Distributional Scaling Laws for Emergent Capabilities

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we explore the nature of sudden breakthroughs in language model performance at scale, which stands in contrast to smooth improvements governed by scaling laws. While advocates of "emergence" argue that abrupt performance gains arise from acquiring new capabilities at specific scales, recent work has suggested that these are illusions caused by thresholding effects. We propose an alternative explanation: that breakthroughs are driven by random variation, particularly multimodal performance distributions across random seeds. Using a length generalization task as a case study, we show that different random seeds lead to both highly linear or emergent behavior. We further demonstrate that the probability of a model acquiring a breakthrough capability increases continuously with scale, despite apparent discontinuities in performance. Additionally, we find that scaling models in width versus depth has distinct effects: depth impacts the likelihood of sampling from a successful distribution, while width improves the average performance of successful models. These insights suggest a need to consider the role of random variation in scaling and emergent capabilities in LMs.

1 Introduction

On most benchmarks, language model (LM) performance is determined by a scaling law [12, 8] that responds smoothly to parameter size and overall training compute. There are, however, a number of celebrated exceptions in which performance abruptly improves on specific benchmarks [15].

Sudden breakthroughs at scale provide the backdrop of one of the most heated debates in modern machine learning. On one side, advocates of *emergence* claim that performance abruptly improves at particular scales because those scales allow the LM to acquire specific concepts that permit out-of-distribution generalization [17]. On the other side, skeptics argue that these sudden improvements are a *mirage* driven by thresholding effects and alleviated by more appropriate continuous metrics—though a few **breakthrough capabilities** remain stubbornly emergent [13]. Here, we argue that such discontinuities are driven by predictable changes in the *probability* of a breakthrough at each scale.

We posit that a breakthrough capability is distinguished not by direct responses to scale, but by *multimodal* random variation. Such effects are currently undetected because scaling laws are generally measured across independent training runs, and resources are rarely committed to correct for random variation by reusing the same hyperparameter settings with different random seeds. Although random variation may be insignificant when model performance is measured in-distribution [6], previous work already suggests that settings which require compositional reasoning may be prone to having performance vary widely across random seeds [19, 20] which can also be seen at larger scales [10]. Since training numerous seeds becomes prohibitively expensive at large scales, we study the ability for models to *length generalize* on compositional tasks. Even in these synthetic settings, previous works train on only at most tens of seeds and report summary statistics of these runs. In this work, we seek to characterize the resulting distribution— particularly demonstrating that multimodal variation is present for the studied task. Our contributions can be summarized as follows:

Submitted to Workshop on Scientific Methods for Understanding Deep Learning, NeurIPS 2024.

- **Breakthrough scaling curves can result from bimodal performance distributions.** Using length generalization as a case study, we demonstrate that different random seeds can exhibit either highly linear or highly emergent scale behavior. We connect these differences to the bimodal distribution of this compositional skill across random seeds, a property that materializes at many parameter scales. At these scales, "emergence" is a stochastic property.
- **Although a given scale curve can exhibit discontinuity (emergence), the *probability* of a model learning a skill actually responds *continuously* with respect to scale.** Modeling the bimodal distribution as a mix of failure and success distributions, we illustrate that improvements with scale can come from changes in the probability of success or in the average performance of a successful model.
- **The random variation distribution changes differently when scaling up models with respect to width vs. depth.** Although the scaling laws literature generally treats parameter scale as a single property, we find that scaling in depth changes the probability of sampling from the successful distribution—thereby changing the probability of emergence—whereas scaling in width only changes the average performance of the successful distribution. With respect to emergent capabilities, these are significant differences.

2 Methodology

Breakthrough capabilities often require compositional reasoning [15, 9, 1]; one such category is the propensity for length generalization [18]. Below we outline our experimental setup, with further details given in Appendix A.

Architecture: In all of our experiments, we train decoder-only transformer models from scratch, using rotary position embeddings (RoPE) [16] at each layer. To observe the random variation distribution at a variety of scales, we train our models on 250 seeds at a range of 5 scales across three variations of scaling— scaling up width, scaling up depth, and scaling with a fixed parameter count.

