Neural Neural Scaling Laws

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Neural scaling laws predict how language model performance improves with increased training inputs. While aggregate metrics like validation loss can follow smooth power-law curves, individual downstream tasks exhibit diverse scaling behaviors: some improve monotonically, others plateau, and some even degrade with scale. We argue that predicting downstream performance from validation loss suffers from two limitations: averaging token-level losses obscures signal, and no simple parametric family can capture the full spectrum of scaling behaviors. To address this, we propose Neural Neural Scaling Laws (NEUNEU), a neural network that frames scaling law prediction as time-series extrapolation. NEUNEU combines temporal context from observed accuracy trajectories with token-level validation losses, learning to predict future performance without the limitations inherent in assuming a specific functional form. Trained entirely on open-source model checkpoints from HuggingFace, NEUNEU achieves 1.99% mean absolute error in predicting model accuracy on 66 downstream tasks—a 44% reduction compared to logistic scaling laws (3.56% MAE)—and generalizes zero-shot to unseen model families, architectures, and parameter counts.

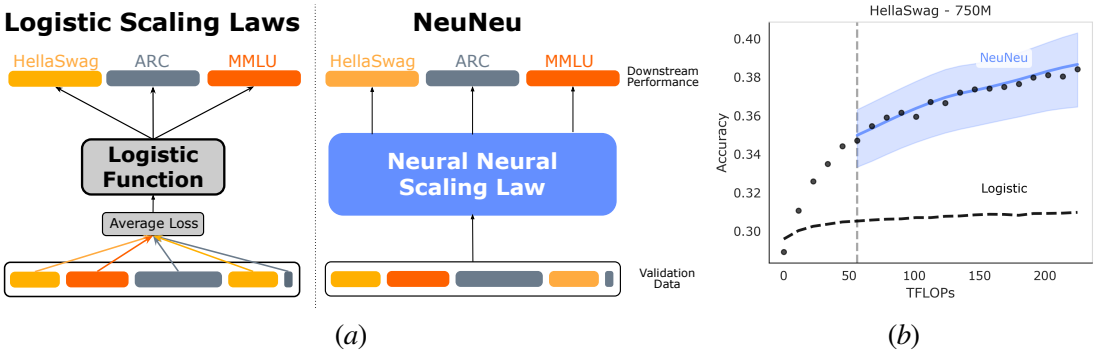
1. Introduction

Neural scaling laws characterize how language model performance improves with increased compute, data, and parameters [15, 17]. These laws typically take the form of power-law relationships such as $L(C) = \alpha C^{-\beta}$, where L is the reducible loss and C is an input to training, like compute. Such simple functional forms have proven remarkably useful for predicting training dynamics and optimizing resource allocation.

However, the translation from pretraining loss to downstream performance is far more complex. While aggregate metrics like pretraining loss or task performance averaged over many domains follow smooth scaling curves, individual tasks exhibit diverse behaviors as they scale: some improve monotonically, others plateau, and some even degrade with scale—a phenomenon known as inverse scaling [23]. Taken together, it seems that no single parametric family can capture the full spectrum of scaling behaviors [21].

We hypothesize that predicting future performance from validation loss suffers from two flaws, both of which limit the usefulness of downstream scaling laws: first, validation loss creates a bottleneck by averaging token losses into a single, obscured signal; and second, no simple hypothesis class exists to fit all behaviors of downstream tasks. To fix these issues, we use a neural network that predicts downstream task performance while incorporating token-level loss information.

Our "neural" neural scaling law, or NEUNEU, frames scaling law prediction as a time-series extrapolation problem. Unlike parametric approaches that rely on aggregate metrics, NEUNEU predicts downstream performance by combining observed accuracy trajectories with token-level validation losses. This allows the model to leverage the signal within loss distributions that averaging typically obscures, as well as the trends within accuracy trajectories that pointwise loss-to-accuracy mappings ignore. To ensure generalization across unseen model families and parameter counts, we design inputs to be invariant across



(*a*) (*b*)

Figure 1: Richer signal from token-level losses (a) enables NEUNEU to better forecast accuracies (b).

model scales. We achieve this by abstracting training steps into relative compute intervals and converting unbounded losses into token probabilities, enabling the network to learn patterns in training dynamics decoupled from the specific training configuration.

We train NEUNEU on open-source language model training trajectories [22] on HuggingFace [41], meaning that anyone can fit their own neural neural scaling laws without first performing large numbers of training runs. Our results show that NEUNEU achieves 1.99% mean absolute error (MAE) on 66 downstream tasks, a 44% reduction compared to the widely used logistic scaling laws [11, 22].

2. Neural Neural Scaling Laws

Problem setting. Periodically, we evaluate a language model on a validation set of N tokens and D downstream tasks. Suppose we are at time t and want to predict a language model's downstream accuracy at time $t+K$. We have at our disposal token-level loss vectors $\boldsymbol{\ell}_{1:t} \in \mathbb{R}^{t\times N}$ and observed downstream accuracies $\mathbf{y}_{1:t} \in [0,1]^{t\times D}$.

Logistic scaling laws solve this prediction problem by assuming that 1) the average validation loss $\bar{\ell}_t$ is sufficient to predict all downstream task performances $\mathbf{y}_t$ and 2) the relationship between validation loss and downstream task performance is well-described by a logistic function, which is reasonable for predicting values that transition between chance and a saturating threshold. One then fits the parameters of the logistic function: $y_t^{(j)} = f(\bar{\ell}_t;a,k,L_0,b) = \frac{a}{1+e^{-k(\ell_t-L_0)}} + b$.

These scaling laws have high bias and low expressivity. In moving to a neural network, which is more expressive, we must choose input and output representations that allow it to extrapolate. We present the architecture in §2.1, and finish with training and evaluation details in §2.2 and §2.3.

2.1. Architecture

NEUNEU observes a sequence of accuracies and compute gaps $y_1,g_1,\ldots,y_t,g_t$ and use this information to predict the next accuracy y_{t+g_t} in the sequence. Gaps are units of compute between evaluation steps. For example, the sequence $(0.5,1,0.6)$ means that the LM's accuracy was 0.5, and after one unit of training compute (*e.g.*, 500 TFLOPs), the accuracy is now 0.6. Abstracting compute as gaps induces invariance across LM training scales, enabling generalization.

