# Revisiting the Scaling Properties of Downstream Metrics in Large Language Model Training

**Jakub Krajewski**[2,3*†]
gim.jakubk@gmail.com

**Amitis Shidani**[1*]
amitis_shidani@apple.com

**Dan Busbridge**[1]
dbusbridge@apple.com

**Sam Wiseman**[1]
s_wiseman@apple.com

**Jason Ramapuram**[1]
jramapuram@apple.com

[1] **Apple** [2] **University of Warsaw** [3] **IDEAS NCBR**

## Abstract

While scaling laws for Large Language Models (LLMs) traditionally focus on proxy metrics like pretraining loss, predicting downstream task performance has been considered unreliable. This paper challenges that view by proposing a direct framework to model the scaling of benchmark performance from the training budget. We find that for a fixed token-to-parameter ratio, a simple power law can accurately describe the scaling behavior of log accuracy on multiple popular downstream tasks. Furthermore, we introduce functional forms that predict accuracy across token-to-parameter ratios and account for inference compute under repeated sampling. Our results show that the direct approach extrapolates better than the previously proposed two-stage procedure, which is prone to compounding errors. We validate our findings on models with up to 17B parameters trained on up to 350B tokens across two dataset mixtures. To support reproducibility and encourage future research, we release the complete set of pretraining losses and downstream evaluation results.

## 1 Introduction

Large Language Models (OpenAI et al., 2024; Team et al., 2025; DeepSeek-AI et al., 2025) based on the Transformer (Vaswani et al., 2023) architecture have achieved impressive results, approaching or exceeding human-level performance across multiple domains. Scaling Laws (Hestness et al., 2017; Kaplan et al., 2020) are an established method for modeling the performance of these networks, enabling researchers to plan large-scale training runs based on curated sets of smaller experiments. Traditionally, these laws focus on predicting proxy metrics for model quality, such as pre-training log-perplexity. This has proven invaluable for optimizing training hyperparameters, like the optimal ratio of tokens to parameters.

Another important direction in understanding the scaling of LLMs is tracking the behavior of more interpretable indicators of model capabilities, like accuracy on downstream benchmarks measuring the performance on general knowledge, reasoning, math and coding tasks. Despite early attempts to solve this problem (Grattafiori et al., 2024; Isik et al., 2025; Chen et al., 2025), scaling downstream metrics have been often referred to as noisy and unreliable (Schaeffer et al., 2025b; Lourie et al., 2025).

[†]Work done as an intern at Apple. [*] Core contributors.

Current approaches to modeling the downstream performance performance of LLMs (Grattafiori et al., 2024; Chen et al., 2025; Bhagia et al., 2025) typically rely on a two-stage approach, where the training budget is first mapped to a proxy metric like mean log-probability of the correct answer, and then another dependence is established, mapping to benchmark accuracy.

In this paper, we propose a framework for directly predicting downstream performance from the pre-training budget. We demonstrate that when holding the token-to-parameter ratio fixed, the scaling of downstream accuracy is accurately captured by a simple power law. We validate this law across an extensive suite of 130 experiments, spanning models up to 17B parameters trained on 350B tokens, and evaluate its predictive power on twelve popular benchmarks used for evaluating LLM capabilities. Our contributions can be summarized as follows:

1. We show that the downstream model performance on multiple popular benchmarks scales predictably with respect to the pre-training budget. We train our models either using C4 or a more modern mixture dataset (Section 3) to highlight the influence of data composition.
2. We propose a *direct*, two-parameter scaling law for downstream performance. We show that, contrary to previous claims of unreliability, it accurately predicts model capabilities from the training budget and offers a simpler alternative to prior two-stage methods.
3. We extend the scaling law to model downstream performance across different token-to-parameter ratios. We further derive the formula for predicting pass rates with repeated sampling in code metrics.

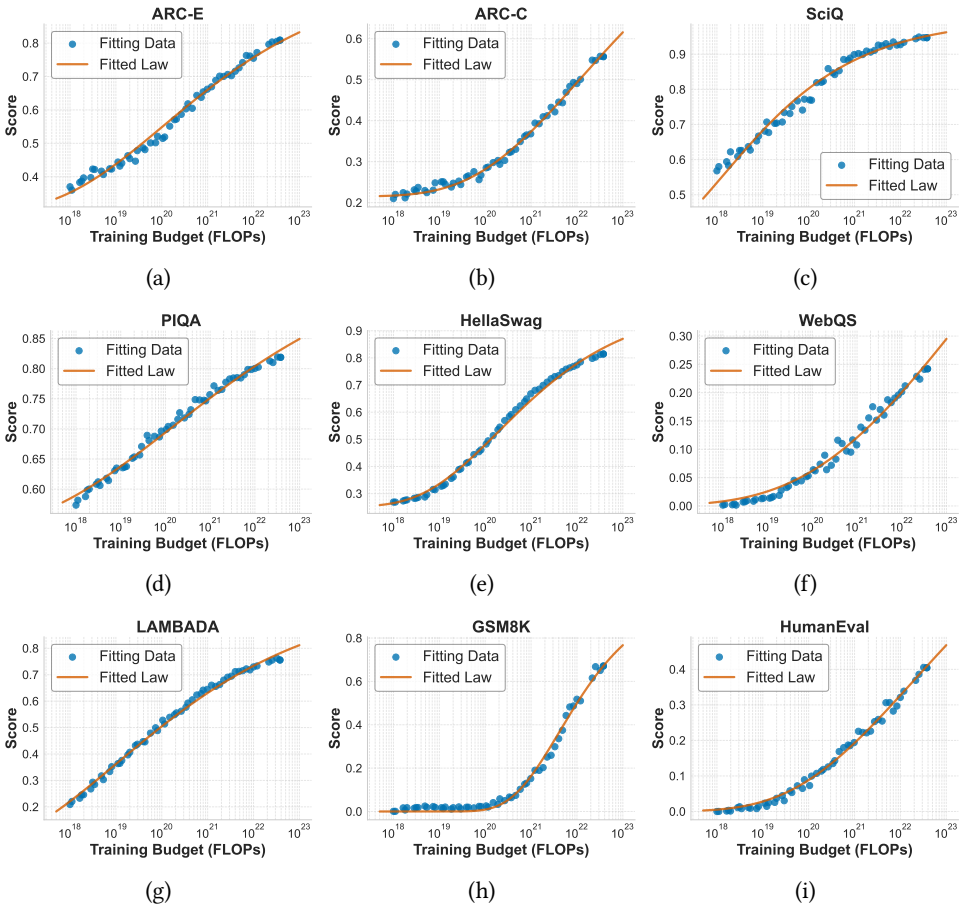



Figure 1: Benchmark accuracy can be described using a direct scaling law based on training FLOPs. The solid line represents the scaling law fit using Equation 2, and each point corresponds to accuracy measured for the final checkpoint at a given training budget.

## 2 Related Work

**Downstream scaling and what it predicts.** A large body of work proposes to forecast downstream task performance from small, cheaper experiments, but they differ in the signal one should extrapolate and in the tasks/metrics used for validation. Two-stage methods posit an intermediate proxy (typically pretraining loss or task-specific negative log-likelihood) and then map that proxy to accuracy on a benchmark. For instance, Chen et al. (2024) fit a compute→loss power law and a subsequent loss→performance map post-emergence.

LLaMA 3 (Dubey et al., 2024) adopts the same two-stage template, correlating training FLOPs with the normalized NLL of the correct choice and then using a sigmoidal link from NLL to accuracy to forecast final scores for ranked multiple-choice tasks. In parallel, loss-centric analyses formalize when and why such mappings can work: Du et al. (2024) explain emergent "breaks" through loss thresholds; Brandfonbrener et al. (2024) show that losses across datasets can be predicted from each other (loss-to-loss), offering a unifying surrogate for downstream ability. (Mayilvahanan et al., 2025) analyze the contribution of various parts of the training procedure and find the crucial importance of the dataset used. (Li et al., 2025b) provides a broad survey of different approaches and discrepancies in the scaling law literature, providing recommendations for reproducibility.

Closer to practice, Gadre et al. (2024a) report that aggregate downstream error is near-linear when plotting log-error vs. validation loss, while Bhagia et al. (2024) and Chen et al. (2024) give systematic recipes for fitting loss-to-accuracy transforms. Our study is technically aligned with these efforts in treating downstream performance as a predictable function of a small number of observable quantities, but we deliberately place the emphasis on the evaluation side: we analyze predictability across metric families (ranked classification, exact-match generation, chain-of-thought math, and code pass rates), and we stress model-agnostic procedures that require no access to internal losses beyond what the evaluation harness already.

**Compute-efficient scaling measurements.** Compute-efficient ladders (Bhagia et al., 2024) advocate measuring task scaling on a small "ladder" of models to infer the trend for a target scale, showing strong fits on ranked classification benchmarks where each item is scored by a model's log-probability over choices. By design, this restricts metrics to ranked classification and uses the model's log-likelihood of the correct answer as the proxy; it is not obvious how to choose or calibrate the proxy for two-stage prediction when the downstream metric is not ranked classification, e.g., code pass rates or exact-match generation.

DataDecide (Magnusson et al., 2025) addresses a complementary question: choosing pretraining data by running small experiments and extrapolating which data mixture will win at larger scales; methodologically, it shares our emphasis on early, low cost decisions, but it operates on the data axis rather than on evaluation predictability.

Complementing these observations, GPT-4 documents predictable scaling on evaluation tasks by fitting a power law to smaller-run models and extrapolating mean log pass rate on HumanEval by using forecasts of $10^3\times$ less compute; they also note that individual items can display non-monotonicities even when aggregate trends are smooth (OpenAI, 2023). Relative to these works, our paper focuses on forecasting evaluation outcomes directly from training FLOPs, rather than inducing possible failure modes that can arise from metric design.

**Reliability, emergence, and scope.** The literature also cautions that downstream scaling is not universally smooth: apparent "emergent abilities" can be artifacts of metric thresholds (Schaeffer et al., 2023) or true structural breaks tied to loss regimes (Du et al., 2024). These effects complicate two-stage fits that assume a single global link from proxy to accuracy (Chen et al., 2024; Bhagia et al., 2024). Our results echo this caution while aiming to be practically useful: we make conservative, metric-aware predictions, and we surface when a metric is likely to induce nonmonotonic or thresholded behavior. We view our contribution as incremental and complementary: rather than proposing a new universal law, we clarify when evaluation is predictable in practice across the metrics researchers actually use, and we provide a lightweight recipe and diagnostics that can be reproduce for a fixed set of pretraining datasets and evaluations.

# 3 Describing the Scaling Behavior of Downstream Metrics

## 3.1 Experimental Setup

We characterize the relationship between training budget and downstream performance by pre-training a comprehensive grid of models. Our experimental setup spans 48 distinct training budgets and five token-to-parameter ratios (10, 20, 40, 80, and 160), allowing us to systematically map the scaling landscape. We start by presenting the details of this setup first, and then exploring the connection between the compute and downstream performance.

**Model architecture and training hyperparameters.** We use a standard modern pre-norm decoder-only Transformer architecture. We employ the RoPE positional embedding (Su et al., 2023), and SwiGLU activation (Shazeer, 2020). We use the tokenizer from (Li et al., 2025a) with vocabulary size of 150k tokens. The sequence length is set to 4096. Further details on the training hyperparameters are provided in Appendix G.

**Dataset.** We train the models on a mixture of data covering general web data, math and code, to be able to track model capabilities in various domains. More specifically, we sample 75% of tokens from DCLM (Li et al., 2025a), 15% from Stack v2 (Lozhkov et al., 2024) (cleaned using the OpenCoder (Huang et al., 2025) heuristics and filtered to remove licenced data), and 10% from OpenMathReasoning dataset (Moshkov et al., 2025).

**Code.** Our training code is based on the open-source AXLearn repository (Lee et al., 2025).

**Evaluation benchmarks.** We consider the following benchmarks, covering various model capabilities: ARC-Easy (Clark et al., 2018), ARC-Challenge (Clark et al., 2018), SqiQ (Welbl et al., 2017), PIQA (Bisk et al., 2019), Hellaswag (Zellers et al., 2019), WebQS (Berant et al., 2013), Winogrande (Sakaguchi et al., 2019), LAMBADA (Paperno et al., 2016), TriviaQA (Joshi et al., 2017), GSM8k (Cobbe et al., 2021), HumanEval (Chen et al., 2021), LBPP (Matton et al., 2024).

## 3.2 Scaling of Accuracy with Training FLOPs

We first analyze models trained at a fixed token-to-parameter ratio (TPR) of 20, which approximates the compute-optimal point suggested by the Chinchilla scaling laws (Hoffmann et al., 2022). For this fixed TPR, we sweep across a range of compute budgets to model the relationship between downstream task performance and training FLOPs. A broader analysis, including overtrained models with different token multipliers, is presented in Section 3.3.

Modeling the scaling of downstream metrics is inherently challenging. This complexity arises because a benchmark's overall accuracy is an aggregation of scaling behaviors on individual examples, making the observed trend highly dependent on the benchmark's specific composition. Given this, we adopt an *empirical* methodology: we investigate whether a single functional form can accurately model performance scaling across a diverse set of popular downstream benchmarks.

Caballero et al. (2023) showed that a wide range of evaluation metrics can be modeled by a smooth approximation of a piecewise linear function in log-log space. We examine the fit of this function, referred to as the Broken Neural Scaling Laws (BNSL). More specifically, we model the downstream accuracy $Q$ based on the training budget $C$ using the following equation:

$$Q = a + b\,C^{-c_0}\left(1 + \left(\frac{C}{d_1}\right)^{1/f_1}\right)^{-c_1 f_1}, \tag{1}$$

which is a BNSL with one transition point. We fit the parameters $a, c_0, d_1, f_1$ based on data.

We further examine the possibility of modeling the curves using a simplified formula with fewer parameters. The first natural idea would be to eliminate the breaking points from Equation 1, resulting in a simple power law. However, such a functional form would be strictly concave as a function of $C$, an assumption not justified by the observed data. For example, ARC-Easy and HellaSwag in Figure 1 are convex in the lower compute range and concave for higher budgets.

Therefore, we consider modeling the *log accuracy* as a power law, a functional form that has the ability of describing the observed $S-$shaped behavior. In the preliminary exploration, we also noticed an approximately linear relationship between log accuracy and training FLOPs on a log-log scale. More precisely, we consider the following functional form:

$$-\log(Q) = \frac{A}{C^{\alpha}}, \tag{2}$$

where $A > 0$ and $\alpha > 0$ are benchmark-specific coefficients fitted on the data.

**Lower asymptote for Accuracy.** Note that Equation 2 indicates that $Q$ takes values from a fixed range, with the lower and upper asymptotes set to zero and one, respectively. This cannot accurately describe the behavior of multiple-choice benchmarks, where even choosing random option results in $Q > 0$. Therefore, for such tasks, we normalize accuracy to take values from the interval $[0, 1]$. More specifically, we first apply the transformation

$$Q' := (Q - Q_{\text{random}})/(1 - Q_{\text{random}}), \tag{3}$$

where $Q_{\text{random}}$ is the random accuracy for a given benchmark and metric type. For multiple choice tasks, we estimate $Q_{\text{random}}$ by training a set of 10 small models in the compute scale below $1 \times 10^{17}$ FLOPs and calculate the average of their performance.

**Fitting and validation.** We empirically confirm the validity of the functional form of Equation 2 by examining the quality of the fit presented in Figure 1. We further evaluate our scaling law by fitting coefficients to experiments with the training budget below $6 \times 10^{21}$ FLOPs, leaving the remaining runs as the validation set. The aggregated absolute and relative errors averaged across the metrics are shown in Table 1.

