range correlations in texts," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 29, pp. 11582–11587, 2012.

[15] R. Ferrer-i Cancho, V. M. Longa, and G. Lorenzo, "Long-distance dependencies are not uniquely human," in *The Evolution of Language*, Singapore: World Scientific, 2008, pp. 115–122.

[16] A. Kershenbaum, A. E. Bowles, T. M. Freeberg, D. Z. Jin, A. R. Lameira, and K. Bohn, "Animal vocal sequences: Not the Markov chains we thought they were," *Proc. Roy. Soc. B: Biol. Sci.*, vol. 281, no. 1792, 2014, Art. no. 20141370.

[17] S. Melnyk, O. Usatenko, V. Yampol'skii, and V. Golick, "Competition between two kinds of correlations in literary texts," *Phys. Rev. E*, vol. 72, no. 2, 2005, Art. no. 026140.

[18] C. D. Manning, K. Clark, J. Hewitt, U. Khandelwal, and O. Levy, "Emergent linguistic structure in artificial neural networks trained by self-supervision," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 48, pp. 30046–30054, 2020.

[19] J. Vig and Y. Belinkov, "Analyzing the structure of attention in a transformer language model," 2019, *arXiv:1906.04284*.

[20] J. Weinberg, Philosophers on GPT-3 (updated with replies by GPT-3). Section: Public philosophy and outreach, 2020. [Online]. Available: https://dailynous.com/2020/07/30/philosophers-gpt-3/
[21] D. M. Powers, "Applications and explanations of Zipf's law," in *Proc. Joint Conf. New Methods Lang. Process. Comput. Natural Lang. Learn.*, 1998, pp. 151–160.
[22] G. K. Zipf, *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Evanston, IL, USA: Routledge, 2013.
[23] J. W. Kantelhardt, S. A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, and H. E. Stanley, "Multifractal detrended fluctuation analysis of nonstationary time series," *Physica A: Stat. Mechanics Appl.*, vol. 316, no. 1, pp. 87–114, 2002.
[24] J. W. Kantelhardt, E. Koscielny-Bunde, H. H. Rego, S. Havlin, and A. Bunde, "Detecting long-range correlations with detrended fluctuation analysis," *Physica A: Stat. Mechanics Appl.*, vol. 295, no. 3/4, pp. 441–454, 2001.
[25] E. De Santis, P. Naraei, A. Martino, A. Sadeghian, and A. Rizzi, "Multifractal characterization and modeling of blood pressure signals," *Algorithms*, vol. 15, no. 8, 2022, Art. no. 259.
[26] E. De Santis, A. Sadeghian, and A. Rizzi, "A smoothing technique for the multifractal analysis of a medium voltage feeders electric current," *Int. J. Bifurcation Chaos*, vol. 27, no. 14, 2017, Art. no. 1750211.
[27] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, "Recurrence plots for the analysis of complex systems," *Phys. Rep.*, vol. 438, no. 5/6, pp. 237–329, 2007.
[28] S. M. Pincus, I. M. Gladstone, and R. A. Ehrenkranz, "A regularity statistic for medical data analysis," *J. Clin. Monit.*, vol. 7, no. 4, pp. 335–345, 1991.
[29] S. M. Pincus, "Approximate entropy as a measure of system complexity.," *Proc. Nat. Acad. Sci. USA*, vol. 88, no. 6, pp. 2297–2301, 1991.
[30] E. De Santis, G. De Santis, and A. Rizzi, "Multifractal characterization of texts for pattern recognition: On the complexity of morphological structures in modern and ancient languages," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10143–10160, Aug. 2023.
[31] A. Schenkel, J. Zhang, and Y.-C. Zhang, "Long range correlation in human writings," *Fractals*, vol. 1, no. 01, pp. 47–57, 1993.
[32] P. Allegrini, P. Grigolini, and L. Palatella, "Intermittency and scale-free networks: A dynamical model for human language complexity," *Chaos, Solitons Fractals*, vol. 20, no. 1, pp. 95–105, 2004.
[33] K. Kosmidis, A. Kalampokis, and P. Argyrakis, "Language time series analysis," *Physica A: Stat. Mechanics Appl.*, vol. 370, no. 2, pp. 808–816, 2006.
[34] S. Drożdż et al., "Quantifying origin and character of long-range correlations in narrative texts," *Inf. Sci.*, vol. 331, pp. 32–44, 2016.
[35] I. Grabska-Gradzinska, A. Kulig, J. Kwapien, P. Oswiecimka, and S. Drozdz, "Multifractal analysis of sentence lengths in English literary texts," 2012, *arXiv:1212.3171*.
[36] M. A. Montemurro and P. A. Pury, "Long-range fractal correlations in literary corpora," *Fractals*, vol. 10, no. 04, pp. 451–461, 2002.
[37] M. Ausloos, "Generalized hurst exponent and multifractal function of original and translated texts mapped into frequency and length time series," *Phys. Rev. E*, vol. 86, no. 3, 2012, Art. no. 031108.
