How Do Large Language Monkeys Get Their Power (Laws)?

Rylan Schaeffer [1] **Joshua Kazdan** [2] **John Hughes** [3 4] **Jordan Juravsky** [1] **Sara Price** [4] **Aengus Lynch** [4 5]
Erik Jones [6] **Robert Kirk** [5] **Azalia Mirhoseini** [1] **Sanmi Koyejo** [1]

Abstract

Recent research across mathematical problem solving, proof assistant programming and multimodal jailbreaking documents a striking finding: when (multimodal) language model tackle a suite of tasks with multiple attempts per task – succeeding if any attempt is correct – then the negative log of the average success rate scales a power law in the number of attempts. In this work, we identify an apparent puzzle: a simple mathematical calculation predicts that on each problem, the failure rate should fall exponentially with the number of attempts. We confirm this prediction empirically, raising a question: from where does aggregate polynomial scaling emerge? We then answer this question by demonstrating per-problem exponential scaling can be made consistent with aggregate polynomial scaling if the distribution of single-attempt success probabilities is heavy tailed such that a small fraction of tasks with extremely low success probabilities collectively warp the aggregate success trend into a power law - even as each problem scales exponentially on its own. We further demonstrate that this distributional perspective explains previously observed deviations from power law scaling, and provides a simple method for forecasting the power law exponent with an order of magnitude lower relative error, or equivalently, $\sim 2-4$ orders of magnitude less inference compute. Overall, our work contributes to a better understanding of how neural language model performance improves with scaling inference compute and the development of scaling-predictable evaluations of (multimodal) language models.

[1]Stanford Computer Science [2]Stanford Statistics [3]Speechmatics [4]ML Alignment & Theory Scholars [5]University College London [6]Anthropic. Correspondence to: Rylan Schaeffer <rschaef@cs.stanford.edu>, Sanmi Koyejo <sanmi@cs.stanford.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

Scaling behaviors of large neural language models have surprised and fascinated engineers, scientists and society alike (Hestness et al., 2017; Kaplan et al., 2020; Brown et al., 2020a; Hoffmann et al., 2022; Ganguli et al., 2022; Sorscher et al., 2022; Wei et al., 2022b; Schaeffer et al., 2023; OpenAI et al., 2024), shaping engineering, economic and governmental interests in frontier AI systems (Bommasani et al., 2021; Eloundou et al., 2023; Anderljung et al., 2023; Wang et al., 2023; Reuel et al., 2024; Besiroglu et al., 2024a; Maslej et al., 2024). For a more thorough exposition of relevant literature, please see Related Work (Section 6).

One direction of renewed interest is inference-time compute scaling, whereby compute is controllably increased at inference to improve the performance of a model, e.g., Pachocki et al. (2024). In this direction, recent research discovered that language model success rates scale predictably with the number of independent attempts made at accomplishing a task. Specifically, in a paper titled, "Large Language Monkeys: Scaling Inference Compute with Repeated Sampling," Brown et al. (2024) studied how language model performance changes at mathematical problem solving and coding problems when k independent attempts are sampled per problem. Performance on the i-th problem was measured using the expected (over attempts) success rate (Kulal et al., 2019; Chen et al., 2021), defined as:

$$\text{pass}_\text{i}@\text{k} \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{k \text{ Attempts}}\Big[\mathbb{I}[\text{Any attempt on } i\text{-th problem succeeds}]\Big]. \quad (1)$$

Using the unbiased and numerically stable estimator of Chen et al. (2021) (for details, see Appendix B), Brown et al. (2024) found that the negative log averaged-over-P-problems success rate falls as a power law with the number of independent attempts per problem k:

$$-\log\left(\frac{1}{P}\sum_{i=1}^{P}\text{pass}_\text{i}@\text{k}\right) \approx ak^{-b}, \quad (2)$$

for model-specific and benchmark-specific constants $a, b > 0$ (Fig. 1 Top). Soon after, on a separate topic of jailbreaking multimodal language models via text, image and audio

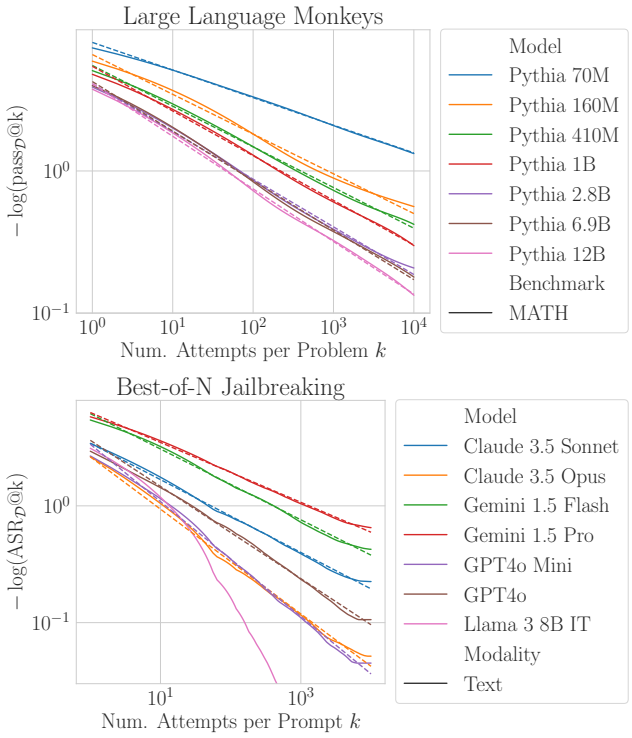

Figure 1: **Power Law Scaling in Language Models from Repeat Sampling.** Top: Brown et al. (2024) found the negative log average pass rate $-\log(\text{pass}_\mathcal{D}@\text{k})$ at solving mathematical problems scales polynomially (i.e., as a power law) with the number of independent attempts per problem k. Bottom: Hughes et al. (2024) similarly found the negative log average attack success rate $-\log(\text{ASR}_\mathcal{D}@\text{k})$ when jailbreaking multimodal language models scales polynomially with the number of jailbreak attempts per prompt. Should such power law scaling be expected? From where do large language monkeys obtain their power (laws)?

attacks, independent work by Hughes et al. (2024) studied jailbreaking success rates when k independent attempts are made per harmful prompt. Performance was measured using Attack Success Rate (ASR) at k:

$$\text{ASR}_\text{i}@\text{k} \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{k \text{ Attempts}}\Big[\mathbb{I}[\text{Any attack on } i\text{-th prompt succeeds}]\Big]. \tag{3}$$

This "Best-of-N Jailbreaking" attack similarly discovered that the negative log averaged-over-P-prompts attack success rate fell as a power law with the number of jailbreak attempts per prompt k:

$$-\log\left(\frac{1}{P}\sum_{i=1}^{P}\text{ASR}_\text{i}@\text{k}\right) \approx ak^{-b}, \tag{4}$$

for model-specific and modality-specific constants $a, b > 0$ (Fig. 1 Bottom). For the specific coefficients from both papers, see Appendix. C. As a minor matter of terminology, both papers frame their results in terms of "coverage" – the fraction of problems that can be solved after k attempts per problem – but as Brown et al. (2024) pointed out, coverage is equivalent to the average success rate (Appendix D); we prefer this latter framing as it avoids the binary implication that each problem either is or is not solved after k attempts.

2. Should Power Law Scaling Be Expected?

Should we expect large language monkeys to have such power (laws)? That is, should the negative log of the average success rate scale polynomially with the number of independent attempts k? As we now explain mathematically and demonstrate empirically, such polynomial scaling with k is perhaps surprising because, for any single problem, the negative log success rate at k should fall exponentially with k; the intuition is that $\text{pass}_\text{i}@\text{k}$ is 1 unless *all* attempts fail, and since attempts are independent, the probability that all fail is exponentially unlikely with the number of attempts.

Mathematically, on any given attempt, the model has probability $\text{pass}_\text{i}@1$ of solving the i-th problem. Recalling that $\text{pass}_\text{i}@\text{k}$ is defined as 1 if *any* of the k attempts succeed, 0 otherwise, by linearity of expectation and by independence of the k attempts, we can rewrite $\text{pass}_\text{i}@\text{k}$ as:

$$\text{pass}_\text{i}@\text{k} = \mathop{\mathbb{E}}_{k \text{ Attempts}}\Big[1 - \mathbb{I}[\text{All } k \text{ Attempts Fail}]\Big] \tag{5}$$

$$= 1 - \prod_{j=1}^{k}\mathop{\mathbb{E}}_{1 \text{ Attempt}}\Big[\mathbb{I}[j\text{-th Attempt Fails}]\Big]. \tag{6}$$

The probability that the j-th attempt fails is one minus the probability that the j-th attempt succeeds. Since each attempt is i.i.d. with success probability $\text{pass}_\text{i}@1$, we find

$$\text{pass}_\text{i}@\text{k} = 1 - (1 - \text{pass}_\text{i}@1)^k. \tag{7}$$

For large k, $(1 - \text{pass}_\text{i}@1)^k$ will be small. Recalling that the Taylor Series expansion of $\log(1 + x)$ for small x is $\sum_{i=1}^{\infty}(-1)^{i-1}x^i/i \approx x$, we have:

$$-\log(\text{pass}_\text{i}@\text{k}) = -\log\Big(1 - (1 - \text{pass@1})^k\Big) \tag{8}$$

$$\approx (1 - \text{pass}_\text{i}@1)^k. \tag{9}$$

Thus, *for any single problem*, we should expect the negative log expected (over attempts) success rate to fall *exponentially* with k, not polynomially with k.

To confirm this claim, we plotted the scaling of model performance on each problem – measured either by $-\log(\text{pass}_\text{i}@\text{k})$ or by $-\log(\text{ASR}_\text{i}@\text{k})$ – against the number of independent attempts k. We specifically used Brown

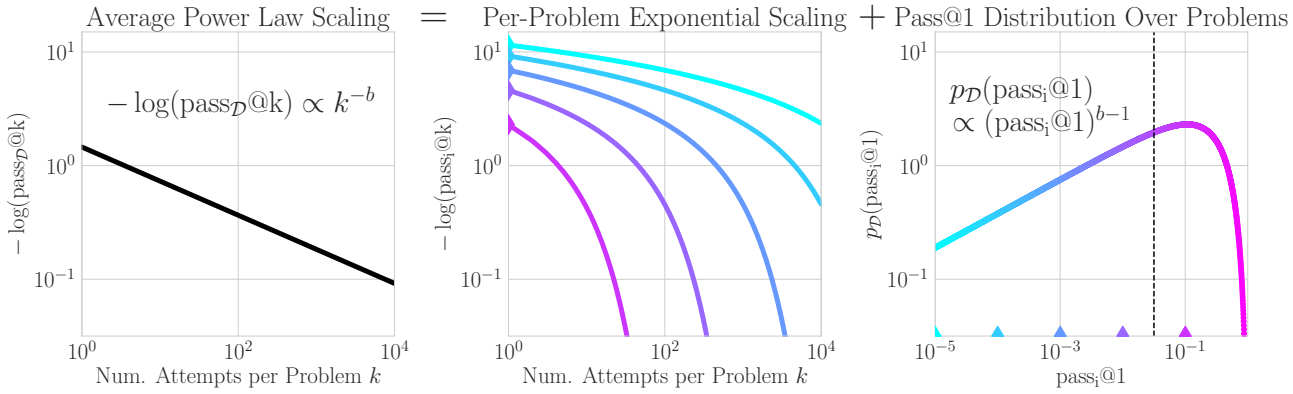

Figure 2: **Schematic: The Origin of Power Laws from Scaling Inference Compute via Repeat Sampling.** The $-\log(\text{pass}_{\mathcal{D}}@k)$ scales as a power law with the number of attempts per problem k (left). This arises from a combination of two factors: (1) for each problem, $-\log(\text{pass}_i@k)$ scales exponentially with k (center), and (2) the distribution (over problems in the dataset) of single-attempt success rates $\text{pass}_i@1$ itself has a left power-law tail of small values (right).

et al. (2024)'s data of the Pythia language model family (Biderman et al., 2023) solving 128 mathematical problems from MATH Hendrycks et al. (2021) as well as Hughes et al. (2024)'s data from jailbreaking frontier AI systems – Claude, GPT4 (OpenAI et al., 2024), Gemini (Team et al., 2024a;b) and Llama 3 8B Instruction Tuned (IT) (Grattafiori et al., 2024) – on 159 prompts from HarmBench (Mazeika et al., 2024). For each individual mathematical problem and jailbreaking prompt, we found the negative log expected (over attempts) success rates fall exponentially with k as expected (Fig. 3), including on Llama 3 8B IT which does not exhibit an aggregate power law (Fig. 1).

3. Distribution of Per-Problem Single-Attempt Success Rates Creates Power Law Scaling

How does polynomial scaling of the negative log *average* success rate emerge from exponential scaling of the negative log *per-problem* success rate? The answer to this question *must* lie in the distribution $\mathcal{D}$ over benchmark problems of single attempt (i.e., $k = 1$) success rates because this distribution's density $p_{\mathcal{D}}(\text{pass}_i@1)$ links the per-problem scaling behavior to the aggregate scaling behavior via the definition of the aggregate success rate $\text{pass}_{\mathcal{D}}@k$:

$$\begin{aligned}\text{pass}_{\mathcal{D}}@k &\overset{\text{def}}{=} \mathop{\mathbb{E}}_{\text{pass}_i@1\sim\mathcal{D}}\Big[\text{pass}_i@k(\text{pass}_i@1)\Big] \\ &= 1 - \int_0^1 (1-\text{pass}_i@1)^k \, p_{\mathcal{D}}(\text{pass}_i@1)\, \mathrm{d}\,\text{pass}_i@1\,.\end{aligned} \tag{10}$$

Based on a known result that power laws can originate from an appropriately weighted sum of exponential functions (Appendix E.1), we begin by considering simple distributions for the single-attempt success probabilities and asking which yield power law scaling between $-\log(\text{pass}_{\mathcal{D}}@k)$ and k, as well as what properties of the distributions set the scaling exponent. In Appendices E.3-E.8, we derive that several simple distributions yield power law scaling with different exponents whereas others do not:

$$-\log\Big(\text{pass}_{\text{Uniform}(0,\,\beta\le 1)}@k\Big) \propto k^{-1}.$$
$$-\log\Big(\text{pass}_{\text{Beta}(\alpha,\beta)}@k\Big) \propto k^{-\alpha}.$$
$$-\log\Big(\text{pass}_{\text{Kumaraswamy}(\alpha,\,\beta)}@k\Big) \propto k^{-\alpha}.$$
$$-\log\Big(\text{pass}_{\text{ContinuousBernoulli}(\lambda<1/2)}@k\Big) \propto k^{-1}.$$
$$-\log\Big(\text{pass}_{\text{Reciprocal}(0<\alpha<\beta<1)}@k\Big) \propto \frac{(1-\alpha)^k}{k}.$$

To test this understanding, we examined whether the data of Brown et al. (2024) and Hughes et al. (2024) had per-problem single-attempt success rate distributions that matched one of these simple distributions (Fig. 4). We found that the distributions could indeed be well fit by a 3-parameter $\text{Kumaraswamy}(\alpha, \beta, a = 0, c)$ distribution with scale parameter c (Fig. 4, black dashed lines); we found the scale parameter was critical to obtain good fits because the standard 2-parameter Kumaraswamy distribution is supported on $(0, 1)$ whereas most single-attempt success distributions have a smaller maximum such as 0.01 or 0.1.

More generally, what are the distributional properties that create such power law scaling and that set the specific power law exponent? As we now show, the negative log average success rate will exhibit power law scaling in k with exponent b if and only if the distribution over problems of single-attempt success probabilities itself behaves like a power law near 0 with exponent $b - 1$:

Theorem 3.1 (Sufficiency of Power-Law Left Tail in Dis-

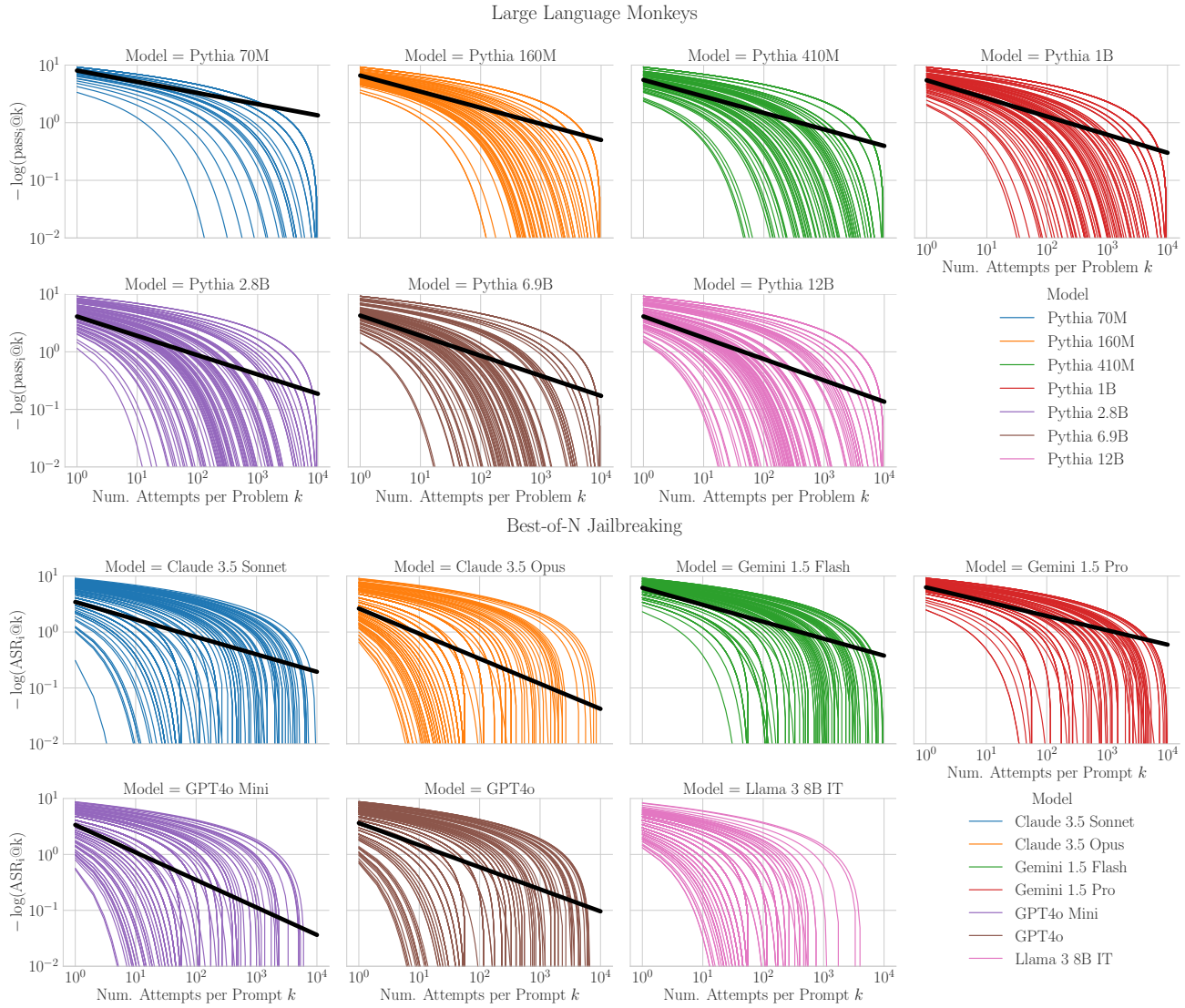

Figure 3: **Per-problem performance scales exponentially with the number of attempts per problem** k. Top: Pythia language models on 128 problems from MATH, with performance on the i-th problem measured as $-\log(\text{pass}_i@k)$. Bottom: Frontier AI models on jailbreaking prompts from HarmBench, with performance on the i-th problem measured as $-\log(\text{ASR}_i@k)$. In both settings, on each problem, the negative log *per-problem* success rate falls exponentially with the number of independent attempts k. However, the negative log *average* success rate falls as a power law with k (black).

tribution of Single-Attempt Success Rates). *Let $\mathcal{D}$ be a probability distribution on $[0,1]$ with PDF $p_{\mathcal{D}}(\text{pass}_i@1)$. Suppose there exist constants $b > 0$, $C > 0$, $\theta > 0$ and $\delta > 0$ such that, for all $0 < \text{pass}_i@1 < \delta$, we have*

$$p_{\mathcal{D}}(\text{pass}_i@1) \;=\; C{\cdot}(\text{pass}_i@1)^{b-1} + O\big((\text{pass}_i@1)^{b-1+\theta}\big).$$

Then, for large k,

$$-\log\big(\,\text{pass}_{\mathcal{D}}@k\,\big) \;\sim\; C\,\Gamma(b)\;k^{-b}.$$

Theorem 3.2 (Necessity of Power-Law Left Tail in Distribution of Single-Attempt Success Rates). *Let $\mathcal{D}$ be a distribution over $\text{pass}_i@1 \in [0,1]$ with PDF $p_{\mathcal{D}}(\text{pass}_i@1)$. Suppose there exist constants $b > 0$ and $A > 0$ such that for large k,*

$$-\log\big(\,\text{pass}_{\mathcal{D}}@k\,\big) \sim A\,k^{-b}.$$

Then, under mild regularity assumptions, the probability density must satisfy

$$p_{\mathcal{D}}(\text{pass}_i@1) \;\sim\; \frac{A}{\Gamma(b)}\,(\text{pass}_i@1)^{b-1} \quad \textit{as}\ \ \text{pass}_i@1 \to 0^{+}.$$

In Fig. 2, we illustrate this connection schematically. For proofs, see Appendices E.8 and E.9. These results clarify

