RELATIVE-BASED SCALING LAW FOR NEURAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling laws aim to accurately predict model performance across different scales. Existing scaling-law studies almost exclusively rely on cross-entropy as the evaluation metric. However, cross-entropy provides only a partial view of performance: it measures the absolute probability assigned to the correct token, but ignores the relative ordering between correct and incorrect tokens. Yet, relative ordering is crucial for language models, such as in greedy-sampling scenario. To address this limitation, we investigate scaling from the perspective of relative ordering. We first propose the Relative-Based Probability (RBP) metric, which quantifies the probability that the correct token is ranked among the top predictions. Building on this metric, we establish the Relative-Based Scaling Law, which characterizes how RBP improves with increasing model size. Through extensive experiments on four datasets and four model families spanning five orders of magnitude, we demonstrate the robustness and accuracy of this law. Finally, we illustrate the broad application of this law with two examples, namely providing a deeper explanation of emergence phenomena and facilitating finding fundamental theories of scaling laws. In summary, the Relative-Based Scaling Law complements the cross-entropy perspective and contributes to a more complete understanding of scaling large language models. Thus, it offers valuable insights for both practical development and theoretical exploration.

1 INTRODUCTION

Scaling laws are an important tool in the era of large language models. Their primary goal is to predict how model performance changes as the model size increases. (Hestness et al., 2017; Kaplan et al., 2020; Rosenfeld et al., 2020; Henighan et al., 2020) The key challenge of scaling law studies is identifying reliable performance metrics that can be accurately predicted. (Hoffmann et al., 2022; Bergsma et al., 2025) To date, cross-entropy has proven to be the most reliable metric for this purpose, and consequently it has become the dominant choice in scaling law research. Cross-entropy-based scaling laws not only guide the training of large language models, but also provide insights about model mechanisms and artificial intelligence theories. (Kaplan et al., 2020; Hoffmann et al., 2022; Henighan et al., 2020) Cross-entropy has even been adopted beyond language modeling to search for scaling laws in new domains, such as multimodal learning and information retrieval. (Shukor et al., 2025; Edwards et al., 2024; Aghajanyan et al., 2023; Lourie et al., 2025; Fang et al., 2024; Shukor et al., 2025)

However, focusing solely on cross-entropy as the metric provides an incomplete picture of a model's scaling behavior. This limitation arises because cross-entropy primarily measures the absolute probability assigned to the correct answer, while ignoring the relative ordering of predictions. (Chung et al., 2022; Xu et al., 2024a;b) In fact, absolute-based and relative-based perspectives capture two distinct aspects of model performance, and neither can substitute for the other. As illustrated in Figure 1, a model assigns a probability of 0.28 to the correct token. Yet the rank of this token can be different. In one case, two incorrect candidates may still outrank the correct one; in another, all incorrect candidates may receive lower scores, placing the correct token at the top. Thus, the same probability score may correspond to different rankings. Consequently, cross-entropy–based scaling laws fail to capture how the relative position of the correct answer changes with model size. This shortcoming is particularly severe because relative ordering plays a central role in practical applications of language models, such as greedy decoding and top-k sampling. (Noarov et al., 2025;
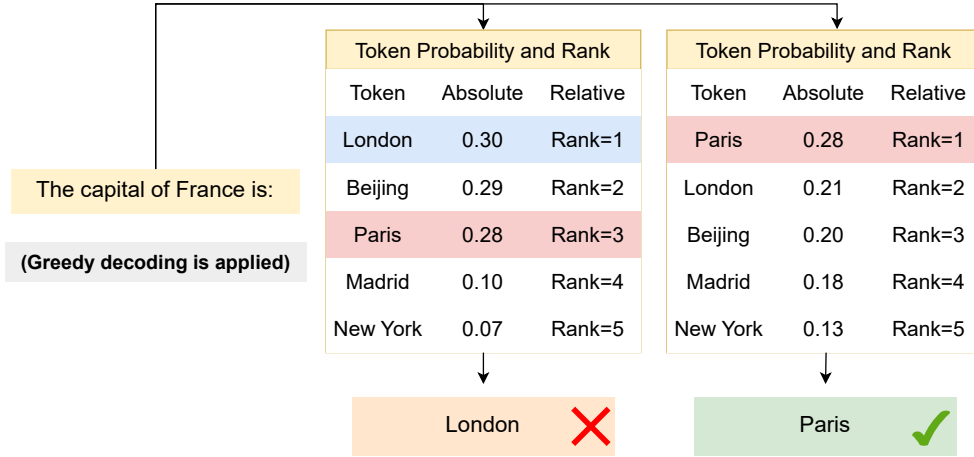
Figure 1: Illustration of the limitation of absolute-based metrics in evaluating generation performance. **(Left)** The ground-truth token has a high absolute probability but is ranked below competitors, leading greedy decoding to fail. **(Right)** With the same absolute probability yet a higher rank, the ground-truth token is correctly chosen. This shows that absolute-based metrics cannot capture the crucial relative ranking among tokens.

Lai et al., 2024; Bruch, 2019; Tang et al., 2024) The inability to account for such order-sensitive sampling strategies highlights a significant gap between cross-entropy metric and the performance observed in real-world usage.

To address this limitation[1], we introduce a new metric, Relative-Based Probability (RBP), to capture a model's ability to rank the correct token among the top candidates. Given a parameter k, RBP_k measures the probability that the correct token appears within model's top-k predictions. The computation of RBP_k proceeds as follows: First, we select a corpus. Then, for each target token in the corpus, we check whether it is included among the top-k tokens predicted by the model. Finally, we calculate the proportion of such cases over the entire dataset, which yields the RBP value. Unlike cross-entropy that is concerned with the absolute prediction value, RBP focuses on the relative ordering of predictions. As a result, RBP_k provides a complementary perspective on model performance.

Next, we examine how RBP_k changes as the model size increases. Our results reveal a clear scaling law when k is much smaller than the vocabulary size. Specifically, we find that

$$-\log\left(\text{RBP}_k\right) \;\propto\; S^{-\alpha}, \quad k \ll \text{Vocab Size}, \tag{1}$$

where S denotes the number of non-embedding model parameters and α is a positive constant. We refer to this relationship as the Relative-Based Scaling Law. To investigate this relationship, we conduct experiments across four datasets. We use 4 model families, covering a total of 24 models whose sizes span over 5 orders of magnitude. The results show that when k lies in the range of roughly 1 to 100, the relationship between $-\log(\text{RBP}_k)$ and model size can be fitted very well by a power-law. The fitted power-law achieves an R^2 of approximately 0.99, which is comparable to the R^2 values obtained from cross-entropy–based scaling laws. Relative-Based Scaling Law deepens our understanding of scaling behavior: enlarging model size fundamentally reshapes the ranking of tokens. As models grow, a greater proportion of correct tokens are placed among the top predictions, and this improvement follows a power-law trend. Such an understanding goes beyond what cross-entropy–based scaling laws alone can capture.

