Generative Criticality in Large Language Model Temperature Scaling

Huajian Ruan[1,2,3] **Jinyang Li**[4,5,6] **Xingyu Guo**[1,2,3,*] **Lingxiao Wang**[5,7,*]

[1]State Key Laboratory of Nuclear Physics and Technology, Institute of Quantum Matter, South China Normal University, Guangzhou 510006, China
[2]Key Laboratory of Atomic and Subatomic Structure and Quantum Control (MOE), Guangdong-Hong Kong Joint Laboratory of Quantum Matter, Guangzhou 510006, China
[3]Guangdong Basic Research Center of Excellence for Structure and Fundamental Interactions of Matter, Guangdong Provincial Key Laboratory of Nuclear Science, Guangzhou 510006, China
[4]KEK Theory Center, Institute of Particle and Nuclear Studies, Tsukuba, Japan
[5]RIKEN Center for Interdisciplinary Theoretical and Mathematical Sciences (iTHEMS), Wako, Saitama 351-0198, Japan
[6]Graduate University for Advanced Studies (SOKENDAI), Oho 1-1, Tsukuba, Ibaraki 305-0801, Japan
[7]Institute for Physics of Intelligence, The University of Tokyo, Hongo, Tokyo 113-0033, Japan
`guoxy@m.scnu.edu.cn, lingxiao.wang@riken.jp`
*Corresponding authors
Code: https://github.com/Lastail/Critical

Abstract

We propose a statistical-field framework for text generated by large language models (LLMs), treating token embeddings as continuous spin variables on a one-dimensional chain. Defining a susceptibility from the connected two-point correlator and an order parameter from the ensemble-averaged embedding field, we vary the `softmax` temperature T and observe a sharp susceptibility peak near a characteristic T_c with power-law-like scaling, a concurrent rapid change in the order parameter, and a collapse onto a single semantic direction below T_c. The intrinsic dimension estimated by the two nearest neighbor (TwoNN) method independently corroborates these findings, reaching a minimum near T_c. Results are robust across model scales (Qwen3: 0.6B–32B) and prompt categories. While the phenomenology closely resembles a continuous phase transition, the non-equilibrium nature of autoregressive generation warrants further investigation. Our framework provides quantitative tools for probing the collective statistical structure of LLM outputs and suggests connections between decoding strategies and critical phenomena.

1 Introduction

Understanding the structure of language remains a central challenge in linguistics [1, 2] and artificial intelligence. While large language models (LLMs) have enabled text modeling at unprecedented scale [3], most analyses rely on information-theoretic quantities such as entropy and mutual information [4, 5, 6], which do not capture microstructural or emergent macroscopic behavior. Recent work has begun mapping LLM-generated texts to statistical field theories [7, 8], but rigorous definitions of the relevant physical quantities are still lacking.

We address this gap by constructing a statistical-mechanics framework for LLM-generated text [9]. Within the token embedding space, we define susceptibility and an order parameter [10], and study their behavior as a function of the `softmax` temperature T [11, 12]. We find that these quantities exhibit critical behavior [13, 14]: near a critical temperature T_c, the susceptibility diverges and the order parameter undergoes a rapid change. To corroborate this from a geometric perspective, we apply the two nearest neighbor (TwoNN) intrinsic-dimension estimator [15, 16], which independently identifies the same critical region through non-monotonic features in $I_d(T)$.

Our main contributions are: (i) a statistical-field framework for LLM outputs with well-defined physical observables; (ii) evidence of critical behavior driven by the temperature parameter; and (iii) independent geometric validation via the TwoNN-estimated intrinsic dimension.