Table 1: Fit quality of the power law functional form.

| Train MAE | Valid MAE | Train MRE (%) | Valid MRE (%) |
|---|---|---|---|
| 0.0086 | 0.0195 | 0.86% | 1.95% |

### 3.3 Extending Across Token-To-Parameter Ratios

In this section, we extend the previously proposed results, which were dependent solely on training compute, to different token-to-parameter ratios.

Our model for log accuracy scaling (Equation 2) parallels the standard power laws used for log perplexity (Henighan et al., 2020; Hoffmann et al., 2022). A key distinction is the absence of an irreducible term in our model. In perplexity scaling, this term accounts for the inherent entropy of the data, creating a performance floor. For accuracy, however, the theoretical performance ceiling is one. We therefore assume that with an infinite compute budget, a model can achieve perfect accuracy, making an irreducible error term unnecessary.

Building on this analogy, we model the negative log accuracy, as a function of model parameters $N$ and dataset size $D$. We adapt the functional form for pretraining loss from Hoffmann et al. (2022). However, consistent with our earlier reasoning, we exclude the irreducible error term. We fit the coefficients $A$, $\alpha$, $B$, and $\beta$ in the resulting equation:

$$-\log Q = \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}. \tag{4}$$

**Fitting and validation.** We fit the coefficients of Equation 4 by minimizing the Huber loss ($\delta = 1e-3$) with the L-BFGS-B algorithm. To evaluate the model's predictive power, we established a hold-out validation set comprising all runs with either a training budget over $6 \times 10^{21}$ FLOPs or a token-to-parameter ratio (TPR) greater than 80. The model coefficients were then fit exclusively on the remaining training data. Table 2 reports the absolute and relative errors on this validation set, demonstrating a strong predictive fit.

Figure 2 provides a visual confirmation of the model's quality. For the Lambada benchmark, panels (a)-(c) show the close alignment between observed data and the fitted curves across different TPRs. Panels (d)-(f) further illustrate the model's effectiveness on three additional benchmarks.

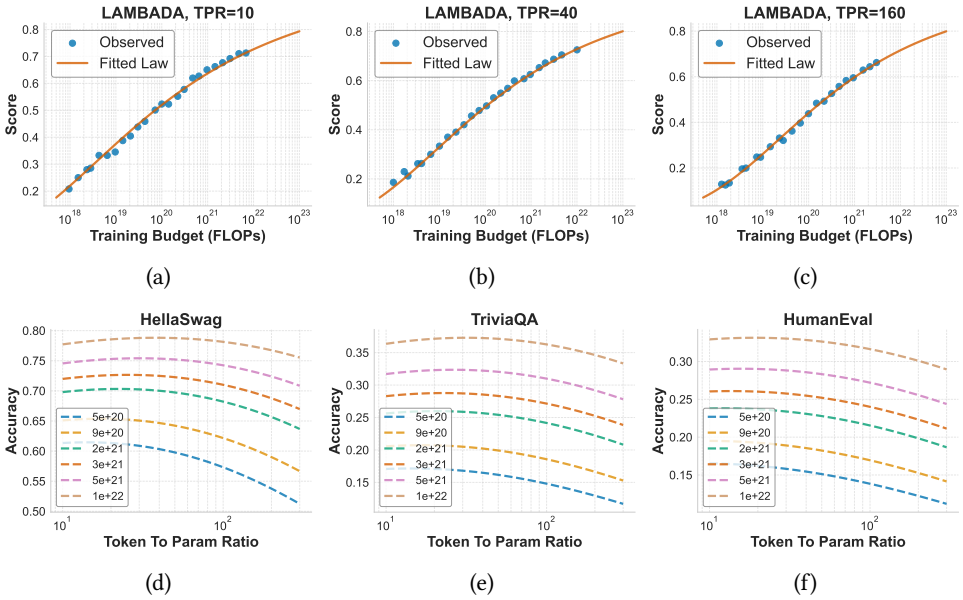



Figure 2: Comparison of scaling the downstream accuracy on different token-to-parameter ratios.

Table 2: Fit quality of Equation 4.

| Train MAE | Valid MAE | Train MRE (%) | Valid MRE (%) |
|---|---|---|---|
| 0.0103 | 0.0191 | 1.03% | 1.91% |

### 3.4 Considering the Effect of Repeated Sampling with Inference Compute

We also study the effect of increasing the number of samples in pass@k for coding benchmarks across different training budgets.

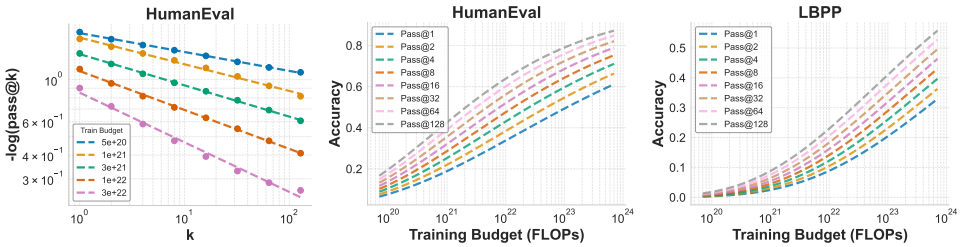



Figure 3: Comparison of pass@k behaviour across tasks. (a) Intuition for the functional form. (b) Predicted pass rate curves for HumanEval. (c) Predicted pass rate curves for LBPP.

Figure 3 plots the negative log pass rate against k for the HumanEval benchmark. The plot highlights two primary observations:

1. For a fixed training budget, the relationship is approximately linear in log scale, indicating that the rate follows a power law with respect to $k$. This observation is consistent with findings from related work (Brown et al., 2024; Hughes et al., 2024; Kwok et al., 2025), observing power law scaling with the number of attempts $k$ for coding, math, and adversarial robustness.
2. The slope of this linear relationship depends on the training compute budget, becoming steeper for larger compute. Concurrently to this work, similar finding is presented in (Schaeffer et al., 2025b).

Connecting these observations with scaling law in Equation 2 for $k = 1$, we propose the following equation for modeling the pass@k rate $Q$ based on the training budget $C$ and number of trials $k$:

$$\log(-\log Q(C,k)) = \log A + \alpha \log C + \beta \log k + \delta \log C \log k \tag{5}$$

**Fitting and validation.** We fit the scaling law in Equation 5 to HumanEval pass@k data (see Figure 3 for visualizations). For validation, we established a hold-out set, fitting the model's coefficients on experiments with FLOPs below $6 \times 10^{21}$ and $k \leq 32$. The model's absolute and relative errors on the remaining validation data, reported in Table 3, confirm its predictive accuracy.

Table 3: Fit quality of Equation 5.

| Train MAE | Valid MAE | Train MRE (%) | Valid MRE (%) |
|---|---|---|---|
| 0.0111 | 0.0284 | 6.30% | 7.94% |

**Theoretical justification.** In Appendix C we derive analytical upper and lower bounds for the pass rate given $k$ independent trials. These results can serve as complementary for deriving the functional form of Equation 5. Theoretical analysis of the mechanisms behind power law scaling of repeated sampling is also explored in related work (Schaeffer et al., 2025a; Levi, 2024; Kazdan et al., 2025).

## 4 Predicting Accuracy of Large Models

### 4.1 Directly Predicting Accuracy from Training Compute

We leverage the power-law scaling law from Section 3.2 to extrapolate model performance from smaller to larger compute budgets. To test the aproach, we fit the scaling law coefficients using only experiments trained with $3 \times 10^{18}$ to $6 \times 10^{21}$ FLOPs. When fitting coefficients in Equation 2, we only use runs with accuracy at least 5% points above the random performance, as we notice that these small accuracy results tend to have larger variance disturbing the fit. We then use this model to predict the accuracy for larger, held-out models, whose training budget is up to 6.7x larger. For comparison, we also fit the coefficients in BNSL (Equation (1)) and assess the prediction accuracy on the same validation set of experiments. The results are presented in Table 4. We notice that while both functional forms achieve a good extrapolation quality (below 3% points MAE on average), our proposed power law scaling achieves a slightly lower validation error.

Table 4: MRE and MAE in downstream accuracy prediction for BNSL and Power Law scaling strategies.

| | | Power Law | | BNSL | |
|---|---|---|---|---|---|
| Benchmark | Metric Type | MAE | MRE (%) | MAE | MRE (%) |
| ARC-E | acc_norm | 0.0186 | 2.37 | 0.0555 | 7.00 |
| ARC-C | acc_norm | 0.0068 | 1.30 | 0.0155 | 2.86 |
| SciQ | acc_norm | 0.0051 | 0.55 | 0.0045 | 0.49 |
| PIQA | acc_norm | 0.0151 | 1.86 | 0.0087 | 1.06 |
| HellaSwag | acc_norm | 0.0298 | 3.74 | 0.0173 | 2.15 |
| Winogrande | acc | 0.0119 | 1.64 | 0.0363 | 4.93 |
| WebQS | exact_match | 0.0166 | 7.28 | 0.0033 | 1.46 |
| TriviaQA | exact_match | 0.0259 | 6.17 | 0.0424 | 10.13 |
| LAMBADA | acc | 0.0258 | 3.45 | 0.0043 | 0.58 |
| GSM8K | exact_match | 0.0559 | 9.67 | 0.0807 | 13.90 |
| HumanEval | pass@1 | 0.0171 | 5.11 | 0.0293 | 8.29 |
| LBPP | pass@1 | 0.0154 | 13.43 | 0.0066 | 6.70 |
| Average | | 0.0203 | 4.72 | 0.0254 | 4.96 |

### 4.2 Predicting Downstream Accuracy in a Two-Stage Approach

Previous works (Bhagia et al., 2025; Chen et al., 2025) use a *two-stage* approach to predict downstream task performance based on model characteristics such as FLOPs, size, and training data. The intuition behind this approach is that the model's loss acts as a proxy for predicting downstream task accuracy. However, we argue that while this approach provides useful interpretability in some cases, the multi-stage nature of this method compounds errors from each stage, ultimately resulting in scaling laws with higher variance and reduced predictive accuracy.

To demonstrate this, we explore the correlation between various proxy metrics, such as log-probability, evaluation loss, Brier Score, and final accuracy on downstream benchmarks. We then implement the two-stage approach on our dataset and compare its performance against our proposed direct scaling law, which directly predicts downstream accuracy from FLOPs.

#### 4.2.1 Correlation Between Proxy Metrics and Downstream Accuracy

The two-stage framework relies on the critical assumption that an intermediate proxy metric, such as a model's training loss, strongly correlates with its final downstream accuracy. In this section, we analyze the correlation between downstream task accuracy and several candidate proxy metrics including log-probability, evaluation loss, and the Brier Score. A strong and consistent correlation is a prerequisite for the two-stage method to be effective, as any poor choice in this relationship will propagate errors and reduce the accuracy of the final scaling predictions.

To capture the potentially non-linear dependency between proxy metrics and downstream task performance, we fit logistic functions of the form $\text{Acc} = 1/(1 + e^{-a \cdot \text{proxy} + b})$ and compare the goodness of fit across metrics. We evaluate fit quality using Root Mean Squared Error (RMSE) and the coefficient of determination ($R^2$). Note that the choice of proxy metric does not affect the first stage of the two-stage approach, as Brandfonbrener et al. (2024) demonstrated that all metrics exhibit similar scaling behavior with respect to model size and compute.

Our analysis shows that most proxy metrics demonstrate strong predictive power for downstream task performance, with goodness-of-fit statistics falling within a similar range across all metrics ($R^2 > 0.95$). This suggests that practitioners can select proxy metrics based on computational convenience or availability rather than predictive accuracy. Figure 4 shows a representative example for the `arc_easy` benchmark; we observe comparable correlation strength across other benchmarks as well (see Appendix D for comprehensive results and detailed fit statistics for each proxy metric). These findings validate the assumption underlying the two-stage framework and demonstrate that the choice of proxy metric is unlikely to be a limiting factor in prediction accuracy.

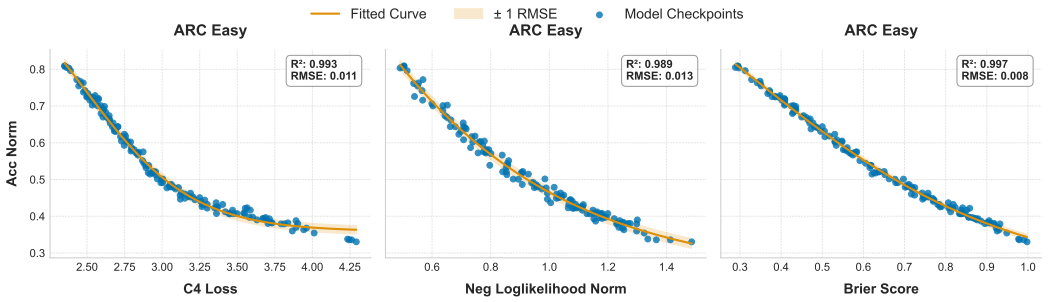



Figure 4: Dependency between downstream task and proxy metric candidates. All metrics demonstrate strong prediction power, i.e. high $R^2$ and low RMSE.

### 4.3 Two-Stage Approach is not as Strong as Direct Approach

#### 4.3.1 Two-Stage Approach is not as Strong as Direct Approach

Following prior work Bhagia et al. (2025); Chen et al. (2025), we evaluate two-stage scaling laws that use negative log-likelihood (NLL) as a proxy metric to predict downstream task performance. This approach is based on the hypothesis that language modeling capability correlates with task-

specific accuracy. As established in the previous section, our analysis showed that the choice of proxy metric does not significantly impact the final prediction; therefore, we adopt NLL to maintain consistency with these baselines.

The two-stage methods map the proxy metric, $L$, to accuracy using two transition formulas: (1) a *linear* transition, Acc = $a + bL$, and (2) a *logistic* transition, Acc = $[a/(1 + e^{-k(L-L_0)})] + b$ , where $a$, $b$, $k$, and $L_0$ are fitted parameters. We implement both formulas and compare them against our direct scaling approach. For all models, we fit the scaling laws on data up to a compute budget of $6 \times 10^{21}$ FLOPs and validate on models trained with greater compute.

To evaluate predictive performance, we use Mean Relative Error (MRE) and Mean Absolute Error (MAE). To assess goodness-of-fit on the training data, we use RMSE and $R^2$. These metrics, averaged over all benchmarks, are reported in Table 5. An illustrative comparison of the fits is shown in Figure 5, with additional examples in Section D, Figure 7.

Our findings indicate that the direct approaches (BNSL and the simple scaling law) consistently outperform the two-stage methods in prediction. This holds true even though the two-stage models often exhibit a superior goodness-of-fit (higher $R^2$ and lower RMSE) on the training data. For example, Figure 5 shows the two-stage logistic model achieving a better fit than BNSL, yet failing to extrapolate accurately. We attribute this discrepancy to the *compounding of errors*: inaccuracies in the first stage (FLOPs-to-NLL) are amplified by the second stage (NLL-to-Accuracy), leading to poor overall predictive power compare to the direct approach.

Moreover, we conduct an additional analysis to examine the effect of the FLOPs threshold on the quality of scaling fits. This can be interpreted as a form of sensitivity analysis: we ask what minimum FLOPs threshold is required for each benchmark and scaling law to achieve a reliable fit with an error below a given rate. The results are reported in Section E. In summary, we find that BNSL and the simple power law exhibit relatively stable behaviour across a wide range of FLOPs, whereas the two-stage linear and logistic models are less robust and often yield higher error rates.