We believe Relative-Based Scaling Law holds important implications for the development of large language models as well as for advancing the theoretical foundations of artificial intelligence. In this work, we illustrate its utility through two concrete examples. (1) Relative-Based Scaling Law

---

[1]Code is available at `https://anonymous.4open.science/r/relative-based-scaling-law-7F3E`.

enhances our understanding of emergence phenomena: Prior studies have attempted to explain emergence through cross-entropy–based scaling laws, but such explanations fail to generalize to decoding strategies like greedy sampling or top-k sampling, where relative ordering is crucial. (Wei et al., 2022; McKenzie et al., 2023; Schaeffer et al., 2023; Lu et al., 2023; Krakauer et al., 2025) By contrast, the Relative-Based Scaling Law naturally accounts for these scenarios and resolves this issue. (2) Relative-Based Scaling Law opens new directions for theoretical research: Although RBP and cross-entropy capture complementary aspects of model performance, we find that their scaling behaviors are surprisingly similar. They share not only identical mathematical form but also very close fitted exponents. This confusing coincidence suggests the existence of a more fundamental theory of intelligence that can unify both scaling laws. At present, no established theory can fully explain this phenomenon. In this paper, we put forward a conjecture as a step toward such unification.

In summary, Relative-Based Scaling Law provides a relative-ordering perspective on scaling behavior. It complements the cross-entropy view and contributes to a more complete picture of scaling in neural language models. Together, the two laws shall offer both practical guidance for scaling up language models and theoretical insight into the fundamental theories of artificial intelligence.

2 RELATED WORK

Scaling laws for neural language models, often based on absolute metrics like cross-entropy loss, provide a framework for understanding performance evolution with model size, dataset size, and compute budget (Kaplan et al., 2020; Henighan et al., 2020; Hoffmann et al., 2022; He et al., 2024; Edwards et al., 2024; Aghajanyan et al., 2023; Lourie et al., 2025).

Absolute-based metrics, however, do not capture the relative ranking of the ground-truth token among candidates, which can misalign metric improvements with generative performance (Tang et al., 2024; Chatzi et al., 2024; Freitag & Al-Onaizan, 2017; Acosta et al., 2024; Song et al., 2024; Prabhu, 2024). To address this, relative-based metrics have been proposed, improving rank-sensitive accuracy in both small- and large-scale models (Petersen et al., 2022; Xiao et al.).

Despite these advances, relative-based evaluation has not been systematically studied under scaling laws for neural language models. Our work fills this gap by introducing a relative-based metric and empirically analyzing its scaling behavior across architectures and datasets.

3 RELATIVE-BASED METRIC

In this section, we introduce our metric for evaluating model performance. Prior work has commonly relied on an *absolute-based* approach, which measures the probability assigned to the ground-truth token. In contrast, we propose a new *relative-based* metric, RBP_k, which evaluates whether the ground-truth token appears within the model's top-k predictions. This provides a perspective that is complementary to cross-entropy and is more directly aligned with practical decoding and sampling strategies.
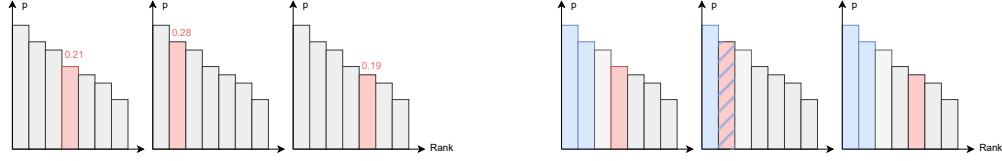
3.1 PRIOR ABSOLUTE-BASED METRIC

In previous studies, model performance has typically been evaluated using cross-entropy loss ($\mathcal{L}_{\text{CE}}$), which exhibits clear scaling behavior across model sizes. Let t be the ground-truth token, and $\text{p}_{\text{A}}(t)$ be the probability score output by the model. Cross-entropy equals

$$\mathcal{L}_{\text{CE}} = \mathbb{E}[-\log \text{p}_{\text{A}}(t)], \tag{2}$$

However, $\text{p}(t)$ alone does not reflect the *relative ordering* of the ground-truth token among all candidate tokens. As shown in Figure 1, the ground-truth token may be assigned relatively high probability but not ranked as the top-1 candidate. If greedy-decoding or top-k sampling are used during inference, such a probability score cannot reflect models' real-world performance.

3.2 OUR RELATIVE-BASED METRIC

To address this problem, we introduce a *relative-based* metric, namely Relative-based Probability. It is computed based on a hyper-parameter k. Therefore, we denote this metric as RBP_k. RBP_k

(a) Calculation of the absolute-based metric $\mathrm{p_A}$. Ground-truth is highlighted in red. The metric $\mathrm{p_A}$ is computed as the average of the probabilities assigned to the ground-truth across different instances, $\mathrm{p_A} = (0.21 + 0.28 + 0.19)/3 \approx 0.227$.

(b) Calculation of the relative-based metric RBP_k with $k = 2$. The ground truth is highlighted in red. The red rectangle with blue hatching indicates cases where the ground truth appears within the top-2 ranked predictions. Since the ground truth is included in the top-2 predictions for only one out of the three instances, we obtain $\mathrm{RBP}_2 = 1/3 \approx 0.333$.

Figure 2: Illustration of two distinct metric calculation methods. Figure 2a shows the absolute-based metric $\mathrm{p_A}$, which directly quantifies the probability assigned to the ground-truth. Figure 2b shows the relative-based metric RBP_k, which measures how often the ground-truth falls within the top-k predictions. These figures highlight how the two metrics capture different aspects of model behavior.

measures the probability that the ground-truth token appears within the model's top-k predictions. Compared to cross-entropy, it is an independent metric that complements the relative-ordering perspective to model performance. For example, if RBP_1 is 30%, it means that greedy sampling works for 30% cases. Yet this cannot be deduced from the cross-entropy value.

RBP_k is formally defined as follows. Let $\mathcal{V}$ denote the vocabulary and t be the ground-truth token. We use $\mathrm{p}(v)$ be the score of token v. Then, the rank of ground-truth token R equals:

$$R = \sum_{v \in \mathcal{V}} \mathbf{1}\{\mathrm{p}(\mathrm{v}) \geq \mathrm{p}(\mathrm{t})\}. \tag{3}$$

R is a random variable. We define RBP_k as its cumulative distribution function. Formally, it equals:

$$\mathrm{RBP}_k = \Pr(R \leq k), \tag{4}$$

We can see that RBP_k is the probability that the ground-truth token is within the top-k scored tokens.

