



HOKKAIDO UNIVERSITY

Title	Berezinskii-Kosterlitz-Thouless transition in a context-sensitive random language model
Author(s)	Toji, Yuma; Takahashi, Jun; Roychowdhury, Vwani et al.
Citation	Physical Review E, 113(1), 015305 https://doi.org/10.1103/s7nf-bwzd
Issue Date	2026-01-21
Doc URL	https://hdl.handle.net/2115/99110
Rights	©[2026] American Physical Society
Type	journal article
File Information	paper_language_model_long_000_001_main.pdf



Berezinskii–Kosterlitz–Thouless transition in a context-sensitive random language model

Yuma Toji,¹ Jun Takahashi,² Vwani Roychowdhury,³ and Hideyuki Miyahara^{1,*}

¹ *Graduate School of Information Science and Technology,
Hokkaido University, Sapporo, Hokkaido 060-0814, Japan*

² *Institute for Solid State Physics, The University of Tokyo,
5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8581, Japan*

³ *Henry Samueli School of Engineering and Applied Science,
University of California, Los Angeles, California 90095*

(Dated: March 25, 2026)

Several power-law critical properties involving different statistics in natural languages—reminiscent of scaling properties of physical systems at or near phase transitions—have been documented for decades. The recent rise of large language models has added further evidence and excitement by providing intriguing similarities with notions in physics such as scaling laws and emergent abilities. However, specific instances of classes of generative language models that exhibit phase transitions, as understood by the statistical physics community, are lacking. In this work, inspired by the one-dimensional Potts model in statistical physics, we construct a simple probabilistic language model that falls under the class of context-sensitive grammars, which we call the context-sensitive random language model, and numerically demonstrate an unambiguous phase transition in the framework of a natural language model. We explicitly show that a precisely defined order parameter—that captures symbol frequency biases in the sentences generated by the language model—changes from strictly zero to a strictly nonzero value (in the infinite-length limit of sentences), implying a mathematical singularity arising when tuning the parameter of the stochastic language model we consider. Furthermore, we identify the phase transition as a variant of the Berezinskii–Kosterlitz–Thouless (BKT) transition, which is known to exhibit critical properties not only at the transition point but also in the entire phase. This finding leads to the possibility that critical properties in natural languages may not require careful fine-tuning nor self-organized criticality, but are generically explained by the underlying connection between language structures and the BKT phases. From a statistical physics perspective, our paper captures a rare phenomenon of the BKT phase in a one-dimensional system, as the BKT phase is widely studied and believed to exist only in two-dimensional systems.

I. INTRODUCTION

Natural languages exhibit striking statistical regularities, such as Zipf’s law, which describes a power-law relationship between word frequencies and their rank [1]. Similarly, it has been shown that the number of bits of information provided by a symbol about another symbol drops roughly as a power law as a function of their distance [2]. These critical properties, observed across diverse languages, have inspired explanations rooted in statistical physics, drawing parallels between linguistic phenomena and phase transitions. For example, the scaling behavior inherent in Zipf’s law suggests underlying mechanisms that resemble those driving critical phenomena in physical systems [3].

Recent advances in generative language models, including large language models (LLMs) [4–7], have further highlighted intriguing connections between language and physics. Notably, studies have uncovered scaling laws and emergent capabilities in these models, which apparently become evident as the system size increases [8–10].

These findings have prompted the question of whether the statistical properties of natural languages and the behavior of their generative models can be understood as manifestations of genuine phase transitions in physical terms.

These observed phenomena in LLMs, combined with the fact that they become apparent when some parameters of the system are large enough (network depth, width, training data set size, etc.), evoke parallels to phase transitions in the field of statistical physics. The fact that in critical (continuous) phase transitions many quantities, such as the correlation function, exhibit power-law scaling also fuels the approach of studying these emergent phenomena in languages from a statistical physics perspective. The notion of phase transitions in the thermodynamic limit, i.e., in an idealized limit where some parameter governing the system size tends to infinity, enables us to study qualitative differences in the system in a rigorous way.

Such a formal qualitative study is essential because once we have a parameterized language model, it will always have a “trivial regime” corresponding to a trivial distribution of characters/words/tokens, and an objectively defined hallmark of nontriviality—such as phase transitions—will be a theoretically powerful tool to distinguish meaningful sentences from those generated in

*Electronic address: miyahara@ist.hokudai.ac.jp, hmiyahara512@gmail.com; corresponding author.

the trivial regimes. Just as there is a clear and mathematically well-defined boundary between water and ice, the existence of a phase transition in language models could lead us to understand the intrinsic difference separating, e.g., a language-like structure that actually conveys information versus generated text that is just statistical gibberish.

The potential existence of a spin-glass-like phase transition in probabilistic context-free grammars (CFGs) has recently attracted attention [11, 12] [see Fig. 1(a) for a schematic diagram of CFGs]. However, detailed analyses of CFGs have revealed the absence of such claimed phase transitions in some naturally assumed thermodynamic limits [13]. While consensus is yet to be reached on this matter [14, 15], it would be fair to say that numerical and analytical support for a true phase transition, i.e., a mathematically singular behavior in a clearly defined thermodynamic limit, is lacking. Furthermore, as a probabilistic CFG is one of the simplest probabilistic models of natural language, this conundrum with the existence of phase transitions naturally raises the question of how complex a natural language model needs to be to exhibit phase transitions.

Phase transitions in statistical physics are precisely defined mathematical concepts that require careful analysis. We provide a schematic figure of the types of phase transitions investigated in this paper in Fig. 1(b). In particular, to demonstrate a phase transition, one should be able to construct an *order parameter*, which is strictly zero in the disordered phase and is nonzero in the ordered phase, thus allowing a rigorous characterization of different phases in a system. This switch from zero to nonzero is inevitably accompanied by a mathematical singularity (a divergence or discontinuity in the derivative, etc.) of the order parameter. Note that these features are only realized in the thermodynamic limit. In practice, one can only observe finite-size systems and extrapolate the data to see consistency with the existence of phase transitions in the thermodynamic limit. In our paper, we will consider the relative frequencies of letters in the alphabet as the order parameter, and explore whether it shows a second-order phase transition, where the derivative of the order parameter has a singularity as a function of temperature. Temperature in the context of a language model represents the probability with which a particular rule is applied in text generation.

There are, however, different types of second-order phase transitions. In a conventional transition, the critical behavior characterized by power-law scaling phenomena emerges only at the point of the phase transition. That is, the system parameters have to be accurately tuned—so that the system is very close to the phase transition point—for any such scaling laws to be observed. Both LLMs and natural languages, however, show robust scaling behaviors without requiring extensive fine-tuning of parameters. This raises the possibility that for language models to exhibit phase transitions, they might need to go through a different type of phase tran-

sition, referred to as the Berezinskii–Kosterlitz–Thouless (BKT) [16–20] transition. One of the most exceptional features of the BKT phase is that the system becomes critical throughout the entire phase, and not only at the transition point [see Fig. 1(b)]. The presence of any BKT transition could explain the robust presence of scaling laws in language models.

The BKT transition, however, is most commonly discussed in two-dimensional systems with continuous symmetries. However, since language models generate strings of text any reasonable physical modeling would lead to one-dimensional systems. While the BKT transition is also realizable in long-range interacting one-dimensional systems, such one-dimensional phenomena are less studied and less well known. One method to determine whether a phase transition is BKT is to compute the kurtosis of the order parameter, which is referred to as the *Binder parameter* in the physics literature. The Binder parameter also nicely captures phase transitions—showing a singularity in its derivative—while being able to distinguish between different types of phases: it is strictly 0 in the disordered phase, and is exactly 1 in a conventionally ordered phase, while having a nontrivial value (between 0 and 1) at critical points where correlations have power-law decays; see Fig. 1(b). The Binder parameter becomes an especially powerful tool for studying our language model since the model turns out to exhibit a type of the BKT phase.

In this study, we construct a new language model that falls under the class of context-sensitive grammars (CSGs), which we call the context-sensitive random language model, by introducing context-sensitivity to the random language model [11, 12]. Our construction is based on the long-range interacting Potts model, and utilizes the additional complexity that CSGs possess from being one level higher in the Chomsky hierarchy [21, 22] compared to CFGs.

CSGs allow the application of a rewrite rule to depend on the symbols surrounding the one being rewritten. Such rules naturally introduce effective long-range couplings and capture linguistic phenomena—agreement, feature matching, and structural constraints—that CFGs cannot express. In this work, we extract and formalize these essential ingredients to construct a probabilistic, CSG-inspired random language model that is sufficiently expressive to generate extended correlations while remaining amenable to thermodynamic analysis.

The model dynamics combine three interacting processes: (1) [Growth] A nonterminal symbol expands via rules of the form $X \rightarrow YZ$, increasing string length and enabling a thermodynamic limit as $\text{length} \rightarrow \infty$. A growth parameter q controls the balance between expansion and rewriting, with different regimes giving rise to qualitatively distinct global behaviors. (2) [Context-sensitive rewrites with Metropolis acceptance] A substring $\alpha_- X \alpha_+$ may be rewritten as $\alpha_- Y \alpha_+$, with acceptance probability $p = \min(1, e^{-\frac{\Delta E}{k_B T}})$ where ΔE is computed from an energy functional coupling symbol pairs at

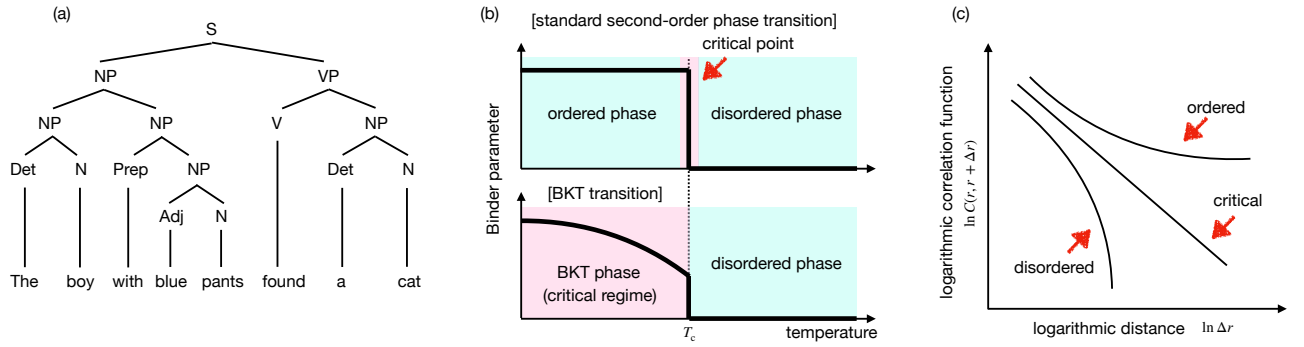


FIG. 1: (a) Schematic of a CFG. This diagram shows the syntactic structure of “The boy with blue pants found a cat.” Natural languages are, however, context-sensitive and in this paper we define a simple CSG that demonstrates a phase transition. (b) We consider the relative frequencies of letters or symbols in the alphabet—as captured by the net magnetization of the corresponding physical system—as an order parameter in our CSG. The Binder parameter, which is the normalized kurtosis of the underlying order parameter [see Eq. 8], can capture phase transitions effectively, especially for one-dimensional models. A schematic of the Binder parameter is shown for the standard second-order phase transition (upper) and the BKT transition (lower). For the BKT transition the Binder parameter’s dependence on temperature—which determines the probability with which a rule is applied in the underlying CSG—does not become a step function in the thermodynamic limit. (c) Correlation functions for the disordered, critical, and ordered phases. When a system is critical, the correlation function shows a polynomial decay.

distance $|i - j|$ through a long-range kernel $|i - j|^{-(1+s)}$. This structure is analogous to a one-dimensional long-range Potts model but with crucial differences: the domain itself grows, interactions depend on context, and the rewriting operations are not simple spin flips but grammar-inspired substitutions. (3) [Transition to terminal symbols] A nonterminal symbol may transition into a terminal symbol, after which it no longer grows or undergoes rewriting. Note that the notion of nonterminal symbols has no direct counterpart in physics and is specific to language models. For simplicity, we neglect this process in the present study; as a result, we can take the thermodynamic limit straightforwardly.

Languages have a one-dimensional structure naturally associated with them: namely, the direction from the beginning to the end of a sentence. Together with the observation that the degrees of freedom in a sentence are essentially discrete (such as words or tokens), we focus our attention on the archetypal toy model in statistical physics with discrete degrees of freedom: the one-dimensional Potts model. We focus on the order parameter that captures the relative frequency of the letters or symbols in the string. We also compute the higher-order cumulants (variance and kurtosis), which correspond to the susceptibility and Binder parameter from a physics perspective, and are the most widely used measures for detecting phase transitions. From the perspective of language models, e.g., the variance/susceptibility corresponds to how diverse the produced strings are, and therefore could be viewed as a measure of the average nontriviality of the sentences. We compute the system-size dependence of the susceptibility and perform the finite-size scaling method [23–26], which allows us to confidently conclude the existence of a phase transition in the thermodynamic

(infinite-size) limit. For example, we can observe a clear divergence of the susceptibility in the low-temperature BKT phase in our model, which indicates an entire parameter range where nontrivially correlated strings are produced.

In our numerical simulations, we first compute the system-size dependence of the quantity corresponding to magnetization. This quantifies how much a specific symbol is favored in the generating process of a sentence. We also compute the corresponding susceptibility and the Binder parameter, which play an important role in distinguishing between the standard second-order phase transition and the BKT transition [27, 28]. We then perform the finite-size scaling method to confirm the existence of phase transitions. In addition to these, we also investigate the correlation functions¹ between symbols [2]. These quantities play an essential role in discussing the scaling law in at least natural languages as the power-law scaling in correlation functions serves as the smoking gun for critical phase transitions in physics [see Fig. 1(c) for correlation functions]².

It should be noted that the context-sensitive random language model is *not* a mere translation of the physical Potts model into a language model setting. The unique property of language models in general, absent

¹ In addition to the correlation function, we also compute mutual information in Appendix D.

² Furthermore, we analyze the histogram of the magnetization and show typical configurations of symbols for parameters at which the histogram has a peak and relative frequencies of letters in the alphabet, which are often discussed in the context of natural language processing in Appendix D.

in standard statistical mechanics models, is the increase in symbols over time, and there is an additional degree of freedom in how the symbols grow; see the parameters q, t and the specifics of the $X \rightarrow YZ$ rule in Eq. 1b. *These properties of the context-sensitive random language model render the existence of a phase transition nontrivial*, and indeed we find that for certain settings of the language parameters, the phase transition can vanish. Furthermore, we compare the context-sensitive random language model with multiple alphabets to the one-dimensional long-range Potts model [29]. As for phase transitions in one-dimensional models, it is well known that the Ising model—which is equivalent to the Potts model in the case of an alphabet of size two (i.e., two letters)—in one dimension shows no phase transition with short-range interactions but exhibits phase transitions with long-range interactions [30–37]. For the one-dimensional long-range Potts model with two-level spins ($\sigma = \pm 1$), which is also known as the Ising model, it is known to exhibit an ordered phase at low temperatures and to undergo a BKT transition [19, 38–41]. However, such a BKT phase is observed when the decay exponent s in the power-law interaction with distance is set to $s = 1$ in the exponent [see Eq. 3]. In our language model, we observe a BKT transition even when $s = 0.9$, which is far from the setting studied in Refs. [19, 39]. We also observe a BKT transition in our systems with multi-level spins, i.e., alphabets with $K > 2$ letters. *We hypothesize that our language model, where the system size grows in size following a set of probabilistic rules, behaves differently from the physics models that inspired their conception and BKT transitions might be much more robust in our systems.*

The structure of this paper is as follows. In Sec. II, we describe in detail the language model we constructed, inspired by the Potts model. Section III introduces the physical quantities computed to analyze the layer-wise behavior of the language model. In Sec. IV, we present the results for a representative parameter setting, namely the case of $K = 2$ and $X \rightarrow YZ$. Based on these results, Sec. V discusses the implications of the observed phase transition phenomenon in the context of language models. Finally, we conclude the paper in Sec. VI by summarizing the findings and discussing their relevance to large LLMs. In the appendices, we describe the numerical setup in detail and show additional numerical simulations, and they are written in a self-contained manner.

II. LANGUAGE MODEL

We construct a probabilistic language model based on the long-range Potts model, which we call the context-sensitive random language model³. First of all, the generation process of a sentence begins with the start

symbol, which is a nonterminal symbol. Until all the nonterminal symbols (lexical elements that specify the production rules constituting a formal grammar) in the sentence are replaced by terminal symbols (outputs of the production rules), a nonterminal symbol is uniformly randomly chosen and one of the following three processes is applied to it:

$$X \rightarrow x \quad (\text{probability: } qt), \quad (1a)$$

$$X \rightarrow YZ \quad (\text{probability: } q(1-t)), \quad (1b)$$

$$\alpha_- X \alpha_+ \rightarrow \alpha_- Y \alpha_+ \quad (\text{probability: } (1-q)), \quad (1c)$$

where uppercase and lowercase letters are nonterminal and terminal symbols, respectively, and α_- and α_+ are strings. To take the thermodynamic limit, we set $t = 0$; in other words, we do not consider Eq. (1a)⁴. For Eq. (1b), we consider two branching processes: (i) $X \rightarrow YZ$, where Y and Z are uniformly randomly chosen, and (ii) $X \rightarrow XX$. (In the main text, we focus on (i); see Appendix D for (ii).) Furthermore, when Eq. (1c) is chosen, the corresponding update is accepted with the following probability:

$$p = \min(1, e^{-\frac{\Delta E}{k_B T}}), \quad (2)$$

where ΔE for $\sigma_0 \rightarrow \tilde{\sigma}_0$ is given by

$$\Delta E := J \left(\sum_{\ell=1}^{\lfloor Nr_- \rfloor} \frac{\delta_{\sigma_0, \sigma_{-\ell}} - \delta_{\tilde{\sigma}_0, \sigma_{-\ell}}}{\ell^{1+s}} + \sum_{\ell=1}^{\lfloor Nr_+ \rfloor} \frac{\delta_{\sigma_0, \sigma_{\ell}} - \delta_{\tilde{\sigma}_0, \sigma_{\ell}}}{\ell^{1+s}} \right). \quad (3)$$

Here, $J > 0$ is a hyperparameter that determines the strength of the interaction, and $\sigma_i \in \{0, 1, \dots, K-1\}$ is an integer corresponding to the symbol at the i -th character in the sentence. Here, $r_+, r_- \in [0, 1]$ specify the range of interaction. Furthermore, we consider the open boundary condition; so terms beyond the boundary are ignored. Note that Eq. (2) is the acceptance rate of the Metropolis-Hastings algorithm [42–44], and Eq. (3) is the energy gap of the one-dimensional long-range Potts model with respect to a one-spin flip.

III. METHODS

We summarize the physical quantities of interest used to discuss phase transitions⁵. First, we begin with the magnetization, which is the order parameter, and it reads

$$M := \| \mathbf{M} \|, \quad (4)$$

⁴ This thermodynamic limit is the simplest one while there can be different definitions of thermodynamic limits, which are beyond our scope.

⁵ The details of the physical quantities computed in this paper are given in Appendix C.

³ The details of our model are given in Appendix B.

where $\mathbf{M} \in \mathbb{R}^{K-1}$ is given by

$$\mathbf{M} := \frac{1}{N} \sum_{i=1}^N \mathbf{e}_{\sigma_i}. \quad (5)$$

Here, $\{\mathbf{e}_k\}_{k=1}^K$ satisfies $\|\mathbf{e}_k\| = 1$ for $k = 1, 2, \dots, K$, $\mathbf{e}_k \cdot \mathbf{e}_l = \text{const.}$ for $k \neq l$, and $\sum_{k=1}^K \mathbf{e}_k = \mathbf{0}$ ⁶. Note that $\|\mathbf{M}\| := \sqrt{\mathbf{M} \cdot \mathbf{M}}$.

We then define the magnetic susceptibility by

$$\chi := N(\langle M^2 \rangle - \langle M \rangle^2). \quad (6)$$

For later convenience, we also define

$$\tilde{\chi} := N\langle M^2 \rangle. \quad (7)$$

As a well-used method to discern phase transitions, the Binder parameter is often used, and it reads

$$U := -\frac{K-1}{2} \left(\frac{\langle M^4 \rangle}{\langle M^2 \rangle^2} - \frac{K+1}{K-1} \right). \quad (8)$$

From the viewpoint of statistics, the Binder parameter, Eq. (8), is the normalized kurtosis, which becomes zero for the normal distribution. Thus, it can be used to distinguish the disordered phase, in which the *generalized central limit theorem* empirically holds, and the ordered phase.

The correlation function is given by

$$\tilde{G}(i, j) := \langle \mathbf{e}_{\sigma_i} \cdot \mathbf{e}_{\sigma_j} \rangle. \quad (9)$$

It is well known that in a critical regime, the correlation function exhibits a polynomial decay, while it shows an exponential decay outside a critical regime.

Finite-size scaling⁷ is one of the most popular methods to discuss the existence of phase transitions. By applying the scaling assumption to $\tilde{\chi}$, Eq. (7), we obtain

$$\frac{\tilde{\chi}(T, N)}{N^{\frac{2}{\nu}}} = \tilde{f}_{\tilde{\chi}}(N^{\frac{1}{\nu}} t). \quad (10)$$

Note that the scaling assumption implies the existence of a universal function $\tilde{f}_A(\cdot)$ for any thermodynamic quantity A . We will confirm Eq. (10) via numerical simulations.

IV. NUMERICAL SIMULATIONS

We show numerical simulations restricting ourselves to the case of $K = 2$ and $X \rightarrow YZ$ ⁸. We run the Monte Carlo simulations 80×80 times to compute expected values and standard errors.

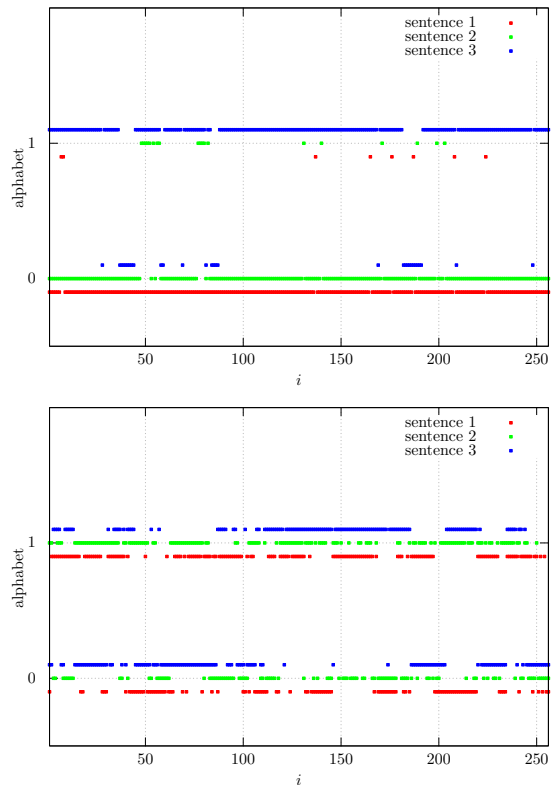


FIG. 2: Typical configurations of symbols for $X \rightarrow YZ$. We plot (upper) states whose magnetization lies between 0.9000 and 0.9200 at $k_B T = 0.760$, and (lower) states whose magnetization lies between 0.0000 and 0.0200 at $k_B T = 1.160$. We set $K = 2$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We plot the first 256 symbols out of 4096. Note that the estimated critical temperature is $k_B T_* = 0.960$.

