

---

# A Hitchhiker’s Guide to Scaling Law Estimation

---

Leshem Choshen<sup>1 2 3</sup> Yang Zhang<sup>2 3</sup> Jacob Andreas<sup>1</sup>

## Abstract

Scaling laws predict the loss of a target language model by extrapolating from easier-to-train models with fewer parameters or smaller training sets. This provides an efficient way for practitioners and researchers alike to choose optimizers, datasets, and model architectures. Despite the widespread use of scaling laws to model the dynamics of language model training, there has been little work on understanding how to best estimate and interpret them. We collect (and release) a large-scale dataset containing losses and downstream evaluations for 485 previously published pretrained models. We use these to estimate more than 1,000 scaling laws, then derive a set of best practices for estimating scaling laws in new model families. We find that fitting scaling laws to intermediate checkpoints of training runs (and not just their final losses) substantially improves accuracy, and that—all else equal—estimates of performance are generally most accurate when derived from other models of similar sizes. However, because there is a significant degree of variability across model seeds, training multiple small models is sometimes more useful than training a single large one. Moreover, while different model families differ in scaling behavior, they are often similar enough that a target model’s behavior can be predicted from a single model with the same architecture, along with scaling parameter estimates derived from other model families.<sup>1</sup>

## 1. Introduction

Substantial effort and cost are required to train even a single large language model (LLM), and even greater cost and effort are required to evaluate proposed changes to language

---

<sup>1</sup>MIT <sup>2</sup>MIT-IBM Watson AI Lab <sup>3</sup>IBM Research. Correspondence to: Leshem Choshen <leshem.choshen@mit.edu>.

*Proceedings of the 42<sup>nd</sup> International Conference on Machine Learning*, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

<sup>1</sup>See our [repository](#) for code, data, and experimental results.

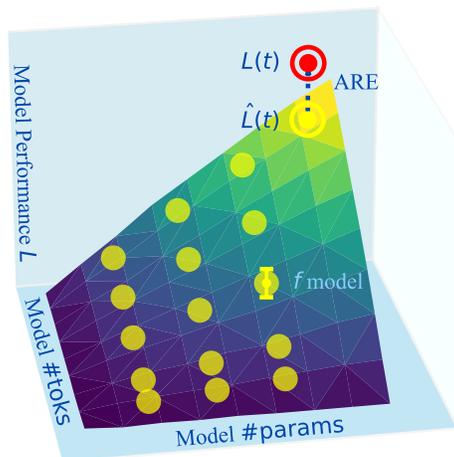


Figure 1: Illustration of a scaled family, an estimated scaling law, and its prediction error for a target model.

models’ architecture or training data. There is thus an acute need for efficient decision-making aids that can evaluate new methods without full-scale training. A large body of work motivates or evaluates these changes using small models (Warstadt et al., 2023; Hu et al., 2024a; Hillier et al., 2024), synthetic tasks (Akyürek et al., 2024; Wortsman et al., 2023) or theory (Jelassi et al., 2024). But one of the most important tools for current practitioners is the estimation of **scaling laws** for LLMs (Ivgy et al., 2022; Dubey et al., 2024).

A scaling law extrapolates the performance of a target model from the performance of a set of models with fewer parameters or smaller training sets. Typically, this requires models to belong to the same *model family*, differing only in parameter count and training size, but using the same architecture and training distribution. A high-quality scaling law accurately predicts the target model’s test performance (Rosenfeld et al.; Kaplan et al., 2020; Hoffmann et al., 2022). Most past work fixes a model family and exhaustively trains models to relate scale and performance through a new scaling law. One question that has received comparatively little attention is how to create such laws in the first place.

This paper offers a practical guide to when, how and **which small models** to use, to efficiently obtain meaningful predictions about large models’ behavior—maximizing prediction reliability while minimizing the budget for preliminary ex-

perimentation, which necessarily involves tradeoffs between the number of preliminary models trained, the size of the largest preliminary model, and size of the dataset used to train it. To answer those questions one requires analysis **across model families** and scaling law procedures.

We begin by collecting data to perform a large-scale meta-analysis of scaling laws (§3). Usually, scaling law research relies on a single collection of closely related models, or alters only a minimal aspect of pretraining (e.g. data size; Muennighoff et al., 2024). We instead gather a large and diverse set of scaled families, to allow this and future meta-analysis of scaling laws that generalize across architectures, datasets and settings.

The rest of the paper uses this data to analyze a number of key questions around scaling law estimation:

1. **What reliability is achievable and expected from scaling laws?** Variation between LLM initializations produce unpredictable changes of up to 4% in loss. Most published controlled experiments on pretraining decisions, report changes between 4% and 50% (§4).
2. **How much does the shape of scaling laws vary across model families?** Different model families have scaling laws with a different functional dependence on model size. However, transformer LLMs are similar enough that, with a single model from a target family and a scaling law from a different model family, it is sometimes possible to accurately estimate target model performance. (§5).
3. **Must scaling laws be estimated only from fully trained models?** Even though optimization procedures are typically sensitive to the full size of a training run, (unprincipled) estimation of scaling laws from intermediate training checkpoints greatly improves scaling law fit (§6). It is generally possible to estimate a model’s final loss beginning roughly  $1/3$  of the way through training.
4. **How large must models be to produce reliable scaling laws?** All else equal, experimenting with large models is typically more useful than with small models (§7), but may be outweighed by the benefits of reduced variance from training more, smaller models (§8).
5. Taken together, **cost-effective estimation** of a scaling law should consider the **number** of models, the **size** of the models, and the number of **training tokens** for each model. We highlight those size, tokens and number of models effects in Fig. 2.

Our experiments also provide insight into the functional form of scaling laws themselves, suggesting that they may have fewer degrees of freedom than typically assumed (§9). We conclude with discussion of other work on scaling law estimation that may be of interest to practitioners (§10).

## 2. Defining a scaling law

A scaling law estimates the loss of a costly model by training cheaper ones (see Fig. 1) typically with fewer parameters ( $\#\text{params}$ ) and training tokens ( $\#\text{toks}$ ). A scaling law is a function that predicts a target model’s loss on held-out data when setting the value of one hyperparameter (Kaplan et al., 2020) or both (Rosenfeld et al.; Hoffmann et al., 2022). Comparing laws’ predictions about different pretraining choices (e.g. data Ge et al., 2024) allows informed decisions about which large-scale model to train. A scaling law also enables finding the optimal choice of hyperparameters under computational constraints on pretraining (Hoffmann et al., 2022) or inference (Touvron et al., 2023; Sardana et al.).

Formally, we call a **model**  $f$  any single neural language model with a specific set of parameters. Different seeds, or even different checkpoints from the same training run, correspond to different models. We define a **scaled model family**  $F$  as a set of models, with each  $f \in F$  differing only in size  $\#\text{params}(f)$  and number of tokens  $\#\text{toks}(f)$ . We note that a change in size is usually applied in a systematic manner that affects the number of attention heads, width, depth and such network characteristics that in all our data changes with it in a one-to-one mapping.

Two subsets of scaled model families will be especially important in our analysis. First, the **maximal parameter family**  $F_{\text{max}P}$  contains only models with the largest number of parameters. Formally, define  $m = \max_{f \in F} \#\text{params}(f)$ ; then  $F_{\text{max}P} = \{f \in F : \#\text{params}(f) = m\}$ .  $F_{\text{max}P}$  will generally contain the **target** model(s) whose behavior we wish to predict  $t \in F_{\text{target}}$ . Second, the **q-maximal token family**  $F_{\#\text{tok} > q}$  contains all models trained on at least a  $q$ -sized fraction of the training set. Formally, define  $t = q \cdot (\max_{f \in F} \#\text{toks}(f))$ ; then  $F_{\#\text{tok} > q} = \{f \in F : \#\text{toks}(f) \geq t\}$ . Note that this definition does not distinguish between partially trained models, and models trained to convergence on a subset of the training set. These two types of models generally differ, but as current theory does not predict in what manner and the former are cheap substitutes, we test empirically if those are similar enough to make good predictions in Section 6.