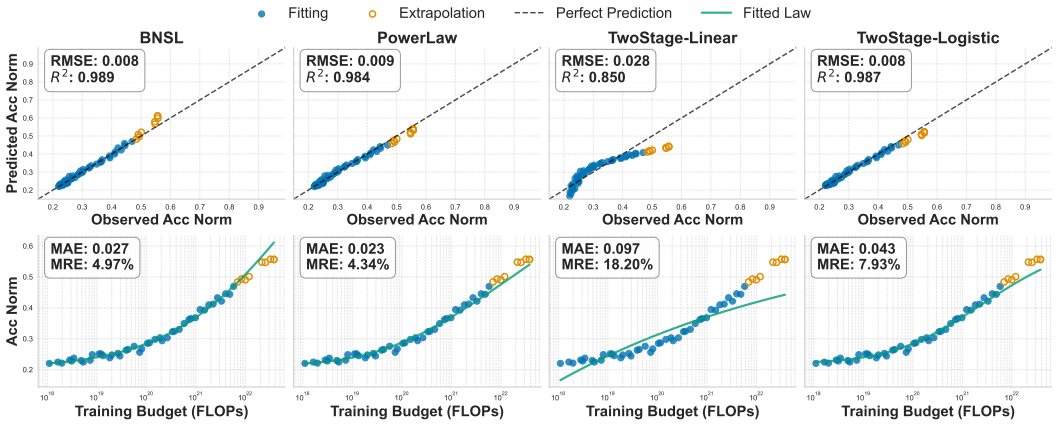


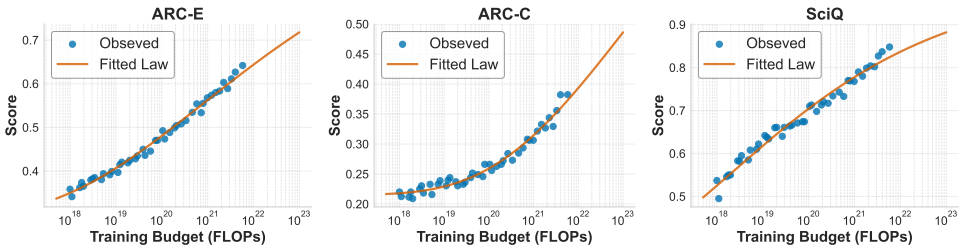

Figure 5: Scaling law fits. Comparing the direct approaches (BNSL, simple power law) from Section 3.2 with two-stage approaches (Linear and Logistic) for ARC Challenge.

Table 5: Performance comparison of scaling law strategies across different error metrics.

| | MRE (%) | | MAE | | RMSE | | $R^2$ | |
|---|---|---|---|---|---|---|---|---|
| Scaling Law Strategy | mean | std | mean | std | mean | std | mean | std |
| PowerLaw | 1.963 | 1.201 | 0.015 | 0.010 | 0.011 | 0.006 | 0.986 | 0.011 |
| BNSL | 2.713 | 2.569 | 0.020 | 0.020 | 0.007 | 0.003 | 0.993 | 0.004 |
| TwoStage-Linear | 6.667 | 6.958 | 0.044 | 0.034 | 0.023 | 0.011 | 0.943 | 0.054 |
| TwoStage-Logistic | 6.351 | 3.278 | 0.047 | 0.024 | 0.017 | 0.006 | 0.974 | 0.013 |

# 5 Discussion

## 5.1 Effect of the Data Mixture




(a) ARC-Easy. (b) ARC-Challenge. (c) SciQ.

Figure 6: Example downstream scaling curves when changing the pre-training dataset to C4. All metrics are shown in Appendix F.

A natural question to consider is whether the previously presented scaling curves are specific to one particular data mixture. To address this concern, we train a suite of models with the same setup as described in Section 3, but change the dataset to C4 (Raffel et al., 2023). We fix the token-to-parameter ratio to 20 for these experiments and consider 44 training budgets between $1 \times 10^{18}$ and $6 \times 10^{21}$ FLOPs. We note random performance of these models on code and math specific downstream tasks, due to the lack of alignment of the dataset towards these domains. Overall, we observe accuracy of more than 10 percentage points above random chance on 8 downstream tasks. We fit coefficients in Equation (2) to each of them and present example plots in Figure 10. We observe good fit quality, similar to observed in Section 3, indicating that the validity of the presented scaling trends are not restricted to one particular data mixture.

# 6 Conclusions

In this work, we demonstrate that downstream benchmark accuracy scales predictably with training compute. We introduce a simple, direct scaling law that accurately models the relationship between training FLOPs and final benchmark performance. By establishing this predictable scaling behavior, our work helps make the development of large-scale models more systematic and efficient, presenting the scaling of downstream capabilities as a direct and measurable consequence of scale.

# 7 Reproducibility

We aim to support reproducibility and encourage future research based on the results of this work. We detail training hyperparameters and model training details in Section 3.1 and Appendix G. All of the models were trained using a codebase based on the open-source AXLearn repository, using publicly available datasets. We release complete set of pretraining losses and evaluation results of the models used in this study, and scaling law fitting code in a repository available under the link: https://github.com/apple/ml-scaling-downstream-metrics.

## References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1533–1544, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1160.

Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Power lines: Scaling laws for weight decay and batch size in llm pre-training, 2025. URL https://arxiv.org/abs/2505.13738.

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. *CoRR*, abs/2412.04403, 2024. doi: 10.48550/ARXIV.2412.04403. URL https://doi.org/10.48550/arXiv.2412.04403.

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, 2025. URL https://arxiv.org/abs/2412.04403.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019. URL https://arxiv.org/abs/1911.11641.

David Brandfonbrener, Nikhil Anand, Nikhil Vyas, Eran Malach, and Sham M. Kakade. Loss-to-loss prediction: Scaling laws for all datasets. *CoRR*, abs/2411.12925, 2024. doi: 10.48550/ARXIV.2411.12925. URL https://doi.org/10.48550/arXiv.2411.12925.

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

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws, 2023. URL https://arxiv.org/abs/2210.14891.

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

Yangyi Chen, Binxuan Huang, Yifan Gao, Zhengyang Wang, Jingfeng Yang, and Heng Ji. Scaling laws for predicting downstream performance in llms. *CoRR*, abs/2410.08527, 2024. doi: 10.48550/ARXIV.2410.08527. URL https://doi.org/10.48550/arXiv.2410.08527.

Yangyi Chen, Binxuan Huang, Yifan Gao, Zhengyang Wang, Jingfeng Yang, and Heng Ji. Scaling laws for predicting downstream performance in llms, 2025. URL https://arxiv.org/abs/2410.08527.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

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

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. Understanding emergent abilities of language models from the loss perspective. *CoRR*, abs/2403.15796, 2024. doi: 10.48550/ARXIV.2403.15796. URL https://doi.org/10.48550/arXiv.2403.15796.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, 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, 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, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, 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, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Oleg Filatov, Jan Ebert, Jiangtao Wang, and Stefan Kesselheim. Time transfer: On optimal learning rate and batch size in the infinite data limit, 2025. URL https://arxiv.org/abs/2410.05838.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Jenia Jitsev, Alexandros G. Dimakis, Gabriel Ilharco, 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. *CoRR*, abs/2403.08540, 2024a. doi: 10.48550/ARXIV.2403.08540. URL https://doi.org/10.48550/arXiv.2403.08540.

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

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.

Tom Gunter, Zirui Wang, Chong Wang, Ruoming Pang, Andy Narayanan, Aonan Zhang, Bowen Zhang, Chen Chen, Chung-Cheng Chiu, David Qiu, Deepak Gopinath, Dian Ang Yap, Dong Yin, Feng Nan, Floris Weers, Guoli Yin, Haoshuo Huang, Jianyu Wang, Jiarui Lu, John Peebles, Ke Ye, Mark Lee, Nan Du, Qibin Chen, Quentin Keunebroek, Sam Wiseman, Syd Evans, Tao Lei, Vivek Rathod, Xiang Kong, Xianzhi Du, Yanghao Li, Yongqiang Wang, Yuan Gao, Zaid Ahmed, Zhaoyang Xu, Zhiyun Lu, Al Rashid, Albin Madappally Jose, Alec Doane, Alfredo Bencomo, Allison Vanderby, Andrew Hansen, Ankur Jain, Anupama Mann Anupama, Areeba Kamal, Bugu Wu, Carolina Brum, Charlie Maalouf, Chinguun Erdenebileg, Chris Dulhanty, Dominik Moritz, Doug Kang, Eduardo Jimenez, Evan Ladd, Fangping Shi, Felix Bai, Frank Chu,

Fred Hohman, Hadas Kotek, Hannah Gillis Coleman, Jane Li, Jeffrey Bigham, Jeffery Cao, Jeff Lai, Jessica Cheung, Jiulong Shan, Joe Zhou, John Li, Jun Qin, Karanjeet Singh, Karla Vega, Kelvin Zou, Laura Heckman, Lauren Gardiner, Margit Bowler, Maria Cordell, Meng Cao, Nicole Hay, Nilesh Shahdadpuri, Otto Godwin, Pranay Dighe, Pushyami Rachapudi, Ramsey Tantawi, Roman Frigg, Sam Davarnia, Sanskruti Shah, Saptarshi Guha, Sasha Sirovica, Shen Ma, Shuang Ma, Simon Wang, Sulgi Kim, Suma Jayaram, Vaishaal Shankar, Varsha Paidi, Vivek Kumar, Xin Wang, Xin Zheng, Walker Cheng, Yael Shrager, Yang Ye, Yasu Tanaka, Yihao Guo, Yunsong Meng, Zhao Tang Luo, Zhi Ouyang, Alp Aygar, Alvin Wan, Andrew Walkingshaw, Andy Narayanan, Antonie Lin, Arsalan Farooq, Brent Ramerth, Colorado Reed, Chris Bartels, Chris Chaney, David Riazati, Eric Liang Yang, Erin Feldman, Gabriel Hochstrasser, Guillaume Seguin, Irina Belousova, Joris Pelemans, Karen Yang, Keivan Alizadeh Vahid, Liangliang Cao, Mahyar Najibi, Marco Zuliani, Max Horton, Minsik Cho, Nikhil Bhendawade, Patrick Dong, Piotr Maj, Pulkit Agrawal, Qi Shan, Qichen Fu, Regan Poston, Sam Xu, Shuangning Liu, Sushma Rao, Tashweena Heeramun, Thomas Merth, Uday Rayala, Victor Cui, Vivek Rangarajan Sridhar, Wencong Zhang, Wenqi Zhang, Wentao Wu, Xingyu Zhou, Xinwen Liu, Yang Zhao, Yin Xia, Zhile Ren, and Zhongzheng Ren. Apple intelligence foundation language models, 2024. URL https://arxiv.org/abs/2407.21075.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling, 2020. URL https://arxiv.org/abs/2010.14701.

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

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

Siming Huang, Tianhao Cheng, J. K. Liu, Jiaran Hao, Liuyihan Song, Yang Xu, J. Yang, Jiaheng Liu, Chenchen Zhang, Linzheng Chai, Ruifeng Yuan, Zhaoxiang Zhang, Jie Fu, Qian Liu, Ge Zhang, Zili Wang, Yuan Qi, Yinghui Xu, and Wei Chu. Opencoder: The open cookbook for top-tier code large language models, 2025. URL https://arxiv.org/abs/2411.04905.

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

Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Koyejo. Scaling laws for downstream task performance in machine translation, 2025. URL https://arxiv.org/abs/2402.04177.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, 2017. URL https://arxiv.org/abs/1705.03551.

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, 2020. URL https://arxiv.org/abs/2001.08361.

Joshua Kazdan, Rylan Schaeffer, Youssef Allouah, Colin Sullivan, Kyssen Yu, Noam Levi, and Sanmi Koyejo. Efficient prediction of pass@k scaling in large language models, 2025. URL https://arxiv.org/abs/2510.05197.

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

Mark Lee, Tom Gunter, Chang Lan, John Peebles, Hanzhi Zhou, Kelvin Zou, Sneha Bangalore, Chung-Cheng Chiu, Nan Du, Xianzhi Du, Philipp Dufter, Ruixuan Hou, Haoshuo Huang, Dongseong Hwang, Xiang Kong, Jinhao Lei, Tao Lei, Meng Li, Li Li, Jiarui Lu, Zhiyun Lu, Yiping Ma, David Qiu, Vivek Rathod, Senyu Tong, Zhucheng Tu, Jianyu Wang, Yongqiang Wang, Zirui Wang, Floris Weers, Sam Wiseman, Guoli Yin, Bowen Zhang, Xiyou Zhou, Danyang Zhuo, Cheng Leong, and Ruoming Pang. Axlearn: Modular large model training on heterogeneous infrastructure, 2025. URL https://arxiv.org/abs/2507.05411.

Noam Levi. A simple model of inference scaling laws, 2024. URL https://arxiv.org/abs/2410.16377.