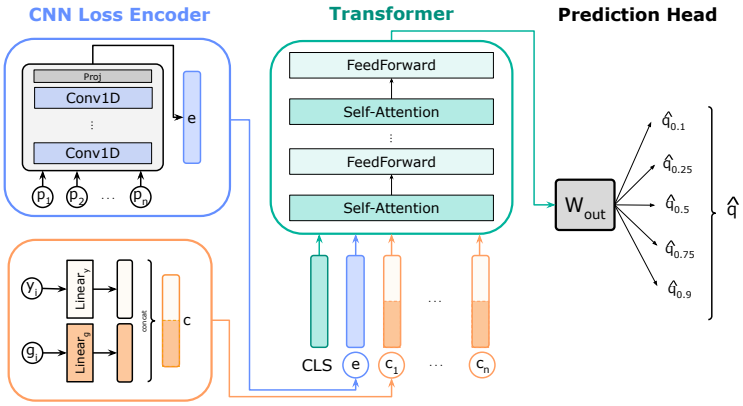


Figure 2: NEUNEU encodes and processes token-level validation probabilities alongside a sequence of historical downstream accuracies and compute gaps, which are projected into context tokens. The BERT-style transformer [9] backbone uses this information to predict a distribution over the downstream accuracy via quantile regression on the [CLS] token representation.

NEUNEU consists of three components: loss encoder, transformer, and prediction head. See Figure 2. We show the forward pass for NEUNEU first, then explain each component.

$$\mathbf{e} = \text{LossEncoder}(\mathbf{p}_t) \tag{1}$$

$$\mathbf{c}_i = \text{concat}\big(\text{Linear}_y(y_i), \text{Linear}_g(g_i)\big) \quad \text{for } i = 1,\ldots,t \tag{2}$$

$$\mathbf{H} = \text{Transformer}([\text{CLS};\mathbf{e};\mathbf{c}_1;\ldots;\mathbf{c}_t]) \tag{3}$$

$$\hat{\mathbf{q}} = \mathbf{W}_{\text{out}} \cdot \mathbf{H}_0 \tag{4}$$

Loss Encoder. The loss encoder in Equation (1) produces an embedding $\mathbf{e}$. To process N token probabilities, we use L layers of learned 1D convolutions for hierarchical downsampling:

$$\begin{aligned}
\mathbf{x}_0 &= \mathbf{p}_t \\
\mathbf{x}_i &= \text{GELU}(\text{GroupNorm}(\text{Conv1D}(\mathbf{x}_{i-1}))) \\
\mathbf{e} &= \mathbf{W}_{\text{proj}} \cdot \text{flatten}(\mathbf{x}_L)
\end{aligned}$$

Transformer. The main sequence model is a standard transformer encoder with bidirectional attention [9, 36] and rotary embeddings (RoPE, Su et al. [33]). The first half of $\mathbf{c}_i$ contains information about y_i, and the second half about g_i. The input sequence to the transformer is $[\text{CLS};\mathbf{e};\mathbf{c}_1;\ldots;\mathbf{c}_t]$. The transformer processes this with 6 layers of pre-norm self-attention, and the output is predicted from the CLS token position.

Prediction Head. Unlike logistic scaling laws, our model also captures prediction uncertainty via quantile regression [18]. The output embedding $\mathbf{H}_0$ from the CLS position is projected to $Q = 5$ quantile predictions: $\hat{\mathbf{q}} = [\hat{q}_{0.1}, \hat{q}_{0.25}, \hat{q}_{0.5}, \hat{q}_{0.75}, \hat{q}_{0.9}] = \mathbf{W}_{\text{out}} \cdot \mathbf{H}_0$. Training uses pinball loss [18] summed across quantiles. The output $\hat{\mathbf{q}}$ provides a calibrated distribution over predicted accuracy at the target compute scale.

2.2. Training Data

A critical step in training such a neural predictor is obtaining diverse LM pretraining trajectories. In this work, we demonstrate that such data is **freely available on HuggingFace**: anyone can train our model

from open-source data. In particular, we train NEUNEU using training runs of 6 model sizes from the DataDecide model suite [22]. Each training trajectory contains three random seeds, a variable number of checkpoints, and each checkpoint is evaluated on the 66 downstream tasks including the OLMES evaluation suite [13]. We use random seed 0 from the{90M, 150M, 300M, 530M, 750M, 1B} parameter models, saving other seeds for evaluation. In total, we use 6 model sizes trained on 24 different pretraining datasets, yielding 144 unique pretraining trajectories, or $144 \times 66 = 9{,}504$ unique accuracy trajectories. For all training and evaluation details, see Appendix B.

2.3. Evaluation and Baselines

We test generalization of all scaling laws on four kinds of unseen language model pretraining runs: new 1) random seeds, 2) pretraining datasets, 3) model families, and 4) downstream tasks. We use:

- Random seed 2 from DCLM-Baseline training runs with parameter sizes{90M, 150M, 300M, 530M, 750M, 1B} [22].
- C4 training runs with parameter sizes{90M, 150M, 300M, 530M, 750M, 1B}, seed 0. We withhold all C4 runs from the scaling laws' training data [22].
- Pythia [4] runs with parameter sizes{70M, 1.4B, 2.8B, 6.9B, 12B} and OLMo-Hybrid-7B [24]. These model sizes lie outside of the training data distribution of our predictive models and represent challenging shifts in pretraining dataset, architecture, and parameter size. In particular, OLMo-Hybrid-7B is not a transformer—it mixes attention and recurrent neural network layers.

Baselines. We train all predictors on the training data described in §2.2. At inference time, we condition all transformer models on accuracies from the first 20% of each heldout training run and compute mean absolute error (MAE) for all methods on the unobserved 80%.