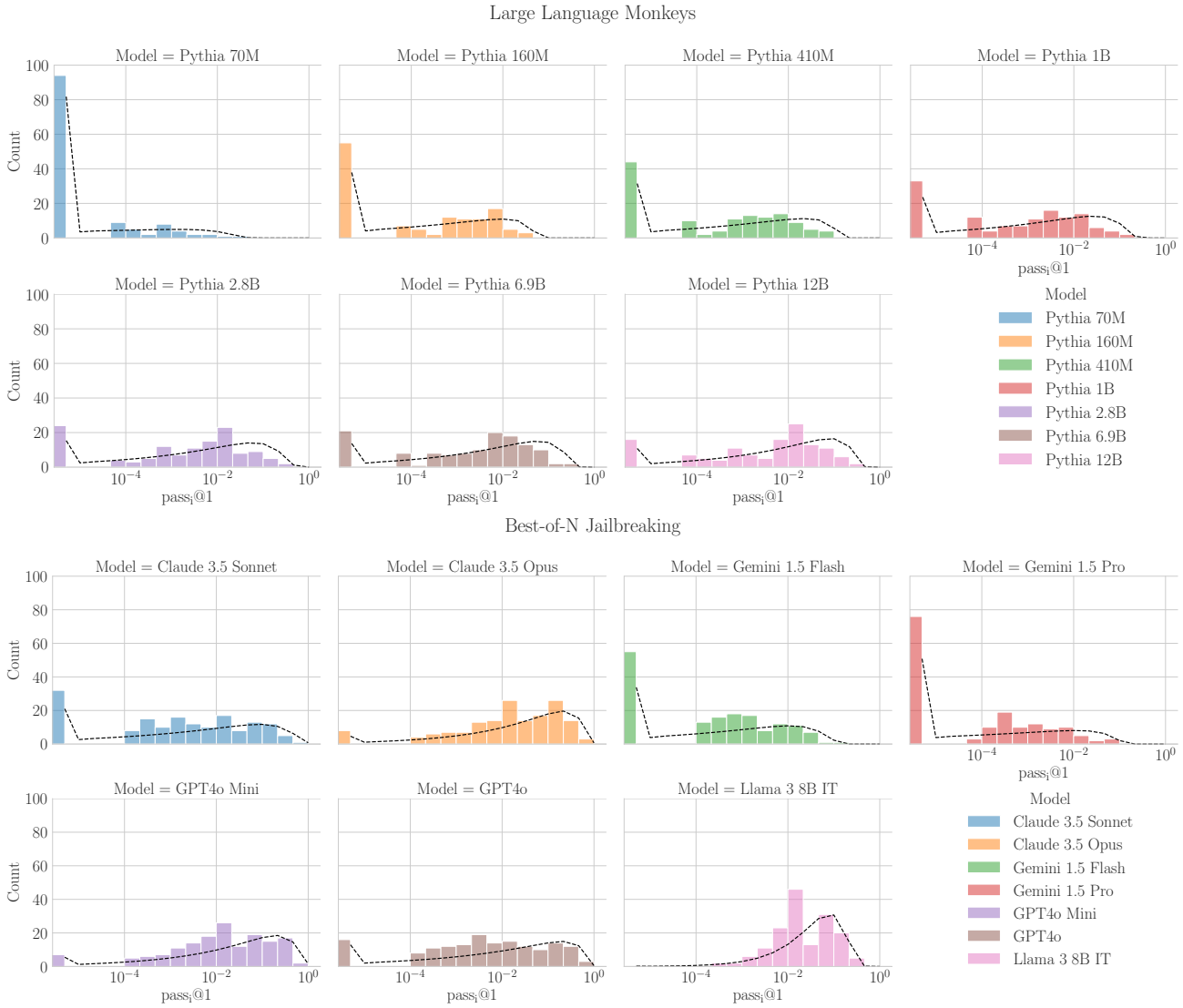

Figure 4: **Single-Attempt Success Rates Distributions Possess Power Law-Like Left Tails.** Pythia language models on 128 MATH problems (top) and frontier AI systems on 159 HarmBench prompts (bottom) exhibit distributions (over problems) of $\text{pass}_\text{i}@1$ and $\text{ASR}_\text{i}@1$ with power law-like tails that are well fit by scaled Beta-Binomial distributions (black dashed lines), which produce aggregate power law scaling. Note that Llama 3 8B Instruction Tuned (IT) does not possess a power law tail, explaining why the model did not exhibit aggregate power law scaling under Best-of-N jailbreaking (Sec. 4).

that whenever $-\log(\text{pass}_\mathcal{D}@\text{k})$ exhibits power-law decay in k with exponent b, the distribution over problems of single-attempt success rates *must* have "polynomial weight" near $\text{pass}_\text{i}@1 = 0$, i.e. $p_\mathcal{D}(p) = \Theta(p^{b-1})$.

To offer intuition, we know that each problem is being solved by the model (or equivalently, each prompt is jailbreaking the model) exponentially quickly. If one looks across all problems in the benchmark, some have $\text{pass}_\text{i}@1$ so small that they remain unsolved for many, many attempts. Whether these "tiny-$\text{pass}_\text{i}@1$" problems still matter at large k depends on how *many* such problems there are. Polynomial density near 0 "piles up" enough hard problems in just the right way such that even though each of those problems is being solved exponentially quickly, the *aggregate* success rate over problems decreases at only a power-law rate in k. A more succinct mathematical summary is that, for a compound binomial distribution, the lower tail probability controls the upper tail of the marginal survivor function.

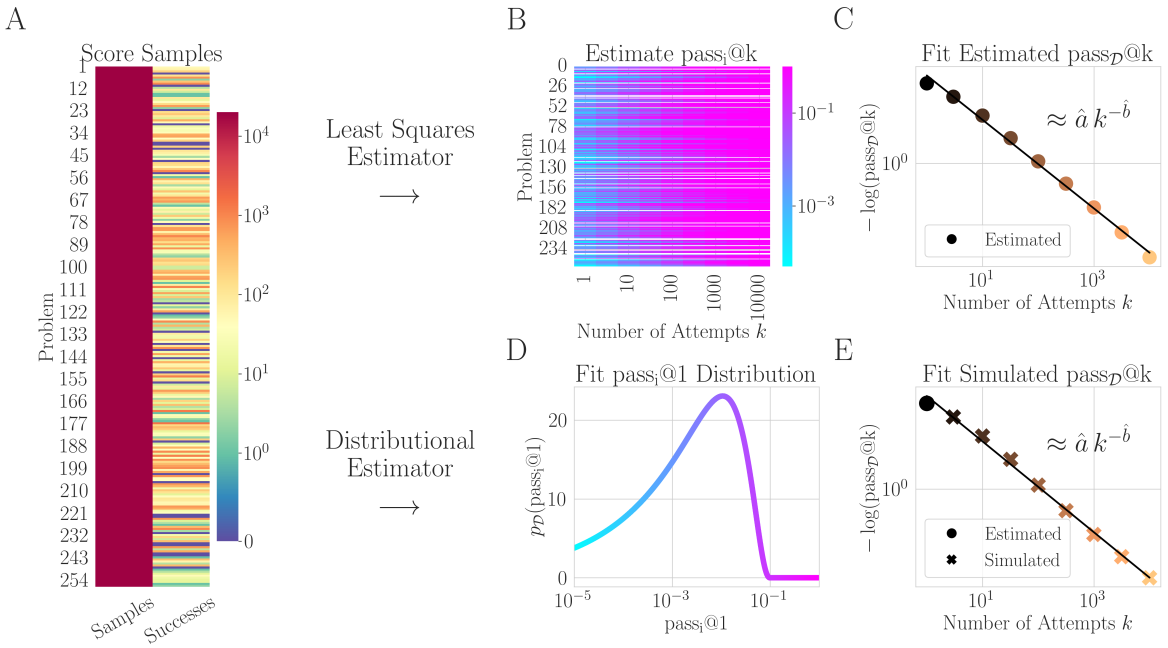

Figure 5: **Schematic: Two Estimators of Power Law Parameters for Scaling Inference Compute via Repeat Sampling.** (A) Both estimators begin by generating many samples per prompt, then computing the number of successes per prompt. In the standard least squares power law parameter estimator (top), (B) $\text{pass}_i@k$ is estimated for each i-th problem at multiple k values, then (C) averaged over problems and fit with linear regression in log-log space. In the distributional power law parameter estimator (bottom), (D) a distribution $\mathcal{D}$ is fit to estimates of $\text{pass}_i@1$, then (E) the single-attempt success probability distribution is used to simulate $\text{pass}_\mathcal{D}@k$ at arbitrary k values for linear regression in log-log space.

4. Lack of Distributional Structure Explains Deviations from Power Law Scaling

Notably, previous papers observed that not every model exhibits power law scaling in every setting. To highlight one, Hughes et al. (2024) observed that when jailbreaking Meta's Llama 3 8B Instruction Tuned (IT) model (Grattafiori et al., 2024), the $-\log(\text{ASR}_\mathcal{D}@k)$ fell faster than any power law (Fig. 1), i.e., the $\text{ASR}_\mathcal{D}@k$ rose much more quickly than the other frontier AI systems. Based on our mathematical insights and the empirical per-problem single-attempt attack success rates (Fig. 4), we can understand why: Llama 3 8B IT could be successfully jailbroken on every prompt within the permitted sampling budget and thus had no heavy left tail necessary to create the aggregate power law scaling.

5. A New Distributional Estimator for Predicting Power Law Scaling

A natural consequence of this connection between the scaling of $-\log(\text{pass}_\mathcal{D}@k)$ and the left tail of the distribution $p_\mathcal{D}(\text{pass}_i@1)$ is that the distribution of single-attempt success rates can be used to predict whether power-law scaling will appear and if so, what the intercept and exponent of the power law will be. To do this, one can fit the distribution $\hat{p}_\mathcal{D}(\text{pass}_i@1)$ and then *simulate* how $\text{pass}_\mathcal{D}@k$ will scale with k (Fig. 5) using the relationship:

$$\widehat{\text{pass}_\mathcal{D}@k} \overset{\text{def}}{=} 1 - \int_0^1 (1 - \text{pass}_i@1)^k \, \hat{p}_\mathcal{D}(\text{pass}_i@1) \, \mathrm{d}\, \text{pass}_i@1 \,. \tag{11}$$

To empirically test this claim, we compared the standard least squares regression estimator (in log-log space) (Hoffmann et al., 2022; Caballero et al., 2022; Besiroglu et al., 2024b) against a *distributional estimator*. To motivate our distributional estimator, we first need explain a key obstacle and how the distributional estimator overcomes it. The obstacle is that there are problems or prompts whose single-attempt success probabilities $\text{pass}_i@1$ lie between $(0, 1/\text{Number of Samples})$ such that, due to finite sampling, we lack the resolution to measure. While we do not know the true single-attempt success probability for the problems that lie in this interval, we *do know how many problems fall into this left tail bucket*, and we can fit a distribution's parameters such that the distribution's probability mass in the interval $(0, 1/\text{Number of Samples})$ matches the empirical fraction of problems in this tail bucket. Thus, our distributional estimator works by first selecting a distribution (e.g., a scaled 3-parameter Beta distribution), discretizing the distribution

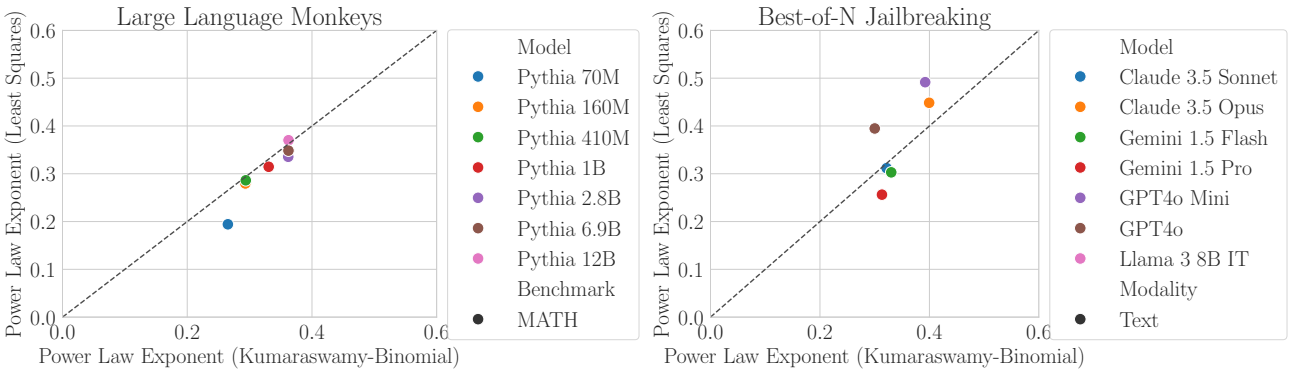

Figure 6: **Comparing Estimators of Power Law Exponents.** We compare two estimators of the power law exponent b in $-\log(\text{pass}_{\mathcal{D}}@\text{k}) \approx ak^{-b}$: (1) the standard least-squares estimator between k and $-\log(\text{pass}_{\mathcal{D}}@\text{k})$ in log-log space, and (2) the distributional estimator of $\text{pass}_{\text{i}}@1$ assuming a scaled Kumaraswamy-Binomial distribution. Using all available data to fit both estimators, we find agreement between the least-squares estimate (ordinate) and the distribution-derived estimate (abscissa) for both Pythia models on MATH (left) and for frontier AI systems on HarmBench (right). For an explanation of why the two estimators match more closely for Large Language Monkeys than for Best-of-N Jailbreaking, see Appendix A.

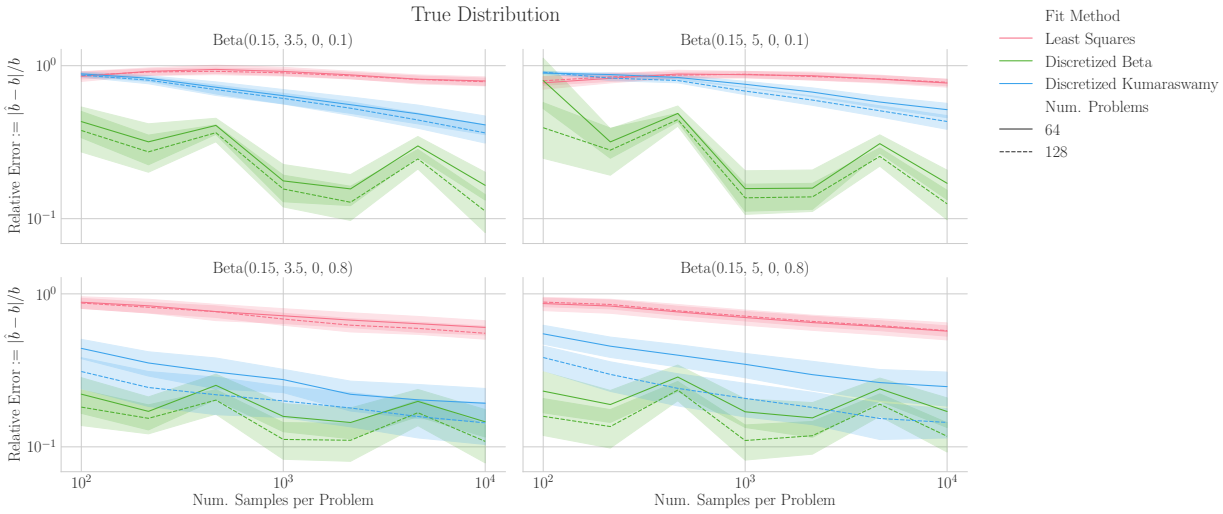

Figure 7: **Comparing Two Estimators of Power Law Exponents via Backtesting.** On synthetic data with known ground-truth power law $a\,k^{-b}$, we compare how well the least squares and the distributional estimator recover the scaling exponent b as measured by the relative error $|\hat{b} - b|/b$ by backtesting: subsampling the number of problems and the number of samples per problem. We find that the distributional estimator obtains significantly better sample efficiency.

according to the sampling resolution $1/\text{Number of Samples}$ and performing maximum likelihood estimation under the discretized distribution's probability mass function.

We tested this distributional estimator in two different ways. First, focusing on Large Language Monkeys, we used all available real data from all problems and all samples per problem to compare the standard least squares regression estimator against the distributional estimator. We found close agreement between the two estimators (Fig. 6), giving us a sense that the two estimators yield reasonably consistent estimates under large sampling budgets.

Second, the distributional estimator also comes with another benefit: it directly provides an estimate of the power law's exponent b in $a\,k^{-b}$. Estimating the power law's exponent is especially valuable because the exponent dictates how success rates are improving with increasing inference compute. To test how the distributional estimator and least squares estimator compare at recovering the true asymptotic power law exponent, we generated synthetic data so that we would have ground-truth knowledge of the true power law exponent, then backtested how the two scaling estimators compare at recovering the true exponent (Alabdulmohsin et al., 2022a; Owen, 2024) by subsampling data with fewer

problems and fewer samples per problem. We found that the distributional estimator obtains significantly better sample efficiency, with approximately an order of magnitude lower relative error $\stackrel{\text{def}}{=} |\hat{b} - b|/b$ compared with the least squares estimator (Fig. 7), or equivalently, $\sim 2 - 4$ orders of magnitude less inference-compute. The distributional estimator performs well even under distributional mismatch.

6. Related Work

Research into scaling laws of deep neural networks has a rich history spanning theoretical foundations, empirical validations, and diverse applications. The earliest investigations discovered power law scaling in simple machine learning settings (Barkai et al., 1993; Mhaskar, 1996; Pinkus, 1999). However, the modern era of scaling laws began with breakthrough studies in neural language models (Hestness et al., 2017; Kaplan et al., 2020; Brown et al., 2020b), catalyzing extensive research across multiple directions. The theoretical understanding of scaling laws has advanced significantly (Spigler et al., 2020; Bousquet et al., 2020; Hutter, 2021; Sharma & Kaplan, 2022; Maloney et al., 2022; Roberts et al., 2022; Bahri et al., 2024; Michaud et al., 2024; Paquette et al., 2024; Atanasov et al., 2024; Bordelon et al., 2024a;b; Lin et al., 2024; Brill, 2024), complemented by comprehensive empirical studies (Rosenfeld et al., 2020; Henighan et al., 2020; Gordon et al., 2021; Tay et al., 2021; Ghorbani et al., 2021; Tay et al., 2022b; Zhai et al., 2022; Alabdulmohsin et al., 2022b; Dehghani et al., 2023; Bachmann et al., 2023). In the context of language models, researchers have explored scaling behaviors in various aspects: context length (Xiong et al., 2023), in-context learning (Chan et al., 2022; Agarwal et al., 2024; Arora et al., 2024), vocabulary size (Tao et al., 2024), and jailbreaking attempts (Anil et al., 2024; Hughes et al., 2024). Studies have also investigated scaling dynamics in fine-tuning (Kalajdzievski, 2024; Zhang et al., 2024), transfer learning (Hernandez et al., 2021), and the impact of repeated data (Hernandez et al., 2022; Muennighoff et al., 2023). Architectural considerations have been extensively studied, including network design (Tay et al., 2022a; Clark et al., 2022), nested models (Kudugunta et al., 2023), pruning strategies (Rosenfeld et al., 2021), and precision requirements (Dettmers & Zettlemoyer, 2023; Kumar et al., 2024; Sun et al., 2025). Research has also addressed multimodal extensions (Aghajanyan et al., 2023; Cherti et al., 2023) and inference optimization (Sardana et al., 2023; Brown et al., 2024; Snell et al., 2024a; Wu et al., 2024; Chen et al., 2024). The field has expanded to encompass diverse domains including reinforcement learning (both single-agent (Jones, 2021; Hilton et al., 2023; Neumann & Gros, 2024) and multi-agent (Neumann & Gros, 2022)), graph networks (Liu et al., 2024), diffusion models (Mei et al., 2024; Liang et al., 2024), and associative memory models (Romani et al., 2013; Cabannes et al., 2024; Schaeffer et al., 2024c). Recent work has explored emerging phenomena such as inverse scaling (McKenzie et al., 2024), unique functional forms (Caballero et al., 2022), scaling patterns across model families (Ruan et al., 2024; Polo et al., 2024), and downstream capabilities (Srivastava et al., 2023; Wei et al., 2022a; Hu et al., 2024; Schaeffer et al., 2024b; Snell et al., 2024b; Wu & Lo, 2024). Researchers have also investigated critical challenges including data contamination (Schaeffer, 2023; Jiang et al., 2024; Dominguez-Olmedo et al., 2024), model-data feedback loops (Dohmatob et al., 2024; Gerstgrasser et al., 2024; Kazdan et al., 2024), and overtraining effects (Gao et al., 2023; Gadre et al., 2024). Additional contributions include studies in sparse autoencoders (Gao et al., 2024), biologically-plausible backpropagation (Filipovich et al., 2022), and self-supervised learning for vision (Schaeffer et al., 2024a). Recent efforts have also focused on reconciling apparent contradictions in scaling behaviors (Besiroglu et al., 2024b; Porian et al., 2024).

7. Discussion and Future Directions

This work advances our mathematical understanding of how and why language model performance improves with additional inference compute through repeat sampling. By establishing rigorous theoretical foundations for these empirically-observed power laws, our work provides practitioners with principled ways to understand and predict model performance when scaling inference compute. The distributional perspective we develop explains previously puzzling deviations from power law scaling and enables more efficient estimation of scaling parameters.

Two related questions are *why* such distributional structure exists in the single-attempt success rates and whether one should expect such structure to appear in future benchmarks. We conjecture there are at least two reasons: (1) benchmark design, in that benchmarks are intentionally crafted that problems have a spread of difficulty without being too easy or too hard, and (2) selection bias, in that more interesting patterns such as power law scaling are more likely to garner more interest from the research community.

Despite focusing on scaling inference compute, our paper contributes a new hypothesis for an open question in scaling pretraining compute: *why are neural scaling laws power laws?* Just as the scaling behavior of $-\log(\text{pass}_{\mathcal{D}}@\text{k})$ only becomes clear for large k, so too might the scaling behavior of pretraining cross entropy with pretraining compute C. Specifically, suppose the pretraining cross entropy $\mathcal{L}$ as a function of pretraining compute C is a sum of many functions which decay at different rates:

$$\mathcal{L}(C) = \omega\Big(\frac{1}{C^{\alpha}}\Big) + \frac{A}{C^{\alpha}} + o\Big(\frac{1}{C^{\alpha}}\Big),$$

where α is the smallest (positive) polynomial exponent and

$\omega(1/C^\alpha)$ represents functions that decay more slowly than any polynomial. Initially, for small C, the dominant term may be unclear, but as pretraining compute is scaled up across $8-10$ orders of magnitude, the leading order term dominates and an approximate power law emerges:

$$\mathcal{L}(C) \approx \text{const} + \frac{A}{C^\alpha} + 0 \quad \text{as} \quad C \to \infty.$$

Thus, a power law relationship may only be reasonable for sufficiently large pretraining compute C, which in turn may require excluding the lowest pretraining compute models in order to obtain good predictions, justifying a widespread empirical practice (Kaplan et al., 2020). We designate possible functions hiding in $\omega(1/C^\alpha)$ and $o(1/C^\alpha)$ as *the dark matter of neural scaling laws*.