Task: We primarily consider an algorithmic task previously studied in Zhou et al. [20]: learning to count. Given two numbers in increasing order, the model is tasked with outputting a sequence which counts consecutively from the first number to the second number. Examples are given in the form `"5, 9 >, 5, 6, 7, 8, 9"`, with the training length representing the length of the counting sequence. It has been previously shown that models can length generalize on this counting task to more than twice their training length; however, we will see in the next section that inspecting the distribution of performance across random variation gives a more nuanced picture about the model's capacity to length generalize.

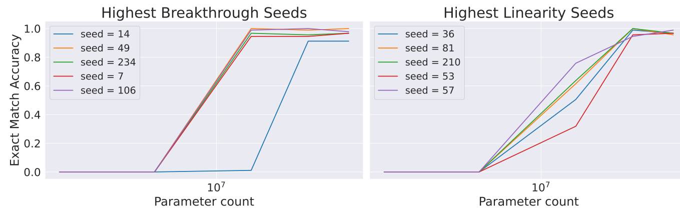
Dataset: During training, we sample sequences i.i.d from the train set and fill the context with examples, following previous work [5, 20]. The length of examples are sampled uniformly from 1 to the maximum training length, which we fix at 30.

3 Results

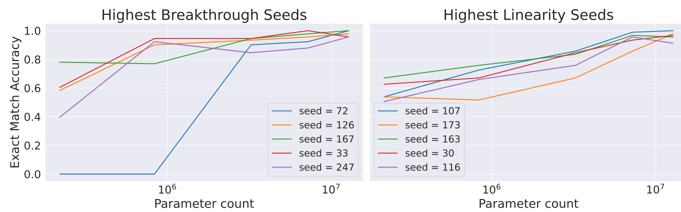
3.1 Bimodality leads to emergence

Following Srivastava et al. [15], we take the vector of model performances for a fixed seed at test length 60 across different scales and calculate their *breakthroughness* and *linearity* metric (for their definitions, refer to Appendix B). We plot the performance across scale for the top 5 seeds with the highest breakthrough metric and highest linearity metric in Figure 1. Here we fix the random seed which has become the standard practice for reporting benchmark performance for LLMs; although the same random seed has no relation across different scales, this is often a fixed parameter overlooked in practice. In Figure 2, we provide violin plots of the resulting EM accuracy distribution across our three variations of scaling models for test lengths from 30 to 100, with additional figures given in Appendix C.

As is clear in Figure 2, many parameter scales exhibit specifically bimodal distributions of length generalization capabilities. The impact of this variability is depicted in Figure 1, which illustrates that we can easily find fixed seeds that show varying levels of emergence and linearity, due to random variation in breakthroughs. Although work on emergence often makes claims as to the specific model



(a) Fixing width to 512 hidden dimension and scaling depth.



(b) Fixing depth to 4 layers and scaling width.

Figure 1: Scaling curves that exhibit emergence (according to the breakthroughness metric in [15]) and those that don't, all sampled from the same task and hyperparameters, fixing random seed for each scale.

size required for particular tasks, we show that one can find different levels of emergence and different breakthrough scales depending on random experimental conditions—particularly in our fixed depth setting (Figure 1b), where curves range from clear breakthroughs to nearly linear.

3.2 Scaling depth versus width

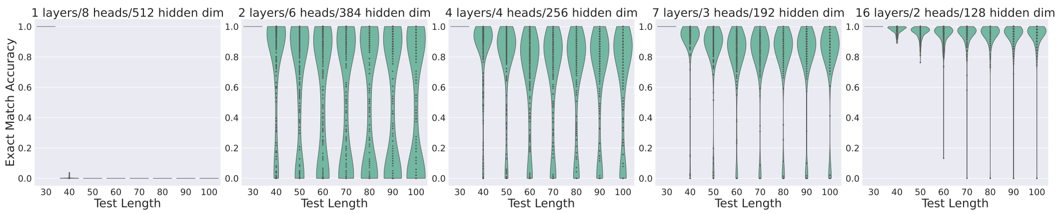
Classically, scaling laws treat parameter size as a single scalar value without differentiating between an increase in width or depth. This difference, though of limited consequence for unimodal capabilities, becomes crucial when considering breakthrough capabilities. In Figure 2a, we see that, as depth increases, a fixed parameter count goes from all model runs failing to almost all models scoring over 80% accuracy even at the longest test lengths. When focusing on parameter growth through depth (Figure 2c), we see mass move from the failed runs to the successful runs; meanwhile, if parameters grow through hidden width (Figure 2b), the probability of a run failing completely does not decrease, but the average performance of a successful run increases.