A **scaling law**  $\hat{L}(f | F)$  estimates the performance of a new model  $f$  given a model family  $F$ . (We will simply write  $\hat{L}(f)$  when the family is clear from the context.) All experiments in this paper use the widely used functional form proposed by Hoffmann et al. (2022):

$$\hat{L}(f) := e^E + \frac{e^A}{\#\text{params}(f)^\alpha} + \frac{e^B}{\#\text{toks}(f)^\beta}. \quad (1)$$

Here  $E$  captures the scaled family’s general performance;  $A$ ,  $\alpha$  and  $B$ ,  $\beta$  describe the scaling effect of  $\#\text{params}$  and  $\#\text{toks}$  respectively. The parameters are in an exponent to ensure positivity, i.e., more training data improves the

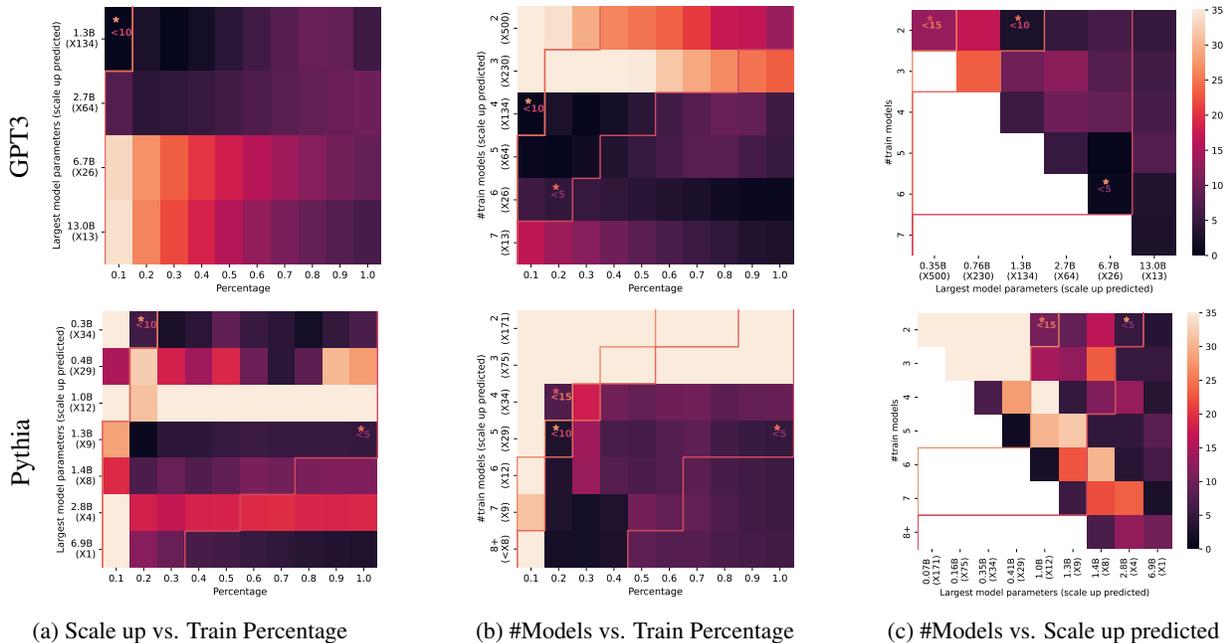


Figure 2: The effects of 3 variables on scaling law accuracy. Each cell corresponds to a single scaling law estimated from a set of model checkpoints  $F_{\text{train}}$ , with the color denoting the error when predicting the largest model. Each column shows a subset of the three axes along which these training sets differ: (1) the number of tokens used to train each LM in  $F_{\text{train}}$  (expressed as a fraction of the full training corpus), (2) the number of distinct models trained; and (3) the size of the largest model trained (expressed as a scale-up factor—the ratio between the target model and the largest model in  $F_{\text{train}}$ ). In (a), all laws are estimated from four models. In (c) all laws use the full corpus. Orange lines show iso-FLOP contours (sets of scaling laws whose training sets require the same computational cost to produce). ★ represent the most efficient ways to obtain 15%, 10% and 5% ARE. One of the most immediate conclusions from these plots is that scaling law estimation is quite noisy—the inclusion of a single badly-behaved model in the estimation procedure can produce large errors, and in small model families error does not reliably decrease with additional computation. However—because of noise—it is often preferable to extrapolate from a large number of small, partially trained models rather than a small number of large models.

scaling.<sup>2</sup> These parameters are estimated by first collecting a set of **training models**  $F_{\text{train}}$ , then minimizing the reconstruction error

$$\arg \min_{E, A, \alpha, B, \beta} \sum_{f \in F_{\text{train}}} (\hat{L}(f) - L(f))^2$$

where  $L(f)$  denotes the empirical negative log-likelihood of some held-out data under the model  $f$ .

In this sense, a scaling law is an ordinary parametric model, and we may ask many of the questions about  $\hat{L}$  that we ask about LLMs themselves—what training data ( $F_{\text{train}}$ ) to collect? How to estimate accuracy? However, to provide empirical answers to these questions, we first require data.

<sup>2</sup>We believe many of our findings apply to other functional forms (Caballero et al.), and even suggest new ones in §9.

### 3. Data for 1000+ scaling laws and more

As part of this work, we have collected and released the largest-scale public dataset describing scaling behavior across model families. This dataset aggregates information from a large number of LLM training efforts that have released information about the behavior of multiple models of different sizes or scales. While experiments in this paper focus on scaling laws that measure loss, the dataset also includes information about model performance on downstream evaluation benchmarks where available. We have focused on language models where the largest one is more than 3B parameters and where data was shared publicly or in private correspondence. Our [repository](#) accepts further contributions and requests for additions. In addition to those, we have manually extracted some data from papers that did not release models but reported losses in figures.

**Other data sources.** We want to highlight other sources for data on model results. Resources on the training check-

points and dynamics are scarce and perhaps the only other collection of such will be in the data-limited babyLM models of 2025 (Charpentier et al., 2025). There are some efforts to collect large sets of downstream evaluations for models that have been useful in other works. Those include DoVE, which aims to collect all LLM evaluations (Habba et al., 2025), and data collected to create observational scaling laws (Maia Polo et al., 2024; Ruan et al., 2024).

### 3.1. Data sources

For each model in this dataset, we collect any downstream evaluation and loss during training that was reported, as well as #toks for each, links to matching checkpoints when available, links to data sources, and information about computational cost (in FLOPs) and number of training epochs (i.e. passes over the training set). Each model is identified by a unique name, a type (e.g. llama), #toks, #params, architecture type (e.g. encoder-decoder), and seed.

Models (families) in this dataset include Pythia (Biderman et al., 2023), OPT (Zhang et al., 2022, collected thanks to Xia et al., 2023; Biderman et al., 2023), OLMO (Groeneveld et al., 2024), Amber (Liu et al., 2023), K2 (LLM360 Team, 2024), Mamba (Liu et al., 2023), RedPajamas(Together, 2023)ModuleFormer mixture of experts (Shen et al., 2023), overtrained models (Gadre et al., 2024), Mamba, Llama and hybrid architecture variations from Poli et al. (2024), transformer architectures (Alabdulmohsin et al., 2022), Bloom (Le Scao et al., 2023), T5-Pile (Sutawika et al., 2024), Pandey (2024) models, GPT-family models with different data regimes (Muennighoff et al., 2024), Gopher (Hoffmann et al., 2022) and GPT-3 (Brown et al., 2020).

The data consists of 1.9M steps of training evaluated on loss or perplexity, usually on multiple data sources belonging to 485 unique pretrained models, and more than 40 scaled families. We hope this will provide a useful resource for the community and plan to extend it further as models get released and their training dynamics are shared. We see such a resource as a facilitator to more research on model development (e.g. A/B testing), scaling laws, downstream scaling laws (Gadre et al., 2024; Ruan et al., 2024; Owen, 2024; Isik et al., 2024), training dynamics (Choshen et al., 2022) and more.

### 3.2. Scaling law estimation

In the rest of the paper, we present findings from estimating hundreds of scaling laws as follows:

**Fitting** For each model family  $F$ , we identify the maximal parameter family  $F_{\max P}$ , and estimate a scaling law  $\hat{L}$  using the remaining models  $F_{\text{train}} = F \setminus F_{\max P}$ . Estimation of scaling law parameters uses the `curve_fit` function in `scikit-learn` (Pedregosa et al., 2011), with square loss.