- LOGISTIC: Logistic scaling law, fitted per task. To give LOGISTIC the best possible chance, we feed it ground truth average loss $\bar{\ell}_{t+K}$ for checkpoint $t+K$: $\hat{y}^{(i)}_{t+K} = f(\bar{\ell}_{t+K}; a, k, L_0, b)$. We assume that whatever prediction method one uses to obtain $\bar{\ell}_{t+K}$ is perfect.
- LC-PFN: A meta-learned, open-source transformer model that performs in-context Bayesian inference over learning curves [1].
- BNSL: Broken neural scaling laws [5], which relax the assumptions of logistic scaling laws by allowing a break or transition between different scaling regimes. Also receives $\bar{\ell}_{t+K}$.
- NOLOSS, AVERAGE: Ablations of NEUNEU that take no loss information or the averaged token probabilities. Discussed in detail in §A. If our hypothesis that token-level information is useful holds, then NEUNEU should outperform NOLOSS and AVERAGE.

3. Results

NEUNEU performs well on all evaluation tasks. In Figure 3, NEUNEU achieves the lowest mean absolute error across all evaluation conditions. NEUNEU also outperforms the NOLOSS and AVERAGE ablations, supporting our hypothesis that average validation loss discards distributional information helpful for downstream prediction.

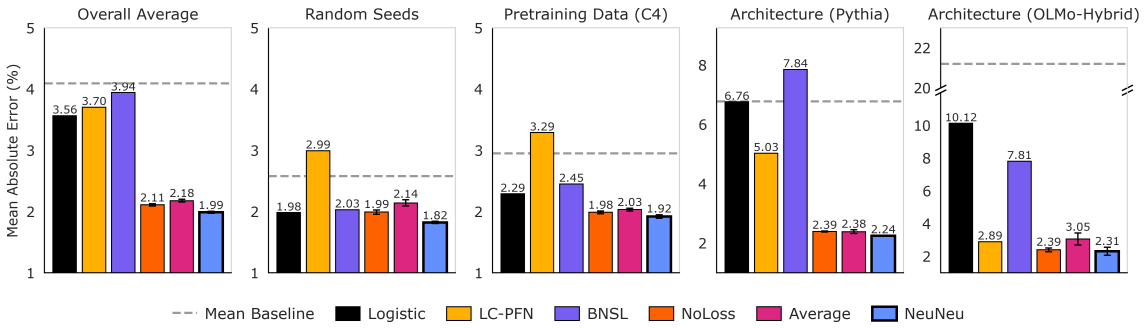


Figure 3: NEUNEU significantly outperforms all other scaling laws at generalizing to new random seeds, pretraining data, and unseen transformer and non-transformer architectures. The gray dashed line denotes mean absolute error (MAE) of always predicting the average accuracy from the training set for each task.

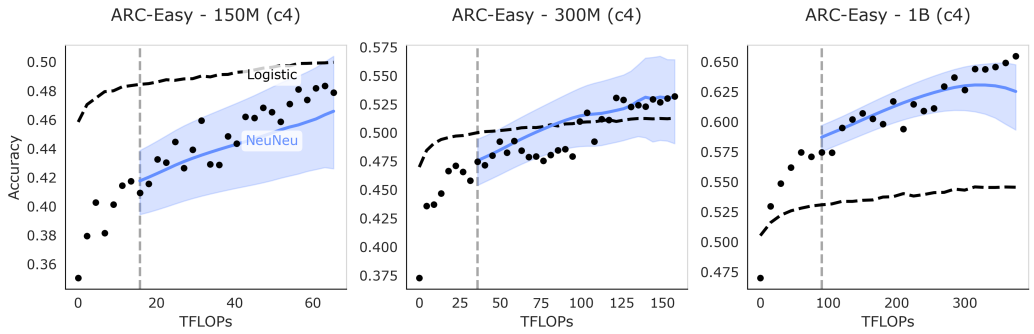


Figure 4: Visualizing NEUNEU's predictions. Black dots are ground truth accuracies, and the grey line marks the beginning of NEUNEU's predictions, after observing the first 20% of the training run. The light-blue band is the 10%-90% interquantile interval predicted by NEUNEU itself.

Parametric scaling laws fail catastrophically on new architectures. Logistic and broken neural scaling laws perform progressively worse on more challenging generalizations. On Pythia and OLMo-Hybrid runs—which differ from the training trajectories in pretraining corpus, parameter count, and architecture—LOGISTIC and BNSL incur around 4 times higher error than on in-distribution (different random seed) evaluations. This suggests that parametric scaling laws like LOGISTIC and BNSL generalize poorly to large changes in training. LC-PFN performs roughly on par with logistic scaling laws, and worse than our neural methods; however, its predictions are not dramatically worse on Pythia and OLMo-Hybrid. Ultimately, NEUNEU has the advantage of being trained specifically to extrapolate language model scaling, whereas LC-PFN is trained to predict learning curves in general. Overall, NEUNEU outperforms all baselines with non-overlapping error bars, the gap widening on the most challenging out-of-distribution evaluations.

Visualization: ARC-Easy across scales. In Figure 4, we contrast predictions from NEUNEU and LOGISTIC on ARC-Easy, a task representing roughly median predictive performance for NEUNEU. Predictions for other tasks can be found in Appendix Figures 7 through 9. Across the 150M to 1B training runs, NEUNEU has tighter fit to the ground truth accuracies and adjusts its predictions based on the model scale. Conversely, we observe that the logistic scaling law overpredicts performance for the 150M model and underpredicts performance for the 1B model.

References

[1] Steven Adriaensen, Herilalaina Rakotoarison, Samuel Müller, and Frank Hutter. Efficient bayesian learning curve extrapolation using prior-data fitted networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 19858–19886. Curran Associates, Inc., 2023.

[2] Ibrahim M Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22300–22312. Curran Associates, Inc., 2022.

[3] 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.

[4] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[5] Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sckjveqlCZ.

[6] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 4 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-24271-9. URL https://doi.org/10.1038/s41598-018-24271-9.

[7] 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.