[38] M. Ausloos, "Measuring complexity with multifractals in texts. Translation effects," *Chaos Solitons Fractals*, vol. 45, no. 11, pp. 1349–1357, 2012.
[39] P. Kokol, V. Podgorelec, M. Zorman, T. Kokol, and T. Njivar, "Computer and natural language texts–A comparison based on long-range correlations," *J. Amer. Soc. Inf. Sci.*, vol. 50, no. 14, pp. 1295–1301, 1999.
[40] M. Lippi, M. A. Montemurro, M. Degli Esposti, and G. Cristadoro, "Natural language statistical features of LSTM-generated texts," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3326–3337, Nov. 2019.
[41] A. Muñoz-Ortiz, C. Gómez-Rodríguez, and D. Vilares, "Contrasting linguistic patterns in human and LLM-generated text," 2023, *arXiv:2308.09067*.
[42] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.
[43] B. Partee, "Lexical semantics and compositionality," *An Invitation Cogn. Sci.*, vol. 1, pp. 311–360, 1995.
[44] D. Hupkes, V. Dankers, M. Mul, and E. Bruni, "Compositionality decomposed: How do neural networks generalise?," *J. Artif. Intell. Res.*, vol. 67, pp. 757–795, 2020.
[45] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *OpenAI*, 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
[46] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, pp. 1–24, 2019.
[47] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.
[48] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA, USA: Addison-Wesley, 1949.
[49] R. Ferrer-i Cancho and R. V. Solé, "Least effort and the origins of scaling in human language," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 3, pp. 788–791, 2003.
[50] S. Yu, C. Xu, and H. Liu, "Zipf's law in 50 languages: Its structural pattern, linguistic interpretation, and cognitive motivation," 2018, *arXiv:1807.01855*.
[51] J. Baixeries, B. Elvevåg, and R. Ferrer-i Cancho, "The evolution of the exponent of Zipf's law in language ontogeny," *PLoS One*, vol. 8, no. 3, 2013, Art. no. e53227.
[52] Z. Yu, L. Yee, and Y. Zu-Guo, "Relationships of exponents in multifractal detrended fluctuation analysis and conventional multifractal analysis," *Chin. Phys. B*, vol. 20, no. 9, 2011, Art. no. 090507.
[53] Q. Cheng, "Generalized binomial multiplicative cascade processes and asymmetrical multifractal distributions," *Nonlinear Processes Geophys.*, vol. 21, no. 2, pp. 477–487, 2014.
[54] S. Drożdż and P. Oświęcimka, "Detecting and interpreting distortions in hierarchical organization of complex time series," *Phys. Rev. E*, vol. 91, Mar. 2015, Art. no. 030902.
[55] T. Schreiber and A. Schmitz, "Surrogate time series," *Physica D: Nonlinear Phenomena*, vol. 142, no. 3/4, pp. 346–382, 2000.
[56] J.-P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *Europhysics Lett.*, vol. 4, no. 9, pp. 973–977, 1987.
[57] G. Ouyang, X. Zhu, Z. Ju, and H. Liu, "Dynamical characteristics of surface EMG signals of hand grasps via recurrence plot," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 1, pp. 257–265, Jan. 2014.
[58] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, "Geometry from a time series," *Phys. Rev. Lett.*, vol. 45, no. 9, pp. 712–716, 1980.
[59] N. Marwan, *Encounters With Neighbours: Current Developments of Concepts Based on Recurrence Plots and their Applications*, Universität Potsdam, 2003.
[60] M. B. Kennel, R. Brown, and H. D. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Phys. Rev. A*, vol. 45, no. 6, pp. 3403–3411, 1992.
[61] C. L. Webber Jr. and J. P. Zbilut, "Recurrence quantification analysis of nonlinear dynamical systems," *Tut. Contemporary Nonlinear Methods Behav. Sci.*, vol. 94, pp. 26–94, 2005.
[62] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," 2019, *arXiv:1904.09751*.
[63] N. H. Timm, *Applied Multivariate Analysis*. Berlin, Germany: Springer, 2002.
[64] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Comput.*, vol. 12, no. 5, pp. 1207–1245, 2000.
[65] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989.
[66] H. V. Foerster, "Thoughts and notes on cognition," in *Understanding Understanding*, Berlin, Germany: Springer, 2003, pp. 169–189.
[67] M. Hutson, "Robo-writers: The rise and risks of language-generating AI," *Nature*, vol. 591, no. 7848, pp. 22–25, 2021.