Acknowledgments

RS acknowledges support from Stanford Data Science and the OpenAI Superalignment Fast Grant. SK acknowledges support by NSF 2046795 and 2205329, IES R305C240046, ARPA-H, the MacArthur Foundation, Schmidt Sciences, OpenAI, and Stanford HAI.

Impact Statement

Our findings have important practical implications for the deployment of large language models, as they can help organizations more accurately forecast compute requirements and make informed trade-offs between model size, inference costs, and performance targets. The mathematical framework we develop could also generalize beyond language models to other domains where similar scaling phenomena emerge. While our work is primarily theoretical, we acknowledge that advances in language model capabilities can have broad societal impacts. We hope that better understanding these fundamental scaling behaviors will help the research community develop more efficient and reliable AI systems.

References

Agarwal, R., Singh, A., Zhang, L. M., Bohnet, B., Rosias, L., Chan, S. C., Zhang, B., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J. D., Chu, E., Behbahani, F., Faust, A., and Larochelle, H. Many-shot in-context learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=AB6XpMzvqH.

Aghajanyan, A., Yu, L., Conneau, A., Hsu, W.-N., Hambardzumyan, K., Zhang, S., Roller, S., Goyal, N., Levy, O., and Zettlemoyer, L. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pp. 265–279. PMLR, 2023.

Alabdulmohsin, I., Neyshabur, B., and Zhai, X. Revisiting neural scaling laws in language and vision, 2022a. URL https://arxiv.org/abs/2209.06640.

Alabdulmohsin, I. M., Neyshabur, B., and Zhai, X. Revisiting neural scaling laws in language and vision. *Advances in Neural Information Processing Systems*, 35:22300–22312, 2022b.

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., Schuett, J., Shavit, Y., Siddarth, D., Trager, R., and Wolf, K. Frontier ai regulation: Managing emerging risks to public safety, 2023. URL https://arxiv.org/abs/2307.03718.

Anil, C., DURMUS, E., Rimsky, N., Sharma, M., Benton, J., Kundu, S., Batson, J., Tong, M., Mu, J., Ford, D. J., Mosconi, F., Agrawal, R., Schaeffer, R., Bashkansky, N., Svenningsen, S., Lambert, M., Radhakrishnan, A., Denison, C., Hubinger, E. J., Bai, Y., Bricken, T., Maxwell, T., Schiefer, N., Sully, J., Tamkin, A., Lanham, T., Nguyen, K., Korbak, T., Kaplan, J., Ganguli, D., Bowman, S. R., Perez, E., Grosse, R. B., and Duvenaud, D. Many-shot jailbreaking. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=cw5mgd71jW.

Arora, A., Jurafsky, D., Potts, C., and Goodman, N. D. Bayesian scaling laws for in-context learning, 2024. URL https://arxiv.org/abs/2410.16531.

Atanasov, A., Zavatone-Veth, J. A., and Pehlevan, C. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

Bachmann, G., Anagnostidis, S., and Hofmann, T. Scaling mlps: A tale of inductive bias, 2023. URL https://arxiv.org/abs/2306.13575.

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., and Sharma, U. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

Barkai, N., Seung, H. S., and Sompolinsky, H. Scaling laws in learning of classification tasks. *Physical review letters*, 70(20):3167, 1993.

Besiroglu, T., Emery-Xu, N., and Thompson, N. Economic impacts of ai-augmented r&d. *Research Policy*, 53(7): 105037, 2024a.

Besiroglu, T., Erdil, E., Barnett, M., and You, J. Chinchilla scaling: A replication attempt, 2024b. URL https://arxiv.org/abs/2404.10102.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Bochud, T. and Challet, D. Optimal approximations of power-laws with exponentials, 2006. URL https://arxiv.org/abs/physics/0605149.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Bordelon, B., Atanasov, A., and Pehlevan, C. A dynamical model of neural scaling laws. *arXiv preprint arXiv:2402.01092*, 2024a.

Bordelon, B., Atanasov, A., and Pehlevan, C. How feature learning can improve neural scaling laws. *arXiv preprint arXiv:2409.17858*, 2024b.

Bousquet, O., Hanneke, S., Moran, S., van Handel, R., and Yehudayoff, A. A theory of universal learning, 2020. URL https://arxiv.org/abs/2011.04483.

Brill, A. Neural scaling laws rooted in the data distribution. *arXiv preprint arXiv:2412.07942*, 2024.

Brown, B., Juravsky, J., Ehrlich, R., Clark, R., Le, Q. V., Ré, C., and Mirhoseini, A. Large language monkeys: Scaling inference compute with repeated sampling, 2024. URL https://arxiv.org/abs/2407.21787.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020a.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020b. URL https://arxiv.org/abs/2005.14165.

Caballero, E., Gupta, K., Rish, I., and Krueger, D. Broken neural scaling laws. *arXiv preprint arXiv:2210.14891*, 2022.

Cabannes, V., Dohmatob, E., and Bietti, A. Scaling laws for associative memories, 2024. URL https://arxiv.org/abs/2310.02984.

Chan, S., Santoro, A., Lampinen, A., Wang, J., Singh, A., Richemond, P., McClelland, J., and Hill, F. Data distributional properties drive emergent in-context learning in transformers. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 18878–18891. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/77c6ccacfd9962e2307fc64680fc5ace-Paper-Conference.pdf.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Chen, Y., Pan, X., Li, Y., Ding, B., and Zhou, J. A simple and provable scaling law for the test-time compute of large language models, 2024. URL https://arxiv.org/abs/2411.19477.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Clark, A., de Las Casas, D., Guy, A., Mensch, A., Paganini, M., Hoffmann, J., Damoc, B., Hechtman, B., Cai, T., Borgeaud, S., et al. Unified scaling laws for routed language models. In *International conference on machine learning*, pp. 4057–4086. PMLR, 2022.

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A. P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pp. 7480–7512. PMLR, 2023.

Dettmers, T. and Zettlemoyer, L. The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning*, pp. 7750–7774. PMLR, 2023.

Dohmatob, E., Feng, Y., Yang, P., Charton, F., and Kempe, J. A tale of tails: Model collapse as a change of scaling laws, 2024. URL https://arxiv.org/abs/2402.07043.

Dominguez-Olmedo, R., Dorner, F. E., and Hardt, M. Training on the test task confounds evaluation and emergence, 2024. URL https://arxiv.org/abs/2407.07890.

Elkies, N. D. Is there a way to express an power law decay as a series of exponentials? MathOverflow, 2016. URL https://mathoverflow.net/q/251661. URL:https://mathoverflow.net/q/251661 (version: 2016-10-08).

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. Gpts are gpts: An early look at the labor market impact potential of large language models, 2023. URL https://arxiv.org/abs/2303.10130.

Filipovich, M. J., Cappelli, A., Hesslow, D., and Launay, J. Scaling laws beyond backpropagation, 2022. URL https://arxiv.org/abs/2210.14593.

Gadre, S. Y., Smyrnis, G., Shankar, V., Gururangan, S., Wortsman, M., Shao, R., Mercat, J., Fang, A., Li, J., Keh, S., et al. Language models scale reliably with over-training and on downstream tasks. *arXiv preprint arXiv:2403.08540*, 2024.

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

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., Korbak, T., Agrawal, R., Pai, D., Gromov, A., Roberts, D. A., Yang, D., Donoho, D. L., and Koyejo, S. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data, 2024. URL https://arxiv.org/abs/2404.01413.

Ghorbani, B., Firat, O., Freitag, M., Bapna, A., Krikun, M., Garcia, X., Chelba, C., and Cherry, C. Scaling laws for neural machine translation. In *International Conference on Learning Representations*, 2021.

Gordon, M. A., Duh, K., and Kaplan, J. Data and parameter scaling laws for neural machine translation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5915–5922, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.478. URL https://aclanthology.org/2021.emnlp-main.478.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H.,

Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Hernandez, D., Kaplan, J., Henighan, T., and McCandlish, S. Scaling laws for transfer, 2021. URL https://arxiv.org/abs/2102.01293.

Hernandez, D., Brown, T., Conerly, T., DasSarma, N., Drain, D., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Henighan, T., Hume, T., et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Hilton, J., Tang, J., and Schulman, J. Scaling laws for single-agent reinforcement learning, 2023. URL https://arxiv.org/abs/2301.13442.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Hu, S., Liu, X., Han, X., Zhang, X., He, C., Zhao, W., Lin, Y., Ding, N., Ou, Z., Zeng, G., Liu, Z., and Sun, M. Predicting emergent abilities with infinite resolution evaluation, 2024. URL https://arxiv.org/abs/2310.03262.

Hughes, J., Price, S., Lynch, A., Schaeffer, R., Barez, F., Koyejo, S., Sleight, H., Jones, E., Perez, E., and Sharma, M. Best-of-n jailbreaking, 2024. URL https://arxiv.org/abs/2412.03556.

Hutter, M. Learning curve theory, 2021. URL https://arxiv.org/abs/2102.04074.

Jiang, M., Liu, K. Z., Zhong, M., Schaeffer, R., Ouyang, S., Han, J., and Koyejo, S. Investigating data contamination for pre-training language models, 2024. URL https://arxiv.org/abs/2401.06059.

Jones, A. L. Scaling scaling laws with board games. *arXiv preprint arXiv:2104.03113*, 2021.

Kalajdzievski, D. Scaling laws for forgetting when fine-tuning large language models, 2024. URL https://arxiv.org/abs/2401.05605.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Kazdan, J., Schaeffer, R., Dey, A., Gerstgrasser, M., Rafailov, R., Donoho, D. L., and Koyejo, S. Collapse or thrive? perils and promises of synthetic data in a self-generating world, 2024. URL https://arxiv.org/abs/2410.16713.

Kudugunta, S., Kusupati, A., Dettmers, T., Chen, K., Dhillon, I., Tsvetkov, Y., Hajishirzi, H., Kakade, S., Farhadi, A., Jain, P., et al. Matformer: Nested transformer for elastic inference. *arXiv preprint arXiv:2310.07707*, 2023.

Kulal, S., Pasupat, P., Chandra, K., Lee, M., Padon, O., Aiken, A., and Liang, P. S. Spoc: Search-based pseudocode to code. *Advances in Neural Information Processing Systems*, 32, 2019.

Kumar, T., Ankner, Z., Spector, B. F., Bordelon, B., Muennighoff, N., Paul, M., Pehlevan, C., Ré, C., and Raghunathan, A. Scaling laws for precision. *arXiv preprint arXiv:2411.04330*, 2024.

Liang, Z., He, H., Yang, C., and Dai, B. Scaling laws for diffusion transformers, 2024. URL https://arxiv.org/abs/2410.08184.

Lin, L., Wu, J., Kakade, S. M., Bartlett, P. L., and Lee, J. D. Scaling laws in linear regression: Compute, parameters, and data. *arXiv preprint arXiv:2406.08466*, 2024.

Liu, J., Mao, H., Chen, Z., Zhao, T., Shah, N., and Tang, J. Towards neural scaling laws on graphs, 2024. URL https://arxiv.org/abs/2402.02054.

Maloney, A., Roberts, D. A., and Sully, J. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., and Clark, J. Artificial intelligence index report 2024, 2024. URL https://arxiv.org/abs/2405.19522.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., and Hendrycks, D. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. URL https://arxiv.org/abs/2402.04249.

McKenzie, I. R., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., McLean, E., Kirtland, A., Ross, A., Liu, A., Gritsevskiy, A., Wurgaft, D., Kauffman, D., Recchia, G., Liu, J., Cavanagh, J., Weiss, M., Huang, S., Droid, T. F., Tseng, T., Korbak, T., Shen, X., Zhang, Y., Zhou, Z., Kim, N., Bowman, S. R., and Perez, E. Inverse scaling: When bigger isn't better, 2024. URL https://arxiv.org/abs/2306.09479.

Mei, K., Tu, Z., Delbracio, M., Talebi, H., Patel, V. M., and Milanfar, P. Bigger is not always better: Scaling properties of latent diffusion models, 2024. URL https://arxiv.org/abs/2404.01367.

Mhaskar, H. N. Neural networks for optimal approximation of smooth and analytic functions. *Neural computation*, 8 (1):164–177, 1996.

Michaud, E., Liu, Z., Girit, U., and Tegmark, M. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36, 2024.

mpmath development team, T. *mpmath: a Python library for arbitrary-precision floating-point arithmetic (version 1.3.0)*, 2023. http://mpmath.org/.

Muennighoff, N., Rush, A., Barak, B., Le Scao, T., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., and Raffel, C. A. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

Neumann, O. and Gros, C. Scaling laws for a multi-agent reinforcement learning model. *arXiv preprint arXiv:2210.00849*, 2022.

Neumann, O. and Gros, C. Alphazero neural scaling and zipf's law: a tale of board games and power laws, 2024. URL https://arxiv.org/abs/2412.11979.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han,

J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Owen, D. How predictable is language model benchmark performance?, 2024.

Pachocki, J., Tworek, J., Fedus, L., Kaiser, L., Chen, M., Sidor, S., and Zaremba, W. Learning to reason with LLMs. Technical report, OpenAI, September 2024. URL https://openai.com/index/learning-to-reason-with-llms. Contributors include the o1 Contributions team, Core Contributors, and multiple research and safety teams.

Paquette, E., Paquette, C., Xiao, L., and Pennington, J. 4+3 phases of compute-optimal neural scaling laws. *arXiv preprint arXiv:2405.15074*, 2024.

Pinkus, A. Approximation theory of the mlp model in neural networks. *Acta numerica*, 8:143–195, 1999.

Polo, F. M., Somerstep, S., Choshen, L., Sun, Y., and Yurochkin, M. Sloth: scaling laws for llm skills to predict multi-benchmark performance across families, 2024. URL https://arxiv.org/abs/2412.06540.

Porian, T., Wortsman, M., Jitsev, J., Schmidt, L., and Carmon, Y. Resolving discrepancies in compute-optimal scaling of language models, 2024. URL https://arxiv.org/abs/2406.19146.

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P., Anderljung, M., Garfinkel, B., Heim, L., Trask, A., Mukobi, G., Schaeffer, R., Baker, M., Hooker, S., Solaiman, I., Luccioni, A. S., Rajkumar, N., Moës, N., Ladish, J., Guha, N., Newman, J., Bengio, Y., South, T., Pentland, A., Koyejo, S., Kochenderfer, M. J., and Trager, R. Open problems in technical ai governance, 2024. URL https://arxiv.org/abs/2407.14981.

Roberts, D. A., Yaida, S., and Hanin, B. *The principles of deep learning theory*, volume 46. Cambridge University Press Cambridge, MA, USA, 2022.

Romani, S., Pinkoviezky, I., Rubin, A., and Tsodyks, M. Scaling laws of associative memory retrieval. *Neural computation*, 25(10):2523–2544, 2013.

Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y., and Shavit, N. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations*, 2020.

Rosenfeld, J. S., Frankle, J., Carbin, M., and Shavit, N. On the predictability of pruning across scales. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9075–9083. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/rosenfeld21a.html.

Ruan, Y., Maddison, C. J., and Hashimoto, T. Observational scaling laws and the predictability of language model performance, 2024. URL https://arxiv.org/abs/2405.10938.

Sardana, N., Portes, J., Doubov, S., and Frankle, J. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. In *Forty-first International Conference on Machine Learning*, 2023.

Schaeffer, R. Pretraining on the test set is all you need, 2023. URL https://arxiv.org/abs/2309.08632.

Schaeffer, R., Miranda, B., and Koyejo, S. Are emergent abilities of large language models a mirage? In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 55565–55581. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/adc98a266f45005c403b8311ca7e8bd7-Paper-Conference.pdf.

Schaeffer, R., Lecomte, V., Pai, D. B., Carranza, A., Isik, B., Unell, A., Khona, M., Yerxa, T., LeCun, Y., Chung, S., Gromov, A., Shwartz-Ziv, R., and Koyejo, S. Towards an improved understanding and utilization of maximum manifold capacity representations, 2024a. URL https://arxiv.org/abs/2406.09366.

Schaeffer, R., Schoelkopf, H., Miranda, B., Mukobi, G., Madan, V., Ibrahim, A., Bradley, H., Biderman, S., and Koyejo, S. Why has predicting downstream capabilities of frontier ai models with scale remained elusive?, 2024b. URL https://arxiv.org/abs/2406.04391.

Schaeffer, R., Zahedi, N., Khona, M., Pai, D., Truong, S., Du, Y., Ostrow, M., Chandra, S., Carranza, A., Fiete, I. R., Gromov, A., and Koyejo, S. Bridging associative memory and probabilistic modeling, 2024c. URL https://arxiv.org/abs/2402.10202.

Sharma, U. and Kaplan, J. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022.

Snell, C., Lee, J., Xu, K., and Kumar, A. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024a.

Snell, C., Wallace, E., Klein, D., and Levine, S. Predicting emergent capabilities by finetuning, 2024b. URL https://arxiv.org/abs/2411.16035.

Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., and Morcos, A. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Spigler, S., Geiger, M., and Wyart, M. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, December 2020. ISSN 1742-5468. doi: 10.1088/1742-5468/abc61d. URL http://dx.doi.org/10.1088/1742-5468/abc61d.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A., La, A., Lampinen, A., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ramírez, C. F., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodola, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G., Jaimovitch-López, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H., Schütze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocoń, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K. D., Gimpel, K., Omondi, K., Mathewson, K., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Colón, L. O., Metz, L., Şenel, L. K.,

Bosma, M., Sap, M., ter Hoeve, M., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Quintana, M. J. R., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N. A., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P., Eckersley, P., Htut, P. M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., LeBras, R., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R., Lee, R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Shyamolima, Debnath, Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S. T., Shieber, S. M., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V., Prabhu, V. U., Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z., and Wu, Z. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023. URL https://arxiv.org/abs/2206.04615.

Sun, X., Li, S., Xie, R., Han, W., Wu, K., Yang, Z., Li, Y., Wang, A., Li, S., Xue, J., Cheng, Y., Tao, Y., Kang, Z., Xu, C., Wang, D., and Jiang, J. Scaling laws for floating point quantization training, 2025. URL https://arxiv.org/abs/2501.02423.

Tao, C., Liu, Q., Dou, L., Muennighoff, N., Wan, Z., Luo, P., Lin, M., and Wong, N. Scaling laws with vocabulary: Larger models deserve larger vocabularies. *arXiv preprint arXiv:2407.13623*, 2024.

Tay, Y., Dehghani, M., Rao, J., Fedus, W., Abnar, S., Chung, H. W., Narang, S., Yogatama, D., Vaswani, A., and Metzler, D. Scale efficiently: Insights from pre-training and fine-tuning transformers. *arXiv preprint arXiv:2109.10686*, 2021.

Tay, Y., Dehghani, M., Abnar, S., Chung, H. W., Fedus, W., Rao, J., Narang, S., Tran, V. Q., Yogatama, D., and Metzler, D. Scaling laws vs model architectures: How does inductive bias influence scaling? In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2022a.