[8] Corinna Cortes, L. D. Jackel, Sara Solla, Vladimir Vapnik, and John Denker. Learning curves: Asymptotic values and rate of convergence. In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann, 1993. URL https://proceedings.neurips.cc/paper_files/paper/1993/file/1aa48fc4880bb0c9b8a3bf979d3b917e-Paper.pdf.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

[10] Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. Understanding emergent abilities of language models from the loss perspective. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 53138–53167. Curran Associates, Inc., 2024. doi: 10.52202/079017-1683.

[11] 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, Thomas Kollar, Yair Carmon, Achal Dave, Reinhard Heckel, Niklas Muennighoff, and Ludwig Schmidt. Language models scale reliably with over-training and on downstream tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iZeQBqJamf.

[12] 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.

[13] Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. OLMES: A standard for language model evaluations. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5005–5033, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.282. URL https://aclanthology.org/2025.findings-naacl.282/.

[14] 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.

[15] 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, Oriol Vinyals, Jack William Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=iBBcRUlOAPR.

[16] Yuzhen Huang, Jinghan Zhang, Zifei Shan, and Junxian He. Compression represents intelligence linearly. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=SHMj84U5SH.

[17] 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. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

[18] Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of economic perspectives*, 15 (4):143–156, 2001.

[19] Jakub Krajewski, Amitis Shidani, Dan Busbridge, Sam Wiseman, and Jason Ramapuram. Revisiting the scaling properties of downstream metrics in large language model training, 2025. URL https://arxiv.org/abs/2512.08894.

[20] Emmy Liu, Amanda Bertsch, Lintang Sutawika, Lindia Tjuatja, Patrick Fernandes, Lara Marinov, Michael Chen, Shreya Singhal, Carolin Lawrence, Aditi Raghunathan, Kiril Gashteovski, and Graham Neubig. Not-just-scaling laws: Towards a better understanding of the downstream impact of language model design decisions. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16396–16427, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.830. URL https://aclanthology.org/2025.emnlp-main.830/.

[21] Nicholas Lourie, Michael Y. Hu, and Kyunghyun Cho. Scaling laws are unreliable for downstream tasks: A reality check. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16167–16180, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.877. URL https://aclanthology.org/2025.findings-emnlp.877/.

[22] 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. Datadecide: How to predict best pretraining data with small experiments. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=p9YlQPF8fE.

[23] 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, Alisa Liu, Jiacheng Liu, Tom Tseng, Tomasz Korbak, Najoung Kim, Samuel R. Bowman, and Ethan Perez. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=DwgRm72GQF. Featured Certification.

[24] William Merrill, Yanhong Li, Tyler Romero, Anej Svete, Caia Costello, Pradeep Dasigi, Dirk Groeneveld, David Heineman, Bailey Kuehl, Nathan Lambert, Chuan Li, Kyle Lo, Saumya Malik, DJ Matusz, Benjamin Minixhofer, Jacob Morrison, Luca Soldaini, Finbarr Timbers, Pete Walsh, Noah A. Smith, Hannaneh Hajishirzi, and Ashish Sabharwal. Olmo hybrid: From theory to practice and back, 2026. URL https://arxiv.org/abs/2604.03444.

[25] 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.

[26] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202/.

[27] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[28] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[29] Herilalaina Rakotoarison, Steven Adriaensen, Neeratyoy Mallik, Samir Garibov, Edward Bergman, and Frank Hutter. In-context freeze-thaw bayesian optimization for hyperparameter optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[30] Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryenvpEKDr.

[31] 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, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 55565–55581. Curran Associates, Inc., 2023.

[32] 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? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=I1NtlLvJal.

[33] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568 (C), February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

[34] Kevin Swersky, Jasper Snoek, and Ryan Prescott Adams. Freeze-thaw bayesian optimization, 2014. URL https://arxiv.org/abs/1406.3896.

[35] Lindia Tjuatja and Graham Neubig. BehaviorBox: Automated discovery of fine-grained performance differences between language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18851–18873, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.923. URL https://aclanthology.org/2025.acl-long.923/.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[37] 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. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

[38] Jason Wei, Najoung Kim, Yi Tay, and Quoc V. Le. Inverse scaling can become u-shaped, 2023. URL https://arxiv.org/abs/2211.02011.

[39] Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the web: Constructing domains enhances pre-training data curation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=boSqwdvJVC.

[40] Ethan G Wilcox, Michael Hu, Aaron Mueller, Tal Linzen, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Ryan Cotterell, and Adina Williams. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics, Jul 2024. URL osf.io/preprints/psyarxiv/rfwgd_v1.

[41] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

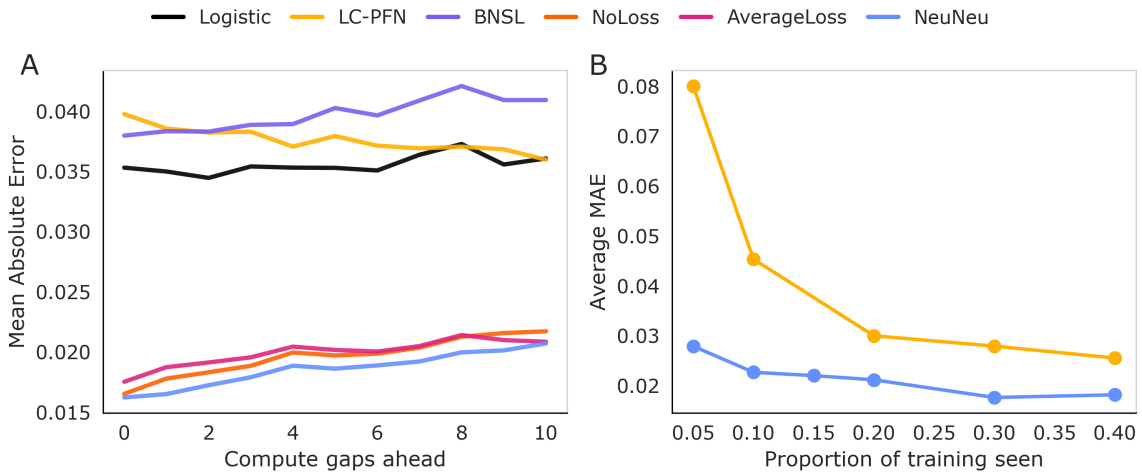


Figure 5: Additional extrapolation results. (A) NEUNEU and the neural ablations maintain low error as the extrapolation horizon increases, while parametric and general-purpose learning-curve baselines remain substantially higher. (B) LC-PFN improves as more of the training trajectory is observed, indicating that it is inferring from context.