In practice, we compute RBP_k as follows. First, we select a corpus. Then, for each token in the corpus, we record 1 if the ground-truth token falls in the top-k predictions and 0 otherwise. The average of these values yields an empirical estimation of RBP_k.

In the next section, we will show that RBP_k also exhibits a power-law scaling behavior with model size when k is small. Therefore it is able to precisely capture performance improvements from a relative-ordering perspective.

4 RELATIVE-BASED SCALING LAW

Based on Relative-based Probability, this section establishes the *Relative-based Scaling Law*. It characterizes how RBP_k evolves with model size. Our finding is that for small k values ($k \ll |\mathcal{V}|$, with $|\mathcal{V}|$ denoting the vocabulary size), RBP_k follows a precise power-law relationship with model size:

$$-\log \mathrm{RBP}_k \; \propto \; S^{-\alpha} \quad \left(k \ll |\mathcal{V}|\right) \tag{5}$$

where S denotes the model size and $\alpha > 0$ is the scaling exponent. We refer to this relationship as the *relative-based scaling law*.

Now we start to investigate the relationship between RBP_k and model size. Since RBP_k depends on the choice of threshold k, experiments are conducted in the following three regimes:

1. $k = 1$ regime: This regime is corresponding to the hardest setting where only the top-ranked token is considered. In this case, RBP_k closely links to greedy decoding strategy.
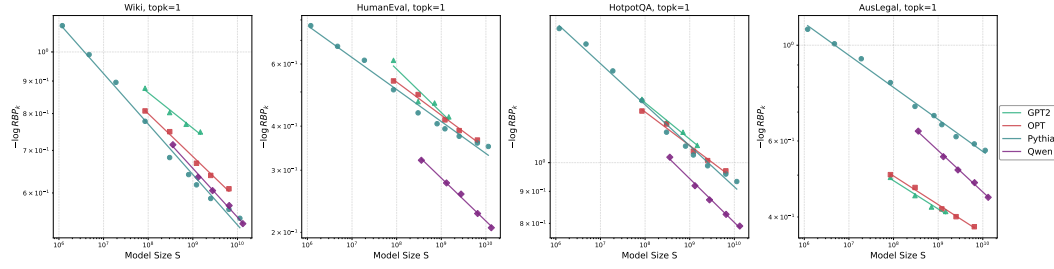
Figure 3: Relative-based scaling laws when $k = 1$. Across all model series and all datasets, RBP_1 exhibits precise power-law relationship with model sizes.

2. Moderate-k regime: In this regime, k is larger than 1 but far smaller than the vocabulary size. In our experiments, we set $1 < k \leq 100$. This setup closely links to top-k sampling strategies.

3. Large-k regime: In this regime, k approaches to the vocabulary size. In our experiments, we set k to 20,000 or 30,000.

These three regimes provide a comprehensive view of how RBP_k scales with model size. We present the experimental results respectively in the following three subsections.

4.1 $k = 1$ REGIME

In the case of $k = 1$, the relative-based metric exhibits a remarkably precise power-law trend across all model families and datasets. Figure 4 illustrates RBP_1 versus model size on four datasets. For data points in the same model family, they form a straight line on the log-log plot. This indicates that RBP_1 precisely follows a power-law relationship with model size.

The result marks the discovery of a novel scaling law. It means that scaling up models increases the probability that the correct token is ranked highest. And the increasing rate follows a power-law rate. This perspective cannot be deduced from prior cross-entropy scaling law. It demonstrates that scaling up models fundamentally changes the relative-ordering and results in more correct tokens ranked at the top. Since RBP_1 closely links to the greedy decoding strategies, it reflects that greedy decoding performance will improve if model is scaled up.

4.2 MODERATE-k REGIME: $1 < k \leq 100$

Moreover, we also observe that the power-law relationship still holds for small k values. We set k to 10, 50, and 100. The results are illustrated in Figure 4. We can see that for all cases, the data points lie on a straight line in the log-log plot. The fitted curves achieve $R^2 \geq 0.97$ across all datasets, indicating that the power-law relationship remains highly robust in this regime.

The result means that Relative-based Scaling Law can generalize to small k values. When model is scaled up, more and more correct tokens are ranked within the top-k predictions. And the rate of improvement follows a power-law manner. Since RBP_k measures the probability that the ground-truth token appears within the top-k candidates, it closely links to top-k sampling strategy. It indicates that top-k sampling result will be improved when model is scaled up, which is another perspective that cannot be deduced from cross-entropy scaling law alone.

4.3 LARGE-k REGIME: $k \to |\mathcal{V}|$

As k approaches the vocabulary size $|\mathcal{V}|$, the scaling behavior of RBP_k breaks down. We set k to $20,000$ and $30,000$, and illustrate the results in Figure 5[2]. We can see that the data points scatter