2 LLM Building Blocks and the O(*N*) Model

An LLM tokenizes input text into subword units via a fixed vocabulary [17, 18, 19], then maps each token to an N-dimensional vector through a trainable embedding matrix [18, 20]. Since token embeddings are high-dimensional vectors whose norms concentrate around a characteristic scale, this motivates a heuristic physical picture, treating the text sequence as a one-dimensional lattice chain of N-dimensional vectors [21, 10, 22], for constructing an effective Hamiltonian over the high-dimensional semantic space,

$$\mathcal{H} = \sum_{\sigma,\tau} J_{\sigma\tau}\, t_\sigma\, t_\tau + \sum_{\sigma} H_\sigma\, t_\sigma\,, \tag{1}$$

where J is the coupling matrix (encoding non-local interactions between all pairs of lattice sites), H is an external field set by the prompt, and Greek indices label all different lattice sites. To simplify the problem, we omit the higher-order interaction terms, e.g., $O(t^2)$, but LLMs can still determine the effective couplings and thus governs how each site responds to H.

Inter-token interactions are mediated by the self-attention mechanism [11]. Given query, key, and value projections $Q_i{=}F_Q(t_i)$, $K_i{=}F_K(t_i)$, $\mathcal{V}_i{=}F_\mathcal{V}(t_i)$:

$$\mathrm{Attention}(Q,K,\mathcal{V}) = \mathrm{softmax}\big(QK^T/\sqrt{d_k}\big)\,\mathcal{V}\,, \tag{2}$$

where d_k is the dimension of the key vectors. Stacking M layers of $\mathrm{FF}\circ\mathrm{Attention}$ with residual connections yields the full Transformer [11]. At generation time, the next-token distribution is controlled by temperature T [23]: $p_i = \exp(z_i/T)/\sum_j \exp(z_j/T)$, interpolating between deterministic ($T{\to}0$) and uniform ($T{\to}\infty$) sampling. Although T is not a thermodynamic temperature in the strict physical sense, its structural role in the softmax mirrors that of temperature in the Boltzmann distribution, and we treat this parallel as a formal correspondence.

3 Generative Criticality

We apply tools from statistical field theory and intrinsic dimension estimation to probe the critical structure of LLM outputs.

Critical Behavior. We define the susceptibility of generated text as,

$$\chi = \frac{1}{L}\sum_{\sigma,\tau}\Big[\frac{1}{N_s}\sum_i t_\sigma^{(i)}\cdot t_\tau^{(i)} - \frac{1}{N_s^2}\sum_i t_\sigma^{(i)}\cdot\sum_j t_\tau^{(j)}\Big], \tag{3}$$

where L is the sequence length, Greek indices σ,τ label lattice sites (token positions), Latin indices i,j label members of an ensemble of size N_s, and the dot products are taken in the N-dimensional embedding space. Although this expression is formally analogous to a variance, the non-trivial content lies in the diverging fluctuations near a critical temperature T_c, where $\chi \sim |T-T_c|^{-\gamma}$ [24], consistent with standard thermodynamic scaling. Similarly, the ensemble-averaged token expectation,

$$\langle t\rangle = \frac{1}{N_s L}\sum_{i,\sigma} t_\sigma^{(i)} \tag{4}$$

serves as an order parameter [25]: in the high-temperature (disordered) phase it tends toward zero; in the low-temperature (ordered) phase, deterministic generation yields a nonzero ensemble average, analogous to the spontaneous breaking of O(N) symmetry below T_c.

Intrinsic Dimension. The TwoNN method estimates the intrinsic dimension I_d of a data manifold from local distance ratios [26]. For each point, defining $\mu = r_2/r_1$, local uniformity implies $f(\mu) = I_d\, \mu^{-I_d-1}$, yielding

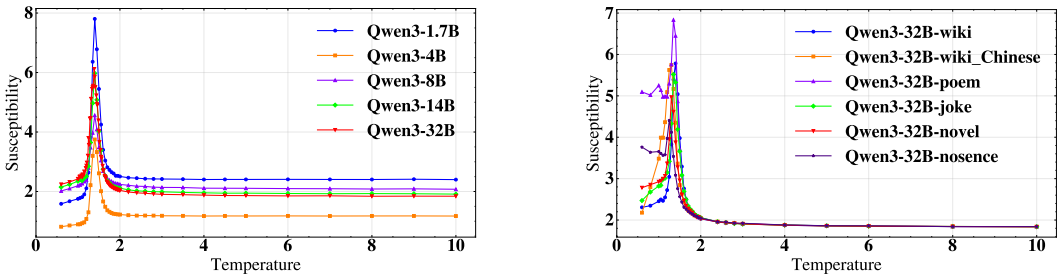
$$-\ln\big(1 - F(\mu)\big) = I_d \cdot \ln \mu\,, \tag{5}$$

where $F(\mu)$ is the cumulative distribution function. The intrinsic dimension acts as an unsupervised phase-transition detector [15]: I_d is small in the ordered phase, approaches the total degrees of freedom in the disordered phase, and exhibits non-monotonic features at criticality.