We first plot the first 256 symbols of typical configurations of symbols in Fig. 2. Here, “typical” configurations refer to those whose magnetization corresponds to the peak of the magnetization histograms⁹. As shown in the figure, typical configurations of letters in the alphabet depend on the temperature. Hereafter, we investigate the statistical properties of these sentences.

In Fig. 3, we plot the temperature dependence of the magnetization, Eq. (4), the susceptibility, Eq. (6), and the Binder parameter, Eq. (8). First of all, for $k_B T \lesssim 1$, the magnetization has a nonzero value ($\langle M \rangle > 0$), and the system seems to be ordered. In the same regime, the susceptibility becomes not merely large but also divergent with respect to the system size, for $k_B T \lesssim 1$, unlike in the second-order phase transition. In other words, the system is critical for a parameter regime of nonzero measure, and this fact implies that the system is in the BKT phase. The Binder parameter also has a crossing point

⁶ See Appendix C 1 for the derivation of $\{\mathbf{e}_k\}_{k=1}^K$.

⁷ See Appendix C 8 for finite-size scaling.

⁸ Numerical simulations for other setups are given in Appendix D.

⁹ The histogram of magnetization for this parameter set is shown in Fig. D11.

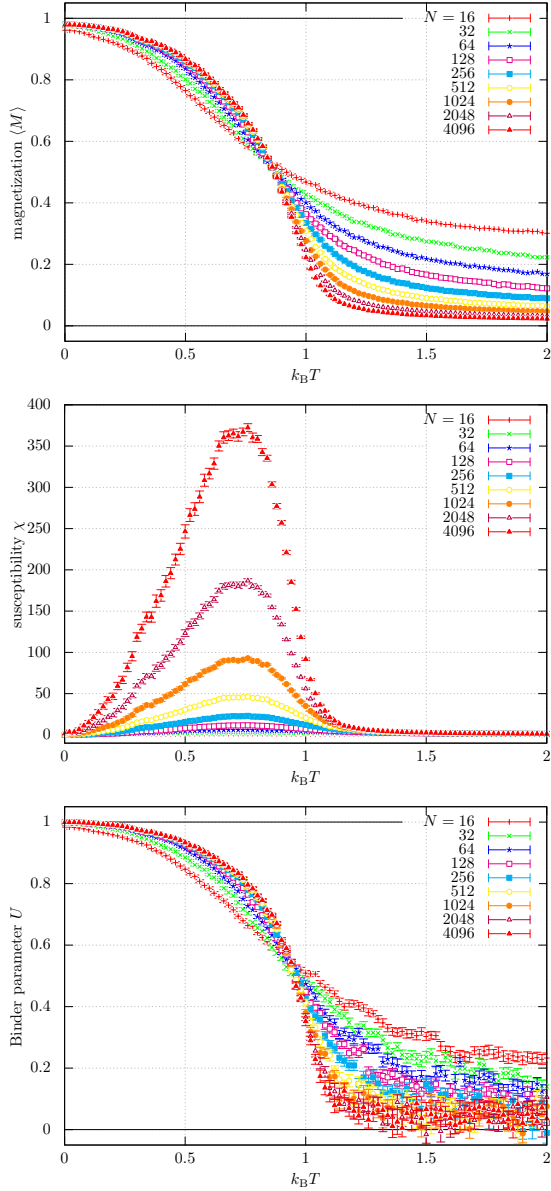


FIG. 3: Temperature dependence of (upper) the magnetization, Eq. (4), (middle) the susceptibility, Eq. (6), and (lower) the Binder parameter, Eq. (8), for $X \rightarrow YZ$. We set $q = 10^{-2.0}$. We also set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The length of the generated sentence by the language model, N , was varied from 16 to 4096.

when the system size is varied, which also supports the existence of a phase transition.

In Fig. 4, we plot the correlation functions for different conditions. For a fixed system size, the correlation function looks straight for $k_B T \sim 1.1$ [see Fig. 4(upper)]. Then, we turn our attention to the system-size dependence of the correlation function. As Fig. 4(lower) depicts, the correlation functions become straight for $k_B T \lesssim 1$.

In Fig. 5(upper), we draw the phase diagram of the

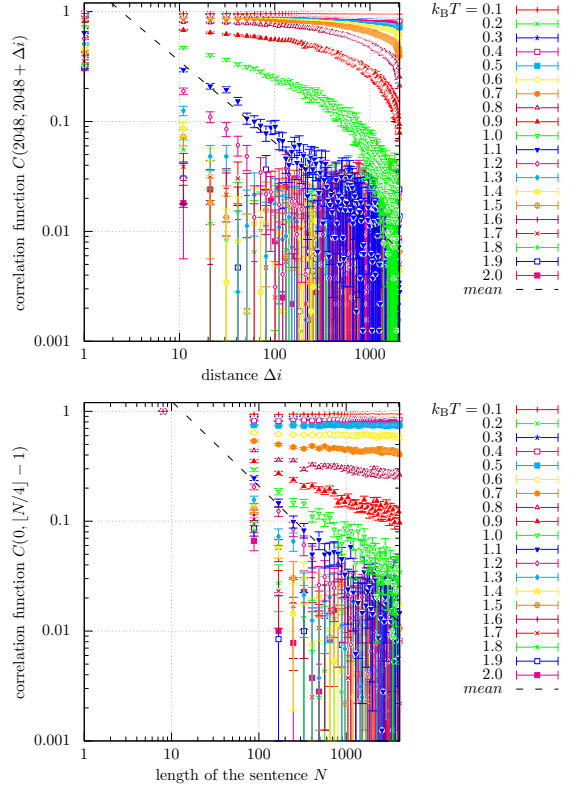


FIG. 4: Correlation functions, Eq. (9), (upper) with $i = 2048$ and $j = 2048 + \Delta i$ and (lower) with $i = 0$ and $j = \lfloor N/4 \rfloor - 1$ for $X \rightarrow YZ$. We set $K = 2$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. Temperature $k_B T$ was varied from 0.1 to 2.0. The dashed lines are fitted lines to the data for (upper) $\Delta i \in [10, 1000]$ and (lower) $N \in [100, 1000]$ at $k_B T = 1.1$.

context-sensitive random language model with $K = 2$ and $X \rightarrow YZ$. In this phase diagram, T_c , below which the susceptibility is divergent with respect to the system size, is plotted by red dots, and the fitting curve is computed via least squares regression with a quadratic function. In other words, the susceptibility is not divergent with respect to the system size above T_c . We conduct finite-size scaling on $\tilde{\chi}$ and plot the result in Fig. 5(middle). Finally, we show the q -dependence of the critical exponents ν and γ . This figure shows that ν increases with q but γ stays unchanged.

V. DISCUSSIONS

The results presented here demonstrate that a simple, context-sensitive generative language model—combining growth dynamics with long-range context-dependent interactions—can exhibit a BKT transition. This finding is notable for several reasons. First, classical discrete one-dimensional spin systems, even with long-range couplings, rarely display BKT behavior; such transitions typically require continuous symmetries or finely tuned

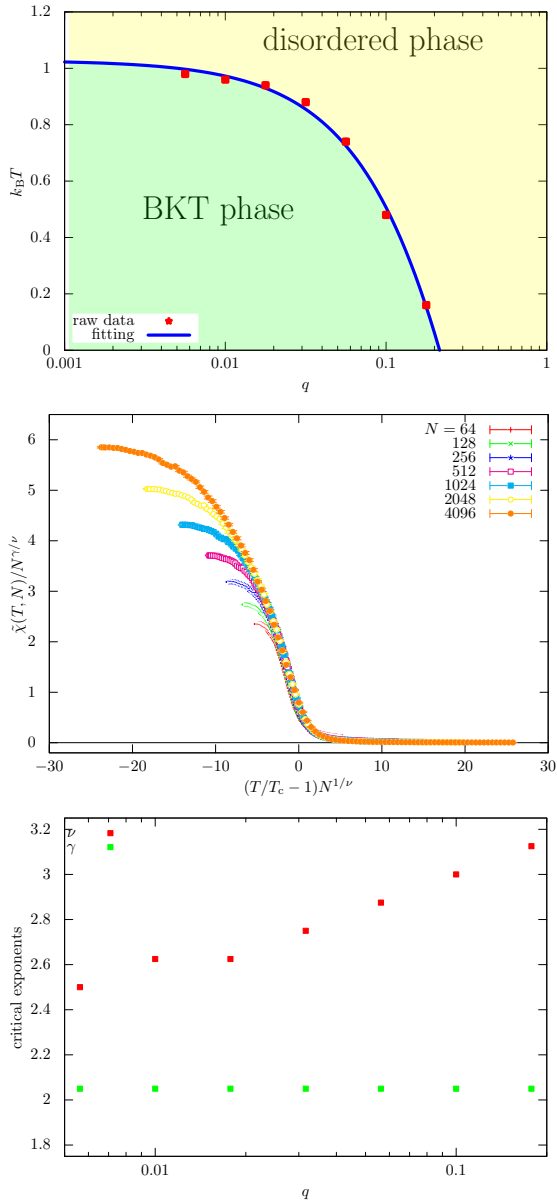


FIG. 5: (upper) Phase diagram of the context-sensitive random language model, where the horizontal and vertical axes are the growth rate of a sentence q and critical temperature $k_B T$, respectively. We consider $X \rightarrow YZ$ and set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. (middle) Finite-size scaling of $\tilde{\chi}$ at $q = 10^{-2.0}$. We set $T_c = 0.960$, $\nu = 2.6250$, and $\gamma = 2.05$, where the values of ν and γ are determined such that the scaling assumption holds. We varied $N = 64, 128, 256, 512, 1024, 2048, 4096$. (lower) q -dependence of the critical exponents ν and γ .

interactions. In our model, the emergence of BKT-like criticality arises not from finely adjusted symmetry but from the inherent structure of the generative process itself: context sensitivity and expansion jointly modify the effective dimensionality of the system. This suggests that linguistic generative processes lie naturally outside the universality classes usually invoked for one-dimensional

discrete models.

Second, the BKT transition observed here provides a principled explanation for why scaling phenomena in language appear robust across corpora, languages, and even across artificial generative systems such as LLMs. In BKT phases, scale-invariant behavior persists across a finite region rather than at a finely tuned critical point. This is consistent with the empirical observation that natural languages display power-law statistics without requiring external tuning, and that LLMs exhibit predictable scaling laws across multiple orders of magnitude. The present model shows that hierarchical growth, long-range dependencies, and context-aware constraints—all essential features of language—create a natural pathway to such extended criticality.

Third, the model highlights the role of grammar-like mechanisms in organizing long-range structure. While neural LLMs implement interactions through attention mechanisms and deep feedforward layers rather than explicit rewrite rules, both systems share the ability to propagate information across long ranges and adapt symbol generation to global context. The success of the simplified CSG-based model in generating BKT-like ordering suggests that long-range coherence in LLMs may arise from universal mechanisms rooted in context-sensitive generativity rather than purely neural architectural choices.

We next discuss in greater detail the numerical findings presented in the previous section, starting with the case of $K = 2$. The identified phase transition, reminiscent of the Ising model, entails the spontaneous breaking of $\mathbb{Z}_2$ symmetry, presenting a unique form of phase transition distinct from those previously discussed in CFGs, as highlighted in prior works such as Refs. [11–13] and similar studies. Our investigation suggests that the phase transition phenomenon in CSGs is particularly interesting compared to that of CFGs, which has led to the ongoing debate. Such a result implies that long-range interactions play a pivotal role in inducing a different type of phase transition in language models.

Furthermore, the critical nature of the discovered phase transition is crucial. When arbitrarily constructing systems that undergo phase transitions, it is common for transitions to be discontinuous (first-order). However, in such discontinuous transitions, the lack of nontrivial scaling laws makes them unsuitable for explaining phenomena like Zipf’s law, which is observed in natural languages.

It is important to note that the rules, such as Eq. (1b), involved in the studied statistical mechanics model are specific to language models and are not present in typical statistical mechanics models where the system does not grow over time. Therefore, the observed phase transition phenomenon in this study is not merely a repetition of calculations for the one-dimensional long-range Potts model in the context of natural language processing. The results of this study may offer new insights into the statistical mechanics of systems that spatially expand over

time, beyond language-specific frameworks.

It is important to examine what occurs in the case of alphabets with $K > 2$ distinct letters, as this is a feature of any natural language. Unlike the Potts model, where $K \geq 5$ exhibits a first-order phase transition, we observed the BKT transition in the context-sensitive random language model with $K = 10$. However, its temperature dependence is heavily influenced by numerical conditions, and the unique processes specific to the context-sensitive random language model, Eq. (1), contribute to a much richer phase diagram compared to conventional statistical-mechanical models.

We have also computed the correlation functions and mutual information for several numerical setups and observed the scaling laws. Due to the length-expanding nature of language models, there is some arbitrariness in the definition of correlation functions. Depending on this definition, the correlation functions exhibit power-law decay. This suggests the possibility of power laws occurring not only at the exact phase transition point but potentially throughout an entire phase.

The relative frequencies of letters in the alphabet are often studied in natural language processing. For $K = 10$ and $q = 10^{-2.0}$, we plot the relative frequencies of letters in the alphabet for typical configurations [see Fig. D52]. Multiple symbols appear in sentences in the case of $X \rightarrow YZ$, while one or two symbols dominate sentences in the case of $X \rightarrow XX$. In natural languages, Zipf's law is believed to hold; therefore, the case of $X \rightarrow YZ$ seems to be more natural, though the phenomenon observed in the case of $X \rightarrow XX$ looks quite similar to that of a first-order phase transition. However, further investigation may be required to reach a definitive conclusion, as there are several different branching processes, including those that exist between $X \rightarrow YZ$ and $X \rightarrow XX$.

In finite-size scaling, we determined the critical exponents such that the scaling assumption holds. It is important to note that the existence of such values is non-trivial, as there is no universal function in general. In Ref. [37], the critical exponents of the one-dimensional long-range Ising model are reported, but our observation is that the scaling assumption does not hold for these quantities. Furthermore, γ and ν do not depend on the choice between $X \rightarrow YZ$ and $X \rightarrow XX$; however, they depend on q and K . The phase diagrams heavily depend on the choice between $X \rightarrow YZ$ and $X \rightarrow XX$, making the independence of γ and ν from these choices an interesting feature. The fact that γ and ν depend on K seems natural since the number of states or flavors often affects renormalization flows. Contrary to the choice between $X \rightarrow YZ$ and $X \rightarrow XX$, γ and ν depend on q , which is a newly introduced parameter that controls the growth rate of a sentence.

The reader may find that the above discussions seem to contradict common understanding regarding the BKT transition. Typically, the BKT transition is studied within the context of the XY model, where it is believed that $\mathbb{Z}_2$ spontaneous symmetry breaking does not

occur [16–18]. In contrast, the one-dimensional long-range Ising model is known to exhibit $\mathbb{Z}_2$ spontaneous symmetry breaking [38, 45–48] alongside the BKT transition [19, 39]. In this sense, our study sheds light on the *unconventional* BKT transition and posits that it may capture the nature of language models.

Additionally, in deep learning models for natural language processing, attention mechanisms are considered crucial [49]. These attention mechanisms facilitate long-range interactions, aligning with the results of our study, which suggest that long-range interactions play a role in stabilizing phase transitions in language models. From the viewpoint of spin glass theory with the Potts model, the attention mechanism is investigated in Ref. [50], but it is yet to be known how it affects phase transitions. More recently, the possibility of a mirage of the emergent ability of LLMs was pointed out in Ref. [51]; thus, a similar investigation of neural language models and LLMs is one of the most important future directions.

It is also worth noting that Refs. [52–56] discuss phase transitions in neural language and diffusion models, which are indeed intriguing. However, the approach differs from ours; while our work focuses on identifying the minimal language model that exhibits phase transitions, Refs. [52–56] directly investigate neural language and diffusion models.

VI. CONCLUSIONS

In this study, we have constructed the context-sensitive random language model by combining the random language model developed by DeGiuli [11, 12] with the one-dimensional long-range Potts model, and we numerically demonstrated the presence and absence of phase transitions depending on numerical conditions. Moreover, we unveiled nontrivial phase transition phenomena specific to the language model rules. In the future, it will be intriguing to explore whether more nontrivial phase transitions, as discussed in Refs. [11, 13], can be realized in CSGs or their probabilistic extensions, akin to the transitions observed in this study. Additionally, by considering more realistic language models, we aim to elucidate the scaling laws observed in LLMs and the mechanisms behind emergent language capabilities.

Indeed, our study has shed light on the significance of long-range interactions in the phase transition of language models. However, the precise relationship with mechanisms such as attention mechanisms remains ambiguous. Future research will seek to elucidate the behavior of attention mechanisms in the long-range limit and establish connections between long-range interactions, attention mechanisms, phase transitions, scaling laws, and the emergence of language capabilities.

Acknowledgments

H.M. thanks Koji Hukushima, Kai Nakaishi, Sho Yokoi, Ryo Ueda, Tatsuki Kuribayashi, Takashi Mori, Masahito Mochizuki, Yusuke Miyajima, and Masaharu Yoshioka for fruitful discussions. J.T. thanks Yoshihiko Nishikawa, K. Nakaishi, K. Hukushima, S. Yokoi, R. Ueda, and T. Kuribayashi for insightful discussions too. H.M. was supported by JSPS KAKENHI Grant Number JP25H01499.

Author contributions

H.M. conceived the original idea. H.M. and J.T. designed the research. Y.T. and H.M. performed numerical simulations. All authors discussed the results. H.M., J.T., and V.R. wrote the paper.

Competing interests

The authors declare no competing financial interests.

Data availability

We will provide all the numerical codes and plot files upon request.

Appendix A: Overview of the appendices

The appendices, which are written in a self-contained manner, provide the supplementary information and are organized as follows. In Appendix B, we formulate the problem setting of this study. Appendix C describes some physical quantities computed in this paper. Appendix D presents supporting numerical simulations.

Appendix B: Language model

This section outlines the model under consideration in this study. With regard to the model, particular emphasis is placed on explaining the interactions between symbols. As for the interaction between symbols, we consider the one-dimensional Potts model with a long-range interaction and construct a new language model that falls under the class of CSGs, which we call the context-sensitive random language model. Note that the Potts model is identical to the Ising model when the number of states that each spin takes is two. Finally, we explain the context dependency of the constructed language model.

1. Proposed model

In this study, we consider a language model that significantly simplifies natural language, which we call the context-sensitive random language model [11–14, 57, 58]. In the language model, we begin with a randomly chosen symbol, which is a nonterminal symbol. Then, a symbol in the sentence is first uniformly chosen at random and one of the following three generation rules is applied to the symbol:

$$X \rightarrow x \quad (\text{probability: } qt), \quad (\text{B1a})$$

$$X \rightarrow YZ \quad (\text{probability: } q(1-t)), \quad (\text{B1b})$$

$$\alpha_- X \alpha_+ \rightarrow \alpha_- Y \alpha_+ \quad (\text{probability: } (1-q)). \quad (\text{B1c})$$

Here, uppercase letters X , Y , and Z represent nonterminal symbols, taking values $A_1, A_2, \dots, A_K$. Lowercase letters x , y , and z represent terminal symbols, taking values $a_1, a_2, \dots, a_K$. First, Eq. (B1a) represents rules where nonterminal symbols correspond to their respective terminal symbols. Second, Eq. (B1b) represents rules where a nonterminal symbol is replaced by two nonterminal symbols. Last, Eq. (B1c) introduces context-dependency to nonterminal symbols, rewriting them based on the context α_- and α_+ , which are strings before and after the symbol of interest, respectively. Furthermore, Y in Eq. (B1c), called the proposed sample in the Monte Carlo (MC) method, is chosen from a probability distribution that satisfies the detailed balance condition. In this paper, we run the Metropolis-Hastings algorithm to determine if this update is executed or not. We describe the details later.

While there may be room for discussion regarding the extent of context-dependency in natural languages, we adopt a CSG with a focus on strong context access in modern neural language models. It is worth noting that excluding Eq. (B1c) leads to a CFG, aligning with models discussed in previous studies [11–14, 57, 58]. The probabilities of adopting rules in Eqs. (B1a), (B1b), and (B1c) are denoted as qt , $q(1-t)$, and $1-q$, respectively. The parameter $t \rightarrow 0$ represents the limit in which terminal symbols do not appear, and this limit enables us to discuss phase transitions easily because the thermodynamic limit $N \rightarrow \infty$ can be taken straightforwardly. Then, in this paper, we focus on this limit for simplicity.

2. One-dimensional long-range Potts model

In Refs. [30–37], phase transitions in the following one-dimensional long-range Potts model¹⁰ are discussed [59]:

$$H = -\frac{1}{2} \sum_{\substack{i,j=1,2,\dots,N: \\ i \neq j}} J_{i,j} \delta_{\sigma_i, \sigma_j}, \quad (\text{B2})$$

¹⁰ When $K = 2$, this model reduces to the Ising model.

where

$$J_{i,j} = \begin{cases} \frac{J}{|i-j|^{1+s}} & (j-i \in [-Nr_-, Nr_+]), \\ 0 & (\text{otherwise}). \end{cases} \quad (\text{B3})$$

Here $r_+, r_- \in [0, 1]$ are parameters that specify the range of interaction. Note that Eq. (B2) is referred to as the ferromagnetic model because symbols are more likely to have the same state for $J \gg 0$. Additionally, when $X_j = A_k, a_k$ for $k = 1, 2, \dots, K$, σ_j is defined as

$$\sigma_j = k. \quad (\text{B4})$$

For pedagogical purposes, we describe the case of $K = 2$ in Appendix B3.

