Scaling Laws Are Unreliable for Downstream Tasks: A Reality Check

Nicholas Lourie[1*] **Michael Y. Hu**[1*] **Kyunghyun Cho**[1,2,3]
[1]New York University [2]Prescient Design [3] CIFAR LMB
{nick.lourie, michael.hu, kyunghyun.cho}@nyu.edu

Abstract

Downstream scaling laws aim to predict task performance at larger scales from the model's performance at smaller scales. Whether such prediction should be possible is unclear: some works discover clear linear scaling trends after simple transformations of the performance metric, whereas others point out fundamental challenges to downstream scaling laws, such as emergence and inverse scaling. In this work, we conduct a meta-analysis of existing data on downstream scaling laws, and we find that predictable scaling only occurs in a minority of cases: 39% of the time. Moreover, seemingly benign changes to the experimental setting can completely change the scaling behavior. Our analysis underscores the need to understand the conditions under which scaling laws succeed. To accurately model the relationship between pretraining loss and task performance, we must embrace the cases in which scaling behavior deviates from linear trends.

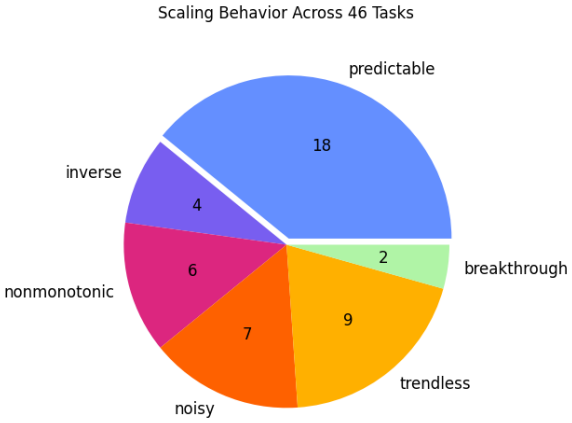

Figure 1: Revisiting the 46 tasks studied in Gadre et al. (2025), we find that only 18 tasks—or 39%—demonstrate smooth, predictable improvement (Figure 5). The other 28 tasks are shown in Figures 6 through 10, where we group them into different degenerate scaling behaviors: inverse, nonmonotonic, noisy, trendless, and breakthrough scaling. See Figure 2 for examples.

1 Introduction

Scaling laws for pretraining establish that the loss improves reliably when increasing the size of the model, training data, or compute (Kaplan et al., 2020; Hoffmann et al., 2022; Pearce and Song, 2024). However, better pretraining loss does not always translate to better downstream performance (Magnusson et al., 2024). This gap can be caused by a variety of issues; among the best known is *emergence*, or the fact that on some tasks models below a certain scale show no trend or near-chance performance (Wei et al., 2022). In addition, model performance can increase then decrease, in what is called *inverse scaling* (McKenzie et al., 2023; Wilcox et al., 2024). Despite these challenges, several works also find downstream performance is roughly linear in the pretraining loss, possibly after a transformation (Huang et al., 2024; Gadre et al., 2025; Chen et al., 2025).

Thus, the extent to which downstream scaling laws work is unclear. How can they follow linear forms when we know of many tasks that exhibit emergence or inverse scaling? Are these just edge cases, or are they common, with no established explanation? Here, we aim to clarify this confusion. We explore and consider three core factors affecting downstream scaling laws: 1) the data used for pretraining and validation, 2) the downstream task, and 3) the experimental setup. Realistic changes to each of these factors can change the relationship to downstream performance, to the point that the scaling law's functional form might no longer hold. The behavior can even change qualitatively—with a decrease in trend under one set of conditions flipping to an increase under another. Such qualitative changes can not be removed by a new functional form. They pose fundamental obstacles to studying large scale models with small scale proxies.

[*]Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 16167–16180
November 4-9, 2025