4 Experiments and Main Results

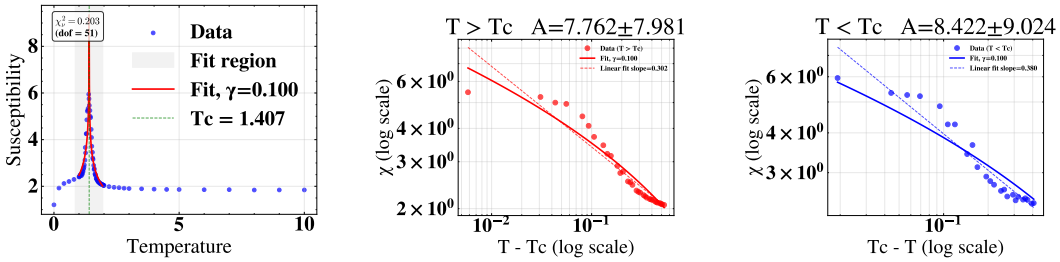
We use the Qwen3 family [27] (0.6B–32B), fixing output length to 300 tokens. Prompts are drawn from English Wikipedia; controls include Chinese Wikipedia, jokes, poems, novels, and nonsensical texts. All generations use "no-think" mode (i.e., extended reasoning is disabled). For each temperature we generate $N_s = 1{,}000$ samples and compute the ensemble average over token positions to obtain the order parameter. It should be noted that the results we presented are embeddings in different dimensions(Table 1).

Susceptibility. Figure 1(a) shows $\chi(T)$ for different model scales; Figure 1(b) shows results across prompt categories. A pronounced peak appears near T_c, with power-law scaling $\chi \sim (T - T_c)^{-\gamma}$ and $\gamma \approx 0.1$ on both sides (Figure 2. We also present the fitting results for different sizes in the appendix(Figure 7). The curves do not collapse across models or prompts, indicating that parameter count and prompt type act as distinct effective fields. Convergence with increasing ensemble size and sequence length is confirmed in Figure 3.



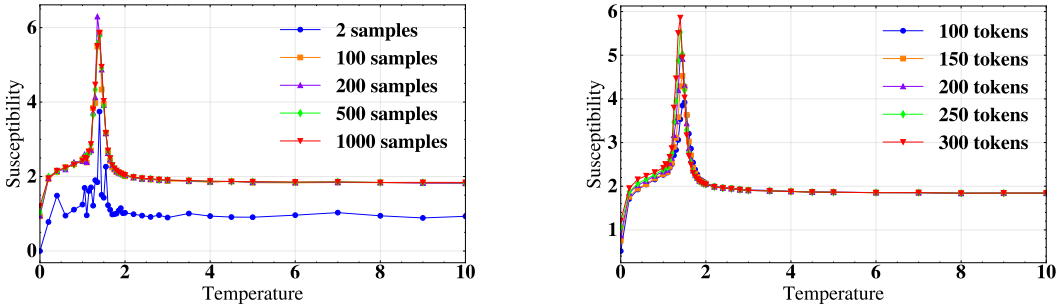
(a) (b)

Figure 1: Susceptibility χ vs. temperature T for (a) different model scales using Wikipedia prompts (1000 samples, 280 tokens). All models exhibit a peak near $T_c \approx 1.4$, with larger models showing higher peak susceptibility. (b) Different prompt categories generated by Qwen3-32B (1000 samples, 300 tokens).