---

[2]For Pythia, OPT, and GPT-2 models, vocabulary sizes are 50,256, 50,272, and 50,257, respectively, such that the chosen k values nearly cover the entire vocabulary; for Qwen models, vocabulary sizes are 151,936 for models up to 7B and 152,064 for models above 7B. We don't report Qwen series in this regime because it doesn't apply to the condition $k \to |\mathcal{V}|$.
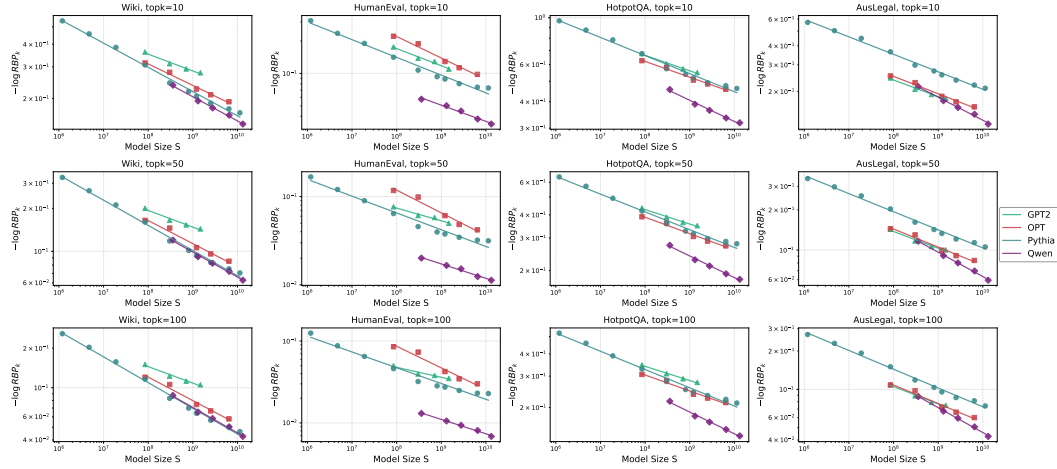
Figure 4: Scaling laws when $1 < k \ll |\mathcal{V}|$. The relative-based metric RBP_k maintains strong power-law scaling behavior, with consistently high R^2 values across datasets and model series. This indicates that the metric reliably captures performance improvements under top-k sampling strategies commonly used in practice.
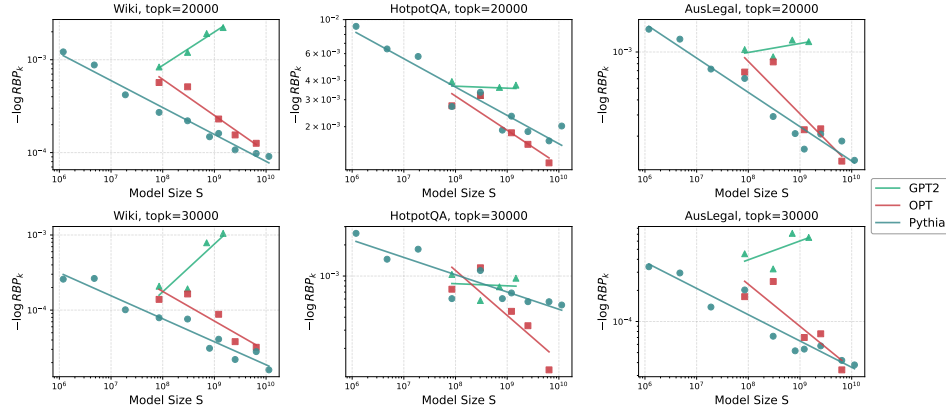

Figure 5: Scaling laws when $k \to |\mathcal{V}|$. In this regime, the power-law behavior deteriorates substantially, with large scatter and inconsistent slopes across models and datasets. This indicates that RBP_k becomes less informative about scaling when the threshold approaches the vocabulary size.

significantly. For example, RBP_k for GPT2 even increases rather than decreases. The case for $k = 30,000$ becomes much more random than the case for $k = 20,000$. Therefore, we believe that Relative-based Scaling Law only holds for k that is small.

The reason for this behaviour remains to be investigated. We conjecture that that reason might lies in random noise. When k nearly covers the entire vocabulary, most ground-truth tokens are included in the top-k set, regardless of model size. In this case, some noisy data may dominate RBP_k, which breaks down its power-law scaling pattern. The noise even makes RBP_k increase when model is scaled up, as shown by the GPT2 results in Figure 5.

4.4 SUMMARY

The results across the three regimes collectively reveal how the choice of threshold k governs the extent to which RBP_k reflects power-law scaling with model size. In Figure 6, we summarize how the threshold k affects the fitting quality and the scaling exponent.

According to the left figure, we can see that the scaling law holds robustly in the small and moderate k regimes, where $k \ll |\mathcal{V}|$. The fitted quality only starts to drop after k grows close to the vocabulary

size. We can see that R^2 values remain high (above 0.9) when $k < 1000$. Therefore, Relative-based Scaling Law is a robust law for a wide range of k values.

According to the right figure, the scaling exponent increases as k increases. A higher exponent means that the performance increases faster when the model is scaled up. Therefore, the result indicates that optimizing RBP_{100} shall be easy, since the exponent is close to 0.2. Yet it is hard to optimize RBP_1, as the exponent is less than 0.1. Such an understanding is a unique contribution of Relative-based Scaling Law.
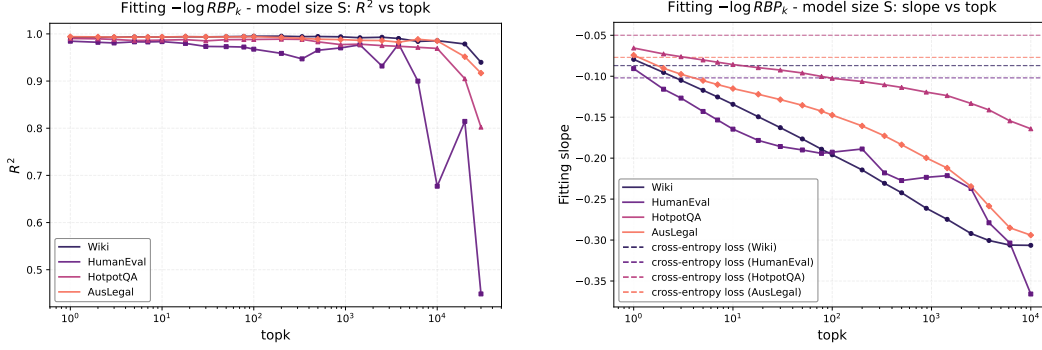
(a) Coefficient of determination (R^2) as a function of k. The scaling law holds robustly for $k < 1000$, where R^2 is very close to 1. It starts to drop when k is too large.

(b) Exponent of the fitted power-law as a function of k. Exponent for $k = 1$ is close to those of cross-entropy-based scaling laws. The exponent value decreases gradually as k increases, indicating that RBP_k is easier to optimize for bigger k.

Figure 6: Scaling fitting results of RBP_k as a function of k. Figure 6a shows the power-law fitting quality. Figure 6b shows how scaling exponent evolves with k.

5 APPLICATION OF RELATIVE-BASED SCALING LAW

In this section, we demonstrate two applications of Relative-based Scaling Law. The first is to explain emergence phenomena, and the other is to explore fundamental principles that drive both prior cross-entropy-based and our Relative-based Scaling Laws.

5.1 EXPLAINING EMERGENCE

A significant challenge for scaling laws is to explain complex macroscopic phenomena, such as the "emergence" of new capabilities in large language models. Emergence refers to the sharp, non-linear jump in performance on a specific task once model size surpasses a threshold. Although (Schaeffer et al., 2023) explain this phenomenon with cross-entropy-based scaling law, this scaling law cannot generalize to greedy or top-k sampling and thus cannot explain emergence phenomenon in these scenarios. We show that the relative-based scaling law can complement this drawback and provide a quantitative explanation.

