Pretraining Scaling Laws for Generative Evaluations of Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Neural scaling laws have played a central role in modern machine learning, driving the field's ever-expanding scaling of parameters, data and compute. While much research has gone into fitting scaling laws and predicting performance on pretraining losses and on *discriminative* evaluations such as multiple-choice question-answering benchmarks, comparatively little research has been done on fitting scaling laws and predicting performance on *generative* evaluations such as mathematical problem-solving or coding. In this work, we propose and evaluate two different pretraining scaling laws for fitting pass-at-k on generative evaluations and for predicting pass-at-k of the most expensive model using the performance of cheaper models. Our three scaling laws differ in the covariates used: (1) model parameters and pretraining tokens, (2) log likelihoods of gold reference solutions. We make four main contributions: First, we show how generative evaluations offer new hyperparameters (in our setting, k) that researchers can use to control the scaling laws parameters and the predictability of performance. Second, in terms of scaling law parameters, we find that the parameters + tokens scaling law stabilize for the last $\sim 1.5 - 2.5$ orders of magnitude, whereas the gold reference likelihood scaling law stabilizes for the last ~ 5 orders of magnitude. Third, in terms of predictive performance, we find both scaling laws perform comparably, although the log likelihoods of gold reference solutions predicts slightly worse for large k. Our framework provides researchers and practitioners with insights and methodologies to forecast generative performance, accelerating progress toward models that can reason, solve, and create.

1 Introduction

Neural scaling laws, which predictably map resources to the performance of neural networks, are foundational for developing frontier AI systems (Kaplan et al., 2020b; Hoffmann et al., 2022b). Such empirical regularities for improving model performance from scaling parameters, data, and compute have driven pursuit of ever-larger models. Much of the research in neural scaling laws has focused on fitting and predicting model performance on pretraining losses, e.g., (Kaplan et al., 2020b; Hoffmann et al., 2022b; Clark et al., 2022; Hernandez et al., 2022; Sardana et al., 2024; Muennighoff et al., 2023; Porian et al., 2024; Pearce & Song, 2024) and on "downstream" *discriminative* evaluations like multiple-choice question-answering, e.g., (Schaeffer et al., 2023; Gadre et al., 2024; Schaeffer et al., 2024; Grattafiori et al., 2024; Chen et al., 2025; Bhagia et al., 2025).

While both pretraining losses and discriminative evaluations are undeniably useful, many capabilities we care about are *generative*: writing books, autoformalizing mathematics, or conducting cutting-edge research. Comparatively little research has investigated scaling laws for generative evaluations (OpenAI et al., 2024; Hu et al., 2024); for a deeper discussion of related work, please see Appendix B. Unlike pretraining losses and discriminative evaluations, generative evaluations

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

introduce new considerations such as which metric is used to quantify performance and how samples are drawn from the model, including the sampling temperature (Ackley et al., 1985) and the sampling algorithm, e.g., top-k (Fan et al., 2018), top-p (Holtzman et al., 2020).

In this work, we propose a framework for fitting scaling laws and predicting performance on generative evaluations, focusing on a particular setting: benchmarks with verifiable binary rewards, with multiple attempts per problem, with performance scored using the "pass-at-k" metric (Kulal et al., 2019; Chen et al., 2021). We chose this setting due to its widespread use in the literature (First et al., 2023; Brown et al., 2024; Hassid et al., 2024; Hughes et al., 2024; Chen et al., 2024; Ehrlich et al., 2025; Kwok et al., 2025). In this setting, we make the following contributions:

1. We propose three scaling laws that fit and predict pass-at-k from two different covariates: (1) model parameters and pretraining tokens (Sec. 3) and (2) likelihoods of gold reference solutions (Sec. 4).
2. We reveal that a hyperparameter specific to this setting – the number of attempts per problem k – offers a new lever to change the parameters of the scaling laws as well as how accurately the scaling laws extrapolate to more expensive models.
3. We quantify how stable each scaling law's parameters are, finding that parameters the parameters + tokens scaling law stabilize by the last $\sim 1.5 - 2.5$ orders of magnitude of pretraining compute, whereas parameters for the gold reference likelihood scaling law are stable over the last ~ 5 orders of magnitude.
4. We quantify how predictive each scaling law is, finding that all three comparably predict the performance of the most expensive model based on cheaper models.

2 Methodology

Benchmark To study how language model performance changes during pretraining on generative evaluations, we used **GSM8K** (Cobbe et al., 2021), a well-established benchmark containing $8,500$ grade school math word problems designed to evaluate a model's ability to perform multi-step mathematical reasoning. Each attempt at solving a problem yields a binary outcome: the attempt is either successful or unsuccessful. Due to limited budget, we used a randomly chosen subset of 128 problems. See Appendix C for a description of how many samples we drew per problem per model.

Metric To evaluate the performance of each model, we used the ubiquitous pass-at-k metric (Kulal et al., 2019). For the i-th problem, a model's pass rate at k attempts is the probability of the event that *any* of k attempts are successful. Because generative evaluations sample from the language model, this per-problem pass rate must be estimated. We use the unbiased, low variance and numerically stable estimator introduced by Chen et al. (2021), which, for each problem, draws $n_i > k$ samples, counts the number of successes $s_i \in [0, n_i]$, and then averages over all subsets of size k:

$$\mathrm{pass_i@k} \quad \stackrel{\text{def}}{=} \quad 1 - \binom{n_i - s_i}{k} \Big/ \binom{n_i}{k}. \tag{1}$$

For a benchmark $\mathcal{B}$ of $|\mathcal{B}|$ problems, a model's benchmark pass rate at k attempts is estimated as:

$$\mathrm{pass}_{\mathcal{B}}\mathrm{@k} \quad \stackrel{\text{def}}{=} \quad \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathrm{pass_i@k} \tag{2}$$

Pass rates are sometimes phrased as the fraction of problems in the benchmark solved ("coverage"), but these two quantities are mathematically equivalent (Brown et al., 2024; Schaeffer et al., 2025b).

For the **sampling** and **language model family** descriptions, see the appendix sectio I.

3 Fitting and Predicting Pass Rates from Parameters and Tokens

We assessed how well the benchmark $\mathrm{pass}_{\mathcal{B}}\mathrm{@k}$ can be (1) fit and (2) predicted as a function of the number of model parameters N and pretraining tokens D.

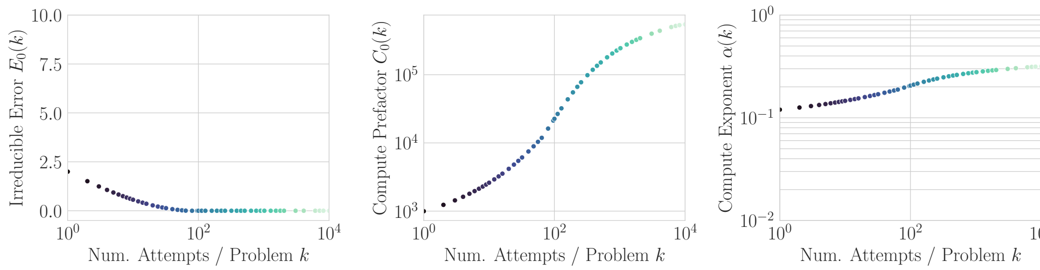

Figure 1: **Scaling Law Parameters Vary with Attempts per Problem (Full Fit).** In comparison with pretraining loss scaling laws and discriminative evaluation scaling laws, the number of attempts per problem k introduces a new hyperparameter that affects the parameters of generative evaluation scaling laws, and also how predictable $\text{pass}_{\mathcal{B}}@k$ is. Here we show the effect of k for the pretraining compute scaling law (Eqn. 6); for a side-by-side comparison with our next two scaling laws, see Appendix D. **Left:** The irreducible error $E_0(k)$ falls roughly exponentially with k and is ≈ 0 by $k \approx 10^2$, consistent with the probability that a problem remains unsolved decaying exponentially in the number of attempts. **Center:** The compute prefactor $C_0(k)$ increases monotonically with k, indicating that once $E_0(k) \to 0$, variation is dominated by the compute-dependent term. **Right:** The compute exponent $\alpha(k)$ increases smoothly, from ~ 0.15 at $k{=}1$ to ~ 0.3 at $k{=}10^4$, implying steeper and more regular scaling at larger k. Hue corresponds to k.