(a) (b) (c)

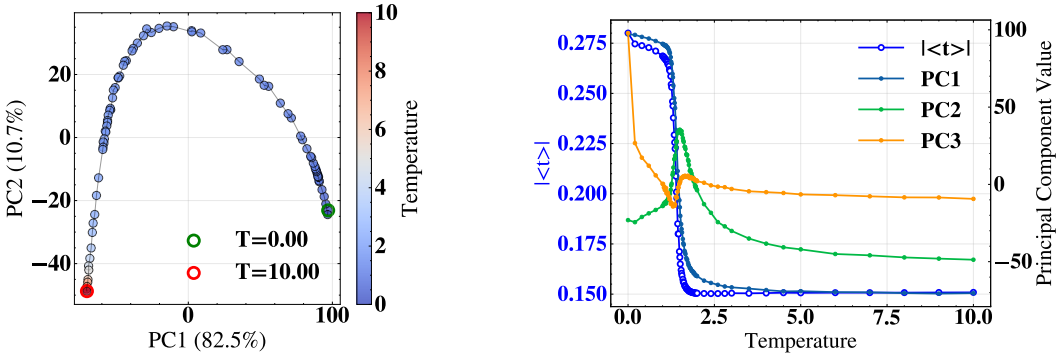
Figure 2: Power-law fits of susceptibility near T_c for Qwen3-32B. (a) Full $\chi(T)$ curve with the fit region highlighted; (b) log-log plot for $T > T_c$; (c) log-log plot for $T < T_c$. Both sides yield a critical exponent $\gamma \approx 0.1$, consistent with power-law divergence.



(a) (b)

Figure 3: Stability analysis of the susceptibility signal. (a) Convergence with increasing ensemble size N_s from 2 to 1000 samples; (b) Convergence with increasing sequence length from 100 to 300 tokens.

Order Parameter Structure. Applying PCA to the ensemble-averaged order-parameter vector across temperatures, we find that generated texts concentrate along a single semantic direction below T_c, with a sharp directional change at criticality (Figure 4). This confirms the phase transition and suggests that, at low temperatures, the LLM effectively selects tokens approximating a unique semantic target.



(a) (b)

Figure 4: Order parameter structure via PCA. (a) Projection onto the first two principal components (PC1 explains 82.5%, PC2 explains 10.7%), colored by temperature. Below T_c, ensemble outputs cluster tightly along a single direction; above T_c, they disperse. (b) Magnitude $|\langle t \rangle|$ and the first three principal component values vs. temperature, showing a sharp transition near T_c.

Intrinsic Dimension. We apply the TwoNN method to sentence-level embedding vectors (obtained via average pooling over token embeddings) of $N_s = 1{,}000$ texts at each temperature (Figure 5). The intrinsic dimension I_d reaches a minimum near T_c and rises rapidly above it: at low T the configuration space is restricted; near T_c it enters a critical regime; at high T the system explores the full configuration space. Fits are excellent at all nonzero temperatures (see Appendix A).

5 Conclusion

We have presented a statistical-field framework for LLM-generated text, defining susceptibility and an order parameter in the token embedding space. Systematic variation of the softmax temperature reveals phase-transition-like behavior: a divergent susceptibility peak near T_c, a sharp change in the order parameter accompanied by dimensional collapse onto a single semantic direction, and a minimum in the TwoNN intrinsic dimension in the same critical region. These mutually consistent signatures validate the feasibility of statistical-field analysis for probing the collective structure of LLM outputs. Future directions include applying renormalization-group methods [28] to capture long-

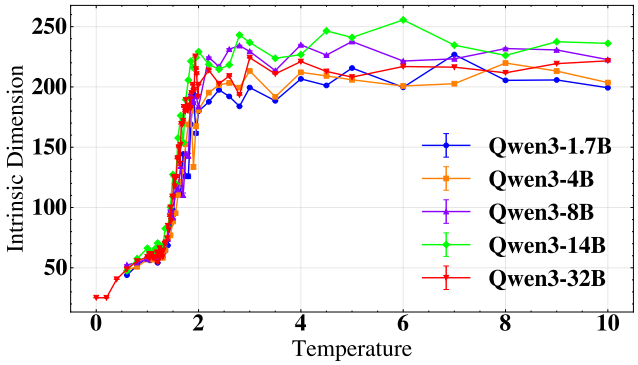


Figure 5: Intrinsic dimension I_d vs. temperature for five Qwen3 model scales. All models exhibit a minimum near $T_c \approx 1.3$–1.5, after which I_d rises sharply and saturates at high T, consistent with the onset of a disordered phase where the configuration space expands.