Appendix A. Validation Loss Representations

One drawback of logistic scaling laws is that the average loss $\bar{\ell}_t$ does not retain distributional information, which we hypothesize is beneficial for predicting downstream performance. Two models can achieve the same validation loss with loss distributions of different skews or variances, which could be indicative of different underlying capabilities.

First, to fix the issue with cross-entropy loss being unbounded, we convert token-wise losses into token-wise probabilities:

$$p_{t,i} = e^{-\ell_{t,i}} \quad \text{for } i = 1,\ldots,N$$

In general, we found that training on probabilities leads to better neural models than training on losses; see Figure 6. To test our hypothesis about distributional information, we consider three neural models, distinguished by how they use token probabilities:

- NEUNEU: Takes the token probability vector $\mathbf{p}_t \in \mathbb{R}^N$, where N denotes the validation set size or a subsample thereof. We use $N = 256{,}000$ in practice, roughly the size of one LM training minibatch.
- AVERAGE: Takes all observed average validation probabilities $\bar{\mathbf{p}}_{\leq t}$, similar to how logistic scaling laws use the average loss.
- NOLOSS: Takes no information about token probabilities or losses. The model makes predictions using the downstream accuracies $\mathbf{y}_{1:t}$ only.

Appendix B. Reproducibility

Code: https://anonymous.4open.science/r/neuneu-144C/

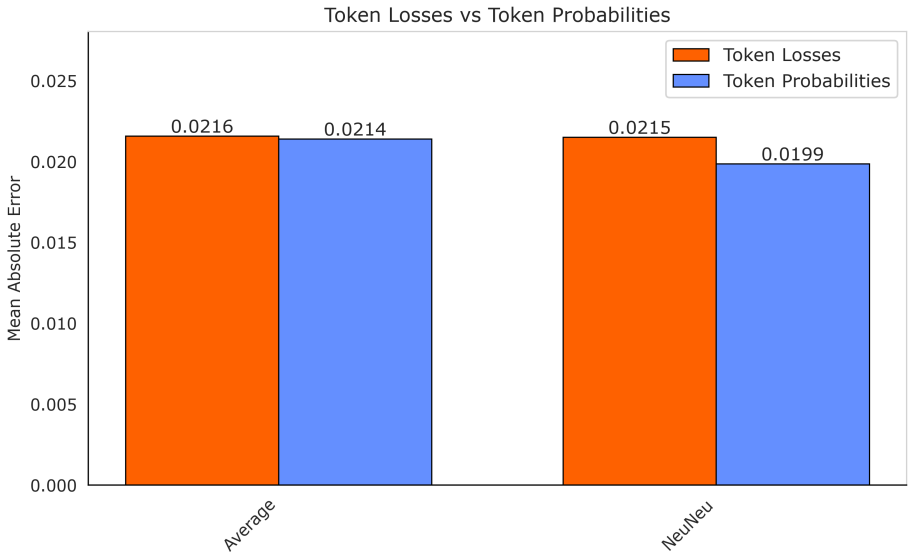


Figure 6: Using token probabilities produces better neural predictors than token losses. Another reason to use probabilities is that the function e^{-x}, or the conversion from loss to probability, has larger derivative for smaller loss values, meaning that small changes in loss near convergence for the language model are amplified. This likely matters more than changes when the loss is large and language model performance on downstream tasks is near-chance.

Data augmentation. For each model and task, we construct training samples from checkpoint accuracies as follows. Let $(y_1,y_2,\ldots,y_T)$ denote the sequence of accuracies at consecutive checkpoints. We first impute unit gaps to form the context sequence: $\mathcal{S}_0=[(y_1,1),(y_2,1),\ldots,(y_T,1)]$.

To create multiple examples from one training trajectory and a model that is robust to missing data, we randomly drop tuples from the sequence with probability $p=0.4$, as inspired by Che et al. [6]. When tuple i is dropped, its gap is absorbed into the preceding tuple:

$$[(y_{i-1},g_{i-1}),(y_i,g_i),(y_{i+1},g_{i+1})] \xrightarrow{\text{drop } y_i} [(y_{i-1},g_{i-1}+g_i),(y_{i+1},g_{i+1})]$$

For a subsequence $\mathcal{S}=[(y_{s_1},g_{s_1}),\ldots,(y_{s_k},1)]$ ending at checkpoint s_k, we generate training targets for all future checkpoints $j, s_k<j\leq T$. Let $g_{\text{target}}=j-s_k$. We simply replace the final 1 with g_{target} to create $\mathcal{S}'$, then add the appropriate input representation:

$$(\mathcal{S}',\mathbf{p}_t)\mapsto y_j \quad (\text{NEUNEU}), \qquad (\mathcal{S}',[\bar{p}_1,\ldots,\bar{p}_{s_k}])\mapsto y_j \quad (\text{AVERAGE})$$

We evaluate all checkpoints on a shard of the WebOrganizer dataset [39] and save the token probabilities. When training NEUNEU, we sample random spans of length 256,000. To handle tokenization differences across models, we tokenize on whitespace and combine probabilities of subwords, following Tjuatja and Neubig [35]. See Table 1 for hyperparameters.

From DataDecide [22], we meta-train on trajectories from the model sizes{90M, 150M, 300M, 530M, 750M, 1B} and all pretraining datasets except C4, which we use for evaluation. In total, this yields 6 model sizes $\times$ 24 pretraining datasets = 144 training configurations, where each training configuration contains a sequence of model checkpoints.

Table 1: Model hyperparameters, shared across our neural models. We chose learning rate and weight decay values based on models of similar size in [22].

