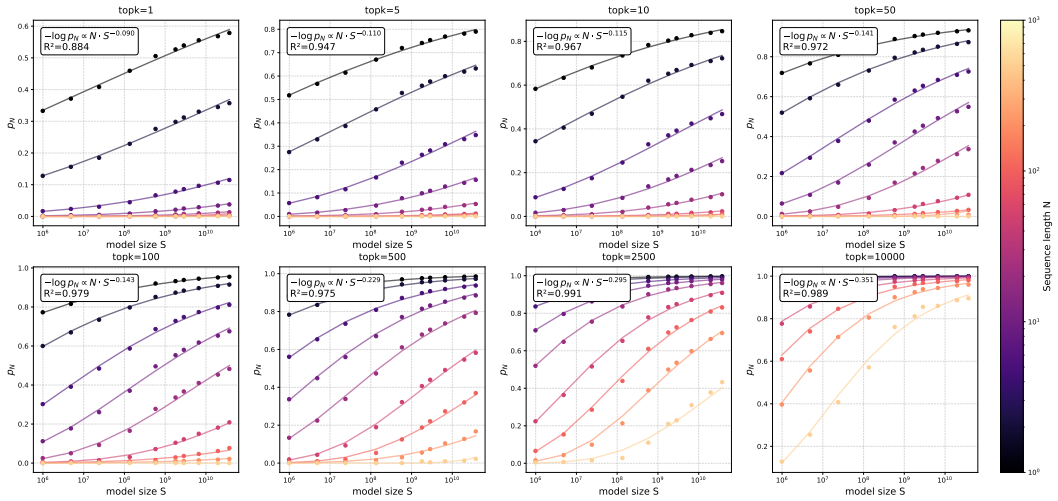


Figure 7: Illustration of the emergence phenomena, $-\log \mathrm{p_{N,k}} \propto \mathrm{N} \cdot \mathrm{S}^{-\alpha}$. Each subplot corresponds to a fixed k, with colored lines for different N. The fitted power-law curves align closely with empirical data, confirming the law across a wide range of conditions. Crucially, the resulting curves naturally exhibit the sigmoidal shape characteristic of observed emergence phenomena, demonstrating that emergence follows directly from smooth microscopic scaling.

for constant C, the probability curve is inherently sigmoidal:

- For small S, the exponent is large and negative, driving $\mathrm{p_{N,k}}$ close to zero.
- As S grows, the exponent approaches zero, and $\mathrm{p_{N,k}}$ rises sharply toward one.

The sequence length N acts as an amplifier: larger N sharpens the transition, creating the observed sudden "knee" in performance.

In summary, emergence is not a breakdown of scaling laws but a predictable macroscopic effect of a smooth power law at the token level. The perceived phase transition is simply the exponential amplification of microscopic trends when mapped to sequence-level tasks under greedy or top-k sampling.

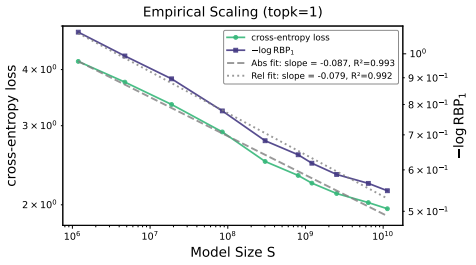
5.2 Connecting Cross-Entropy- and Relative-Based Scaling Laws

The next application is the peculiar connection between the Relative-based scaling law and cross-entropy-based scaling law. As emphasized earlier, the two laws capture fundamentally different aspects of the output distribution: cross-entropy-based law focuses on the probability mass assigned to the ground-truth token, while the relative-based law examines its rank among candidates. By construction, they are not interchangeable. Nevertheless, we observe that both exhibit remarkably similar power-law decay with respect to model size. In particular, when $k = 1$, the decay exponents of cross-entropy (absolute-based) and $-\log \mathrm{p_1}$ (relative-based) are nearly identical, as shown in Figure 8a.

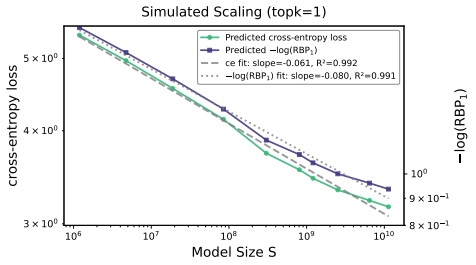
We believe this peculiar coincidence suggests a deeper principles that can derive both laws. Yet this perspective has not been proposed in prior theoretical studies and none of the existing theories can explain this. Therefore, we believe this coincidence shall be the key to the discovery of a deeper scaling theory.