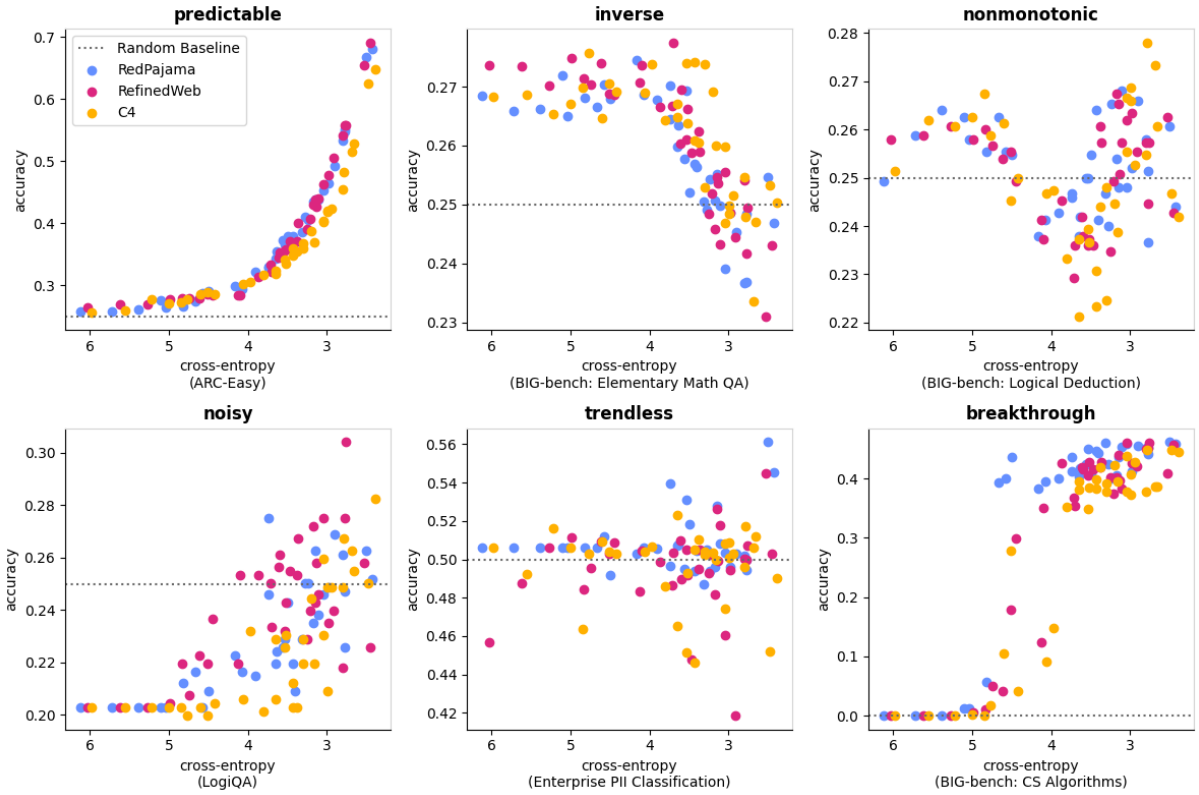

Figure 2: A taxonomy of different scaling behaviors. Predictable scaling fits closely to a linear functional form after, for example, exponentiating the cross-entropy loss. However, depending on the downstream task, models do not always improve with scale (inverse, nonmonotonic, and trendless), or the improvement might be highly noisy. The improvement might also follow a functional form that is difficult to extrapolate like a sigmoid (breakthrough).

In particular, we find that:

1. Choosing a different dataset for the validation perplexity can flip scaling trends. For instance, models pretrained on one corpus might appear to improve faster than another, but this trend can reverse with different validation data (§3).

2. Revisiting a prior study (Gadre et al., 2025), we only find predictable scaling in a minority of cases: 39%. Phenomena like emergence or inverse scaling can actually be quite common, occurring for many downstream tasks (§4).

3. Scaling behavior from one experimental setup can *qualitatively* change under another; a task with predictable scaling could become non-monotonic or even show no trend at all (§5).

Our analysis suggests that scaling laws are *context-specific*. As such, we cannot assume downstream scaling will always be strictly linear. Rather, we need to better understand the failure modes of existing scaling laws and develop a holistic (and perhaps more complex) model of how foundation models improve on downstream tasks.

2 Background

Downstream scaling laws try to extrapolate the performance of large-scale models from small-scale proxies. When successful, these scaling laws enable cost-effective experiments at the small scale that transfer to the large. Scaling laws for pretraining are well established (Rosenfeld et al., 2020; Kaplan et al., 2020; Hoffmann et al., 2022); however, the ultimate goal of language models is to perform well on downstream tasks. As such, downstream scaling laws are also of great interest.

Unlike pretraining, there is little consensus on how to approach downstream scaling laws. Early efforts tried to predict downstream performance directly from parameters, data, or compute (Ivgi et al., 2022; Mahmood et al., 2022; OpenAI, 2023), but downstream performance often showed a noisy relationship to these quantities (Tay et al., 2022). Other efforts sought surrogates for scale, like latent capabilities (Ruan et al., 2024) or task-specific losses (Grattafiori et al., 2024; Bhagia et al., 2024), in the hope of mapping from compute to surrogate, and from surrogate to downstream performance.

Many works have found downstream scaling laws are more stable when stated in terms of pretraining loss, the most widely used surrogate for scale (Xia et al., 2023; Huang et al., 2024; Du et al., 2024; Gadre et al., 2025; Chen et al., 2025). Specifically, if two models differ in their number of parameters or pretraining tokens but still attain the same pretraining loss, then they tend to achieve the same downstream performance (Xia et al., 2023; Du et al., 2024; Gadre et al., 2025). Some authors generalize this principle, viewing pretraining loss as a general way to compare models—not just with different scales but different training recipes (Huang et al., 2024; Chen et al., 2025). However, while pretraining loss correlates with downstream performance to a surprising degree, this relationship is far from absolute (Tay et al., 2022).

In the best case, downstream task performance is roughly linear in some monotonic transformation of validation loss (x) and downstream metric (y):

$$f(y) = a\, g(x) + b$$

For instance, Gadre et al. (2025) relate the error rate to the validation loss via: $y = \epsilon - k\, \exp\{-\gamma\, x\}$. The functions, f and g, must depend on the metric; in some cases, a linear relationship alone could be enough (Huang et al., 2024; Chen et al., 2025).

Such global structure enables extrapolation—the ultimate goal of downstream scaling laws; however, predictable structure might not exist. Phenomena such as emergence, inverse, and U-shaped scaling preclude this kind of global structure (Wei et al., 2022; McKenzie et al., 2023; Wei et al., 2023). They destroy it by creating *structural breaks*, or points where the function describing one part of the scaling curve does not describe another (Andrews, 1993). Without global structure, it is impossible to extrapolate from small to large. Certain tasks, like multiple choice questions, are more susceptible to emergence (Schaeffer et al., 2025). Sometimes emergence can be mitigated by choosing a different downstream metric (Schaeffer et al., 2023), but in other cases breakthrough improvements remain "stubbornly emergent" (Zhao et al., 2025; Du et al., 2024). Li et al. (2025) found scaling laws generally difficult to reproduce and sensitive to the functional form, training setup, data collection, and fitting algorithm, while Magnusson et al. (2025) found that comparing models at a smaller scale—without extrapolation—performed as well or better than scaling laws in determining the best data mix.

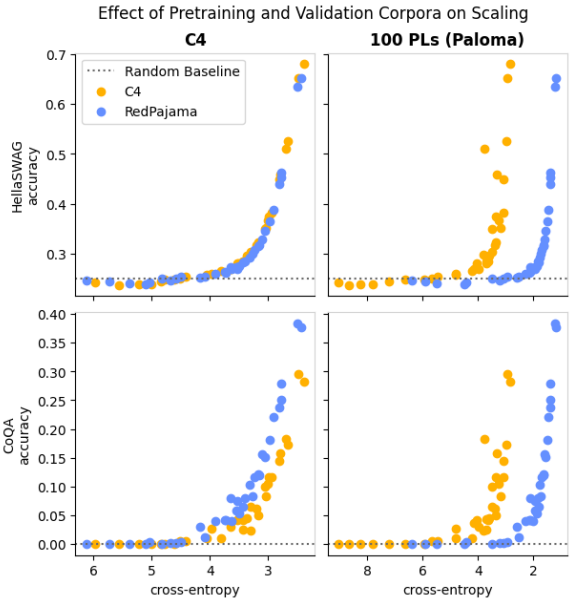

Figure 3: Choosing a different validation corpus can exaggerate or even reverse which pretraining corpus appears superior. On HellaSwag, the C4 corpus seems better than RedPajama when using 100 PLs as the validation set. Conversely, the scaling trends on CoQA for C4 and RedPajama flip when computing validation perplexity on C4 versus 100 PLs.