3.1 Fitting Pass Rates from Model Parameters and Pretraining Tokens

Motivated by Hoffmann et al. (2022b), we posited that the negative log of benchmark pass rates might instead follow a scaling law as a function of model parameters N and pretraining tokens D:

$$-\log\big(\text{pass}_{\mathcal{B}}@k\big)(N, D, k) = \mathcal{E}_0(k) + \frac{N_0(k)}{N^{\beta(k)}} + \frac{D_0(k)}{D^{\gamma(k)}}, \tag{3}$$

where $\mathcal{E}_0(k)$ is the irreducible error, $N_0(k)$ is the parameter scaling prefactor, $\beta(k)$ is the parameter scaling coefficient, $D_0(k)$ is the token scaling prefactor and $\gamma(k)$ is the token scaling coefficient.

We fit this parameters + tokens scaling law. Visually, this law provides a better characterization of performance for the full range of models, with the fitted curves appearing tighter and reducing the residual scatter (Fig.9). The irreducible error term $\widehat{\mathcal{E}}_0(k)$ decreases with the number of attempts per problem k, although it does so more gradually, reaching zero by $k \approx 3 \times 10^2$ (Fig. 5). However, despite the improved global fit, we observed that the largest-compute model checkpoints in our dataset exhibit a comparatively large relative error, suggesting that while this law better explains the performance of less-trained models, it may struggle with extrapolation to the frontier.

3.2 Predicting Pass Rates from Model Parameters and Pretraining Tokens

We evaluated how predictive the parameters + tokens scaling law is using the same backtesting approach (Section F.2). We again used Pythia 12B model trained for 300B tokens as the target, and iteratively fit Eqn. 3. The relative error in predicting the target model's performance decreases and then plateaus as more capable models are included in the fit (Fig. 10, Top Left). For all values of k, these extrapolations become reliable once the largest checkpoint used for fitting is within approximately $\sim 2-2.5$ orders of magnitude of the target's compute. Similarly, the five backtested scaling parameters converge to their full-fit values once models within this same compute range of $\sim 1.5 - 2.5$ orders of magnitude are included (Fig. 10, Other Five Panels). For small k, this law shows a slight predictive advantage over the compute-only baseline (see Appendix K), an effect that vanishes at larger k.

4 Fitting and Predicting Pass Rates from Gold Reference Likelihoods

Generative evaluations oftentimes contain not just problems and correct answers, but also "gold references": high quality responses that reach the correct answer from the given problem. For autoregressive language models, one can straightforwardly compute the likelihood of the gold

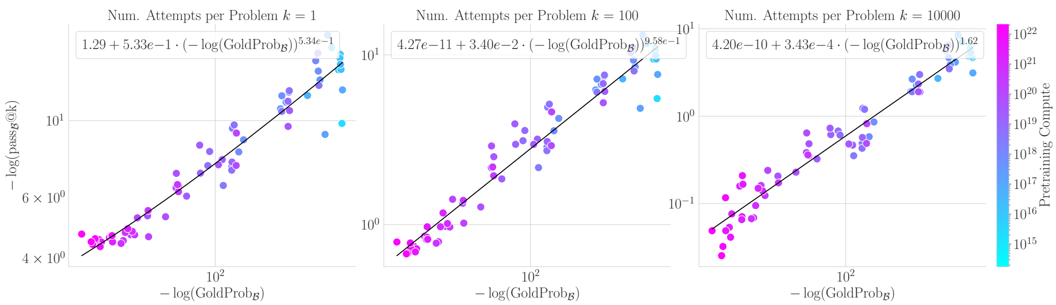

Figure 2: **Scaling of Pass Rates with Gold Reference Likelihood (Full Fit).** Each panel fits Eqn. 5, $-\log(\text{pass}_{\mathcal{B}}@k) = \xi_0 + K_0(k) \cdot [-\log\big(\text{GoldProb}_{\mathcal{B}}\big)]^{\kappa(k)}$, for $k \in \{1, 10^2, 10^4\}$. Regressing the gold reference log-likelihoods against the pass rates produces a visually tighter fit with less residual scatter than the compute scaling law (cf. Fig. 7). As k increases, the irreducible error $\widehat{\xi}_0(k)$ falls toward zero and the exponent $\widehat{\kappa}(k)$ increases, making the relationship a steeper, purer power law.

reference given the problem, as well as the arithmetic mean over the benchmark's gold reference probabilities:

$$\text{GoldProb}_{\mathcal{B}} \overset{\text{def}}{=} \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} p_\theta(\text{Gold Reference}_i | \text{Problem}_i), \tag{4}$$

For our gold reference likelihood scaling law, we assessed how well the benchmark $\text{pass}_{\mathcal{B}}@\text{k}$ can be (1) fit and (2) predicted as a function of the log likelihoods of the gold references.

4.1 Fitting Pass Rates from Gold Reference Likelihoods

Motivated by Schaeffer et al. (2024), we tested whether this average likelihood of the gold responses can be accurately regressed to fit and predict benchmark pass rates $\text{pass}_{\mathcal{B}}@k$.

$$-\log\big(\text{pass}_{\mathcal{B}}@k\big)(\text{GoldProb}_{\mathcal{B}}, k) = \xi_0(k) + K_0(k) \cdot \Big[-\log\big(\text{GoldProb}_{\mathcal{B}}\big)\Big]^{\kappa(k)} \tag{5}$$

Perhaps surprisingly, we discovered that this straightforwardly computable quantity serves as an accurate predictor of benchmark pass rates. As shown in Figure 2, regressing against the negative log likelihood of the gold references produces a visually tighter fit than the pretraining compute law (Fig. 7). The residual scatter is reduced, and the relationship between $-\log(\text{pass}_{\mathcal{B}}@k)$ and $-\log(\text{GoldProb}_{\mathcal{B}})$ has a smaller reducible error and more closely resembles a power law. We next examined how the number of attempts per problem k influences the scaling law's parameters (Fig. 6). Consistent with our findings for the prior two scaling laws, the irreducible error term $\xi_0(k)$ collapses toward zero as k increases and the exponent $\kappa(k)$ increases smoothly.

4.2 Predicting Pass Rates from Gold Reference Likelihoods

Using our backtesting approach once more, we evaluated how well the gold references scaling law can predict the performance of our most expensive model. For small k, the gold reference likelihoods scaling law has slightly lower relative error at predicting the most expensive model's performance, but worsens for larger k (Fig. 11 Upper Left). The most remarkable result, however, is the stability of the scaling law's parameters. In stark contrast to the parameter + token scaling laws whose parameters only stabilized when fit on models within $\sim$2 orders of magnitude of the target, the parameters for the gold reference law are exceptionally robust: the backtested estimates for the three parameters converge to their final values when fitting on models up to $\sim$5 of magnitude cheaper than the target (Fig. 11, Other Three Panels) and for all values of k. This significantly stable finding suggests that the relationship between gold reference likelihoods and pass rates is a consistent signal, enabling long-range forecasts from comparatively inexpensive models.

References

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

Akshita Bhagia, Jiacheng Liu, Alexander Wettig, David Heineman, Oyvind Tafjord, Ananya Harsh Jha, Luca Soldaini, Noah A. Smith, Dirk Groeneveld, Pang Wei Koh, Jesse Dodge, and Hannaneh Hajishirzi. Establishing task scaling laws via compute-efficient model ladders. In *Second Conference on Language Modeling*, 2025. URL `https://openreview.net/forum?id=FeAM2RVO8l`.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, 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.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2024. URL `https://arxiv.org/abs/2407.21787`.

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more llm calls all you need? towards scaling laws of compound inference systems, 2024. URL `https://arxiv.org/abs/2403.02419`.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL `https://arxiv.org/abs/2107.03374`.

Yangyi Chen, Binxuan Huang, Yifan Gao, Zhengyang Wang, Jingfeng Yang, and Heng Ji. Scaling laws for predicting downstream performance in LLMs. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=PJUbMDkQVY`.

Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. Unified scaling laws for routed language models. In *International conference on machine learning*, pp. 4057–4086. PMLR, 2022.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL `https://arxiv.org/abs/2412.19437`.

Ryan Ehrlich, Bradley Brown, Jordan Juravsky, Ronald Clark, Christopher Ré, and Azalia Mirhoseini. Codemonkeys: Scaling test-time compute for software engineering, 2025. URL `https://arxiv.org/abs/2501.14723`.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

Emily First, Markus N. Rabe, Talia Ringer, and Yuriy Brun. Baldur: Whole-proof generation and repair with large language models, 2023. URL `https://arxiv.org/abs/2303.04910`.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Jenia Jitsev, Luca Soldaini, Alexandros G. Dimakis, Gabriel Ilharco, Pang Wei Koh, Shuran Song, Thomas Kollar, Yair Carmon, Achal Dave, Reinhard Heckel, Niklas Muennighoff, and Ludwig Schmidt. Language models scale reliably with over-training and on downstream tasks, 2024. URL `https://arxiv.org/abs/2403.08540`.

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

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. Scaling laws for neural machine translation. In *International Conference on Learning Representations*, 2021.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. Why do small language models underperform? studying language model saturation via the softmax bottleneck, 2024. URL `https://arxiv.org/abs/2404.07647`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet,

Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia

Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Michael Hassid, Tal Remez, Jonas Gehring, Roy Schwartz, and Yossi Adi. The larger the better? improved llm code-generation via budget reallocation. In *First Conference on Language Modeling*, 2024.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically, 2017. URL https://arxiv.org/abs/1712.00409.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

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

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

Shengding Hu, Xin Liu, Xu Han, Xinrong Zhang, Chaoqun He, Weilin Zhao, Yankai Lin, Ning Ding, Zebin Ou, Guoyang Zeng, Zhiyuan Liu, and Maosong Sun. Predicting emergent abilities with infinite resolution evaluation, 2024. URL https://arxiv.org/abs/2310.03262.

John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, and Mrinank Sharma. Best-of-n jailbreaking, 2024. URL `https://arxiv.org/abs/2412.03556`.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020a.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020b. URL `https://arxiv.org/abs/2001.08361`.

Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy S Liang. Spoc: Search-based pseudocode to code. *Advances in Neural Information Processing Systems*, 32, 2019.

Jacky Kwok, Christopher Agia, Rohan Sinha, Matt Foutter, Shulu Li, Ion Stoica, Azalia Mirhoseini, and Marco Pavone. Robomonkey: Scaling test-time sampling and verification for vision-language-action models, 2025. URL `https://arxiv.org/abs/2506.17811`.

Noam Itzhak Levi. A simple model of inference scaling laws. In *Forty-second International Conference on Machine Learning*, 2025.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted

Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

Tim Pearce and Jinyeop Song. Reconciling kaplan and chinchilla scaling laws. *Transactions on Machine Learning Research*, 2024.

Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving discrepancies in compute-optimal scaling of language models. *Advances in Neural Information Processing Systems*, 37:100535–100570, 2024.

Qwen, An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

Daniel A Roberts, Sho Yaida, and Boris Hanin. *The principles of deep learning theory*, volume 46. Cambridge University Press Cambridge, MA, USA, 2022.

Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. In *International Conference on Machine Learning*, pp. 43445–43460. PMLR, 2024.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (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`.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. Why has predicting downstream capabilities of frontier ai models with scale remained elusive?, 2024. URL `https://arxiv.org/abs/2406.04391`.

Rylan Schaeffer, Joshua Kazdan, and Yegor Denisov-Blanch. Min-p, max exaggeration: A critical analysis of min-p sampling in language models, 2025a. URL `https://arxiv.org/abs/2506.13681`.

Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get their power (laws)? In *Forty-second International Conference on Machine Learning*, 2025b. URL `https://openreview.net/forum?id=QqVZ28qems`.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao,

Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kepa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah O Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu,

Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2024. URL `https://arxiv.org/abs/2312.11805`.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models, 2024. URL `https://arxiv.org/abs/2408.00724`.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12104–12113, 2022.

Zhipu-AI, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihan Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguang Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models, 2025. URL `https://arxiv.org/abs/2508.06471`.

A Language Model Usage

Language models were used by the authors to aid or polish the writing of the paper. Authors take full responsibility for the content.

B Related Work

The study of neural scaling laws, which characterize how model performance predictably improves with resources, has become a cornerstone of modern large-scale machine learning (Hestness et al., 2017; Kaplan et al., 2020a; Hoffmann et al., 2022a; Roberts et al., 2022; Bahri et al., 2024). Foundational work demonstrated that pretraining loss follows a power law with respect to model size, dataset size, and training compute (Kaplan et al., 2020a; Henighan et al., 2020; Hoffmann et al., 2022a). These principles now guide the development of nearly all state-of-the-art language models (OpenAI et al., 2024; Team et al., 2024; Grattafiori et al., 2024; DeepSeek-AI et al., 2025; Qwen et al., 2025; Zhipu-AI et al., 2025). This research area has since expanded to explore scaling properties in diverse domains, including computer vision (Zhai et al., 2022), machine translation (Ghorbani et al., 2021), and even for optimizing inference-time compute rather than training (Wu et al., 2024; Snell et al., 2024). Despite this breadth, scaling laws for complex, multi-step *generative tasks*—which often represent the ultimate goal of language modeling—have remained comparatively underexplored due to the challenges of evaluation. Our work builds most directly on recent efforts to forecast performance on such generative evaluations. The *GPT-4 Technical Report* pioneered this direction by showing that the negative log pass rate on a subset of the HumanEval coding benchmark scaled predictably as a power law of pretraining compute (OpenAI et al., 2024). They successfully predicted GPT-4's performance by extrapolating from smaller models over approximately three orders of magnitude (OpenAI et al., 2024). We adopt their proposed functional form for our compute-based law and extend their work by systematically analyzing how the number of attempts, k, reshapes the law's parameters. Furthermore, our rigorous backtesting reveals the brittleness of this compute-only law, finding it reliable only across 1-2 orders of magnitude in our setting. Separately, our work shares a philosophical motivation with Hu et al. (2024), who investigate emergent abilities. They argue that seemingly unpredictable jumps in performance on discriminative tasks can be understood as smooth, continuous improvements when measured by the model's underlying log-likelihoods. We apply a similar insight to a different problem: forecasting. We operationalize the log probability of gold-standard solutions not just as an explanatory tool, but as a direct covariate in a predictive scaling law for generative tasks, and we empirically demonstrate its superior stability for long-range extrapolation.

C Methodology: Number of Samples Per Problem Per Model

Because sampling from language models is expensive and because our compute budget is limited, we needed to judiciously choose how many samples to draw per problem per model. A key consideration is that the number of samples per problem sets the resolution, i.e., if we draw n samples per problem, then any pass rate on that problem below $1/n$ will appear to be 0. Since pass rates typically increase with pretraining compute, we chose a loose heuristic based on the pretraining compute of each model. For a model with C pretraining FLOP, we aimed to sample

$$\text{Num. Samples per Problem } = \text{clip}(100 \cdot 10^{23-\log_{10}(C)}, 3.2e4, 1e6).$$

$3.2e4$ was chosen so that we would have more samples per problem than the largest number of attempts per problem k we considered $(1e4)$. Fig. 3 shows how many samples per problem per model we had at the time our analyses were conducted. Additional sampling would have likely benefited our analyses.