Ethan Li, Anders Boesen Lindbo Larsen, Chen Zhang, Xiyou Zhou, Jun Qin, Dian Ang Yap, Narendran Raghavan, Xuankai Chang, Margit Bowler, Eray Yildiz, John Peebles, Hannah Gillis Coleman, Matteo Ronchi, Peter Gray, Keen You, Anthony Spalvieri-Kruse, Ruoming Pang, Reed Li, Yuli Yang, Emad Soroush, Zhiyun Lu, Crystal Xiao, Rong Situ, Jordan Huffaker, David Griffiths, Zaid Ahmed, Peng Zhang, Daniel Parilla, Asaf Liberman, Jennifer Mallalieu, Parsa Mazaheri, Qibin Chen, Manjot Bilkhu, Aonan Zhang, Eric Wang, Dave Nelson, Michael FitzMaurice, Thomas Voice, Jeremy Liu, Josh Shaffer, Shiwen Zhao, Prasanth Yadla, Farzin Rasteh, Pengsheng Guo, Arsalan Farooq, Jeremy Snow, Stephen Murphy, Tao Lei, Minsik Cho, George Horrell, Sam Dodge, Lindsay Hislop, Sumeet Singh, Alex Dombrowski, Aiswarya Raghavan, Sasha Sirovica, Mandana Saebi, Faye Lao, Max Lam, TJ Lu, Zhaoyang Xu, Karanjeet Singh, Marc Kirchner, David Mizrahi, Rajat Arora, Haotian Zhang, Henry Mason, Lawrence Zhou, Yi Hua, Ankur Jain, Felix Bai, Joseph Astrauskas, Floris Weers, Josh Gardner, Mira Chiang, Yi Zhang, Pulkit Agrawal, Tony Sun, Quentin Keunebroek, Matthew Hopkins, Bugu Wu, Tao Jia, Chen Chen, Xingyu Zhou, Nanzhu Wang, Peng Liu, Ruixuan Hou, Rene Rauch, Yuan Gao, Afshin Dehghan, Jonathan Janke, Zirui Wang, Cha Chen, Xiaoyi Ren, Feng Nan, Josh Elman, Dong Yin, Yusuf Goren, Jeff Lai, Yiran Fei, Syd Evans, Muyang Yu, Guoli Yin, Yi Qin, Erin Feldman, Isha Garg, Aparna Rajamani, Karla Vega, Walker Cheng, TJ Collins, Hans Han, Raul Rea Menacho, Simon Yeung, Sophy Lee, Phani Mutyala, Ying-Chang Cheng, Zhe Gan, Sprite Chu, Justin Lazarow, Alessandro Pappalardo, Federico Scozzafava, Jing Lu, Erik Daxberger, Laurent Duchesne, Jen Liu, David Güera, Stefano Ligas, Mary Beth Kery, Brent Ramerth, Ciro Sannino, Marcin Eichner, Haoshuo Huang, Rui Qian, Moritz Schwarzer-Becker, David Riazati, Mingfei Gao, Bailin Wang, Jack Cackler, Yang Lu, Ransen Niu, John Dennison, Guillaume Klein, Jeffrey Bigham, Deepak Gopinath, Navid Shiee, Darren Botten, Guillaume Tartavel, Alex Guillen Garcia, Sam Xu, Victoria MönchJuan Haladjian, Zi-Yi Dou, Matthias Paulik, Adolfo Lopez Mendez, Zhen Li, Hong-You Chen, Chao Jia, Dhaval Doshi, Zhengdong Zhang, Raunak Manjani, Aaron Franklin, Zhile Ren, David Chen, Artsiom Peshko, Nandhitha Raghuram, Hans Hao, Jiulong Shan, Kavya Nerella, Ramsey Tantawi, Vivek Kumar, Saiwen Wang, Brycen Wershing, Bhuwan Dhingra, Dhruti Shah, Ob Adaranijo, Xin Zheng, Tait Madsen, Hadas Kotek, Chang Liu, Yin Xia, Hanli Li, Suma Jayaram, Yanchao Sun, Ahmed Fakhry, Vasileios Saveris, Dustin Withers, Yanghao Li, Alp Aygar, Andres Romero Mier Y Teran, Kaiwei Huang, Mark Lee, Xiujun Li, Yuhong Li, Tyler Johnson, Jay Tang, Joseph Yitan Cheng, Futang Peng, Andrew Walkingshaw, Lucas Guibert, Abhishek Sharma, Cheng Shen, Piotr Maj, Yasutaka Tanaka, You-Cyuan Jhang, Vivian Ma, Tommi Vehvilainen, Kelvin Zou, Jeff Nichols, Matthew Lei, David Qiu, Yihao Qian, Gokul Santhanam, Wentao Wu, Yena Han, Dominik Moritz, Haijing Fu, Mingze Xu, Vivek Rathod, Jian Liu, Louis D'hauwe, Qin Ba, Haitian Sun, Haoran Yan, Philipp Dufter, Anh Nguyen, Yihao Feng, Emma Wang, Keyu He, Rahul Nair, Sanskruti Shah, Jiarui Lu, Patrick Sonnenberg, Jeremy Warner, Yuanzhi Li, Bowen Pan, Ziyi Zhong, Joe Zhou, Sam Davarnia, Olli Saarikivi, Irina Belousova, Rachel Burger, Shang-Chen Wu, Di Feng, Bas Straathof, James Chou, Yuanyang Zhang, Marco Zuliani, Eduardo Jimenez, Abhishek Sundararajan, Xianzhi Du, Chang Lan, Nilesh Shahdadpuri, Peter Grasch, Sergiu Sima, Josh Newnham, Varsha Paidi, Jianyu Wang, Kaelen Haag, Alex Braunstein, Daniele Molinari, Richard Wei, Brenda Yang, Nicholas Lusskin, Joanna Arreaza-Taylor, Meng Cao, Nicholas Seidl, Simon Wang, Jiaming Hu, Yiping Ma, Mengyu Li, Kieran Liu, Hang Su, Sachin Ravi, Chong Wang, Xin Wang, Kevin Smith, Haoxuan You, Binazir Karimzadeh, Rui Li, Jinhao Lei, Wei Fang, Alec Doane, Sam Wiseman, Ismael Fernandez, Jane Li, Andrew Hansen, Javier Movellan, Christopher Neubauer, Hanzhi Zhou, Chris Chaney, Nazir Kamaldin, Valentin Wolf, Fernando Bermúdez-Medina, Joris Pelemans, Peter Fu, Howard Xing, Xiang Kong, Wayne Shan, Gabriel Jacoby-Cooper, Dongcai Shen, Tom Gunter, Guillaume Seguin, Fangping Shi, Shiyu Li, Yang Xu, Areeba Kamal, Dan Masi, Saptarshi Guha, Qi Zhu, Jenna Thibodeau, Changyuan Zhang, Rebecca Callahan, Charles Maalouf, Wilson Tsao, Boyue Li, Qingqing Cao,

Naomy Sabo, Cheng Leong, Yi Wang, Anupama Mann Anupama, Colorado Reed, Kenneth Jung, Zhifeng Chen, Mohana Prasad Sathya Moorthy, Yifei He, Erik Hornberger, Devi Krishna, Senyu Tong, Michael, Lee, David Haldimann, Yang Zhao, Bowen Zhang, Chang Gao, Chris Bartels, Sushma Rao, Nathalie Tran, Simon Lehnerer, Co Giang, Patrick Dong, Junting Pan, Biyao Wang, Dongxu Li, Mehrdad Farajtabar, Dongseong Hwang, Grace Duanmu, Eshan Verma, Sujeeth Reddy, Qi Shan, Hongbin Gao, Nan Du, Pragnya Sridhar, Forrest Huang, Yingbo Wang, Nikhil Bhendawade, Diane Zhu, Sai Aitharaju, Fred Hohman, Lauren Gardiner, Chung-Cheng Chiu, Yinfei Yang, Alper Kokmen, Frank Chu, Ke Ye, Kaan Elgin, Oron Levy, John Park, Donald Zhang, Eldon Schoop, Nina Wenzel, Michael Booker, Hyunjik Kim, Chinguun Erdenebileg, Nan Dun, Eric Liang Yang, Priyal Chhatrapati, Vishaal Mahtani, Haiming Gang, Kohen Chia, Deepa Seshadri, Donghan Yu, Yan Meng, Kelsey Peterson, Zhen Yang, Yongqiang Wang, Carina Peng, Doug Kang, Anuva Agarwal, Albert Antony, Juan Lao Tebar, Albin Madappally Jose, Regan Poston, Andy De Wang, Gerard Casamayor, Elmira Amirloo, Violet Yao, Wojciech Kryscinski, Kun Duan, and Lezhi L. Apple intelligence foundation language models: Tech report 2025, 2025a. URL https://arxiv.org/abs/2507.13575.

Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. (mis)fitting: A survey of scaling laws, 2025b. URL https://arxiv.org/abs/2502.18969.

Nicholas Lourie, Michael Y. Hu, and Kyunghyun Cho. Scaling laws are unreliable for downstream tasks: A reality check, 2025. URL https://arxiv.org/abs/2507.00885.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, Zhuang Li, Wen-Ding Li, Megan Risdal, Jia Li, Jian Zhu, Terry Yue Zhuo, Evgenii Zheltonozhskii, Nii Osae Osae Dade, Wenhao Yu, Lucas Krauß, Naman Jain, Yixuan Su, Xuanli He, Manan Dey, Edoardo Abati, Yekun Chai, Niklas Muennighoff, Xiangru Tang, Muhtasham Oblokulov, Christopher Akiki, Marc Marone, Chenghao Mou, Mayank Mishra, Alex Gu, Binyuan Hui, Tri Dao, Armel Zebaze, Olivier Dehaene, Nicolas Patry, Canwen Xu, Julian McAuley, Han Hu, Torsten Scholak, Sebastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, Mostofa Patwary, Nima Tajbakhsh, Yacine Jernite, Carlos Muñoz Ferrandis, Lingming Zhang, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder 2 and the stack v2: The next generation, 2024. URL https://arxiv.org/abs/2402.19173.

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. *CoRR*, abs/2504.11393, 2025. doi: 10.48550/ARXIV.2504.11393. URL https://doi.org/10.48550/arXiv.2504.11393.

Alexandre Matton, Tom Sherborne, Dennis Aumiller, Elena Tommasone, Milad Alizadeh, Jingyi He, Raymond Ma, Maxime Voisin, Ellen Gilsenan-McMahon, and Matthias Gallé. On leakage of code generation evaluation datasets, 2024. URL https://arxiv.org/abs/2407.07565.

Prasanna Mayilvahanan, Thaddäus Wiedemer, Sayak Mallick, Matthias Bethge, and Wieland Brendel. Llms on the line: Data determines loss-to-loss scaling laws, 2025. URL https://arxiv.org/abs/2502.12120.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset, 2025. URL https://arxiv.org/abs/2504.16891.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

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.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context, 2016. URL https://arxiv.org/abs/1606.06031.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019. URL https://arxiv.org/abs/1907.10641.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *CoRR*, abs/2304.15004, 2023. doi: 10.48550/ARXIV.2304.15004. URL https://doi.org/10.48550/arXiv.2304.15004.

Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get their power (laws)?, 2025a. URL https://arxiv.org/abs/2502.17578.

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?, 2025b. URL https://arxiv.org/abs/2406.04391.

Noam Shazeer. Glu variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg

Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen

Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu

Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Ko-

ray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. Do large language model benchmarks test reliability?, 2025. URL https://arxiv.org/abs/2502.03461.

David J. Wales and Jonathan P. K. Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116, July 1997. ISSN 1520-5215. doi: 10.1021/jp970984n. URL http://dx.doi.org/10.1021/jp970984n.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions, 2017. URL https://arxiv.org/abs/1707.06209.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities, 2023. URL https://arxiv.org/abs/2309.14322.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer, 2022. URL https://arxiv.org/abs/2203.03466.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019. URL https://arxiv.org/abs/1905.07830.

Hanlin Zhang, Depen Morwani, Nikhil Vyas, Jingfeng Wu, Difan Zou, Udaya Ghai, Dean Foster, and Sham Kakade. How does critical batch size scale in pre-training?, 2025. URL https://arxiv.org/abs/2410.21676.

# Appendices

## A Limitations and Future Work

Our claims are empirical and conditional. We model downstream accuracy as a simple function of training compute (Equation 2) and extend to parameters–tokens (Equation 4) and pass@k (Equation 5), but we do not yet offer a mechanistic account of these forms; connecting them to item-difficulty mixtures or error-decay processes remains open.

- **Metric Dependence.** Aggregate trends are smooth on the benchmarks we study, but composite or thresholded tasks can look step-like or non-monotone (e.g., BIG-bench–style mixes). A principled "metric audit" that partitions by latent difficulty would clarify when predictability should be expected.
- **Data Mixture.** The composition of the pretraining data significantly affects scaling behavior. For example, a C4-only model preserves scaling on general QA while its performance on code and math reverts toward chance (Figure 10). Scaling claims should therefore be understood as conditional on the specific data mixture and filtering pipeline used.
- **Scale Thresholds.** Reliable extrapolation appears only after task-specific FLOPs thresholds; below them, fits can be brittle (Figures 8 and 9). Future protocols should declare the threshold used and report success probabilities vs. that threshold.
- **Training Recipe Scope.** Results are from decoder-only Transformers with a fixed modern recipe; we did not test MoE, retrieval, or multimodal models, nor alternative optimizers/schedules. Exponents may be recipe-dependent.
- **Beyond Pretraining.** We fit laws on pretraining checkpoints; continued pretraining, instruction tuning, and preference optimization can reshape downstream accuracy and may require extended forms or new covariates.
- **Uncertainty and compute trade-offs** We emphasize mean errors but not calibrated prediction intervals; bootstrap-based intervals and floor estimation (Equation 3) would aid decision-making. For code, train vs. inference compute (pass@k) trade-offs deserve explicit Pareto analyses.

In summary, our framework helps reconcile predictable and unpredictable scaling phenomena. Predictability emerges beyond a task and metric specific scale threshold and under a fixed data mixture. Making these preconditions explicit should improve reproducibility across studies.

## B Experiment List

Below we present a full list of models trained as a part of this study. In the *datasets* column, DCLM+code+math indicates the mixture described in Section 3. A subset of models was also trained on the C4 dataset, as part of the ablation in Section 5.1.

Table 6: List of experiments (Part 1/3).

| params | flops | tokens | layers | hidden dim | heads | datasets |
|---|---|---|---|---|---|---|
| 17.58B | 3.77e+22 | 357.4B | 44 | 5632 | 44 | DCLM+code+math |
| 17.58B | 3.77e+22 | 357.4B | 44 | 5632 | 44 | DCLM+code+math |
| 17.58B | 3.77e+22 | 357.4B | 44 | 5632 | 44 | DCLM+code+math |
| 16.51B | 3.22e+22 | 324.7B | 43 | 5504 | 43 | DCLM+code+math |
| 14.33B | 2.54e+22 | 295.0B | 41 | 5248 | 41 | DCLM+code+math |
| 13.40B | 2.15e+22 | 268.0B | 40 | 5120 | 40 | DCLM+code+math |
| 10.71B | 6.85e+21 | 106.6B | 37 | 4736 | 37 | DCLM+code+math |
| 9.88B | 1.19e+22 | 201.0B | 36 | 4608 | 36 | DCLM+code+math |
| 9.10B | 9.97e+21 | 182.6B | 35 | 4480 | 35 | DCLM+code+math |
| 9.10B | 4.80e+21 | 88.0B | 35 | 4480 | 35 | DCLM+code+math |
| 8.41B | 8.37e+21 | 165.9B | 34 | 4352 | 34 | DCLM+code+math |
| 7.71B | 6.98e+21 | 150.7B | 33 | 4224 | 33 | DCLM+code+math |
| 7.05B | 3.07e+21 | 72.6B | 32 | 4096 | 32 | DCLM+code+math |
| 7.05B | 5.80e+21 | 136.9B | 32 | 4096 | 32 | DCLM+code+math, C4 |
| 6.48B | 4.84e+21 | 124.4B | 31 | 3968 | 31 | DCLM+code+math |
| 6.48B | 1.00e+22 | 258.2B | 31 | 3968 | 31 | DCLM+code+math |
| 5.90B | 4.00e+21 | 113.0B | 30 | 3840 | 30 | DCLM+code+math, C4 |
| 5.90B | 2.12e+21 | 59.9B | 30 | 3840 | 30 | DCLM+code+math |
| 5.35B | 3.30e+21 | 102.7B | 29 | 3712 | 29 | DCLM+code+math, C4 |
| 4.88B | 1.45e+21 | 49.5B | 28 | 3584 | 28 | DCLM+code+math |
| 4.88B | 2.73e+21 | 93.3B | 28 | 3584 | 28 | DCLM+code+math, C4 |
| 4.40B | 4.64e+21 | 175.9B | 27 | 3456 | 27 | DCLM+code+math |
| 4.40B | 2.24e+21 | 84.7B | 27 | 3456 | 27 | DCLM+code+math, C4 |
| 3.96B | 7.15e+21 | 301.4B | 26 | 3328 | 26 | DCLM+code+math |
| 3.96B | 9.69e+20 | 40.8B | 26 | 3328 | 26 | DCLM+code+math |
| 3.96B | 1.83e+21 | 77.0B | 26 | 3328 | 26 | DCLM+code+math, C4 |
| 3.57B | 1.50e+21 | 69.9B | 25 | 3200 | 25 | DCLM+code+math, C4 |
| 3.57B | 3.11e+21 | 145.2B | 25 | 3200 | 25 | DCLM+code+math |
| 3.19B | 6.45e+20 | 33.7B | 24 | 3072 | 24 | DCLM+code+math |
| 3.19B | 1.22e+21 | 63.5B | 24 | 3072 | 24 | DCLM+code+math, C4 |
| 3.19B | 4.76e+21 | 248.8B | 24 | 3072 | 24 | DCLM+code+math |
| 2.84B | 9.82e+20 | 57.7B | 23 | 2944 | 23 | DCLM+code+math, C4 |
| 2.84B | 2.04e+21 | 119.8B | 23 | 2944 | 23 | DCLM+code+math |
| 2.84B | 4.73e+20 | 27.8B | 23 | 2944 | 23 | DCLM+code+math |
| 2.53B | 7.25e+20 | 47.7B | 22 | 2816 | 22 | DCLM+code+math, C4 |
| 2.53B | 1.50e+21 | 98.9B | 22 | 2816 | 22 | DCLM+code+math |
| 2.53B | 3.12e+21 | 205.3B | 22 | 2816 | 22 | DCLM+code+math |
| 2.53B | 7.97e+20 | 52.5B | 22 | 2816 | 22 | DCLM+code+math, C4 |
| 2.23B | 5.80e+20 | 43.3B | 21 | 2688 | 21 | DCLM+code+math, C4 |
| 2.23B | 2.27e+21 | 169.5B | 21 | 2688 | 21 | DCLM+code+math |
| 2.23B | 3.08e+20 | 23.0B | 21 | 2688 | 21 | DCLM+code+math |
| 1.96B | 9.60e+20 | 81.6B | 20 | 2560 | 20 | DCLM+code+math |
| 1.96B | 4.62e+20 | 39.3B | 20 | 2560 | 20 | DCLM+code+math, C4 |
| 1.96B | 2.23e+20 | 18.9B | 20 | 2560 | 20 | DCLM+code+math |
| 1.73B | 6.99e+20 | 67.4B | 19 | 2432 | 19 | DCLM+code+math |
| 1.73B | 3.71e+20 | 35.7B | 19 | 2432 | 19 | DCLM+code+math, C4 |