We define task success as the model correctly predicting N consecutive ground-truth tokens within its top-k candidates. Assuming independence and stationarity across positions, the success probability is:

$$p_{\mathrm{N,k}} = (\mathrm{RBP_k})^{\mathrm{N}}. \tag{6}$$

When $\mathrm{RBP}_k \propto S^{-\alpha}$, the sequence-level relation becomes:

$$-\log p_{\mathrm{N,k}} = -\mathrm{N}\log(\mathrm{RBP_k}) \propto \mathrm{N} \cdot \mathrm{S}^{-\alpha}. \tag{7}$$

Thus, sequence-level log-error scales smoothly with model size S, linearly amplified by N. Figure 7 confirms this prediction empirically.

This explains the apparent "emergence." Since

$$p_{\mathrm{N,k}} \approx \exp(-\mathrm{C} \cdot \mathrm{N} \cdot \mathrm{S}^{-\alpha}), \tag{8}$$
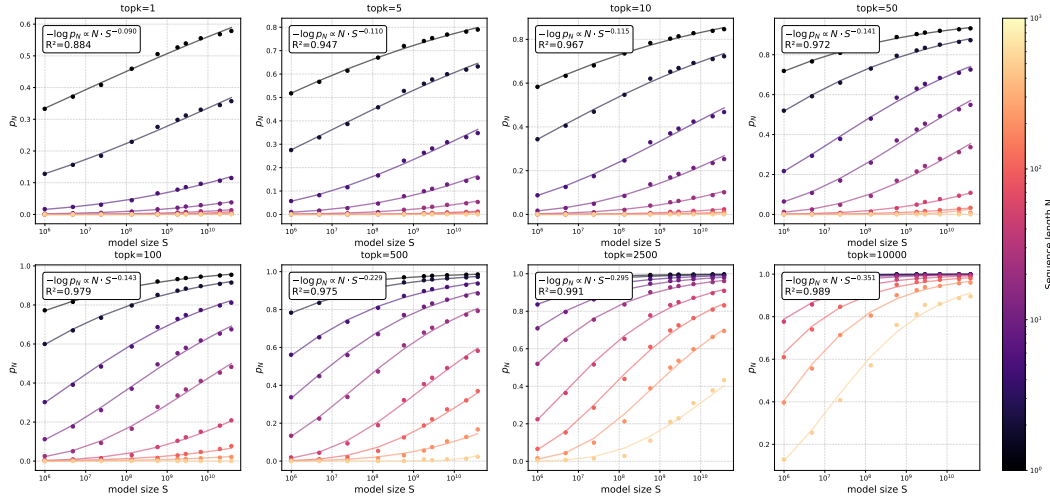
Figure 7: Illustration of the emergence phenomena, $-\log p_{N,k} \propto N \cdot S^{-\alpha}$. Each subplot corresponds to a fixed k, with colored lines for different N. The fitted power-law curves align closely with empirical data, confirming the law across a wide range of conditions. Crucially, the 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 $p_{N,k}$ close to zero.
- As S grows, the exponent approaches zero, and $p_{N,k}$ 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 $-\log p_1$ (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:

$$P(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0, \tag{9}$$
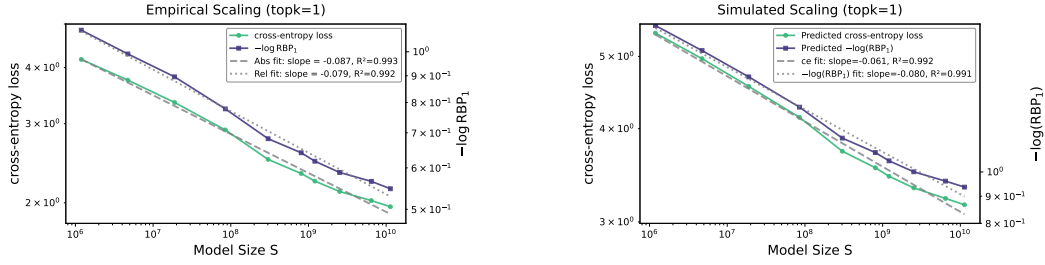
(a) Empirical results: CE and $-\log \mathrm{RBP}_k$ exhibit similar scaling forms and slopes.

(b) Simulated results: CE and $-\log \mathrm{RBP}_k$ show similar scaling patterns.
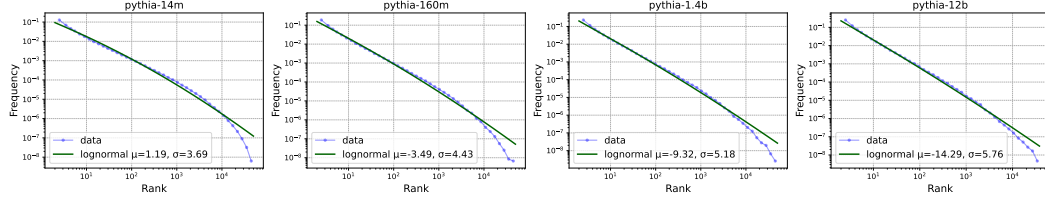
(c) Hypothesis: Ground-truth ranking frequency follows the lognormal distribution.

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

training and scaling. *arXiv preprint arXiv:2304.01373*, 2023. Pythia suite (16 models, controlled training data/checkpoints) — code and checkpoints available from EleutherAI / authors' repo.

Sebastian Bruch. An alternative cross entropy loss for learning-to-rank. *arXiv preprint arXiv:1911.09798*, 2019.

Umar Butler and contributors. Open australian legal corpus (oalc). `https://huggingface.co/datasets/umarbutler/open-australian-legal-corpus` (dataset) and https://github.com/isaacus-dev/open-australian-legal-corpus-creator (creator repo). Multijurisdictional open corpus of Australian legislative and judicial documents; access / code repository on GitHub and Hugging Face. Accessed 2025-09-25.

Lixin Cai, Zhangchi Zhu, Jianyong Wang, Wei Zhang, and Cong Xu. Understanding the role of cross-entropy loss in fairly evaluating large language models for recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 2024. UNVERIFIED: could not find a RecSys paper with this exact author list/title. The arXiv/RecSys work on this topic is by Cong Xu et al. (arXiv:2402.06216). Please check.

Ivi Chatzi, Eleni Straitouri, Suhas Thejaswi, and Manuel Gomez-Rodriguez. Prediction-powered ranking of large language models. *arXiv preprint arXiv:2402.17826*, 2024.

C. Y. Chung, T. Kudo, and N. Kanda. How well does perplexity measure the predictive performance of language models? *arXiv preprint arXiv:2208.10091*, 2022. UNVERIFIED: arXiv id 2208.10091 belongs to a different paper (see notes). Please check author list / eprint.

Thomas D. P. Edwards, James Alvey, Justin Alsing, Nam H. Nguyen, and Benjamin D. Wandelt. Scaling-laws for large time-series models. *arXiv preprint arXiv:2405.13867*, 2024.