In this paper, we propose a conjecture that can derive both scaling laws. We analyze the distribution of ground-truth token ranks across models of different sizes. Empirical evidence shows that these rank distributions are long-tailed. (Ma et al., 2025; Chatzi et al., 2024; Cai et al., 2024; Zhan et al., 2025). Therefore, our hypothesis is that the rank distribution follows a lognormal distribution:

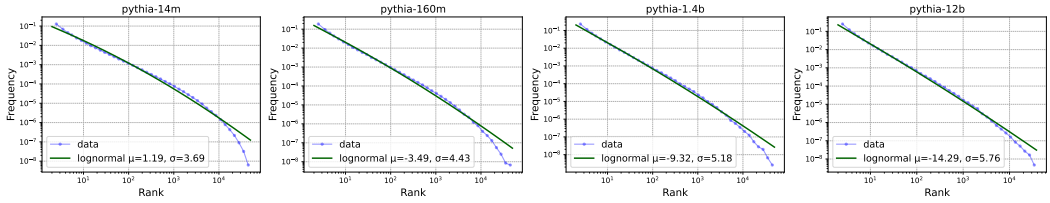
$$P(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0, \tag{9}$$



(a) Empirical results: CE and $-\log \mathrm{RBP}_k$ exhibit similar scaling forms and slopes.



(b) Simulated results: CE and $-\log \mathrm{RBP}_k$ show similar scaling patterns.



(c) Hypothesis: Ground-truth ranking frequency follows the lognormal distribution.

where x denotes the rank of the ground-truth token, and μ and σ are parameters that systematically depend on model size S. Figure 8c shows this is a good fit, which has also been observed in prior studies (Zhan et al., 2025). Under our assumption, we can derive the two scaling laws, as shown in Figure 8b, which is similar to the real results, as shown in Figure 8a. Detailed derivation has been provided in Appendix D. This conjecture aims to show that the two laws together may indicate a deeper theory. For future scaling law theories, they shall not only aim to derive cross-entropy-based scaling law but also Relative-based Scaling Law with the similar scaling exponent.

6 Conclusion

In this paper, we propose Relative-based Probability (RBP) and establish the Relative-Based Scaling Law to study model performance from the perspective of relative ordering. Unlike cross-entropy, RBP captures the probability that the correct token ranks among the top predictions, which is not only a nature performance metric but also a critical in practical scenarios such as greedy decoding or top-k sampling. Extensive experiments across multiple datasets and model families show that this law reliably characterizes performance improvement with scale. Our results demonstrate that Relative-based Scaling Law complements cross-entropy scaling laws. It provides new insights into emergence phenomena and offers an intriguing open question for theoretical studies about scaling laws.

References

Andres Acosta, Dmytro Galkin, Andreas Førde, Svein Arne Huseøy, Tor-Morten Grønli, and Håkon Stensland. A thorough examination of decoding methods in the era of llms. *arXiv preprint arXiv:2407.12467*, 2024.

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. *arXiv preprint arXiv:2301.03728*, 2023.

Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Power lines: Scaling laws for weight decay and batch size in llm pre-training. *arXiv preprint arXiv:2505.13738*, 2025.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, and ... Pythia: A suite for analyzing large language models across

D Connection Between Absolute-Based and Relative-based Scaling Law

We assume that the token probabilities $\mathrm{p_k}$ follow a lognormal distribution:

$$\mathrm{p_k} = \frac{\psi(\mathrm{k})}{\mathrm{c}(\mu,\sigma)}, \quad \text{where} \quad \psi(\mathrm{x}) = \frac{1}{\sqrt{2\pi}\,\sigma\,\mathrm{x}} \exp\Big(-\frac{(\ln \mathrm{x}-\mu)^2}{2\sigma^2}\Big), \quad \mathrm{x}>0, \tag{11}$$

and $c(\mu,\sigma)$ is the normalizing factor:

$$c(\mu,\sigma) = \sum_{k=1}^{\infty} \psi(k). \tag{12}$$

Define the standard normal pdf and cdf as

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \qquad \Phi(t) = \int_{-\infty}^{t} \varphi(u)\,du, \tag{13}$$

and let

$$a = -\frac{\mu}{\sigma}, \qquad P = \int_1^{\infty} \psi(x)\,dx = \Phi\Big(\frac{\mu}{\sigma}\Big). \tag{14}$$

Conveniently, $\psi(1) = \varphi(a)/\sigma$.