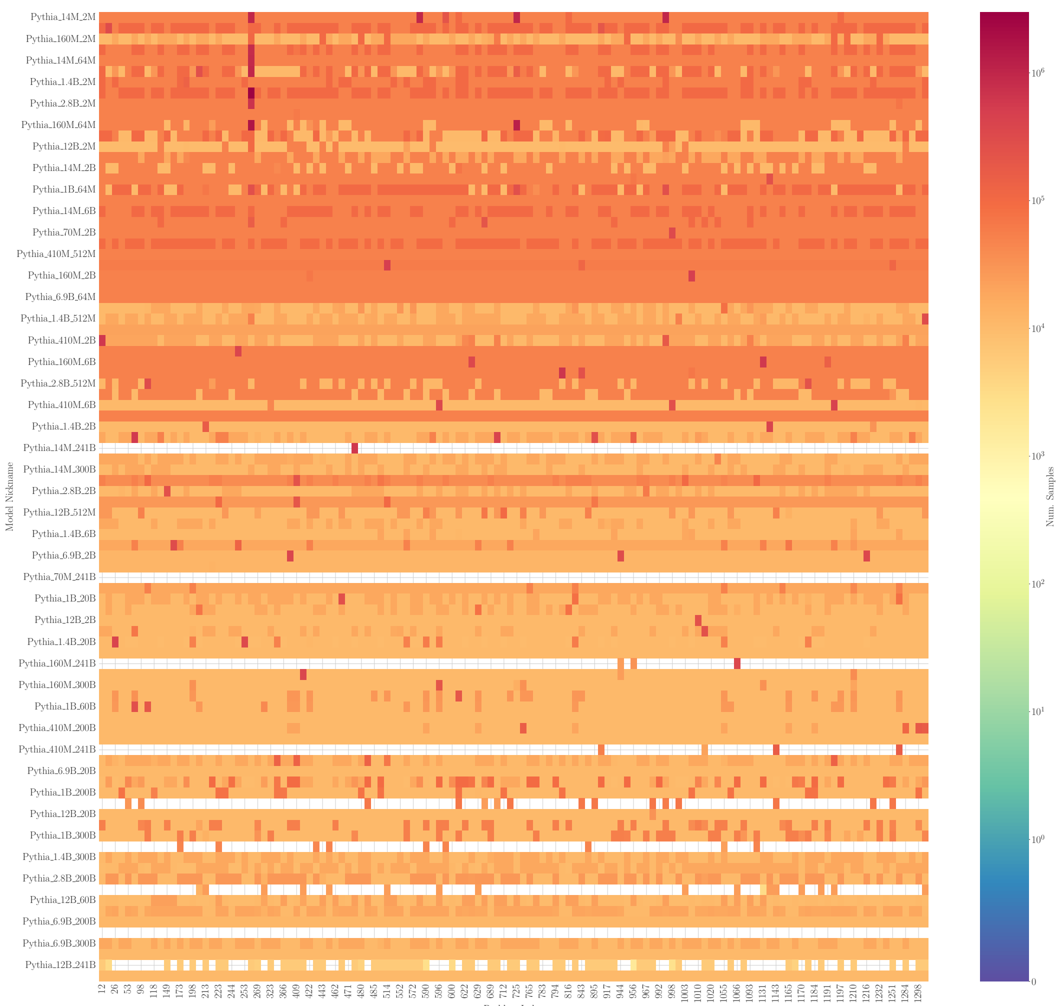

Figure 3: **Number of Samples per Model per Problem.** This heatmap visualizes how many samples we had drawn per model (y-axis) per problem (x-axis) at the time our analyses were conducted.

D Scaling Law Parameters by Number of Attempts Per Problem k

For the three scaling laws we consider, we visualize how the fit parameters change as a function of the number of attempts per problem k (Fig. 4, Fig. 5, Fig. 6).

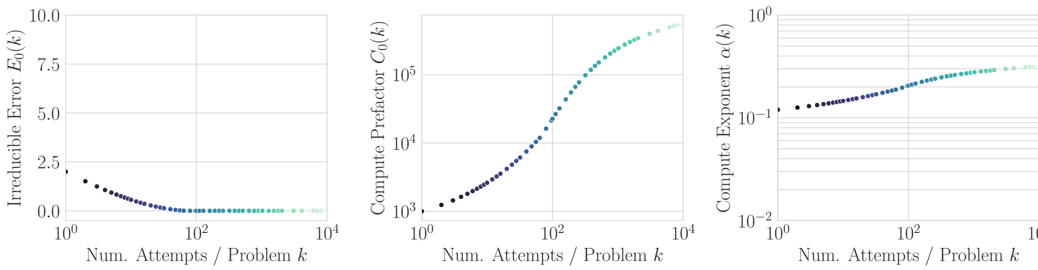

Figure 4: **Pretraining Compute Scaling Law Parameters by Number of Attempts per Problem (Full Fits).** We visualize how the parameters of Eqn. 6 change with respect to k. Hue corresponds to k.

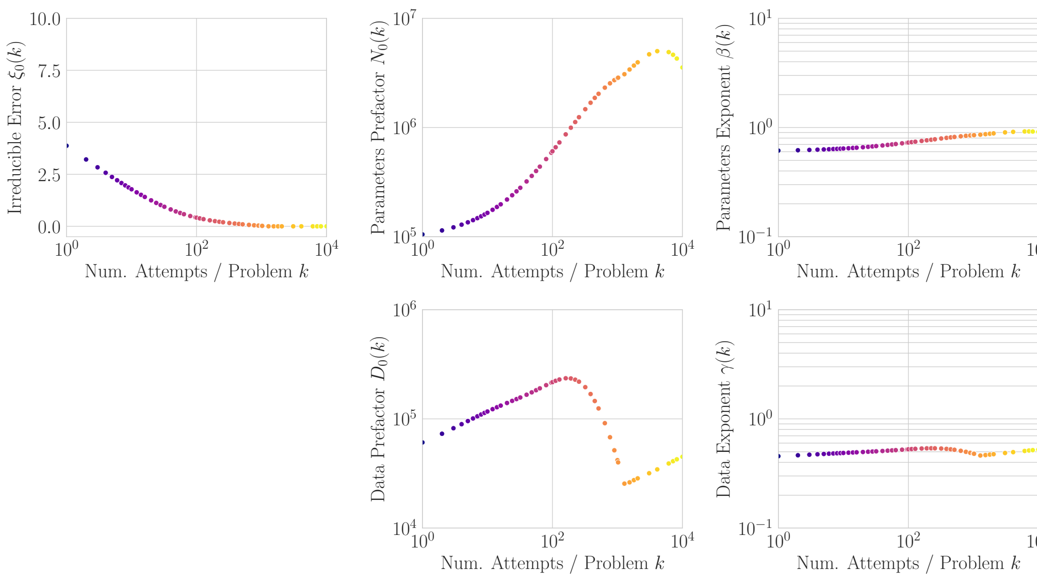

Figure 5: **Model Parameters and Tokens Scaling Law Parameters by Number of Attempts per Problem (Full Fits).** We visualize how the parameters of Eqn. 3 change with respect to k. Hue corresponds to k.

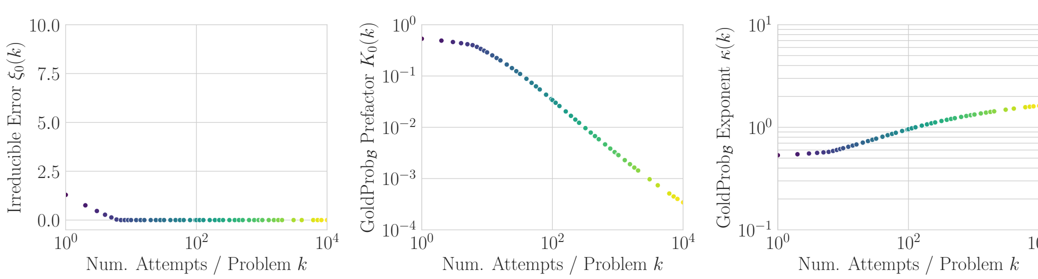

Figure 6: **Gold Reference Log Likelihoods Scaling Law Parameters by Number of Attempts per Problem (Full Fits).** We visualize how the parameters of Eqn. 4 change with respect to k. Hue corresponds to k.