Yan Fang, Jingtao Zhan, Qingyao Ai, Jiaxin Mao, Weihang Su, Jia Chen, and Yiqun Liu. Scaling laws for dense retrieval. *arXiv preprint arXiv:2403.18684*, 2024.

Allen Institute for AI (AllenAI). allenai/c4: A processed version of the colossal clean crawled corpus. `https://huggingface.co/datasets/allenai/c4`, 2022. Derived from Raffel et al. (2020).

Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In *Proceedings of the Workshop on Neural Machine Translation and Generation*, pp. 15–24. Association for Computational Linguistics, 2017.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Yifei He, Alon Benhaim, Barun Patra, Praneetha Vaddamanu, Sanchit Ahuja, Parul Chopra, Vishrav Chaudhary, Han Zhao, and Xia Song. Scaling laws for multilingual language models. *arXiv preprint arXiv:2410.12883*, 2024.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory F. Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*, 2020. doi: 10.48550/arXiv.2001.08361.

David C. Krakauer, John W. Krakauer, and Melanie Mitchell. Large language models and emergence: A complex systems perspective. *arXiv preprint arXiv:2506.11135*, 2025.

Wei Lai, Zhiquan Tan, Chenghai Li, Jindong Wang, and Weiran Huang. Diff-erank: A novel rank-based metric for evaluating large language models. *arXiv preprint arXiv:2401.17139*, 2024.

Nicholas Lourie, Michael Y. Hu, and Kyunghyun Cho. Scaling laws are unreliable for downstream tasks: A reality check. *arXiv preprint arXiv:2507.00885*, 2025.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning? *arXiv preprint arXiv:2309.01809*, 2023. Accepted / presented at ACL 2024 (long paper).

Huan Ma, Jingdong Chen, Joey Tianyi Zhou, Guangyu Wang, and Changqing Zhang. Estimating llm uncertainty with logits. *arXiv preprint arXiv:2502.00290*, 2025.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, and et al. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*, 2023.

monology / Hugging Face. pile-uncopyrighted: A cleaned version of the pile with copyrighted content removed. `https://huggingface.co/datasets/monology/pile-uncopyrighted`, 2024. Derived from Gao et al. (2020).

Georgy Noarov, Soham Mallick, Tao Wang, Sunay Joshi, Yan Sun, Yangxinyu Xie, Mengxin Yu, and Edgar Dobriban. Foundations of top-k decoding for language models. *arXiv preprint arXiv:2505.19371*, 2025.

OpenAI. HumanEval: Code for "evaluating large language models trained on code" (humaneval dataset). `https://github.com/openai/human-eval`, 2021. GitHub repository / evaluation harness for HumanEval (164 Python problems).

Felix Petersen, Hilde Kühne, Christian Borgelt, and Oliver Deussen. Differentiable top-k classification learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*, pp. 1765–1775, 2022. Introduces differentiable top-k loss functions for classification.

Sumanth Prabhu. Pedal: Enhancing greedy decoding with large language models using diverse exemplars. *arXiv preprint arXiv:2408.08869*, 2024.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, and ... Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. Qwen2.5 model family technical report (multi-author; open-weight offerings and hosted MoE variants).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners (gpt-2). `https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`, 2019. OpenAI technical report / blog post introducing GPT-2 family.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 8748–8760, 2020.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations (ICLR)*, 2020.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Proceedings of the NeurIPS 2023 Workshop / Papers*, 2023. NeurIPS 2023 paper / PDF available in NeurIPS proceedings.

Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. *arXiv preprint arXiv:2504.07951*, 2025.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. *arXiv preprint arXiv:2407.10457*, 2024.

Chenxia Tang, Jianchun Liu, Hongli Xu, and Liusheng Huang. Top-$n\sigma$: Not all logits are you need. *arXiv preprint arXiv:2411.07641*, 2024.

Jason Wei, Yi Tay, Rishi Bommasani, Catherine Jean, Hyung Won Chung, Da Hou, Barret Zoph, Quoc V Le, Sharan Narang, Aakanksha Chowdhery, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Wikimedia Foundation. Wikimedia database dumps. `https://dumps.wikimedia.org/`. Official Wikimedia data dumps (wikipedia/Wikimedia content used widely as training/benchmark data). Accessed 2025-09-25.

Rufeng Xiao, Yuze Ge, Rujun Jiang, and Yifan Yan. A unified framework for rank-based loss minimization. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, pp. —.

Cong Xu, Zhangchi Zhu, Jun Wang, Jianyong Wang, and Wei Zhang. Understanding the role of cross-entropy loss in fairly evaluating large language model-based recommendation. *arXiv preprint arXiv:2402.06216*, 2024a.

Cong Xu, Zhangchi Zhu, Jun Wang, Jianyong Wang, and Wei Zhang. Understanding the role of cross-entropy loss in fairly evaluating large language model-based recommendation. *arXiv preprint arXiv:2402.06216*, 2024b. duplicate/corrected entry (same arXiv:2402.06216). Consolidated authors to match arXiv record.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *arXiv / ACL (HotpotQA dataset)*, 2018. Dataset of 113k Wikipedia question-answer pairs with supporting facts.

Jingtao Zhan, Jiahao Zhao, Jiayu Li, Yiqun Liu, Bo Zhang, Qingyao Ai, Jiaxin Mao, Hongning Wang, Min Zhang, and Shaoping Ma. Evaluating intelligence via trial and error. *arXiv preprint arXiv:2502.18858*, 2025.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, ..., and Luke Zettlemoyer. Opt: Open pretrained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. Meta AI's OPT suite (125M–175B) with released models and logbook.

A EXPERIMENTAL SETUP

Models. We select several representative model families covering a wide range of scales. Specifically, we include the `Pythia` series (Biderman et al., 2023) (14M–12B), the `GPT-2` series (Radford et al., 2019), the `OPT` family released by Meta (Zhang et al., 2022), and the more recent `Qwen2.5` series (Qwen Team et al., 2024) (0.5B–14B). Together these span over four orders of magnitude in parameter count and provide multiple independently trained scaling series.

Datasets. To ensure robustness across domains, we evaluate on four representative benchmarks spanning diverse linguistic and reasoning tasks: `Wikipedia` for open-domain natural text and general language modeling ability (Wikimedia Foundation), `HumanEval` for Python programming tasks focusing on code generation (OpenAI, 2021), `HotpotQA` for multi-hop question answering that requires handling complex queries (Yang et al., 2018), and the `Open Australian Legal Corpus` for long-form legal documents that test domain-specific expertise (Butler & contributors).