3. One-dimensional long-range Ising model

In the literature [60], each σ_i takes $\{1, -1\}$ rather than $\{0, 1\}$ for $K = 2$, i.e., the Ising model; then, we follow the notation in this subsection. Therefore, Eq. (B2) with $K = 2$ reads

$$H = -\frac{1}{4} \sum_{\substack{i,j=1,2,\dots,N: \\ i \neq j}} J_{i,j} \sigma_i \sigma_j, \quad (\text{B5})$$

and Eq. (B4) is transformed into

$$\sigma_j := \begin{cases} 1 & (X_j = A_1, a_1), \\ -1 & (X_j = A_2, a_2). \end{cases} \quad (\text{B6})$$

For Monte Carlo updates, the change of the energy of a system by a spin flip is crucially important; Eq. (B10) for $\sigma_0 \rightarrow \tilde{\sigma}_0 = -\sigma_0$ becomes

$$\Delta E := -\frac{J}{2} \left(\sum_{\ell=1}^{\lfloor Nr_- \rfloor} \frac{(\tilde{\sigma}_0 - \sigma_0) \sigma_{-\ell}}{\ell^{1+s}} + \sum_{\ell=1}^{\lfloor Nr_+ \rfloor} \frac{(\tilde{\sigma}_0 - \sigma_0) \sigma_\ell}{\ell^{1+s}} \right) \quad (\text{B7})$$

$$= J \left(\sum_{\ell=1}^{\lfloor Nr_- \rfloor} \frac{\sigma_0 \sigma_{-\ell}}{\ell^{1+s}} + \sum_{\ell=1}^{\lfloor Nr_+ \rfloor} \frac{\sigma_0 \sigma_\ell}{\ell^{1+s}} \right). \quad (\text{B8})$$

Next, we review the studies on phase transitions of the one-dimensional long-range Ising model (the one-dimensional long-range Potts model with $K = 2$) (B2). For this model, phase transitions occur for $s \in (0, 1)$ while phase transitions do not occur for $s > 1$ [30, 32, 33, 36, 37]. In Ref. [31], the existence of a phase transition for $s = 1$ was reported. Furthermore, when $s = 1$, it is known that the BKT transition occurs [19, 39].

There are numerous studies on critical exponents for various s , which is called the decay exponent in this paper [61]. For $s \in (0, 0.5)$, it has been reported that the phase transition follows the mean-field approximation and exhibits a jump in the specific heat [62]. Additionally, the critical exponent is noted to be independent of

temperature in this range [40]. For $s \in (0.5, 1.0)$, a sharp peak in the specific heat suggests a phase transition distinct from that predicted by the mean-field approximation [62]. In this paper, we consider the decay exponent s within this range for simplicity. Furthermore, at $s = 1$, there is a jump in magnetization [33, 34, 41]. It is also worth noting that an effective MC method was proposed in Ref. [63].

4. Context-dependency

Let us elaborate on Eq. (B1c). Let N be the length of the entire string, and define $\alpha_- := X_{-\lfloor Nr_- \rfloor} X_{-(\lfloor Nr_- \rfloor - 1)} \dots X_{-2} X_{-1}$ and $\alpha_+ := X_1 X_2 \dots X_{\lfloor Nr_+ \rfloor - 1} X_{\lfloor Nr_+ \rfloor}$. Here, X represents some character in the alphabet, and the subscript values denote the relative positions with respect to $A_1, A_2, \dots, A_K$ in Eq. (B1c). The floor function $\lfloor x \rfloor$ indicates the greatest integer that is equal to or smaller than $x \in \mathbb{R}$, and r_- and r_+ are parameters corresponding to the interaction range of the one-dimensional long-range Potts model.

Let $i = 0$ be the location of X . Motivated by the Metropolis-Hastings algorithm, when Eq. (B1c) is chosen, symbols are changed with the following probability:

$$p = \min(1, e^{-\frac{\Delta E}{k_B T}}), \quad (\text{B9})$$

where ΔE for $\sigma_0 \rightarrow \tilde{\sigma}_0$ is defined as

$$\Delta E := J \left(\sum_{\ell=1}^{\lfloor Nr_- \rfloor} \frac{\delta_{\sigma_0, \sigma_{-\ell}} - \delta_{\tilde{\sigma}_0, \sigma_{-\ell}}}{\ell^{1+s}} + \sum_{\ell=1}^{\lfloor Nr_+ \rfloor} \frac{\delta_{\sigma_0, \sigma_\ell} - \delta_{\tilde{\sigma}_0, \sigma_\ell}}{\ell^{1+s}} \right). \quad (\text{B10})$$

Note that Eq. (B10) is computed from Eq. (B2): $\Delta E = \tilde{H} - H$.

According to Eq. (B9), symbols are more likely to flip when ΔE is small. Furthermore, from Eq. (B10), ΔE decreases when the numerators of the first and second terms of the right-hand side of Eq. (B10) take negative values. Considering Eq. (B4), this means symbols are flipped when the surrounding symbols are of a different type. The parameter $k_B T$ controls the tendency of the same symbols to align and corresponds to the strength of the desire in statistical mechanics to align with the absolute temperature.

Equation (B10) is termed a long-range interaction due to the power-law decay ($1/\ell^{1+s}$) of interactions between symbols. The decay exponent s determines the strength of the power-law decay, and in this paper, we fix $s = 0.9$. In summary, the rule in Eq. (B1c) is statistically equivalent to the Metropolis-Hastings algorithm for flipping a single symbol in the one-dimensional long-range Potts model, as applied in MC simulations.

Appendix C: Methods: physical quantities for discussing phase transitions and finite-size scaling

Phase transitions are discerned through the singularity of order parameters, which represent the expectation value of a function of random variables. In models like the Potts model, magnetization, denoting the ratio of the difference between the numbers of up spins and down spins to the sum of these numbers, is commonly employed as an order parameter. Subsequently, we examine the second-order moment of the random variable, known as the susceptibility. At critical points, the susceptibility typically diverges, making it crucial to observe its behavior to confirm phase transitions. More interestingly, the susceptibility is always divergent in the BKT phase. Last, the Binder parameter, akin to the normalized version of kurtosis, plays a crucial role in discussions of phase transitions. According to a generalized version of the *central limit theorem*, the Binder parameter is equal to one below the critical temperature and zero above it. Furthermore, in the case of second-order phase transitions, the lines depicting the temperature dependence of the Binder parameter for fixed system sizes intersect at the critical temperature.

To explore the scaling law, which is known to apply at critical points, we also compute the correlation functions and mutual information. It is important to note that due to the inherent growth property of language models, there is some arbitrariness in defining the correlation functions and mutual information.

At the end of this section, we review the finite-size scaling method, which is a powerful method based on the scaling assumption to discuss and quantify phase transitions. In particular, critical exponents are used to classify phase transitions into universality classes.

1. Magnetization

Although this study does not deal with magnetic materials, an order parameter similar to that in the Potts model can be defined when considering two symbols. We term this order parameter “magnetization,” and define it as follows:

$$M := \|\mathbf{M}\|, \quad (\text{C1})$$

where $\mathbf{M} \in \mathbb{R}^{K-1}$ is given by

$$\mathbf{M} := \frac{1}{N} \sum_{i=1}^N \mathbf{e}_{\sigma_i}. \quad (\text{C2})$$

Note that $\|\mathbf{M}\| := \sqrt{\mathbf{M} \cdot \mathbf{M}}$. The construction of $\{\mathbf{e}_k\}_{k=1}^K$ in Eq. (C1) is shown in Algorithm 1. Due to its construction in Algorithm 1, $\{\mathbf{e}_k\}_{k=1}^K$ satisfy $\|\mathbf{e}_k\| = 1$ for $k = 1, 2, \dots, K$, $\mathbf{e}_k \cdot \mathbf{e}_l = \text{const.}$ for $k \neq l$, and $\sum_{k=1}^K \mathbf{e}_k = \mathbf{0}$. Then the l -th element of $\{\mathbf{e}_k\}_{k=1}^K$ is com-

ALGO. 1: Construction of $\{\mathbf{e}_k\}_{k=1}^K$ in Eq. (C1).

```

1 set  $\mathbf{e}_1 := [1, 0, \dots, 0]^\top$ 
2 for  $k = 2, 3, 4, \dots, K$  do
3   initialize  $\mathbf{e}_k := [0, 0, \dots, 0]^\top$ 
4   for  $l = 1, 2, 3, \dots, k-1$  do
5      $[\mathbf{e}_k]_l = -\frac{1}{K-l}[\mathbf{e}_l]_l$ 
6    $[\mathbf{e}_k]_k = \sqrt{1 - \sum_{l=1}^k ([\mathbf{e}_k]_l)^2}$ 

```

puted as

$$[\mathbf{e}_k]_l := \begin{cases} \sqrt{\frac{K(K-k)}{(K-1)(K-k+1)}} & (k = l), \\ -\sqrt{\frac{K}{(K-1)(K-l)(K-l+1)}} & (k > l), \\ 0 & (k < l). \end{cases} \quad (\text{C3})$$

The detailed derivation of Eq. (C3) is shown in Appendix C2. Note that $\langle \mathbf{M} \rangle = \mathbf{0}$ because of the construction of $\{\mathbf{e}_k\}_{k=1}^K$.

Finally, we consider the meaning of M , as defined in Eq. (C1). It signifies the degree of frequency bias among nonterminal symbols $A_1, A_2, \dots, A_K$. In the subsequent analysis, by observing the variations of M , Eq. (C1), with respect to $k_B T$, we aim to ascertain whether a critical point exists where the properties of the language, constructed based on the one-dimensional Potts model, undergo a qualitative and abrupt change. At the critical point of the phase transition, the magnetization, as defined in Eq. (C1), becomes non-differentiable with respect to a control parameter, which is temperature in this paper, in the thermodynamic limit $N \rightarrow \infty$.

2. Derivation of Eq. (C3)

The derivation of Eq. (C3) can be shown via mathematical induction. That is, $[\mathbf{e}_{k+1}]_{k+1}$ is computed as follows:

$$[\mathbf{e}_{k+1}]_{k+1} = \sqrt{1 - \sum_{l=1}^k [\mathbf{e}_{k+1}]_l^2} \quad (\text{C4})$$

$$= \sqrt{1 - \sum_{l=1}^k \left(-\sqrt{\frac{K}{(K-(l-1))(K-1)(K-l)}} \right)^2} \quad (\text{C5})$$

$$= \sqrt{1 - \sum_{l=1}^k \frac{K}{(K-(l-1))(K-1)(K-l)}} \quad (\text{C6})$$

$$= \sqrt{1 - \frac{K}{K-1} \sum_{l=1}^k \frac{1}{(K-(l-1))(K-l)}} \quad (\text{C7})$$

$$= \sqrt{1 - \frac{K}{K-1} \sum_{l=1}^k \left(\frac{1}{K-l} - \frac{1}{K-(l-1)} \right)} \quad (\text{C8})$$

$$= \sqrt{1 - \frac{K}{K-1} \left(\frac{1}{K-k} - \frac{1}{K} \right)} \quad (\text{C9})$$

$$= \sqrt{1 - \frac{k}{(K-1)(K-k)}} \quad (\text{C10})$$

$$= \sqrt{\frac{K(K-k-1)}{(K-1)(K-k)}}. \quad (\text{C11})$$

Using the value of $[\mathbf{e}_{k+1}]_{k+1}$, Eq. (C11), and the two conditions of $\{\mathbf{e}_k\}_k$, we can compute $[\mathbf{e}_{k+1}]_l$ for $k+1 > l$, and the other elements are zero by definition. Thus, Eq. (C3) holds for $k+1$.

3. Susceptibility

Next, we define the susceptibility as follows:

$$\chi := N(\langle M^2 \rangle - \langle M \rangle^2). \quad (\text{C12})$$

The susceptibility, Eq. (C12), is known to diverge just above the critical point and is well-studied due to its ease of experimental measurement, e.g., in the q -state clock model [64, 65]. We also note that the following form is another candidate for defining the susceptibility:

$$\tilde{\chi} := N(\langle \mathbf{M} \cdot \mathbf{M} \rangle - \langle \mathbf{M} \rangle \cdot \langle \mathbf{M} \rangle) \quad (\text{C13})$$

$$= N\langle \mathbf{M} \cdot \mathbf{M} \rangle \quad (\text{C14})$$

$$= N\langle M^2 \rangle. \quad (\text{C15})$$

However, Eq. (C15) is known to be valid only for the disordered phase.

4. Binder parameter

The non-differentiability of the magnetization, as expressed in Eq. (C1), and the divergence of susceptibility, as defined in Eq. (C12), are fundamental characteristics of phase transitions. However, in numerical experiments, encountering challenges in taking the thermodynamic limit $N \rightarrow \infty$ due to finite-system sizes is common. Consequently, analysis using the Binder parameter is often preferred [42, 66]. The Binder parameter is defined as follows:

$$U := -\frac{K-1}{2} \left(\frac{\langle (\mathbf{M} \cdot \mathbf{M})^2 \rangle}{\langle \mathbf{M} \cdot \mathbf{M} \rangle^2} - \frac{K^2-1}{(K-1)^2} \right) \quad (\text{C16})$$

$$= -\frac{K-1}{2} \left(\frac{\langle M^4 \rangle}{\langle M^2 \rangle^2} - \frac{K+1}{K-1} \right). \quad (\text{C17})$$

It is noteworthy that when $\mathbf{M}$ is sampled from a Gaussian distribution with K degrees of freedom, $\langle M^4 \rangle / \langle M^2 \rangle^2$

evaluates to $K^2 - 1 / (K - 1)^2$. Consequently, Eq. (C17) equals zero for the disordered state and one when all the spin variables assume the same value. It is known that at the critical point, the temperature dependence of the Binder parameter, as defined in Eq. (C17), for different system sizes intersects. Moreover, it is also recognized that it diverges negatively at non-critical phase transition points, that is, the first-order phase transition. Hence, the method of investigating the system size dependence of the Binder parameter, as defined in Eq. (C17), has become standard for identifying phase transition points in numerical computations. From a statistical perspective, the Binder parameter is analogous to kurtosis, as it quantifies the tailedness of probability distributions.

5. Correlation function

By using $\{\mathbf{e}_k\}_{k=1}^K$ in Eq. (C3), we define the following correlation function:

$$G(i, j) := \langle (\mathbf{e}_{\sigma_i} - \langle \mathbf{e}_{\sigma_i} \rangle) \cdot (\mathbf{e}_{\sigma_j} - \langle \mathbf{e}_{\sigma_j} \rangle) \rangle \quad (\text{C18})$$

$$= \langle \mathbf{e}_{\sigma_i} \cdot \mathbf{e}_{\sigma_j} \rangle - \langle \mathbf{e}_{\sigma_i} \rangle \cdot \langle \mathbf{e}_{\sigma_j} \rangle. \quad (\text{C19})$$

In Eq. (C19), the following quantity is an essential part:

$$\tilde{G}(i, j) := \langle \mathbf{e}_{\sigma_i} \cdot \mathbf{e}_{\sigma_j} \rangle. \quad (\text{C20})$$

Similar to Eq. (C15), we compute Eq. (C20) instead of Eq. (C19). After some algebra, we obtain the following correlation function from Eq. (C20) [67]:

$$\tilde{G}(i, j) = \frac{K \langle \delta_{\sigma_i, \sigma_j} \rangle - 1}{K - 1}. \quad (\text{C21})$$

The correlation function is considered to be important for the research on phase transitions since it decays polynomially at critical points, at which second-order phase transitions take place, and decays exponentially otherwise.

6. Mutual information

Mutual information is a measure that quantifies the correlation between two random variables [2, 15]:

$$I(i, j) = \sum_{\sigma_i, \sigma_j=1}^K p(\sigma_i, \sigma_j) \ln \frac{p(\sigma_i, \sigma_j)}{p(\sigma_i)p(\sigma_j)}. \quad (\text{C22})$$

In Ref. [2], the scaling law of mutual information for CFGs was reported.

7. Histogram of the magnetization

For ΔM denoting the discretization interval of the magnetization and n an integer, the histogram of the magnetization, Eq. (C1), is defined as

$$\text{hist}(M) := \#\{\mathbf{M}_i | \mathbf{M}_i \in [n\Delta M, (n+1)\Delta M)\}, \quad (\text{C23})$$

where M_i is the magnetization, Eq. (C1), of the i -th MC sample for $i = 1, 2, \dots, N_{\text{MC}}$ and N_{MC} is the number of MC samples.

In second-order phase transitions, the histogram of the magnetization, as defined in Eq. (C23), transitions from a distribution with a peak at $M = 0$ for high $k_B T$ above the critical temperature to a uniform distribution at the critical point. Subsequently, it transitions to a distribution with a peak at $M = 1$ for low $k_B T$ below the critical temperature. However, in first-order phase transitions, the histogram of the magnetization, as defined in Eq. (C23), exhibits two peaks at the transition temperature. This phenomenon is attributed to hysteresis.

8. Finite-size scaling

In thermodynamics, we are interested in physical quantities in the thermodynamic limit $N \rightarrow \infty$; however, it is obviously impossible to take this limit in numerical simulations. Finite-size scaling is one of the numerical methods based on the scaling assumption to infer the behavior of physical quantities in the thermodynamic limit from numerical results of finite-size systems by extrapolating physical quantities of several system sizes.

Let us consider the system size and temperature dependence of a certain physical quantity A : $A(N, T)$. In general, we can consider any parameter instead of temperature T or even multiple parameters at the same time. At $T \sim T_c$, $A(N, T)$ is empirically known to behave in the thermodynamic limit as follows:

$$A(N, T) = t^{-x_A}, \quad (\text{C24})$$

where x_A is the critical exponent of A and

$$t := \left| \frac{T_c - T}{T_c} \right|. \quad (\text{C25})$$

Similarly, the correlation length ξ_∞ diverges as

$$\xi_\infty \propto t^{-\nu}. \quad (\text{C26})$$

where ν is the critical exponent of the correlation length. The scaling assumption states that there exists a function $f_A(\cdot)$ for $A(N, T)$ such that

$$A(N, T) = N^{\frac{x_A}{\nu}} f_A\left(\frac{N}{\xi_\infty}\right). \quad (\text{C27})$$

Let us take $\tilde{\chi}$, Eq. (C15), as an example. Equation (C24) for it reads

$$\tilde{\chi}(T) \propto t^{-\gamma}, \quad (\text{C28})$$

where γ is the critical exponent for $\tilde{\chi}$, Eq. (C15). The scaling assumption, Eq. (C27), for $\tilde{\chi}$, Eq. (C15), becomes

$$\frac{\tilde{\chi}(T, N)}{N^{\frac{\gamma}{\nu}}} = f_{\tilde{\chi}}(N t^\nu) \quad (\text{C29})$$

$$= \tilde{f}_{\tilde{\chi}}(N^{\frac{1}{\nu}} t). \quad (\text{C30})$$

Note that $\tilde{f}_{\tilde{\chi}}(\cdot)$ is also a universal function for $\tilde{\chi}$, Eq. (C15). We verify Eq. (C30) in the main text.

Appendix D: Numerical simulations

We conduct Monte Carlo simulations 80×80 times and present the numerical results of the physical quantities defined in Appendix C for the proposed model, as described in Eq. (B1). Additionally, considering the rule $X \rightarrow YZ$, in which symbols are added with equal probability, from Eq. (B1b), we also examine the case $X \rightarrow XX$, in which symbols are duplicated, because we aim to investigate how the rule, in which a sentence grows, unique to language models affects physical quantities in the thermodynamic limit. Here, we investigate the cases of $K = 2$ and $K = 10$, setting $J = 1.0$. This is because the Potts model is known to exhibit a second-order phase transition for $K \leq 4$ and a first-order phase transition for $K \geq 5$, suggesting that the proposed model is also expected to show qualitatively different behavior for $K = 2$ and $K = 10$.

First, we show the magnetization, the susceptibility, and the Binder parameter. Second, we plot the correlation functions and mutual information. Third, we depict the histogram of the magnetization.

Next, we show the system-size dependence of the magnetization, the two susceptibilities, and the Binder parameter. We then draw the phase diagrams for several numerical conditions. We see that the phase diagram drastically changes depending on $X \rightarrow YZ$ and $X \rightarrow XX$. We then plot the results of the finite-size scaling method. We see that the critical exponents are independent of the choice of $X \rightarrow YZ$ or $X \rightarrow XX$ but vary depending on q . Finally, we plot the q -dependence of the critical exponents ν and γ .

For $K = 10$, we show the relative frequencies of letters in the alphabet. In the literature on natural language processing, the relative frequencies of letters in the alphabet are known to often satisfy Zipf's law.

1. Case of $K = 2$ and $X \rightarrow YZ$

Let us consider the case of $K = 2$ and $X \rightarrow YZ$, in which symbols are added with equal probability. Since there are two nonterminal symbols and they increase randomly, there are eight processes to consider:

$$\begin{aligned} A &\rightarrow AA, A \rightarrow AB, A \rightarrow BA, A \rightarrow BB, \\ B &\rightarrow AA, B \rightarrow AB, B \rightarrow BA, B \rightarrow BB. \end{aligned} \quad (\text{D1})$$

That is, it is assumed that these processes occur with equal probability for A and B .

a. Magnetization, susceptibilities, and Binder parameter

In Fig. D6, the temperature dependence of the magnetization, Eq. (C1), for $q = 10^{-2.0}, 10^{-0.5}$ is shown. In the case of $q = 10^{-2.0}$, as the system size N increases below the transition point $T \lesssim 1$, the magnetization converges

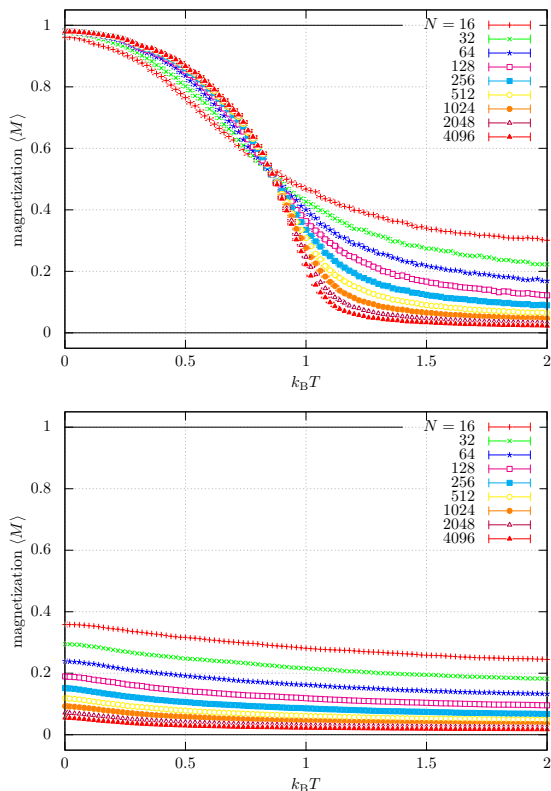


FIG. D6: Temperature dependence of the magnetization, Eq. (C1), for $X \rightarrow YZ$. We set (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-0.5}$. We also set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The length of the generated sentence by the language model, N , was varied from 16 to 4096. The upper panel is identical to Fig. 3(upper).

to a nonzero value, while above the transition point, it converges to zero. This behavior is consistent with the critical $\mathbb{Z}_2$ symmetry breaking observed alongside the intersection of the Binder parameters. However, in the case of $q = 10^{-0.5}$, $\langle M \rangle$ increases toward $k_B T = 0$, but it seems to hover around zero for any temperature. Therefore, Fig. D6 implies that the critical behaviors of the context-sensitive random language model, as described in Eq. (B1), strongly depend on the value of q .