range correlations, and leveraging the critical behavior identified here as a diagnostic for trustworthy AIfor instance, the sharp onset of the disordered phase may serve as a quantitative indicator of regimes in which LLM outputs become unreliable.

References

[1] Steven Pinker. *The language instinct: How the mind creates language*. Penguin uK, 2003.

[2] Noam Chomsky. *Syntactic structures*. Walter de Gruyter, 2002.

[3] Melanie Mitchell and David C Krakauer. The debate over understanding in ais large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.

[4] Colin Scheibner, Lindsay M. Smith, and William Bialek. Large language models and the entropy of english. *ArXiv*, abs/2512.24969, 2025.

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

[6] Wentian Li. Mutual information functions of natural language texts. Santa Fe Institute Santa Fe, NM, USA, 1989.

[7] Kai Nakaishi, Yoshihiko Nishikawa, and Koji Hukushima. Critical phase transition in large language models. *arXiv preprint arXiv:2406.05335*, 2024.

[8] Youran Sun and Babak Haghighat. Phase transitions in large language models and the $o(n)$ model. *arXiv preprint arXiv:2501.16241*, 2025.

[9] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.

[10] Daniel J Amit and Victor Martin-Mayor. *Field theory, the renormalization group, and critical phenomena: graphs to computers*. World Scientific, 2005.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[12] Yuqi Zhu, Jia Li, Ge Li, YunFei Zhao, Zhi Jin, and Hong Mei. Hot or cold? adaptive temperature sampling for code generation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 437–445, 2024.

[13] Henry W Lin and Max Tegmark. Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7):299, 2017.

[14] Cheng-Shang Chang. A simple explanation for the phase transition in large language models with list decoding. *arXiv preprint arXiv:2303.13112*, 2023.

[15] T. Mendes-Santos, X. Turkeshi, M. Dalmonte, and Alex Rodriguez. Unsupervised learning universal critical behavior via the intrinsic dimension. *Physical Review X*, 11(1), February 2021.

[16] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

[17] Christopher M Bishop and Hugh Bishop. *Deep learning: Foundations and concepts*. Springer Nature, 2023.

[18] Zhen Yang, Yingxue Zhang, Fandong Meng, and Jie Zhou. Teal: Tokenize and embed all for multi-modal large language models. *arXiv preprint arXiv:2311.04589*, 2023.

[19] Sandeep Mehta, Darpan Shah, Ravindra Kulkarni, and Cornelia Caragea. Semantic tokenizer for enhanced natural language processing. *arXiv preprint arXiv:2304.12404*, 2023.

[20] Roman Egger. Text representations and word embeddings: Vectorizing textual data. In *Applied data science in tourism: Interdisciplinary approaches, methodologies, and applications*, pages 335–361. Springer, 2022.

[21] Sidney Coleman, Roman Jackiw, and HDavid Politzer. Spontaneous symmetry breaking in the o (n) model for large n. *Physical Review D*, 10(8):2491, 1974.

[22] Bertrand Eynard and Charlotte Kristjansen. Exact solution of the o (n) model on a random lattice. *Nuclear Physics B*, 455(3):577–618, 1995.

[23] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[24] Leo P Kadanoff. Spin-spin correlations in the two-dimensional ising model. *Il Nuovo Cimento B (1965-1970)*, 44(2):276–305, 1966.

[25] Jean Zinn-Justin. *Quantum field theory and critical phenomena*, volume 171. Oxford university press, 2021.