Thresholds. We estimate RBP_k across a wide spectrum of threshold values k, spanning over five orders of magnitude, from very small thresholds up to values approaching $|\mathcal{V}|$. Each data point is obtained by averaging over at least $5 \cdot 10^5$ tokens to ensure statistical stability.

B ANALYSIS OF RBP_k AS A FUNCTION OF TOP-k

To obtain a more comprehensive understanding of RBP_k, we analyze its behavior across commonly used ranges of top-k values. As shown in Figure 9, RBP_k exhibits a power-law relationship with respect to k, analogous to the scaling observed with model size S. However, as reported in Table 1, the fitted coefficient with k is substantially faster than that with S.

This observation leads to the following implications:

1. To quickly increase the probability of the ground-truth appearing within the top-k predictions (RBP_k), increasing the parameter k is more effective than enlarging the model size.

2. To maintain a fixed level of RBP_k, the relative changes in k and S can be quantified using the ratio of their corresponding log-log slopes. Let α_k and α_S denote the slopes obtained from the linear regression of RBP_k against $\log k$ and $\log S$, respectively. Then, for small relative changes, the required change in model size $\Delta S/S$ to compensate a change $\Delta k/k$ in k can be estimated as:

$$\frac{\Delta S}{S} \approx -\frac{\alpha_k}{\alpha_S} \cdot \frac{\Delta k}{k}. \tag{10}$$

This relation provides a quantitative guideline for jointly adjusting k and S to achieve a desired RBP_k level.
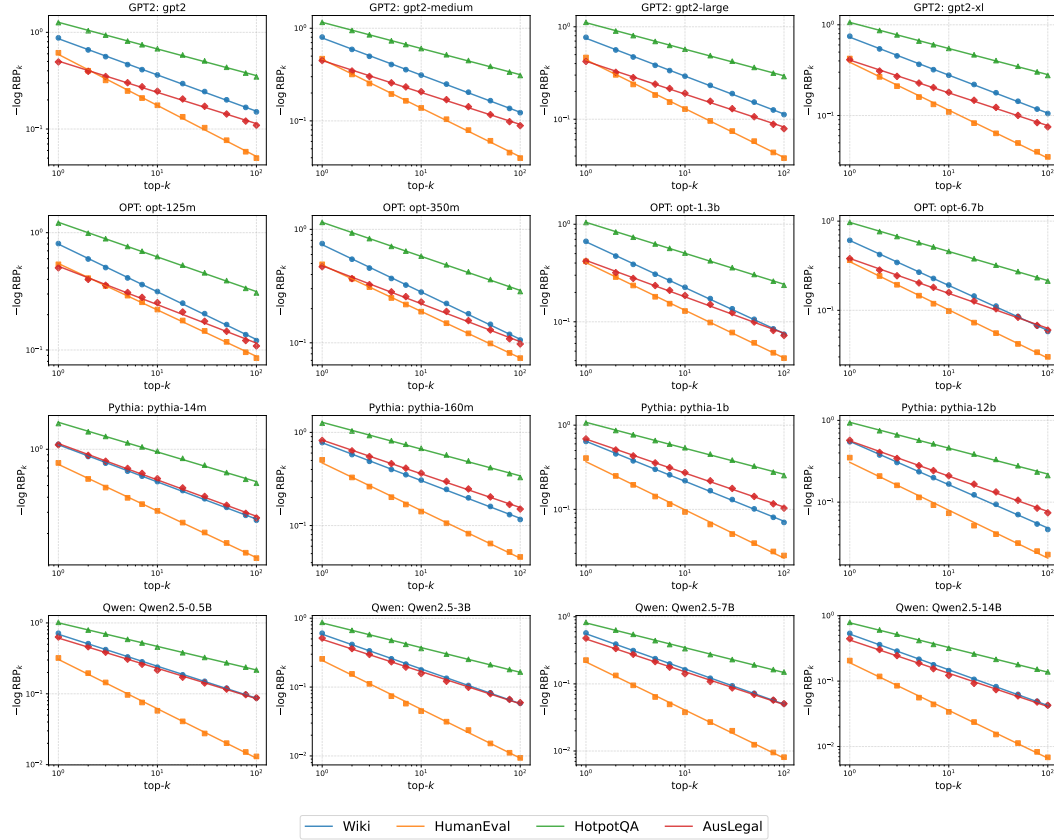

Figure 9: The plots illustrate a clear power-law relationship of RBP_k with increasing top-k values across various language models and datasets. This consistent trend is observed within the common top-k sampling range, particularly for $k \leq 100$.

Series	Model	R^2				Slope			
		Wiki	HEval	HpQA	AusLeg	Wiki	HEval	HpQA	AusLeg
GPT2	gpt2	0.999	0.998	0.999	0.998	-0.375	-0.528	-0.276	-0.323
	medium	1.000	0.999	1.000	0.999	-0.401	-0.526	-0.283	-0.345
	large	1.000	0.999	0.999	0.998	-0.410	-0.529	-0.287	-0.356
	xl	1.000	0.997	0.999	0.999	-0.418	-0.528	-0.288	-0.361
OPT	125m	1.000	0.999	0.999	0.995	-0.407	-0.396	-0.295	-0.328
	350m	1.000	1.000	0.999	0.997	-0.417	-0.410	-0.300	-0.335
	1.3b	1.000	1.000	1.000	0.997	-0.470	-0.489	-0.318	-0.373
	2.7b	1.000	1.000	1.000	0.998	-0.485	-0.524	-0.322	-0.383
	6.7b	1.000	0.999	1.000	0.998	-0.504	-0.541	-0.326	-0.392
Pythia	14m	0.999	0.999	0.998	0.998	-0.308	-0.387	-0.246	-0.302
	31m	0.999	0.999	0.999	0.999	-0.337	-0.435	-0.261	-0.318
	70m	0.999	0.999	0.997	0.998	-0.370	-0.482	-0.278	-0.338
	160m	0.999	0.998	0.998	0.999	-0.406	-0.511	-0.287	-0.362
	410m	0.999	0.998	0.999	0.999	-0.451	-0.559	-0.297	-0.384
	1b	0.999	0.996	0.999	0.999	-0.475	-0.573	-0.309	-0.404
	1.4b	0.999	0.995	0.999	0.999	-0.487	-0.572	-0.309	-0.408
	2.8b	1.000	0.994	0.999	0.999	-0.503	-0.581	-0.310	-0.416
	6.9b	0.999	0.994	0.999	0.999	-0.520	-0.588	-0.313	-0.420
	12b	0.999	0.993	0.998	0.999	-0.530	-0.586	-0.316	-0.432
Qwen	0.5B	0.999	0.999	1.000	0.998	-0.451	-0.697	-0.333	-0.428
	1.5B	1.000	0.999	0.999	0.998	-0.490	-0.701	-0.352	-0.455
	3B	0.999	0.998	0.999	0.998	-0.504	-0.710	-0.359	-0.468
	7B	1.000	0.998	0.999	0.998	-0.525	-0.716	-0.370	-0.489
	14B	1.000	0.998	0.999	0.997	-0.546	-0.729	-0.377	-0.508

Table 1: Fitted R^2 and slopes of $-\log \mathrm{RBP}_k$ as a function of top-k for different models and datasets. The fits are highly accurate (all $R^2 \geq 0.99$), and the slopes with respect to k are considerably steeper than those with respect to model size S.