Next, we show the susceptibility, Eq. (C12), in Fig. D7. Figure D7 seems to be consistent with Fig. D6 since it illustrates a diverging behavior of the susceptibility, typical for a second-order phase transition, at a specific temperature in the case of $q = 10^{-2.0}$. However, it does not show a cusp at its peak, which is a typical signal of the second-order phase transition. However, it monotonically increases toward zero temperature without exhibiting any peaks for $q = 10^{-0.5}$. This result also implies that there is no phase transition for $q = 10^{-0.5}$.

In Fig. D8, the Binder parameter, Eq. (C17), is shown. In Fig. D8(upper), a clear intersection of the Binder parameter is visible, indicating the presence of a critical point around $k_B T \sim 1.0$, while there is no crossing of the

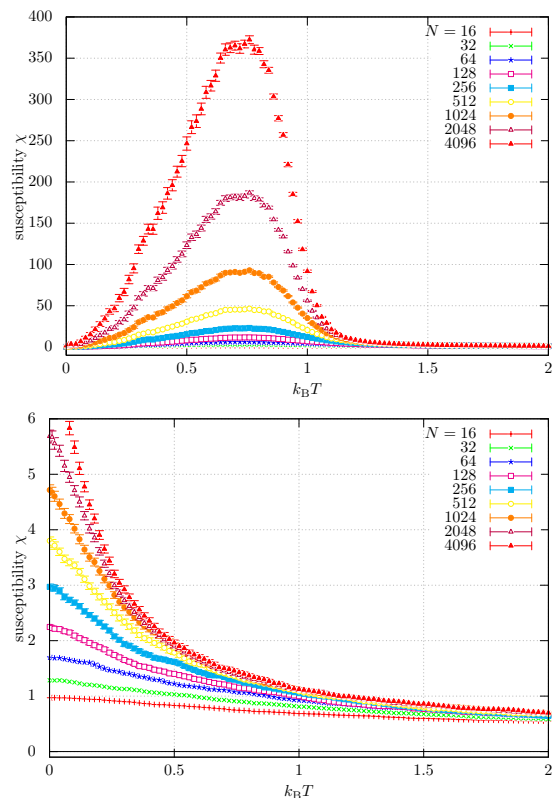


FIG. D7: Temperature dependence of the susceptibility, Eq. (C12), for $X \rightarrow YZ$. We set (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-0.5}$. We also set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The length of the generated sentence by the language model, N , was varied from 16 to 4096. The upper panel is identical to Fig. 3(middle).

Binder parameter in Fig. D8(lower).

b. Correlation functions, mutual information, and histogram of magnetization for $q = 10^{-2.0}$

In Fig. D9, we display two types of correlation functions. The first one, defined by Eq. (C20) with $i = 2048$ and $j = 2048 + \Delta i$, quantifies the correlation between symbols i and j within a given temporal sentence. It is equivalent to the conventional correlation function studied in research on spin systems. The second type, defined by Eq. (C20) with $i = 0$ and $j = \lfloor N/4 \rfloor - 1$, measures how symbols correlate with each other throughout the generation process. This type of correlation function is especially pertinent to models, in which the length of a generated sentence N grows, and cannot be defined for conventional spin models. Figure D9(upper) illustrates that for high $k_B T$, the correlation function decays rapidly with Δi , while for small $k_B T$, there is still correlation even for large Δi . More interestingly, the correlation function appears to satisfy the scaling behavior around $k_B T \sim 1.1$. In Fig. D9(lower), the scaling law seems to

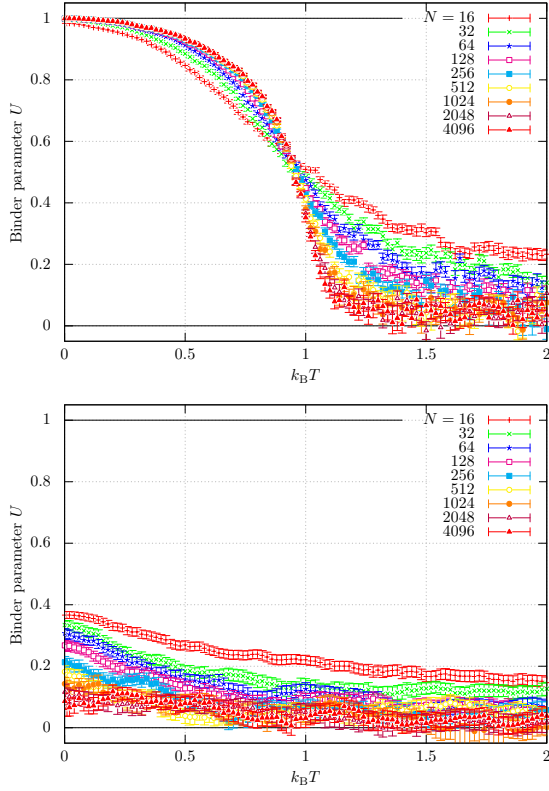


FIG. D8: Temperature dependence of the Binder parameter, Eq. (C17), for $X \rightarrow YZ$. We set (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-0.5}$. We also set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The length of the generated sentence by the language model, N , was varied from 16 to 4096. The upper panel is identical to Fig. 3(lower).

hold for $T \in [0.7, 1.1]$.

In Fig. D10, we plot mutual information, Eq. (C22). Similar to Fig. D9(upper), the scaling law seems to hold for $k_B T \sim 1.0$.

Finally, we show the histogram of the magnetization, Eq. (C23), in Fig. D11. The histogram of the magnetization, Eq. (C23), in Fig. D11 is almost uniform at $k_B T \sim 0.9$, and this behavior is typical for a second-order phase transition. For smaller $k_B T \sim 0.7$, it has a sharp peak around 1, and for larger $k_B T \sim 1.4$, it has a sharp peak around zero. Up to now, we can numerically conclude that this phase transition is not a first-order phase transition. As discussed later, we will conclude that this is a BKT transition, which is similar to a second-order phase transition.

c. Correlation functions, mutual information, and histogram of magnetization for $q = 10^{-0.5}$

In Figs. D12, D13, and D14, we present the correlation functions, Eq. (C20), mutual information, Eq. (C22), and the magnetization, Eq. (C23), respectively. These fig-

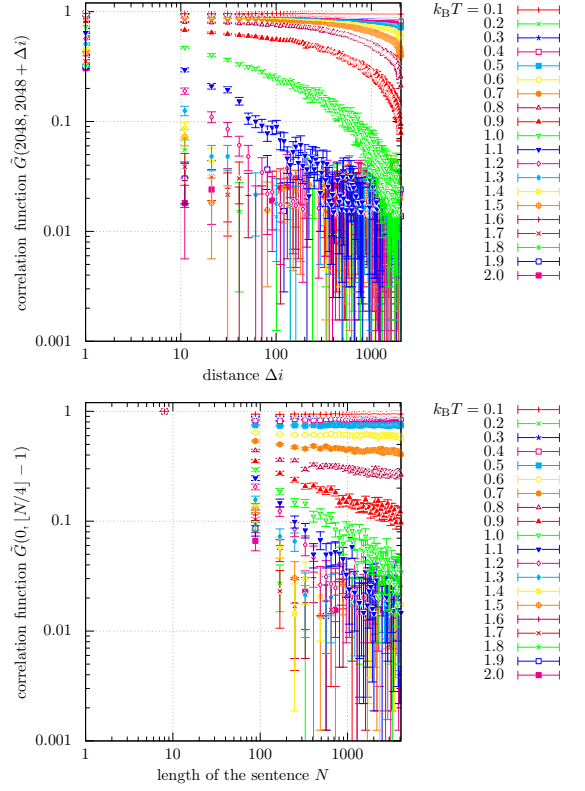


FIG. D9: Correlation functions, Eq. (C20), (upper) with $i = 2048$ and $j = 2048 + \Delta i$ and (lower) with $i = 0$ and $j = \lfloor N/4 \rfloor - 1$ for $X \rightarrow YZ$. We set $K = 2$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0. This figure is identical to Fig. 4.

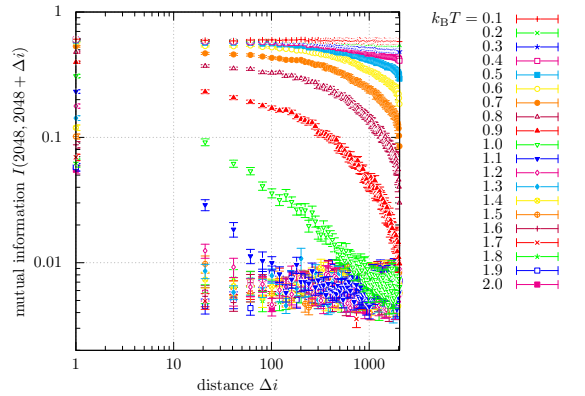


FIG. D10: Mutual information, Eq. (C22), with $i = 2048$ and $j = 2048 + \Delta i$ for $X \rightarrow YZ$. We set $K = 2$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

ures are all consistent with the estimate that the system is in the trivial “disordered” state.

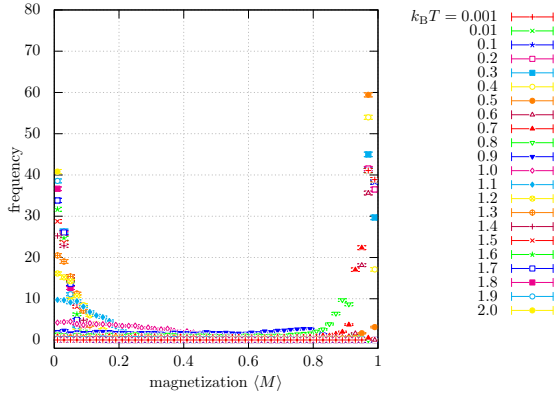


FIG. D11: Histogram of the magnetization, Eq. (C1), for $X \rightarrow YZ$. We set $K = 2$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

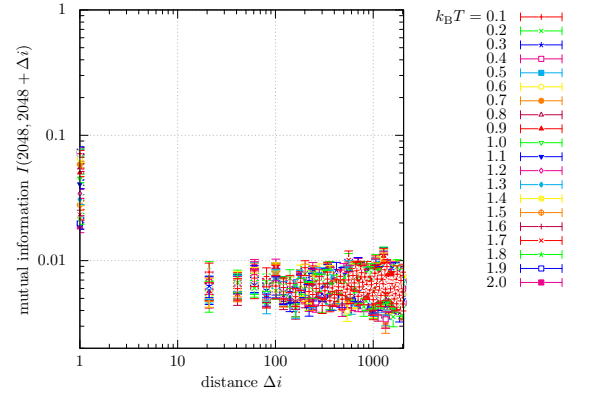


FIG. D13: Mutual information, Eq. (C22), with $i = 2048$ and $j = 2048 + \Delta i$ for $X \rightarrow YZ$. We set $K = 2$, $J = 1.0$, $q = 10^{-0.5}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

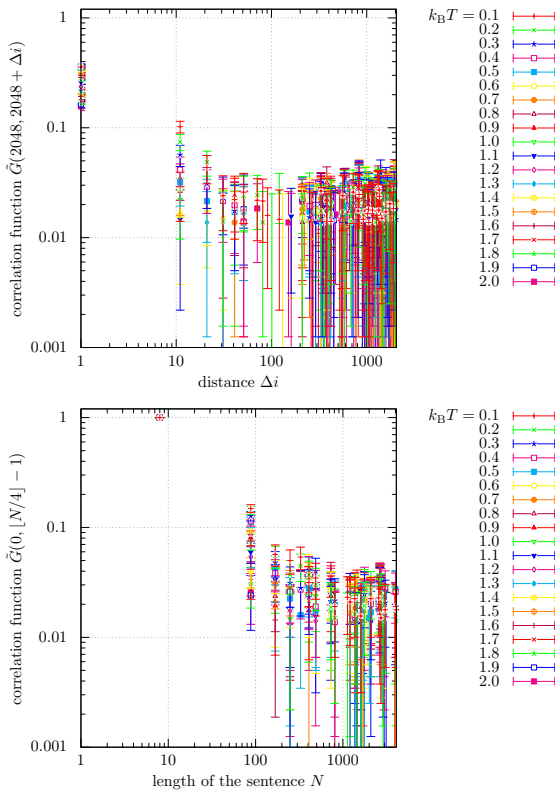


FIG. D12: Correlation functions, Eq. (C20), (upper) with $i = 2048$ and $j = 2048 + \Delta i$ and (lower) with $i = 0$ and $j = \lfloor N/4 \rfloor - 1$ for $X \rightarrow YZ$. We set $K = 2$, $J = 1.0$, $q = 10^{-0.5}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

d. System-size dependence of the magnetization, the susceptibilities, and the Binder parameter

In Figs. D15, D16, and D17, we show the system-size dependence of the magnetization, Eq. (C1), the two sus-

ceptibilities, Eqs. (C12) and (C15), and the Binder parameter, Eq. (C17), at $q = 10^{-2.0}$, respectively. The critical temperature was determined by assessing whether the line of the inverse of the susceptibilities in Fig. D16 is linear or not. This estimate is consistent with the temperature at which the Binder parameter takes a non-zero value in the thermodynamic limit, as shown in Fig. D17.

e. Phase diagram, finite-size scaling, and q -dependence of critical exponents

In Fig. D18, we plot the phase diagram of our language model with $K = 2$ and $X \rightarrow YZ$. We determined the boundary between the two phases by fitting a quadratic function via least squares regression. This figure shows that the BKT phase exists for $k_B T \lesssim 0.20$.

In Fig. D19, we perform finite-size scaling on $\tilde{\chi}$ at $q = 10^{-2.0}$. We set $T_c = 0.960$, $\nu = 2.6250$, and $\gamma = 2.05$, where we determine the values of ν and γ such that the

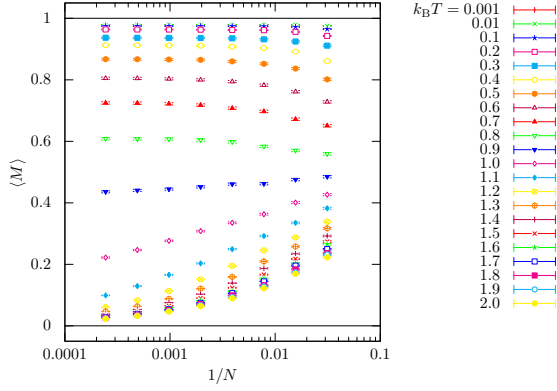


FIG. D15: System-size dependence of the magnetization, Eq. (C1), at $q = 10^{-2.0}$. We consider $X \rightarrow YZ$ and set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 32, 64, 128, 256, 512, 1024, 2048, 4096$.

scaling assumption holds; this figure shows that the scaling assumption holds.

In Fig. D20, we set $T_c = 0.960$, $\nu = 2.29$, and $\gamma = 2.05$, where ν and γ are taken from Ref. [37]; the critical exponents are different from those of the one-dimensional long-range Ising model. More specifically, the critical exponents depend on the value of q .

In Fig. D21, we plot the q -dependence of the critical exponents ν and γ . By observing these robust critical phenomena over a wide range of the parameter space, we conclude that the BKT transition is occurring in the proposed model, Eq. (B1), rather than a second-order phase transition.

2. Case of $K = 2$ and $X \rightarrow XX$

We consider the case of $K = 2$ and $X \rightarrow XX$, in which symbols are duplicated. Since there are two nonterminal symbols, there are two processes:

$$A \rightarrow AA, \quad B \rightarrow BB. \quad (\text{D2})$$

Furthermore, we assume that these two processes occur with equal probability. Note that this model can be regarded as a generalization of the Pólya urn model since Eq. (B1) reduces to it in the limit of $t \rightarrow 0$ and $q \rightarrow 1$.

a. Magnetization, susceptibilities, and Binder parameter

In Fig. D22, the magnetization, Eq. (C1), for $q = 10^{-2.0}$ and $q = 10^{-0.5}$ is shown. For $q = 10^{-2.0}$, the behavior of the magnetization, Eq. (C1), is similar to the case of $X \rightarrow YZ$ shown in Fig. D6. However, the behavior of the magnetization, Eq. (C1), is quite different for $X \rightarrow XX$ and $X \rightarrow YZ$. That is, in Fig. D22(lower), the magnetization seems to have a nontrivial value for small $k_B T$.

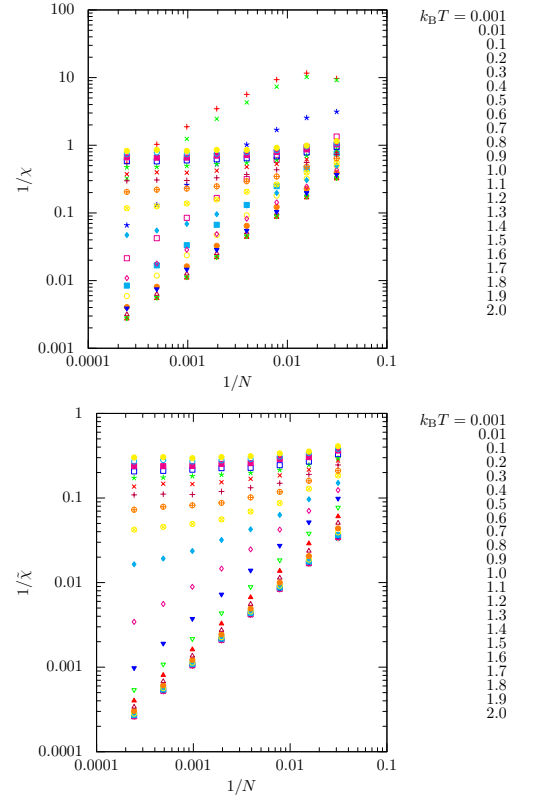


FIG. D16: System-size dependence of the susceptibilities, (upper) Eq. (C12) and (lower) Eq. (C15) at $q = 10^{-2.0}$. We consider $X \rightarrow YZ$ and set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 32, 64, 128, 256, 512, 1024, 2048, 4096$.

In Fig. D23, we show the susceptibility, Eq. (C12). As expected, Fig. D23 shows diverging behavior for finite temperatures in both cases.

In Fig. D24, we plot the Binder parameter. In Fig. D24, we can observe the crossings of the Binder parameter. This fact implies that, in the case of $X \rightarrow XX$, there exists second-order phase transitions or BKT transitions for a wide range of q .

b. Correlation functions, mutual information, and histogram of magnetization for $q = 10^{-2.0}$

In Figs. D25, D26, and D27, we show the correlation functions, Eq. (C20), mutual information, Eq. (C22), and the histogram of the magnetization, Eq. (C23), respectively. These figures are consistent with either second-order phase transitions or BKT transitions.

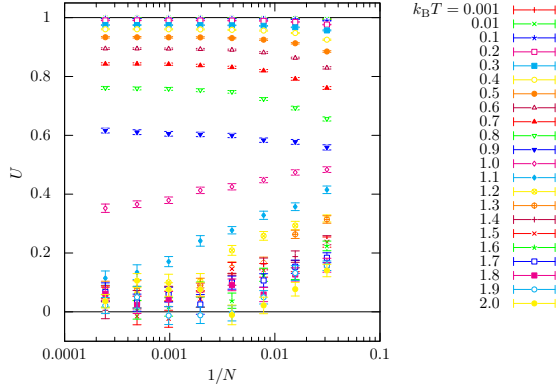


FIG. D17: System-size dependence of the Binder parameter, Eq. (C17), at $q = 10^{-2.0}$. We consider $X \rightarrow YZ$ and set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 32, 64, 128, 256, 512, 1024, 2048, 4096$.

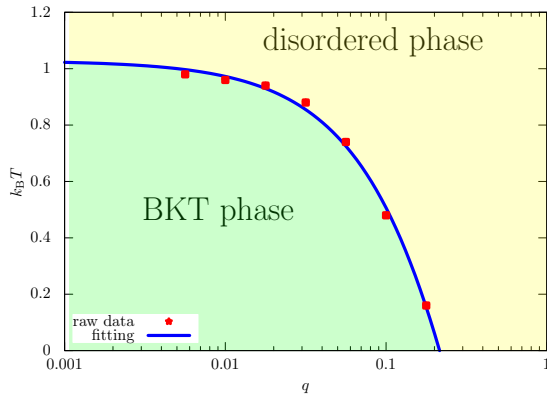


FIG. D18: Phase diagram of the context-sensitive random language model where the horizontal and vertical axes are the growth rate of a sentence q and temperature $k_B T$, respectively. We consider $X \rightarrow YZ$ and set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. This figure is identical to Fig. 5(upper).

c. Correlation functions, mutual information, and histogram of magnetization for $q = 10^{-0.5}$

In Figs. D28, D29, and D30, we show the correlation functions, Eq. (C20), mutual information, Eq. (C22), and the histogram of the magnetization, Eq. (C23), respectively. Similar to the case of $q = 10^{-2.0}$, these figures are also consistent with either second-order phase transitions or BKT transitions.

d. System-size dependence of the magnetization, the susceptibilities, and the Binder parameter

In Figs. D31, D32, and D33, we show the system-size dependence of the magnetization, Eq. (C1), the two susceptibilities, Eqs. (C12) and (C15), and the Binder pa-

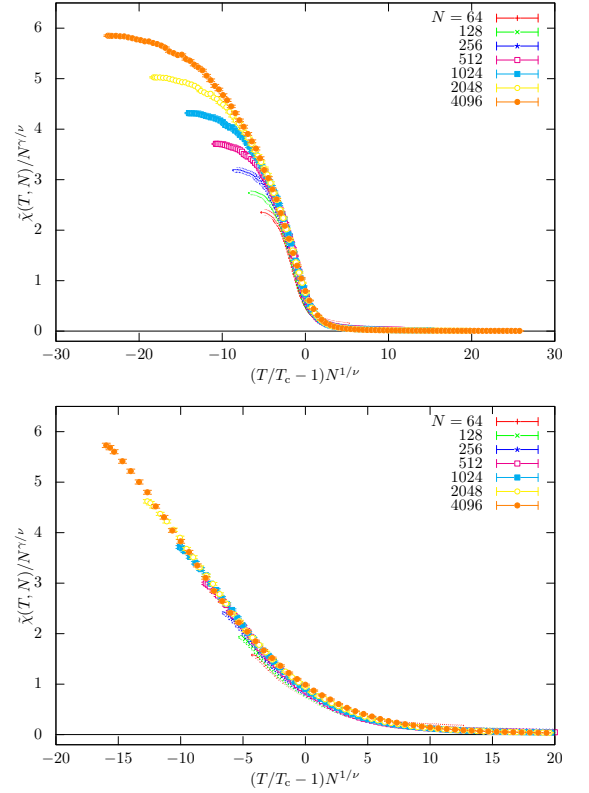


FIG. D19: Finite-size scaling of $\tilde{\chi}$ at (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-1.0}$. We set $T_c = 0.960$, $\nu = 2.6250$, and $\gamma = 2.05$ for $q = 10^{-2.0}$ and $T_c = 0.48$, $\nu = 3.0$, and $\gamma = 2.05$ for $q = 10^{-1.0}$, where the values of ν and γ are determined such that the scaling assumption holds. We consider $X \rightarrow YZ$ and set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 64, 128, 256, 512, 1024, 2048, 4096$. The upper panel is identical to Fig. 5(middle).

rameter, Eq. (C17), at $q = 10^{-2.0}$, respectively. The critical temperature was again determined by assessing whether the line is linear or not. This estimate is consistent with the temperature at which the Binder parameter takes a non-zero value in the thermodynamic limit, as shown in Fig. D33.

e. Phase diagram, finite-size scaling, and q -dependence of critical exponents

In Fig. D34, we plot the phase diagram of our language model with $K = 2$ and $X \rightarrow XX$. We determined the boundary between the two phases by fitting a quadratic function via least squares regression. In the case of $X \rightarrow XX$, the regime of the BKT phase is drastically expanded, especially for large q compared with the case of $X \rightarrow YZ$.