The specifics of how these distributions shift is shown in Figure 3. Here, we plot the fraction of runs which achieve EM accuracy below 10% and above 50% across model scales, as well as the mean of the runs which achieve above 50% accuracy across scales. First, note that the average accuracy for a "successful" run (any run with above 50% EM accuracy) increases persistently and monotonically with width, but not depth. Second, as we increase depth, the proportion of successful runs increases and the proportion of failed runs (below 10% match accuracy) falls—but as we increase width, neither trend appears. We conclude that both depth and width can improve performance, but depth improves performance primarily by shifting the likelihood of a breakthrough, whereas width only improves the expected performance of a model with the breakthrough capability.

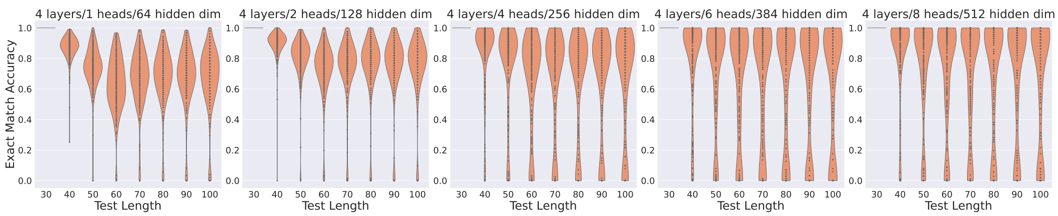
4 Discussion and conclusions

Zhou et al. [20] first documented the variability of length generalization across random seeds, which we take advantage of in our work. In general, out-of-distribution behavior like compositional rules [11] or associative biases [14] often exhibit extreme variation compared to in-distribution performance. We are also not the first to note bimodal distributions like these, which are also present in text classifier performance metrics [7] and even in the timing of generalization breakthroughs during training [4].

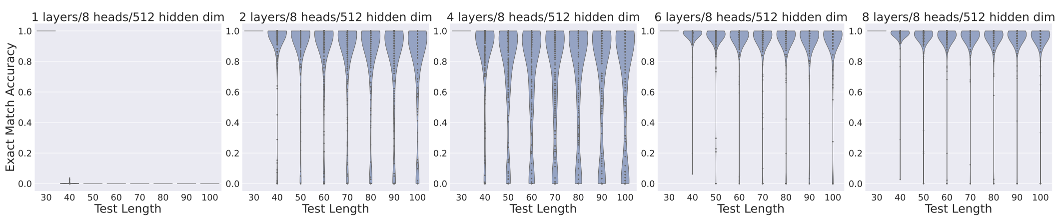
Our findings are highly suggestive, but they leave much to future work. Studying other emergent synthetic tasks, especially other compositional or length generalization settings, would confirm how general these findings are. Most crucially, we intend to study whether breakthrough capabilities in language models are associated with bimodal or outlier behavior.



(a) Fixing parameter count to 3.2m parameters.



(b) Fixing network depth to 4 layers.



(c) Fixing network width to 512 hidden dimension.

Figure 2: Violin plots illustrating how the EM accuracy distributions across test lengths vary across a) increasing depth while fixing parameter count, b) increasing width while fixing depth, and c) increasing depth while fixing width. Although the number of parameters remains the same, the distribution in deeper networks exhibits significantly less bimodality. This is further exemplified where at fixed depths, bimodality remains present even at greater widths.

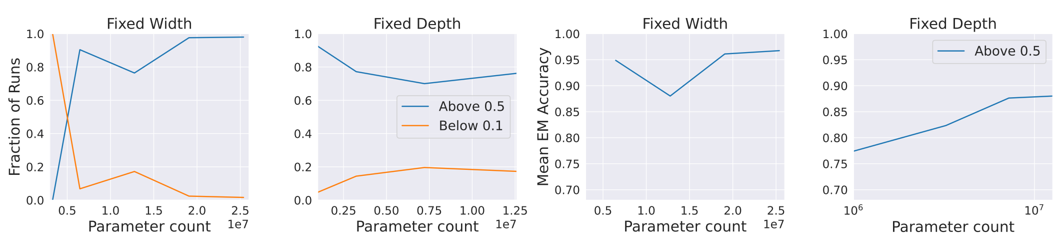


Figure 3: Across fixed width and fixed depth, on the two leftmost plots we show the fraction of runs which achieve above 50% and below 10% EM accuracy, and on the two rightmost plots we show the mean performance of the runs which achieve EM accuracy above 50%.