Table 7: List of experiments (Part 2/3).

| params | flops | tokens | layers | hidden dim | heads | datasets |
|---|---|---|---|---|---|---|
| 1.73B | 1.45e+21 | 139.9B | 19 | 2432 | 19 | DCLM+code+math |
| 1.73B | 3.01e+21 | 290.4B | 19 | 2432 | 19 | DCLM+code+math |
| 1.73B | 3.37e+20 | 32.5B | 19 | 2432 | 19 | DCLM+code+math, C4 |
| 1.50B | 2.66e+20 | 29.5B | 18 | 2304 | 18 | DCLM+code+math, C4 |
| 1.50B | 2.16e+21 | 239.7B | 18 | 2304 | 18 | DCLM+code+math |
| 1.50B | 1.41e+20 | 15.6B | 18 | 2304 | 18 | DCLM+code+math |
| 1.50B | 1.04e+21 | 115.5B | 18 | 2304 | 18 | DCLM+code+math |
| 1.30B | 1.00e+20 | 12.9B | 17 | 2176 | 17 | DCLM+code+math |
| 1.30B | 2.08e+20 | 26.8B | 17 | 2176 | 17 | DCLM+code+math, C4 |
| 1.30B | 1.54e+21 | 197.8B | 17 | 2176 | 17 | DCLM+code+math |
| 1.30B | 1.89e+20 | 24.3B | 17 | 2176 | 17 | DCLM+code+math, C4 |
| 1.30B | 4.32e+20 | 55.6B | 17 | 2176 | 17 | DCLM+code+math |
| 1.12B | 6.43e+20 | 95.3B | 16 | 2048 | 16 | DCLM+code+math |
| 1.12B | 3.10e+20 | 45.9B | 16 | 2048 | 16 | DCLM+code+math |
| 1.12B | 1.49e+20 | 22.1B | 16 | 2048 | 16 | DCLM+code+math, C4 |
| 1.12B | 7.19e+19 | 10.7B | 16 | 2048 | 16 | DCLM+code+math |
| 0.96B | 2.18e+20 | 37.9B | 15 | 1920 | 15 | DCLM+code+math |
| 0.96B | 4.52e+20 | 78.7B | 15 | 1920 | 15 | DCLM+code+math |
| 0.96B | 1.16e+20 | 20.1B | 15 | 1920 | 15 | DCLM+code+math, C4 |
| 0.96B | 9.39e+20 | 163.3B | 15 | 1920 | 15 | DCLM+code+math |
| 0.96B | 1.05e+20 | 18.3B | 15 | 1920 | 15 | DCLM+code+math, C4 |
| 0.81B | 1.52e+20 | 31.3B | 14 | 1792 | 14 | DCLM+code+math |
| 0.81B | 4.28e+19 | 8.8B | 14 | 1792 | 14 | DCLM+code+math |
| 0.81B | 6.56e+20 | 134.8B | 14 | 1792 | 14 | DCLM+code+math |
| 0.81B | 3.16e+20 | 64.9B | 14 | 1792 | 14 | DCLM+code+math |
| 0.81B | 8.08e+19 | 16.6B | 14 | 1792 | 14 | DCLM+code+math, C4 |
| 0.81B | 7.34e+19 | 15.1B | 14 | 1792 | 14 | DCLM+code+math, C4 |
| 0.69B | 2.22e+20 | 53.6B | 13 | 1664 | 13 | DCLM+code+math |
| 0.69B | 5.67e+19 | 13.7B | 13 | 1664 | 13 | DCLM+code+math, C4 |
| 0.69B | 1.07e+20 | 25.8B | 13 | 1664 | 13 | DCLM+code+math |
| 0.69B | 4.61e+20 | 111.3B | 13 | 1664 | 13 | DCLM+code+math |
| 0.69B | 3.01e+19 | 7.3B | 13 | 1664 | 13 | DCLM+code+math |
| 0.58B | 1.53e+20 | 44.2B | 12 | 1536 | 12 | DCLM+code+math |
| 0.58B | 2.07e+19 | 6.0B | 12 | 1536 | 12 | DCLM+code+math |
| 0.58B | 4.30e+19 | 12.4B | 12 | 1536 | 12 | DCLM+code+math, C4 |
| 0.58B | 3.90e+19 | 11.3B | 12 | 1536 | 12 | DCLM+code+math, C4 |
| 0.58B | 3.17e+20 | 91.8B | 12 | 1536 | 12 | DCLM+code+math |
| 0.58B | 7.36e+19 | 21.3B | 12 | 1536 | 12 | DCLM+code+math |
| 0.48B | 5.02e+19 | 17.6B | 11 | 1408 | 11 | DCLM+code+math |
| 0.48B | 1.04e+20 | 36.5B | 11 | 1408 | 11 | DCLM+code+math |
| 0.48B | 1.41e+19 | 4.9B | 11 | 1408 | 11 | DCLM+code+math |
| 0.48B | 2.16e+20 | 75.8B | 11 | 1408 | 11 | DCLM+code+math |
| 0.48B | 2.93e+19 | 10.3B | 11 | 1408 | 11 | DCLM+code+math, C4 |
| 0.48B | 2.66e+19 | 9.3B | 11 | 1408 | 11 | DCLM+code+math, C4 |
| 0.39B | 3.44e+19 | 14.5B | 10 | 1280 | 10 | DCLM+code+math |
| 0.39B | 1.48e+20 | 62.6B | 10 | 1280 | 10 | DCLM+code+math |
| 0.39B | 9.67e+18 | 4.1B | 10 | 1280 | 10 | DCLM+code+math |
| 0.39B | 7.14e+19 | 30.1B | 10 | 1280 | 10 | DCLM+code+math |
| 0.39B | 2.01e+19 | 8.5B | 10 | 1280 | 10 | DCLM+code+math, C4 |
| 0.39B | 1.82e+19 | 7.7B | 10 | 1280 | 10 | DCLM+code+math, C4 |
| 0.32B | 6.48e+18 | 3.4B | 9 | 1152 | 9 | DCLM+code+math |

Table 8: List of experiments (Part 3/3).

| params | flops | tokens | layers | hidden dim | heads | datasets |
|---|---|---|---|---|---|---|
| 0.32B | 1.22e+19 | 6.4B | 9 | 1152 | 9 | DCLM+code+math, C4 |
| 0.32B | 1.34e+19 | 7.0B | 9 | 1152 | 9 | DCLM+code+math, C4 |
| 0.32B | 1.11e+19 | 5.8B | 9 | 1152 | 9 | DCLM+code+math, C4 |
| 0.32B | 9.92e+19 | 51.6B | 9 | 1152 | 9 | DCLM+code+math |
| 0.32B | 4.78e+19 | 24.9B | 9 | 1152 | 9 | DCLM+code+math |
| 0.32B | 2.30e+19 | 12.0B | 9 | 1152 | 9 | DCLM+code+math |
| 0.26B | 4.29e+18 | 2.8B | 8 | 1024 | 8 | DCLM+code+math |
| 0.26B | 1.52e+19 | 9.9B | 8 | 1024 | 8 | DCLM+code+math |
| 0.26B | 6.57e+19 | 42.6B | 8 | 1024 | 8 | DCLM+code+math |
| 0.26B | 8.08e+18 | 5.2B | 8 | 1024 | 8 | DCLM+code+math, C4 |
| 0.26B | 3.17e+19 | 20.5B | 8 | 1024 | 8 | DCLM+code+math |
| 0.26B | 7.35e+18 | 4.8B | 8 | 1024 | 8 | DCLM+code+math, C4 |
| 0.21B | 1.01e+19 | 8.2B | 7 | 896 | 7 | DCLM+code+math |
| 0.21B | 4.84e+18 | 3.9B | 7 | 896 | 7 | DCLM+code+math, C4 |
| 0.21B | 2.83e+18 | 2.3B | 7 | 896 | 7 | DCLM+code+math |
| 0.21B | 5.33e+18 | 4.3B | 7 | 896 | 7 | DCLM+code+math, C4 |
| 0.21B | 2.09e+19 | 16.9B | 7 | 896 | 7 | DCLM+code+math |
| 0.21B | 4.33e+19 | 35.2B | 7 | 896 | 7 | DCLM+code+math |
| 0.21B | 2.33e+18 | 1.9B | 7 | 896 | 7 | DCLM+code+math |
| 0.16B | 2.31e+19 | 24.0B | 6 | 768 | 6 | DCLM+code+math |
| 0.16B | 1.11e+19 | 11.5B | 6 | 768 | 6 | DCLM+code+math |
| 0.16B | 2.80e+19 | 29.0B | 6 | 768 | 6 | DCLM+code+math |
| 0.16B | 1.35e+19 | 14.0B | 6 | 768 | 6 | DCLM+code+math |
| 0.16B | 2.84e+18 | 2.9B | 6 | 768 | 6 | DCLM+code+math, C4 |
| 0.16B | 3.13e+18 | 3.2B | 6 | 768 | 6 | DCLM+code+math, C4 |
| 0.16B | 3.44e+18 | 3.6B | 6 | 768 | 6 | DCLM+code+math, C4 |
| 0.16B | 1.51e+18 | 1.6B | 6 | 768 | 6 | DCLM+code+math |
| 0.16B | 6.49e+18 | 6.7B | 6 | 768 | 6 | DCLM+code+math |
| 0.12B | 7.02e+18 | 9.5B | 5 | 640 | 5 | DCLM+code+math |
| 0.12B | 1.63e+18 | 2.2B | 5 | 640 | 5 | DCLM+code+math, C4 |
| 0.12B | 1.79e+18 | 2.4B | 5 | 640 | 5 | DCLM+code+math, C4 |
| 0.12B | 1.97e+18 | 2.7B | 5 | 640 | 5 | DCLM+code+math, C4 |
| 0.12B | 9.51e+17 | 1.3B | 5 | 640 | 5 | DCLM+code+math |
| 0.12B | 4.10e+18 | 5.6B | 5 | 640 | 5 | DCLM+code+math |
| 0.12B | 3.38e+18 | 4.6B | 5 | 640 | 5 | DCLM+code+math |
| 0.12B | 1.46e+19 | 19.8B | 5 | 640 | 5 | DCLM+code+math |
| 0.09B | 8.94e+18 | 16.3B | 4 | 512 | 4 | DCLM+code+math |
| 0.09B | 1.10e+18 | 2.0B | 4 | 512 | 4 | DCLM+code+math, C4 |
| 0.09B | 7.38e+18 | 13.5B | 4 | 512 | 4 | DCLM+code+math |
| 0.09B | 3.56e+18 | 6.5B | 4 | 512 | 4 | DCLM+code+math |
| 0.09B | 1.00e+18 | 1.8B | 4 | 512 | 4 | DCLM+code+math, C4 |
| 0.09B | 2.08e+18 | 3.8B | 4 | 512 | 4 | DCLM+code+math |
| 0.09B | 4.31e+18 | 7.9B | 4 | 512 | 4 | DCLM+code+math |
| 0.09B | 1.71e+18 | 3.1B | 4 | 512 | 4 | DCLM+code+math |
| 0.06B | 9.96e+17 | 2.6B | 3 | 384 | 3 | DCLM+code+math |
| 0.06B | 4.29e+18 | 11.1B | 3 | 384 | 3 | DCLM+code+math |
| 0.06B | 2.07e+18 | 5.4B | 3 | 384 | 3 | DCLM+code+math |
| 0.06B | 3.54e+18 | 9.2B | 3 | 384 | 3 | DCLM+code+math |
| 0.06B | 1.71e+18 | 4.4B | 3 | 384 | 3 | DCLM+code+math |
| 0.04B | 1.27e+18 | 5.2B | 2 | 256 | 2 | DCLM+code+math |
| 0.04B | 8.96e+17 | 3.7B | 2 | 256 | 2 | DCLM+code+math |
| 0.04B | 1.53e+18 | 6.3B | 2 | 256 | 2 | DCLM+code+math |
| 0.04B | 1.86e+18 | 7.6B | 2 | 256 | 2 | DCLM+code+math |

## C Upper and Lower Bounds for Pass@k

We derive analytical bounds for the pass@k probability, i.e. the probability that at least one of $k$ independently sampled attempts succeeds. Let each trial succeed with probability $q \in [0, 1]$ independently. Then the probability that all $k$ trials fail is

$$p(\text{All Failures up to } k) = (1 - q)^k,$$

and consequently, the probability of at least one success (the *pass@k*) is

$$p(\text{pass@}k) = 1 - (1 - q)^k. \tag{6}$$

For $x > -1$ and integer $k \geq 1$, Bernoulli's inequality states that

$$(1 + x)^k \geq 1 + kx.$$

Setting $x = -q$ (so that $1 + x = 1 - q$), we obtain

$$(1 - q)^k \geq 1 - kq.$$

Substituting this into Equation 6 gives the *upper bound*:

$$p(\text{pass@}k) \leq \max(kq, 1). \tag{7}$$

Using the classical inequality $1 - q \leq e^{-q}$ for $q \geq 0$, we have

$$(1 - q)^k \leq e^{-kq}.$$

Substituting this into Equation 6, and noting that for any $x > 0$ with Taylor expansion we have $e^x \geq 1 = x$, thus $xe^{-x} \leq 1 - e^{-x}$, yields the *lower bound* by setting $x = kq$:

$$p(\text{pass@}k) \geq 1 - e^{-kq} \geq kqe^{-kq}. \tag{8}$$

Combining Equation 7 and Equation 8, we obtain

$$kqe^{-kq} \;\leq\; p(\text{pass@}k) \;\leq\; \max(kq, 1). \tag{9}$$

For a model with compute budget $C$, we can now use our approximation of $q \approx e^{-\frac{A}{C^\alpha}}$. The lower bound here motivates our formula of Equation 5. We have

$$p(\text{pass@}k) = Q(C, k) \approx kqe^{-kq} = k \exp\left(-k \exp\left(-\frac{A}{C^\alpha}\right) - \frac{A}{C^\alpha}\right),$$

$$-\log Q(C, K) \approx \log k + \frac{A}{C^\alpha} + k \exp(-\frac{A}{C^\alpha}),$$

$$log\left(-\log Q(C, K)\right) \approx \log(A) + \alpha \log(C) + \frac{A}{C^\alpha} \log k + \log\log k$$

where in Equation 5, we ignore the effect of $\log\log k$ and we modify the interaction term of $k$ and $C$ to a simpler form of $\log C \log k$. We additionally learn each of the coefficient of the equation rather than assuming they all follow the same $A$ and $\alpha$.