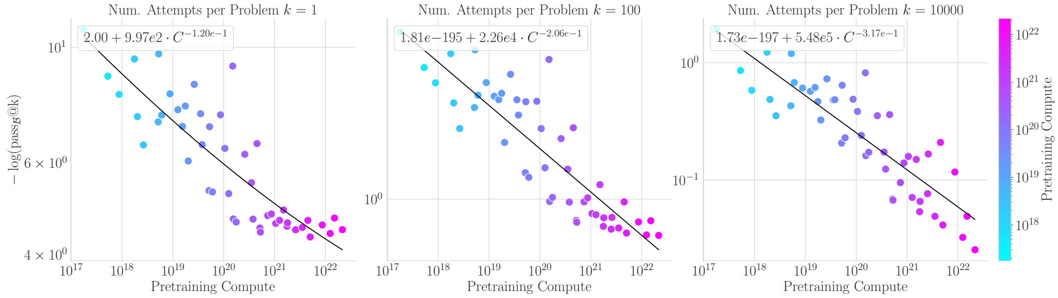

Figure 7: **Scaling of Pass Rates with Pretraining Compute (Full Fit).** Each panel fits Eqn. 6 $-\log\big(\text{pass}_{\mathcal{B}}@k\big)(C, k) = E_0(k) + C_0(k) \cdot C^{-\alpha(k)}$ to GSM8K pass rates for Pythia checkpoints across ∼5 orders of magnitude of pretraining compute. Panels show $k \in \{1, 10^2, 10^4\}$. The fit curves capture the trend moderately well and reveal two effects of increasing k: (i) the irreducible error $\widehat{E}_0(k)$ falls from $2 \rightarrow 0$ as k increases from $1 \rightarrow 100$; and (ii) the compute-dependent terms increase with k: the prefactor grows from $\widehat{C}_0(k = 1) \approx 1e3$ to $\widehat{C}_0(k = 10000) \sim 7.8e5$ and the exponent rises from $\widehat{\alpha}(k = 1) \approx 0.120$ to $\widehat{\alpha}(k = 100) \approx 0.325$. **Takeaway:** Larger k makes the relationship a pure power law and makes pass rates a steeper function of pretraining compute.

E Compute Scaling Law

We took the standard approximation for pretraining compute of $C \approx 6 \cdot N \cdot D$ (Kaplan et al., 2020b; Hoffmann et al., 2022b; Pearce & Song, 2024; Gadre et al., 2024; Schaeffer et al., 2024; Porian et al., 2024).

F Fitting and Predicting Pass Rates from Pretraining Compute

For our first scaling law, we assessed how well the benchmark $\text{pass}_{\mathcal{B}}$@k can be (1) *fit* as a function of pretraining compute C and (2) *predicted* from pretraining compute.

F.1 Fitting Pass Rates from Pretraining Compute

Motivated by the GPT-4 Technical Report (OpenAI et al., 2024), we posited that the negative log of benchmark pass rates follows a scaling law as a function of model pretraining compute C:

$$-\log\big(\text{pass}_{\mathcal{B}}@k\big)(C, k) = E_0(k) + \frac{C_0(k)}{C^{\alpha(k)}}, \tag{6}$$

where $E_0(k) \geq 0$ is the irreducible error, $C_0(k) > 0$ is the compute scaling prefactor, and $\alpha(k) > 0$ is the compute scaling exponent. In contrast with prior research on scaling laws, parameters that were previously constant are now parameterized by the number of attempts per problem k. Thus, k can be viewed as a hyperparameter that offers a new lever to change the predictability of scaling laws in a way that is not available for pretraining loss or discriminative evaluation scaling laws.

We fit the scaling law parameters based on the model checkpoints' pretraining compute and their corresponding $\text{pass}_{\mathcal{B}}@k$. Visually, the fit scaling laws captured the trend reasonably well for different values of number of attempts per problem k (Fig. 7). We next evaluated what role the number of attempts per problem k has on the scaling law parameters: As k increases, the irreducible error term $E_0(k)$ falls roughly exponentially with k and is effectively 0 by $k \approx 1 \times 10^2$ (Fig. 1 Left). This is consistent with the probability a problem goes unsolved falling exponentially with the number of attempts per problem (Levi, 2025; Schaeffer et al., 2025b). Thus, by increasing k, Eqn. 6 becomes a pure power law with no irreducible error. In comparison, as k increases, the compute prefactor and compute exponent also increase, with the compute prefactor rising ∼4 orders of magnitude (Fig. 1 Center) and the compute exponent rising moderately from 1.20×10^{-1} to 3.25×10^{-1} (Fig. 1 Right).

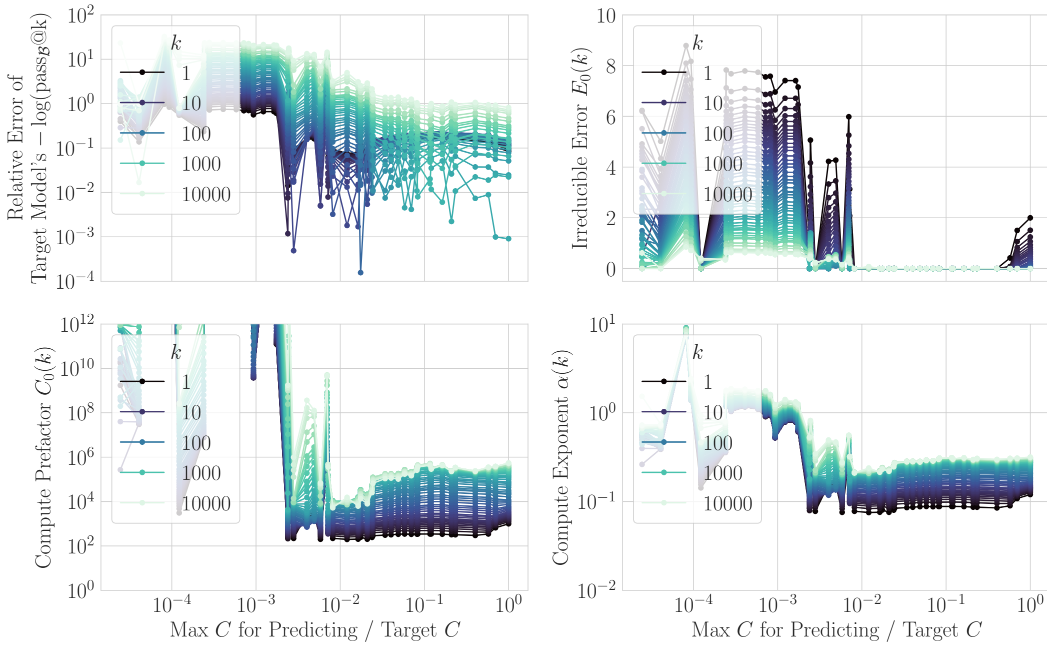

Figure 8: **Predicting Pass Rates from Scaling Pretraining Compute (Backtesting).** We evaluate predictability via *backtesting*: we iteratively fit Eq. 6 on subsets of models up to a maximum pretraining compute, then extrapolate to predict the most expensive model's $-\log\big(\text{pass}_{\mathcal{B}}@k\big)$. The most expensive model is Pythia 12B-parameter 300B-token ($\approx 2.16 \times 10^{22}$ FLOP). The x-axis in all panels is the ratio of the largest-compute checkpoint included in the fit to the target model's compute. **Top Left:** Relative error decreases as higher-compute checkpoints are included and then plateaus: for $k \in \{1, 10^2\}$ the plateau occurs once the fit includes checkpoints within $\sim$2 orders of magnitude of the target; for $k = 10^4$ it occurs within $\sim$1.5 order. **Other Three Panels:** Backtested estimates of the scaling law parameters converge as the fit includes higher-compute checkpoints, stabilizing to full-fit values once included models are within $\sim$2 orders of magnitude of the target.

F.2 Predicting Pass Rates from Pretraining Compute

Previously, we fit Eqn. 6 using all available checkpoints. However, a key motivation for scaling laws is *prediction*: can we forecast the performance of an expensive model using cheaper models?