Enrico De Santis (Member, IEEE) received the MASc (Hons.) and PhD degrees in information and communication engineering from "Sapienza" University of Rome, Italy, and worked as an assistant researcher and successively as a postdoc with the Department of Computer Science, Toronto Metropolitan University, Toronto, Canada. Currently, he holds a researcher position with the Department of Information Engineering, Electronics and Telecommunications (DIET), "Sapienza". His research interests include artificial intelligence, complex systems and data-driven modeling, natural language processing, computational intelligence, neural networks and fuzzy systems with application in smart grids and predictive maintenance. Since 2017, he has joined an innovative startup with "Sapienza" University as CTO, dealing with the management of Artificial Intelligence projects in production environments.

Alessio Martino (Member, IEEE) received the graduated degree summa cum laude in communications engineering from the University of Rome "La Sapienza", Italy, 2016. From 2016 to 2019, he served as PhD research fellow in Information and Communications Technologies with the University of Rome "La Sapienza" (Department of Information Engineering, Electronics and Telecommunications). During his PhD, he also served as scientific collaborator with Consortium for Research in Automation and Telecommunication, Rome, Italy. After obtaining the PhD, he has been granted a 1-year postdoctoral research Fellowship with the University of Rome "La Sapienza" and a 1-year postdoctoral research Fellowship with the Italian National Research Council. Currently, he is assistant professor of computer science with LUISS University. His research interests include machine learning, computational intelligence and knowledge discovery. Currently, he's focusing on large-scale machine learning, advanced pattern recognition systems, Big Data analysis, parallel & distributed computing, granular computing and complex systems modelling, in applications including bioinformatics and computational biology, natural language processing and energy distribution networks.

Antonello Rizzi (Senior Member, IEEE) since July 2010 has been with the Department of Information Engineering, Electronics and Telecommunications (DIET), "Sapienza" University of Rome as an Assistant professor. Currently, he serves as an associate professor with DIET. Since 2008, he is the scientific coordinator and R&D technical director in the Intelligent Systems Laboratory within the Research Center for Sustainable Mobility of Lazio region, Italy. His major research interests are in computational intelligence and pattern recognition, including supervised and unsupervised machine learning techniques, neural networks, fuzzy systems, and evolutionary algorithms, with application in smart grids and microgrids modeling and control, intelligent systems for sustainable mobility, battery management systems. He has (co)-authored more than 220 international journal/conference papers and book chapters.

Open Access funding provided by 'Università degli Studi di Roma "La Sapienza" 2' within the CRUI CARE Agreement