# D More Details for the Choice of Proxy Metric in Two-Stage Approach

## D.1 Most Proxy Metrics Show Strong Predictive Power

We evaluate the efficacy of various proxy metrics in predicting final downstream performance on several benchmarks. Following the methodology in Section 4.2.1, we model the relationship between each proxy and downstream accuracy using a logistic function. As shown in Table 9, the low RMSE and high $R^2$ values indicate that nearly all proxy metrics are strong predictors of the final outcome.

This key result implies that the specific choice of proxy metric is less critical than previously assumed, affording practitioners the flexibility to choose based on convenience or computational cost. More importantly, it suggests that if many proxies are equally predictive, a direct optimization of the downstream task might be a more effective approach. We validate this hypothesis experimentally: our direct approach not only performs well but consistently outperforms all proxy-based methods.

Additionally, we analyze the predictive power of individual MMLU sub-tasks for overall MMLU accuracy (Table 10). We find considerable variance, indicating that certain sub-tasks are far better predictors of aggregate performance than others.

## D.2 Comparing Scaling Laws across Benchmarks

We evaluate the performance of two-stage (linear, logistic) and direct (BNSL, simple power law) modeling approaches on several benchmarks. As illustrated in Figure 7, while all methods demonstrate a high goodness-of-fit, their performance differs markedly during extrapolation.

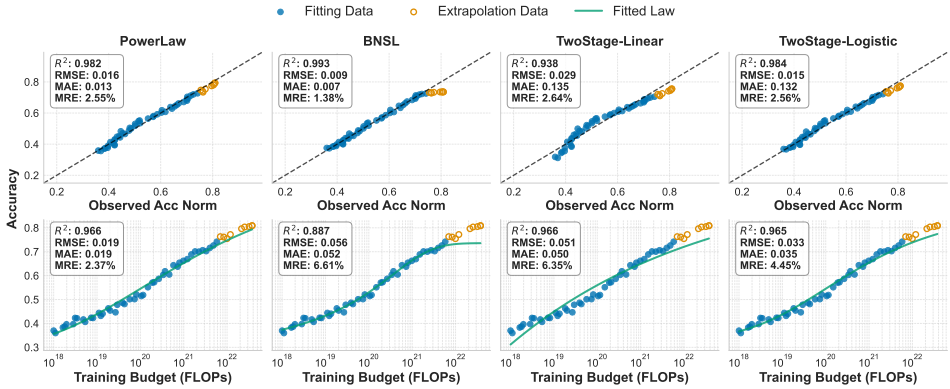
The direct approaches consistently outperform their two-stage counterparts. We attribute this performance gap to the compounding errors inherent in two-stage pipelines, where inaccuracies from the initial fitting stage are propagated and amplified. This finding is robust across all tested benchmarks, highlighting the benefits of a direct, end-to-end modeling strategy for this task.

Table 9: Predictive performance of various metrics across all benchmarks. For each benchmark, we report the coefficient of determination ($R^2$), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Relative Error (MRE) when predicting the final task accuracy.

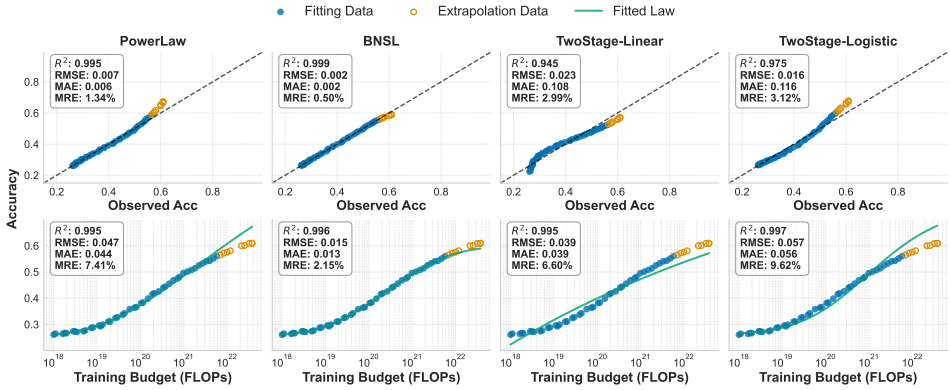
| Benchmark | Predictor Metric | $R^2$ | RMSE | MAE | MRE |
|---|---|---|---|---|---|
| arc_easy | arc_easy_acc | 0.996 | 0.008 | 0.006 | 0.013 |
| | arc_easy_brier | 0.997 | 0.008 | 0.006 | 0.012 |
| | arc_easy_loglikelihood | 0.995 | 0.010 | 0.007 | 0.014 |
| | arc_easy_neg_loglikelihood_norm | 0.989 | 0.013 | 0.010 | 0.021 |
| | c4_loss | 0.993 | 0.011 | 0.009 | 0.018 |
| | dclm_loss | 0.993 | 0.011 | 0.009 | 0.018 |
| | fineweb_loss | 0.993 | 0.011 | 0.009 | 0.018 |
| | mmlu_5s_loss | 0.992 | 0.012 | 0.009 | 0.018 |
| | openwebtext_2_loss | 0.992 | 0.012 | 0.009 | 0.019 |
| arc_challenge | arc_challenge_acc | 0.993 | 0.008 | 0.006 | 0.021 |
| | arc_challenge_brier | 0.994 | 0.007 | 0.006 | 0.020 |
| | arc_challenge_loglikelihood | 0.987 | 0.011 | 0.008 | 0.028 |
| | arc_challenge_neg_loglikelihood_norm | 0.980 | 0.012 | 0.009 | 0.032 |
| | c4_loss | 0.991 | 0.009 | 0.007 | 0.025 |
| | dclm_loss | 0.991 | 0.009 | 0.007 | 0.025 |
| | fineweb_loss | 0.991 | 0.009 | 0.007 | 0.024 |
| | mmlu_5s_loss | 0.990 | 0.009 | 0.007 | 0.025 |
| | openwebtext_2_loss | 0.989 | 0.010 | 0.008 | 0.027 |
| sciq | c4_loss | 0.985 | 0.015 | 0.012 | 0.017 |
| | dclm_loss | 0.985 | 0.016 | 0.012 | 0.017 |
| | fineweb_loss | 0.985 | 0.015 | 0.012 | 0.017 |
| | mmlu_5s_loss | 0.987 | 0.015 | 0.012 | 0.016 |
| | openwebtext_2_loss | 0.985 | 0.016 | 0.012 | 0.016 |
| | sciq_acc | 0.992 | 0.011 | 0.009 | 0.013 |
| | sciq_brier | 0.994 | 0.010 | 0.008 | 0.011 |
| | sciq_loglikelihood | 0.984 | 0.016 | 0.013 | 0.018 |
| | sciq_neg_loglikelihood_norm | 0.978 | 0.018 | 0.014 | 0.021 |
| piqa | c4_loss | 0.994 | 0.005 | 0.004 | 0.006 |
| | dclm_loss | 0.994 | 0.005 | 0.004 | 0.006 |
| | fineweb_loss | 0.994 | 0.005 | 0.004 | 0.006 |
| | mmlu_5s_loss | 0.994 | 0.006 | 0.005 | 0.007 |
| | openwebtext_2_loss | 0.994 | 0.006 | 0.004 | 0.007 |
| | piqa_acc | 0.992 | 0.007 | 0.005 | 0.008 |
| | piqa_brier | 0.995 | 0.005 | 0.004 | 0.006 |
| | piqa_loglikelihood | 0.993 | 0.006 | 0.005 | 0.007 |
| | piqa_neg_loglikelihood_norm | 0.986 | 0.008 | 0.006 | 0.010 |
| hellaswag | c4_loss | 0.974 | 0.016 | 0.012 | 0.039 |
| | dclm_loss | 0.975 | 0.015 | 0.011 | 0.039 |
| | fineweb_loss | 0.974 | 0.016 | 0.012 | 0.039 |
| | mmlu_5s_loss | 0.969 | 0.017 | 0.012 | 0.041 |
| | openwebtext_2_loss | 0.977 | 0.015 | 0.011 | 0.037 |
| hellaswag | c4_loss | 1.000 | 0.002 | 0.001 | 0.004 |
| | dclm_loss | 1.000 | 0.002 | 0.002 | 0.004 |
| | fineweb_loss | 1.000 | 0.002 | 0.001 | 0.004 |
| | hellaswag_brier | 0.000 | 0.108 | 0.096 | 0.260 |
| | hellaswag_loglikelihood | 1.000 | 0.002 | 0.002 | 0.004 |
| | hellaswag_neg_loglikelihood_norm | 1.000 | 0.002 | 0.002 | 0.004 |
| | mmlu_5s_loss | 0.999 | 0.004 | 0.003 | 0.007 |
| | openwebtext_2_loss | 0.998 | 0.004 | 0.003 | 0.009 |

Table 10: Predictive performance of in-domain sub-task accuracy for MMLU.

| **Predictor Metric** | $R^2$ | **RMSE** | **MAE** | **MRE** |
|---|---|---|---|---|
| Abstract Algebra | 0.521 | 0.068 | 0.042 | 0.125 |
| Anatomy | 0.932 | 0.025 | 0.017 | 0.052 |
| Astronomy | 0.969 | 0.017 | 0.013 | 0.043 |
| Business Ethics | 0.938 | 0.024 | 0.016 | 0.050 |
| Clinical Knowledge | 0.972 | 0.016 | 0.012 | 0.040 |
| College Biology | 0.950 | 0.022 | 0.016 | 0.051 |
| College Chemistry | 0.400 | 0.076 | 0.052 | 0.158 |
| College Computer Science | 0.838 | 0.039 | 0.024 | 0.073 |
| College Mathematics | 0.759 | 0.048 | 0.029 | 0.088 |
| College Medicine | 0.909 | 0.029 | 0.020 | 0.062 |
| College Physics | 0.442 | 0.073 | 0.050 | 0.152 |
| Computer Security | 0.972 | 0.016 | 0.012 | 0.039 |
| Conceptual Physics | 0.921 | 0.027 | 0.019 | 0.062 |
| Econometrics | 0.525 | 0.067 | 0.043 | 0.131 |
| Electrical Engineering | 0.940 | 0.024 | 0.014 | 0.042 |
| Elementary Mathematics | 0.899 | 0.031 | 0.021 | 0.065 |
| Formal Logic | 0.495 | 0.069 | 0.046 | 0.142 |
| Global Facts | 0.240 | 0.085 | 0.059 | 0.179 |
| High School Biology | 0.978 | 0.014 | 0.011 | 0.038 |
| High School Chemistry | 0.874 | 0.035 | 0.023 | 0.073 |
| High School Computer Science | 0.892 | 0.032 | 0.022 | 0.069 |
| High School European History | 0.974 | 0.016 | 0.011 | 0.037 |
| High School Geography | 0.978 | 0.014 | 0.011 | 0.036 |
| High School Government And Politics | 0.982 | 0.013 | 0.010 | 0.033 |
| High School Macroeconomics | 0.957 | 0.020 | 0.014 | 0.047 |
| High School Mathematics | 0.000 | 0.098 | 0.070 | 0.209 |
| High School Microeconomics | 0.948 | 0.022 | 0.017 | 0.056 |
| High School Physics | 0.704 | 0.053 | 0.034 | 0.107 |
| High School Psychology | 0.986 | 0.012 | 0.009 | 0.029 |
| High School Statistics | 0.704 | 0.053 | 0.034 | 0.105 |
| High School Us History | 0.963 | 0.019 | 0.015 | 0.052 |
| High School World History | 0.969 | 0.017 | 0.014 | 0.048 |
| Human Aging | 0.933 | 0.025 | 0.018 | 0.060 |
| Human Sexuality | 0.970 | 0.017 | 0.012 | 0.036 |
| International Law | 0.952 | 0.021 | 0.015 | 0.048 |
| Jurisprudence | 0.958 | 0.020 | 0.015 | 0.050 |
| Logical Fallacies | 0.951 | 0.022 | 0.017 | 0.057 |
| Machine Learning | 0.573 | 0.064 | 0.040 | 0.123 |
| Management | 0.947 | 0.023 | 0.017 | 0.055 |
| Marketing | 0.970 | 0.017 | 0.013 | 0.047 |
| Medical Genetics | 0.917 | 0.028 | 0.020 | 0.065 |
| Miscellaneous | 0.984 | 0.012 | 0.010 | 0.037 |
| Moral Disputes | 0.979 | 0.014 | 0.011 | 0.038 |
| Moral Scenarios | 0.000 | 0.098 | 0.070 | 0.209 |
| Nutrition | 0.982 | 0.013 | 0.009 | 0.030 |
| Philosophy | 0.969 | 0.017 | 0.013 | 0.043 |
| Prehistory | 0.979 | 0.014 | 0.011 | 0.037 |
| Professional Accounting | 0.914 | 0.029 | 0.020 | 0.064 |
| Professional Law | 0.000 | 0.098 | 0.070 | 0.209 |
| Professional Medicine | 0.873 | 0.035 | 0.021 | 0.066 |
| Professional Psychology | 0.970 | 0.017 | 0.014 | 0.048 |
| Public Relations | 0.946 | 0.023 | 0.016 | 0.053 |
| Security Studies | 0.967 | 0.018 | 0.012 | 0.039 |
| Sociology | 0.987 | 0.011 | 0.009 | 0.029 |
| Us Foreign Policy | 0.974 | 0.016 | 0.012 | 0.040 |
| Virology | 0.947 | 0.023 | 0.015 | 0.048 |
| World Religions | 0.957 | 0.020 | 0.015 | 0.052 |

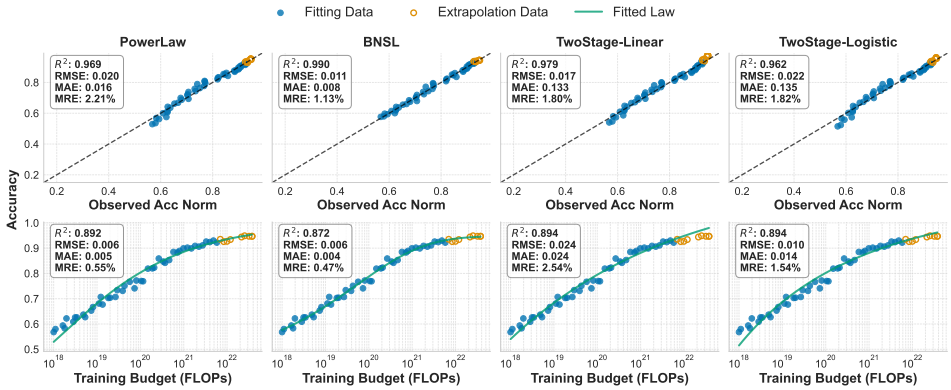



(a) ARC Easy. Legend depicts error on fitted points (top row) and validation points (bottom row).




(b) HellaSwag. Legend depicts error on fitted points (top row) and validation points (bottom row).


(c) PIQA. Legend depicts error on fitted points (top row) and validation points (bottom row).