References

[1] A. Chen, R. Shwartz-Ziv, K. Cho, M. L. Leavitt, and N. Saphra. Sudden Drops in the Loss: Syntax Acquisition, Phase Transitions, and Simplicity Bias in MLMs. 2024. URL https://openreview.net/forum?id=MO5PiKHELW.

[2] D. Ganguli, D. Hernandez, L. Lovitt, A. Askell, Y. Bai, A. Chen, T. Conerly, N. Dassarma, D. Drain, N. Elhage, et al. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764, 2022.

[3] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[4] M. Y. Hu, A. Chen, N. Saphra, and K. Cho. Latent State Models of Training Dynamics. *Transactions on Machine Learning Research*, Aug. 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=NE2xXWo0LF.

[5] S. Jelassi, D. Brandfonbrener, S. M. Kakade, et al. Repeat after me: Transformers are better than state space models at copying. In *Forty-first International Conference on Machine Learning*.

[6] K. Jordan. On the variance of neural network training with respect to test sets and distributions. In *The Twelfth International Conference on Learning Representations*, 2024.

[7] J. Juneja, R. Bansal, K. Cho, J. Sedoc, and N. Saphra. Linear Connectivity Reveals Generalization Strategies. 2023. URL https://openreview.net/forum?id=hY6M0JHl3uL.

[8] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[9] A. T. Löwe, L. Touzo, P. S. Muhle-Karbe, A. M. Saxe, C. Summerfield, and N. W. Schuck. Abrupt and spontaneous strategy switches emerge in simple regularised neural networks, 2024. URL https://arxiv.org/abs/2302.11351.

[10] L. Madaan, A. K. Singh, R. Schaeffer, A. Poulton, S. Koyejo, P. Stenetorp, S. Narang, and D. Hupkes. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024.

[11] R. T. McCoy, J. Min, and T. Linzen. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv:1911.02969 [cs]*, Nov. 2019. URL http://arxiv.org/abs/1911.02969. arXiv: 1911.02969.

[12] J. S. Rosenfeld, A. Rosenfeld, Y. Belinkov, and N. Shavit. A constructive prediction of the generalization error across scales, 2019. URL https://arxiv.org/abs/1909.12673.

[13] R. Schaeffer, B. Miranda, and S. Koyejo. Are Emergent Abilities of Large Language Models a Mirage?, May 2023. URL http://arxiv.org/abs/2304.15004. arXiv:2304.15004 [cs].

[14] T. Sellam, S. Yadlowsky, J. Wei, N. Saphra, A. D'Amour, T. Linzen, J. Bastings, I. Turc, J. Eisenstein, D. Das, I. Tenney, and E. Pavlick. The MultiBERTs: BERT Reproductions for Robustness Analysis. *arXiv:2106.16163 [cs]*, Mar. 2022. URL http://arxiv.org/abs/2106.16163. arXiv: 2106.16163.

[15] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

[16] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[17] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent Abilities of Large Language Models, June 2022. URL http://arxiv.org/abs/2206.07682. Number: arXiv:2206.07682 arXiv:2206.07682 [cs].

[18] H. Zhou, A. Bradley, E. Littwin, N. Razin, O. Saremi, J. Susskind, S. Bengio, and P. Nakkiran. What algorithms can transformers learn? a study in length generalization, 2023. URL https://arxiv.org/abs/2310.16028.

[19] H. Zhou, A. Bradley, E. Littwin, N. Razin, O. Saremi, J. M. Susskind, S. Bengio, and P. Nakkiran. What algorithms can transformers learn? a study in length generalization. In *The Twelfth International Conference on Learning Representations*, 2024.

[20] Y. Zhou, U. Alon, X. Chen, X. Wang, R. Agarwal, and D. Zhou. Transformers can achieve length generalization but not robustly. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

A Additional Experimental Details

Count task: For all of our training runs, we fix the vocabulary size to 150. For evaluation, we compute the exact match (EM) accuracy across all consecutive subsequences of the test length.