In Fig. D35, we perform finite-size scaling on $\tilde{\chi}$ at $q = 10^{-2.0}$. We set $T_c = 1.0$, $\nu = 2.6250$, and $\gamma = 2.05$, where we determine the values of ν and γ such that the scaling assumption holds. This figure shows that the scal-

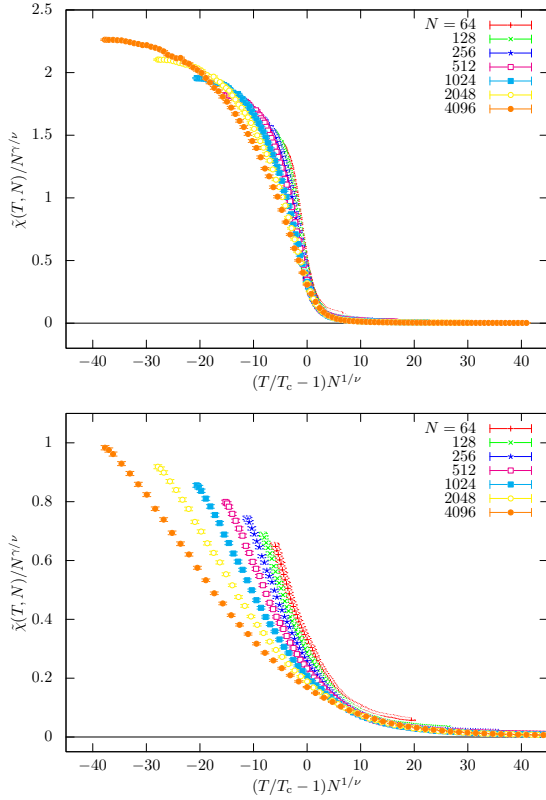


FIG. D20: Finite-size scaling of $\tilde{\chi}$ at (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-1.0}$. We set $T_c = 0.960$ for $q = 10^{-2.0}$ and $T_c = 0.48$ for $q = 10^{-1.0}$. We also set $\nu = 2.29$, and $\gamma = 2.05$, where the values of ν and γ are taken from Ref. [37]. We consider $X \rightarrow YZ$ and set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 64, 128, 256, 512, 1024, 2048, 4096$.

ing assumption holds. In Fig. D36, we set $T_c = 1.0$, $\nu = 2.29$, and $\gamma = 2.05$, where ν and γ are taken from Ref. [37]. Similar to the case of $X \rightarrow YZ$, the critical exponents depend on the value of q .

In Fig. D37, we plot the q -dependence of the critical exponents ν and γ . As already discussed above, we conclude that the BKT transition occurs since critical phenomena are observed over a wide range of the parameter space.

3. Case of $K = 10$ and $X \rightarrow YZ$

Here we consider the case of $K = 10$ and $X \rightarrow YZ$, where X is replaced by two distinct symbols Y and Z .

a. Magnetization, susceptibilities, and Binder parameter

In Fig. D38, we plot the magnetization, Eq. (C1). Figure D38 shows that the magnetization grows for small $k_B T$ in the case of $q = 10^{-2.0}$, but it remains in a trivial state for both $q = 10^{-2.0}$ and $q = 10^{-0.5}$.

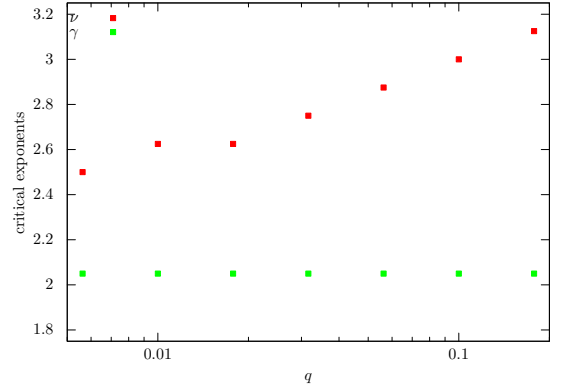


FIG. D21: q -dependence of the critical exponents ν and γ . We consider $X \rightarrow YZ$ and set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. This figure is identical to Fig. 5(lower).

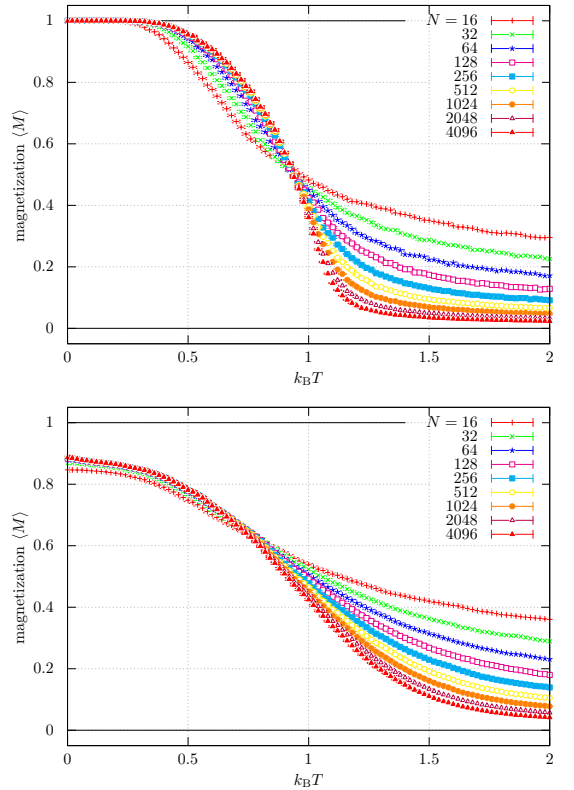


FIG. D22: Temperature dependence of the magnetization, Eq. (C1), for $X \rightarrow XX$. We set (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-0.5}$. We also set $K = 2$, $J = 1.0$, $q = 10^{-0.5}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The length of the generated sentence by the language model, N , was varied from 16 to 4096.

In Fig. D39, we show the susceptibility, Eq. (C12). This figure is consistent with Figure D38 because there is no peak for finite temperature while the susceptibility grows or even appears to diverge for $q = 10^{-2.0}$.

In Fig. D40, we show the Binder parameter. As ex-

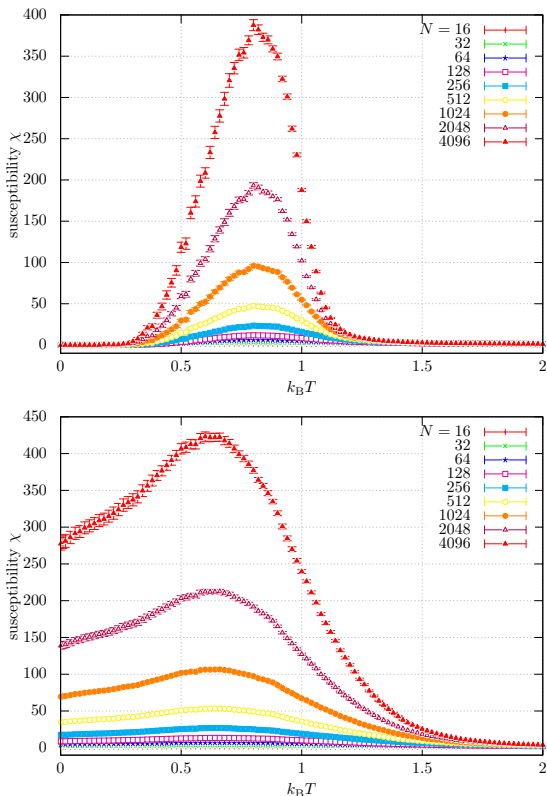


FIG. D23: Temperature dependence of the susceptibility, Eq. (C12), for $X \rightarrow XX$. We set (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-0.5}$. We also set $K = 2$, $J = 1.0$, $q = 10^{-0.5}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The length of the generated sentence by the language model, N , was varied from 16 to 4096.

pected, the Binder parameter does not go to 1 for small $k_B T$, but we can see the dip for $q = 10^{-2.0}$. This fact suggests that the phase transition occurring here is akin to the first-order phase transition at $T \rightarrow 0$, rather than the second-order phase transition.

b. Correlation functions, mutual information, and histogram of magnetization for $q = 10^{-2.0}$

In Fig. D41, we show the correlation functions, Eq. (C20). The behavior of the correlation function with $i = 2048$ and $j = 2048 + \Delta i$ shown in Fig. D41(upper) looks similar to that of the first-order phase transition, since it changes from concave functions to convex functions as the temperature is lowered without becoming a straight line in the log-log plot. However, the correlation function with $i = 0$ and $j = \lfloor N/4 \rfloor - 1$ shown in Fig. D41(lower) shows straight lines in the log-log plot for a wide range of $k_B T$.

In Fig. D42, the histogram of the magnetization, Eq. (C23), is shown. The novel point of Fig. D42 is that the histogram has two peaks toward $k_B T = 0$, and this also suggests that there is a phase transition similar to a

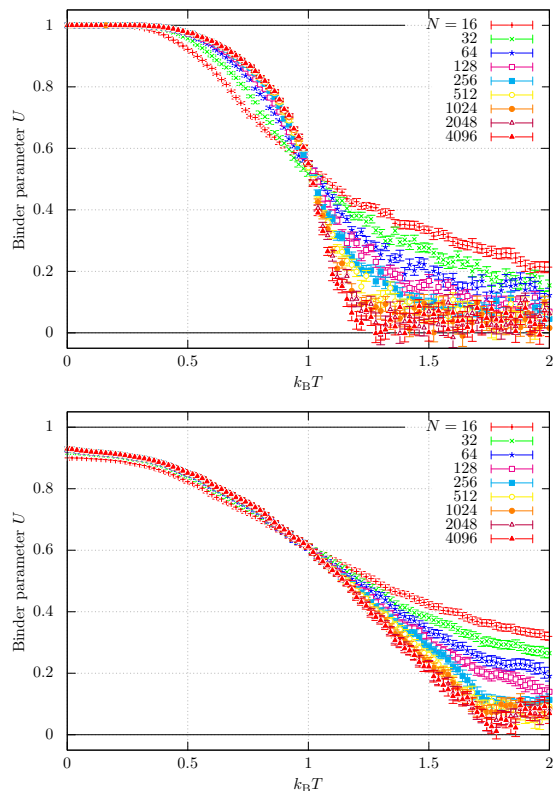


FIG. D24: Temperature dependence of the Binder parameter, Eq. (C17), for $X \rightarrow XX$. We set (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-0.5}$. We also set $K = 2$, $J = 1.0$, $q = 10^{-0.5}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The length of the generated sentence by the language model, N , was varied from 16 to 4096.

first-order phase transition at $k_B T = 0$.

c. Correlation functions, mutual information, and histogram of magnetization for $q = 10^{-0.5}$

In Fig. D43, the correlation function, Eq. (C20), is shown.

In Fig. D44, we show the histogram of the magnetization, Eq. (C23). The above figures consistently imply that the state is always trivial without any singularities.

d. System-size dependence of the magnetization, the susceptibilities, and the Binder parameter

In Figs. D45, D46, and D47, we show the system-size dependence of the magnetization, Eq. (C1), the two susceptibilities, Eqs. (C12) and (C15), and the Binder parameter, Eq. (C17), at $q = 10^{-2.0}$, respectively. Similarly, the critical temperature was determined by assessing whether the line is linear or not. This estimate is consistent with the temperature at which the Binder parameter takes a non-zero value in the thermodynamic

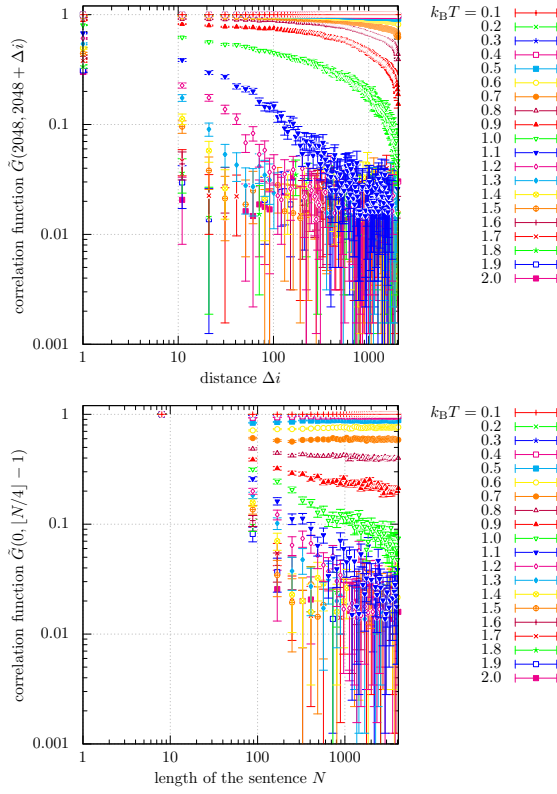


FIG. D25: Correlation functions, Eq. (C20), (upper) with $i = 2048$ and $j = 2048 + \Delta i$ and (lower) with $i = 0$ and $j = \lfloor N/4 \rfloor - 1$ for $X \rightarrow XX$. We set $K = 2$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

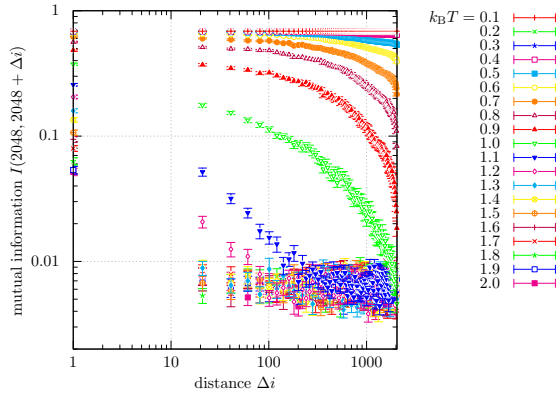


FIG. D26: Mutual information, Eq. (C22), with $i = 2048$ and $j = 2048 + \Delta i$ for $X \rightarrow XX$. We set $K = 2$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

limit, as shown in Fig. D47. Figure D47 suggests that phase transitions take place twice, but we do not discuss the details in this paper.

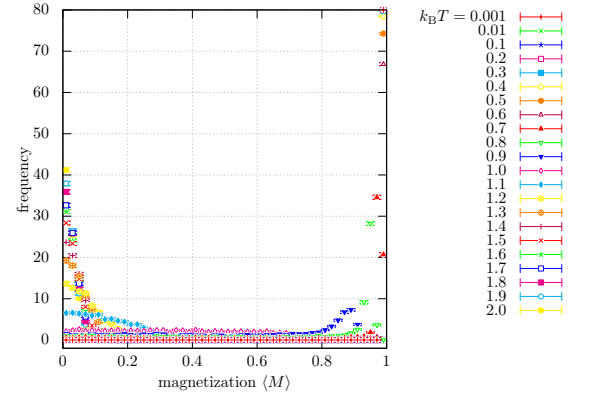


FIG. D27: Histogram of the magnetization, Eq. (C1), for $X \rightarrow XX$. We set $K = 2$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

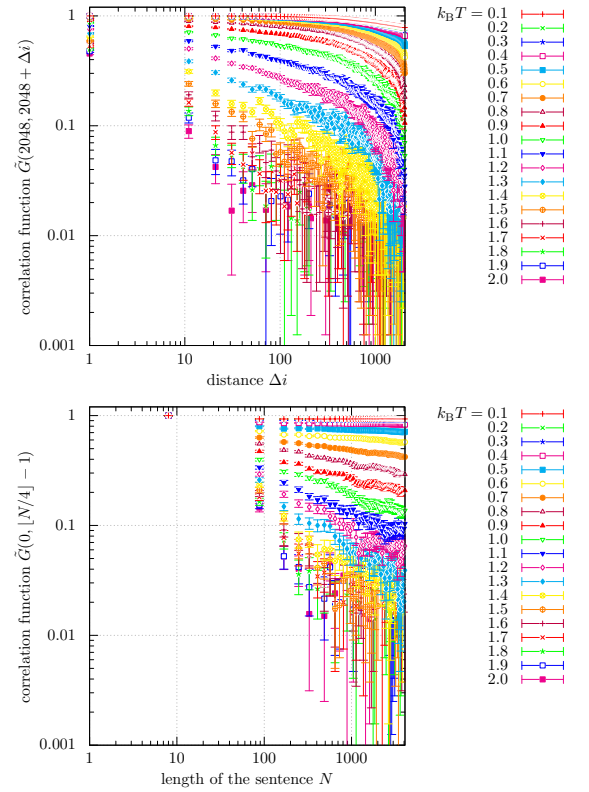


FIG. D28: Correlation functions, Eq. (C20), (upper) with $i = 2048$ and $j = 2048 + \Delta i$ and (lower) with $i = 0$ and $j = \lfloor N/4 \rfloor - 1$ for $X \rightarrow XX$. We set $K = 2$, $J = 1.0$, $q = 10^{-0.5}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

e. Phase diagram, finite-size scaling, and q -dependence of critical exponents

In Fig. D48, we plot the phase diagram of our language model with $K = 10$ and $X \rightarrow YZ$. We determined the

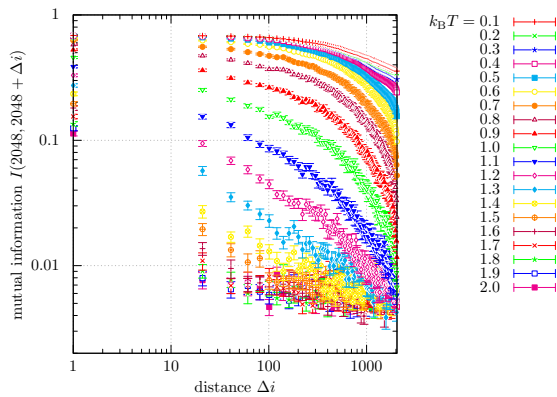


FIG. D29: Mutual information, Eq. (C22), with $i = 2048$ and $j = 2048 + \Delta i$ for $X \rightarrow XX$. We set $K = 2$, $J = 1.0$, $q = 10^{-0.5}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

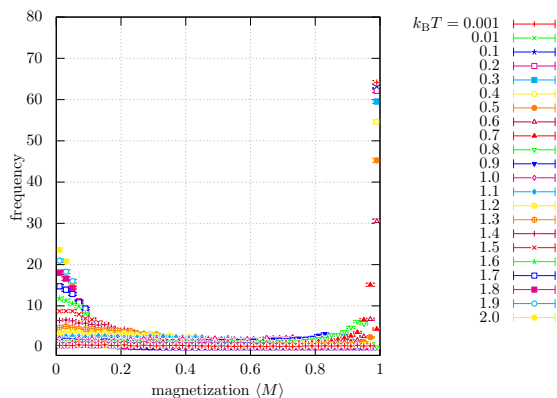


FIG. D30: Histogram of the magnetization, Eq. (C1), for $X \rightarrow XX$. We set $K = 2$, $J = 1.0$, $q = 10^{-0.5}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

boundary between the two phases by fitting a quadratic function via least squares regression. Compared to the case of $K = 2$, the regime of the BKT phase is reduced.

In Fig. D49, we carry out finite-size scaling for $\tilde{\chi}$ at $q = 10^{-2.0}$. We set $T_c = 0.40$, $\nu = 3.0$, and $\gamma = 2.05$, where ν and γ are determined such that the scaling assumption holds.

In Fig. D50, we plot the q -dependence of the critical exponents ν and γ . We have observed the robust critical behavior, concluding that the BKT transition occurs.

f. Configurations and relative frequencies for $q = 10^{-2.0}$

As shown in Fig. D42, the histogram of magnetization has peaks at $\langle M \rangle = 0.550, 0.790$. Then, in Fig. D51, we show typical configurations whose magnetization values are 0.550 and 0.790. In both cases, symbols are scattered, but their distributions have different features. Around $\langle M \rangle = 0.790$, one character dominates, whereas near

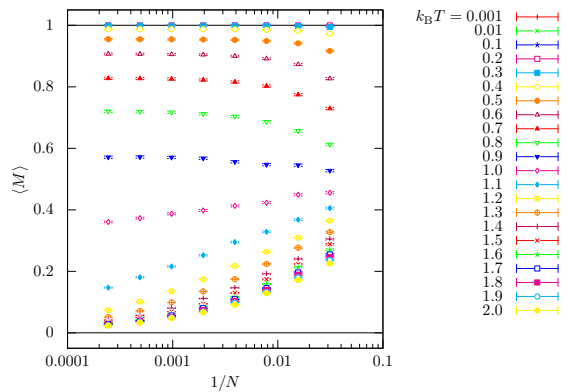


FIG. D31: System-size dependence of the magnetization, Eq. (C1), at $q = 10^{-2.0}$. We consider $X \rightarrow XX$ and set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 32, 64, 128, 256, 512, 1024, 2048, 4096$.

$\langle M \rangle = 0.550$, two major characters are comparatively dominant.

In Fig. D52, we plot the relative frequencies of letters in the alphabet. Symbols on the horizontal axis are sorted by their frequencies. In natural languages, relative frequencies of letters in the alphabet often show Zipf's law; that is, they form a straight line in the log-log plot. The relative frequencies at $k_B T = 0.3000$ seem to be a straight line in Fig. D52(upper). However, the histogram of letters in the alphabet is almost uniformly random at $k_B T = 0.1000$, as shown in Fig. D52(lower). In Fig. D53, we also plot the relative frequencies of letters in the alphabet at $k_B T = 0.0010$. For characters whose relative frequencies are small, we can see Zipf's law.