We additionally experimented with an L-BFGS-based solver but found it to be less stable. It converged to similar results, but often did not converge, was slow and required multiple trials. Some past work has Huber loss to improve the robustness of estimates; we repeat the analysis from the main paper with Huber loss in §E and find the same trends as in our main analysis. We only estimate scaling laws for families that contain at least three models.

**Evaluation** To evaluate estimated scaling laws reliably, we need to account for loss fluctuations during large-scale model training. Thus, we test against a few checkpoints near the end of training: we choose as target models  $F_{\text{target}}$  the 30%-maximal token family from the set  $F_{\#toks>30\%}$  defined in the previous paragraph—that is, we take  $F_{\text{target}} = F_{P, \#toks>30\%}$ . We then report the mean **absolute relative error (ARE)**  $\mathbb{E}_{f \in F_{\text{target}}} |L(f) - \hat{L}(f | F_{\text{train}})| / L(f)$  between the empirical loss  $L$  and the loss  $\hat{L}$  predicted by the scaling law.

## 4. How well should scaling laws predict?

*4% is the best ARE typically obtained; ARE up to 20% can still distinguish between many modeling choices.*

To establish how accurate a scaling law must be to be *useful* to practitioners, we first assess what changes in model accuracy have been considered meaningful in past work. We have surveyed experiments in the literature where an A/B test was performed, i.e., two models were trained similarly, manipulating one attribute to see how it affects scores. Empirically, we found no widely adopted modeling changes that were motivated with less than a 4% relative difference between models. Additionally, reported variance across random restarts of the same model architecture reaches up to 3.5% (c.f., §8; Sellam et al., 2021). We take this to mean that this is approximately the minimal effect size experimenters care, as well the minimal effect that can be reliably measured. Accordingly, this bounds the best goodness of fit we should expect or require of scaling laws.

To offer several concrete points of comparison: Pythia 6.9B models fixed inconsistencies in their code and hence have two versions (c.f. App. B; Biderman et al., 2023) which differ in loss by 40%. They also provide a data deduplication A/B test that decreased loss by roughly 5%. Gadre et al. (2024) tested the effect of training 400M parameter models for different #toks. The smallest modification (doubling the number of training tokens) decreased loss by roughly 4%, while 30× more training tokens produced a 50% loss difference. Instead of varying the amount of data or epochs, Ge et al. (2024) found that training on a different kind of data incurred ARE of approximately 10% and different data mixes led to 6% changes or less.

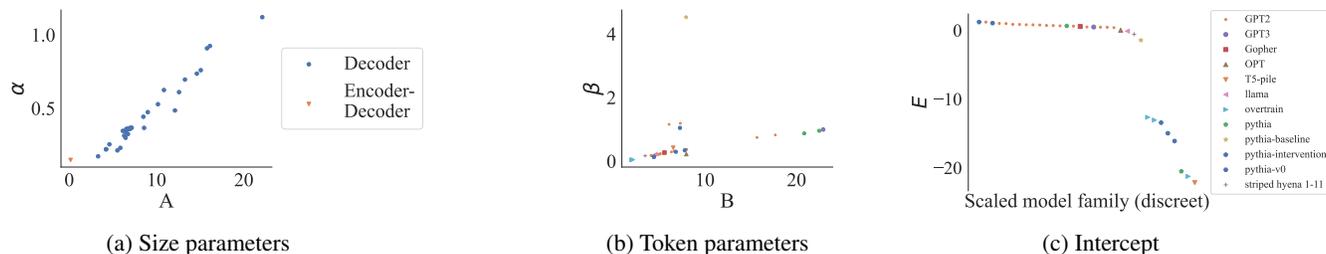


Figure 3: Parameters differ between scaled model families. Surprisingly, however, the pairs of parameters controlling the influence of model and training set size have similar ratios. The legend shows model architecture (left), scaling families (center) and per-family intercept (right).

## 5. When I train a new model, do I even need a new scaling law?

*Different model families exhibit different scaling behavior, but performance can sometimes be estimated using a single model in a new family.*

Scaling laws relate performance to *scalar* training parameters like model or dataset size. For *discrete* decisions (e.g. the choice of nonlinearity), it is unclear how to pool information across models that differ in these traits (see Ruan et al., 2024; Maia Polo et al., 2024, for concurrent work that performs this pooling based on downstream task behavior). Clearly, discrete choices affect loss of pretrained models with the same  $\#\text{params}$  and  $\#\text{toks}$ . But how do they affect the form of scaling laws?

One way to answer this question is to look at the parameter estimates for scaling law parameters  $E$ ,  $\alpha$ ,  $A$ ,  $\beta$  and  $B$  differ across model families. These results are shown in Figure 3, where it can be seen that there are often dramatic differences in all five parameters across families. In this sense, even the rate at which additional data or parameters improve model performance depend on underlying architectural details, suggesting that understanding the behavior of a new model family may require a new scaling law.

But another way to answer this is to ask how reliably we can predict final model accuracy when borrowing (or pooling) some parameters of scaling laws between families—even if these result in poor parameter estimates, they may predict large-scale model behavior within the range of meaningful differences identified in Section 4. To do so, we set the  $\#\text{params}$  scaling parameters ( $A$ ,  $\alpha$ ) to fixed values reported in past work, and estimate remaining parameters for individual model families. We take the variable values found by Muennighoff et al. (2024) (see Besiroglu et al., 2024; Porian et al., 2024 for a discussion of estimates from earlier work including Hoffmann et al., 2022). We find (see Fig. 6 in App. A) that in some cases only a single training run in a new model family is necessary to obtain accurate scaling law predictions. In the OLMo family, for example, we obtain

less than 1% error estimating the accuracy of a 7B model from a collection of 1B model checkpoints. We find that predictions generalize, and a constant  $\#\text{params}$  scaling factor is enough for most models (except the encoder-decoder T5-Pile). However, error rates are larger than in the source family, and predictions for larger models are worse (most conspicuous in OPT’s error of 37%, 25% and 15% when extrapolating from 8.7B, 13B and 30B to 175B).

### 5.1. Can I train the target model a bit instead of many small models?

*Yes, but obtaining reliable estimates in this way requires up to 30% of the full training run.*

The above results (last row of Fig. 6 in App. A) also suggest the possibility of predicting losses not with just smaller models, but with partially trained versions of the target model itself. When predicting inside the same  $\#\text{params}$  family—that is, estimating  $\hat{L}(f \mid F_{\text{target}} \setminus \{f\})$ —the  $\#\text{params}$  term in Equation (1) is constant, and extrapolation is only required for  $\#\text{toks}$ . As seen in the figures, this form of estimation is informative if permitted by computational constraints. Beyond the immediate usefulness of this approach, it is a promising avenue for future research on alternatives to scaling the number of layers.

### 5.2. Are even simpler baselines enough?

*Some extrapolation is necessary: scaling laws can produce accurate estimates even when the target model vastly outperforms any training model.*

To provide another form of comparison for the predicted scaling laws, we compute two baselines. Both baselines adopt a pessimistic evaluation assuming that the target model is no better than the best model in the small model family used to estimate a scaling law. Specifically, the baselines are the *best performance*  $\hat{L}(\emptyset \mid F_{\text{train}}) = \min_{f \in F_{\text{train}}} L(f)$  and the performance of the *most-trained model*, consuming the most compute for train-

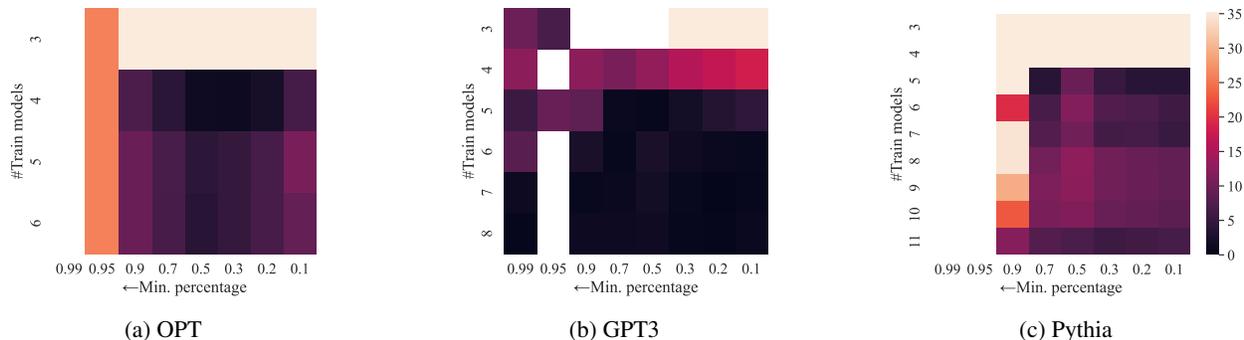


Figure 4: The effect of fitting on more of the training trajectory. Each cell represents the absolute relative error estimating scaling laws from a given number of models (vertical axis) trained on a subset of the *final* checkpoints from a training run (so scaling laws on the left are estimated using all checkpoints, and on the right using only the final 10% of checkpoints). White cells failed to fit. As long as the first  $\approx 10\%$  of checkpoints are discarded, final loss can often be predicted accurately.

ing, i.e.  $\hat{L}(\emptyset | F_{\text{train}}) = \arg \max_{f \in F_{\text{train}}} \# \text{params}(f) \times \# \text{toks}(f)$ . Those baselines might be the best one can expect without fitting a law to scaling.