We evaluated the predictability of pass rates via backtesting, a cross-validation-like process for evaluating forecasts in which cheaper models are used to predict the performance of the most expensive model. We fixed a target model checkpoint pretrained on C_{target} FLOP; in our work, we used Pythia-12B trained for 300B tokens, pretrained with approximately 2.16×10^{22} FLOP. For a sequence of compute caps $C_{\max} \leq C_{\text{target}}$, we: (i) fit Eq. 6 on all model checkpoints with compute $C \leq C_{\max}$ to obtain $\widehat{E}_0(k)$, $\widehat{C}_0(k)$, and $\widehat{\alpha}(k)$; (ii) extrapolate to the target compute horizon, and (iii) measure the absolute relative error to the target model's $-\log\big(\text{pass}_{\mathcal{B}}@k_{\text{target}}\big)$:

$$\text{RelativeError}(k, C_{\max}) \overset{\text{def}}{=} \frac{\text{abs}| -\log\big(\text{pass}_{\mathcal{B}}@k_{\text{target}}\big) - \widehat{E}_0(k) \ - \ \widehat{C}_0(k) \cdot C_{\text{target}}^{-\widehat{\alpha}(k)}\big)|}{-\log\big(\text{pass}_{\mathcal{B}}@k_{\text{target}}\big)}.$$

We report results as a function of the compute ratio $C_{\max}/C_{\text{target}}$.

Relative errors typically decrease and then plateau as higher-compute checkpoints are included in the fit (Fig. 8, top left). For $k \in \{1, 1e2\}$, reliable extrapolation requires that the largest checkpoint used for fitting is within $\sim$2 orders of magnitude of the target's compute, and the relative errors plateau at 1×10^{-1}. In comparison, for $k = 1e4$, reliable extrapolation requires $\sim$1.5 order of magnitude, and the relative errors plateau below 1. Similarly, the backtested estimates of the parameters $\widehat{E}_0(k)$, $\widehat{C}_0(k)$ and $\widehat{\alpha}(k)$ converge to their full-fit values once models within $\sim$1.5$-$2.5 orders of magnitude of C_{target} are included (Fig. 8, other three panels). For these models on this benchmark, the compute scaling law provides useful forecasts so long as forecasts are made using models within $\sim$2 orders of magnitude of compute as the target model; when only smaller checkpoints are available, the predictive error increases and the fitted parameters are unreliable.

G Relation Between the Compute Scaling Law and the Parameters + Tokens Scaling Law

Here, we provide a concrete analytic relation between Eqn. 6 and Eqn. 3 provided in the main text. For a fixed benchmark $\mathcal{B}$ and number of attempts per problem k, we model

$$-\log\bigl(\text{pass}_{\mathcal{B}}\ @k\bigr)(N, D, k) = \mathcal{E}_0(k) + \frac{N_0(k)}{N^{\beta(k)}} + \frac{D_0(k)}{D^{\gamma(k)}}, \tag{7}$$

and, when training under a compute budget $C = c\,ND$,

$$-\log\bigl(\text{pass}_{\mathcal{B}}\ @k\bigr)(C, k) = E(k) + \frac{C_0(k)}{C^{\alpha(k)}}. \tag{8}$$

Compute-optimal envelope. Assume the excess loss *beyond the irreducible term* is additive and separable in (N, D) as in Eq. 7, with $N_0(k), D_0(k) > 0$ and exponents $\beta(k), \gamma(k) > 0$. For each fixed k, define

$$A = N_0(k), \quad B = D_0(k), \quad \beta = \beta(k), \quad \gamma = \gamma(k), \tag{9}$$

and

$$F(N, D) = \frac{A}{N^{\beta}} + \frac{B}{D^{\gamma}}, \qquad \mathcal{F}(C) = \min_{\substack{N,D>0:\\ cND=C}} F(N, D). \tag{10}$$

Then the compute law Eq. 8 is the compute-optimal envelope of Eq. 7, with

$$E(k) = \mathcal{E}_0(k), \qquad \alpha(k) = \frac{\beta(k)\,\gamma(k)}{\beta(k) + \gamma(k)} = \Bigl(\tfrac{1}{\beta(k)} + \tfrac{1}{\gamma(k)}\Bigr)^{-1}, \tag{11}$$

$$C_0(k) = (\beta + \gamma)\,\beta^{-\frac{\beta}{\beta+\gamma}}\,\gamma^{-\frac{\gamma}{\beta+\gamma}}\,A^{\frac{\gamma}{\beta+\gamma}}\,B^{\frac{\beta}{\beta+\gamma}}\,c^{\alpha(k)}. \tag{12}$$

The compute-optimal allocation along the envelope is

$$N^*(C, k) = \Bigl(\frac{\beta A}{\gamma B}\Bigr)^{\frac{1}{\beta+\gamma}} \Bigl(\frac{C}{c}\Bigr)^{\frac{\gamma}{\beta+\gamma}}, \qquad D^*(C, k) = \Bigl(\frac{\gamma B}{\beta A}\Bigr)^{\frac{1}{\beta+\gamma}} \Bigl(\frac{C}{c}\Bigr)^{\frac{\beta}{\beta+\gamma}}. \tag{13}$$

Derivation Sketch. Form the Lagrangian $\mathcal{L} = AN^{-\beta} + BD^{-\gamma} + \lambda\,(ND - C/c)$. First-order conditions give $\beta AN^{-\beta} = \gamma BD^{-\gamma}$ and $ND = C/c$, from which (A.5) follows. Substituting N^*, D^* into $F(N, D)$ yields $\mathcal{F}(C) = C_0(k)\,C^{-\alpha(k)}$ with $\alpha(k)$ and $C_0(k)$ as in Eqs. 11-12, and $E(k) = \mathcal{E}_0(k)$.

Remarks.

1. The compute exponent is the parallel sum of the (N, D) exponents: $\alpha = (1/\beta + 1/\gamma)^{-1}$.
2. If $\beta = \gamma$, then $\alpha = \beta/2$ and $N^*, D^* \propto C^{1/2}$ with $C_0 = 2\,\beta^{-1}(AB)^{1/2}c^{\beta/2}$.
3. If one scales N and D at a fixed ratio (i.e., without re-optimizing with C), the effective compute exponent degrades to $\alpha_{\text{path}} = \min\{\beta, \gamma\}/2$.
4. All quantities may depend on k; the mapping in Eqs. 11-13 applies point-wise in k. The conversion constant c enters only via C_0, not α.

G.1 Multiplicative deviation from the optimal compute envelope

Let $(N^*(C), D^*(C))$ denote the compute–optimal allocation in Eq. 13 and define the *misallocation ratio*

$$r(C) \coloneqq \frac{N(C)}{N^*(C)} = \frac{D^*(C)}{D(C)}. \tag{14}$$

A direct algebraic manipulation using the first-order optimality condition $\beta A(N^*)^{-\beta} = \gamma B(D^*)^{-\gamma}$ gives the exact factorization

$$\frac{-\log\bigl(\text{pass}_{\mathcal{B}} \text{ @}k\bigr)(C) - \mathcal{E}_0}{-\log\bigl(\text{pass}_{\mathcal{B}} \text{ @}k\bigr)_{\text{opt}}(C) - \mathcal{E}_0} = \Phi(r;\beta,\gamma) \coloneqq \frac{\gamma}{\beta+\gamma}\, r^{-\beta} + \frac{\beta}{\beta+\gamma}\, r^{\gamma} \quad \geq 1, \tag{15}$$

with equality iff $r \equiv 1$. Equivalently, this produces a multiplicative correction to Eq. 8

$$-\log(\text{pass}_{\mathcal{B}} \text{ @}k)(C) = \mathcal{E}_0 + C_0\, C^{-\alpha} \quad \underbrace{\Phi\bigl(\tfrac{N}{N^*};\beta,\gamma\bigr)}_{\text{misallocation factor } \geq 1} \quad . \tag{16}$$

Thus the deviation from equality is *fully parameterized* by r via the dimensionless penalty $\Phi \geq 1$.

Local and asymptotic behavior of the penalty. Let $t = \log r$. A Taylor expansion of equation 15 around $r = 1$ gives

$$\Phi(r;\beta,\gamma) = 1 + \frac{\beta\gamma}{2}\, t^2 + O(t^3), \tag{17}$$

so small misallocations incur only a *quadratic* penalty in log-space. Along a power-law path $N \propto C^p$ (with constants absorbed into r), $t = \bigl(p - \frac{\gamma}{\beta+\gamma}\bigr) \log C + O(1)$; hence if $p \neq \frac{\gamma}{\beta+\gamma}$ the penalty grows polynomially:

$$\Phi(r(C)) \asymp \begin{cases} \frac{\beta}{\beta+\gamma}\, r(C)^{\gamma}, & r(C) \to \infty \quad (\text{over-allocating } N), \\ \frac{\gamma}{\beta+\gamma}\, r(C)^{-\beta}, & r(C) \to 0 \quad (\text{under-allocating } N). \end{cases} \tag{18}$$

This isolates the compute–optimal scaling $C^{-\alpha}$ and expresses all off-ridge effects through a *single*, dimensionless factor depending only on the allocation ratio N/N^* (or equivalently D^*/D).