Model scales: As mentioned in Section 2, we scale up our models in three ways: fixing width and scaling depth, fixing depth and scaling width, and fixing parameter count and scaling depth. The precise parameters for each variation are as follows:

- **Fixed depth:** We fix the network depth to 4 layers and vary width by taking hidden dimensions $\{64, 128, 256, 384, 512\}$. The head dimension is fixed to 64.
- **Fixed width:** We fix the hidden dimension to be 512 and vary the depth from $\{1, 2, 4, 6, 8\}$ layers.
- **Fixed parameter count:** We fix the parameter count to 3.2m and vary the depth from $\{1, 2, 4, 7, 16\}$ layers, with the appropriate hidden dimension.

Hyperparameters: We train all of our models to convergence on the train distribution and use a learning rate of $1e-3$ with a cosine decay scheduler and weight decay 0.1. We set the maximum training duration to be 10000 steps, with batch size 128 and context length 256.

B Breakthroughness and Linearity

Srivastava et al. [15] introduced *breakthroughness* and *linearity* metrics to capture model performance improving suddenly or reliably with scale. Given a model's performances y_i at model scales x_i sorted by ascending model scale, the linearity metric L and breakthroughness metric B are respectively calculated as

$$L = \frac{I(y)}{\mathrm{RootMeanSquare}(\{y_{i+1} - y_i\}_i)},$$
$$B = \frac{I(y)}{\mathrm{RootMedianSquare}(\{y_{i+1} - y_i\}_i)}$$

where $I(y) = \mathrm{sign}(\arg\max_i y_i - \arg\min_i y_i)(\max_i y_i - \min_i y_i)$.

C Additional Figures

In Figures 4, 5, and 6 we provide raw histograms for EM accuracy across test lengths in $\{50, 60, 70, 80, 90, 100\}$ and across model scales fixing parameter count, fixing depth, and fixing width respectively. As highlighted in Section 3.2, for a fixed parameter count we see a transition from all model runs failing to almost all models scoring over 80% accuracy even at the longest test lengths. When fixing depth, there remains a fraction of runs failing even at larger widths, but the average performance of a successful run increases. Finally, when fixing width, mass seems to shift from the failed runs to successful runs, again due to scaling depth.

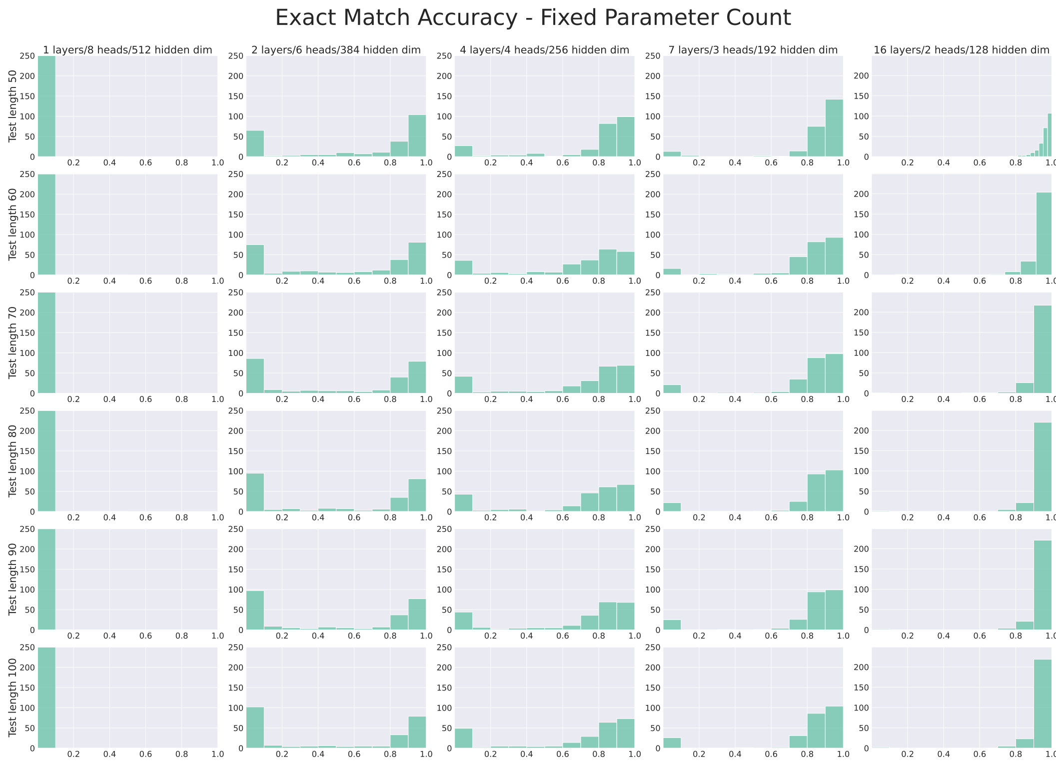


Figure 4: Histogram across model scales fixing parameter count at different test lengths.