We find (See App. 5.2) that the *best performance* baseline is closer to  $L(F_{\text{target}})$ , which is to be expected, as the target model performance is better than any other model in  $F$  and this is the better of the two. In both cases, even with the full  $F$ , the baselines suffer more than 15% error, mostly above 10%, almost never get below 5%, and 18% ARE on average across all scaled families we study.

## 6. I have some data; what portions should I use?

*Estimate scaling laws from intermediate checkpoints, not just fully trained models!*

Most past work on scaling behavior of language models (e.g., Gadre et al., 2024; Muennighoff et al., 2024) has trained a *separate* model for each value of  $\# \text{toks}$  studied. This is based on the assumption that changes in the learning rate schedule, which depend on the size of the full dataset that will be used for training, render losses from intermediate checkpoints uninformative.

However, some recent work has demonstrated the effectiveness of learning schedules that do not require prior access to the size of the training set (Hu et al., 2024b), and some work has questioned whether careful choice of the learning rate decay is even necessary for reliable scaling laws (Porian et al., 2024). Together, these findings motivate revisiting the assumption that only a single useful datapoint may be obtained from each training run. In §5.1, we observed the value of intermediate checkpoints when only a single  $\# \text{params}$  family is used to fit a scaling law. In general, there may be differences between models trained on the same number of

tokens and sizes depending on the choice of learning rate schedule. But it is unknown whether these differences are large enough to impact the estimation of scaling laws. We now test whether this finding extends to larger families—i.e. whether including intermediate checkpoints from all models in a model family reduces ARE.

Results are shown in Figure 4, which plots ARE for scaling laws estimated from data subsets of the form  $F_{\# \text{toks} > q}$  for varying  $q$ . We find that including full training curves in scaling law estimation can predict losses well. In fact, relying merely on the end of training (left in Figure 4) produces significantly worse performance across the board. Our remaining experiments thus fit scaling laws using all these intermediate checkpoints, and not final performance alone.

### 6.1. Should I use all intermediate checkpoints?

*Almost all, but drop very early checkpoints.*

In Fig. 4, we plot ARE for different  $F_{\# \text{toks} > q}$ -maximal token families serving as  $F$ , i.e., when fitting only with the end of training runs. There is not a clear trend indicating whether we should use all data (as might be suggested by GPT-3 results alone) or only some of it. But it is rarely the case that best estimates are obtained from the end of training alone.

There is, however, a distinctly uninformative phase at the beginning of training, as can be seen in the loss curves (App. B) and noted in the literature (e.g., Chen et al.). We observe that this period is more likely to contain significant spikes or an increase in loss (worse performance) despite additional training. We hence hypothesize this part should always be removed from data used to estimate scaling laws.

Indeed, our experiments depicted in Fig. 5 compare scaling law AREs with and without including models trained on less than 10B tokens in  $F$ . Evidently, the very beginning

of training (often not even reported in logs and graphs) is sometimes harmful to the prediction. Specifically, we run the same experiments with and without ignoring the first 10B tokens seen. We find that for some models (e.g., OPT and Pythia) the ARE exceeds 15% even when using the whole data, but drops to 4-10% when ignoring those tokens. In preliminary experiments, we found that cutting fewer tokens gave noisier results, and cutting more had a negligible effect.

### 7. How big a model should I train?

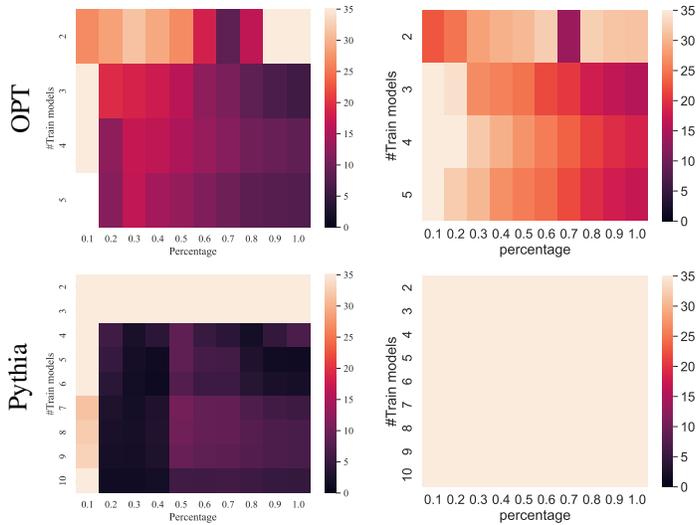
*Larger models are better, but not necessary. Mainly, beware of specific models that might give noisy results.*

In Figure 2 we compare scaling laws when controlling the amount, percentage, or size of the models (2 at a time). We find that choosing models closer in #params to the target model is generally effective (e.g., Fig. 2a, 2c), but the effect is neither strong nor monotonic. For example, in all cases fitting on all  $F$  provides one of the lowest ARE. However, in GPT, Gopher and OPT, predicting with the smallest 4 models available is already enough to achieve less than 10% error. In Pythia, the smallest models are not predictive but the rest of the models provide a similar fit. While relying on a larger model is beneficial, predicting many scales up (e.g., the behavior of a  $34\times$  larger model in Pythia) is still reliable, especially if accounting for other factors we discuss next.

In fact, training additional, larger models before fitting a scaling law may sometimes decrease accuracy due to increased variance in large model performance—see, for example, Pythia 2.8B in Fig. 1. Unfortunately, it is difficult to identify whether a seed is exceptionally high or low-performing without additional information. For example, cross-validation on  $F$  fails to detect it (see App. D).

Instead, this instability can be addressed by accounting for seed variability. A wasteful way to do so would be to train every model several times. A better alternative is to diversify and train each model on as differing hyperparameters as possible and to maximize the information gained (a common practice in efficiency-coverage scenarios, e.g., Perlitz et al., 2024). Hence, we suggest training more models of differing sizes each accounting for both size and seed changes, rather than training multiple seeds. We further discuss the effects of number of models ( $|F|$ ) in §8.

Given the choice of the largest model and the number of models, it is unclear how to optimally space the model sizes, whether linearly, log-scale, or otherwise. We leave that optimization problem for future work.



(a) Cut 10B first tokens (b) Fit all data

Figure 5: The effect of fitting with/without the beginning 10B tokens seen. Each cell represents the absolute relative error when estimating a scaling law from a given number of models (vertical axis) trained on a given subset of checkpoints from the beginning of training (horizontal axis).

### 8. How many models for reliable predictions?

*5 models is a safe bet, more would improve the results’ robustness. These models can be small.*

We have seen that predicting with larger models and hence extrapolating less yields better results. However, given compute constraints (and additional hardware constraints like memory), practitioners may generally wish to use smaller models when possible. Consider for example Fig. 2b where we compare fitting on 4 models but vary their size. We find that *more* models reduce ARE even without being *big-ger* models. As discussed in §7, adding a larger model to a current scaled family serves two goals, it increases the proximity to the predicted model, as well as increases the number of models seen.

We separate the contribution of size and number of models. In Fig. 2c, we predict with the largest model being held constant and add (at minimal cost) smaller models. We see again that larger models do benefit predictions. For example, the small models part (left) of the graph indicates large errors (bright). However, we also see again the unwanted effects a single model may have on the overall prediction. Consider for example the figure’s diagonal in Pythia. Cells in a diagonal share a group of models and each row adds another one to  $F$ . Evidently this specific group hurts results, even when larger models are added to  $F$ . With enough models (bottom of diagonal), the negative decreases. Switching the model (next column) also removes the negative effect.