Hyperparameter	**Value**
Transformer Encoder Architecture	
Hidden dimension	512
Transformer layers	6
Attention heads	8
Feed-forward dimension	2048
Max sequence length	512
Convolutional Neural Network	
CNN channels	[8, 16, 32, 64]
CNN kernel size	64
CNN stride	16
Number of groups for GroupNorm	min(8, number of channels)
Padding	kernel size / 2
Training	
Optimizer	AdamW
Batch size	256
Learning rate	6×10^{-4}
Weight decay	0.033
Epochs	3
Warmup ratio	0.1
Max gradient norm	1.0
Drop gap probability (see §2.2)	0.4
Random seed	{0,1,2,3,4}
Data	
Max encoder tokens	256,000
Quantiles	[0.1, 0.25, 0.5, 0.75, 0.9]

To get validation losses for these model checkpoints, we evaluated them on shard 141 from the WebOrganizer dataset [39], which we chose randomly from all shards and can be accessed at https://huggingface.co/datasets/WebOrganizer/Corpus-200B. NEUNEU takes random token probability spans of length 256,000 as input during training. During evaluation, we take the first span from shard 141 for simplicity.

Compute. All experiments run on a cluster with a mix of L40S and H200 GPUs. We trained NOLOSS, AVERAGE, and NEUNEU on L40S GPUs for roughly 2 GPU hours. Inference for larger models like Pythia-12B and OLMo-Hybrid were done on H200 GPUs.

We compare our meta-model against two baselines: LC-PFN [1], a learning-curve extrapolator, and Broken Neural Scaling Laws (BNSL) [5], fit as a zero-shot loss-to-accuracy mapping in the same regime as our logistic baseline.

B.1. LC-PFN

We use the public `lcpfn` package without retraining. LC-PFN is a prior-data fitted network with fixed weights at release. The model is loaded once per evaluation run via `LCPFN()` in `eval` mode.

For each (model, task) trajectory we condition on the first 20% of checkpoints and predict accuracy at every remaining checkpoint. Training steps are normalized to $[0,1]$ by dividing by the maximum step in the *full* trajectory (not the context window) so that target positions on the curve are preserved; accuracies are passed through unchanged. We query `LCPFN` with

$$\hat{y}_{\text{test}} = \texttt{predict\_quantiles}(x_{\text{train}}, y_{\text{train}}, x_{\text{test}}, q),$$

$q = \{0.1, 0.25, 0.5, 0.75, 0.9\}$. The median is used as the point prediction; the other quantiles supply the predictive intervals reported in the uncertainty plots.

Figure 5B suggests that LC-PFN is working as intended. Like NeuNeu, LC-PFN is a transformer that performs in-context inference, and is not tuned after pretraining. When giving more context to LC-PFN, its prediction error over the remaining accuracies also decreases. Thus, we conclude that LC-PFN is indeed inferring from the existing trajectory, but begins from higher error because it is not specifically trained on the task of downstream scaling prediction.

B.2. BNSL

We treat BNSL as a zero-shot scaling-law baseline analogous to the logistic scaling laws. A one-break BNSL curve is fit per task on the training corpus of (average loss, accuracy) pairs and then applied to the eval model's ground-truth average losses to produce predicted accuracies at every checkpoint. No accuracy from the eval trajectories are observed.

Functional form. We use the one-break form from Caballero et al. [5],

$$y(x) = a + bx^{-c_0}\left(1 + (x/d_1)^{1/f_1}\right)^{-c_1 f_1},$$

fitted to error $y = 1 - \text{acc}$ rather than accuracy, since BNSL models a positive decreasing quantity. Predictions are mapped back via $\hat{\text{acc}} = \text{clip}(1 - \hat{y}, 0, 1)$. For the input x, average loss ℓ is decreasing, so we transform it to a positive progress measure $x = s/\ell$, where s is the median observed token-level loss. To train BNSL, we follow the curve fitting advice in Caballero et al. [5]. BNSL is a deterministic curve fit and does not contribute uncertainty estimates.

Appendix C. Related Work

Over the decades, a number of works discovered and (rediscovered) power law scaling with respect to data [8, 14]. With the advent of language model pretraining [26–28], data became abundant, and the focus shifted towards scaling compute. Rosenfeld et al. [30] discovered that loss exhibits power law scaling with respect to parameters as well as data, and soon after Kaplan et al. [17] named this phenomenon *scaling laws*, popularizing the idea by investigating its implications for language models. Later, Hoffmann et al. [15] refined the method by proposing different ways to estimate scaling laws and recommending the most widely used power law functional form: $L(N,D) = e + \frac{a}{N^\alpha} + \frac{b}{D^\beta}$. This approach remains the basis for how scaling laws are applied to pretraining.

Unfortunately, translation from pretraining to downstream tasks has proven more difficult. Wei et al. [37] documented how models display *emergent capabilities*, or capabilities that appear suddenly at scale.

Such capabilities are hard to predict, because model performance appears the same at smaller scales. The choice of evaluation metric can ease or exacerbate the problem of emergence [31, 32], but even with carefully constructed metrics, extrapolating downstream performance remains a challenge [10]. As Liu et al. [20] note, factors beyond compute or loss impact scaling laws; we embrace this fact by providing NEUNEU richer input representations.

A major obstacle for extrapolation is the diversity of scaling behaviors. While many tasks improve with scale, others exhibit *inverse scaling*, where performance gets worse [23, 40] or does so at first only to become U-shaped [38]. To capture these behaviors, researchers have tried creating more complex parametric forms [2, 5], predicting performance directly from data, parameters, and compute [19, 25], and predicting performance from intermediate task losses such as pretraining loss or the probability of the correct answer [3, 7, 11, 12, 16]. Despite these efforts, reliably predicting downstream scaling remains a challenge [21].

Our work attempts to move scaling laws beyond parametric assumptions. In doing so, it relates closely to a precursor of scaling laws: trajectory forecasting. Trajectory forecasting has long been studied in hyperparameter optimization; for example, Freeze-Thaw Bayesian Optimization forecasts the asymptotic performance of partially trained models to dynamically allocate compute [34]. This perspective has recently evolved into meta-learning approaches such as LC-PFN [1] and FT-PFN [29], which use transformers trained on synthetic functions to perform in-context inference over learning curves. Our work shares the motivation of these methods but diverges in methodology: rather than using synthetic priors, we learn to extrapolate directly from open-source training runs using quantile regression.