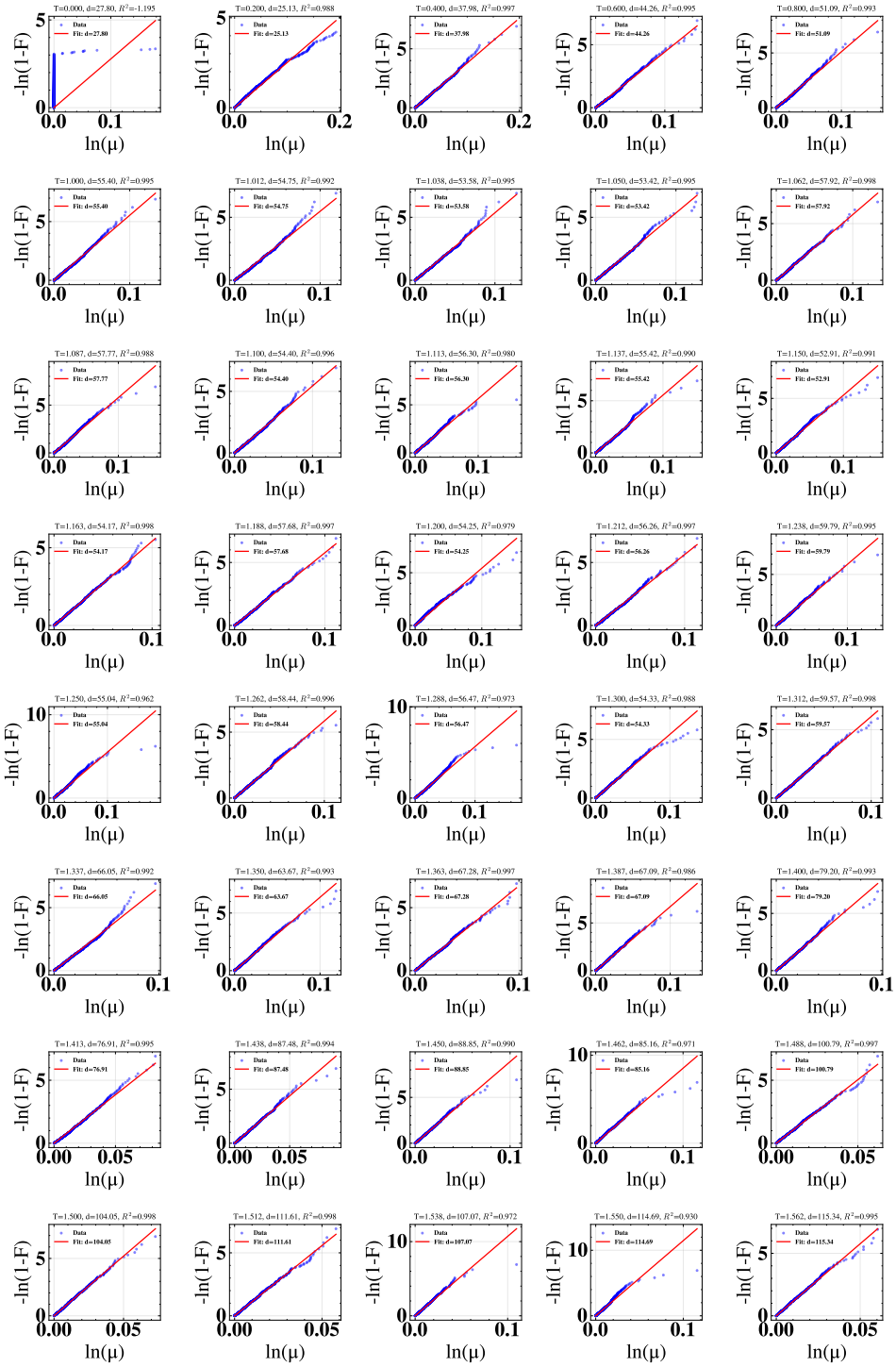
[26] Elena Facco, Maria dErrico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

[27] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[28] Kenneth G Wilson and John Kogut. The renormalization group and the ϵ expansion. *Physics Reports*, 12(2):75–199, 1974.

A TwoNN Fitting

We present the data and fitting plot for the TwoNN method applied to Qwen332B in the appendix to show more details.