3 Scaling Laws Are Specific to the Data

Downstream scaling laws depend on several factors, including the pretraining data, validation data, and downstream task. If you vary the pretraining data, then you must fix a validation corpus to compare the loss across models. Once it is fixed, you might hope to focus on the validation loss alone and simplify your research. It is unclear how to choose the correct validation data but, what is more, its choice can entirely reverse which pretraining setup appears superior for a downstream task if we do not consider the full context. The pretraining data, validation data, and downstream task all interact in forming the scaling law—you can not consider one without the other two. In particular, you can not compare data mixes by their validation losses or downstream scaling laws based upon them.

To evince these claims, we reexamine the results from Gadre et al. (2025), who pretrained models over different corpora. We observe how pretraining data, validation data, and task all interact in forming the scaling law (Figure 3). The colors correspond to pretraining corpora: C4 (Raffel et al., 2020) and RedPajama (Weber et al., 2024); the columns correspond to validation data: C4 (Raffel et al., 2020) and Paloma's 100 Program-

ming Languages (Magnusson et al., 2024); and the rows correspond to tasks: CoQA (Reddy et al., 2019) and HellaSwag (Zellers et al., 2019). If you ignore the other factors, then choice of validation data can completely change the scaling trend:

Exaggerating differences. For HellaSwag with C4 as the validation corpus (top left), pretraining on either C4 or RedPajama produces the same scaling law. However, when using 100 Programming Languages (100 PLs, top right), the scaling laws for C4 and RedPajama no longer superimpose—pretraining on C4 appears to achieve much better performance even for a worse validation loss.

Flipping scaling trends. Changing the task to CoQA (bottom left) also changes the scaling laws, with RedPajama now achieving better performance sooner. Even worse, changing the validation corpus from C4 to 100 PLs reverses this relationship *again* (bottom right).

Thus, whether or not better perplexity translates to better downstream performance depends on the task, the pretraining corpus, and the validation loss. Changes to any one of these three factors can reverse which pretraining setup appears superior.

4 Irregular Scaling Is Common

Phenomena like emergence and inverse scaling suggest that linear scaling laws do not capture all task scaling behavior. However, it is also unclear how prevalent these phenomena are in practice. Thus, we re-examine scaling behavior on the 46 tasks tested by Gadre et al. (2025), classifying them into six categories visualized in Figure 2. We find that linear scaling actually occurs in a *minority* of cases in their setting: 39% of the time (Figure 1). For some experimental setups, non-linear scaling is actually the norm. Are the experimental choices of Gadre et al. (2025) abnormal? On the contrary, all three of pretraining corpora, validation datasets, and downstream tasks are from well-known sources in the literature. Downstream tasks are comprised of established evaluations like BoolQ, HellaSwag, and BIG-Bench (Clark et al., 2019; Zellers et al., 2019; Srivastava et al., 2023). Irregular scaling occurs within popular tasks and is easy to find.

5 Scaling Behavior Is Not Always Robust

Finally, we show that conclusions about scaling laws may not generalize across settings: setups with the same validation data and downstream tasks may observe entirely different scaling trends. To show this, we take the 10 overlapping tasks between Gadre et al. (2025) and Magnusson et al. (2025). These authors consider several of the same pretraining corpora and downstream tasks; however, their implementation details differ (see Appendix A). For example, Gadre et al. (2025) use fewer answer choices for Commonsense QA. Thus, we might expect to see quantitative differences in some of their scaling laws, but what is more surprising is that we also see *qualitative* changes in their scaling behavior.

To enable comparison, we evaluate all models from Gadre et al. (2025) and Magnusson et al. (2025) on the same validation corpus (C4). For Magnusson et al. (2025), which does not use C4, we evaluate 200 released checkpoints, which vary in their pretraining setup:[1]

1. Parameters: {20M, 60M, 150M, 300M, 1B}.
2. Pretraining corpus: {Dolma, C4, DCLM-Baseline, RefinedWeb (Falcon), FineWeb-Pro} (Soldaini et al., 2024; Raffel et al., 2020; Li et al., 2024; Penedo et al., 2023b; Zhou et al., 2025).
3. Training steps: from 5,000 to 40,000, in intervals of 5,000.

Out of the 10 overlapping downstream tasks, Figure 4 shows three that produce different scaling trends between setups. In both setups, MMLU shows a positive trend; however, the noisiness and shape of that trend differs greatly. At an even greater extreme, CommonsenseQA's scaling behavior qualitatively changes: while CommonsenseQA shows nonmonotonic scaling in Gadre et al.'s (2025) results, it exhibits a clean scaling law with Magnusson et al.'s (2025). On the other hand, in Magnusson et al.'s (2025) results, BoolQ's scaling law appears trendless; however, this lack of trend comes from the models spanning too small a validation loss range. Since Gadre et al.'s (2025) models cover a wider range of validation losses, the trend more clearly emerges. Thus, scaling behavior can fluctuate, even between controlled studies for the same task with the same pretraining corpora.

6 Discussion

Given the cost of training modern foundation models, scaling laws have become an invaluable tool for making informed modeling decisions. Scal-

[1]The 20M models are the smallest models in Magnusson et al. (2025) and have the highest cross-entropy loss.