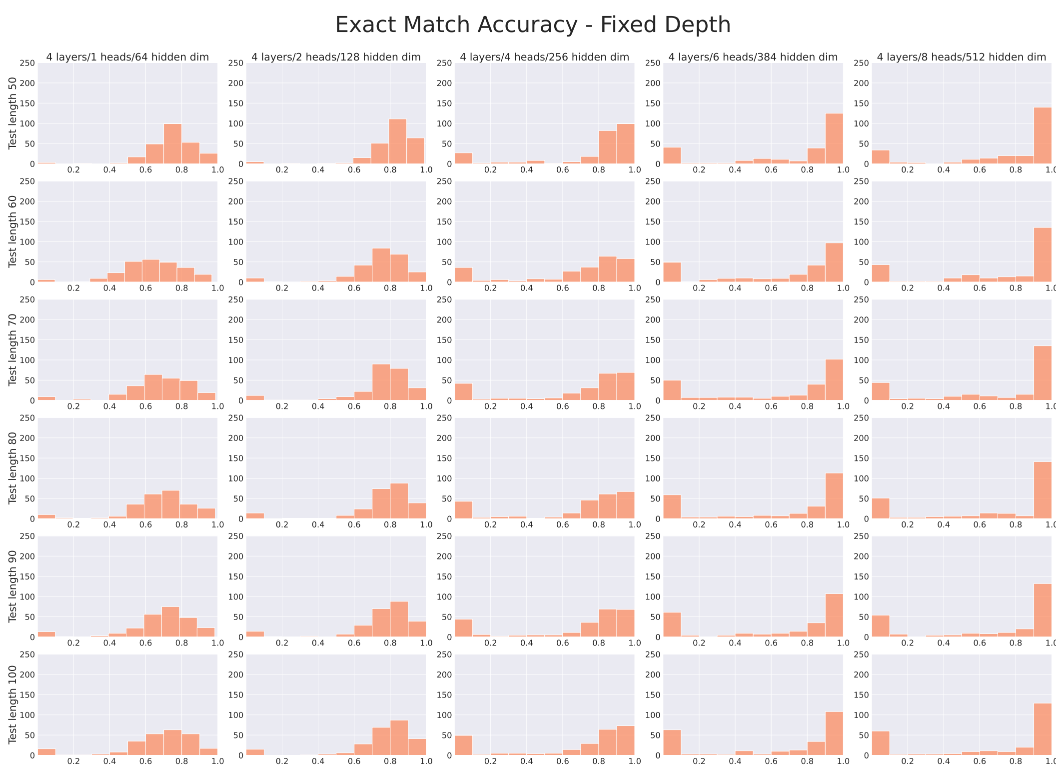


Figure 5: Histogram across model scales fixing network depth at different test lengths.