Tay, Y., Wei, J., Chung, H. W., Tran, V. Q., So, D. R., Shakeri, S., Garcia, X., Zheng, H. S., Rao, J., Chowdhery, A., Zhou, D., Metzler, D., Petrov, S., Houlsby, N., Le, Q. V., and Dehghani, M. Transcending scaling laws with 0.1 URL https://arxiv.org/abs/2210.11399.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., Molloy, J., Isard, M., Barham, P. R., Hennigan, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Doherty, R., Collins, E., Meyer, C., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Krawczyk, J., Du, C., Chi, E., Cheng, H.-T., Ni, E., Shah, P., Kane, P., Chan, B., Faruqui, M., Severyn, A., Lin, H., Li, Y., Cheng, Y., Ittycheriah, A., Mahdieh, M., Chen, M., Sun, P., Tran, D., Bagri, S., Lakshminarayanan, B., Liu, J., Orban, A., Güra, F., Zhou, H., Song, X., Boffy, A., Ganapathy, H., Zheng, S., Choe, H., Ágoston Weisz, Zhu, T., Lu, Y., Gopal, S., Kahn, J., Kula, M., Pitman, J., Shah, R., Taropa, E., Merey, M. A., Baeuml, M., Chen, Z., Shafey, L. E., Zhang, Y., Sercinoglu, O., Tucker, G., Piqueras, E., Krikun, M., Barr, I., Savinov, N., Danihelka, I., Roelofs, B., White, A., Andreassen, A., von Glehn, T., Yagati, L., Kazemi, M., Gonzalez, L., Khalman, M., Sygnowski, J., Frechette, A., Smith, C., Culp, L., Proleev, L., Luan, Y., Chen, X., Lottes, J., Schucher, N., Lebron, F., Rrustemi, A., Clay, N., Crone, P., Kocisky, T., Zhao, J., Perz, B., Yu, D., Howard, H., Bloniarz, A., Rae, J. W., Lu, H., Sifre, L., Maggioni, M., Alcober, F., Garrette, D., Barnes, M., Thakoor, S., Austin, J., Barth-Maron, G., Wong, W., Joshi, R., Chaabouni, R., Fatiha, D., Ahuja, A., Tomar, G. S., Senter, E., Chadwick, M., Kornakov, I., Attaluri, N., Iturrate, I., Liu, R., Li, Y., Cogan, S.,

Chen, J., Jia, C., Gu, C., Zhang, Q., Grimstad, J., Hartman, A. J., Garcia, X., Pillai, T. S., Devlin, J., Laskin, M., de Las Casas, D., Valter, D., Tao, C., Blanco, L., Badia, A. P., Reitter, D., Chen, M., Brennan, J., Rivera, C., Brin, S., Iqbal, S., Surita, G., Labanowski, J., Rao, A., Winkler, S., Parisotto, E., Gu, Y., Olszewska, K., Addanki, R., Miech, A., Louis, A., Teplyashin, D., Brown, G., Catt, E., Balaguer, J., Xiang, J., Wang, P., Ashwood, Z., Briukhov, A., Webson, A., Ganapathy, S., Sanghavi, S., Kannan, A., Chang, M.-W., Stjerngren, A., Djolonga, J., Sun, Y., Bapna, A., Aitchison, M., Pejman, P., Michalewski, H., Yu, T., Wang, C., Love, J., Ahn, J., Bloxwich, D., Han, K., Humphreys, P., Sellam, T., Bradbury, J., Godbole, V., Samangooei, S., Damoc, B., Kaskasoli, A., Arnold, S. M. R., Vasudevan, V., Agrawal, S., Riesa, J., Lepikhin, D., Tanburn, R., Srinivasan, S., Lim, H., Hodkinson, S., Shyam, P., Ferret, J., Hand, S., Garg, A., Paine, T. L., Li, J., Li, Y., Giang, M., Neitz, A., Abbas, Z., York, S., Reid, M., Cole, E., Chowdhery, A., Das, D., Rogozińska, D., Nikolaev, V., Sprechmann, P., Nado, Z., Zilka, L., Prost, F., He, L., Monteiro, M., Mishra, G., Welty, C., Newlan, J., Jia, D., Allamanis, M., Hu, C. H., de Liedekerke, R., Gilmer, J., Saroufim, C., Rijhwani, S., Hou, S., Shrivastava, D., Baddepudi, A., Goldin, A., Ozturel, A., Cassirer, A., Xu, Y., Sohn, D., Sachan, D., Amplayo, R. K., Swanson, C., Petrova, D., Narayan, S., Guez, A., Brahma, S., Landon, J., Patel, M., Zhao, R., Villela, K., Wang, L., Jia, W., Rahtz, M., Giménez, M., Yeung, L., Keeling, J., Georgiev, P., Mincu, D., Wu, B., Haykal, S., Saputro, R., Vodrahalli, K., Qin, J., Cankara, Z., Sharma, A., Fernando, N., Hawkins, W., Neyshabur, B., Kim, S., Hutter, A., Agrawal, P., Castro-Ros, A., van den Driessche, G., Wang, T., Yang, F., yiin Chang, S., Komarek, P., McIlroy, R., Lučić, M., Zhang, G., Farhan, W., Sharman, M., Natsev, P., Michel, P., Bansal, Y., Qiao, S., Cao, K., Shakeri, S., Butterfield, C., Chung, J., Rubenstein, P. K., Agrawal, S., Mensch, A., Soparkar, K., Lenc, K., Chung, T., Pope, A., Maggiore, L., Kay, J., Jhakra, P., Wang, S., Maynez, J., Phuong, M., Tobin, T., Tacchetti, A., Trebacz, M., Robinson, K., Katariya, Y., Riedel, S., Bailey, P., Xiao, K., Ghelani, N., Aroyo, L., Slone, A., Houlsby, N., Xiong, X., Yang, Z., Gribovskaya, E., Adler, J., Wirth, M., Lee, L., Li, M., Kagohara, T., Pavagadhi, J., Bridgers, S., Bortsova, A., Ghemawat, S., Ahmed, Z., Liu, T., Powell, R., Bolina, V., Iinuma, M., Zablotskaia, P., Besley, J., Chung, D.-W., Dozat, T., Comanescu, R., Si, X., Greer, J., Su, G., Polacek, M., Kaufman, R. L., Tokumine, S., Hu, H., Buchatskaya, E., Miao, Y., Elhawaty, M., Siddhant, A., Tomasev, N., Xing, J., Greer, C., Miller, H., Ashraf, S., Roy, A., Zhang, Z., Ma, A., Filos, A., Besta, M., Blevins, R., Klimenko, T., Yeh, C.-K., Changpinyo, S., Mu, J., Chang, O., Pajarskas, M., Muir, C., Cohen, V., Lan, C. L., Haridasan, K., Marathe, A., Hansen, S., Douglas, S., Samuel, R., Wang, M., Austin, S., Lan, C., Jiang, J., Chiu, J., Lorenzo, J. A., Sjösund, L. L., Cevey, S., Gleicher, Z., Avrahami, T., Boral, A., Srinivasan, H., Selo, V., May, R., Aisopos, K., Hussenot, L., Soares, L. B., Baumli, K., Chang, M. B., Recasens, A., Caine, B., Pritzel, A., Pavetic, F., Pardo, F., Gergely, A., Frye, J., Ramasesh, V., Horgan, D., Badola, K., Kassner, N., Roy, S., Dyer, E., Campos, V. C., Tomala, A., Tang, Y., Badawy, D. E., White, E., Mustafa, B., Lang, O., Jindal, A., Vikram, S., Gong, Z., Caelles, S., Hemsley, R., Thornton, G., Feng, F., Stokowiec, W., Zheng, C., Thacker, P., Çağlar Ünlü, Zhang, Z., Saleh, M., Svensson, J., Bileschi, M., Patil, P., Anand, A., Ring, R., Tsihlas, K., Vezer, A., Selvi, M., Shevlane, T., Rodriguez, M., Kwiatkowski, T., Daruki, S., Rong, K., Dafoe, A., FitzGerald, N., Gu-Lemberg, K., Khan, M., Hendricks, L. A., Pellat, M., Feinberg, V., Cobon-Kerr, J., Sainath, T., Rauh, M., Hashemi, S. H., Ives, R., Hasson, Y., Noland, E., Cao, Y., Byrd, N., Hou, L., Wang, Q., Sottiaux, T., Paganini, M., Lespiau, J.-B., Moufarek, A., Hassan, S., Shivakumar, K., van Amersfoort, J., Mandhane, A., Joshi, P., Goyal, A., Tung, M., Brock, A., Sheahan, H., Misra, V., Li, C., Rakićević, N., Dehghani, M., Liu, F., Mittal, S., Oh, J., Noury, S., Sezener, E., Huot, F., Lamm, M., Cao, N. D., Chen, C., Mudgal, S., Stella, R., Brooks, K., Vasudevan, G., Liu, C., Chain, M., Melinkeri, N., Cohen, A., Wang, V., Seymore, K., Zubkov, S., Goel, R., Yue, S., Krishnakumaran, S., Albert, B., Hurley, N., Sano, M., Mohananey, A., Joughin, J., Filonov, E., Kepa, T., Eldawy, Y., Lim, J., Rishi, R., Badiezadegan, S., Bos, T., Chang, J., Jain, S., Padmanabhan, S. G. S., Puttagunta, S., Krishna, K., Baker, L., Kalb, N., Bedapudi, V., Kurzrok, A., Lei, S., Yu, A., Litvin, O., Zhou, X., Wu, Z., Sobell, S., Siciliano, A., Papir, A., Neale, R., Bragagnolo, J., Toor, T., Chen, T., Anklin, V., Wang, F., Feng, R., Gholami, M., Ling, K., Liu, L., Walter, J., Moghaddam, H., Kishore, A., Adamek, J., Mercado, T., Mallinson, J., Wandekar, S., Cagle, S., Ofek, E., Garrido, G., Lombriser, C., Mukha, M., Sun, B., Mohammad, H. R., Matak, J., Qian, Y., Peswani, V., Janus, P., Yuan, Q., Schelin, L., David, O., Garg, A., He, Y., Duzhyi, O., Älgmyr, A., Lottaz, T., Li, Q., Yadav, V., Xu, L., Chinien, A., Shivanna, R., Chuklin, A., Li, J., Spadine, C., Wolfe, T., Mohamed, K., Das, S., Dai, Z., He, K., von Dincklage, D., Upadhyay, S., Maurya, A., Chi, L., Krause, S., Salama, K., Rabinovitch, P. G., M, P. K. R., Selvan, A., Dektiarev, M., Ghiasi, G., Guven, E., Gupta, H., Liu, B., Sharma, D., Shtacher, I. H., Paul, S., Akerlund, O., Aubet, F.-X., Huang, T., Zhu, C., Zhu, E., Teixeira, E., Fritze, M., Bertolini, F., Marinescu, L.-E., Bölle, M., Paulus, D., Gupta, K., Latkar, T., Chang, M., Sanders, J., Wilson, R., Wu, X., Tan, Y.-X., Thiet, L. N., Doshi, T., Lall, S., Mishra, S., Chen, W., Luong, T., Benjamin, S., Lee, J., Andrejczuk, E., Rabiej, D., Ranjan, V., Styrc, K., Yin, P., Simon, J., Harriott, M. R., Bansal, M., Robsky, A., Bacon, G., Greene, D., Mirylenka, D.,

Zhou, C., Sarvana, O., Goyal, A., Andermatt, S., Siegler, P., Horn, B., Israel, A., Pongetti, F., Chen, C.-W. L., Selvatici, M., Silva, P., Wang, K., Tolins, J., Guu, K., Yogev, R., Cai, X., Agostini, A., Shah, M., Nguyen, H., Donnaile, N. O., Pereira, S., Friso, L., Stambler, A., Kurzrok, A., Kuang, C., Romanikhin, Y., Geller, M., Yan, Z., Jang, K., Lee, C.-C., Fica, W., Malmi, E., Tan, Q., Banica, D., Balle, D., Pham, R., Huang, Y., Avram, D., Shi, H., Singh, J., Hidey, C., Ahuja, N., Saxena, P., Dooley, D., Potharaju, S. P., O'Neill, E., Gokulchandran, A., Foley, R., Zhao, K., Dusenberry, M., Liu, Y., Mehta, P., Kotikalapudi, R., Safranek-Shrader, C., Goodman, A., Kessinger, J., Globen, E., Kolhar, P., Gorgolewski, C., Ibrahim, A., Song, Y., Eichenbaum, A., Brovelli, T., Potluri, S., Lahoti, P., Baetu, C., Ghorbani, A., Chen, C., Crawford, A., Pal, S., Sridhar, M., Gurita, P., Mujika, A., Petrovski, I., Cedoz, P.-L., Li, C., Chen, S., Santo, N. D., Goyal, S., Punjabi, J., Kappaganthu, K., Kwak, C., LV, P., Velury, S., Choudhury, H., Hall, J., Shah, P., Figueira, R., Thomas, M., Lu, M., Zhou, T., Kumar, C., Jurdi, T., Chikkerur, S., Ma, Y., Yu, A., Kwak, S., Ähdel, V., Rajayogam, S., Choma, T., Liu, F., Barua, A., Ji, C., Park, J. H., Hellendoorn, V., Bailey, A., Bilal, T., Zhou, H., Khatir, M., Sutton, C., Rzadkowski, W., Macintosh, F., Shagin, K., Medina, P., Liang, C., Zhou, J., Shah, P., Bi, Y., Dankovics, A., Banga, S., Lehmann, S., Bredesen, M., Lin, Z., Hoffmann, J. E., Lai, J., Chung, R., Yang, K., Balani, N., Bražinskas, A., Sozanschi, A., Hayes, M., Alcalde, H. F., Makarov, P., Chen, W., Stella, A., Snijders, L., Mandl, M., Kärrman, A., Nowak, P., Wu, X., Dyck, A., Vaidyanathan, K., R, R., Mallet, J., Rudominer, M., Johnston, E., Mittal, S., Udathu, A., Christensen, J., Verma, V., Irving, Z., Santucci, A., Elsayed, G., Davoodi, E., Georgiev, M., Tenney, I., Hua, N., Cideron, G., Leurent, E., Alnahlawi, M., Georgescu, I., Wei, N., Zheng, I., Scandinaro, D., Jiang, H., Snoek, J., Sundararajan, M., Wang, X., Ontiveros, Z., Karo, I., Cole, J., Rajashekhar, V., Tumeh, L., Ben-David, E., Jain, R., Uesato, J., Datta, R., Bunyan, O., Wu, S., Zhang, J., Stanczyk, P., Zhang, Y., Steiner, D., Naskar, S., Azzam, M., Johnson, M., Paszke, A., Chiu, C.-C., Elias, J. S., Mohiuddin, A., Muhammad, F., Miao, J., Lee, A., Vieillard, N., Park, J., Zhang, J., Stanway, J., Garmon, D., Karmarkar, A., Dong, Z., Lee, J., Kumar, A., Zhou, L., Evens, J., Isaac, W., Irving, G., Loper, E., Fink, M., Arkatkar, I., Chen, N., Shafran, I., Petrychenko, I., Chen, Z., Jia, J., Levskaya, A., Zhu, Z., Grabowski, P., Mao, Y., Magni, A., Yao, K., Snaider, J., Casagrande, N., Palmer, E., Suganthan, P., Castaño, A., Giannoumis, I., Kim, W., Rybiński, M., Sreevatsa, A., Prendki, J., Soergel, D., Goedeckemeyer, A., Gierke, W., Jafari, M., Gaba, M., Wiesner, J., Wright, D. G., Wei, Y., Vashisht, H., Kulizhskaya, Y., Hoover, J., Le, M., Li, L., Iwuanyanwu, C., Liu, L., Ramirez, K., Khorlin, A., Cui, A., LIN, T., Wu, M., Aguilar, R., Pallo, K., Chakladar, A., Perng, G., Abellan, E. A., Zhang, M., Dasgupta, I., Kushman, N., Penchev, I., Repina, A., Wu, X., van der Weide, T., Ponnapalli, P., Kaplan, C., Simsa, J., Li, S., Dousse, O., Yang, F., Piper, J., Ie, N., Pasumarthi, R., Lintz, N., Vijayakumar, A., Andor, D., Valenzuela, P., Lui, M., Paduraru, C., Peng, D., Lee, K., Zhang, S., Greene, S., Nguyen, D. D., Kurylowicz, P., Hardin, C., Dixon, L., Janzer, L., Choo, K., Feng, Z., Zhang, B., Singhal, A., Du, D., McKinnon, D., Antropova, N., Bolukbasi, T., Keller, O., Reid, D., Finchelstein, D., Raad, M. A., Crocker, R., Hawkins, P., Dadashi, R., Gaffney, C., Franko, K., Bulanova, A., Leblond, R., Chung, S., Askham, H., Cobo, L. C., Xu, K., Fischer, F., Xu, J., Sorokin, C., Alberti, C., Lin, C.-C., Evans, C., Dimitriev, A., Forbes, H., Banarse, D., Tung, Z., Omernick, M., Bishop, C., Sterneck, R., Jain, R., Xia, J., Amid, E., Piccinno, F., Wang, X., Banzal, P., Mankowitz, D. J., Polozov, A., Krakovna, V., Brown, S., Bateni, M., Duan, D., Firoiu, V., Thotakuri, M., Natan, T., Geist, M., tan Girgin, S., Li, H., Ye, J., Roval, O., Tojo, R., Kwong, M., Lee-Thorp, J., Yew, C., Sinopalnikov, D., Ramos, S., Mellor, J., Sharma, A., Wu, K., Miller, D., Sonnerat, N., Vnukov, D., Greig, R., Beattie, J., Caveness, E., Bai, L., Eisenschlos, J., Korchemniy, A., Tsai, T., Jasarevic, M., Kong, W., Dao, P., Zheng, Z., Liu, F., Yang, F., Zhu, R., Teh, T. H., Sanmiya, J., Gladchenko, E., Trdin, N., Toyama, D., Rosen, E., Tavakkol, S., Xue, L., Elkind, C., Woodman, O., Carpenter, J., Papamakarios, G., Kemp, R., Kafle, S., Grunina, T., Sinha, R., Talbert, A., Wu, D., Owusu-Afriyie, D., Du, C., Thornton, C., Pont-Tuset, J., Narayana, P., Li, J., Fatehi, S., Wieting, J., Ajmeri, O., Uria, B., Ko, Y., Knight, L., Héliou, A., Niu, N., Gu, S., Pang, C., Li, Y., Levine, N., Stolovich, A., Santamaria-Fernandez, R., Goenka, S., Yustalim, W., Strudel, R., Elqursh, A., Deck, C., Lee, H., Li, Z., Levin, K., Hoffmann, R., Holtmann-Rice, D., Bachem, O., Arora, S., Koh, C., Yeganeh, S. H., Põder, S., Tariq, M., Sun, Y., Ionita, L., Seyedhosseini, M., Tafti, P., Liu, Z., Gulati, A., Liu, J., Ye, X., Chrzaszcz, B., Wang, L., Sethi, N., Li, T., Brown, B., Singh, S., Fan, W., Parisi, A., Stanton, J., Koverkathu, V., Choquette-Choo, C. A., Li, Y., Lu, T., Ittycheriah, A., Shroff, P., Varadarajan, M., Bahargam, S., Willoughby, R., Gaddy, D., Desjardins, G., Cornero, M., Robenek, B., Mittal, B., Albrecht, B., Shenoy, A., Moiseev, F., Jacobsson, H., Ghaffarkhah, A., Rivière, M., Walton, A., Crepy, C., Parrish, A., Zhou, Z., Farabet, C., Radebaugh, C., Srinivasan, P., van der Salm, C., Fidjeland, A., Scellato, S., Latorre-Chimoto, E., Klimczak-Plucińska, H., Bridson, D., de Cesare, D., Hudson, T., Mendolicchio, P., Walker, L., Morris, A., Mauger, M., Guseynov, A., Reid, A., Odoom, S., Loher, L., Cotruta, V., Yenugula, M., Grewe, D., Petrushkina, A., Duerig, T., Sanchez, A., Yadlowsky, S., Shen, A., Globerson, A., Webb, L., Dua, S., Li, D., Bhupatiraju, S., Hurt, D., Qureshi, H., Agarwal, A., Shani, T., Eyal,

M., Khare, A., Belle, S. R., Wang, L., Tekur, C., Kale, M. S., Wei, J., Sang, R., Saeta, B., Liechty, T., Sun, Y., Zhao, Y., Lee, S., Nayak, P., Fritz, D., Vuyyuru, M. R., Aslanides, J., Vyas, N., Wicke, M., Ma, X., Eltyshev, E., Martin, N., Cate, H., Manyika, J., Amiri, K., Kim, Y., Xiong, X., Kang, K., Luisier, F., Tripuraneni, N., Madras, D., Guo, M., Waters, A., Wang, O., Ainslie, J., Baldridge, J., Zhang, H., Pruthi, G., Bauer, J., Yang, F., Mansour, R., Gelman, J., Xu, Y., Polovets, G., Liu, J., Cai, H., Chen, W., Sheng, X., Xue, E., Ozair, S., Angermueller, C., Li, X., Sinha, A., Wang, W., Wiesinger, J., Koukoumidis, E., Tian, Y., Iyer, A., Gurumurthy, M., Goldenson, M., Shah, P., Blake, M., Yu, H., Urbanowicz, A., Palomaki, J., Fernando, C., Durden, K., Mehta, H., Momchev, N., Rahimtoroghi, E., Georgaki, M., Raul, A., Ruder, S., Redshaw, M., Lee, J., Zhou, D., Jalan, K., Li, D., Hechtman, B., Schuh, P., Nasr, M., Milan, K., Mikulik, V., Franco, J., Green, T., Nguyen, N., Kelley, J., Mahendru, A., Hu, A., Howland, J., Vargas, B., Hui, J., Bansal, K., Rao, V., Ghiya, R., Wang, E., Ye, K., Sarr, J. M., Preston, M. M., Elish, M., Li, S., Kaku, A., Gupta, J., Pasupat, I., Juan, D.-C., Someswar, M., M., T., Chen, X., Amini, A., Fabrikant, A., Chu, E., Dong, X., Muthal, A., Buthpitiya, S., Jauhari, S., Hua, N., Khandelwal, U., Hitron, A., Ren, J., Rinaldi, L., Drath, S., Dabush, A., Jiang, N.-J., Godhia, H., Sachs, U., Chen, A., Fan, Y., Taitelbaum, H., Noga, H., Dai, Z., Wang, J., Liang, C., Hamer, J., Ferng, C.-S., Elkind, C., Atias, A., Lee, P., Listík, V., Carlen, M., van de Kerkhof, J., Pikus, M., Zaher, K., Müller, P., Zykova, S., Stefanec, R., Gatsko, V., Hirnschall, C., Sethi, A., Xu, X. F., Ahuja, C., Tsai, B., Stefanoiu, A., Feng, B., Dhandhania, K., Katyal, M., Gupta, A., Parulekar, A., Pitta, D., Zhao, J., Bhatia, V., Bhavnani, Y., Alhadlaq, O., Li, X., Danenberg, P., Tu, D., Pine, A., Filippova, V., Ghosh, A., Limonchik, B., Urala, B., Lanka, C. K., Clive, D., Sun, Y., Li, E., Wu, H., Hongtongsak, K., Li, I., Thakkar, K., Omarov, K., Majmundar, K., Alverson, M., Kucharski, M., Patel, M., Jain, M., Zabelin, M., Pelagatti, P., Kohli, R., Kumar, S., Kim, J., Sankar, S., Shah, V., Ramachandruni, L., Zeng, X., Bariach, B., Weidinger, L., Vu, T., Andreev, A., He, A., Hui, K., Kashem, S., Subramanya, A., Hsiao, S., Hassabis, D., Kavukcuoglu, K., Sadovsky, A., Le, Q., Strohman, T., Wu, Y., Petrov, S., Dean, J., and Vinyals, O. Gemini: A family of highly capable multimodal models, 2024a. URL https://arxiv.org/abs/2312.11805.

Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., Mariooryad, S., Ding, Y., Geng, X., Alcober, F., Frostig, R., Omernick, M., Walker, L., Paduraru, C., Sorokin, C., Tacchetti, A., Gaffney, C., Daruki, S., Sercinoglu, O., Gleicher, Z., Love, J., Voigtlaender, P., Jain, R., Surita, G., Mohamed, K., Blevins, R., Ahn, J., Zhu, T., Kawintiranon, K., Firat, O., Gu, Y., Zhang, Y., Rahtz, M., Faruqui, M., Clay, N., Gilmer, J., Co-Reyes, J., Penchev, I., Zhu, R., Morioka, N., Hui, K., Haridasan, K., Campos, V., Mahdieh, M., Guo, M., Hassan, S., Kilgour, K., Vezer, A., Cheng, H.-T., de Liedekerke, R., Goyal, S., Barham, P., Strouse, D., Noury, S., Adler, J., Sundararajan, M., Vikram, S., Lepikhin, D., Paganini, M., Garcia, X., Yang, F., Valter, D., Trebacz, M., Vodrahalli, K., Asawaroengchai, C., Ring, R., Kalb, N., Soares, L. B., Brahma, S., Steiner, D., Yu, T., Mentzer, F., He, A., Gonzalez, L., Xu, B., Kaufman, R. L., Shafey, L. E., Oh, J., Hennigan, T., van den Driessche, G., Odoom, S., Lucic, M., Roelofs, B., Lall, S., Marathe, A., Chan, B., Ontanon, S., He, L., Teplyashin, D., Lai, J., Crone, P., Damoc, B., Ho, L., Riedel, S., Lenc, K., Yeh, C.-K., Chowdhery, A., Xu, Y., Kazemi, M., Amid, E., Petrushkina, A., Swersky, K., Khodaei, A., Chen, G., Larkin, C., Pinto, M., Yan, G., Badia, A. P., Patil, P., Hansen, S., Orr, D., Arnold, S. M. R., Grimstad, J., Dai, A., Douglas, S., Sinha, R., Yadav, V., Chen, X., Gribovskaya, E., Austin, J., Zhao, J., Patel, K., Komarek, P., Austin, S., Borgeaud, S., Friso, L., Goyal, A., Caine, B., Cao, K., Chung, D.-W., Lamm, M., Barth-Maron, G., Kagohara, T., Olszewska, K., Chen, M., Shivakumar, K., Agarwal, R., Godhia, H., Rajwar, R., Snaider, J., Dotiwalla, X., Liu, Y., Barua, A., Ungureanu, V., Zhang, Y., Batsaikhan, B.-O., Wirth, M., Qin, J., Danihelka, I., Doshi, T., Chadwick, M., Chen, J., Jain, S., Le, Q., Kar, A., Gurumurthy, M., Li, C., Sang, R., Liu, F., Lamprou, L., Munoz, R., Lintz, N., Mehta, H., Howard, H., Reynolds, M., Aroyo, L., Wang, Q., Blanco, L., Cassirer, A., Griffith, J., Das, D., Lee, S., Sygnowski, J., Fisher, Z., Besley, J., Powell, R., Ahmed, Z., Paulus, D., Reitter, D., Borsos, Z., Joshi, R., Pope, A., Hand, S., Selo, V., Jain, V., Sethi, N., Goel, M., Makino, T., May, R., Yang, Z., Schalkwyk, J., Butterfield, C., Hauth, A., Goldin, A., Hawkins, W., Senter, E., Brin, S., Woodman, O., Ritter, M., Noland, E., Giang, M., Bolina, V., Lee, L., Blyth, T., Mackinnon, I., Reid, M., Sarvana, O., Silver, D., Chen, A., Wang, L., Maggiore, L., Chang, O., Attaluri, N., Thornton, G., Chiu, C.-C., Bunyan, O., Levine, N., Chung, T., Eltyshev, E., Si, X., Lillicrap, T., Brady, D., Aggarwal, V., Wu, B., Xu, Y., McIlroy, R., Badola, K., Sandhu, P., Moreira, E., Stokowiec, W., Hemsley, R., Li, D., Tudor, A., Shyam, P., Rahimtoroghi, E., Haykal, S., Sprechmann, P., Zhou, X., Mincu, D., Li, Y., Addanki, R., Krishna, K., Wu, X., Frechette, A., Eyal, M., Dafoe, A., Lacey, D., Whang, J., Avrahami, T., Zhang, Y., Taropa, E., Lin, H., Toyama, D., Rutherford, E., Sano, M., Choe, H., Tomala, A., Safranek-Shrader, C., Kassner, N., Pajarskas, M., Harvey, M., Sechrist, S., Fortunato, M., Lyu, C., Elsayed, G., Kuang, C., Lottes, J., Chu, E., Jia, C., Chen, C.-W., Humphreys, P., Baumli, K., Tao, C., Samuel, R., dos Santos, C. N., Andreassen, A., Rakićević, N., Grewe, D., Kumar, A., Winkler, S., Caton, J., Brock, A., Dalmia,

S., Sheahan, H., Barr, I., Miao, Y., Natsev, P., Devlin, J., Behbahani, F., Prost, F., Sun, Y., Myaskovsky, A., Pillai, T. S., Hurt, D., Lazaridou, A., Xiong, X., Zheng, C., Pardo, F., Li, X., Horgan, D., Stanton, J., Ambar, M., Xia, F., Lince, A., Wang, M., Mustafa, B., Webson, A., Lee, H., Anil, R., Wicke, M., Dozat, T., Sinha, A., Piqueras, E., Dabir, E., Upadhyay, S., Boral, A., Hendricks, L. A., Fry, C., Djolonga, J., Su, Y., Walker, J., Labanowski, J., Huang, R., Misra, V., Chen, J., Skerry-Ryan, R., Singh, A., Rijhwani, S., Yu, D., Castro-Ros, A., Changpinyo, B., Datta, R., Bagri, S., Hrafnkelsson, A. M., Maggioni, M., Zheng, D., Sulsky, Y., Hou, S., Paine, T. L., Yang, A., Riesa, J., Rogozinska, D., Marcus, D., Badawy, D. E., Zhang, Q., Wang, L., Miller, H., Greer, J., Sjos, L. L., Nova, A., Zen, H., Chaabouni, R., Rosca, M., Jiang, J., Chen, C., Liu, R., Sainath, T., Krikun, M., Polozov, A., Lespiau, J.-B., Newlan, J., Cankara, Z., Kwak, S., Xu, Y., Chen, P., Coenen, A., Meyer, C., Tsihlas, K., Ma, A., Gottweis, J., Xing, J., Gu, C., Miao, J., Frank, C., Cankara, Z., Ganapathy, S., Dasgupta, I., Hughes-Fitt, S., Chen, H., Reid, D., Rong, K., Fan, H., van Amersfoort, J., Zhuang, V., Cohen, A., Gu, S. S., Mohananey, A., Ilic, A., Tobin, T., Wieting, J., Bortsova, A., Thacker, P., Wang, E., Caveness, E., Chiu, J., Sezener, E., Kaskasoli, A., Baker, S., Millican, K., Elhawaty, M., Aisopos, K., Lebsack, C., Byrd, N., Dai, H., Jia, W., Wiethoff, M., Davoodi, E., Weston, A., Yagati, L., Ahuja, A., Gao, I., Pundak, G., Zhang, S., Azzam, M., Sim, K. C., Caelles, S., Keeling, J., Sharma, A., Swing, A., Li, Y., Liu, C., Bostock, C. G., Bansal, Y., Nado, Z., Anand, A., Lipschultz, J., Karmarkar, A., Proleev, L., Ittycheriah, A., Yeganeh, S. H., Polovets, G., Faust, A., Sun, J., Rrustemi, A., Li, P., Shivanna, R., Liu, J., Welty, C., Lebron, F., Baddepudi, A., Krause, S., Parisotto, E., Soricut, R., Xu, Z., Bloxwich, D., Johnson, M., Neyshabur, B., Mao-Jones, J., Wang, R., Ramasesh, V., Abbas, Z., Guez, A., Segal, C., Nguyen, D. D., Svensson, J., Hou, L., York, S., Milan, K., Bridgers, S., Gworek, W., Tagliasacchi, M., Lee-Thorp, J., Chang, M., Guseynov, A., Hartman, A. J., Kwong, M., Zhao, R., Kashem, S., Cole, E., Miech, A., Tanburn, R., Phuong, M., Pavetic, F., Cevey, S., Comanescu, R., Ives, R., Yang, S., Du, C., Li, B., Zhang, Z., Iinuma, M., Hu, C. H., Roy, A., Bijwadia, S., Zhu, Z., Martins, D., Saputro, R., Gergely, A., Zheng, S., Jia, D., Antonoglou, I., Sadovsky, A., Gu, S., Bi, Y., Andreev, A., Samangooei, S., Khan, M., Kocisky, T., Filos, A., Kumar, C., Bishop, C., Yu, A., Hodkinson, S., Mittal, S., Shah, P., Moufarek, A., Cheng, Y., Bloniarz, A., Lee, J., Pejman, P., Michel, P., Spencer, S., Feinberg, V., Xiong, X., Savinov, N., Smith, C., Shakeri, S., Tran, D., Chesus, M., Bohnet, B., Tucker, G., von Glehn, T., Muir, C., Mao, Y., Kazawa, H., Slone, A., Soparkar, K., Shrivastava, D., Cobon-Kerr, J., Sharman, M., Pavagadhi, J., Araya, C., Misiunas, K., Ghelani, N., Laskin, M., Barker, D., Li, Q., Briukhov, A., Houlsby, N., Glaese, M., Lakshminarayanan, B., Schucher, N., Tang, Y., Collins, E., Lim, H., Feng, F., Recasens, A., Lai, G., Magni, A., Cao, N. D., Siddhant, A., Ashwood, Z., Orbay, J., Dehghani, M., Brennan, J., He, Y., Xu, K., Gao, Y., Saroufim, C., Molloy, J., Wu, X., Arnold, S., Chang, S., Schrittwieser, J., Buchatskaya, E., Radpour, S., Polacek, M., Giordano, S., Bapna, A., Tokumine, S., Hellendoorn, V., Sottiaux, T., Cogan, S., Severyn, A., Saleh, M., Thakoor, S., Shefey, L., Qiao, S., Gaba, M., yiin Chang, S., Swanson, C., Zhang, B., Lee, B., Rubenstein, P. K., Song, G., Kwiatkowski, T., Koop, A., Kannan, A., Kao, D., Schuh, P., Stjerngren, A., Ghiasi, G., Gibson, G., Vilnis, L., Yuan, Y., Ferreira, F. T., Kamath, A., Klimenko, T., Franko, K., Xiao, K., Bhattacharya, I., Patel, M., Wang, R., Morris, A., Strudel, R., Sharma, V., Choy, P., Hashemi, S. H., Landon, J., Finkelstein, M., Jhakra, P., Frye, J., Barnes, M., Mauger, M., Daun, D., Baatarsukh, K., Tung, M., Farhan, W., Michalewski, H., Viola, F., de Chaumont Quitry, F., Lan, C. L., Hudson, T., Wang, Q., Fischer, F., Zheng, I., White, E., Dragan, A., baptiste Alayrac, J., Ni, E., Pritzel, A., Iwanicki, A., Isard, M., Bulanova, A., Zilka, L., Dyer, E., Sachan, D., Srinivasan, S., Muckenhirn, H., Cai, H., Mandhane, A., Tariq, M., Rae, J. W., Wang, G., Ayoub, K., FitzGerald, N., Zhao, Y., Han, W., Alberti, C., Garrette, D., Krishnakumar, K., Gimenez, M., Levskaya, A., Sohn, D., Matak, J., Iturrate, I., Chang, M. B., Xiang, J., Cao, Y., Ranka, N., Brown, G., Hutter, A., Mirrokni, V., Chen, N., Yao, K., Egyed, Z., Galilee, F., Liechty, T., Kallakuri, P., Palmer, E., Ghemawat, S., Liu, J., Tao, D., Thornton, C., Green, T., Jasarevic, M., Lin, S., Cotruta, V., Tan, Y.-X., Fiedel, N., Yu, H., Chi, E., Neitz, A., Heitkaemper, J., Sinha, A., Zhou, D., Sun, Y., Kaed, C., Hulse, B., Mishra, S., Georgaki, M., Kudugunta, S., Farabet, C., Shafran, I., Vlasic, D., Tsitsulin, A., Ananthanarayanan, R., Carin, A., Su, G., Sun, P., V, S., Carvajal, G., Broder, J., Comsa, I., Repina, A., Wong, W., Chen, W. W., Hawkins, P., Filonov, E., Loher, L., Hirnschall, C., Wang, W., Ye, J., Burns, A., Cate, H., Wright, D. G., Piccinini, F., Zhang, L., Lin, C.-C., Gog, I., Kulizhskaya, Y., Sreevatsa, A., Song, S., Cobo, L. C., Iyer, A., Tekur, C., Garrido, G., Xiao, Z., Kemp, R., Zheng, H. S., Li, H., Agarwal, A., Ngani, C., Goshvadi, K., Santamaria-Fernandez, R., Fica, W., Chen, X., Gorgolewski, C., Sun, S., Garg, R., Ye, X., Eslami, S. M. A., Hua, N., Simon, J., Joshi, P., Kim, Y., Tenney, I., Potluri, S., Thiet, L. N., Yuan, Q., Luisier, F., Chronopoulou, A., Scellato, S., Srinivasan, P., Chen, M., Koverkathu, V., Dalibard, V., Xu, Y., Saeta, B., Anderson, K., Sellam, T., Fernando, N., Huot, F., Jung, J., Varadarajan, M., Quinn, M., Raul, A., Le, M., Habalov, R., Clark, J., Jalan, K., Bullard, K., Singhal, A., Luong, T., Wang, B., Rajayogam, S., Eisenschlos, J., Jia, J., Finchelstein, D., Yakubovich, A., Balle, D., Fink, M., Agarwal, S., Li, J., Dvijotham, D., Pal, S., Kang, K.,

Konzelmann, J., Beattie, J., Dousse, O., Wu, D., Crocker, R., Elkind, C., Jonnalagadda, S. R., Lee, J., Holtmann-Rice, D., Kallarackal, K., Liu, R., Vnukov, D., Vats, N., Invernizzi, L., Jafari, M., Zhou, H., Taylor, L., Prendki, J., Wu, M., Eccles, T., Liu, T., Kopparapu, K., Beaufays, F., Angermueller, C., Marzoca, A., Sarcar, S., Dib, H., Stanway, J., Perbet, F., Trdin, N., Sterneck, R., Khorlin, A., Li, D., Wu, X., Goenka, S., Madras, D., Goldshtein, S., Gierke, W., Zhou, T., Liu, Y., Liang, Y., White, A., Li, Y., Singh, S., Bahargam, S., Epstein, M., Basu, S., Lao, L., Ozturel, A., Crous, C., Zhai, A., Lu, H., Tung, Z., Gaur, N., Walton, A., Dixon, L., Zhang, M., Globerson, A., Uy, G., Bolt, A., Wiles, O., Nasr, M., Shumailov, I., Selvi, M., Piccinno, F., Aguilar, R., McCarthy, S., Khalman, M., Shukla, M., Galic, V., Carpenter, J., Villela, K., Zhang, H., Richardson, H., Martens, J., Bosnjak, M., Belle, S. R., Seibert, J., Alnahlawi, M., McWilliams, B., Singh, S., Louis, A., Ding, W., Popovici, D., Simicich, L., Knight, L., Mehta, P., Gupta, N., Shi, C., Fatehi, S., Mitrovic, J., Grills, A., Pagadora, J., Munkhdalai, T., Petrova, D., Eisenbud, D., Zhang, Z., Yates, D., Mittal, B., Tripuraneni, N., Assael, Y., Brovelli, T., Jain, P., Velimirovic, M., Akbulut, C., Mu, J., Macherey, W., Kumar, R., Xu, J., Qureshi, H., Comanici, G., Wiesner, J., Gong, Z., Ruddock, A., Bauer, M., Felt, N., GP, A., Arnab, A., Zelle, D., Rothfuss, J., Rosgen, B., Shenoy, A., Seybold, B., Li, X., Mudigonda, J., Erdogan, G., Xia, J., Simsa, J., Michi, A., Yao, Y., Yew, C., Kan, S., Caswell, I., Radebaugh, C., Elisseeff, A., Valenzuela, P., McKinney, K., Paterson, K., Cui, A., Latorre-Chimoto, E., Kim, S., Zeng, W., Durden, K., Ponnapalli, P., Sosea, T., Choquette-Choo, C. A., Manyika, J., Robenek, B., Vashisht, H., Pereira, S., Lam, H., Velic, M., Owusu-Afriyie, D., Lee, K., Bolukbasi, T., Parrish, A., Lu, S., Park, J., Venkatraman, B., Talbert, A., Rosique, L., Cheng, Y., Sozanschi, A., Paszke, A., Kumar, P., Austin, J., Li, L., Salama, K., Perz, B., Kim, W., Dukkipati, N., Baryshnikov, A., Kaplanis, C., Sheng, X., Chervonyi, Y., Unlu, C., de Las Casas, D., Askham, H., Tunyasuvunakool, K., Gimeno, F., Poder, S., Kwak, C., Miecnikowski, M., Mirrokni, V., Dimitriev, A., Parisi, A., Liu, D., Tsai, T., Shevlane, T., Kouridi, C., Garmon, D., Goedeckemeyer, A., Brown, A. R., Vijayakumar, A., Elqursh, A., Jazayeri, S., Huang, J., Carthy, S. M., Hoover, J., Kim, L., Kumar, S., Chen, W., Biles, C., Bingham, G., Rosen, E., Wang, L., Tan, Q., Engel, D., Pongetti, F., de Cesare, D., Hwang, D., Yu, L., Pullman, J., Narayanan, S., Levin, K., Gopal, S., Li, M., Aharoni, A., Trinh, T., Lo, J., Casagrande, N., Vij, R., Matthey, L., Ramadhana, B., Matthews, A., Carey, C., Johnson, M., Goranova, K., Shah, R., Ashraf, S., Dasgupta, K., Larsen, R., Wang, Y., Vuyyuru, M. R., Jiang, C., Ijazi, J., Osawa, K., Smith, C., Boppana, R. S., Bilal, T., Koizumi, Y., Xu, Y., Altun, Y., Shabat, N., Bariach, B., Korchemniy, A., Choo, K., Ronneberger, O., Iwuanyanwu, C., Zhao, S., Soergel, D., Hsieh, C.-J., Cai, I., Iqbal, S., Sundermeyer, M., Chen, Z., Bursztein, E., Malaviya, C., Biadsy, F., Shroff, P., Dhillon, I., Latkar, T., Dyer, C., Forbes, H., Nicosia, M., Nikolaev, V., Greene, S., Georgiev, M., Wang, P., Martin, N., Sedghi, H., Zhang, J., Banzal, P., Fritz, D., Rao, V., Wang, X., Zhang, J., Patraucean, V., Du, D., Mordatch, I., Jurin, I., Liu, L., Dubey, A., Mohan, A., Nowakowski, J., Ion, V.-D., Wei, N., Tojo, R., Raad, M. A., Hudson, D. A., Keshava, V., Agrawal, S., Ramirez, K., Wu, Z., Nguyen, H., Liu, J., Sewak, M., Petrini, B., Choi, D., Philips, I., Wang, Z., Bica, I., Garg, A., Wilkiewicz, J., Agrawal, P., Li, X., Guo, D., Xue, E., Shaik, N., Leach, A., Khan, S. M., Wiesinger, J., Jerome, S., Chakladar, A., Wang, A. W., Ornduff, T., Abu, F., Ghaffarkhah, A., Wainwright, M., Cortes, M., Liu, F., Maynez, J., Terzis, A., Samangouei, P., Mansour, R., Kepa, T., Aubet, F.-X., Algymr, A., Banica, D., Weisz, A., Orban, A., Senges, A., Andrejczuk, E., Geller, M., Santo, N. D., Anklin, V., Merey, M. A., Baeuml, M., Strohman, T., Bai, J., Petrov, S., Wu, Y., Hassabis, D., Kavukcuoglu, K., Dean, J., and Vinyals, O. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024b. URL https://arxiv.org/abs/2403.05530.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models, 2022a. URL https://arxiv.org/abs/2206.07682.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022b.

Wu, T.-Y. and Lo, P.-Y. U-shaped and inverted-u scaling behind emergent abilities of large language models, 2024. URL https://arxiv.org/abs/2410.01692.

Wu, Y., Sun, Z., Li, S., Welleck, S., and Yang, Y. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models, 2024. URL https://arxiv.org/abs/2408.00724.

Xiong, W., Liu, J., Molybog, I., Zhang, H., Bhargava, P., Hou, R., Martin, L., Rungta, R., Sankararaman, K. A., Oguz, B., Khabsa, M., Fang, H., Mehdad, Y., Narang, S., Malik, K., Fan, A., Bhosale, S., Edunov, S., Lewis, M., Wang, S., and Ma, H. Effective long-context scaling

of foundation models, 2023. URL https://arxiv.org/abs/2309.16039.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12104–12113, 2022.

Zhang, B., Liu, Z., Cherry, C., and Firat, O. When scaling meets llm finetuning: The effect of data, model and finetuning method, 2024. URL https://arxiv.org/abs/2402.17193.

A. Clarification of How Large Language Monkeys and Best-of-N Jailbreaking Sampled Data

In this manuscript, we used the phrasing of "independent attempts," which is not fully correct. In this appendix section, we clarify why we chose this terminology, what likely impacts we believe this inaccuracy may have had on our results, and how to correct the paper accordingly.

Large Language Monkeys (Brown et al., 2024) indeed drew $10,000$ independent attempts per problem, but Best-of-N Jailbreaking (Hughes et al., 2024) sampled data slightly different: for each problem, jailbreaking attempts were drawn until either a successful jailbreak was obtained or until a maximum limit of $10,000$ attempts was hit. Samples were also drawn in minibatches of size 60, making the (in)dependence of samples a bit tricky.

We omitted this nuance because it offers a second-order correction to our paper's main story while offering little additional insight. Neither of our theorems and none of our main text figures change. We suspect that this slightly different sampling procedure explains why, in Fig. 6, the estimated power law exponents between the least squares power law estimator and the distributional power law estimator deviate more significantly from identity for Best-of-N Jailbreaking than for Large Language Monkeys. A natural way to correct for this is to use a beta-negative binomial distribution rather than a beta-binomial distribution, with an additional correction for the maximum number of attempts. For more information, please see Appendix H.

B. Estimating Success Rates Using Chen et al. (2021)'s Estimator

In this manuscript, we defined $\text{pass}_\text{i}@\text{k}$ and $\text{ASR}_\text{i}@\text{k}$ as:

$$\text{pass}_\text{i}@\text{k} \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{k \text{ Attempts}} \left[\mathbb{I}[\text{At least 1 attempt by the model solves the } i\text{-th problem}]\right]$$

$$\text{ASR}_\text{i}@\text{k} \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{k \text{ Attempts}} \left[\mathbb{I}[\text{At least 1 attempt jailbreaks the model on the } i\text{-th prompt}]\right]$$

Throughout this manuscript, to estimate $\text{pass}_\text{i}@\text{k}$ and $\text{ASR}@\text{k}$, we used the unbiased and lower variance estimator introduced by Chen et al. (2021): for the i-th problem, we sampled $n \gg k$ attempts per problem, counted the number of successful attempts c, and then swept k to compute an estimate of $\text{pass}_\text{i}@\text{k}$ for different k values:

$$\widehat{\text{pass}_\text{i}@\text{k}} = 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}} \tag{12}$$

Two comments: Firstly, n as used here has no relationship with the number of problems in the benchmark (Sec. 1), and secondly, our notation differs slightly from that of Chen et al. (2021), but the ideas are consistent. A numerically stable Python implementation of the estimator is provided in Fig. 8:

```
def estimate_success_rate_at_k_per_problem(n: int, c: int, k: int) -> float:
    """
    :param n: number of total attempts on this problem.
    :param c: number of correct attempts on this problem.
    :param k: k in pass_i@$k$.
    """
    if n - c < k: return 1.0
    return 1.0 - np.prod(1.0 - k / np.arange(n - c + 1, n + 1))
```

Figure 8: A numerically stable unbiased estimator of $\text{pass}_\text{i}@\text{k}$, introduced by Chen et al. (2021).