(d) SciQ. Legend depicts error on fitted points (top row) and validation points (bottom row).

Figure 7: Comparing the direct approaches (BNSL, simple power law) with two-stage approaches (Linear and Logistic) for various benchmarks.

## E Critical FLOPs Threshold for Predicting Downstream Performance

The choice of the FLOPs threshold, which partitions data for training and validation, significantly impacts the reported performance and reliability of scaling law models. We conduct an analysis to quantify this sensitivity and determine the minimum FLOPs threshold required for each model to achieve a robust extrapolation, which we define as achieving a Mean Relative Error (MRE) below 10%. In Figure 8, we plot the extrapolation MRE and MAE as a function of the FLOPs threshold. The results show that the BNSL and simple power law models exhibit stable performance, maintaining low MRE across a wide range of thresholds. In contrast, the two-stage linear and logistic models are less robust, demonstrating higher error rates and greater sensitivity to the choice of threshold.

To formalize this robustness analysis, we introduce a binary success criterion: a model "succeeds" at a given threshold if its MRE is below 10%. As illustrated in Figure 9, we sweep the FLOPs threshold from $6e19$ to $5e22$ and fit a logistic regression to the binary success outcomes. This allows us to estimate the FLOPs threshold at which each model is likely to fail. Our findings confirm that the simple power law is the most resilient model, followed closely by BNSL. The two-stage models are shown to be suboptimal, requiring a much higher FLOPs threshold to consistently achieve reliable fits. This analysis provides a way to assess model stability for scaling law extrapolation.

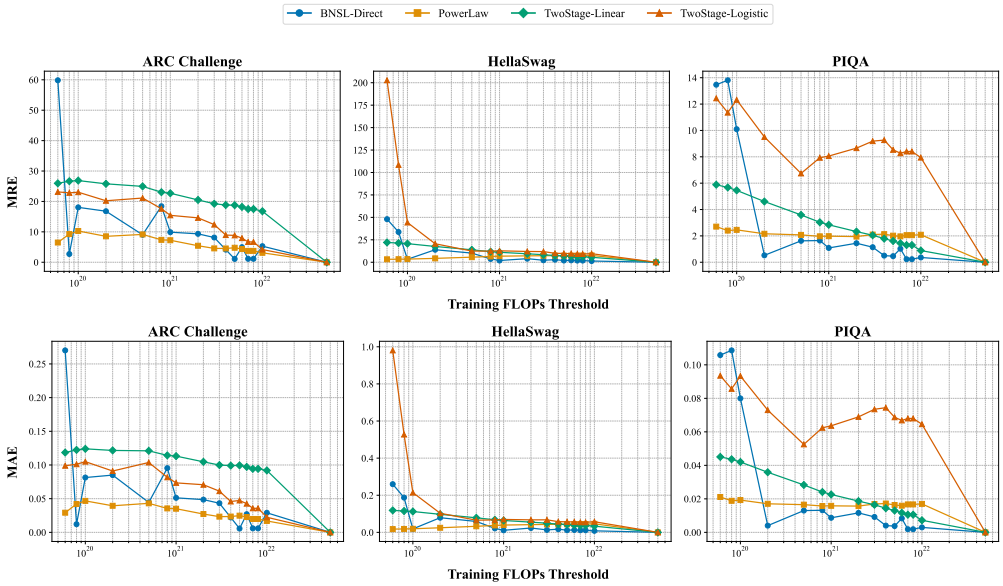



Figure 8: Trend of two metrics MRE and MAE across benchmarks and scaling laws.

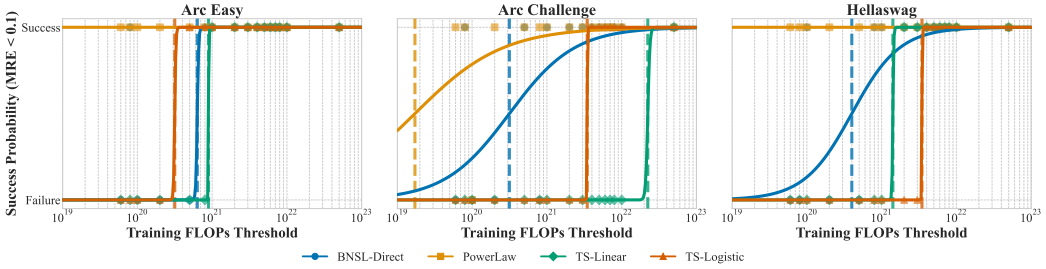
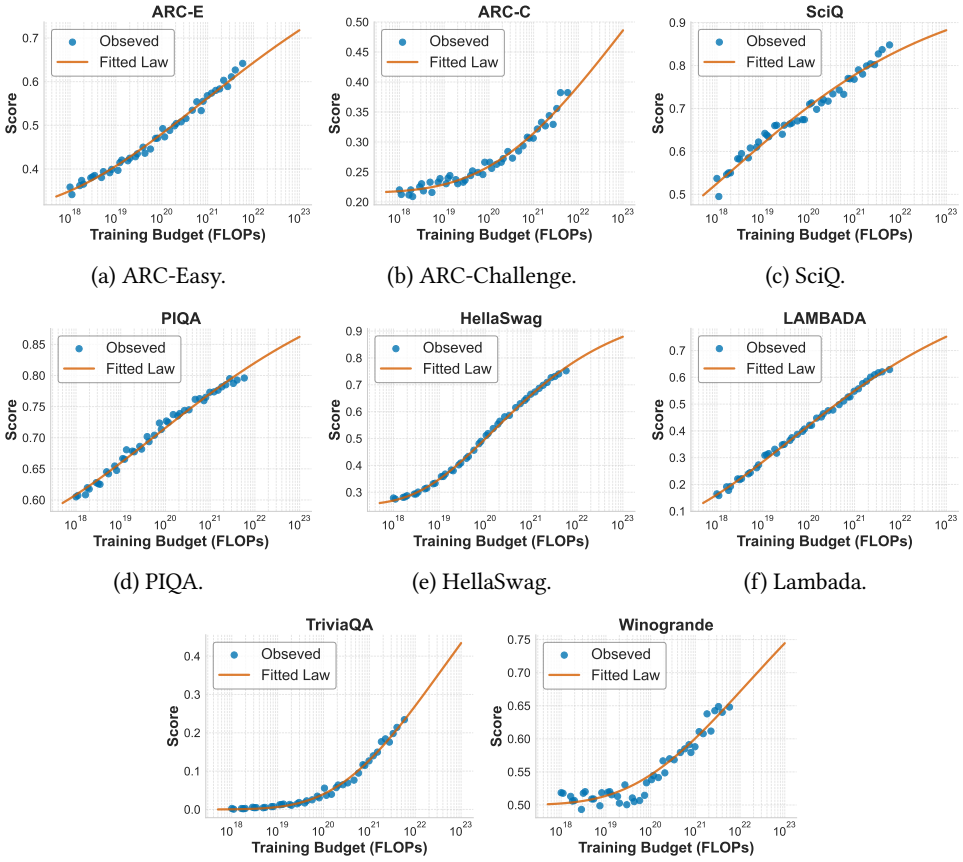



Figure 9: Minimum FLOPs Threshold for Reliable Extrapolation. We define 'Success' at a given FLOPs threshold as achieving MRE below 10%. A logistic regression is fitted to these binary outcomes to model the probability of success as a function of the threshold. The vertical dashed line indicates the estimated threshold where the probability of success surpasses 50%, effectively separating the failure and success regimes.

## F Downstream Scaling Curves for C4 Dataset




(a) ARC-Easy. (b) ARC-Challenge. (c) SciQ.

(d) PIQA. (e) HellaSwag. (f) Lambada.

(g) TriviaQA. (h) Winogrande.

Figure 10: Downstream scaling curves when changing the pre-training dataset to C4.

## G Hyperparameters

We determined the maximum batch size with near-optimal performance at 1.8B training tokens as 64 and scaled it proportionally to $D^{0.5}$, based on the recommendations from the literature (Filatov et al., 2025; Bergsma et al., 2025; Zhang et al., 2025). We set the maximum global learning rate to 5e-3, derived by tuning a proxy model and later transferring with $\mu$-parametrization (Yang et al., 2022) in its simplified form, as described in Wortsman et al. (2023). We follow Gunter et al. (2024) in the setup of optimizer and weight decay (decoupled weight decay of 3.16e-4), since it has been ablated for quality and stability across a range of compute scales.

## H Predicting the Average Across Benchmarks

Gadre et al. (2024b) propose to predict the average score across tasks rather than metrics on specific benchmarks. They use the two stage approach in this problem. Here we present the initial examination on the possibility of describing and extrapolating the value the average score using the direct method.

We note that is not immediately clear which functional form is theoretically correct, as we take the mean of the scores on multiple individual benchmarks. Taking the empirical perspective, in this pilot study we examine whether the functional form of Equation 2 can be used in this case.

We consider models trained with the to token-to-param ratio 20. For each experiment, we calculate the mean of the model scores across Arc-E, Arc,C, SciQ, PIQA, HellaSwag, WinoGrande, WebQS, TriviaQA, Lambada and HumanEval. We do not consider LBPP, as this metric only gives signal on a relatively large scale, and would not contribute meaningfully. We exclude models where any of the benchmarks achieved a score of less than 5% points above the random performance to reduce the points with random noise. For all metrics, where the scores are not by default in the range [0, 1] (i.e. multiple-choice tasks), before taking the mean, we first normalize the scores as described in Section 3.2 (we apply the transformation $Q' := (Q - Q_{\text{random}})/(1 - Q_{\text{random}})$).

The results are illustrated in Figure 11. We exclude expeirments with more than 6e21 training FLOPs as the set for validating the extrapolation. We observe good quality of the fit, as outlined in Table 11. As the next step, we could consider modeling this relationship using an equation with more fitted parameters, like BNSL with multiple breaking points. We leave further examination of this setup for future work.

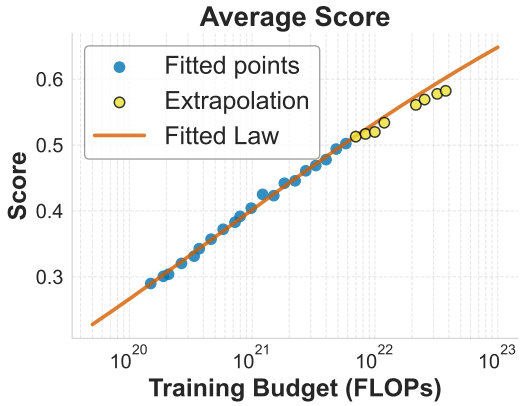



Figure 11: Illustration of the fit quality for predicting the average across benchmarks.

Table 11: Fit quality of predicting the average across benchmarks.

| Train MAE | Valid MAE | Train MRE (%) | Valid MRE (%) |
|---|---|---|---|
| 0.0030 | 0.0116 | 0.76% | 2.08% |

## I Modeling Accuracy with Irreducible Error

In Section 3 we assume the perfect scenario, where the maximum achievable accuracy on each task is equal to 1. This assumption may not hold in practice: it has been observed that the benchmark score plateaus on a value below 1, due to a certain number of incorrectly labeled or ambiguous questions (Vendrow et al., 2025). Here we examine the possibility of adding the asymptote of maximum achievable accuracy to the functional form.

We replicate the approach secribed in Section 3.2, but with Equation 2 modified to incorporate the irreducible error:

$$-\log(Q) = \frac{A}{C^{\alpha}} + E, \tag{10}$$

where $Q_{\max} \coloneqq \exp(-E) \in (0, 1)$ represents the estimated value of maximum achievable accuracy. We fit the coefficients $A, \alpha$ and $E$, using only models with training FLOPs less than 6e21, leaving the remaining ones as the validation set. The results are illustrated in Figure 12, showing a good quality of the fit with this functional form.

Table 12 outlines the estimated values of $Q_{\max}$ and error rates on the validation points. For seven benchmarks, the fitted value of $Q_{\max}$ remains unchanged from Equation 2, with $Q_{\max} = 1$. In three cases (PIQA, HellaSwag, Lambada), we can see meaningful values of maximum accuracy, with $Q_{\max}$

estimated as 0.903, 0.912 and 0.947, respectively. For two benchmarks, where the accuracy observed in all experiments used for the fit is relatively far from perfect score (Winogrande, WebQS), the estimated values of $Q_{\text{max}}$ (0.776, 0.51) are unlikely to be correct.

To summarize our findings, using the functional form with the $Q_{\text{max}}$ can be a useful strategy, especially when the target models are expected to approach the maximum achievable accuracy. However, this approach must be applied with caution, to make sure that the estimated perfect score aligns with the number of incorrect or ambiguous questions. In all cases, it is crucial to carefully examine the evaluation datasets, and apply both manual and automatic quality filtering (for example, as descibed in Vendrow et al. (2025)), to ensure that the test scores reliably measure model quality.

We can incorporate the maximum achievable score in other scaling law forms considered in this work, similarly to how we did it for Equation 2. For example, in Equation 4, we can analogously incorporate the additional summand, resulting in the following functional form: $-\log Q = \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} + E$.

Table 12: Estimated maximum accuracy and error rates for fitting the scaling law with maximum achievable accuracy.

| Benchmark | Metric Type | Fitted $Q_{\text{max}}$ | Valid MRE | Valid MAE |
|---|---|---|---|---|
| ARC-E | acc norm | 1.000 | 1.81% | 0.0143 |
| ARC-C | acc norm | 1.000 | 1.26% | 0.0066 |
| SciQ | acc norm | 1.000 | 0.56% | 0.0053 |
| PIQA | acc norm | 0.903 | 0.30% | 0.0024 |
| HellaSwag | acc norm | 0.912 | 0.99% | 0.0079 |
| Winogrande | acc | 0.776 | 1.65% | 0.0122 |
| WebQS | exact match | 0.510 | 1.26% | 0.0028 |
| TriviaQA | exact match | 1.000 | 6.93% | 0.0290 |
| LAMBADA | acc | 0.947 | 2.27% | 0.0170 |
| GSM8K | exact match | 1.000 | 10.99% | 0.0635 |
| HumanEval | pass@1 | 1.000 | 3.82% | 0.0125 |
| LBPP | pass@1 | 1.000 | 6.17% | 0.0070 |

## J Additional Details on Scaling Law Fitting and Validation

We fit Equation 1 by optimizing the Huber Loss ($\delta = 1e-3$) using the basin-hopping (Wales & Doye, 1997) algorithm. We use only points with accuracy $Q > Q_{\text{random}}$ for the fit of BNSL (note that we apply a less strict filtering than $Q_{\text{random}} + 0.05$ considered for Equation 2, since this helps in estimating the lower asymptote and improves extrapolation performance for BNSL). In the case of LBPP, we filter the fitted points to $Q_{\text{random}} + 0.02$, since all of the models at our scale achieve relatively low scores on this benchmark and we want to ensure there are enough points to consider. In fitting Equation 10 and in Appendix H, we adopt the L-BFGS-B algorithm with the Huber Loss ($\delta = 1e-3$) objective, sampling a grid of initalizations and choosing the one with the best score on the fitted points. Throughout the paper, we use least squares for fitting the coefficients in all other cases (Equation 2 and in TwoStage-Linear procedure), due to the closed form of the solution and lack of the dependency on hyperparameters in optimization. We detail the estimated values of $Q_{\text{random}}$ for the acc_norm in Table 13. For all other metric types, we adopt the standard random-guess baseline.