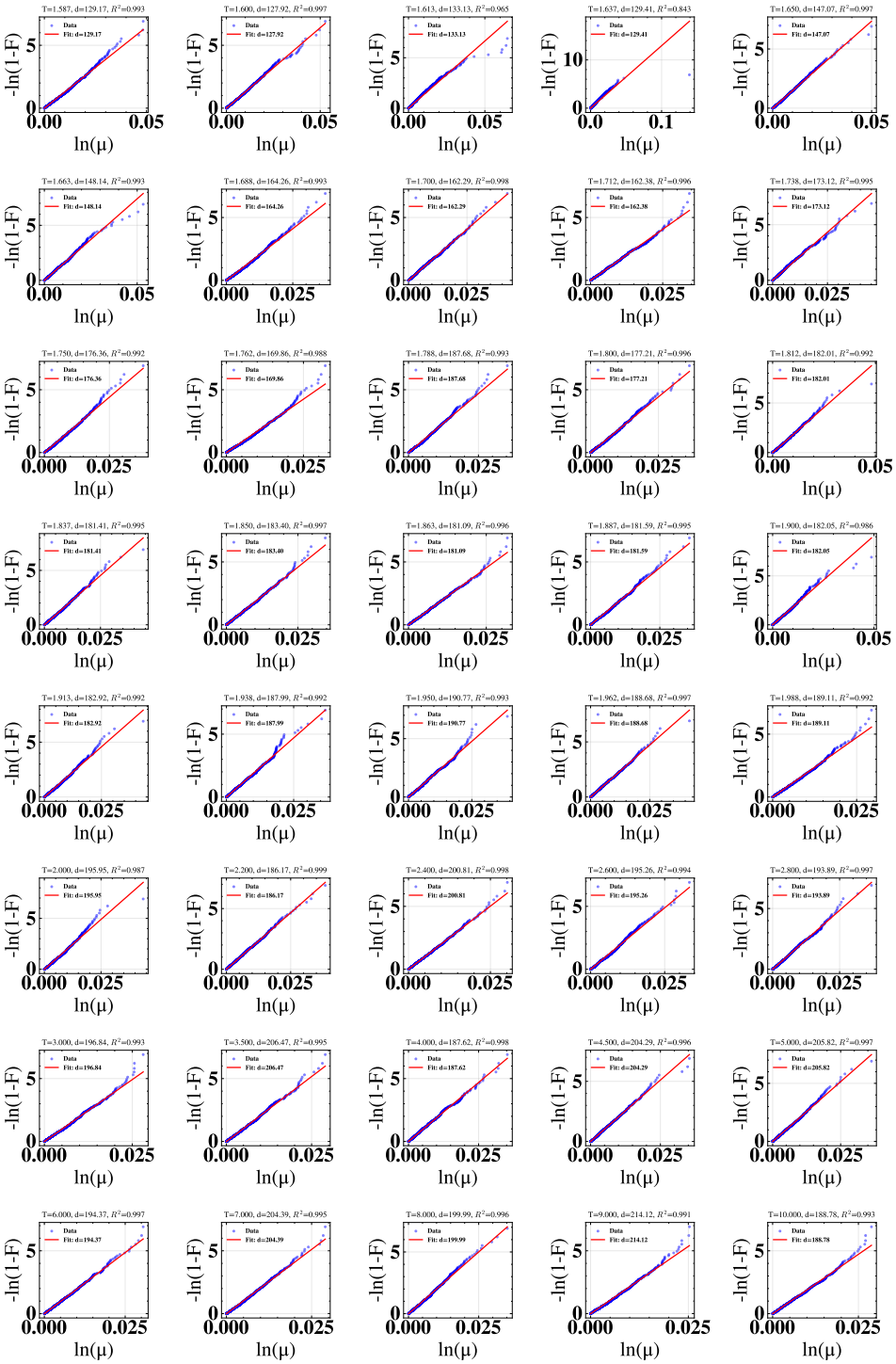


Figure 6: By evaluating the ratio of the second-nearest neighbor distance to the nearest neighbor distance among these data points and fitting its distribution, the derived slope yields the intrinsic dimension.