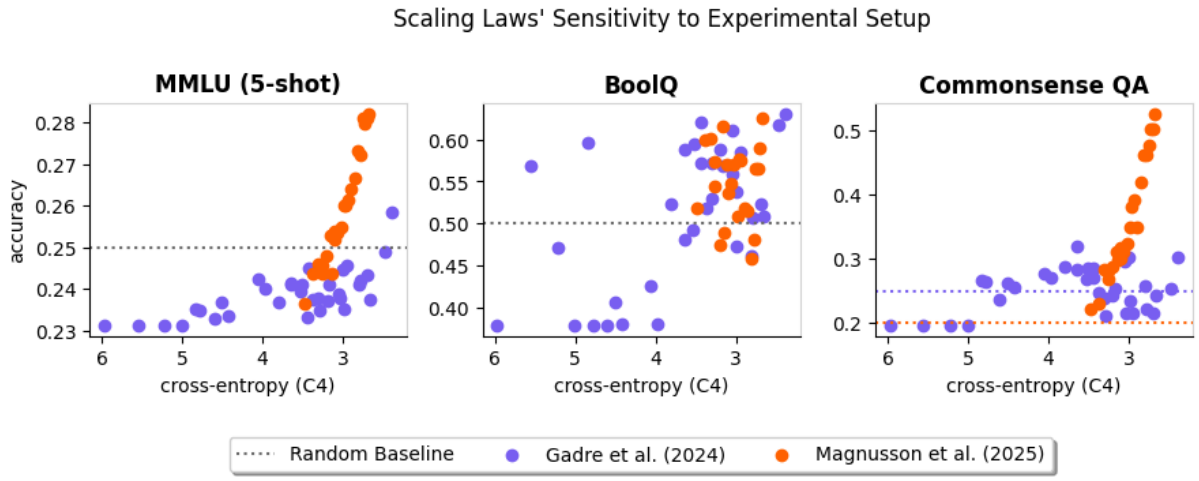

Figure 4: Scaling behavior changes depending on the experimental setting. Gadre et al. (2025) and Magnusson et al. (2025) both train language models on C4 and evaluate on MMLU, BoolQ, and Commonsense QA. Still, they differ in their details, such as model architecture, task formatting, or the number of answer choices (in the case of Commonsense QA). *Even with the same corpora and downstream task, scaling trends can be dramatically different.*

ing laws enable us to extrapolate results where compute costs would otherwise make extensive experimentation infeasible. However, extrapolations are only worthwhile when their assumptions are faithful to the data.

As practitioners of scaling laws, we must realize that predictable scaling laws often exist, but one cannot assume that they hold for all contexts. Even if scaling is stable on the same task for the same validation data, other aspects of the experimental setup might change the scaling behavior (§5). To some extent, scaling laws are *investigator-specific* (Li et al., 2025), and so each investigator must verify the scaling law's presence with visualizations and regression diagnostics (Shalizi, 2015).

For researchers, downstream scaling laws offer many fascinating new directions. Empirically, we need better ways to stabilize scaling laws and detect when irregular scaling might occur. We must understand the factors affecting scaling laws, and what parts of the experimental setup must remain constant for linear scaling laws to hold. Theoretically, we need a model for why predictable scaling occurs (Hutter, 2021), and a core goal of such a theory should be explaining exactly the cases in which it does not.

7 Conclusion

In this work, we surveyed where downstream scaling laws break down. Depending on the pretraining corpus, validation corpus, or downstream task, scaling laws can change. Better perplexity does not always translate to better downstream performance (§3); *perplexity is not all you need*. Even when holding pretraining and validation data the same, more often than not a predictable scaling law does not exist at all (§4). Irregular behaviors like nonmonotonic, trendless, or breakthrough scaling are all common; one must establish predictable scaling for the given task before relying on it. Finally, seeing predictable scaling in one experimental setup does not guarantee it for another (§5). Until we better understand why predictable scaling arises and its sufficient conditions, investigators must verify scaling laws in their own settings.

Limitations

Our work uses data and model checkpoints from existing studies (Gadre et al., 2025; Magnusson et al., 2025). While this is sufficient for the counterexamples featured in this work, there may be unknown biases shared between these two projects that we have missed.

In addition, our study establishes that downstream scaling laws are unreliable *under current practice*. Neural networks used to be notoriously difficult to train; however, as understanding developed around how to architect, initialize, and optimize them, training neural networks became routine and reliable. In a similar way, it is possible that the difficulties discussed here may be overcome by better techniques for measuring and estimating scaling laws. We hope the challenges we identify here inspire researchers to search for them.

Acknowledgments

MYH is supported by the NSF Graduate Research Fellowship. This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) with a grant funded by the Ministry of Science and ICT (MSIT) of the Republic of Korea in connection with the Global AI Frontier Lab International Collaborative Research. This work was also supported by the Samsung Advanced Institute of Technology (under the project Next Generation Deep Learning: From Pattern Recognition to AI) and the National Science Foundation (under NSF Award 1922658). This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

Donald W. K. Andrews. 1993. Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856.

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. 2024. Establishing task scaling laws via compute-efficient model ladders. *Preprint*, arXiv:2412.04403.

Yangyi Chen, Binxuan Huang, Yifan Gao, Zhengyang Wang, Jingfeng Yang, and Heng Ji. 2025. Scaling laws for predicting downstream performance in LLMs. *Transactions on Machine Learning Research*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. 2024. Understanding emergent abilities of language models from the loss perspective. In *Advances in Neural Information Processing Systems*, volume 37, pages 53138–53167. Curran Associates, Inc.

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, Luca Soldaini, Jenia Jitsev, Alex Dimakis, Gabriel Ilharco, Pang Wei Koh, Shuran Song, and 6 others. 2025. Language models scale reliably with over-training and on downstream tasks. In *The Thirteenth International Conference on Learning Representations*.

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, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. 2025. OLMES: A standard for language model evaluations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5005–5033, Albuquerque, New Mexico. Association for Computational Linguistics.

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, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*.

Yuzhen Huang, Jinghan Zhang, Zifei Shan, and Junxian He. 2024. Compression represents intelligence linearly. In *First Conference on Language Modeling*.

Marcus Hutter. 2021. Learning curve theory. *arXiv preprint arXiv:2102.04074*.

Maor Ivgi, Yair Carmon, and Jonathan Berant. 2022. Scaling laws under the microscope: Predicting transformer performance from small scale experiments. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7354–7371, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee F Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2024. Datacomp-LM: In search of the next generation of training sets for language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. 2025. (mis)fitting scaling laws: A survey of scaling

law fitting techniques in deep learning. In *The Thirteenth International Conference on Learning Representations*.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Pete Walsh, Yanai Elazar, Kyle Lo, Dirk Groeneveld, Iz Beltagy, Hannaneh Hajishirzi, Noah A. Smith, Kyle Richardson, and Jesse Dodge. 2024. Paloma: A benchmark for evaluating language model fit. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. 2025. Datadecide: How to predict best pretraining data with small experiments. *Preprint*, arXiv:2504.11393.