4. Case of $K = 10$ and $X \rightarrow XX$

Lastly, we consider the case of $K = 10$ and $X \rightarrow XX$.

a. Magnetization, susceptibilities, and Binder parameter

In Fig. D54, we plot the magnetization, Eq. (C1). In both cases, the magnetization, Eq. (C1), approaches a nontrivial value when $k_B T$ is decreased. The magnetization seems to jump near $k_B T \sim 0.7$ for $q = 10^{-2.0}$, whereas it gradually increases for $q = 10^{-0.5}$.

In Fig. D55, the susceptibility, Eq. (C12) is plotted. In Fig. D55, the susceptibility has a peak at a specific temperature. This fact also supports the existence of a phase transition.

In Fig. D56, the Binder parameter, Eq. (C17), is shown. For $q = 10^{-2.0}$, the Binder parameter exhibits a sharp dip before reaching 1 as $k_B T$ decreases. This behavior is characteristic of a first-order phase transition. Similarly, for $q = 10^{-0.5}$, the Binder parameter displays a similar trend, albeit without the width of the dip. In

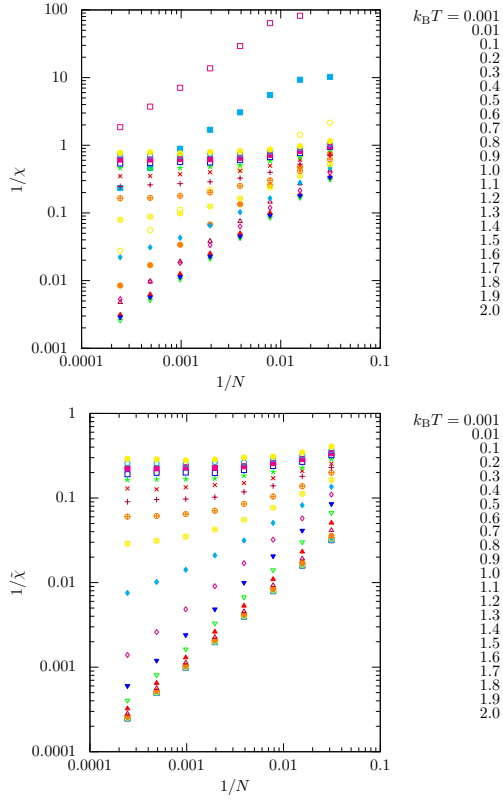


FIG. D32: System-size dependence of the susceptibilities, (upper) Eq. (C12) and (lower) Eq. (C15) at $q = 10^{-2.0}$. We consider $X \rightarrow XX$ and set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 32, 64, 128, 256, 512, 1024, 2048, 4096$.

both cases, the existence of a first-order phase transition appears evident. However, the exact cause of the wide dip in the Binder parameter is not clear.

b. Correlation functions, mutual information, and histogram of magnetization for $q = 10^{-2.0}$

In Fig. D57, the correlation function, Eq. (C20), is shown. In Fig. D57(upper), the correlation functions, Eq. (C20), transition from concave to convex functions as $k_B T$ decreases. This observation supports the existence of a first-order phase transition. However, Fig. D57(lower) displays straight lines over a wide range of $k_B T$.

In Fig. D58, we plot the histogram of the magnetization, Eq. (C23). At $k_B T \sim 0.4$, there are two peaks. This fact strongly supports the existence of a phase transition similar to that of the first-order phase transition.

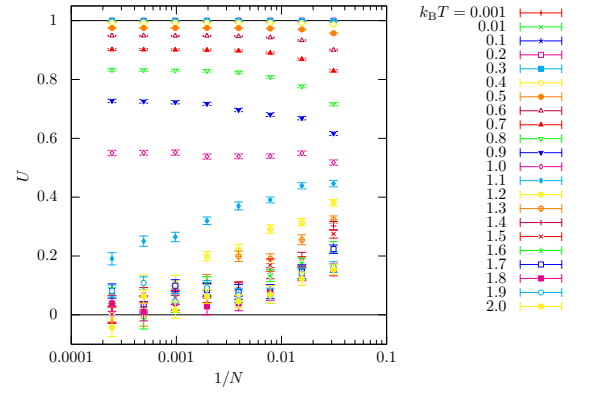


FIG. D33: System-size dependence of the Binder parameter, Eq. (C17), at $q = 10^{-2.0}$. We consider $X \rightarrow XX$ and set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 32, 64, 128, 256, 512, 1024, 2048, 4096$.

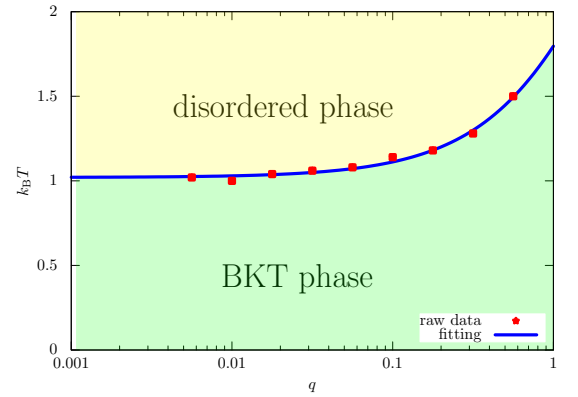


FIG. D34: Phase diagram of the context-sensitive random language model, where the horizontal and vertical axes are the growth rate of a sentence q and temperature $k_B T$, respectively. We consider $X \rightarrow XX$ and set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$.

c. Correlation functions, mutual information, and histogram of magnetization for $q = 10^{-0.5}$

In Fig. D59, the correlation functions, Eq. (C20), are shown. For $q = 10^{-0.5}$, the correlation function varies similar to that of the second-order phase transition.

In Fig. D60, the histogram of the magnetization, Eq. (C23), is shown. Figure D60 shows two peaks toward $k_B T = 0$. This fact implies that a phase transition akin to that of the first-order phase transition occurs at $k_B T = 0$.

d. System-size dependence of the magnetization, the susceptibilities, and the Binder parameter

In Figs. D61, D62, and D63, we show the system-size dependence of the magnetization, Eq. (C1), the two sus-

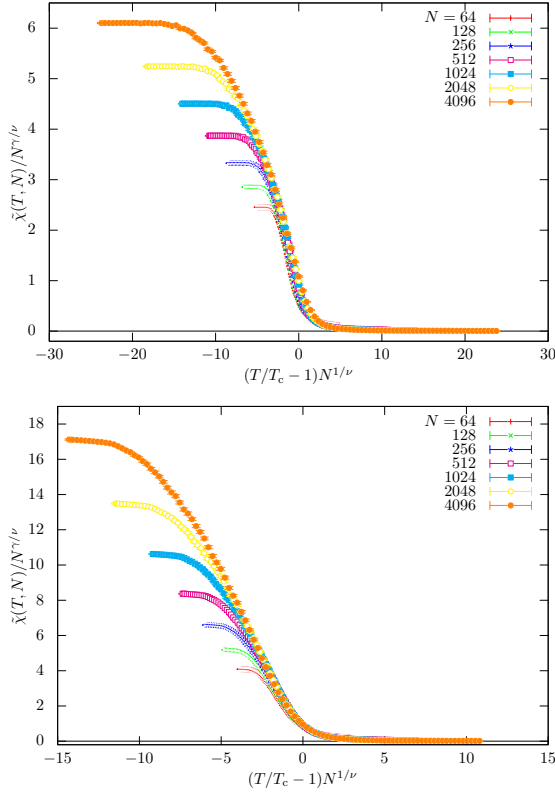


FIG. D35: Finite-size scaling of $\tilde{\chi}$ at (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-1.0}$. We set $T_c = 1.0$, $\nu = 2.6250$, and $\gamma = 2.05$ for $q = 10^{-2.0}$ and $T_c = 1.14$, $\nu = 3.1250$, and $\gamma = 2.05$ for $q = 10^{-1.0}$, where the values of ν and γ are determined such that the scaling assumption holds. We consider $X \rightarrow XX$ and set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 64, 128, 256, 512, 1024, 2048, 4096$.

ceptibilities, Eqs. (C12) and (C15), and the Binder parameter, Eq. (C17), at $q = 10^{-2.0}$, respectively. The critical temperature was again determined by assessing whether the line is linear or not. This estimate is consistent with the temperature at which the Binder parameter takes a non-zero value in the thermodynamic limit, as shown in Fig. D63. Figure D63 also suggests that phase transitions occur twice, but we do not discuss the details in this paper.

e. Phase diagram, finite-size scaling, and q -dependence of critical exponents

In Fig. D64, we plot the phase diagram of our language model with $K = 10$ and $X \rightarrow XX$. We determined the boundary between the two phases by fitting a quadratic function via least squares regression.

In Fig. D65, we carry out finite-size scaling for $\tilde{\chi}$ at $q = 10^{-2.0}$. We set $T_c = 0.640$, $\nu = 2.8750$, and $\gamma = 2.05$, where ν and γ are determined so that the scaling assumption holds.

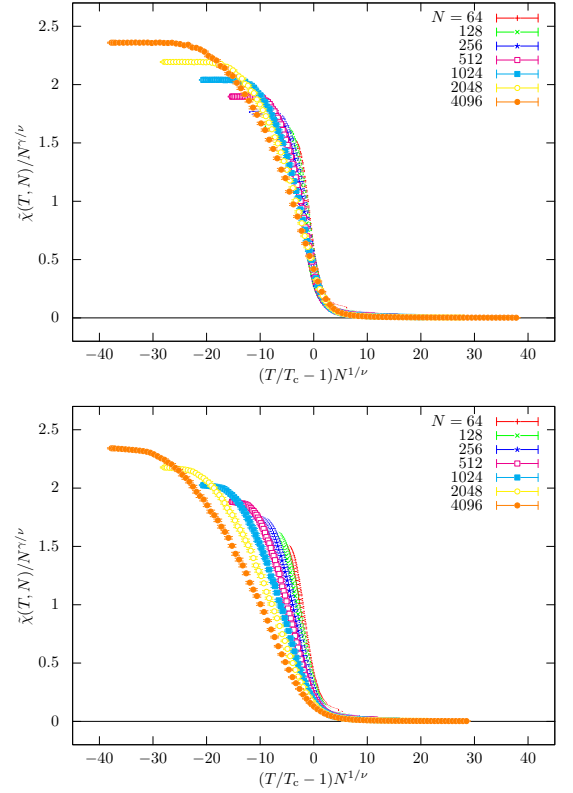


FIG. D36: Finite-size scaling of $\tilde{\chi}$ at (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-1.0}$. We set $T_c = 1.0$ for $q = 10^{-2.0}$ and $T_c = 1.14$ for $q = 10^{-1.0}$. We also set $\nu = 2.29$, and $\gamma = 2.05$, where the values of ν and γ are taken from Ref. [37]. We consider $X \rightarrow XX$ and set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 64, 128, 256, 512, 1024, 2048, 4096$.

In Fig. D66, we plot the q -dependence of the critical exponents ν and γ . We have again observed the robust critical behavior, concluding that the BKT transition occurs.

f. Configurations and relative frequencies for $q = 10^{-2.0}$

In Fig. D67, we plot typical configurations whose magnetization value is 0.670 at $k_B T = 0.440$ and configurations whose magnetization value is 0.990 at $k_B T = 0.0010$. At $k_B T = 0.440$, the configuration with $M = 0.670$ is chosen because it corresponds to the peak of the magnetization histogram at $\langle M \rangle = 0.670$. When the magnetization is around 0.670, two characters out of ten are dominant and two states coexist. As expected, the system is ordered when the magnetization is near 0.990.

In Fig. D68, relative frequencies of letters in the alphabet are shown. When the magnetization is around 0.670, two major characters are dominant, and other characters rarely appear in sentences. As a result, the relative frequencies do not fit Zipf's law. In the case where the magnetization is near 0.990, that is, the ordered state is

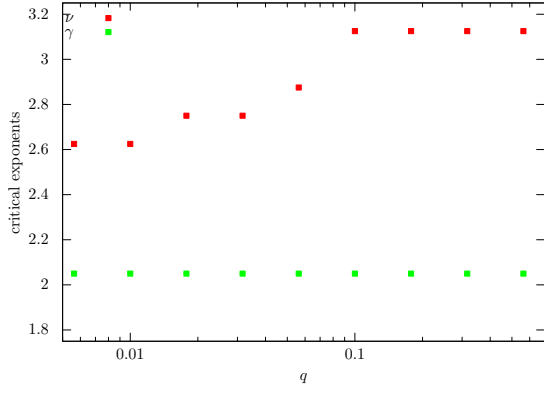


FIG. D37: q -dependence of the critical exponents ν and γ . We consider $X \rightarrow XX$ and set $K = 2$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$.

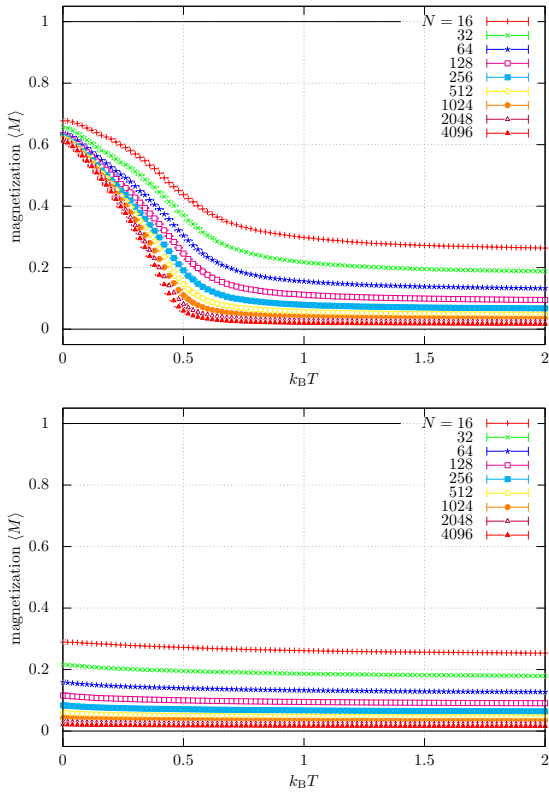


FIG. D38: Temperature dependence of the magnetization, Eq. (C1), for $X \rightarrow YZ$. We set (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-0.5}$. We also set $K = 10$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The length of the generated sentence by the language model, N , was varied from 16 to 4096.

realized, only a single character appears in a sentence.

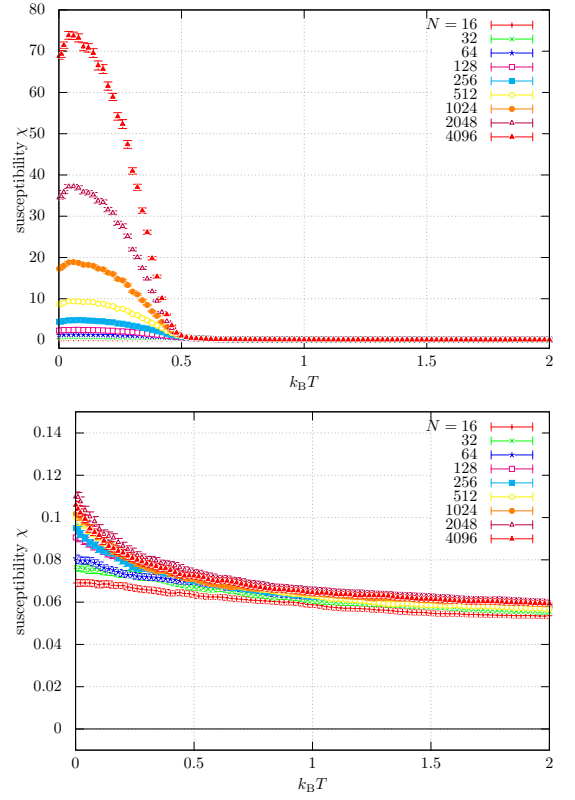


FIG. D39: Temperature dependence of the susceptibility, Eq. (C12), for $X \rightarrow YZ$. We set (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-0.5}$. We also set $K = 10$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The length of the generated sentence by the language model, N , was varied from 16 to 4096.

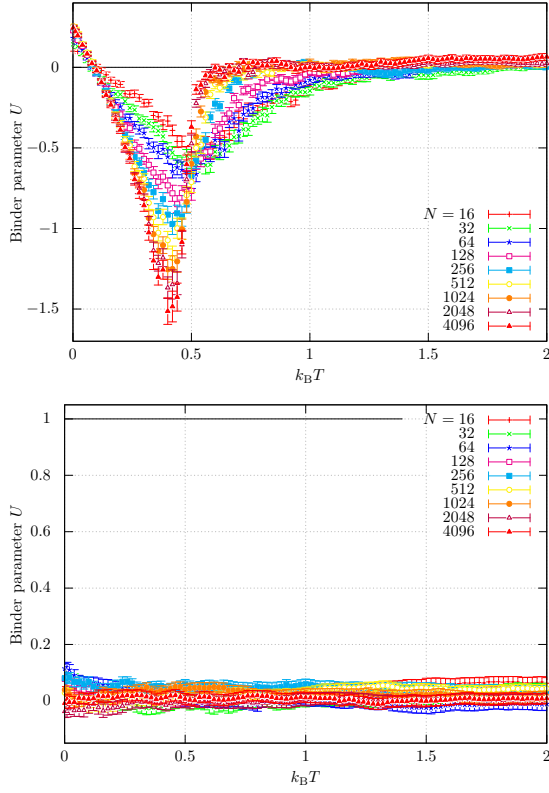


FIG. D40: Temperature dependence of the Binder parameter, Eq. (C17), for $X \rightarrow YZ$. We set (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-0.5}$. We also set $K = 10$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The length of the generated sentence by the language model, N , was varied from 16 to 4096.

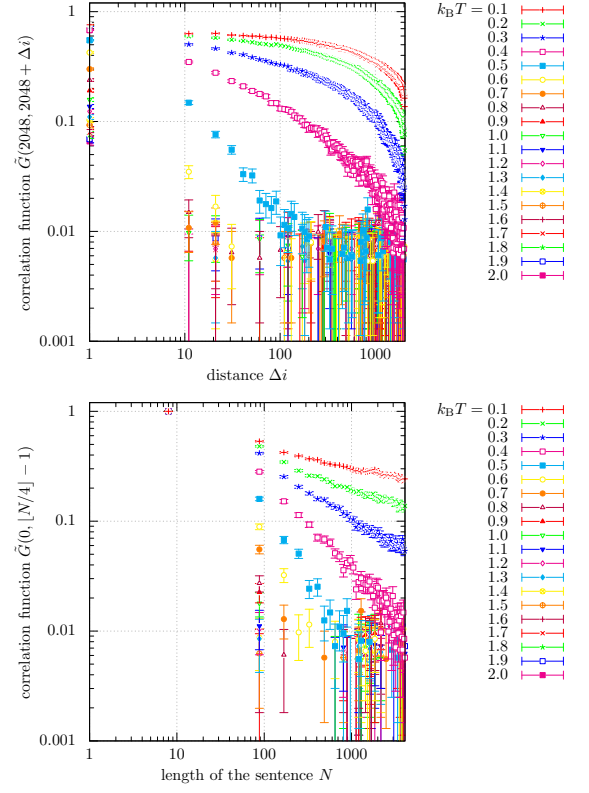


FIG. D41: Correlation functions, Eq. (C20), (upper) with $i = 2048$ and $j = 2048 + \Delta i$ and (lower) with $i = 0$ and $j = \lfloor N/4 \rfloor - 1$ for $X \rightarrow YZ$. We set $K = 10$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

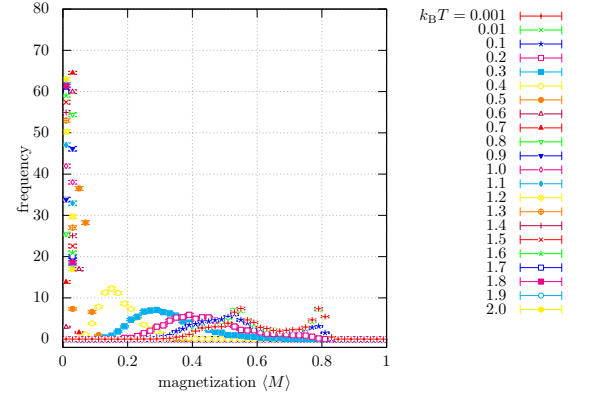


FIG. D42: Histogram of the magnetization, Eq. (C1), for $X \rightarrow YZ$. We set $K = 10$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

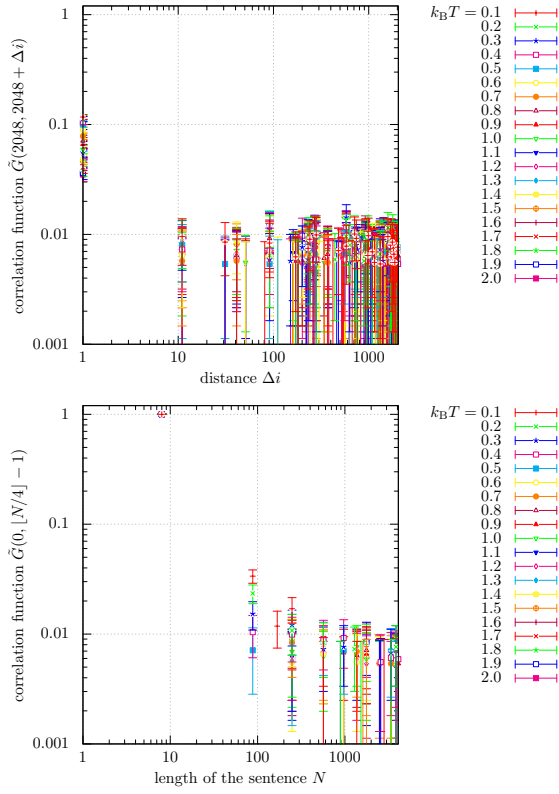


FIG. D43: Correlation functions, Eq. (C20), (upper) with $i = 2048$ and $j = 2048 + \Delta i$ and (lower) with $i = 0$ and $j = \lfloor N/4 \rfloor - 1$ for $X \rightarrow YZ$. We set $K = 10$, $J = 1.0$, $q = 10^{-0.5}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

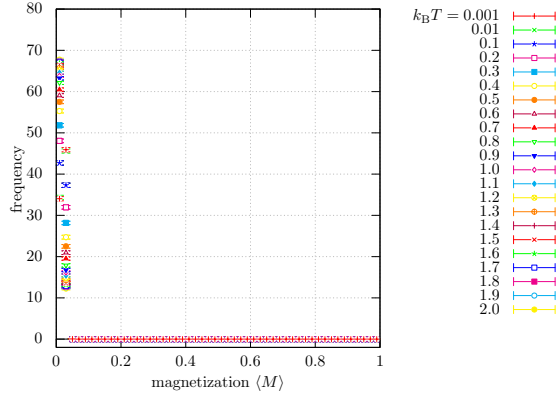


FIG. D44: Histogram of the magnetization, Eq. (C1), for $X \rightarrow YZ$. We set $K = 10$, $J = 1.0$, $q = 10^{-0.5}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

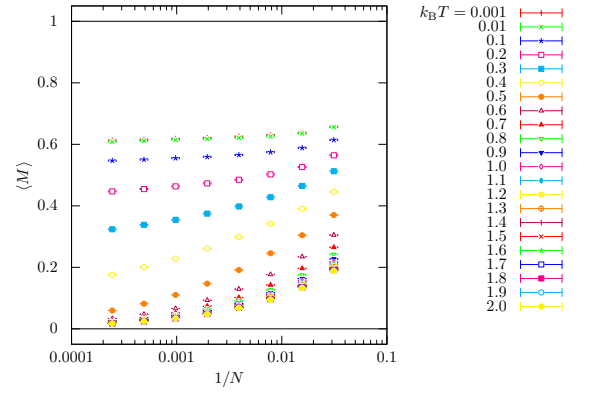


FIG. D45: System-size dependence of the magnetization, Eq. (C1), at $q = 10^{-2.0}$. We consider $X \rightarrow YZ$ and set $K = 10$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 32, 64, 128, 256, 512, 1024, 2048, 4096$.