B Configurations of Qwen3 models

We include a table to illustrate their varying semantic spaces across different embedding dimensions.

Model	Hidden size	Layers	Attention heads (Q / KV)
Qwen3-0.6B	1024	28	16 / 8
Qwen3-1.7B	2048	28	16 / 8
Qwen3-4B	2560	36	32 / 8
Qwen3-8B	4096	36	32 / 8
Qwen3-14B	5120	48	40 / 8
Qwen3-32B	8192	64	64 / 8

Table 1: Architectural parameters of the Qwen3 series models.

C Testing scaling laws at different sizes

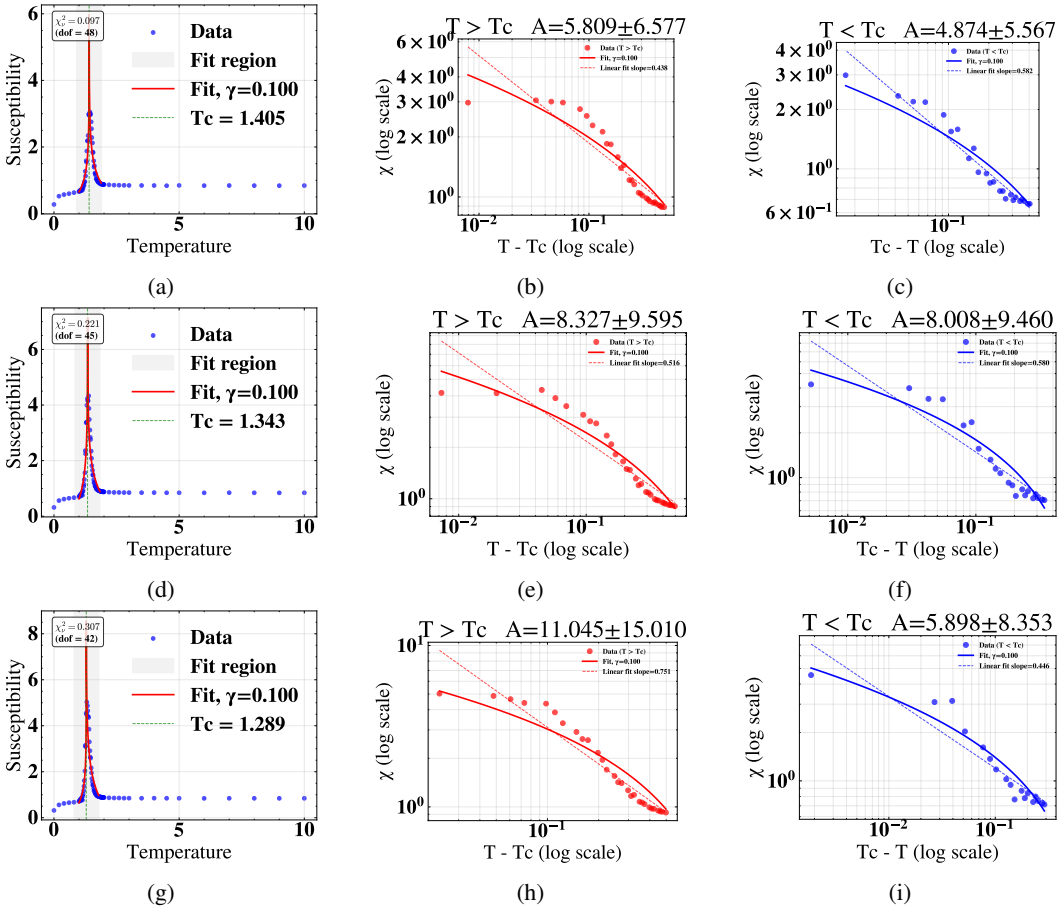


Figure 7: Scaling laws tests at different sizes. Panels (a), (b), and (c) show exponent fitting for a system size of 200 tokens; panels (d), (e), and (f) for 400 tokens; panels (g), (h), and (i) for 500 tokens.