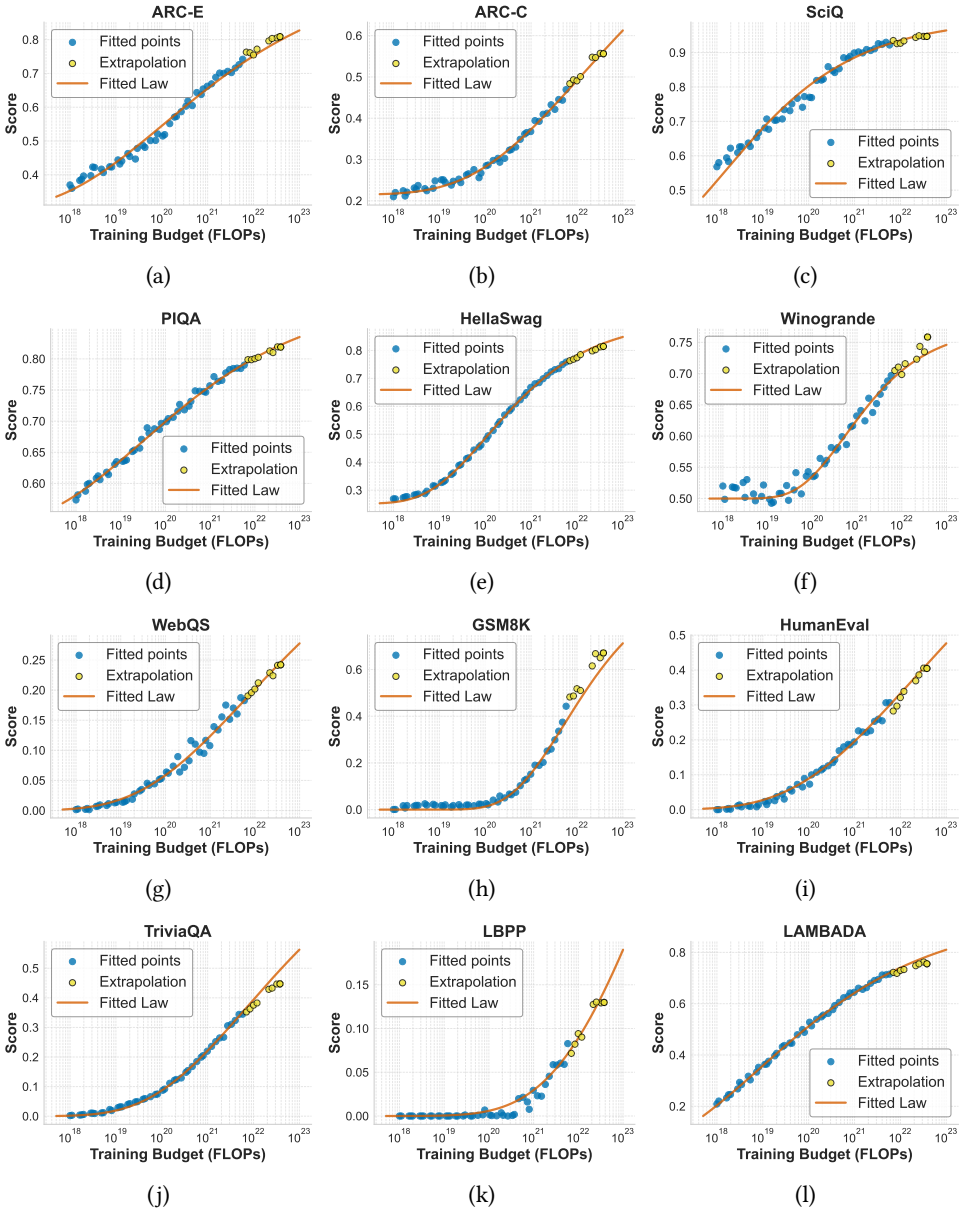



Figure 12: Results of the fitting of the equation with irreducible error for accuracy.

Table 13: Values of $Q_{\text{random}}$.

| Benchmark | Metric Type | $Q_{\text{random}}$ |
|---|---|---|
| Arc-E | acc_norm | 0.291 |
| Arc-C | acc_norm | 0.215 |
| SciQ | acc_norm | 0.304 |
| PIQA | acc_norm | 0.53 |
| HellaSwag | acc_norm | 0.252 |

## K Fit Quality Plots for Varying Token to Param Ratio

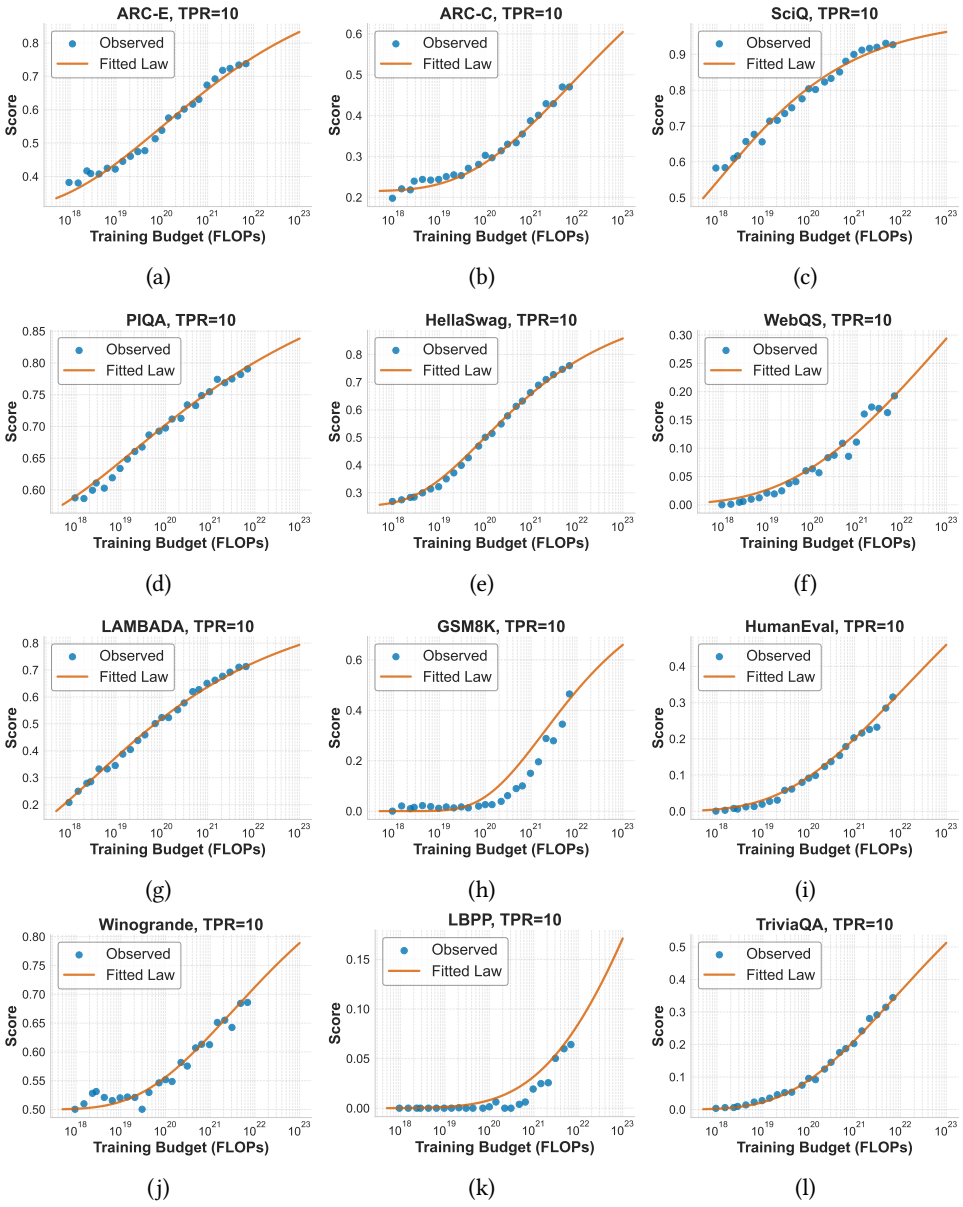



Figure 13: Fit of Eq. (4) for Token to Param Ratio 10.

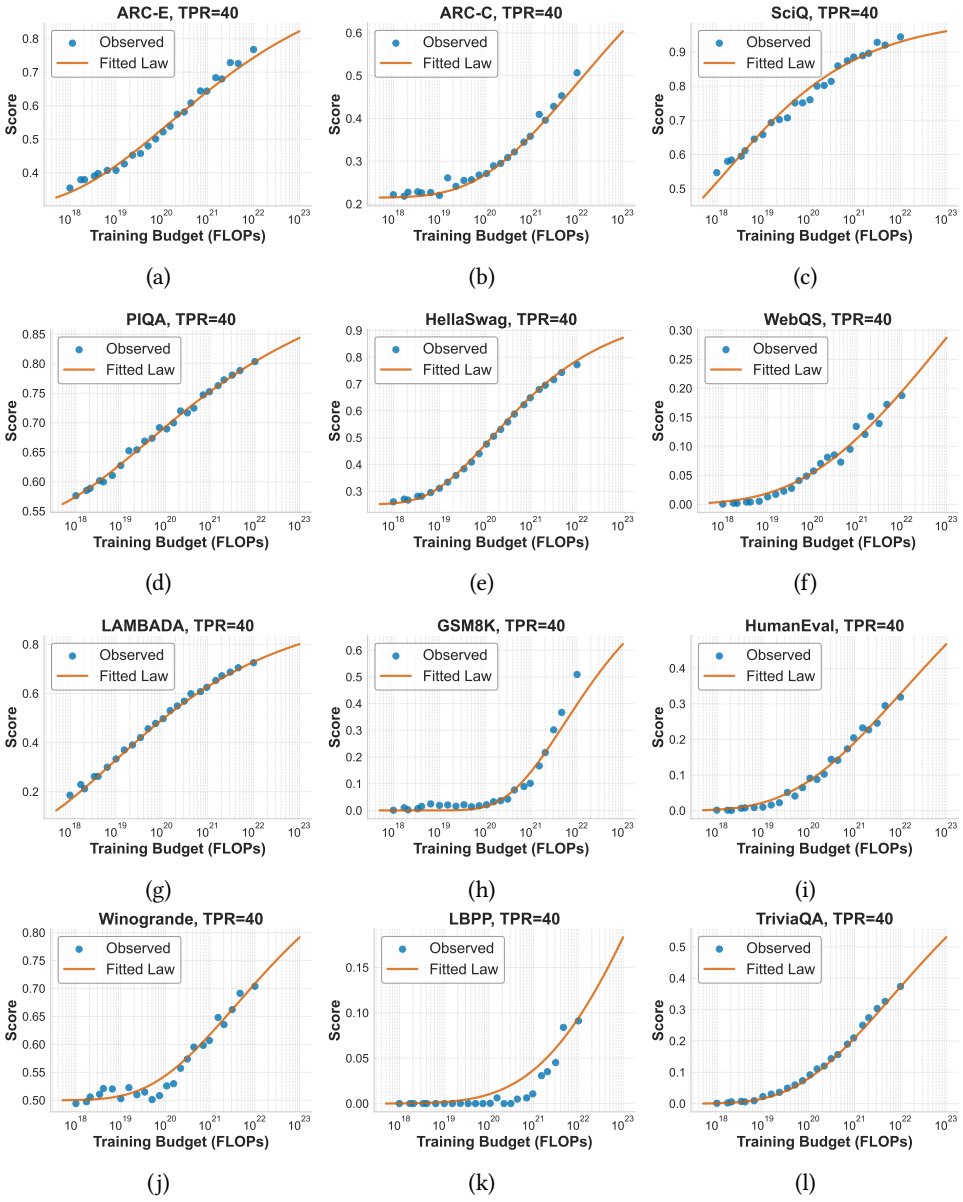



Figure 14: Fit of Eq. (4) for Token to Param Ratio 40.

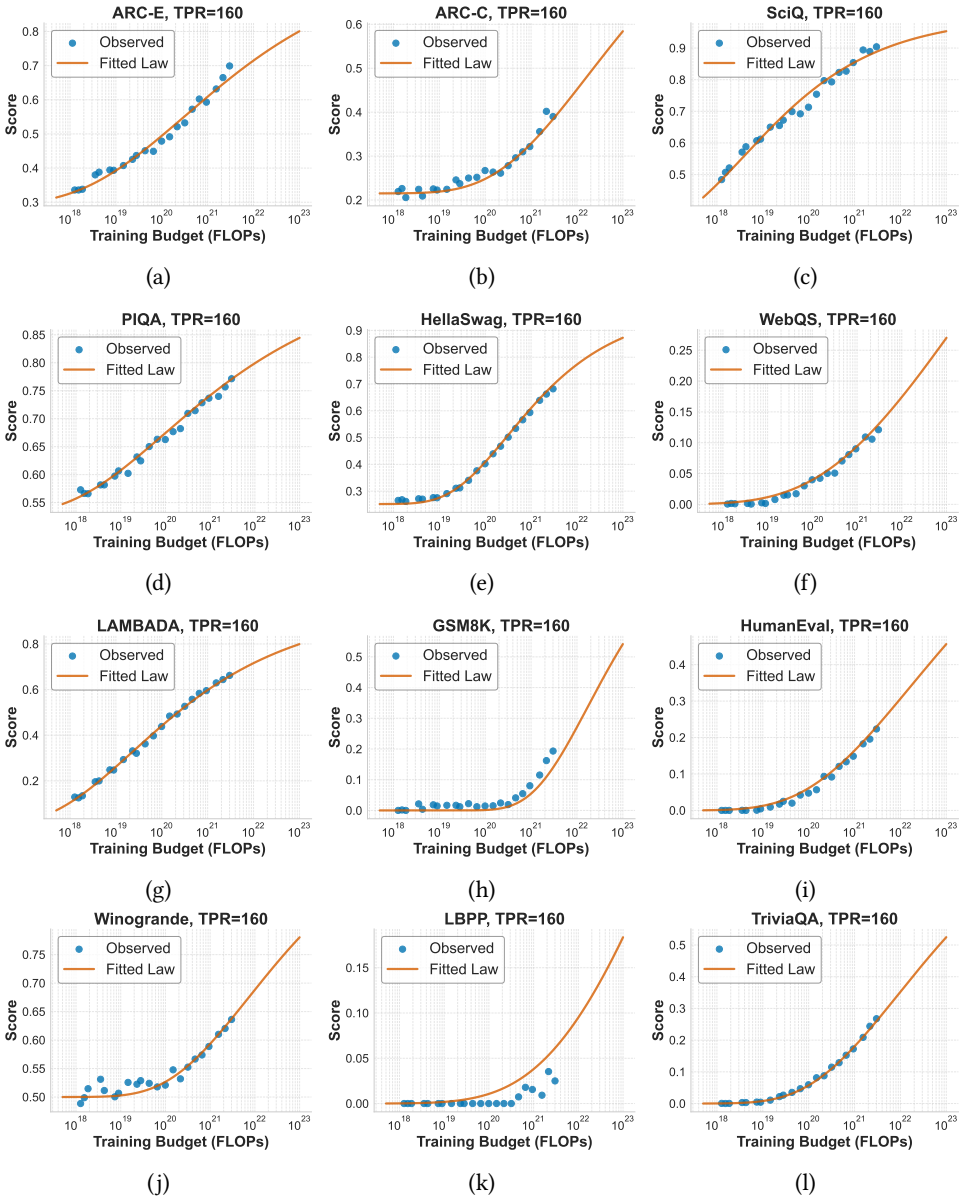



Figure 15: Fit of Eq. (4) for Token to Param Ratio 160.