Rafid Mahmood, James Lucas, David Acuna, Daiqing Li, Jonah Philion, Jose M. Alvarez, Zhiding Yu, Sanja Fidler, and Marc T. Law. 2022. How much more data do i need? estimating requirements for downstream tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 275–284.

Ian R. McKenzie, Alexander Lyzhov, Michael Martin Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Xudong Shen, Joe Cavanagh, Andrew George Gritsevskiy, Derik Kauffman, Aaron T. Kirtland, Zhengping Zhou, Yuhui Zhang, Sicong Huang, Daniel Wurgaft, Max Weiss, Alexis Ross, Gabriel Recchia, and 7 others. 2023. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research*. Featured Certification.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023a. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. In *Advances in Neural Information Processing Systems*, volume 36, pages 79155–79172. Curran Associates, Inc.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023b. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

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

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. 2024. Observational scaling laws and the predictability of langauge model performance. In *Advances in Neural Information Processing Systems*, volume 37, pages 15841–15892. Curran Associates, Inc.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. 2025. Why has predicting downstream capabilities of frontier AI models with scale remained elusive? In *Forty-second International Conference on Machine Learning*.

Cosma Rohilla Shalizi. 2015. The truth about linear regression. *Online Manuscript. http:///www. stat. cmu. edu/˜ cshalizi/TALR*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2022. Scale efficiently: Insights from pretraining and finetuning transformers. In *International Conference on Learning Representations*.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: an open dataset for training large language models. *NeurIPS Datasets and Benchmarks Track*.

Jason Wei, Najoung Kim, Yi Tay, and Quoc Le. 2023. Inverse scaling can become U-shaped. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15580–15591, Singapore. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Ethan G Wilcox, Michael Hu, Aaron Mueller, Tal Linzen, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Ryan Cotterell, and Adina Williams. 2024. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. 2023. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13711–13738, Toronto, Canada. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Rosie Zhao, Tian Qin, David Alvarez-Melis, Sham Kakade, and Naomi Saphra. 2025. Distributional scaling laws for emergent capabilities. *Preprint*, arXiv:2502.17356.

Fan Zhou, Zengzhi Wang, Qian Liu, Junlong Li, and Pengfei Liu. 2025. Programming every example: Lifting pre-training data quality like experts at scale.

A Data Sources

The first set of results comes from Gadre et al. (2025) and is available[2] under the MIT License. Gadre et al. (2025) pretrain transformer language models across different scales on several different corpora (e.g., C4 (Raffel et al., 2020), RedPajama (Weber et al., 2024), and RefinedWeb (Penedo et al., 2023a)). For each model, they compute its validation loss on these and other corpora, and evaluate the model with few-shot prompting across 46 tasks using LLM Foundry.[3]

The second set of results comes from Magnusson et al. (2025) and is available[4] under the ODC-By License. Magnusson et al. (2025) also pretrain transformer language models but they choose different architectural details than Gadre et al. (2025) and use the C4 and RefinedWeb versions from Dolma 1.7 (Soldaini et al., 2024). They pretrain models across these and other corpora and evaluate them with few-shot prompting on 10 different tasks via OLMES (Gu et al., 2025).[5]

The evaluation harnesses, LLM Foundry and OLMES, have some important differences. LLM Foundry ships with its own versions of tasks' datasets. For some tasks (e.g., Commonsense QA and SIQA), it changes the number of answer choices, thus the random baseline can change between the different versions and we have indicated this in the appropriate figures. LLM Foundry also varies the number of shots depending on the task, whereas OLMES uses 5 curated examples for each one. Other differences include the task formulation (whether to use multiple-choice or cloze format) and implementation details such as the prompts.

Finally, LLM Foundry's version of SIQA had an error where the gold labels were incorrect.[6] This issue was fixed in LLM Foundry v0.5.0;[7] however, Gadre et al. (2025) used v0.4.0 for at least some of their experiments.[8] As a result, we do not examine SIQA in our analyses, although Figure 11 in Appendix D includes SIQA for completeness.

[2] https://github.com/mlfoundations/scaling

[3] https://github.com/mosaicml/llm-foundry

[4] https://huggingface.co/datasets/allenai/DataDecide-eval-results

[5] https://github.com/allenai/olmes

[6] https://github.com/mosaicml/llm-foundry/pull/774.

[7] https://github.com/mosaicml/llm-foundry/releases/tag/v0.5.0

[8] https://wandb.ai/samir/dcnlp/runs/rezso5ec/files/requirements.txt

B Reproducibility

We augment the results from Magnusson et al. (2025) by evaluating their models on C4's validation data. This additional information enables us to compare scaling laws from Magnusson et al. (2025) and Gadre et al. (2025).

To compute the validation loss, we ran the DataDecide (Magnusson et al., 2025) models from HuggingFace using `ai2-olmo` (`https://github.com/allenai/OLMo`), computing the perplexity on C4's validation split using a batch size of 64. For inference, we used a combination of A100 and H100 GPUs with 32 GB of CPU RAM. Running inferences over the C4 validation set took approximately 10 minutes.

C Scaling Behaviors

Our analysis centers on *qualitative* scaling behavior in order to avoid having its conclusions depend on a particular form of the scaling law. Without formal criteria, researchers might disagree on particular examples, but then researchers might also disagree on formal criteria. Thus we define the scaling behaviors informally, as follows:

predictable Performance increases, without too much variation around the trend.

inverse Performance decreases as loss improves.

nonmonotonic Performance switches between increasing and decreasing.

noisy Performance increases but varies greatly around the trend.

trendless Performance is flat or there is too much noise to discern a trend.

breakthrough Performance starts flat, abruptly increases, and then plateaus.

Figures 5 through 10 present the tasks studied in Gadre et al. (2025) sorted by scaling behavior. The figures include scaling laws fitted using the functional form from Gadre et al. (2025) that relates the error rate to the validation loss: $y = \epsilon - k\ \exp\{-\gamma\, x\}$. We apply this functional form to the accuracy using the fact that accuracy is one minus the error rate. The scaling laws were fitted by minimizing mean squared error with `differential_evoluation` from SciPy (Virtanen et al., 2020). For the optimization, we used bounds of 0 to 1 for ϵ, 0 to 20 for $\ln(k)$ (we searched k on a log scale to widen the range), and 0 to 20 for γ. In addition, we used 60 for the `popsize` multiplier. Figures 5 through 10 also include the R^2 of each

scaling law's fit as a measure of how closely the scaling behavior adheres to the functional form.