We conclude that if one matches the compute law without optimizing (N, D), the result is *not* a single power law in general: it is a compute–optimal power multiplied by a misallocation penalty $\Phi(N/N^*) \geq 1$.

H How the Compute Scaling Law Emerges from the Parameters + Tokens Scaling Law

The approximately equivalent performance between the compute scaling law (Eqn. 6) and the parameters + tokens scaling law (Eqn. 3) led us to ask whether a deeper connection might exist. We discovered that *the compute scaling law is the compute-optimal envelope of the parameters+tokens scaling law*. We summarize the insight here and defer a full explanation to Appendix G.

Fix a benchmark $\mathcal{B}$ and attempts per problem k. Consider a compute budget $C \approx c\,ND$. The best achievable error is obtained by minimizing the right-hand side of Eqn. 3 over all (N, D) with $ND = C/c$. Evaluating at the optimizer yields Eqn. 6. The mapping of exponents and constants is:

$$\alpha(k) = \Bigl(\tfrac{1}{\beta(k)} + \tfrac{1}{\gamma(k)}\Bigr)^{-1}, \qquad E_0(k) = \mathcal{E}_0(k),$$

with a closed-form $C_0(k)$.

Compute-Optimal Allocation. Along the envelope (i.e., when Eqn. 6 is tight), the optimal split of compute between parameters and tokens is

$$N^*(C,k) = \Bigl(\tfrac{\beta\, N_0(k)}{\gamma\, D_0(k)}\Bigr)^{\frac{1}{\beta+\gamma}} \Bigl(\tfrac{C}{c}\Bigr)^{\frac{\gamma}{\beta+\gamma}}, \qquad D^*(C,k) = \Bigl(\tfrac{\gamma\, D_0(k)}{\beta\, N_0(k)}\Bigr)^{\frac{1}{\beta+\gamma}} \Bigl(\tfrac{C}{c}\Bigr)^{\frac{\beta}{\beta+\gamma}}.$$

Thus the parameter/token ratio *shift with scale* according to β, γ; when $\beta = \gamma$, both grow like $C^{1/2}$.

Departures from the Envelope. If training does not follow the compute-optimal allocation (N^*, D^*), define the misallocation ratio $r(C) = \frac{N}{N^*(C)} = \frac{D^*(C)}{D}$. The departure from the envelope is *multiplicative* on top of Eqn. 6:

$$-\log\left(\text{pass}_{\mathcal{B}}\ @k\right)(C, k) = E_0(k) + C_0(k) \cdot C^{-\alpha(k)} \cdot \underbrace{\left[\tfrac{\gamma}{\beta+\gamma}\, r^{-\beta} + \tfrac{\beta}{\beta+\gamma}\, r^{\gamma}\right]}_{\Phi(r;\beta,\gamma)\ \geq 1},$$

with equality iff $r \equiv 1$. Small deviations are benign ($\Phi = 1 + \frac{\beta\gamma}{2}(\log r)^2 + O(\log r)^3)$), but persistent off-ridge scaling yields a growing penalty. A useful case: if one scales N and D at a *fixed* ratio (i.e., without re-optimizing with C), the effective compute slope degrades to $\alpha_{\text{path}} = \frac{\min\{\beta,\gamma\}}{2}$.

Takeaways. (i) The compute law is not an independent phenomenology; it is the optimal-allocation shadow of the parameters + tokens law. (ii) Use the allocation above to stay on-envelope when planning larger models. (iii) Implementation constants relating C to ND shift $C_0(k)$, but do not change $\alpha(k)$. All statements hold pointwise in k.

I Methodology (Cont.)

Sampling We used temperature-only sampling (Ackley et al., 1985) at τ=1.0 and did not sweep due to prior work (Schaeffer et al., 2025a).

Language Model Family For our experiments, we used the Pythia family (Biderman et al., 2023), which contains 8 models from 14M to 12B parameters pretrained on 300B tokens from The Pile (Gao et al., 2020). We used 8 checkpoints per parameter size of the non-deduplicated variants, but excluded extremely overtrained checkpoints identified as abnormal by prior work (Godey et al., 2024). Each model checkpoint has a specific number of model parameters (N) and has been trained on a certain number of tokens (D). We use intermediate checkpoints in lieu of fully trained models, and note that our results would likely improve with greater experimental control.

J Discussion

In this work, we investigated pretraining scaling laws for generative evaluations of language models. Our work was motivated most directly by OpenAI et al. (2024)'s success at predicting pass rates at HumanEval (Chen et al., 2021), as well as by Hu et al. (2024).

Key Findings Our work's main contribution is the proposal and rigorous backtesting of three distinct scaling laws for pass-at-k, using (1) pretraining compute, (2) model parameters and pretraining tokens, and (3) gold reference log likelihoods as covariates. We demonstrate that a key hyperparameter in generative evaluation – the number of attempts per problem k – offers a novel lever for controlling the scaling law parameters and predictability. Empirically, we find that while all three laws have comparable predictive accuracy, the parameters of the gold reference likelihood law are uniquely stable, converging across nearly five orders of magnitude of compute. Finally, we establish a theoretical connection between the compute law and the parameters + tokens law, proving that the compute law emerges as the compute-optimal envelope of the parameters-and-tokens law.

Limitations The main limitation of this work is that the empirical results are based on a single benchmark (GSM8K) and a single model family (Pythia). Additionally, we consider only pass-at-k generative evaluations, whereas other are of course worth considering as well.

Related Work Due to space constraints, we defer related work to Appendix B.

Future Directions Multiple next steps are clear. First, subsequent work should determine whether predictions can be made more accurately using either additional sampling or controlled pretrained models. Second, OpenAI et al. (2024) claimed subsetting and/or stratifying problems enabled better predictions; what exactly did this approach entail, and how much did it help? Third, the utility of the gold reference log likelihoods was perhaps surprising because if we think of next-token sampling as a branching process, there are likely exponentially many paths from the problem to the correct answer,

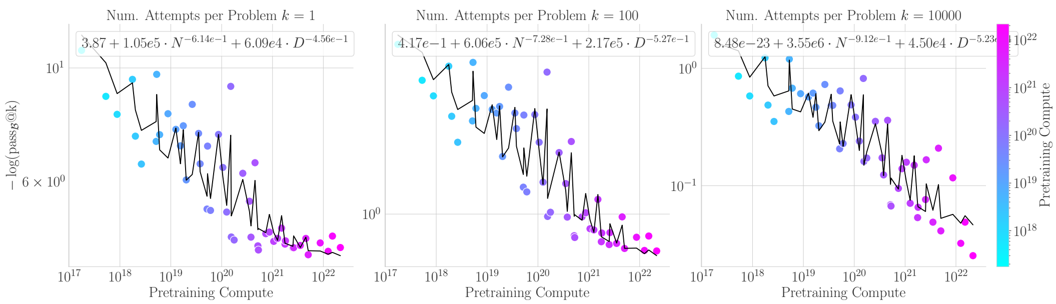

Figure 9: **Scaling of Pass Rates with Model Parameters and Pretraining Data (Full Fit).** Each panel fits Eqn. 3, $-\log\big(\text{pass}_{\mathcal{B}}@k\big)(N, D, k) = \mathcal{E}_0(k) + N_0(k)\, N^{-\beta(k)} + D_0(k)\, D^{-\gamma(k)}$. Relative to the compute scaling law, using parameters N and tokens D instead yields tighter in-range fits for all k. Consistent with Fig. 1, the irreducible error decreases sharply with k (e.g., $\widehat{\mathcal{E}}_0:\ 3.87 \to 0$ by $k \approx 10^2$), after which variation is dominated by the (N, D) terms. Despite the better global fit, the largest-compute model checkpoint in each panel exhibits comparatively large relative error.

and it is unclear that the benchmark-provided gold references have any meaningful relationship with likely samples from the model; this raises at least two future questions: (i) how general is this relationship across benchmarks and models, and (ii) to what extent can this indicator be trusted for predicting the performance of more expensive models, especially under optimization pressure?