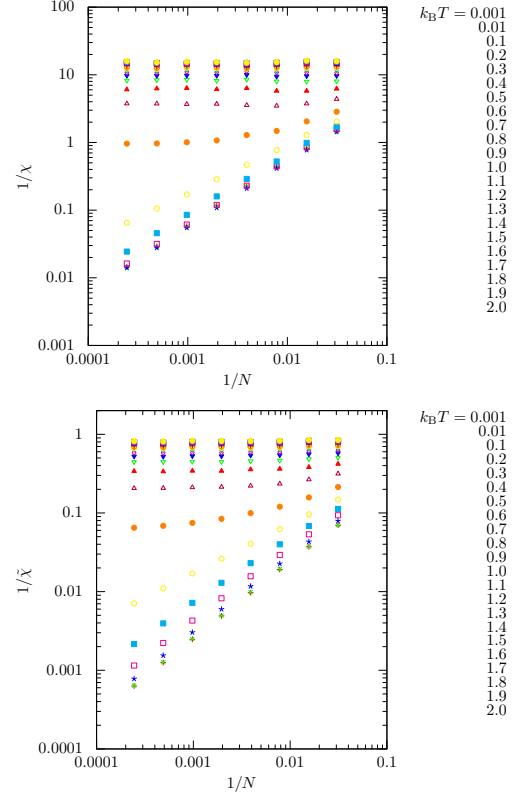


FIG. D46: System-size dependence of the susceptibilities, (upper) Eq. (C12) and (lower) Eq. (C15) at $q = 10^{-2.0}$. We consider $X \rightarrow YZ$ and set $K = 10$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 32, 64, 128, 256, 512, 1024, 2048, 4096$.

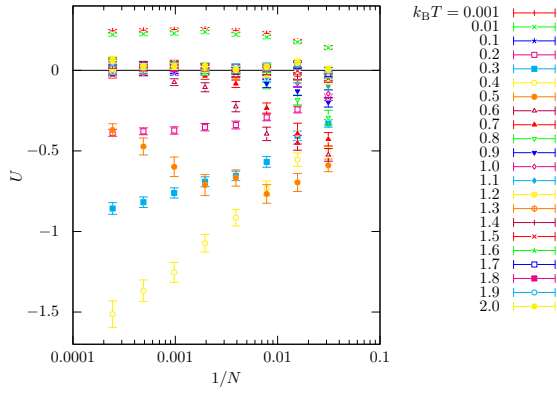


FIG. D47: System-size dependence of the Binder parameter, Eq. (C17), at $q = 10^{-2.0}$. We consider $X \rightarrow YZ$ and set $K = 10$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 32, 64, 128, 256, 512, 1024, 2048, 4096$.

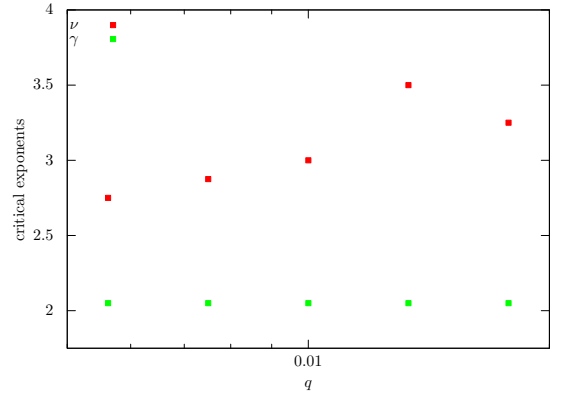


FIG. D50: q -dependence of the critical exponents ν and γ . We consider $X \rightarrow YZ$ and set $K = 10$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$.

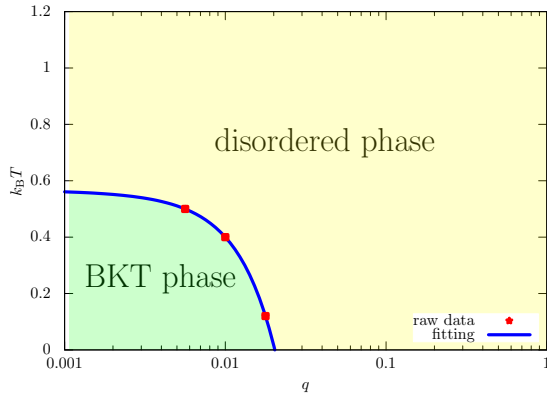


FIG. D48: Phase diagram of the context-sensitive random language model, where the horizontal and vertical axes are the growth rate of a sentence q and temperature $k_B T$, respectively. We consider $X \rightarrow YZ$ and set $K = 10$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$.

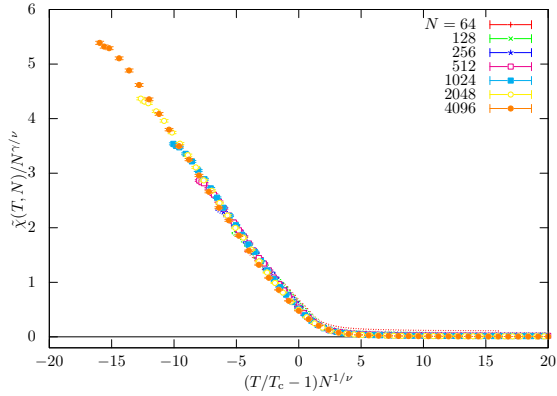


FIG. D49: Finite-size scaling of $\tilde{\chi}$ at $q = 10^{-2.0}$. We set $T_c = 0.40$, $\nu = 3.0$, and $\gamma = 2.05$, where the values of ν and γ are determined such that the scaling assumption is satisfied. We consider $X \rightarrow YZ$ and set $K = 10$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 64, 128, 256, 512, 1024, 2048, 4096$.

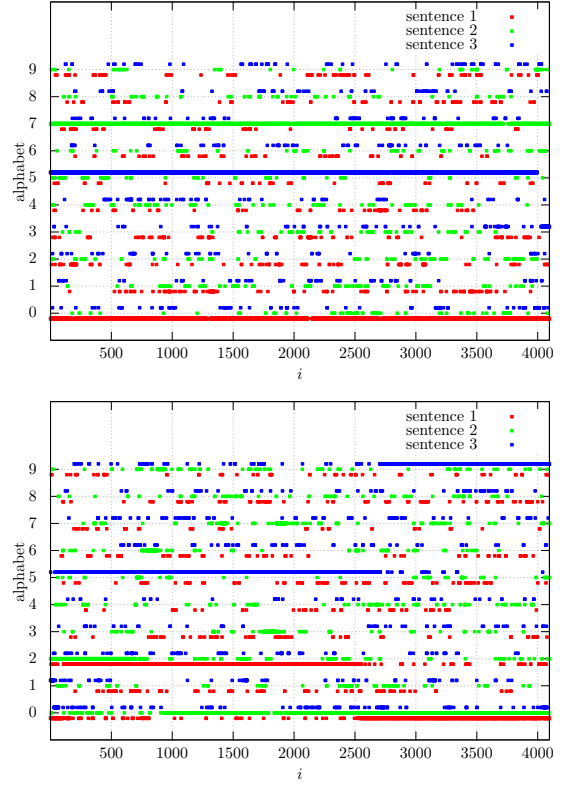


FIG. D51: Typical configurations of symbols for $X \rightarrow YZ$. (upper) We considered states whose magnetization lies between 0.5400 and 0.5600 at $k_B T = 0.0010$, and (lower) states whose magnetization lies between 0.7800 and 0.8000 at $k_B T = 0.0010$. We set $K = 10$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$.

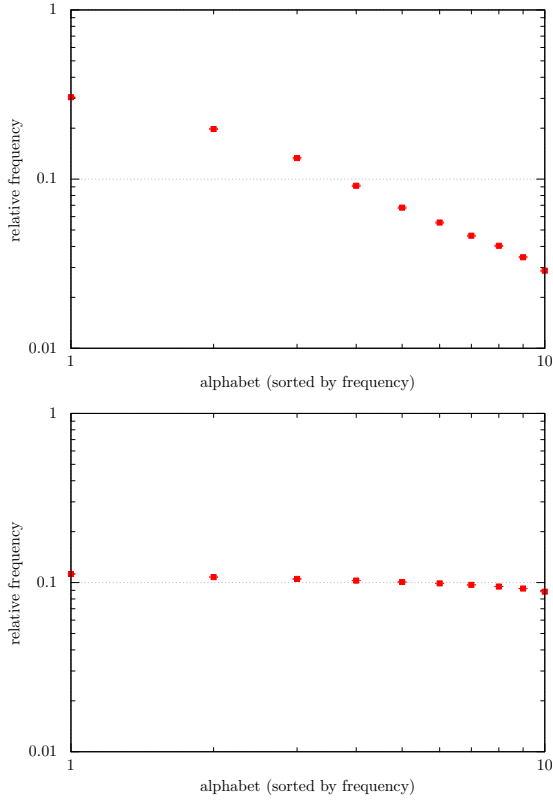


FIG. D52: Relative frequencies of letters in the alphabet for $X \rightarrow YZ$. (upper) We considered states whose magnetization lies between 0.2800 and 0.3000 at $k_B T = 0.3000$, and (lower) states whose magnetization lies between 0.0200 and 0.0400 at $k_B T = 0.1000$. We set $K = 10$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. Symbols on the horizontal axis are sorted by frequency and incremented by unity. The scales of the vertical and horizontal axes are logarithmic.

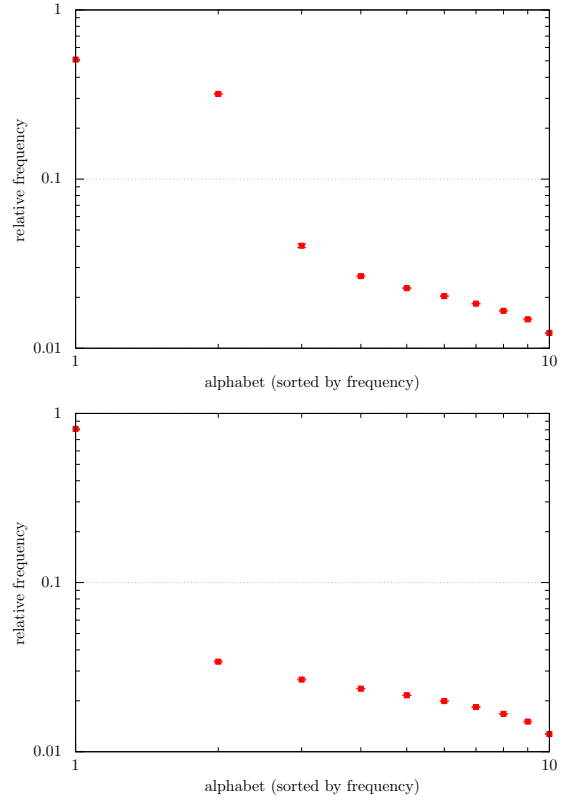


FIG. D53: Relative frequencies of letters in the alphabet for $X \rightarrow YZ$. (upper) We considered states whose magnetization lies between 0.5400 and 0.5600 at $k_B T = 0.0010$, and (lower) states whose magnetization lies between 0.7800 and 0.8000 at $k_B T = 0.0010$. We set $K = 10$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. Symbols on the horizontal axis are sorted by frequency and incremented by unity. The scales of the vertical and horizontal axes are logarithmic.

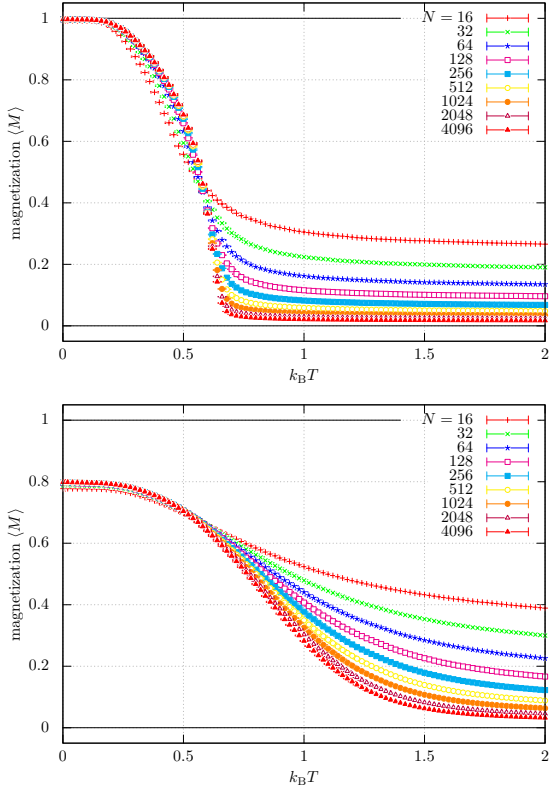


FIG. D54: Temperature dependence of the magnetization, Eq. (C1), for $X \rightarrow XX$. We set (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-0.5}$. We also set $K = 10$, $J = 1.0$, $q = 10^{-0.5}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The length of the generated sentence by the language model, N , was varied from 16 to 4096.

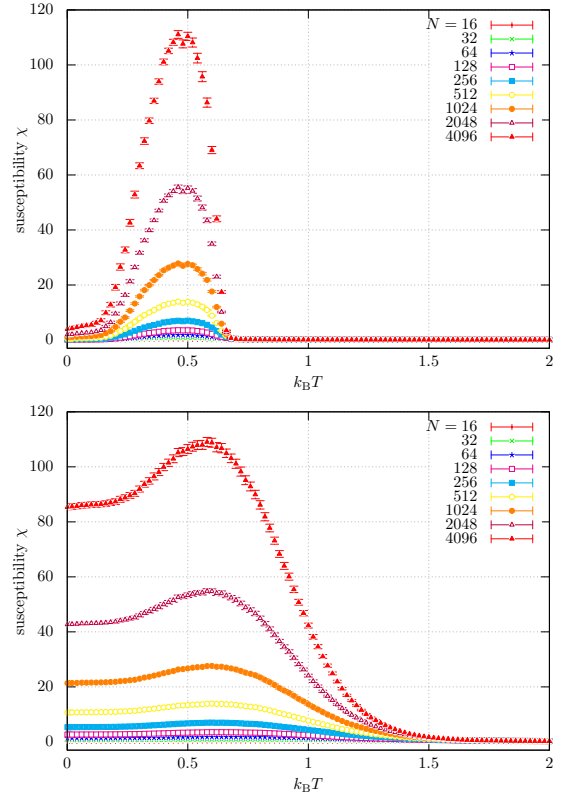


FIG. D55: Temperature dependence of the susceptibility, Eq. (C12), for $X \rightarrow XX$. We set (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-0.5}$. We also set $K = 10$, $J = 1.0$, $q = 10^{-0.5}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The length of the generated sentence by the language model, N , was varied from 16 to 4096.

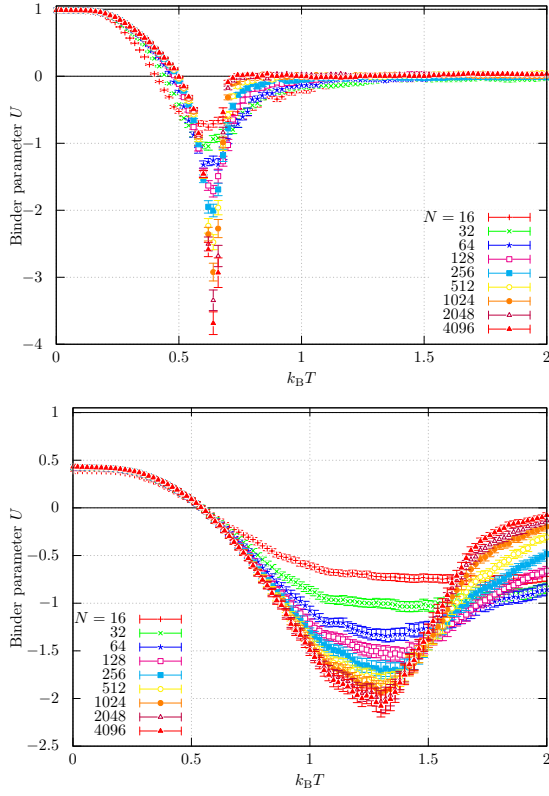


FIG. D56: Temperature dependence of the Binder parameter, Eq. (C17), for $X \rightarrow XX$. We set (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-0.5}$. We also set $K = 10$, $J = 1.0$, $q = 10^{-0.5}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The length of the generated sentence by the language model, N , was varied from 16 to 4096.

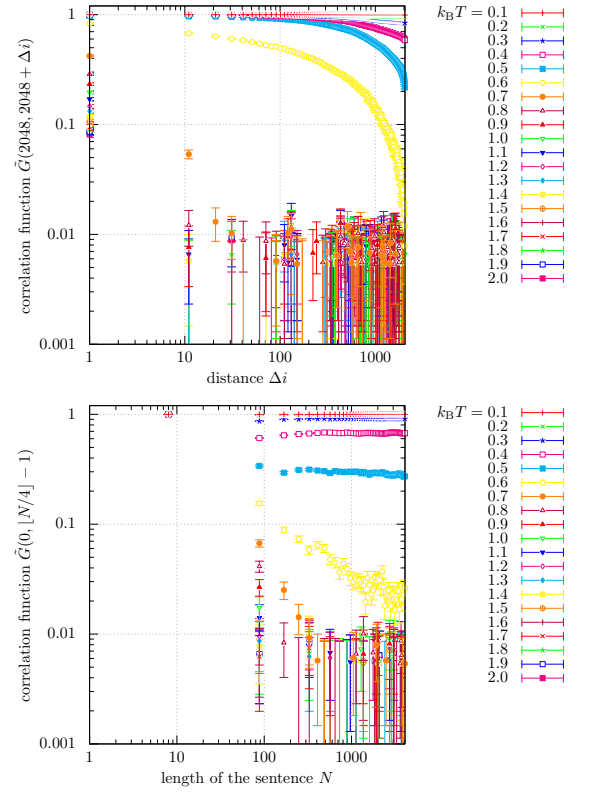


FIG. D57: Correlation functions, Eq. (C20), (upper) with $i = 2048$ and $j = 2048 + \Delta i$ and (lower) with $i = 0$ and $j = \lfloor N/4 \rfloor - 1$ for $X \rightarrow XX$. We set $K = 10$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

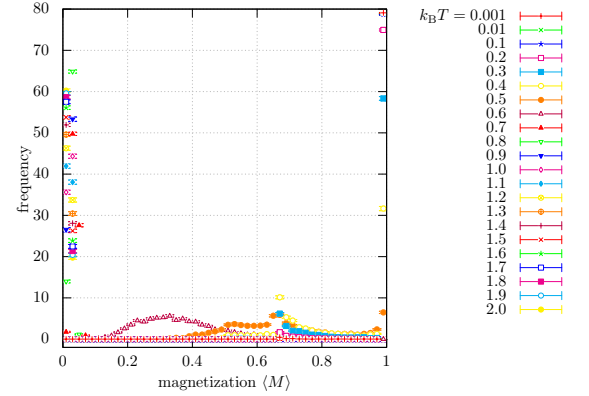


FIG. D58: Histogram of the magnetization, Eq. (C1), for $X \rightarrow XX$. We set $K = 10$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

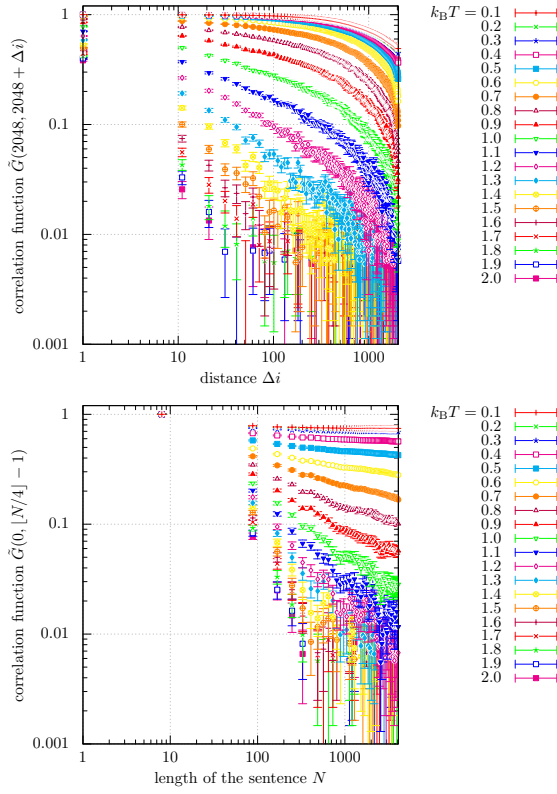


FIG. D59: Correlation functions, Eq. (C20), (upper) with $i = 2048$ and $j = 2048 + \Delta i$ and (lower) with $i = 0$ and $j = \lfloor N/4 \rfloor - 1$ for $X \rightarrow XX$. We set $K = 10$, $J = 1.0$, $q = 10^{-0.5}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

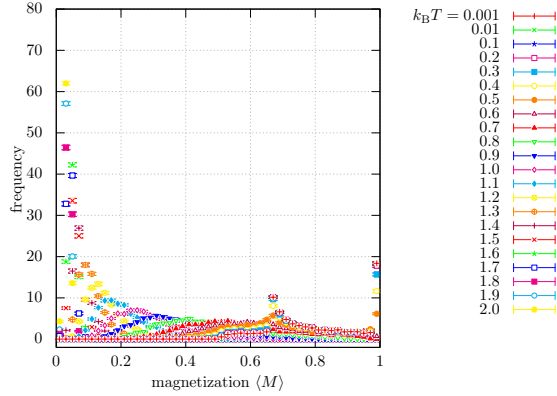


FIG. D60: Histogram of the magnetization, Eq. (C1), for $X \rightarrow XX$. We set $K = 10$, $J = 1.0$, $q = 10^{-0.5}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. The temperature $k_B T$ was varied from 0.1 to 2.0.