D Sensitivity to Experimental Setups

Figure 11 shows how the scaling laws change between the experimental setups of Gadre et al. (2025) and Magnusson et al. (2025) across 10 tasks.

E The Effect of the Pretraining and Validation Corpora

Figure 12 compares how the task, pretraining, and validation corpora affect the scaling curve. For ease of visualization, we only show the effect on tasks with the cleanest, most predictable scaling laws.

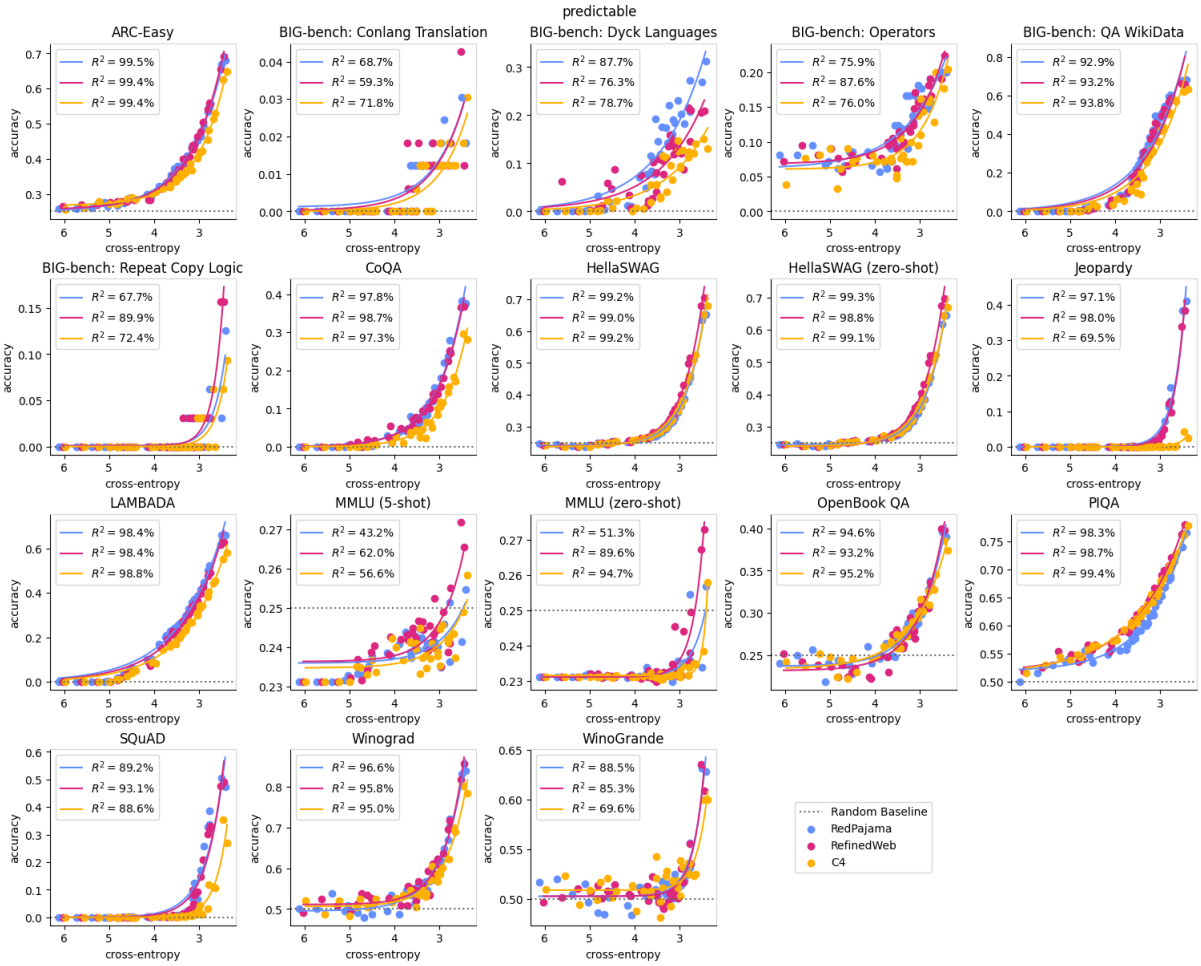

Figure 5: The 18 tasks in Gadre et al. (2025) with scaling behavior well-described by a linear scaling law after transforming the cross-entropy loss.

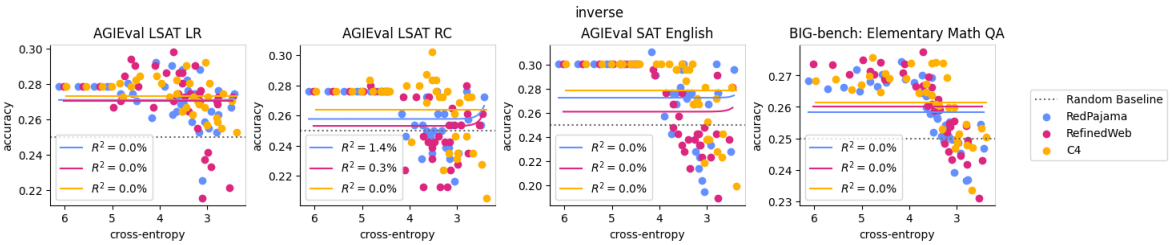

Figure 6: Tasks in Gadre et al. (2025) with inverse scaling.