K Predictive Performance (Backtesting Results)

To evaluate the predictive power of each scaling law, we performed backtesting: fitting the law on cheaper checkpoints and extrapolating to the 12B-parameter model trained on 300B tokens. For the compute scaling law, predictions become reliable once models within roughly two orders of magnitude of the target are included, after which errors plateau. The parameters+tokens scaling law provides tighter in-range fits and slightly lower relative error for small k, but still requires models within $\sim$2 orders of magnitude of the target for stable predictions. The gold reference likelihood law yields comparable predictive accuracy overall but exhibits exceptional stability: its parameters converge even when fit using models up to five orders of magnitude cheaper. Together, these results show that all three laws can forecast performance, with the gold reference law offering the most robust long-range signal. The following figures illustrate these findings in detail.

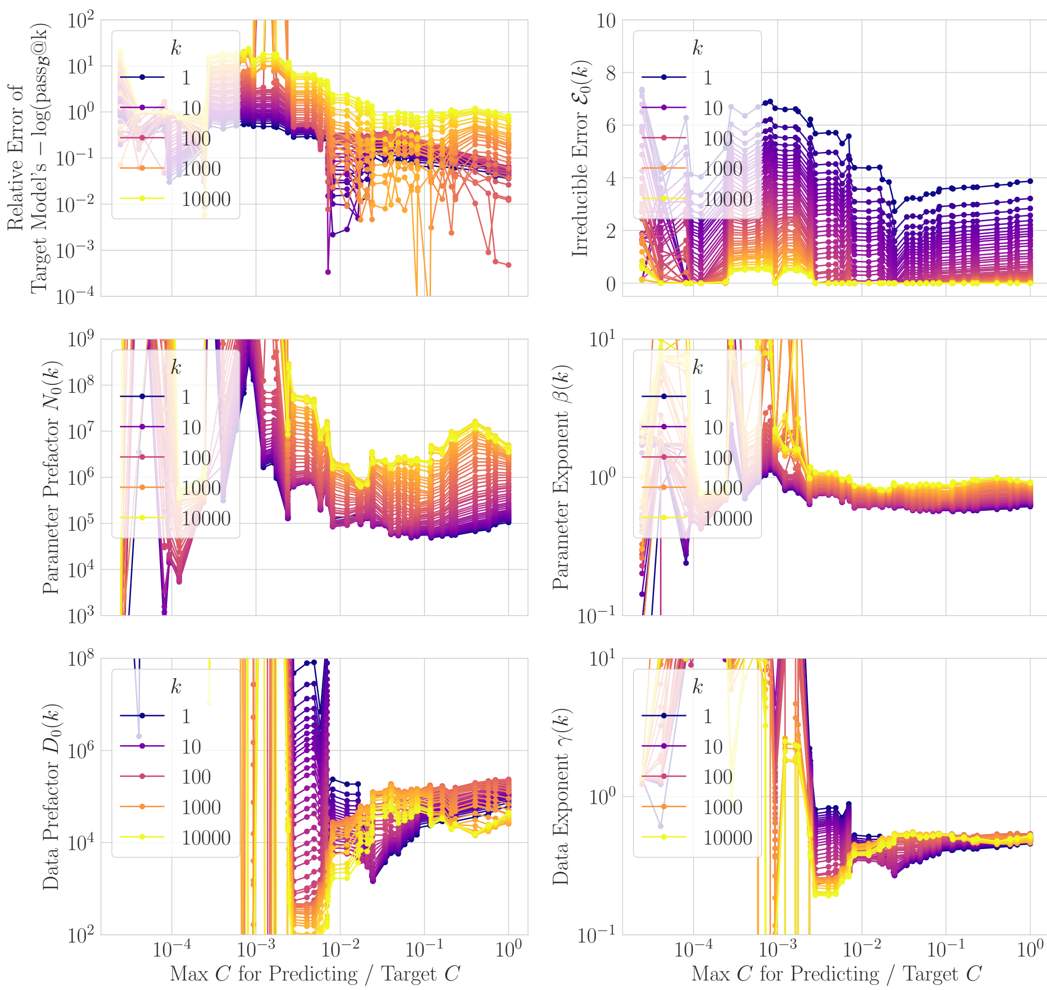

Figure 10: **Predicting Pass Rates from Scaling Parameters and Tokens (Backtesting).** We evaluated how accurately the parameters + tokens scaling law (Eqn. 3) predicts the most expensive model's $-\log(\text{pass}_{\mathcal{B}}@k)$. **Top Left:** The relative error decreases as higher-compute checkpoints are included, plateauing once the fit includes models within $\sim$2 orders of magnitude of the target. **Other Five Panels:** The estimates of the five scaling law parameters are initially unstable but converge to their full-fit values once included models are within $\sim$2 orders of magnitude of the target.

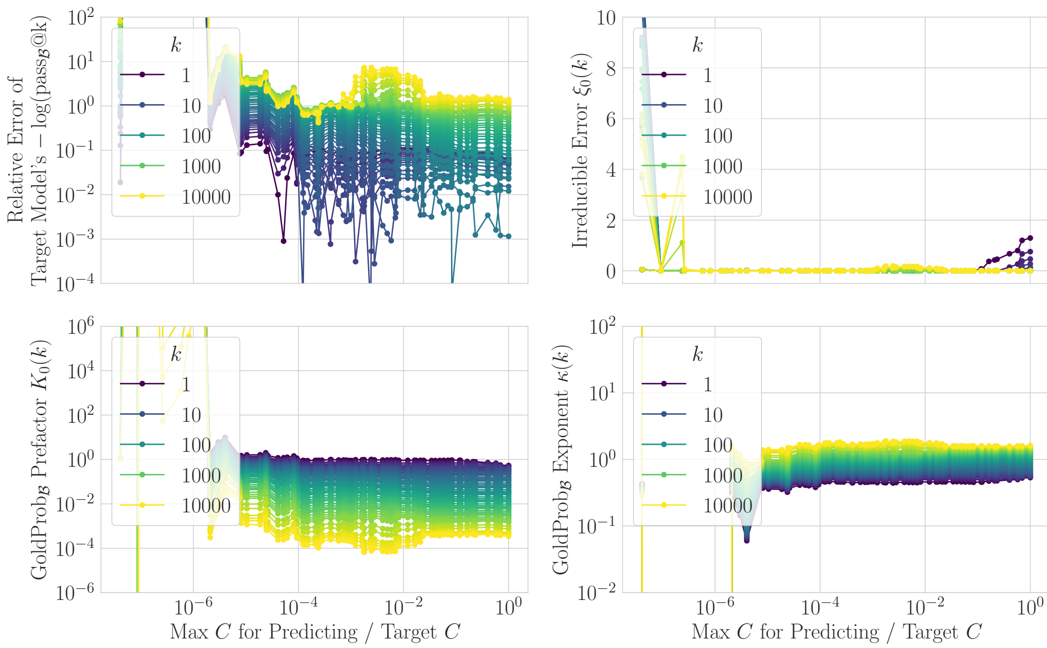

Figure 11: **Predicting Pass Rates from Gold Reference Likelihoods (Backtesting).** While the overall predictive error is comparable to the other laws, it is slightly lower for small k and slightly higher for large k. **Top Left:** The most striking result is the exceptional stability of the scaling law's parameters. In stark contrast to the compute and parameters+tokens laws, whose parameters stabilized only when fit on models within $\sim$2 orders of magnitude of the target, the backtested estimates for $\widehat{\xi}_0(k)$, $\widehat{K}_0(k)$, and $\widehat{\kappa}(k)$ converge to their final values using models up to $\sim$5 orders of magnitude cheaper than the target. **Other Three Panels:** This remarkable stability suggests that the relationship between gold reference likelihoods and pass rates provides a robust signal, enabling reliable, long-range forecasts from much cheaper models.