Moreover, across all rows the tendency is never monotonic, implying larger models do not ensure better predictions.

But in general, we see that increasing the number of models tends to improve prediction. For example, in GPT3 the best predictions are with many models. Perhaps intuitively, adding a larger model and improving both #params and number of models aspects improves quite consistently (Fig. 2b and diagonals of Fig. 2c).

## 9. Are all scaling law parameters crucial?

*Scaling laws might have fewer degrees of freedom than described in the literature.*

Assuming we do not try to account for aspects other than #toks and #params (see §10), one might wonder if some of the observed errors come from model misspecification—an incorrect functional form for  $\hat{L}$ , which (with a small number of exceptions including Caballero et al.) has generally gone uncontested since it was first proposed (Rosenfeld et al.; Hoffmann et al., 2022). Here we specifically evaluate whether scaling laws empirically exhibit fewer degrees of freedom than has been proposed. First, we compute the principal components of the 5 learned parameters and find that 3 components explain 99.49% of the variance between the 5 parameters. Inspection reveals that two of these components tightly couple the pairs of parameters dealing with the same training parameter (#params and #toks). Plotting values of  $A$  against  $\alpha$  and of  $B$  against  $\beta$  (Fig. 3), we see a clear linear relationship between these variables despite their non-linear interaction in Eq. 1. There are a few exceptions: the Encoder-Decoder model T5-Pile shows a different behavior from the rest of the scaled families, and four additional scaled families show a different relationship between  $B$  and  $\beta$ . In fact, all these families share the common feature that they were trained using multiple passes over a single training set Gadre et al. (2024). The outlier point with  $\beta > 4$  is a 70m baseline of Pythia for a continual training intervention experiment (Biderman et al., 2023). Future work may consider different function forms tying some of the parameters or introducing other ones instead.

Another function form change that future work should consider is accounting for the learning rate schedule, as our experiments assumed it was negligible, and previous works disregarded the training trajectory. A mismatch between the form and the real dependence might explain the inconsistencies in using the beginning of training. As noted in §6.1, the beginning is not fitting as well as later on.

## 10. Related work

This work builds on a large number of recent studies relating scaling law estimation and decision-making about

model training. Among the aspects studied are total training costs including inference (Sardana et al.), effects of sophisticated data selection (Sorscher et al., 2022; Ge et al., 2024), training time (Inbar & Sernau, 2024), transfer of learned skills (Hernandez et al., 2021), behavior of models in other modalities (Mikami et al., 2022; Abnar et al.; Al-abdulmohsin et al., 2024; Hesslow et al., 2022) mixtures of experts (Ludziejewski et al.), data mixing (Ge et al., 2024), downstream performance (Muennighoff et al., 2024), vocabulary size (Tao et al., 2024), and architecture comparisons (Tay et al., 2023; Poli et al., 2024) including small models (Muckatira et al., 2024) or other phenomena like finetuning (Zhang et al.) and the loss in different positions in the training sequences (Xiong et al., 2024). Especially relevant to our context is Ruan et al. (2024); Maia Polo et al. (2024) that rely on multiple pretraining settings for creating scaling laws that generalize across models or kinds of losses.

Another line of works that can be seen as a scaling law discusses the relation between model width and hyperparameter choices (rather than loss) (Yang et al., 2022; 2021; Blake et al., 2024; Lingle, 2024).

## 11. Limitations

Our primary metric, ARE, does not distinguish between over- or under-estimation of performance. When using scaling laws to choose between candidate models to train, these error estimates may be unnecessarily conservative (e.g. if both families’ laws are biased in the same direction).

Another difficulty is aggregating information across model families. As most published families evaluate models of incomparable scales, often over incomparable ranges, we were unable to produce an informative version of Figure 2 that aggregated information across all models available, and was thus able to give general recommendations about compute-optimal choice of preliminary experiments.

## 12. Discussion

This paper provides a first study of open questions in the estimation of scaling laws and their relation to large-scale pretraining decisions. We expect that many of these conclusions could be sharpened or extended with the availability of additional information about model training, and we call on other leaders of large-scale training efforts to share training losses and evaluation results from multiple checkpoints during pretraining—even in cases where model parameters themselves cannot be released.

Our findings leave open many important questions, from efficient predictions by fitting on many model families to scaling laws of the deltas between a/b test (e.g. on optimizer choice), and to other methods that efficiently compare archi-

tures without relying on multiple models (e.g. continual learning). In addition, our results in §9 suggest other scaling law parameterizations might better fit data.

Practical recommendations

- §4 Set an estimation goal and a budget.
- §5.1 If budget allows, train the whole model for 30%.
- §A If extremely constrained, predict from one model.
- §6 Use all training losses (except the beginning).
- §7 Train as big as possible, but limit tokens.
- §7 Train more models, not just larger ones.

Impact Statement

This paper presents work whose goal is to advance LLM research. There are many potential societal consequences of our work, most of which lean towards the positive, environmental (efficient), open, and collaborative aspects. Therefore, it is hard to imagine specific issues with the progress allowed by this work.

Acknowledgments

This work was funded by the MIT–IBM Watson AI Lab, and by a Sloan Research Fellowship to JA.

References

Abnar, S., Dehghani, M., Neyshabur, B., and Sedghi, H. Exploring the limits of large scale pre-training. In *International Conference on Learning Representations*.

Akyürek, E., Wang, B., Kim, Y., and Andreas, J. In-context language learning: Architectures and algorithms. *arXiv preprint arXiv:2401.12973*, 2024.

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

Alabdulmohsin, I. M., Zhai, X., Kolesnikov, A., and Beyer, L. Getting vit in shape: Scaling laws for compute-optimal model design. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O’Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and Van Der Wal, O. Pythia: A suite for analyzing large language models across training and scaling. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International*

*Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/biderman23a.html>.

Blake, C., Eichenberg, C., Dean, J., Balles, L., Prince, L. Y., Deiseroth, B., Cruz-Salinas, A. F., Luschi, C., Weinbach, S., and Orr, D. u-mu p: The unit-scaled maximal update parametrization. *arXiv preprint arXiv:2407.17465*, 2024.

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

Caballero, E., Gupta, K., Rish, I., and Krueger, D. Broken neural scaling laws. In *The Eleventh International Conference on Learning Representations*.

Charpentier, L., Choshen, L., Cotterell, R., Gul, M. O., Hu, M., Jumelet, J., Linzen, T., Liu, J., Mueller, A., Ross, C., Shah, R. S., Warstadt, A., Wilcox, E., and Williams, A. BabyLM turns 3: Call for papers for the 2025 babyLM workshop, 2025. URL <https://arxiv.org/abs/2502.10645>.

Chen, A., Shwartz-Ziv, R., Cho, K., Leavitt, M. L., and Saphra, N. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in mlms. In *The Twelfth International Conference on Learning Representations*.

Choshen, L., Hacoheh, G., Weinsall, D., and Abend, O. The grammar-learning trajectories of neural language models. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8281–8297, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.568. URL <https://aclanthology.org/2022.acl-long.568>.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Al-lonsius, D., Song, D., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova,

E., Dinan, E., Smith, E. M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Papsupuleti, M., Singh, M., Paluri, M., Kardas, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Tan, X. E., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Grattafiori, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Vaughan, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Franco, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejjia, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Wyatt, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Brinkman, E., Arcaute, E., Dunbar,