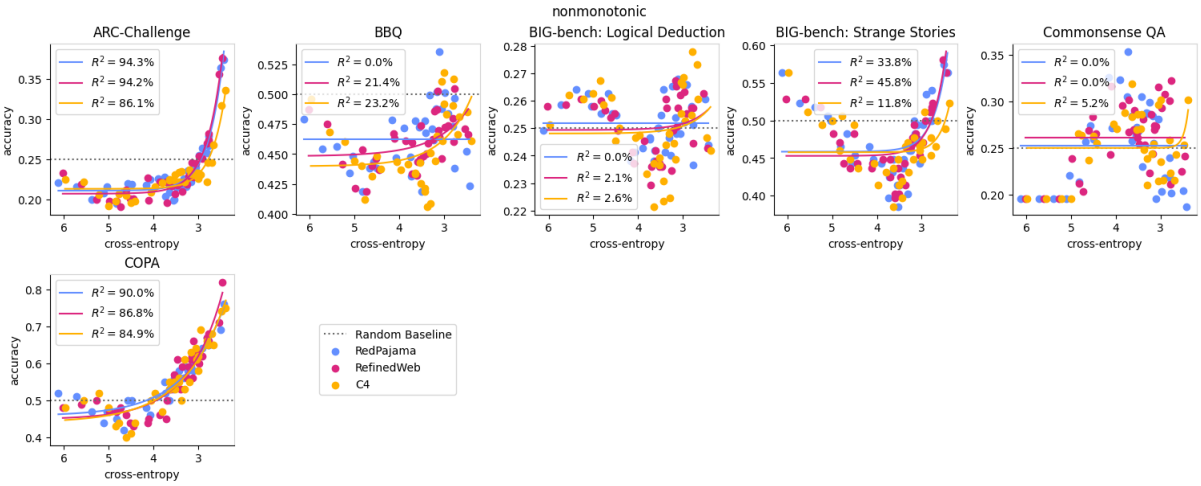

Figure 7: Tasks in Gadre et al. (2025) with nonmonotonic scaling.

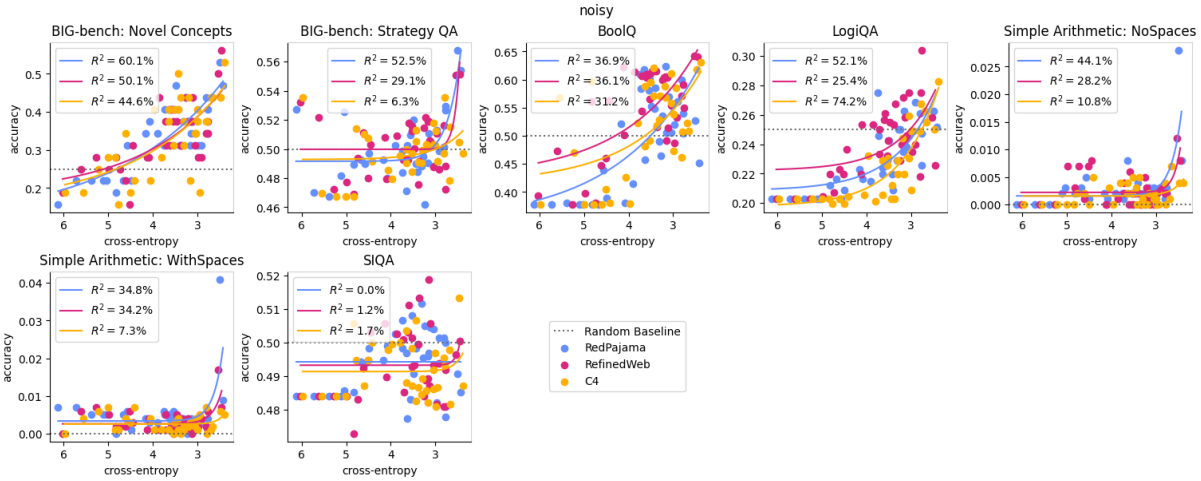

Figure 8: Tasks in Gadre et al. (2025) with noisy scaling.

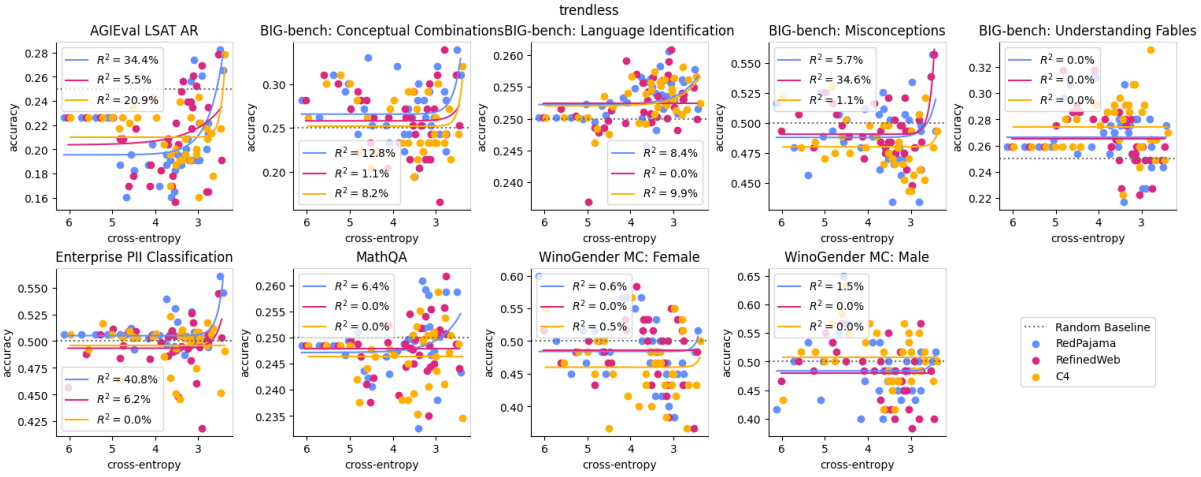

Figure 9: Tasks in Gadre et al. (2025) with no clear scaling trend.

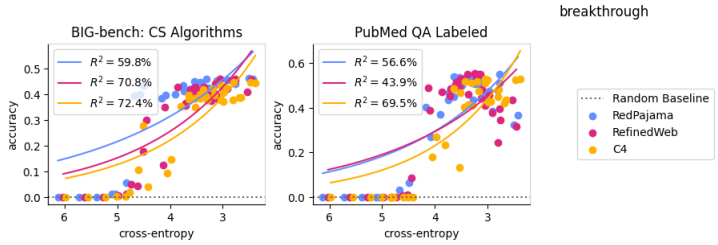

Figure 10: Tasks in Gadre et al. (2025) demonstrating breakthrough scaling.