C ADDITIONAL RESULTS ON SCALING BEHAVIOR

As shown in Table 2, we present detailed numerical results for the scaling behavior of different model series across multiple datasets, reported in terms of the fitted R^2 and slopes as threshold k increases. To further strengthen the generality of our proposed scaling law, we additionally include results on two large-scale pretraining datasets: `allenai/C4`(Raffel et al., 2020; for AI , AllenAI) and `monology/pile-uncopyrighted`(Gao et al., 2020; monology / Hugging Face, 2024). We observe clear scaling patterns on all these corpora, which highlights the robustness of the scaling phenomenon. Finally, we note that in most cases the fitted R^2 and slope for cross-entropy are very close to those obtained for $-\log \mathrm{RBP}_1$, which motivates our discussion in 5.2.

Dataset	Top-k	R^2				Slope			
		GPT2	OPT	Pythia	Qwen	GPT2	OPT	Pythia	Qwen
AusLegal	CE	0.961	0.989	0.993	0.996	-0.071	-0.081	-0.077	-0.103
	1	0.980	0.994	0.994	0.997	-0.066	-0.066	-0.074	-0.096
	10	0.984	0.987	0.994	0.994	-0.109	-0.115	-0.115	-0.152
	100	0.979	0.976	0.994	0.997	-0.135	-0.146	-0.147	-0.196
	500	0.982	0.970	0.989	0.996	-0.150	-0.181	-0.184	-0.218
	2500	0.905	0.945	0.986	0.971	-0.184	-0.219	-0.235	-0.260
	10000	0.822	0.924	0.985	0.901	-0.077	-0.340	-0.294	-0.230
C4	CE	0.987	0.986	0.991	0.992	-0.052	-0.061	-0.066	-0.068
	1	0.985	0.993	0.991	0.991	-0.049	-0.055	-0.063	-0.062
	10	0.985	0.987	0.992	0.992	-0.087	-0.100	-0.108	-0.110
	100	0.985	0.977	0.992	0.993	-0.121	-0.148	-0.157	-0.163
	500	0.989	0.962	0.992	0.994	-0.145	-0.185	-0.193	-0.203
	2500	0.985	0.946	0.992	0.986	-0.156	-0.222	-0.230	-0.228
	10000	0.974	0.911	0.988	0.953	-0.162	-0.260	-0.258	-0.237
Github	CE	0.091	0.987	0.995	0.998	-0.027	-0.108	-0.146	-0.133
	1	0.959	0.980	0.995	0.998	-0.076	-0.096	-0.134	-0.118
	10	0.885	0.990	0.996	0.999	-0.084	-0.140	-0.189	-0.185
	100	0.770	0.975	0.995	0.998	-0.072	-0.175	-0.239	-0.242
	500	0.671	0.936	0.995	0.998	-0.062	-0.199	-0.277	-0.269
	2500	0.083	0.895	0.994	0.998	-0.012	-0.242	-0.323	-0.326
	10000	0.045	0.862	0.980	0.976	-0.002	-0.244	-0.363	-0.338
HotpotQA	CE	0.997	0.983	0.992	0.996	-0.044	-0.046	-0.050	-0.063
	1	0.993	0.987	0.990	0.995	-0.058	-0.054	-0.066	-0.070
	10	0.987	0.990	0.987	0.991	-0.073	-0.075	-0.086	-0.100
	100	0.995	0.981	0.988	0.994	-0.080	-0.089	-0.103	-0.122
	500	0.988	0.954	0.983	0.992	-0.073	-0.096	-0.114	-0.145
	2500	0.950	0.954	0.975	0.992	-0.067	-0.113	-0.133	-0.166
	10000	0.989	0.868	0.969	0.973	-0.073	-0.109	-0.164	-0.150
HumanEval	CE	0.905	0.984	0.983	0.988	-0.121	-0.127	-0.102	-0.120
	1	0.895	0.989	0.985	0.995	-0.123	-0.094	-0.091	-0.122
	10	0.979	0.984	0.984	0.984	-0.159	-0.199	-0.165	-0.150
	100	0.947	0.966	0.968	0.994	-0.122	-0.264	-0.193	-0.176
	500	0.930	0.859	0.965	0.938	-0.145	-0.253	-0.227	-0.221
	2500	0.947	0.857	0.932	0.846	-0.202	-0.153	-0.237	-0.188
	10000	0.630	0.951	0.677	1.000	-0.217	-0.394	-0.366	0.000
Wiki	CE	0.965	0.988	0.993	0.993	-0.058	-0.077	-0.087	-0.085
	1	0.983	0.994	0.992	0.994	-0.056	-0.067	-0.079	-0.077
	10	0.982	0.991	0.993	0.995	-0.093	-0.119	-0.134	-0.134
	100	0.979	0.981	0.995	0.995	-0.126	-0.179	-0.196	-0.193
	500	0.950	0.968	0.994	0.993	-0.133	-0.228	-0.242	-0.234
	2500	0.826	0.942	0.993	0.985	-0.084	-0.273	-0.292	-0.266
	10000	0.915	0.917	0.986	0.942	0.090	-0.331	-0.306	-0.325

Table 2: Fitted R^2 and slopes of $-\log \mathrm{RBP}_k$ as a function of model size S for different datasets and top-k thresholds.

16

D   CONNECTION BETWEEN ABSOLUTE-BASED AND RELATIVE-BASED SCALING LAW

We assume that the token probabilities p_k follow a lognormal distribution:

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

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

$$a = -\frac{\mu}{\sigma}, \qquad P = \int_{1}^{\infty} \psi(x)\,dx = \Phi\left(\frac{\mu}{\sigma}\right). \tag{14}$$

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

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

Starting from the definition of cross-entropy,

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

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

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

Substituting (I_1, I_2) gives

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

with

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

Simplifying yields

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

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

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

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 $-\log(p_1)$, which explains why CE and $-\log(p_1)$ behave similarly as the model scales.