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

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

- arXiv:2403.08540*, 2024.
- Ge, C., Ma, Z., Chen, D., Li, Y., and Ding, B. Data mixing made efficient: A bivariate scaling law for language model pretraining, 2024. URL <https://arxiv.org/abs/2405.14908>.
- Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K. R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M. E., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N. A., and Hajishirzi, H. Olmo: Accelerating the science of language models, 2024. URL <https://arxiv.org/abs/2402.00838>.
- Habba, E., Arviv, O., Itzhak, I., Perlitz, Y., Bandel, E., Choshen, L., Shmueli-Scheuer, M., and Stanovsky, G. Dove: A large-scale multi-dimensional predictions dataset towards meaningful llm evaluation, 2025. URL <https://arxiv.org/abs/2503.01622>.
- Hernandez, D., Kaplan, J., Henighan, T., and McCandlish, S. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.
- Hesslow, D., Zanichelli, N., Notin, P., Poli, I., and Marks, D. Rita: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*, 2022.
- Hillier, D., Guertler, L., Tan, C., Agrawal, P., Chen, R., and Cheng, B. Super tiny language models. *ArXiv*, abs/2405.14159, 2024. URL <https://api.semanticscholar.org/CorpusID:269982112>.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- Hu, M. Y., Mueller, A., Ross, C., Williams, A., Linzen, T., Zhuang, C., Cotterell, R., Choshen, L., Warstadt, A., and Wilcox, E. G. Findings of the second babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. *arXiv preprint arXiv:2412.05149*, 2024a.
- Hu, S., Tu, Y., Han, X., He, C., Cui, G., Long, X., Zheng, Z., Fang, Y., Huang, Y., Zhao, W., et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024b.
- Inbar, I. and Sernau, L. Time matters: Scaling laws for any budget. *arXiv preprint arXiv:2406.18922*, 2024.
- Isik, B., Ponomareva, N., Hazimeh, H., Paparas, D., Vasilvitskii, S., and Koyejo, S. Scaling laws for downstream task performance of large language models. *arXiv preprint arXiv:2402.04177*, 2024.
- Ivgi, M., Carmon, Y., and Berant, J. Scaling laws under the microscope: Predicting transformer performance from small scale experiments. *arXiv preprint arXiv:2202.06387*, 2022.
- Jelassi, S., Brandfonbrener, D., Kakade, S. M., and Malach, E. Repeat after me: Transformers are better than state space models at copying. *arXiv preprint arXiv:2402.01032*, 2024.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. Bloom: A 176b-parameter open-access multilingual language model. 2023.
- Lingle, L. A large-scale exploration of mu-transfer. *arXiv preprint arXiv:2404.05728*, 2024.
- Liu, Z., Qiao, A., Neiswanger, W., Wang, H., Tan, B., Tao, T., Li, J., Wang, Y., Sun, S., Pangarkar, O., Fan, R., Gu, Y., Miller, V., Zhuang, Y., He, G., Li, H., Koto, F., Tang, L., Ranjan, N., Shen, Z., Ren, X., Iriando, R., Mu, C., Hu, Z., Schulze, M., Nakov, P., Baldwin, T., and Xing, E. P. Llm360: Towards fully transparent open-source llms, 2023.
- LLM360 Team. Llm360 k2-65b: Scaling up fully transparent open-source llms. 2024.
- Ludziejewski, J., Krajewski, J., Adamczewski, K., Pióro, M., Krutul, M., Antoniak, S., Ciebiera, K., Król, K., Odrzygóźdź, T., Sankowski, P., et al. Scaling laws for fine-grained mixture of experts. In *Forty-first International Conference on Machine Learning*.
- Maia Polo, F., Somerstep, S., Choshen, L., Sun, Y., and Yurochkin, M. Sloth: scaling laws for llm skills to predict multi-benchmark performance across families. *arXiv preprint arXiv:2410.11840*, 2024.
- Mikami, H., Fukumizu, K., Murai, S., Suzuki, S., Kikuchi, Y., Suzuki, T., Maeda, S.-i., and Hayashi, K. A scaling law for syn2real transfer: How much is your pre-training effective? In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 477–492. Springer, 2022.

- Muckatira, S., Deshpande, V., Lialin, V., and Rumshisky, A. Emergent abilities in reduced-scale generative language models. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1242–1257, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.79. URL <https://aclanthology.org/2024.findings-naacl.79>.
- Muennighoff, N., Rush, A., Barak, B., Le Scao, T., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., and Raffel, C. A. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- Owen, D. How predictable is language model benchmark performance? *arXiv preprint arXiv:2401.04757*, 2024.
- Pandey, R. gzip predicts data-dependent scaling laws. *arXiv preprint arXiv:2405.16684*, 2024.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- Perlitz, Y., Bandel, E., Gera, A., Arviv, O., Ein-Dor, L., Shnarch, E., Slonim, N., Shmueli-Scheuer, M., and Choshen, L. Efficient benchmarking (of language models). In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2519–2536, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.139. URL <https://aclanthology.org/2024.naacl-long.139>.
- Poli, M., Thomas, A. W., Nguyen, E., Ponnusamy, P., Deiseroth, B., Kersting, K., Suzuki, T., Hie, B., Ermon, S., Ré, C., Zhang, C., and Massaroli, S. Mechanistic design and scaling of hybrid architectures, 2024. URL <https://arxiv.org/abs/2403.17844>.
- Porian, T., Wortsman, M., Jitsev, J., Schmidt, L., and Carmon, Y. Resolving discrepancies in compute-optimal scaling of language models. *arXiv preprint arXiv:2406.19146*, 2024.
- Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y., and Shavit, N. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations*.
- Ruan, Y., Maddison, C. J., and Hashimoto, T. Observational scaling laws and the predictability of language model performance, 2024.
- Sardana, N., Portes, J., Doubov, S., and Frankle, J. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. In *Forty-first International Conference on Machine Learning*.
- Sellam, T., Yadlowsky, S., Tenney, I., Wei, J., Saphra, N., D’Amour, A., Linzen, T., Bastings, J., Turc, I. R., Eisenstein, J., et al. The multiberts: Bert reproductions for robustness analysis. In *International Conference on Learning Representations*, 2021.
- Shen, Y., Zhang, Z., Cao, T., Tan, S., Chen, Z., and Gan, C. Moduleformer: Learning modular large language models from uncurated data. *arXiv preprint arXiv:2306.04640*, 2023.
- Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., and Morcos, A. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.
- Sutawika, L., Komatsuzaki, A., and Raffel, C. Pile-t5, 2024. URL <https://blog.eleuther.ai/pile-t5/>. Blog post.
- Tao, C., Liu, Q., Dou, L., Muennighoff, N., Wan, Z., Luo, P., Lin, M., and Wong, N. Scaling laws with vocabulary: Larger models deserve larger vocabularies. *arXiv preprint arXiv:2407.13623*, 2024.
- Tay, Y., Dehghani, M., Abnar, S., Chung, H., Fedus, W., Rao, J., Narang, S., Tran, V., Yogatama, D., and Metzler, D. Scaling laws vs model architectures: How does inductive bias influence scaling? In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12342–12364, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.825. URL <https://aclanthology.org/2023.findings-emnlp.825>.
- Together. Releasing 3b and 7b redpajama-incite family of models including base, instruction-tuned & chat models, May 2023. URL <https://www.together.ai/blog/redpajama-models-v1>.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., et al. Findings of the babylm challenge:

Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pp. 1–34, 2023. URL <https://aclanthology.org/2023.conll-babylm.1/>.

Wortsman, M., Liu, P. J., Xiao, L., Everett, K. E., Alemi, A. A., Adlam, B., Co-Reyes, J. D., Gur, I., Kumar, A., Novak, R., et al. Small-scale proxies for large-scale transformer training instabilities. In *The Twelfth International Conference on Learning Representations*, 2023.

Xia, M., Artetxe, M., Zhou, C., Lin, X. V., Pasunuru, R., Chen, D., Zettlemoyer, L., and Stoyanov, V. Training trajectories of language models across scales. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13711–13738, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.767. URL <https://aclanthology.org/2023.acl-long.767>.

Xiong, Y., Chen, X., Ye, X., Chen, H., Lin, Z., Lian, H., Niu, J., and Ding, G. Temporal scaling law for large language models. 2024. URL <https://api.semanticscholar.org/CorpusID:269449894>.

Yang, G., Hu, E., Babuschkin, I., Sidor, S., Liu, X., Farhi, D., Ryder, N., Pachocki, J., Chen, W., and Gao, J. Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems*, 34:17084–17097, 2021.

Yang, G., Hu, E. J., Babuschkin, I., Sidor, S., Liu, X., Farhi, D., Ryder, N., Pachocki, J., Chen, W., and Gao, J. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

Zhang, B., Liu, Z., Cherry, C., and Firat, O. When scaling meets llm finetuning: The effect of data, model and finetuning method. In *The Twelfth International Conference on Learning Representations*.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. Opt: Open pre-trained transformer language models, 2022. URL <https://arxiv.org/abs/2205.01068>.