To reiterate a point made by Chen et al. (2021), estimating $\text{pass}_\text{i}@\text{k}$ as $1 - (1 - \widehat{\text{pass}_\text{i}@1})^k$ is biased (Fig. 9).

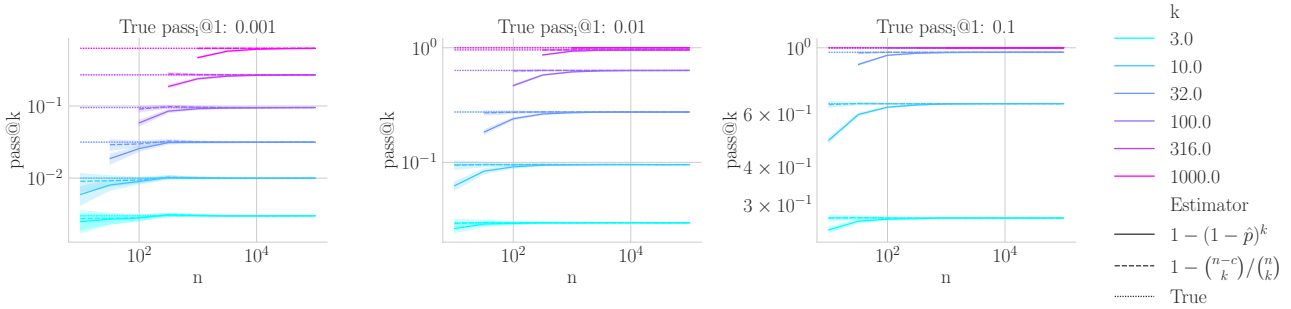

Figure 9: **Bias of Estimators of** $\text{pass}_\text{i}@\text{k}$**.** Numerical simulations show that estimating $\text{pass}_\text{i}@\text{k}$ as $1 - (1 - \widehat{\text{pass}_\text{i}@1})^k$ is biased whereas the estimator of Chen et al. (2021) is not. For a mathematical proof of unbiasedness, see the original paper.

C. Fitting Power Laws to Large Language Monkeys and Best-of-N Jailbreaking

We fit power laws to a subset of data from Large Language Monkeys (Brown et al., 2024) and from Best-of-N Jailbreaking (Hughes et al., 2024), specifically Pythia language models (Biderman et al., 2023) on the MATH benchmark (Hendrycks et al., 2021) and frontier AI models – Claude, GPT4 (OpenAI et al., 2024), Gemini (Team et al., 2024a;b) and Llama 3 (Grattafiori et al., 2024) – on the HarmBench jailbreaking benchmark (Mazeika et al., 2024). We show the functional forms and the fit parameters in Table 1 and Table 2 respectively. To fit the parameters, for Large Language Monkeys, we simply minimized the squared error between the actual and predicted $-\log(\text{pass}_{\mathcal{D}}@\text{k})$, and for Best-of-N Jailbreaking, we similarly minimized the squared error between the actual and predicted $-\log(\text{ASR}_{\mathcal{D}}@\text{k}))$.

Note: Llama 3 8B IT does not exhibit power law scaling under Best-of-N Jailbreaking (shown in Fig. 1, bottom).

Model	Benchmark	a	b
Pythia 70M	MATH	8.026	0.194
Pythia 160M	MATH	6.591	0.280
Pythia 410M	MATH	5.524	0.286
Pythia 1B	MATH	5.452	0.315
Pythia 2.8B	MATH	4.104	0.336
Pythia 6.9B	MATH	4.255	0.348
Pythia 12B	MATH	4.113	0.370

Table 1: Large Language Monkeys (Brown et al., 2024) fitted power law parameters on 128 mathematical problems from MATH (Hendrycks et al., 2021).
Functional Form: $-\log(\text{pass}_{\mathcal{D}}@\text{k}) = a\,k^{-b}$.

Model	Modality	a	b
Claude 3.5 Opus	Text	2.630	0.448
Claude 3.5 Sonnet	Text	3.436	0.312
GPT4o	Text	3.639	0.395
GPT4o Mini	Text	3.637	0.492
Gemini 1.5 Flash	Text	6.158	0.303
Gemini 1.5 Pro	Text	6.296	0.256
Llama 3 8B IT	Text	–	–

Table 2: Best-of-N Jailbreaking (Hughes et al., 2024) fitted power law parameters on text jailbreak prompts from HarmBench (Mazeika et al., 2024).
Functional Form: $-\log(\text{ASR}_{\mathcal{D}}@\text{k}) = a\,k^{-b}$.
Note: Llama 3 8B Instruction Tuned (IT) does not exhibit power law scaling.

D. Mathematical Equivalence Between Coverage and Average Success Rate

Brown et al. (2024) and Hughes et al. (2024) phrase their research in terms of "coverage", defined as the fraction of problems that can be solved or the fraction of prompts that can jailbreak a model, but as Brown et al. (2024) comment and we here derive, the coverage is mathematically equivalent to the average $\text{pass}_\text{i}@\text{k}$ (equivalently, $\text{ASR}@\text{k}$. due to two simple probabilistic primitives: (1) linearity of expectation, (2) the expectation of an indictor random variable of some event is the probability of said event and (3) the definition of $\text{pass}_\text{i}@\text{k}$:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\substack{\text{Prompts}\\\text{Attempts}}}\Big[\text{Coverage}\Big] &\overset{\text{def}}{=} \mathop{\mathbb{E}}_{\substack{\text{Problems}\\\text{Attempts}}}\Big[\text{Fraction of Problems Solved After } k \text{ Attempts}\Big] \\
&= \mathop{\mathbb{E}}_{\text{Problems}}\left[\mathop{\mathbb{E}}_{\text{Attempts}|\text{Problem}}\Big[\mathbb{I}\big[\text{Problem Solved After } k \text{ Attempts}\big]\Big]\right] \\
&= \mathop{\mathbb{E}}_{\text{Problems}}\Big[\text{pass}_{\text{problem}}@\text{k}\Big] \\
&= \text{pass}_{\mathcal{D}}@\text{k}
\end{aligned}
$$

In our work, we prefer phrasing along the lines of "success rate" over "coverage" because success rate avoids coverage's binary implication that each problem/prompt is either "solved" or "not solved".

E. Aggregate Power Laws from a Probability Distribution over Exponential Functions

E.1. Preliminaries: Power Laws from Weighted Exponential Functions

A known result is that power laws can emerge from appropriately weighted sums of exponential functions, e.g., (Bochud & Challet, 2006; Elkies, 2016; Bousquet et al., 2020). For a concrete example with a short proof:

$$x^{-r} = \frac{1}{\Gamma(r)} \int_0^\infty p^{r-1}\, e^{-px}\, dp, \tag{13}$$

where $\Gamma(r) \stackrel{\text{def}}{=} \int_0^\infty s^{r-1}\, e^{-s}\, ds$ is the Gamma function. The proof is via u-substitution $u \stackrel{\text{def}}{=} p\,x$:

$$\frac{1}{\Gamma(r)} \int_0^\infty p^{r-1}\, e^{-px}\, dp = \frac{1}{\Gamma(r)} \int_0^\infty (u/x)^{r-1}\, e^{-u}\, \frac{du}{x} \tag{14}$$

$$= \frac{1}{\Gamma(r)}\, x^{-r} \int_0^\infty u^{r-1}\, e^{-u}\, du \tag{15}$$

$$= \frac{1}{\Gamma(r)}\, x^{-r}\, \Gamma(r) \tag{16}$$

$$= x^{-r} \tag{17}$$

In our particular context, we are interested in the scaling with k of the expected success rate over problems sampled from the benchmark's data distribution:

$$\text{pass}_{\mathcal{D}}@\text{k} \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{\text{pass}_\text{i}@1 \sim \mathcal{D}} \Big[\text{pass}_\text{i}@\text{k} \Big] \tag{18}$$

distribution (over problems in a benchmark) of $\text{pass}_\text{i}@\text{k}$ scores that yields power law scaling with respect to the number of attempts k:

$$-\log \left(\frac{1}{n} \sum_{i=1}^{n} \text{pass}_\text{i}@\text{k} \right) \approx a k^{-b}. \tag{19}$$

for constants $a, b > 0$.

E.2. Delta Distribution: $\text{pass}_\text{i}@1 \sim \delta(p), p \in (0, 1)$

To start with a negative result, we will show that not all distributions of the per-problem success probabilities $\text{pass}_\text{i}@1$ yield aggregate power law scaling. Suppose that the model's $\text{pass}_\text{i}@1$ probabilities across the benchmarks' problems are all exactly $p \in (0, 1)$. For brevity, let $p_i \stackrel{\text{def}}{=} \text{pass}_\text{i}@1$. Then the aggregate success rate is:

$$\mathbb{E}_{p_i \sim \delta(p)}[\text{pass}_\text{i}@\text{k}] = 1 - \mathbb{E}_{p_i}[(1 - p_i)^k] \tag{20}$$

$$= \int_0^1 \delta(p)\, (1 - p_i)^k\, dp_i \tag{21}$$

$$= (1 - p)^k. \tag{22}$$

Recalling that the expansion of $\log(\cdot)$ for small x is $-\log(1 - x) = x + O(x^2)$, in our case, we obtain:

$$-\log \Big(1 - \mathbb{E}_{p_i \sim \delta(p)}[\text{pass}@\text{k}] \Big) = (1 - p)^k + O((1 - p)^{2k}) = (1 - p)^k + o((1 - p)^k). \tag{23}$$

Thus, in the large k regime, we find the negative log aggregate success rate exhibits *exponential* scaling with k as we intuitively expect.

E.3. Uniform Distribution: $\text{pass}_\text{i}@1 \sim \text{Uniform}(\alpha, \beta)$

Suppose $\text{pass}_\text{i}@1$ probabilities follow a uniform distribution $\text{Uniform}(\alpha, \beta)$ where $0 \leq \alpha < \beta \leq 1$. The aggregate success rate after k attempts is defined as:

$$\text{pass}_{\text{Uniform}(\alpha,\beta)}@k \stackrel{\text{def}}{=} 1 - \mathbb{E}\big[(1 - p)^k\big].$$

If $p \sim \text{Uniform}(\alpha, \beta)$, the expectation of $(1-p)^k$ is:

$$\mathbb{E}\big[(1-p)^k\big] \;=\; \frac{1}{\beta-\alpha}\int_{\alpha}^{\beta}(1-p)^k\,\mathrm{d}p.$$

Evaluating the integral gives:

$$\mathbb{E}\big[(1-p)^k\big] \;=\; \frac{(1-\alpha)^{k+1}-(1-\beta)^{k+1}}{(\beta-\alpha)\cdot(k+1)}.$$

Thus, the aggregate success rate becomes:

$$\text{pass}_{\text{Uniform}(\alpha,\beta)}\,@k \;=\; 1 \;-\; \frac{(1-\alpha)^{k+1}-(1-\beta)^{k+1}}{(\beta-\alpha)\cdot(k+1)}.$$

Case A: $\alpha > 0$ If $\alpha > 0$, then both $(1-\alpha)$ and $(1-\beta)$ are strictly less than 1. As $k \to \infty$, $(1-\alpha)^{k+1}$ and $(1-\beta)^{k+1}$ decay exponentially. Hence:

$$\mathbb{E}\big[(1-p)^k\big] \sim \frac{(1-\alpha)^{k+1}}{(\beta-\alpha)\cdot(k+1)},$$

and $\text{pass}_{\text{Uniform}(\alpha,\beta)}\,@k$ approaches 1 exponentially fast:

$$\text{pass}_{\text{Uniform}(\alpha,\beta)}\,@k \sim 1 - \frac{(1-\alpha)^{k+1}}{(\beta-\alpha)\cdot(k+1)}.$$

Thus, the negative log of the aggregate success rate decays exponentially:

$$-\log\big(\text{pass}_{\text{Uniform}(\alpha,\beta)}\,@k\big) \sim e^{-\Omega(k)}.$$

Case B: $\alpha = 0$ When $\alpha = 0$, the uniform distribution is over $[0, \beta]$. In this case:

$$\mathbb{E}\big[(1-p)^k\big] \;=\; \frac{1}{\beta}\cdot\frac{1-(1-\beta)^{k+1}}{k+1}.$$

For large k, $(1-\beta)^{k+1}$ becomes exponentially small, and:

$$\mathbb{E}\big[(1-p)^k\big] \sim \frac{1}{\beta}\cdot\frac{1}{k+1}.$$

The aggregate success rate is then:

$$\text{pass}_{\text{Uniform}(0,\beta)}\,@k \sim 1 - \frac{1}{\beta\cdot k}.$$

The negative log exhibits power-law scaling:

$$-\log\big(\text{pass}_{\text{Uniform}(0,\beta)}\,@k\big) \sim \frac{1}{\beta}\cdot\frac{1}{k}.$$

Special Case: $\text{Uniform}(0,1)$ If $\beta = 1$, the distribution is uniform on $[0, 1]$. In this case:

$$\mathbb{E}\big[(1-p)^k\big] \;=\; \frac{1}{k+1},$$

and the success rate becomes:

$$\text{pass}_{\text{Uniform}(0,1)}\,@k \;=\; 1 - \frac{1}{k+1}.$$

For large k:

$$-\log\big(\text{pass}_{\text{Uniform}(0,1)}\,@k\big) \sim \frac{1}{k}.$$

E.4. 2-Parameter Beta Distribution: $\text{pass}_\text{i}@1 \sim \text{Beta}(\alpha, \beta)$

Suppose that the model's $\text{pass}_\text{i}@1$ probabilities across the benchmark problems follow a Beta distribution:

$$\text{pass}_\text{i}@1 \sim \text{Beta}(\alpha, \beta)$$

The probability density function of this distribution over the support $x \in (0, 1)$ is:

$$f(x; \alpha, \beta) \stackrel{\text{def}}{=} \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \tag{24}$$

where $\alpha > 0, \beta > 0$ and $B(\cdot, \cdot)$ is the Beta function. For brevity, let $p_i \stackrel{\text{def}}{=} \text{pass}_\text{i}@1$. Under our assumed Beta distribution:

$$\text{pass}_{\text{Beta}(\alpha,\beta)}@\text{k} \stackrel{\text{def}}{=} 1 - \mathbb{E}_{p_i \sim \text{Beta}(\alpha,\beta)}[(1-p_i)^k] \tag{25}$$

$$= 1 - \int_0^1 \frac{p_i^{\alpha-1}(1-p_i)^{\beta-1}}{B(\alpha,\beta)} (1-p_i)^k \, dp_i \tag{26}$$

$$= 1 - \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha)\Gamma(\beta+k)}{\Gamma(\alpha+\beta+k)} \tag{27}$$

where $\Gamma(\cdot)$ is again the Gamma function. The $\Gamma(\alpha)$ terms cancel, and a standard asymptotic result of the gamma function for large k tells us that:

$$\frac{\Gamma(\beta+k)}{\Gamma(\alpha+\beta+k)} \sim k^{-\alpha}, \tag{28}$$

and thus:

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)} \frac{\Gamma(\beta+k)}{\Gamma(\alpha+\beta+k)} \sim \frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)} k^{-\alpha}. \tag{29}$$

Recalling again that the expansion of $\log(\cdot)$ for small x is $-\log(1-x) = x + O(x^2)$, in our case, we obtain:

$$-\log\Big(\text{pass}_\mathcal{D}@\text{k}\Big) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)} k^{-\alpha} + O(k^{-2\alpha}) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)} k^{-\alpha} + o(k^{-\alpha}). \tag{30}$$

From this final result, we see that under a Beta distribution and in the large k regime, the negative log aggregate success rate exhibits polynomial (power-law) scaling with k for exponent α

E.5. Kumaraswamy Distribution: $\text{pass}_\text{i}@1 \sim \text{Kumaraswamy}(\alpha, \beta)$

Next, suppose the model's $\text{pass}_\text{i}@1$ probabilities follow a Kumaraswamy distribution. The probability density function of this distribution over the support $x \in (0, 1)$ is:

$$f(x; \alpha, \beta) \stackrel{\text{def}}{=} \alpha \, \beta \, x^{\alpha-1} (1-x^\alpha)^{\beta-1} \tag{31}$$

Again for brevity, let $p_i \stackrel{\text{def}}{=} \text{pass}_\text{i}@1$. Under our assumed Kumaraswamy distribution:

$$\text{pass}_{\text{Kumaraswamy}(\alpha,\beta)}@\text{k} \stackrel{\text{def}}{=} 1 - \mathbb{E}_{p_i \sim \text{Kumaraswamy}(\alpha,\beta)}[(1-p_i)^k] \tag{32}$$

$$= 1 - \int_0^1 (1-p)^k \cdot \alpha \, \beta \, p^{\alpha-1} (1-p^\alpha)^{\beta-1} dp. \tag{33}$$

Define the integral

$$I_k \stackrel{\text{def}}{=} \mathbb{E}\big((1-p)^k\big) = \int_0^1 (1-x)^k \, \alpha \, \beta \, x^{\alpha-1} \left(1-x^\alpha\right)^{\beta-1} \mathrm{d}x. \tag{34}$$

We aim to analyze I_k for large k. Notice that $(1-x)^k$ is exponentially small in k unless x is very close to 0. Thus, intuitively, most of the contribution to I_k arises from $x \in [0, \, O(1/k)]$.

Step 1: Split the integral into two parts. Fix a constant $c > 0$. Write

$$I_k \;=\; \int_0^{c/k} [\cdots]\,\mathrm{d}x \;+\; \int_{c/k}^1 [\cdots]\,\mathrm{d}x \;\stackrel{\text{def}}{=}\; I_{k,\text{left}} \;+\; I_{k,\text{right}},$$

where $[\cdots]$ indicates the same integrand. In the region $x \in [c/k, 1]$, we have $(1-x)^k \le e^{-k\,x} \le e^{-c}$. Hence $I_{k,\text{right}} = O\big(e^{-c}\big)$. Since c can be made arbitrarily large, $I_{k,\text{right}}$ becomes negligible compared to any polynomial in $1/k$.

Step 2: Approximate the integrand in the small-x region. On $[0, c/k]$, we use the approximation $\log(1-x) = -x + O(x^2)$. Thus

$$(1-x)^k = \exp\big(k\log(1-x)\big) = \exp\big(-k\,x + O(k\,x^2)\big).$$

Since $x \le c/k$ implies $k\,x^2 \le c^2/k = O(1/k)$, and $\exp(\epsilon) = 1 + O(\epsilon)$, we get

$$(1-x)^k = \exp(-k\,x)\exp(O(1/k)) = \exp(-k\,x)\,\big(1 + O\big(\tfrac{1}{k}\big)\big).$$

Furthermore, since $(1-y)^m = 1 - my + O(y^2)$, for small x

$$(1-x^\alpha)^{\beta-1} = 1 - (\beta-1)x^\alpha + O(x^{2\alpha}) = 1 + O\big(x^\alpha\big).$$

In the region $x \le c/k$, that error is $O\big(k^{-\alpha}\big)$. Hence, within the small-x region, the integrand

$$(1-x)^k\;\alpha\,\beta\;x^{\alpha-1}\;\big(1-x^\alpha\big)^{\beta-1}$$

can be approximated by

$$\alpha\,\beta\;x^{\alpha-1}\;e^{-k\,x} \;+\; O\Big(k^{-\alpha}\,x^{\alpha-1}\,e^{-k\,x}\Big).$$

Thus

$$I_{k,\text{left}} \;=\; \int_0^{c/k} \alpha\,\beta\;x^{\alpha-1}\,e^{-k\,x}\;\mathrm{d}x \;+\; O\Big(k^{-\alpha}\int_0^{c/k} x^{\alpha-1}\,e^{-k\,x}\,\mathrm{d}x\Big) \;+\; O\big(e^{-c}\big).$$

Step 3: Substitution $u \stackrel{\text{def}}{=} k\,x$. To handle $\int_0^{c/k} x^{\alpha-1}e^{-k\,x}\,\mathrm{d}x$, we substitute $u = k\,x$. Then $x = u/k$, $\mathrm{d}x = \mathrm{d}u/k$, and the upper limit $x = c/k$ becomes $u = c$. Hence,

$$\begin{aligned}\int_0^{c/k} x^{\alpha-1}\,e^{-k\,x}\,\mathrm{d}x &= \int_0^c \Big(\tfrac{u}{k}\Big)^{\alpha-1}\,e^{-\,u}\;\frac{\mathrm{d}u}{k}\\ &= k^{-\alpha}\int_0^c u^{\alpha-1}\,e^{-\,u}\,\mathrm{d}u.\end{aligned}$$

As $c \to \infty$, $\int_0^c u^{\alpha-1}e^{-\,u}\,\mathrm{d}u \to \Gamma(\alpha)$, and for finite c the remainder is $O\big(e^{-\,c}\big)$. Therefore,

$$\int_0^1 x^{\alpha-1}\,e^{-k\,x}\,\mathrm{d}x = k^{-\alpha}\,\Gamma(\alpha) \;+\; O\big(k^{-\alpha}\,e^{-\,c}\big),$$

and absorbing the constant c into big-O notation gives

$$\int_0^1 x^{\alpha-1}\,e^{-k\,x}\,\mathrm{d}x = k^{-\alpha}\,\Gamma(\alpha) \;+\; O\big(k^{-\alpha-\epsilon}\big) \quad \text{for some } \epsilon > 0.$$

Multiplying by the factor $\alpha\,\beta$, we deduce that

$$I_k \;=\; \alpha\,\beta\,\Gamma(\alpha)\,k^{-\alpha} \;+\; O\big(k^{-\alpha-\epsilon}\big).$$

Step 4: Final conclusion for the success rate. Recall $\text{pass}_{\text{Kumaraswamy}(\alpha,\beta)}@\text{k} = 1 - I_k$. Hence

$$\text{pass}_{\text{Kumaraswamy}(\alpha,\beta)}@\text{k} \;=\; 1 \;-\; \alpha\,\beta\,\Gamma(\alpha)\,k^{-\alpha} \;+\; O\big(k^{-\alpha-\epsilon}\big).$$

Since this tends to 1, its negative log is governed by the magnitude of $\alpha\,\beta\,\Gamma(\alpha)\,k^{-\alpha}$. Using the expansion $-\log(1-y) = y + O(y^2)$ as $y \to 0$, we get

$$-\log\Big(\,\text{pass}_{\text{Kumaraswamy}(\alpha,\beta)}@\text{k}\,\Big) \;=\; \alpha\,\beta\,\Gamma(\alpha)\;k^{-\alpha} \;+\; o\big(k^{-\alpha}\big).$$

That is precisely polynomial (power-law) decay in the negative log success rate with exponent α.