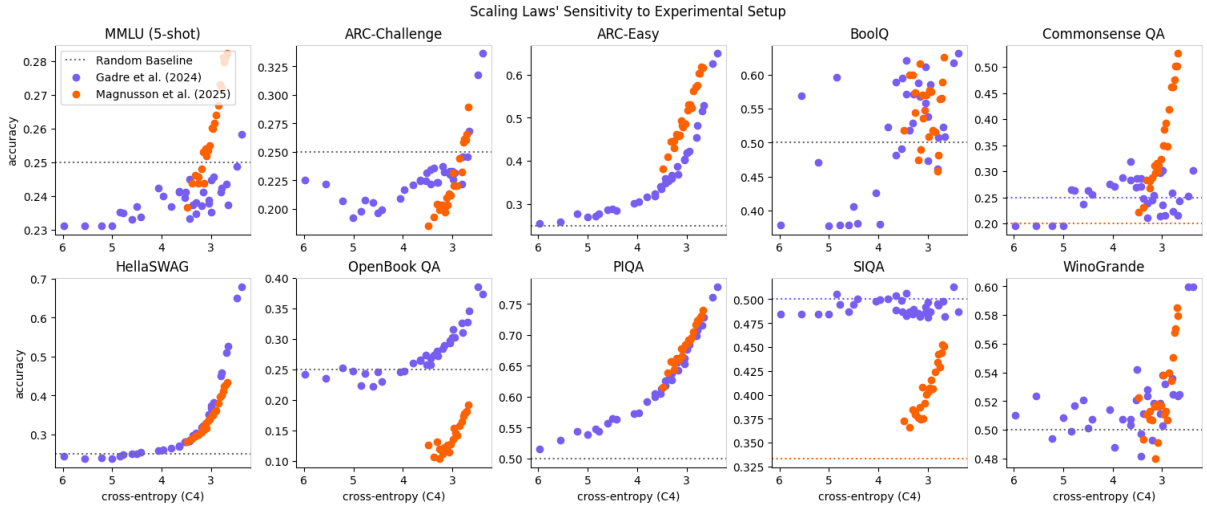

Figure 11: A comparison of scaling in the experimental setups of Gadre et al. (2025) and Magnusson et al. (2025). Both trained language models on C4 and evaluated via few-shot prompting on the tasks above; however, their experimental setups differ: architectural details, prompts, number of shots, task format, and in some cases the number of answer choices (Commonsense QA and SIQA). Such experimental details totally change scaling behavior.

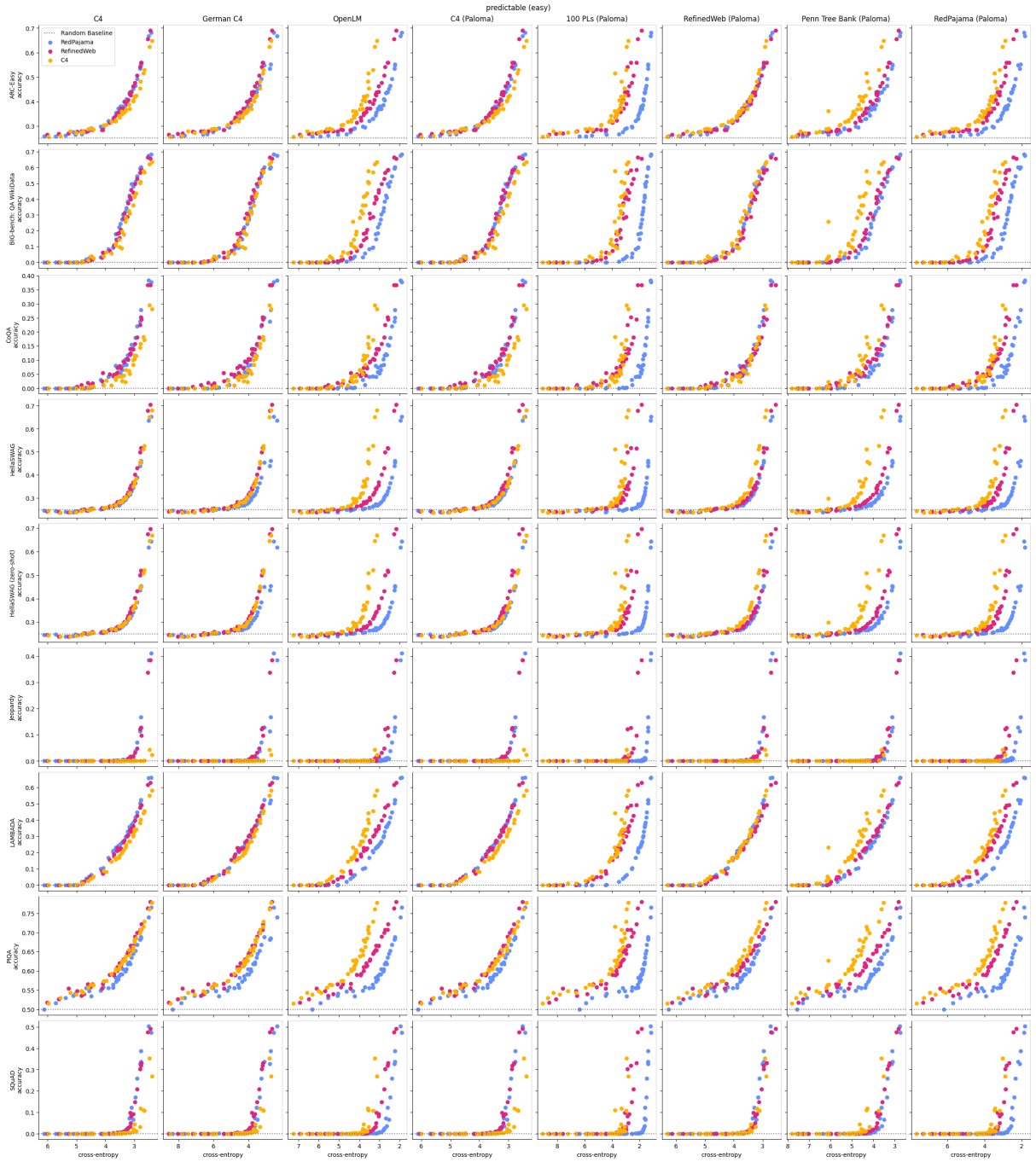

Figure 12: Different pretraining corpora will appear to be the best for downstream tasks, depending on choice of validation dataset.