## A. Scale up with 1 model

We bring errors of data from fitting from a single model on a given percentage of training to the largest model with

full training. Scaling is constant and follows the literature (Muennighoff et al., 2024) and the largest model stands as target model (so the bottom line in each figure represents predicting from the beginning of training). In parallel to this paper, an even more efficient work on predicting with 1 model was suggested, and the two should be incorporated (Maia Polo et al., 2024).

## B. Loss curves and predictions

We provide in Fig. 7 graphs of the loss during training of the target models per originating source (e.g., a paper) together with the predictions by using different percentage of the training.

## C. Is scaling working only upwards?

*No. Small models usually show consistent and predictable performance.*

Usually, one does not use a scaling law to extrapolate to a smaller model as one can just train the small model. However, under observational scaling laws, where one wants to research a phenomenon without scaling at all (Ruan et al., 2024; Maia Polo et al., 2024), or when many models were trained and one wishes to create smaller models for various reasons (Hillier et al., 2024; Warstadt et al., 2023), scaling down might prove useful. Moreover, in the context of traditional scaling laws this may act as a baseline. Such an experiment may shed another light on the number of models  $|F|$  versus their size  $\#params$ . If large models are better because they are more stable or otherwise fit laws more robustly, few models will be enough, if the number of models or scale down difference from the prediction, it will show similar behaviour to scaling up. See more in §8.

To test this we reverse the order of models and predict with the largest models the loss on the smallest models. This means that for example in the case of 3 models, we predict the smallest model’s loss and fit the scaling law relying on the 3 largest models. As before, we break the results by the percentage of training done and do not reverse it.

As shown in Fig. 8, the number of models plays an important role in fitting well and a minimum of 30-40% of the training is necessary for good fit, more than that often improves further.

## D. Can we detect bad models to fit on?

*If so, not through cross validation.*

In §7, we raise the issue of instability of scaling law predictions, with a single model vastly changing the results. We tried to see if, without knowing the ARE, we could remove

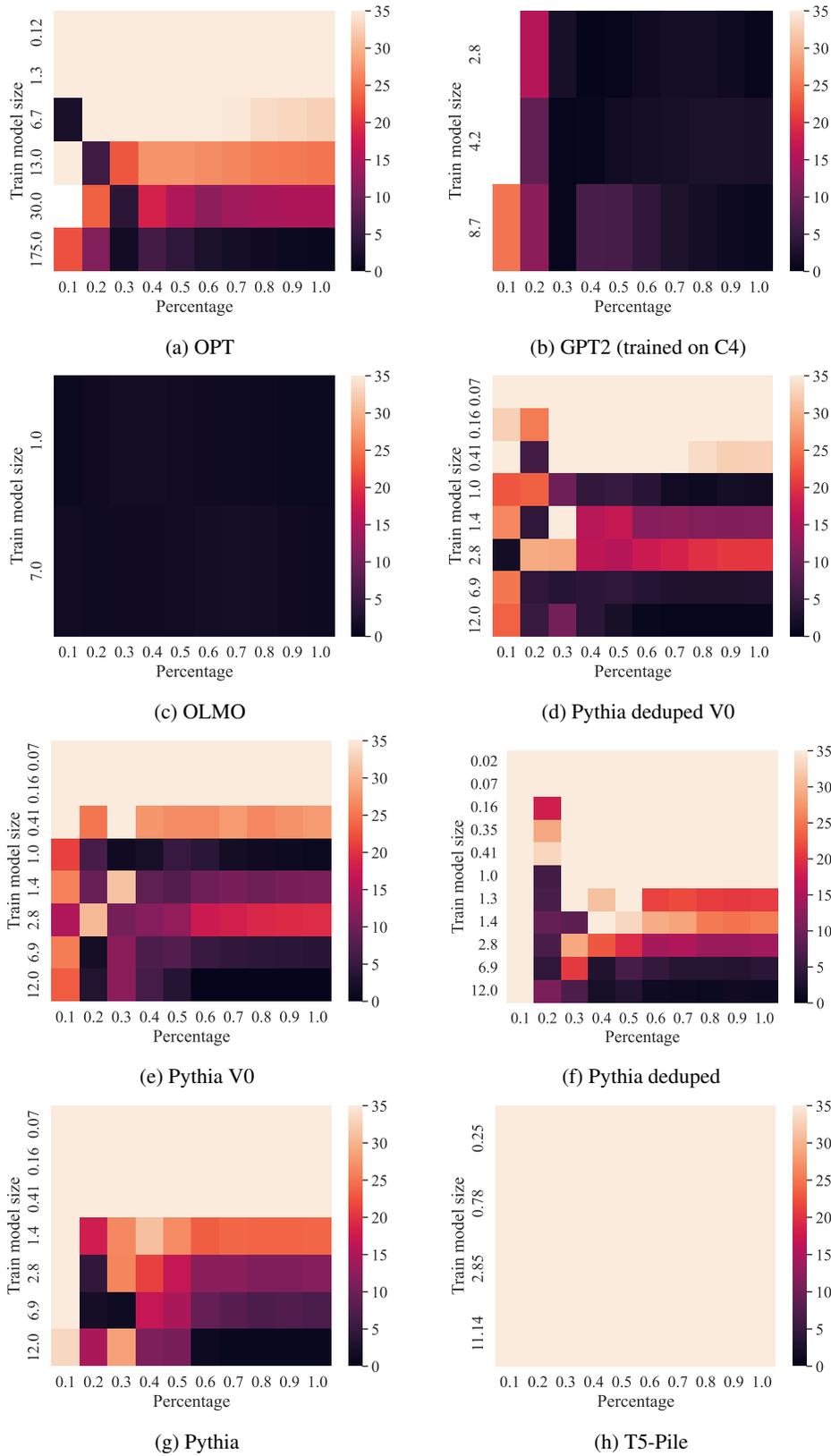


Figure 6: Fitting scaling laws under the assumption that all models scale similarly. Thus, a single model is needed to predict. The last row in each Figure represents predicting a model at the beginning of its training.