We also define the truncated moments (using $t = \ln x$):

$$I_1 \equiv \int_1^{\infty} \psi(x)\ \ln x\,dx = \mu P + \sigma\varphi(a), \tag{15}$$

$$I_2 \equiv \int_1^{\infty} \psi(x)\ \ln^2 x\,dx = (\mu^2+\sigma^2)P + \mu\sigma\varphi(a). \tag{16}$$

Since $\mathrm{p_1} = \psi(1)/\mathrm{c}$, with $\psi(1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\mu^2/(2\sigma^2)}$, we have

$$-\log \mathrm{p_1} = \log \mathrm{c}(\mu,\sigma) + \log(\sqrt{2\pi}\sigma) + \frac{\mu^2}{2\sigma^2}. \tag{17}$$

Starting from the definition of cross-entropy,

$$\mathrm{CE} = -\sum_{k\geq 1} \mathrm{p_k}\log \mathrm{p_k} = \log(\sqrt{2\pi}\sigma) + \log \mathrm{c}(\mu,\sigma) + \frac{1}{\mathrm{c}(\mu,\sigma)}\left[\frac{1}{2\sigma^2}\sum_{\mathrm{k}} \psi(\mathrm{k})(\ln \mathrm{k}-\mu)^2 + \sum_{\mathrm{k}} \psi(\mathrm{k}) \ln \mathrm{k}\right], \tag{18}$$

and using the integral approximations $\sum_k \psi(k)\ln k \approx I_1$, $\sum_k \psi(k)\ln^2 k \approx I_2$, and $c(\mu,\sigma) \approx \frac{1}{2}\psi(1) + P$, we obtain

$$\mathrm{CE} \approx \log(\sqrt{2\pi}\sigma) + \log c + \frac{1}{c}\left[\frac{1}{2\sigma^2}(I_2 - 2\mu I_1 + \mu^2 c) + I_1\right]. \tag{19}$$

Substituting (I_1, I_2) gives

$$\mathrm{CE} \approx \log(\sqrt{2\pi}\sigma) + \log c + \frac{\mu^2}{2\sigma^2} + \frac{1}{c}\left[P\Big(\frac{\mu^2+\sigma^2}{2\sigma^2}+\mu\Big) + \varphi(a)\Big(\frac{\mu}{2\sigma}+\sigma\Big) - \frac{\mu}{\sigma^2}(\mu P + \sigma\varphi(a))\right], \tag{20}$$

with

$$c \approx \tfrac{1}{2}\psi(1) + P = \tfrac{1}{2}\frac{\varphi(a)}{\sigma} + P. \tag{21}$$

Simplifying yields

$$\mathrm{CE} \approx \log(\sqrt{2\pi}\sigma) + \log c + \frac{\mu^2}{2\sigma^2} + \frac{1}{c}\left[P\Big(\frac{\sigma^2-\mu^2}{2\sigma^2}+\mu\Big) + \varphi(a)\Big(\sigma - \frac{\mu}{2\sigma}\Big)\right]. \tag{22}$$

Since

$$\Phi(x) = o(\varphi(x)) \quad \text{as } x \to -\infty, \tag{23}$$

and

$$\psi(1) = \frac{\varphi(a)}{\sigma} = o(\varphi(a)) \quad \text{as } \sigma \to \infty, \tag{24}$$

we have

$$c(\mu, \sigma) = o\big(\varphi(\tfrac{\mu}{\sigma})\big). \tag{25}$$

Thus, as the model scales, eventually

$$-\log c > -\log(\varphi(a)) = \tfrac{1}{2}\log(2\pi) + \tfrac{\mu^2}{2\sigma^2}, \tag{26}$$

So $\log c$ decays faster than the term $\frac{\mu^2}{2\sigma^2}$. In comparison, the factor $\log(\sqrt{2\pi}\sigma)$ increases only slowly with σ, and moreover σ itself grows only mildly with model size. The remaining correction terms in CE, namely the fraction

$$\frac{1}{c}\left[P\Big(\frac{\sigma^2 - \mu^2}{2\sigma^2} + \mu\Big) + \varphi(a)\Big(\sigma - \frac{\mu}{2\sigma}\Big)\right], \tag{27}$$

do not dominate either: the numerator grows at most polynomially in μ and σ, while the denominator c shrinks super-exponentially (since $c = o(\varphi(\mu/\sigma))$). Hence the whole fraction is asymptotically negligible compared to $\log(c)$. Therefore, as $|\mu|$ increases, $\log(c)$ becomes the dominant part of CE. Importantly, this also applies to $-\log(\mathrm{p}_1)$, which explains why CE and $-\log(\mathrm{p}_1)$ behave similarly as the model scales.