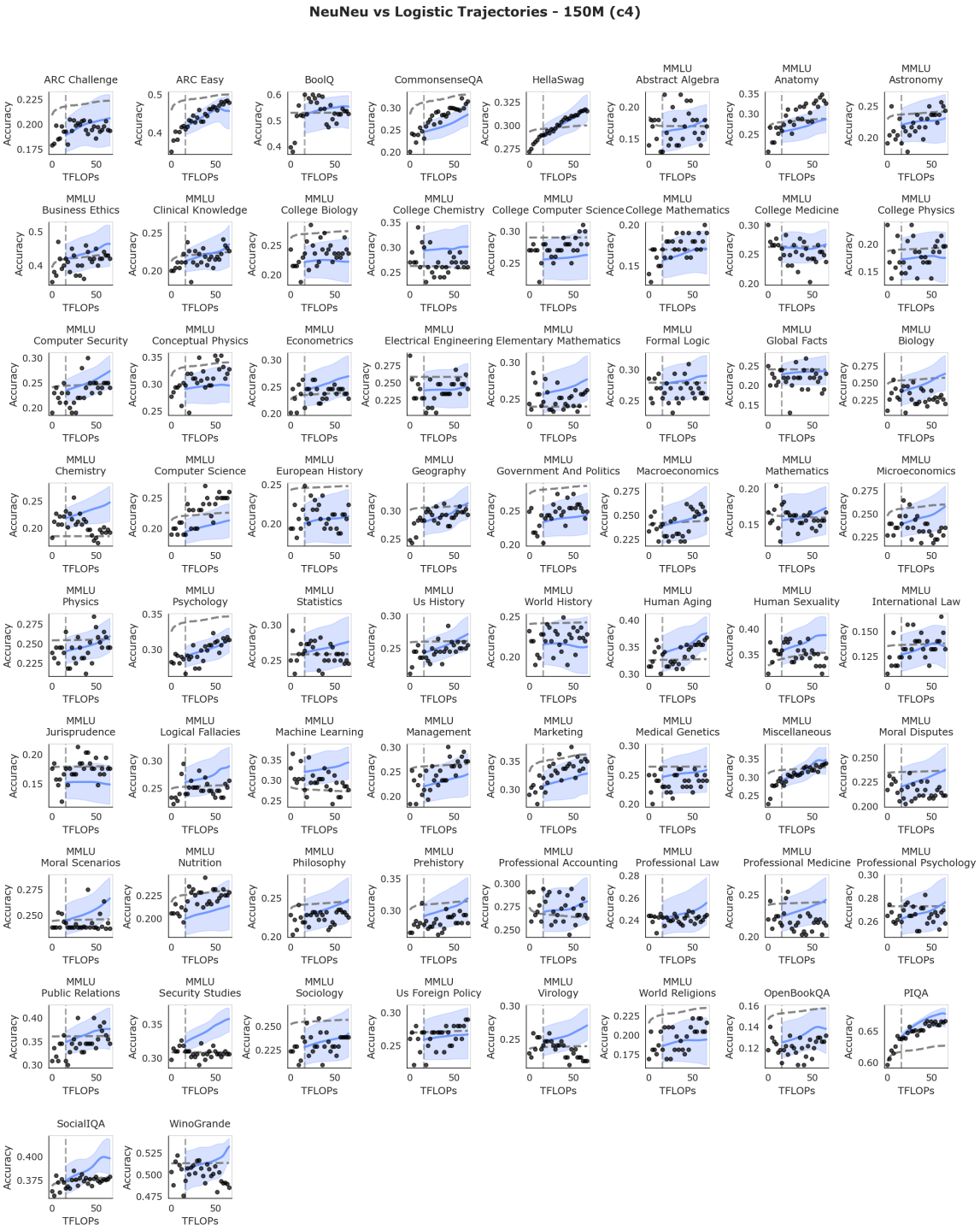


Figure 7: Blue: NEUNEU. Dark grey: Logistic scaling law fitted to the task on the training set (§2.2).