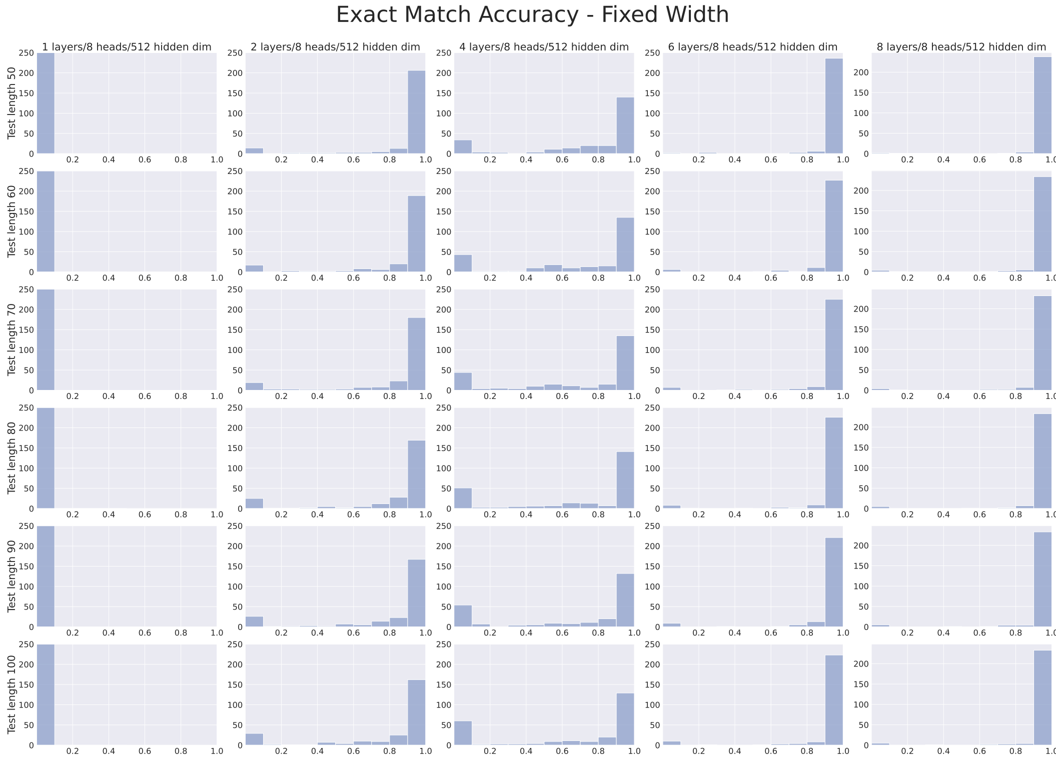


Figure 6: Histogram across model scales fixing network width at different test lengths.

D Debunking Challenge Submission

D.1 What commonly-held position or belief are you challenging?

It is now widely recognized that increasing the scale of language models (e.g., in terms of training computation and model parameters) can result in improved performance and sample efficiency across various downstream NLP tasks. However, there are cases where performance can be predicted smoothly in terms of scale via scaling laws [3, 8] as well as cases where performance cannot be predicted as scale increases due to discontinuous improvements [2, 15, 17]. Our work simultaneously addresses both common positions on emergent capabilities. The first position is that continuous metrics can always make them discrete and smooth continuous scaling laws always hold, even in apparent breakthroughs. The second position is that they represent exceptions to scaling laws determined by capacity at scale. Prior work has called into question the empirical validity of emergence; for instance, Schaeffer et al. [13] argued that more than 92% of proposed emergent skills were caused by thresholding effects in accuracy metrics. This does however, still leave a number of potential emergent abilities. Thus, our work addresses both perspectives by considering the distribution across performance due to random variation, where both emergence and linear improvement can be present for the same task and training hyperparameters.

D.2 How are your results in tension with this commonly-held position?

We propose continuous scaling laws in the *distribution* of capabilities rather than the sampled scalar value associated with that capability for a single trained model. By addressing distributions rather than point samples, we explain emergent behavior using scaling laws that are entirely continuous—but *bimodal*. In other words, the notion of a pointwise scaling "law" leads inevitably to discontinuities, but these discontinuities are stochastic in nature and determined by a continuous underlying distribution. We show the potential issue with only considering point samples in Figure 1, where even on the same hyperparameters, different random seeds exhibit varying degrees of breakthroughness and linearity.

Another element of tension is with the notion of a breakthrough parameter size. While the scaling laws literature tends to treat depth and width equivalently—as they may be in-distribution—we show that depth and width are, in fact, materially different in how they change the probability of a breakthrough. We discuss this in depth in Section 3.2, where upon inspecting the distribution across scale, we see that increasing depth increases the likelihood of sampling from a "successful" distribution, whereas width improves the average performance of "successful" models.

D.3 How do you expect your submission to affect future work?

When targeting emergent properties, our initial results suggest that depth may be prioritized over width. Furthermore, future work on emergence may involve training multiple seeds—though of course, at very large scales this may be resource-intensive. In general, we hope our results provide further evidence to the susceptibility of research on emergence in LLMs to statistical artifacts like the Texas sharpshooter fallacy. Although it becomes computationally prohibitive to run multiple seeds with the same set of hyperparameters to statistically validate empirical observations, we believe the effect of random variation on out-of-distribution or downstream performance should not be disregarded especially if the variation is not Gaussian and can be presented as multimodal. There are several directions for future work, such as investigating the generalizability of our findings and particularly the presence of bimodal distributions across different tasks, investigating influences of various hyperparameters to the random variation distribution, and studying seed performance correlation in the multi-task setting.