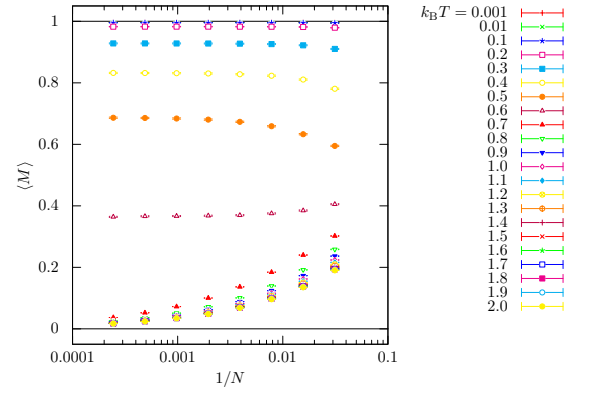


FIG. D61: System-size dependence of the magnetization, Eq. (C1), at $q = 10^{-2.0}$. We consider $X \rightarrow XX$ and set $K = 10$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 32, 64, 128, 256, 512, 1024, 2048, 4096$.

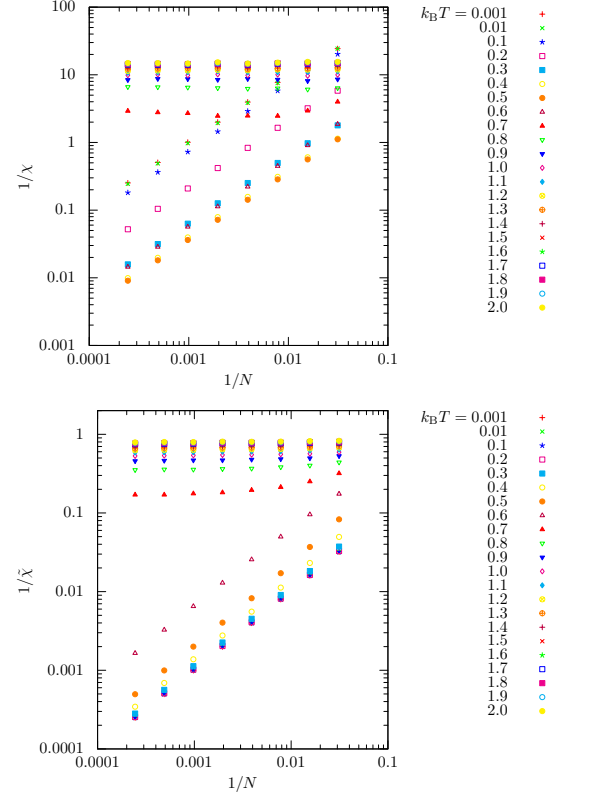


FIG. D62: System-size dependence of the susceptibilities, (upper) Eq. (C12) and (lower) Eq. (C15) at $q = 10^{-2.0}$. We consider $X \rightarrow XX$ and set $K = 10$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 32, 64, 128, 256, 512, 1024, 2048, 4096$.

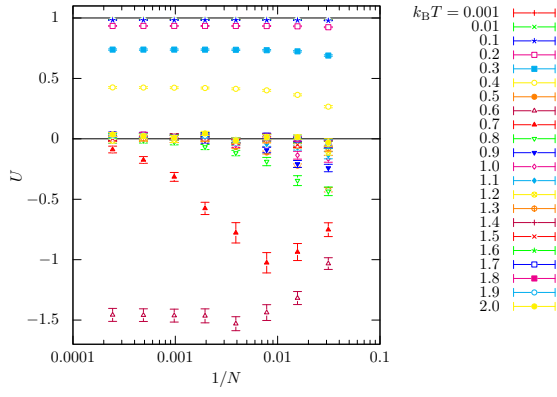


FIG. D63: System-size dependence of the Binder parameter, Eq. (C17), at $q = 10^{-2.0}$. We consider $X \rightarrow XX$ and set $K = 10$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 32, 64, 128, 256, 512, 1024, 2048, 4096$.

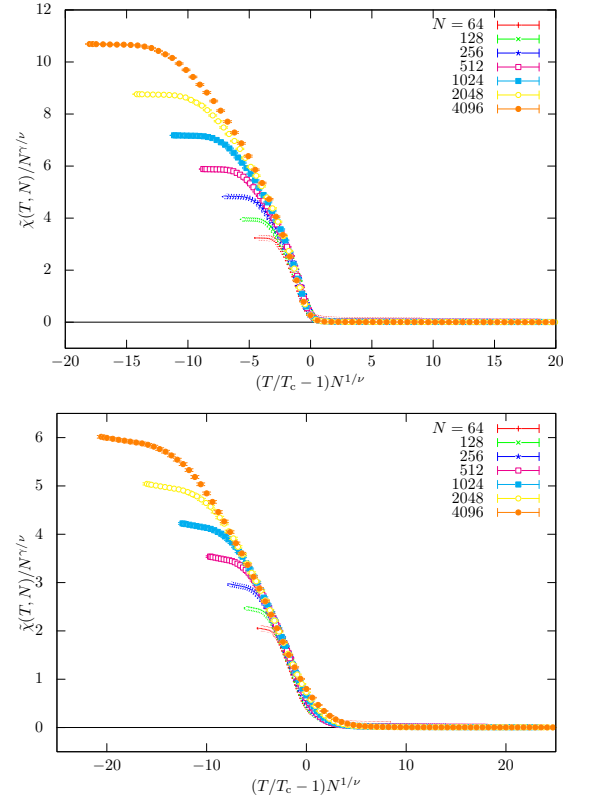


FIG. D65: Finite-size scaling of $\tilde{\chi}$ at (upper) $q = 10^{-2.0}$ and (lower) $q = 10^{-1.0}$. We set $T_c = 0.640$, $\nu = 2.8750$, and $\gamma = 2.05$ for $q = 10^{-2.0}$ and $T_c = 0.70$, $\nu = 2.75$, and $\gamma = 2.05$ for $q = 10^{-1.0}$, where the values of ν and γ are determined such that the scaling assumption is satisfied. We consider $X \rightarrow XX$ and set $K = 10$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. We varied $N = 64, 128, 256, 512, 1024, 2048, 4096$.

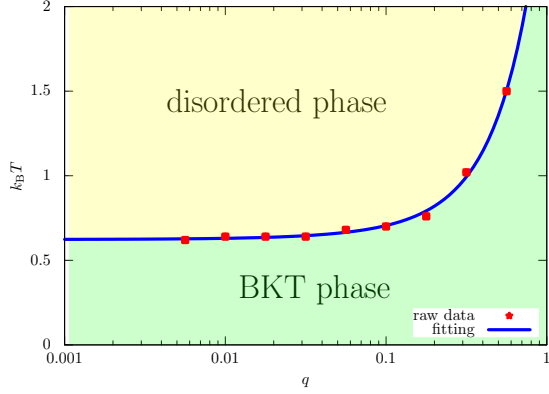


FIG. D64: Phase diagram of the context-sensitive random language model, where the horizontal and vertical axes are the growth rate of a sentence q and temperature $k_B T$, respectively. We consider $X \rightarrow XX$ and set $K = 10$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$.

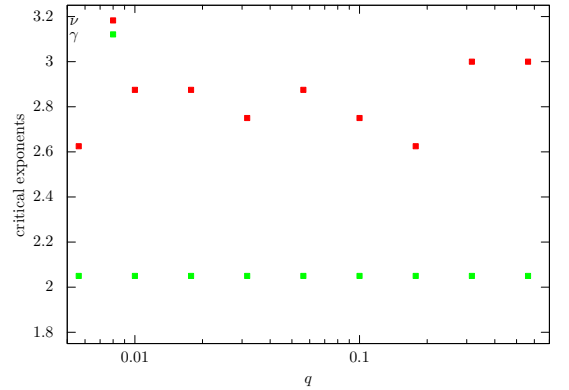


FIG. D66: q -dependence of the critical exponents ν and γ . We consider $X \rightarrow XX$ and set $K = 10$, $J = 1.0$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$.

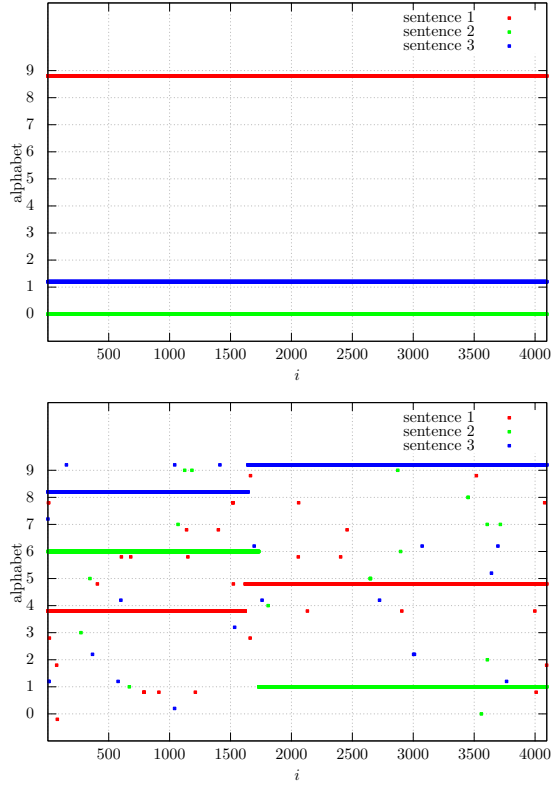


FIG. D67: Typical configurations of symbols for $X \rightarrow XX$. (upper) We considered states whose magnetization lies between 0.6600 and 0.6800 at $k_B T = 0.440$, and (lower) states whose magnetization lies between 0.9800 and 1.0000 at $k_B T = 0.0010$. We set $K = 10$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$.

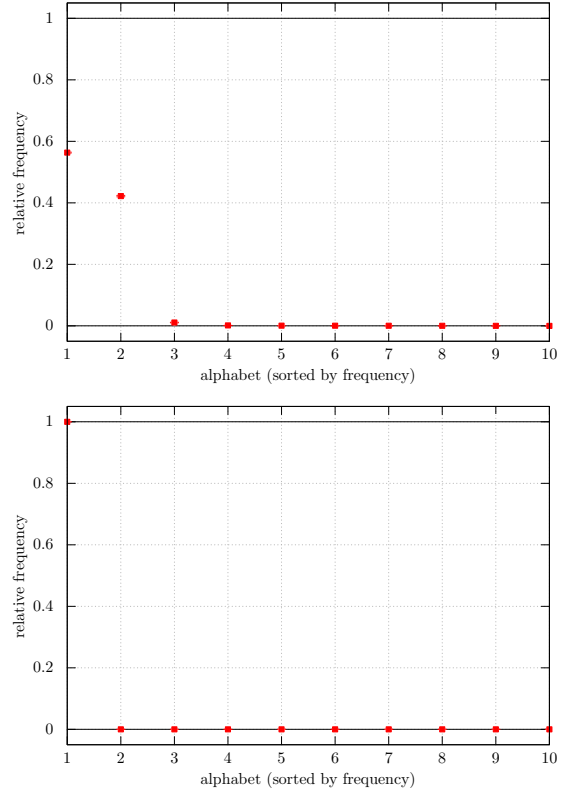


FIG. D68: Relative frequencies of letters in the alphabet for $X \rightarrow XX$. (upper) We considered states whose magnetization lies between 0.6600 and 0.6800 at $k_B T = 0.440$, and (lower) states whose magnetization lies between 0.9800 and 1.0000 at $k_B T = 0.0010$. We set $K = 10$, $J = 1.0$, $q = 10^{-2.0}$, $t = 0$, $s = 0.9$, and $r_- = r_+ = 0.25$. Symbols on the horizontal axis are sorted by frequency and incremented by unity. The scales of the vertical and horizontal axes are linear.

- [1] S. T. Piantadosi, *Psychonomic bulletin & review* **21**, 1112 (2014).
- [2] H. W. Lin and M. Tegmark, *Entropy* **19** (2017), ISSN 1099-4300, URL <https://www.mdpi.com/1099-4300/19/7/299>.
- [3] M. Simkin and V. Roychowdhury, *Physics Reports* **502**, 1 (2011), ISSN 0370-1573, URL <https://www.sciencedirect.com/science/article/pii/S0370157310003339>.
- [4] OpenAI, *Hello GPT-4o* (2022), URL <https://openai.com/index/chatgpt/>.
- [5] S. Pichai and D. Hassabis, *Introducing gemini: our largest and most capable ai model* (2023), URL <https://blog.google/technology/ai/google-gemini-ai/>.
- [6] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, *ACM Transactions on Intelligent Systems and Technology* **16**, 1 (2025).
- [7] M. U. Hadi, Q. Al Tashi, A. Shah, R. Qureshi, A. Muneer, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, et al., *Authorea Preprints* (2024).
- [8] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, *arXiv preprint arXiv:2001.08361* (2020).
- [9] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., *arXiv preprint arXiv:2206.07682* (2022).
- [10] A. Chen, R. Schwartz-Ziv, K. Cho, M. L. Leavitt, and N. Saphra, *arXiv preprint arXiv:2309.07311* (2023).
- [11] E. DeGiuli, *Phys. Rev. Lett.* **122**, 128301 (2019), URL <https://link.aps.org/doi/10.1103/PhysRevLett.122.128301>.
- [12] E. D. Giuli, *Journal of Physics A: Mathematical and Theoretical* **52**, 504001 (2019), URL <https://doi.org/10.1088/1751-8121/ab293c>.
- [13] K. Nakaishi and K. Hukushima, *Phys. Rev. Research* **4**, 023156 (2022), URL <https://link.aps.org/doi/10.1103/PhysRevResearch.4.023156>.
- [14] F. Lalegani and E. De Giuli, *arXiv preprint arXiv:2309.14913* (2023).
- [15] K. Nakaishi and K. Hukushima, *arXiv preprint arXiv:2402.07113* (2024).
- [16] V. Berezinskii, *Sov. Phys. JETP* **32**, 493 (1971).
- [17] V. Berezinskii, *Sov. Phys. JETP* **34**, 610 (1972).
- [18] J. M. Kosterlitz and D. J. Thouless, *Journal of Physics C: Solid State Physics* **6**, 1181 (1973), URL <https://dx.doi.org/10.1088/0022-3719/6/7/010>.
- [19] J. M. Kosterlitz, *Reports on Progress in Physics* **79**, 026001 (2016), URL <https://dx.doi.org/10.1088/0034-4885/79/2/026001>.
- [20] J. M. Kosterlitz, *Journal of Physics C: Solid State Physics* **7**, 1046 (1974), URL <https://dx.doi.org/10.1088/0022-3719/7/6/005>.
- [21] G. Jäger and J. Rogers, *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 1956 (2012).
- [22] G. Delétang, A. Ruoss, J. Grau-Moya, T. Genewein, L. K. Wenliang, E. Catt, C. Cundy, M. Hutter, S. Legg, J. Veness, et al., *arXiv preprint arXiv:2207.02098* (2022).
- [23] V. Ardourel and S. Bangu, *Studies in History and Philosophy of Science* **100**, 99 (2023), ISSN 0039-3681, URL <https://www.sciencedirect.com/science/article/pii/S0039368123000900>.
- [24] Y.-D. Hsieh, Y.-J. Kao, and A. W. Sandvik, *Journal of Statistical Mechanics: Theory and Experiment* **2013**, P09001 (2013), URL <https://dx.doi.org/10.1088/1742-5468/2013/09/P09001>.
- [25] Z. Zuo, S. Yin, X. Cao, and F. Zhong, *Phys. Rev. B* **104**, 214108 (2021), URL <https://link.aps.org/doi/10.1103/PhysRevB.104.214108>.
- [26] H. Ueda, K. Okunishi, K. Harada, R. Krčmár, A. Gendiar, S. Yunoki, and T. Nishino, *Phys. Rev. E* **101**, 062111 (2020), URL <https://link.aps.org/doi/10.1103/PhysRevE.101.062111>.
- [27] M. Hasenbusch, *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P08003 (2008), URL <https://dx.doi.org/10.1088/1742-5468/2008/08/P08003>.
- [28] L. M. Tuan, T. T. Long, D. X. Nui, P. T. Minh, N. D. Trung Kien, and D. X. Viet, *Phys. Rev. E* **106**, 034138 (2022), URL <https://link.aps.org/doi/10.1103/PhysRevE.106.034138>.
- [29] E. Bayong, H. T. Diep, and V. Dotsenko, *Phys. Rev. Lett.* **83**, 14 (1999), URL <https://link.aps.org/doi/10.1103/PhysRevLett.83.14>.
- [30] F. J. Dyson, *Matematika* **16**, 137 (1972).
- [31] J. Fröhlich and T. Spencer (1982).
- [32] F. J. Dyson, *Matematika* **16**, 113 (1972).
- [33] F. J. Dyson, *Communications in Mathematical Physics* **21**, 269 (1971).
- [34] D. J. Thouless, *Phys. Rev.* **187**, 732 (1969), URL <https://link.aps.org/doi/10.1103/PhysRev.187.732>.
- [35] Y. Chang, L. Sun, and X. Cai, *Phys. Rev. E* **76**, 021101 (2007), URL <https://link.aps.org/doi/10.1103/PhysRevE.76.021101>.
- [36] J. Martínez-Herrera, O. Rodríguez-López, and M. Solís, *Physica A: Statistical Mechanics and its Applications* **596**, 127136 (2022), ISSN 0378-4371, URL <https://www.sciencedirect.com/science/article/pii/S0378437122001534>.
- [37] Y. Tomita, *Journal of the Physical Society of Japan* **78**, 014002 (2009), <https://doi.org/10.1143/JPSJ.78.014002>, URL <https://doi.org/10.1143/JPSJ.78.014002>.
- [38] P. W. Anderson and G. Yuval, *Journal of Physics C: Solid State Physics* **4**, 607 (1971), URL <https://dx.doi.org/10.1088/0022-3719/4/5/011>.
- [39] J. M. Kosterlitz, *Phys. Rev. Lett.* **37**, 1577 (1976), URL <https://link.aps.org/doi/10.1103/PhysRevLett.37.1577>.
- [40] M. Aizenman and R. Fernández, *Letters in Mathematical Physics* **16**, 39 (1988).
- [41] M. Aizenman, J. T. Chayes, L. Chayes, and C. M. Newman, *Journal of Statistical Physics* **50**, 1 (1988).
- [42] D. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, 2021), ISBN 9781108490146, URL <https://books.google.co.jp/books?id=WW8yEAAAQBAJ>.
- [43] J. Binney, N. Dowrick, A. Fisher, and M. Newman, *The Theory of Critical Phenomena: An Introduction to the Renormalization Group* (Clarendon Press, 1992), ISBN 9780191660566, URL <https://books.google.co.jp/books?id=BC0QDwAAQBAJ>.
- [44] M. Newman and G. Barkema, *Monte Carlo Methods in Statistical Physics* (Clarendon Press, 1999),

- ISBN 9780198517962, URL <https://books.google.co.jp/books?id=KKL2nQEACAAJ>.
- [45] P. W. Anderson and G. Yuval, Phys. Rev. Lett. **23**, 89 (1969), URL <https://link.aps.org/doi/10.1103/PhysRevLett.23.89>.
 - [46] G. Yuval and P. W. Anderson, Phys. Rev. B **1**, 1522 (1970), URL <https://link.aps.org/doi/10.1103/PhysRevB.1.1522>.
 - [47] P. W. Anderson, G. Yuval, and D. R. Hamann, Phys. Rev. B **1**, 4464 (1970), URL <https://link.aps.org/doi/10.1103/PhysRevB.1.4464>.
 - [48] D. R. Hamann, Phys. Rev. Lett. **23**, 95 (1969), URL <https://link.aps.org/doi/10.1103/PhysRevLett.23.95>.
 - [49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Advances in neural information processing systems **30** (2017).
 - [50] R. Rende, F. Gerace, A. Laio, and S. Goldt, Phys. Rev. Res. **6**, 023057 (2024), URL <https://link.aps.org/doi/10.1103/PhysRevResearch.6.023057>.
 - [51] R. Schaeffer, B. Miranda, and S. Koyejo, Advances in neural information processing systems **36**, 55565 (2023).
 - [52] H. Cui, F. Behrens, F. Krzakala, and L. Zdeborová, Advances in Neural Information Processing Systems **37**, 36342 (2024).
 - [53] F. Cagnetta, L. Petrini, U. M. Tomasini, A. Favero, and M. Wyart, Phys. Rev. X **14**, 031001 (2024), URL <https://link.aps.org/doi/10.1103/PhysRevX.14.031001>.
 - [54] F. Cagnetta, A. Favero, A. Sclocchi, and M. Wyart, arXiv preprint arXiv:2505.07070 (2025).
 - [55] J. Garnier-Brun, M. Mézard, E. Moscatto, and L. Saglietti, arXiv preprint arXiv:2408.15138 (2024).
 - [56] A. Sclocchi, A. Favero, and M. Wyart, Proceedings of the National Academy of Sciences **122**, e2408799121 (2025), <https://www.pnas.org/doi/pdf/10.1073/pnas.2408799121>, URL <https://www.pnas.org/doi/abs/10.1073/pnas.2408799121>.
 - [57] M. Sipser, *Introduction to the Theory of Computation* (Cengage Learning, 2012), ISBN 9781133187790, URL <https://books.google.co.jp/books?id=H94JzgEACAAJ>.
 - [58] S. Stengele, D. Drexel, and G. De las Cuevas, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **479**, 20220553 (2023), <https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.2022.0553>, URL <https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2022.0553>.
 - [59] F. Y. Wu, Rev. Mod. Phys. **54**, 235 (1982), URL <https://link.aps.org/doi/10.1103/RevModPhys.54.235>.
 - [60] Z. Budrikis, Nature Reviews Physics **6**, 530 (2024).
 - [61] E. Luijten, *Interaction range, universality and the upper critical dimension* (1997).
 - [62] M. Krech and E. Luijten, Phys. Rev. E **61**, 2058 (2000), URL <https://link.aps.org/doi/10.1103/PhysRevE.61.2058>.
 - [63] K. Fukui and S. Todo, Journal of Computational Physics **228**, 2629 (2009).
 - [64] Y. Kumano, K. Hukushima, Y. Tomita, and M. Oshikawa, Phys. Rev. B **88**, 104427 (2013), URL <https://link.aps.org/doi/10.1103/PhysRevB.88.104427>.
 - [65] Y. Miyajima, Y. Murata, Y. Tanaka, and M. Mochizuki, Phys. Rev. B **104**, 075114 (2021), URL <https://link.aps.org/doi/10.1103/PhysRevB.104.075114>.
 - [66] A. W. Sandvik, in *AIP Conference Proceedings* (American Institute of Physics, 2010), vol. 1297, pp. 135–338.
 - [67] M. Fujimoto and H. Otsuka, Phys. Rev. E **102**, 032141 (2020), URL <https://link.aps.org/doi/10.1103/PhysRevE.102.032141>.