E.6. Continuous Bernoulli Distribution: $\text{pass}_{\text{i}}@1 \sim \text{ContinousBernoulli}(\lambda)$

Next, suppose the model's $\text{pass}_{\text{i}}@1$ probabilities follow a Continuous Bernoulli distribution. The probability density function of this distribution over the support $x \in [0,1]$ is:

$$f(x;\lambda) \overset{\text{def}}{=} C(\lambda)\lambda^x(1-\lambda)^{1-x} \tag{35}$$

$$C(\lambda) \overset{\text{def}}{=} \begin{cases} 2 & \text{if } \lambda = 1/2 \\ \frac{2\tanh^{-1}(1-2\lambda)}{1-2\lambda} & \text{otherwise} \end{cases}. \tag{36}$$

The density can equivalently be rewritten in a more convenient form for our purposes:

$$f(x;\lambda) = C(\lambda)\lambda^x(1-\lambda)(1-\lambda)^{-x} = C(\lambda)(1-\lambda)\Big(\frac{\lambda}{1-\lambda}\Big)^x \tag{37}$$

Because the individual success probability is low in our data, we shall consider the small $\lambda < 1/2$ regime. We follow the same approach as with the Kumaraswamy distribution.

Step 1: Write the aggregate pass rate. The aggregate pass rate is defined as:

$$\text{pass}_{\text{ContinuousBernoulli}(\lambda)}@\text{k} = 1 - I_k, \quad \text{where} \quad I_k \overset{\text{def}}{=} \int_0^1 (1-p)^k\, f(p;\lambda)\, dp.$$

Substituting the density $f(p;\lambda)$, we get:

$$I_k = \int_0^1 (1-p)^k\, C(\lambda)\, \lambda^p\, (1-\lambda)^{1-p}\, dp.$$

Step 2: Simplify using an exponential form. Using the exponential rewriting:

$$\lambda^p\,(1-\lambda)^{1-p} = (1-\lambda)\exp\Big(p\log\big(\tfrac{\lambda}{1-\lambda}\big)\Big),$$

the integral becomes:

$$I_k = C(\lambda)\,(1-\lambda)\int_0^1 (1-p)^k \exp\Big(p\log\big(\tfrac{\lambda}{1-\lambda}\big)\Big)\, dp.$$

Step 3: Dominance of the small-p region. For large k, $(1-p)^k$ decays exponentially unless p is close to 0. Thus, the main contribution to the integral arises from the region $p \in [0, c/k]$, where $c > 0$ is a constant. Decompose the integral:

$$I_k = \int_0^{c/k} [\cdots]\, dp + \int_{c/k}^1 [\cdots]\, dp \overset{\text{def}}{=} I_{k,\text{left}} + I_{k,\text{right}}.$$

In the region $p \in [c/k, 1]$, we have $(1-p)^k \le e^{-kp} \le e^{-c}$, making $I_{k,\text{right}} = O(e^{-c})$, which is negligible compared to $1/k$. Thus, we focus on $I_{k,\text{left}}$:

$$I_{k,\text{left}} = C(\lambda)\,(1-\lambda)\int_0^{c/k} (1-p)^k \exp\Big(p\log\big(\tfrac{\lambda}{1-\lambda}\big)\Big)\, dp.$$

Step 4: Approximate the integrand. For $p \in [0, c/k]$, use the same approximations from the Kumaraswamy derivation:

$$(1-p)^k = e^{-kp}\left(1 + O(p)\right), \quad \exp\Big(p \log\big(\tfrac{\lambda}{1-\lambda}\big)\Big) = 1 + O(p).$$

Thus, the integrand becomes:

$$(1-p)^k \exp\Big(p \log\big(\tfrac{\lambda}{1-\lambda}\big)\Big) = e^{-kp}\left(1 + O(p)\right).$$

Step 5: Change of variables. Let $u \stackrel{\text{def}}{=} kp$, so $p = u/k$ and $dp = du/k$. The integral becomes:

$$I_{k,\text{left}} = C(\lambda)\,(1-\lambda) \int_0^c e^{-u}\left(1 + O(u/k)\right) \frac{du}{k}.$$

Split the integral:

$$I_{k,\text{left}} = \frac{C(\lambda)\,(1-\lambda)}{k} \int_0^c e^{-u}\, du + O\Big(\frac{1}{k^2}\Big).$$

As $c \to \infty$, $\int_0^c e^{-u}\, du \to 1$. Thus:

$$I_{k,\text{left}} = \frac{C(\lambda)\,(1-\lambda)}{k} + O\Big(\frac{1}{k^2}\Big).$$

Since $I_{k,\text{right}} = O(e^{-c})$ is negligible, we have:

$$I_k = \frac{C(\lambda)\,(1-\lambda)}{k} + O\Big(\frac{1}{k^2}\Big).$$

Step 7: Final conclusion for the success rate. Recall:

$$\text{pass}_{\text{ContinuousBernoulli}(\lambda)}@\text{k} = 1 - I_k.$$

For large k, this implies:

$$\text{pass}_{\text{ContinuousBernoulli}(\lambda)}@\text{k} = 1 - \frac{C(\lambda)\,(1-\lambda)}{k} + O\Big(\frac{1}{k^2}\Big).$$

Using the expansion $-\log(1-y) = y + O(y^2)$ for small y, we find:

$$-\log\big(\text{pass}_{\text{ContinuousBernoulli}(\lambda)}@\text{k}\big) = C(\lambda)\,(1-\lambda)k^{-1} + o(k^{-1}).$$

That is precisely polynomial (power-law) decay in the negative log success rate with exponent -1.

As a side comment, recall that $\tanh^{-1}(x) = \frac{1}{2}\log\left(\frac{1+x}{1-x}\right)$, the normalizing constant $C(\lambda)$ can be rewritten as:

$$C(\lambda) = \frac{2}{1-2\lambda}\frac{1}{2}\log\left(\frac{1+(1-2\lambda)}{1-(1-2\lambda)}\right) = \frac{1}{1-2\lambda}\log\Big(\frac{1-\lambda}{\lambda}\Big). \tag{38}$$

Thus, for small λ, note that $C(\lambda) \approx \log(1/\lambda) = -\log(\lambda)$. For $k \ll -\log(\lambda)$, the $1/k$ formula is valid. However, near $k \approx -\log(\lambda)$, the leading term $-\log(\lambda)/k$ becomes of order 1, and for $k \gg -\log(\lambda)$, the success rate is now very close to 1. Consequently, we see that if λ is very small, there is a soft cutoff scale around $k \approx -\log(\lambda)$.

E.7. Any Continuous Distribution with $p(\text{pass}_\text{i}@1) = c > 0$

Suppose that the distribution over $\text{pass}_\text{i}@1$ is continuous and has constant non-zero density near 0:

$$f(0) = c > 0 \tag{39}$$

Because the density is continuous at 0 with $f(0) = c > 0$, there exist some $\delta > 0$ such that:

$$f(p) = c + O(p) \qquad \text{for all } p \in [0, \delta]. \tag{40}$$

Because the small $\text{pass}_\text{i}@1$ region dominates for large k, a similar argument to the Kumaraswamy argument and Continuous Bernoulli argument yields power law scaling with respect to k with exponent -1:

$$-\log\Big(\,\text{pass}_{\mathcal{D}}@\text{k}\Big) = c\,k^{-1} + o(k^{-1}). \tag{41}$$

This result is consistent with the Continuous Bernoulli, where c is given by $f_{\text{ContinuousBernoulli}(\lambda)}(0;\lambda) = C(\lambda)(1-\lambda)$ for $\lambda < 1/2$. This result reveals that the Continous Bernoulli is just one instance of a larger family: any continuous distribution with non-zero constant density at $\text{pass}_\text{i}@1 = 0$ will exhibit power law scaling with exponent -1.

E.8. Reciprocal Distribution: $\text{pass}_\text{i}@1 \sim \text{Reciprocal}(a, b)$

Next, suppose the model's $\text{pass}_\text{i}@1 \sim \text{Reciprocal(a, b)}$ distribution with $0 < a < b < 1$. The probability density function of this distribution over the support $x \in [a, b]$ is:

$$f(x; a, b) = \frac{1}{(\log(b) - \log(a))\,x} \tag{42}$$

As with the other distributions, the aggregate success rate after k attempts is:

$$\text{pass}_{\text{Reciprocal(a,b)}}@\text{k} \;=\; \mathbb{E}\big[\text{pass}_\text{i}@\text{k}\big] \;=\; 1 \;-\; I_k, \quad \text{where} \quad I_k \;\overset{\text{def}}{=}\; \int_{x=a}^{b} (1-x)^k\,\frac{1}{(\log b - \log a)\,x}\,\mathrm{d}x.$$

We aim to show that I_k is on the order of $\frac{(1-a)^k}{k}$. The main contribution to the integral arises from the vicinity of $x = a$, because $(1-x)^k$ decays rapidly as x grows away from a.

Step 1: Change of variable. Define $y \overset{\text{def}}{=} x - a$, so the domain $x \in [a, b]$ becomes $y \in [0, b-a]$. Then

$$(1-x)^k \;=\; \big((1-a) - y\big)^k,$$

and

$$I_k \;=\; \frac{1}{\log(b/a)} \int_{y=0}^{b-a} \big((1-a) - y\big)^k\,\frac{1}{a+y}\,\mathrm{d}y.$$

Step 2: Expansion near $y = 0$. For small y, write $(1-a) - y = (1-a)\big(1 - \frac{y}{1-a}\big)$; hence

$$\log\big((1-a) - y\big) \;=\; \log(1-a) \;+\; \log\Big(1 - \tfrac{y}{1-a}\Big).$$

Using $\log(1-z) = -z + O(z^2)$ for small z, we get

$$\log\big((1-a) - y\big) \;=\; \log(1-a) \;-\; \frac{y}{1-a} \;+\; O\big(\tfrac{y^2}{(1-a)^2}\big),$$

so

$$(1-a-y)^k \;=\; \exp\Big(k\,\log(1-a) \;-\; k\,\tfrac{y}{1-a} \;+\; O\big(\tfrac{k\,y^2}{(1-a)^2}\big)\Big).$$

In particular, for y up to c/k, the term $k\,y^2 = O(1)$ remains bounded, so

$$(1-a-y)^k \;=\; (1-a)^k\,\exp\Big(-\tfrac{k\,y}{1-a}\Big)\,\big[\,1 + O\big(\tfrac{1}{k}\big)\big].$$

Step 3: The integral is dominated by $y \in [0, O(\frac{1}{k})]$. For large k, $\exp\big(-\frac{k\,y}{1-a}\big)$ decays quickly once y exceeds a multiple of $\frac{1-a}{k}$. Consequently, the integral from $y = c_0/k$ to $b - a$ is exponentially small in k. On $[0, c_0/k]$, we also have $(a+y)^{-1} = \frac{1}{a} + O\big(\frac{1}{k}\big)$. Thus

$$I_k \;=\; \frac{1}{\log(b/a)} \int_{y=0}^{c_0/k} (1-a-y)^k\,\frac{1}{a+y}\,\mathrm{d}y \;+\; \text{(exponentially small tail)}.$$

Substitute our approximation from Step 2 into the integrand:

$$(1-a-y)^k \, \frac{1}{a+y} \;=\; (1-a)^k \, \exp\!\Big(-\tfrac{k\,y}{1-a}\Big) \, \Big[\tfrac{1}{a} + O\big(\tfrac{1}{k}\big)\Big].$$

Step 4: Change variable $u = \frac{k\,y}{1-a}$. Then $y = \frac{(1-a)\,u}{k}$ and $\mathrm{d}y = \frac{1-a}{k}\,\mathrm{d}u$. The upper limit $y = c_0/k$ corresponds to $u = c_0\,\big(\frac{1-a}{1}\big)$, so

$$\int_{y=0}^{c_0/k} \exp\!\Big(-\tfrac{k\,y}{1-a}\Big)\,\mathrm{d}y \;=\; \int_{u=0}^{c_0\,(1-a)} e^{-u}\,\frac{1-a}{k}\,\mathrm{d}u.$$

Letting $c_0 \to \infty$ only contributes an $e^{-c_0(1-a)}$ factor to the tail, which vanishes. Hence

$$\int_{y=0}^{\infty} \exp\!\Big(-\tfrac{k\,y}{1-a}\Big)\,\mathrm{d}y \;=\; \frac{1-a}{k}\int_{u=0}^{\infty} e^{-u}\,\mathrm{d}u \;=\; \frac{1-a}{k}.$$

Putting all factors together,

$$I_k \;=\; \frac{1}{\log(b/a)}\,(1-a)^k\,\Big[\tfrac{1}{a} + O\big(\tfrac{1}{k}\big)\Big]\,\frac{1-a}{k} \;+\; \text{(exponentially small in } k\text{)}.$$

Thus in big-Theta form,

$$I_k \;=\; \Theta\Big(\tfrac{(1-a)^k}{k}\Big).$$

Conclusion. Since $\mathrm{pass}_{\mathrm{Reciprocal(a,b)}}@\mathrm{k} = 1 - I_k$, we get

$$\mathrm{pass}_{\mathrm{Reciprocal(a,b)}}@\mathrm{k} \;=\; 1 \;-\; \Theta\Big(\tfrac{(1-a)^k}{k}\Big).$$

Moreover, using $-\log(1-y) = y + O(y^2)$ for small y, it follows that

$$-\log\Big(\mathrm{pass}_{\mathrm{Reciprocal(a,b)}}@\mathrm{k}\Big) \;=\; \Theta\Big(\tfrac{(1-a)^k}{k}\Big).$$

Hence the negative log aggregate success rate converges to 1 *exponentially fast* in k, which is *not* a power law in k.

Sufficient Condition for Power-Law Scaling in Negative Log of Aggregate Success

Theorem E.1. *Let $\mathcal{D}$ be a probability distribution on $[0,1]$ with PDF $f(p)$. Suppose there exist constants $b > 0$, $C > 0$, $\theta > 0$ and $\delta > 0$ such that, for all $0 < p < \delta$, we have*

$$f(p) \;=\; C\,p^{\,b-1} \;+\; O\big(p^{\,b-1+\theta}\big).$$

Then, for large k,

$$1 \;-\; \mathrm{pass}_{\mathcal{D}}@\mathrm{k} \;=\; C\,\Gamma(b)\,k^{-b} \;+\; O\big(k^{-b-\min(\,1,\theta)}\big),$$

which implies

$$-\log\Big(\mathrm{pass}_{\mathcal{D}}@\mathrm{k}\Big) \;=\; C\,\Gamma(b)\,k^{-b} \;+\; o\big(k^{-b}\big).$$

Equivalently, including the leading constant),

$$-\log\big(\mathrm{pass}_{\mathcal{D}}@\mathrm{k}\big) \;\sim\; C\,\Gamma(b)\;k^{-b}.$$

Proof. **Step 1. Decompose the key integral.**
Define

$$I_k \;\stackrel{\text{def}}{=}\; 1 \;-\; \mathrm{pass}_{\mathcal{D}}@\mathrm{k} \;=\; \int_0^1 (1-p)^k\,f(p)\,\mathrm{d}p.$$

For a positive constant $c > 0$, split I_k:

$$I_k \;=\; \int_0^{c/k} (1-p)^k\,f(p)\,\mathrm{d}p \;+\; \int_{c/k}^1 (1-p)^k\,f(p)\,\mathrm{d}p \;\stackrel{\text{def}}{=}\; I_{k,\mathrm{left}} \;+\; I_{k,\mathrm{right}}.$$

Right Tail Bound ($I_{k,\mathrm{right}}$). For $p \geq c/k$, observe $(1-p)^k \leq e^{-k\,p} \leq e^{-c}$. Hence

$$I_{k,\mathrm{right}} \;=\; \int_{c/k}^1 (1-p)^k\,f(p)\,\mathrm{d}p \;\leq\; e^{-c}\int_0^1 f(p)\,\mathrm{d}p \;=\; e^{-c}.$$

Since c can be made arbitrarily large, e^{-c} can be driven below *any* power of $1/k$. Thus $I_{k,\mathrm{right}} = o\big(k^{-\alpha}\big)$ for any $\alpha > 0$. We may therefore focus on

$$I_{k,\mathrm{left}} \;=\; \int_0^{c/k} (1-p)^k\,f(p)\,\mathrm{d}p,$$

knowing that $I_{k,\mathrm{right}}$ is negligible in polynomial-type estimates.

Step 2. Use the assumed behavior of $f(p)$ near $p = 0$.
By hypothesis, for p up to some $\delta > 0$,

$$f(p) \;=\; C\,p^{\,b-1} + O\big(p^{\,b-1+\theta}\big).$$

Choose $c/k < \delta$, so $p \leq c/k < \delta$ for p in the left integral. Then

$$I_{k,\mathrm{left}} \;=\; \int_0^{c/k} (1-p)^k\Big[\,C\,p^{\,b-1} + O\big(p^{\,b-1+\theta}\big)\Big]\,\mathrm{d}p.$$

Split it into main term and error term:

$$I_{k,\mathrm{left}} \;=\; C\int_0^{c/k} (1-p)^k\,p^{\,b-1}\,\mathrm{d}p \;+\; \int_0^{c/k} (1-p)^k\,O\big(p^{\,b-1+\theta}\big)\,\mathrm{d}p.$$

Denote these T_{main} and T_{err}, respectively.

Step 3. Approximate $(1-p)^k$ by e^{-kp} and control the error.
For p in $[0, c/k]$, expand $\log(1-p) = -p + O(p^2)$. Thus

$$(1-p)^k \;=\; \exp\big(k\log(1-p)\big) \;=\; e^{-k\,p}\,\exp\big(O(k\,p^2)\big) \;=\; e^{-k\,p}\,\big[\,1 + O(k\,p^2)\big].$$

Since $p \le c/k$, we get $k\,p^2 \le c^2/k$, which is bounded for large k. Consequently,

$$(1-p)^k = e^{-k\,p} + O\big(k\,p^2\,e^{-k\,p}\big).$$

We will use this in both T_{main} and T_{err}.

Step 4. Main term T_{main}.

$$T_{\text{main}} \;=\; C\int_0^{c/k} (1-p)^k\,p^{\,b-1}\,\mathrm{d}p.$$

Substituting $(1-p)^k = e^{-k\,p} + O\big(k\,p^2\,e^{-k\,p}\big)$,

$$T_{\text{main}} \;=\; C\int_0^{c/k} e^{-k\,p}\,p^{\,b-1}\,\mathrm{d}p \;+\; C\int_0^{c/k} O\big(k\,p^{\,b+1}\,e^{-k\,p}\big)\,\mathrm{d}p.$$

Call these two integrals T_1 and T_2.

T_1 term.

$$T_1 = C\int_0^{c/k} p^{\,b-1}\,e^{-k\,p}\,\mathrm{d}p.$$

Make the substitution $u \stackrel{\text{def}}{=} k\,p$. Then $p = u/k$, $\mathrm{d}p = \mathrm{d}u/k$, and $p^{\,b-1} = k^{-b+1}\,u^{\,b-1}$. The upper limit $p = c/k$ becomes $u = c$. Thus

$$T_1 = C\int_0^{c} \big(\tfrac{u}{k}\big)^{b-1} e^{-u}\,\tfrac{\mathrm{d}u}{k} = C\,k^{-b}\int_0^{c} u^{\,b-1}\,e^{-u}\,\mathrm{d}u.$$

As $c \to \infty$, $\int_0^c u^{\,b-1}e^{-u}\,\mathrm{d}u \to \Gamma(b)$. So

$$T_1 = C\,k^{-b}\Big(\Gamma(b) - R_c\Big), \quad \text{where } |R_c| = O\big(e^{-c}\big).$$

By choosing c large after $k \to \infty$, we conclude

$$T_1 = C\,\Gamma(b)\,k^{-b} + o\big(k^{-b}\big).$$

T_2 term.

$$T_2 = C\int_0^{c/k} O\big(k\,p^{\,b+1}\,e^{-k\,p}\big)\,\mathrm{d}p.$$

Inside the integral, $k\,p^{\,b+1}\,e^{-k\,p}$ is the main factor. Substituting $u \stackrel{\text{def}}{=} k\,p$ again,

$$p^{\,b+1} = \big(\tfrac{u}{k}\big)^{b+1} = k^{-b-1}\,u^{\,b+1}.$$

Hence

$$T_2 = C\,O(1)\int_0^{c/k} k\,p^{\,b+1}\,e^{-k\,p}\,\mathrm{d}p = O(k)\int_0^{c/k} p^{\,b+1}e^{-k\,p}\,\mathrm{d}p.$$

Substitute $u = k\,p$ and $\mathrm{d}p = \mathrm{d}u/k$. Then

$$T_2 = O(k)\int_0^{c} \big(\tfrac{u}{k}\big)^{b+1}e^{-u}\,\tfrac{\mathrm{d}u}{k} = O(k)\,k^{-b-2}\int_0^{c} u^{\,b+1}\,e^{-u}\,\mathrm{d}u = O\big(k^{-b-1}\big).$$

Thus T_2 is of strictly smaller order than k^{-b}.