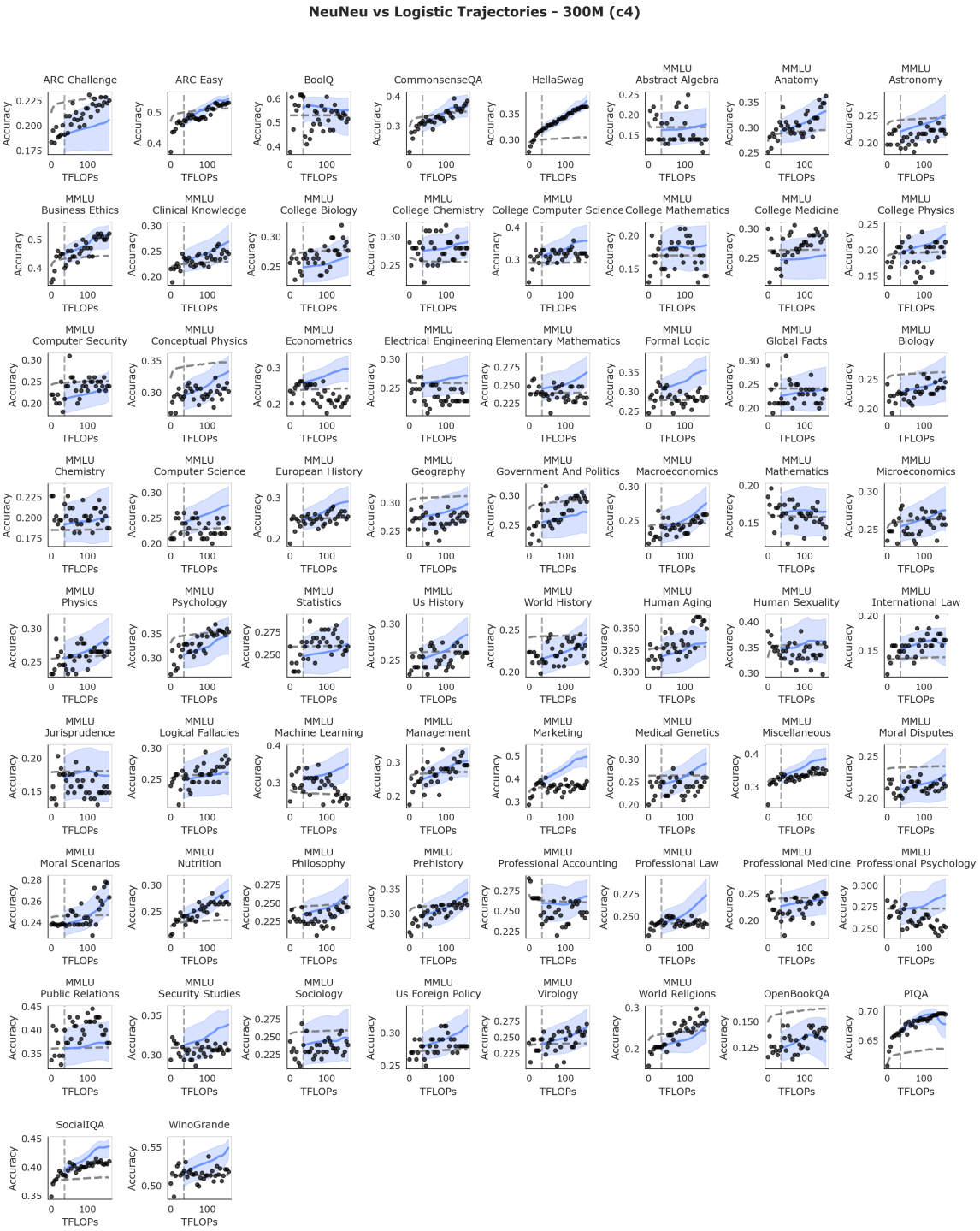


Figure 8: Blue: NEUNEU. Dark grey: Logistic scaling law fitted to the task on the training set (§2.2).

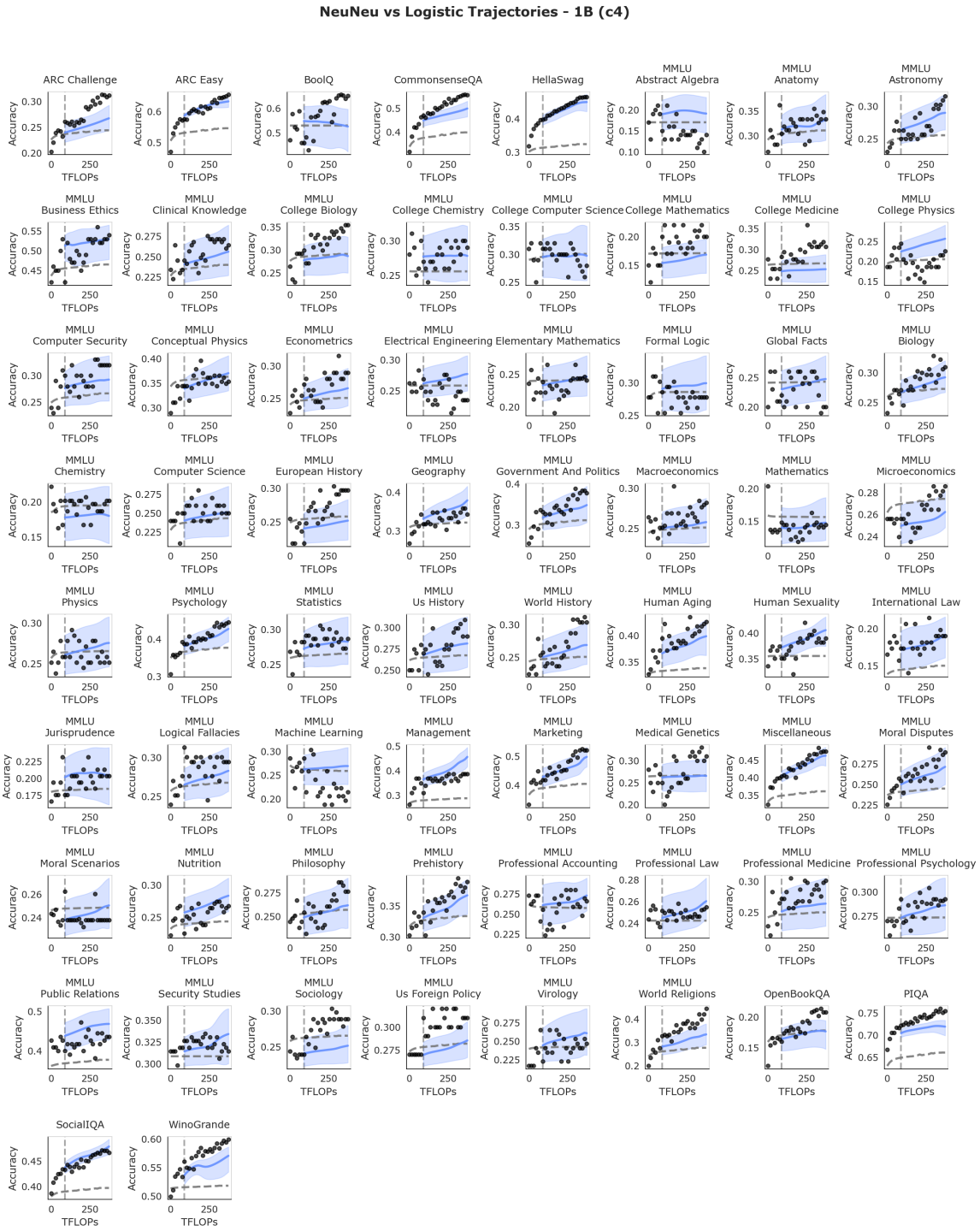


Figure 9: Blue: NEUNEU. Dark grey: Logistic scaling law fitted to the task on the training set (§2.2).