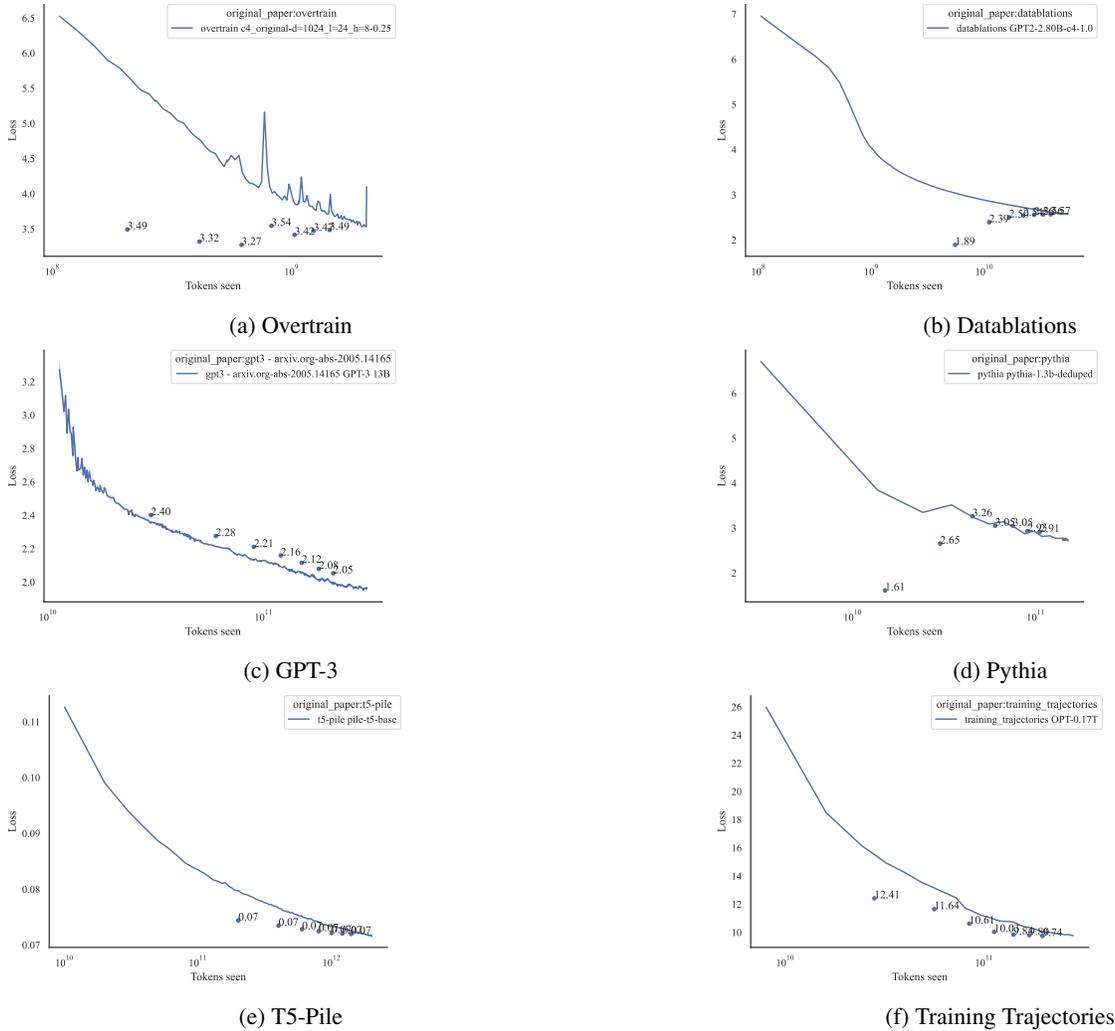


Figure 7: In each figure all losses from a specific source and predictions of the scaling loss with different percentage of the #toks and all models. Predictions are points where the X axis is the available data for prediction and y the prediction value, lines are the actual value. One scaled family per paper was sampled as a representative.

bad models from the prediction. We hypothesized that models that we can't predict would mean models that would skew our predictions when fitted upon. We performed a cross-validation on the #params families in  $F$  each time setting the models with most #toks as target ans excluding the #params family from  $F$ . Our hypothesis was found to be incorrect. Such cases of hard-to-predict models were found to indicate that the models left in  $F$  are bad predictors and not that the target is very dissimilar (a "bad" training). In 58% of the cases removing that model from the scaling law created the worst ARE possible on the actual target, more than removing any other model.

### E. Huber replication

Huber loss is sometimes used instead of ARE (Hoffmann et al., 2022). Huber loss is defined as

$$L_{\delta}(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \leq \delta, \\ \delta \cdot (|a| - \frac{1}{2}\delta), & \text{otherwise.} \end{cases}$$

We use  $\delta = \epsilon 10^{-3}$  as done in (Hoffmann et al., 2022). The overall results are similar but for completeness report them in Fig. 9.

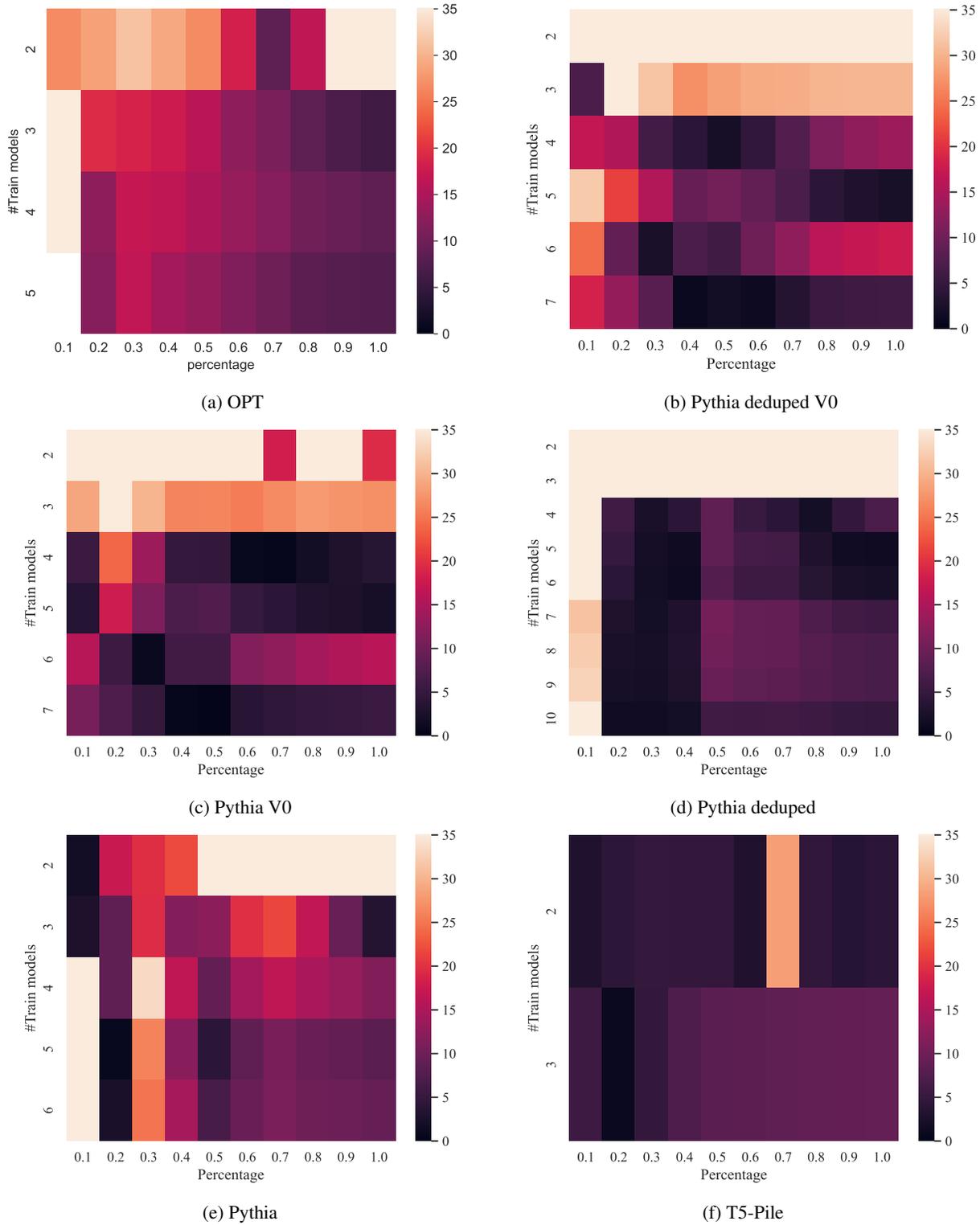


Figure 8: Fitting scaling laws trying to predict the smallest model, with the largest (Y-axis) models trained on a percentage of the data (X-axis).

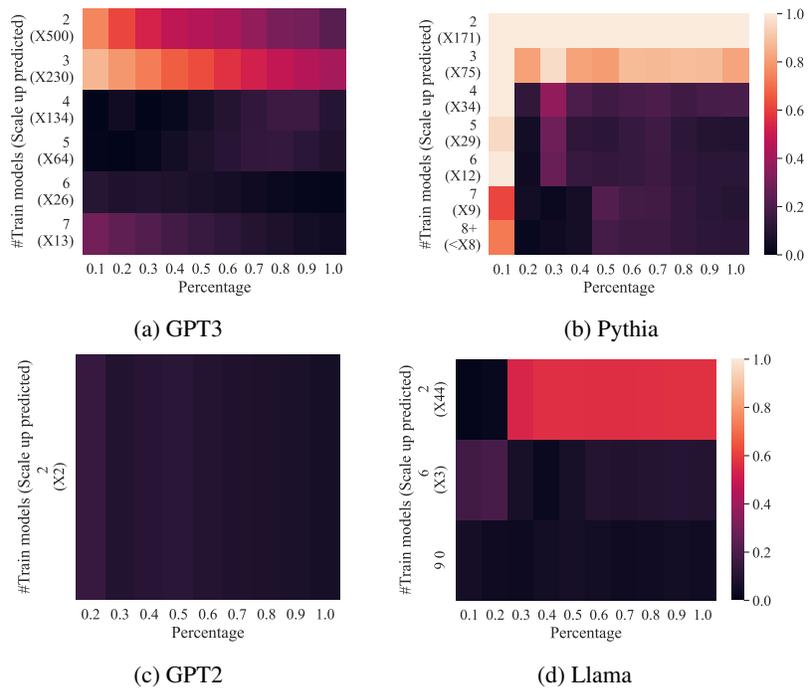


Figure 9: Scaling laws under a Huber loss. The line represents most efficient setting to recieve  $< 0.05$ . Comparison of several models given different amount of models and percentages of training.