Combine T_1 and T_2:

$$T_{\text{main}} = C\,\Gamma(b)\,k^{-b} + O\big(k^{-b-1}\big).$$

Step 5. Error term T_{err}.
Recall

$$T_{\text{err}} = \int_0^{c/k} (1-p)^k \, O\big(p^{\,b-1+\theta}\big) \, \mathrm{d}p.$$

Exactly the same substitution $(1-p)^k = e^{-kp} + O(k\,p^2\,e^{-k\,p})$ plus $u = k\,p$ shows

$$T_{\text{err}} = O\Big(\int_0^{c/k} p^{\,b-1+\theta}\, e^{-k\,p}\, \mathrm{d}p\Big) \;+\; O\Big(\int_0^{c/k} k\,p^{\,b+1+\theta}\, e^{-k\,p}\, \mathrm{d}p\Big).$$

When substituting $u = k\,p$, the exponent on p increases by $+1$ each time if we multiply by k, so each term is of order $k^{-b-\theta}$ or smaller. Concretely,

$$\int_0^{c/k} p^{\,b-1+\theta}\, e^{-k\,p}\, \mathrm{d}p = k^{-b-\theta} \int_0^{c} u^{\,b-1+\theta}\, e^{-u}\, \mathrm{d}u = O\big(k^{-b-\theta}\big),$$

and similarly for the second term, which is even smaller. Hence

$$T_{\text{err}} = O\big(k^{-b-\theta}\big).$$

Step 6. Putting it all together.
Summarize:

$$I_{k,\text{left}} = T_{\text{main}} + T_{\text{err}} = C\,\Gamma(b)\,k^{-b} \;+\; O\big(k^{-b-1}\big) \;+\; O\big(k^{-b-\theta}\big).$$

Thus

$$I_{k,\text{left}} = C\,\Gamma(b)\,k^{-b} \;+\; O\big(k^{-b-\min(1,\theta)}\big).$$

Recalling the tail piece $I_{k,\text{right}} = e^{-c} = o\big(k^{-\alpha}\big)$ for any α, we obtain

$$I_k = I_{k,\text{left}} + I_{k,\text{right}} = C\,\Gamma(b)\,k^{-b} \;+\; O\big(k^{-b-\min(1,\theta)}\big).$$

Hence

$$1 - \text{pass}_{\mathcal{D}}@\text{k} = I_k \;\;\sim\;\; C\,\Gamma(b)\,k^{-b}.$$

Final negative-log argument. Since

$$\text{pass}_{\mathcal{D}}@\text{k} = 1 - I_k = 1 - \big(C\,\Gamma(b)\,k^{-b} + O\big(k^{-b-\min(1,\theta)}\big)\big),$$

for large k it is very close to 1. Then

$$-\log\Big(\text{pass}_{\mathcal{D}}@\text{k}\Big) = -\log\Big(1 - C\,\Gamma(b)\,k^{-b} + \cdots\Big).$$

Using the expansion $-\log(1-x) = x + O(x^2)$ as $x \to 0$, and here $x = C\,\Gamma(b)\,k^{-b}$, we get

$$-\log\Big(\text{pass}_{\mathcal{D}}@\text{k}\Big) = C\,\Gamma(b)\,k^{-b} + o\big(k^{-b}\big).$$

In the "$\sim$" notation including the leading coefficient:

$$-\log\Big(\text{pass}_{\mathcal{D}}@\text{k}\Big) \;\sim\; C\,\Gamma(b)\;k^{-b}.$$

This completes the proof. □

E.9. Necessary Condition for Power Law Scaling from Distribution over $\mathrm{pass_i@1}$

Theorem E.2. *Let $\mathcal{D}$ be a probability distribution over $[0,1]$ with a PDF $f(p)$ satisfying the following regularity near $p=0$:*

- ***No point mass at** $p=0$. So $\int_0^1 f(p)\,\mathrm{d}p = 1$, and f is a genuine PDF on $(0,1]$.*
- ***Continuity and nonnegative behavior near** $p=0$. There exist $\delta > 0$ such that f is continuous on $[0,\delta]$ and has no pathological oscillations or singularities that violate integrability.*

Define the aggregate success rate at k attempts:

$$\mathrm{pass}_{\mathcal{D}}@\mathrm{k} \;\stackrel{def}{=}\; \int_0^1 \Big[\,1-(1-p)^k\Big]\, f(p)\,\mathrm{d}p$$

and relatedly

$$I_k \;\stackrel{def}{=}\; \int_0^1 (1-p)^k\, f(p)\,\mathrm{d}p \;=\; 1-\mathrm{pass}_{\mathcal{D}}@\mathrm{k}\,.$$

***Assume** that there exist constants $A>0$ and $b>0$ such that for large k:*

$$-\log\big(\mathrm{pass}_{\mathcal{D}}@\mathrm{k}\big) \;\sim\; A\,k^{-b}$$

Then

$$I_k \;=\; A\,k^{-b} + o\big(k^{-b}\big),$$

and under the mild regularity assumptions above,

$$f(p) \;\sim\; \frac{A}{\Gamma(b)}\; p^{\,b-1} \quad \textit{as } p \to 0^+.$$

Proof. **Step 1. Relating I_k to $-\log(\mathrm{pass}_{\mathcal{D}}@\mathrm{k})$.**

By definition,

$$\mathrm{pass}_{\mathcal{D}}@\mathrm{k} \;=\; 1-I_k, \quad I_k \;=\; \int_0^1 (1-p)^k\, f(p)\,\mathrm{d}p.$$

Since

$$-\log\Big(\mathrm{pass}_{\mathcal{D}}@\mathrm{k}\Big) \;\sim\; A\,k^{-b},$$

we have, for large k,

$$\mathrm{pass}_{\mathcal{D}}@\mathrm{k} \;=\; \exp\big(-A\,k^{-b}\,(1+o(1))\big).$$

When x is small, $\exp(-x) = 1 - x + O(x^2)$. Thus

$$I_k \;=\; 1-\mathrm{pass}_{\mathcal{D}}@\mathrm{k} \;=\; A\,k^{-b} + o\big(k^{-b}\big).$$

So

$$I_k \;\sim\; A\,k^{-b}.$$

Step 2. Restricting to a small interval near $p=0$.

Since $(1-p)^k$ decays exponentially once p is on the order of $1/k$ or larger, we split:

$$I_k \;\stackrel{\text{def}}{=}\; \int_0^1 (1-p)^k\, f(p)\,\mathrm{d}p \;=\; \int_0^{c/k} (1-p)^k\, f(p)\,\mathrm{d}p \;+\; \int_{c/k}^1 (1-p)^k\, f(p)\,\mathrm{d}p \;\stackrel{\text{def}}{=}\; I_{k,\mathrm{left}} + I_{k,\mathrm{right}},$$

for some positive constant c. In the region $p \geq c/k$, we have $(1-p)^k \leq e^{-k\,p} \leq e^{-c}$, so

$$I_{k,\text{right}} \;\leq\; e^{-c} \int_0^1 f(p)\,\mathrm{d}p \;=\; e^{-c}.$$

Since $c > 0$ can be made large, e^{-c} can be driven below any fixed power of $1/k$. Hence for the $\Theta(k^{-b})$ behavior, the main contribution comes from $[0, c/k]$.

Thus

$$I_k \;=\; I_{k,\text{left}} \;+\; o\big(k^{-m}\big) \;\text{ for every } m > 0$$

Step 3. Change of variables and controlling the ratio of $(1-p)^k$ to e^{-kp}.

(a) Ratio to e^{-kp}. For $p \in \big[0, \frac{c}{k}\big]$, define the ratio

$$R_k(p) \;\stackrel{\text{def}}{=}\; \frac{(1-p)^k}{e^{-k\,p}}.$$

We will show that $R_k(p)$ stays close to 1 uniformly in $p \in [0, c/k]$ for large k. Indeed,

$$(1-p)^k \;=\; \exp\Big[k\,\log(1-p)\Big], \quad \log(1-p) \;=\; -\,p \;-\; \frac{p^2}{2} \;-\; \frac{p^3}{3} \;-\; \ldots\,.$$

Hence

$$\log(1-p) + p \;=\; -\frac{p^2}{2} \;-\; \frac{p^3}{3} \;-\; \ldots \;=\; O\big(p^2\big) \quad \text{as } p \to 0.$$

Multiplying by k, we get

$$k\,\big[\log(1-p) + p\big] \;=\; O\big(k\,p^2\big).$$

Since $0 \leq p \leq \frac{c}{k}$ implies $k\,p^2 \leq \frac{c^2}{k}$, which $\to 0$ as $k \to \infty$, it follows that

$$k\,\log(1-p) \;=\; -k\,p + O\big(\tfrac{1}{k}\big).$$

Exponentiating:

$$(1-p)^k \;=\; e^{-\,k\,p}\,\exp\Big(O\big(\tfrac{1}{k}\big)\Big) \;=\; e^{-\,k\,p}\Big[1 + O\big(\tfrac{1}{k}\big)\Big].$$

Thus

$$R_k(p) \;=\; \frac{(1-p)^k}{e^{-k\,p}} \;=\; 1 + O\big(\tfrac{1}{k}\big),$$

with the $O(\frac{1}{k})$ bound uniform for all $p \in [0, c/k]$. In other words, there is some constant $M > 0$ (independent of k) such that

$$\big|R_k(p) - 1\big| \;\leq\; \frac{M}{k} \quad \text{for all } p \in \Big[0, \frac{c}{k}\Big].$$

(b) Integral expression using $R_k(p)$. Hence on $[0, c/k]$,

$$(1-p)^k\,f(p) \;=\; e^{-k\,p}\,R_k(p)\,f(p).$$

Thus

$$I_{k,\text{left}} \;=\; \int_0^{c/k} e^{-k\,p}\,f(p)\,R_k(p)\,\mathrm{d}p.$$

Define $\Delta_k(p) \stackrel{\text{def}}{=} R_k(p) - 1$, which satisfies $|\Delta_k(p)| \leq M/k$. Then

$$I_{k,\text{left}} = \int_0^{c/k} e^{-k\,p}\,f(p)\,\mathrm{d}p \;+\; \int_0^{c/k} e^{-k\,p}\,f(p)\,\Delta_k(p)\,\mathrm{d}p. \tag{43}$$

Step 4. Substitution $u = k\,p$ and deriving $f(p) \sim p^{\,b-1}$.

(a) The leading part. Focus on the first term of equation 43:

$$\int_0^{c/k} e^{-k\,p}\, f(p)\,\mathrm{d}p.$$

Substitute $u \overset{\text{def}}{=} k\,p$, so $p = \frac{u}{k}$ and $\mathrm{d}p = \frac{1}{k}\,\mathrm{d}u$. The upper limit $p = \frac{c}{k}$ becomes $u = c$. Thus

$$\int_0^{c/k} e^{-k\,p}\, f(p)\,\mathrm{d}p = \int_0^{c} e^{-u}\, f\Big(\frac{u}{k}\Big)\,\frac{\mathrm{d}u}{k}.$$

Hence

$$\int_0^{c/k} e^{-k\,p}\, f(p)\,\mathrm{d}p = \frac{1}{k}\int_0^{c} e^{-u}\, f\Big(\frac{u}{k}\Big)\,\mathrm{d}u.$$

(b) The error part. The second term in equation 43 has $\Delta_k(p) = R_k(p) - 1$ satisfying $|\Delta_k(p)| \le \frac{M}{k}$. So

$$\left|\int_0^{c/k} e^{-k\,p}\, f(p)\,\Delta_k(p)\,\mathrm{d}p\right| \;\le\; \frac{M}{k}\int_0^{c/k} e^{-k\,p}\, f(p)\,\mathrm{d}p.$$

But the integral $\int_0^{c/k} e^{-k\,p}\, f(p)\,\mathrm{d}p$ is precisely the leading part we just considered. Thus the error is bounded by $\frac{M}{k}$ times a term that will turn out to be $\Theta(k^{-b})$. Hence the error is subleading if $b < 1$ is not the case—but even then, we can keep track of it systematically.

Overall, combining both terms, we get

$$I_{k,\text{left}} \;=\; \frac{1}{k}\int_0^{c} e^{-u}\, f\Big(\frac{u}{k}\Big)\,\mathrm{d}u \;+\; O\big(\tfrac{1}{k}\cdot(\text{leading integral})\big). \tag{44}$$

(c) Matching $\Theta(k^{-b})$. Since $I_k = I_{k,\text{left}} + I_{k,\text{right}}$ with $I_{k,\text{right}}$ negligible, we have

$$I_k \;=\; \frac{1}{k}\int_0^{c} e^{-u}\, f\Big(\tfrac{u}{k}\Big)\,\mathrm{d}u \;+\; (\text{small corrections}).$$

But by hypothesis, $I_k \sim \alpha\,k^{-b}$. Thus

$$k\cdot I_k \;=\; \int_0^{c} e^{-u}\, f\Big(\tfrac{u}{k}\Big)\,\mathrm{d}u \;+\; (\text{smaller terms}) \;\sim\; \alpha\,k^{1-b}. \tag{45}$$

Hence the expression

$$\int_0^{c} e^{-u}\, f\Big(\tfrac{u}{k}\Big)\,\mathrm{d}u$$

must be $\Theta\big(k^{\,1-b}\big)$ for large k. Since $\frac{u}{k}$ is small for $0 \le u \le c$, we are effectively sampling f near 0. For the integral to produce $k^{\,1-b}$, we deduce

$$f\Big(\tfrac{u}{k}\Big) \;=\; \Theta\Big(\Big(\tfrac{u}{k}\Big)^{b-1}\Big),$$

i.e. f must behave like $p^{\,b-1}$ near $p = 0$. Rewriting the constant in front, one obtains

$$f\Big(\tfrac{u}{k}\Big) \;=\; \Big(\tfrac{u}{k}\Big)^{b-1}\,\big[\text{some positive constant}\big].$$

(We then identify that constant with $\frac{\alpha}{\Gamma(b)}$ by matching the integral precisely, just as in the prior argument.)

Step 5. Conclusion. We have thus shown that over $p \in [0, c/k]$, one has

$$(1-p)^k \;=\; e^{-k\,p}\,\big[\,1 + O(\tfrac{1}{k})\big],$$

and upon integrating, the required k^{-b} form for I_k forces

$$f(p) \;=\; \frac{A}{\Gamma(b)}\,p^{\,b-1} \;+\; o\big(p^{\,b-1}\big), \quad \text{as } p \to 0^+.$$

This completes the necessity proof.

□

Remark (Mild Regularity). If f had bizarre oscillations or nonintegrable singularities near 0, the integral $\int_0^1 (1-p)^k\, f(p)\,\mathrm{d}p$ might not produce a clean k^{-b}. Typically, we impose monotonicity or at least continuity near $p = 0$, no atom at $p = 0$, and $f(0) = 0$ if $b > 1$ or $f(0) > 0$ if $b = 1$, etc. These assumptions exclude pathological behaviors and guarantee that the local shape of $f(p)$ drives a clean power law.

F. Maximum Likelihood Estimation of Scaled Beta-Binomial Distribution

To model the distribution of $\text{pass}_\text{i}@1$, we can perform maximum likelihood estimation on a scaled three-parameter Beta-Binomial distribution, which we chose because each attempt on the i-th problem is an i.i.d. Bernoulli random variable with success probability $\text{pass}_\text{i}@1$, and we introduced a scale parameter because the largest $\text{pass}_\text{i}@1$ values were typically 1-2 orders of magnitude less than 1.0 (the maximum of the unscaled beta distribution's support).

In greater detail, as background, the 4-parameter Beta distribution has PDF

$$p_Y(y; \alpha, \beta, a, c) \stackrel{\text{def}}{=} \frac{(y-a)^{\alpha-1}(c-y)^{\beta-1}}{(c-a)^{\alpha+\beta-1}\,\mathrm{B}(\alpha,\beta)}, \tag{46}$$

where $\mathrm{B}(\cdot,\cdot)$ is the Beta function. If the minimum a is fixed at 0 and the maximum c is constrained to $a < c < 1$, then the scaled three parameter Beta distribution simplifies to:

$$f_P(p; \alpha, \beta, a=0, c) = \frac{p^{\alpha-1}(c-p)^{\beta-1}}{c^{\alpha+\beta-1}\,\mathrm{B}(\alpha,\beta)}. \tag{47}$$

We want the PMF of a three-parameter Beta-Binomial distribution based on this scaled Beta distribution. For n samples and x successes, the PMF is:

$$P(X=x; \alpha, \beta, c, n) \stackrel{\text{def}}{=} \int_0^c \binom{n}{x} p^x\,(1-p)^{n-x}\,f_P(p; \alpha, \beta, a=0, c)\,dp \tag{48}$$

$$= \binom{n}{x} \frac{1}{c^{\alpha+\beta-1}\,\mathrm{B}(\alpha,\beta)} \int_0^c p^{x+\alpha-1}\,(1-p)^{n-x}\,(c-p)^{\beta-1}\,dp. \tag{49}$$

Using a change of variable $p \stackrel{\text{def}}{=} c\,z$, the PMF can be rewritten as

$$P(X=x; \alpha, \beta, c, n) = \binom{n}{x} \frac{c^x}{\mathrm{B}(\alpha,\beta)} \int_0^1 z^{x+\alpha-1}\,(1-z)^{\beta-1}\,(1-cz)^{n-x}\,dz \tag{50}$$

$$= \binom{n}{x} \frac{c^x \mathrm{B}\big(x+\alpha,\ \beta\big)}{\mathrm{B}(\alpha,\beta)}\ {}_2F_1\Big(-(n-x),\ x+\alpha;\ x+\alpha+\beta;\ c\Big), \tag{51}$$

where ${}_2F_1(\cdot,\cdot;\cdot;\cdot)$ is the (Gauss) hypergeometric function.

G. Maximum Likelihood Estimation of Scaled Kumaraswamy-Binomial Distribution

To model the distribution of $\text{pass}_\text{i}@1$, we can perform maximum likelihood estimation on a scaled three-parameter Kumaraswamy-Binomial distribution, which we chose because each attempt on the i-th problem is an i.i.d. Kumaraswamy random variable with success probability $\text{pass}_\text{i}@1$, and we introduced a scale parameter because the largest $\text{pass}_\text{i}@1$ values were typically 1-2 orders of magnitude less than 1.0 (the maximum of the unscaled beta distribution's support).

In greater detail, the scaled three parameter Kumaraswamy distribution simplifies to:

$$f_P(p; \alpha, \beta, a = 0, c) = \frac{\alpha\beta}{c^\alpha}\, p^{\alpha-1}\, (1 - (p/c)^\alpha)^{\beta-1}, \tag{52}$$

over the support $(0, c)$. The rescaled Kumaraswamy-Binomial distribution then has PMF:

$$P(X = x; \alpha, \beta, c, n) = \binom{n}{x} \frac{\alpha\,\beta}{c^\alpha} \int_0^c p^{\,x+\alpha-1}\, (1-p)^{n-x} \left(1 - \left(\tfrac{p}{c}\right)^\alpha\right)^{\beta-1} dp. \tag{53}$$

One can perform a change of variable $p \overset{\text{def}}{=} cz$, but simplifying yields sums of hypergeometric functions that add little conceptual clarity and so we resort to numerical integration using Python's mpmath library (mpmath development team, 2023).

H. Maximum Likelihood Estimation of Scaled Beta-Negative Binomial Distribution

To model the distribution of $\text{pass}_i@1$, we can perform maximum likelihood estimation on a scaled three-parameter Beta-Negative Binomial distribution. Recall that the scaled three parameter Beta distribution is:

$$f_P(p;\alpha,\beta,a=0,c)=\frac{p^{\alpha-1}(c-p)^{\beta-1}}{c^{\alpha+\beta-1}\,\mathrm{B}(\alpha,\beta)}. \tag{54}$$

We want the PMF of a three-parameter Beta-Negative Binomial distribution based on this scaled Beta distribution. For r desired successes, the PMF that we first draw x failures is:

$$P(X=x;\alpha,\beta,c,r)=\int_0^c \underbrace{\binom{x+r-1}{x}p^r\,(1-p)^x}_{\text{NegBin}(r,p)}\ \underbrace{\frac{p^{\alpha-1}\,(c-p)^{\beta-1}}{c^{\alpha+\beta-1}\,\mathrm{B}(\alpha,\beta)}}_{\text{scaled Beta PDF}}\ dp \tag{55}$$

$$=\binom{x+r-1}{x}\frac{1}{c^{\alpha+\beta-1}\,\mathrm{B}(\alpha,\beta)}\int_0^c p^{\,r+\alpha-1}\,(1-p)^x\,(c-p)^{\beta-1}\,dp. \tag{56}$$

Next, substitute $p = c\,z \Longrightarrow dp = c\,dz$ which rescales the domain $[0,c]$ to $[0,1]$. Under this change:

$$p^{\,r+\alpha-1} = (c\,z)^{\,r+\alpha-1} = c^{\,r+\alpha-1}\,z^{\,r+\alpha-1},$$

$$(c-p)^{\beta-1} = \big(c-c\,z\big)^{\beta-1} = \big(c(1-z)\big)^{\beta-1} = c^{\,\beta-1}\,(1-z)^{\beta-1},$$

$$(1-p)^x = \big(1-c\,z\big)^x.$$

Putting these into the integrand:

$$p^{\,r+\alpha-1}\,(1-p)^x\,(c-p)^{\beta-1}\,dp = \Big(c^{\,r+\alpha-1}\,z^{\,r+\alpha-1}\Big)\Big((1-cz)^x\Big)\Big(c^{\,\beta-1}\,(1-z)^{\beta-1}\Big)\,(c\,dz).$$

Factor out the constants in c:

$$= c^{\,r+\alpha-1}\,c^{\,\beta-1}\,c\ \ z^{\,r+\alpha-1}\,(1-cz)^x\,(1-z)^{\beta-1}\,dz.$$

Since $c^{\,r+\alpha-1}\cdot c^{\,\beta-1}\cdot c = c^{\,r+\alpha+\beta-1}$, we get

$$p^{\,r+\alpha-1}\,(1-p)^x\,(c-p)^{\beta-1}\,dp = c^{\,r+\alpha+\beta-1}\ z^{\,r+\alpha-1}\,(1-z)^{\beta-1}\,(1-cz)^x\ dz.$$

Plugging back into $P(X=x;\alpha,\beta,c,r)$ and simplifying:

$$P(X=x;\alpha,\beta,c,r)=\binom{x+r-1}{x}\frac{c^r}{\mathrm{B}(\alpha,\beta)}\ \int_0^1 z^{\,r+\alpha-1}\,(1-z)^{\beta-1}\,\big(1-c\,z\big)^x\,dz. \tag{57}$$

We can re-express this using the (Gauss) hypergeometric function ${}_2F_1(\cdot,\cdot;\cdot;\cdot)$:

$$P(X=x;\alpha,\beta,c,r) = \binom{x+r-1}{x}\frac{c^r\,\mathrm{B}\big(r+\alpha,\ \beta\big)}{\mathrm{B}\big(\alpha,\ \beta\big)}\ {}_2F_1\Big(-x,\ r+\alpha;\ r+\alpha+\beta;\ c\Big). \tag{58}$$