POLYPYTHIAS: STABILITY AND OUTLIERS ACROSS FIFTY LANGUAGE MODEL PRE-TRAINING RUNS

Oskar van der Wal[*‡] University of Amsterdam
Pietro Lesci[*] University of Cambridge
Max Müller-Eberstein IT University of Copenhagen
Naomi Saphra Harvard University
Hailey Schoelkopf[‡] Anthropic
Willem Zuidema University of Amsterdam
Stella Biderman[†] EleutherAI

ABSTRACT

The stability of language model pre-training and its effects on downstream performance are still understudied. Prior work shows that the training process can yield significantly different results in response to slight variations in initial conditions, e.g., the random seed. Crucially, the research community still lacks sufficient resources and tools to systematically investigate pre-training stability, particularly for decoder-only language models. We introduce the PolyPythias, a set of 45 new training runs for the Pythia model suite: 9 new seeds across 5 model sizes, from 14M to 410M parameters, resulting in about 7k new checkpoints that we release. Using these new 45 training runs, in addition to the 5 already available, we study the effects of different initial conditions determined by the seed—i.e., parameters' initialisation and data order—on (i) downstream performance, (ii) learned linguistic representations, and (iii) emergence of training phases. In addition to common scaling behaviours, our analyses generally reveal highly consistent training dynamics across both model sizes and initial conditions. Further, the new seeds for each model allow us to identify outlier training runs and delineate their characteristics. Our findings show the potential of using these methods to predict training stability.

EleutherAI/pythia EleutherAI/PolyPythias

1 INTRODUCTION

Training deep learning models, including contemporary large-scale transformer-based language models (LMs), is an inherently stochastic process in which **randomness factors** (Pecher et al., 2024), such as the initialisation of model parameters and data shuffling, play a crucial role (Pham et al., 2020; Gundersen et al., 2022). Prior work on LM adaptation (Dodge et al., 2020; McCoy et al., 2020; Agarwal et al., 2021; Mosbach et al., 2021; *inter alia*) and pre-training (D'Amour et al., 2022; Madaan et al., 2024; Alzahrani et al., 2024; *inter alia*) shows that even slight variations in these randomness factors can lead to substantially different outcomes. Specifically, training multiple times using the same implementation, hardware, dataset, and hyperparameters can, nonetheless, lead to large deviations in the final performance. This variability in performance has significant implications, primarily because conclusions drawn from single training runs may be misleading or incomplete. Thus, a systematic investigation into the stability of LM pre-training is essential to ensure robustness, reproducibility, and trustworthiness in applications that use these models (Sellam et al., 2022; Goldfarb-Tarrant et al., 2024). The stability of LM performance to randomness factors in their pre-training is still underexplored, especially for recent decoder-only architectures (e.g., Radford et al., 2019). Moreover, studying learning dynamics while ensuring coverage across these randomness factors is increasingly compute-intensive due to the size of contemporary LMs and datasets.

In this work, we define **stability** as the change in a metric of interest (e.g., validation loss) caused by changes in randomness factors and quantify it using the standard deviation of that metric (see Du and Nguyen, 2023 for other approaches to quantify stability). To provide a basis for analysing the stability of LMs to randomness factors (e.g., their training dynamics or final performance) without

[*]Equal contribution. [†]Senior author. [‡]Work done while at EleutherAI. Correspondence to: Stella Biderman <stella@eleuther.ai>, Oskar van der Wal <oskar.vanderwal@gmail.com>, and Pietro Lesci <pl487@cam.ac.uk>.

incurring the costs to train contemporary LMs, we introduce the **PolyPythias**: an extension of the Pythia model suite (Biderman et al., 2023b) trained on the Pile dataset (Gao et al., 2021), for which we release 9 new training runs for 5 model sizes, from 14M up to 410M parameters. These new 45 training runs—in addition to the 5 already available in the suite—cover approximately 7k checkpoints across pre-training, and enable us to analyse training stability of large-scale transformer-based LM with respect to model size, parameter initialisation, and data order as quantified by metrics along the entire model training pipeline: downstream performance and consistency of predictions (§3), shifts in linguistic representations (§4), and dynamics of the model parameters and training phases (§5).

By studying the PolyPythias, we find that: (i) language modelling is largely stable and follows predictable scaling laws with respect to downstream performance; (ii) across training, we identify consistent learning phases: an initial learning phase between steps 10^3–10^4 and a critical learning phase between steps 10^4–10^5; (iii) using training maps constructed from statistics of the model parameters, we identify the characteristics of stable training runs and the early signals of instability.

In the following sections, we describe the PolyPythias release (§2) and how we use the multiple training runs per model size to study the stability of models across various stages of the model training pipeline (§3–§5). We conclude by combining the insights from the individual analyses (§6).

2 Extending the Pythia Suite: PolyPythias Release Description

The Pythia model suite (Biderman et al., 2023b)—with its open data and weights for multiple model sizes, intermediate checkpoints, and detailed reporting of the training configurations—allows researchers to study the learning dynamics of realistic LMs without the need to train them from scratch. Since its release, the suite has been extensively used to study, e.g., LMs' learning dynamics (Michaelov and Bergen, 2023; Arnold et al., 2024), memorisation patterns (Biderman et al., 2023a; Lesci et al., 2024), and biases (Hu et al., 2024). The suite is composed of 10 models ranging in size from 14M to 12B parameters trained on the Pile dataset (Gao et al., 2021; Biderman et al., 2022), a 300B-token curated collection of English documents.[1] All models are trained using the same data. Specifically, the dataset is shuffled and "packed" into sequences of 2,049[2] tokens. Training was performed using a cosine learning rate schedule with warm-up, and using a batch size of 1,024 sequences, resulting in exactly 143k optimisation steps. In the original Pythia suite, for each model size, a single training run is available and consists of 154 checkpoints: at initialisation (step 0), log-spaced up to step 1k (steps $1, 2, ..., 512$), and every 1k steps afterwards (steps 1k–143k).

We consider models with 14M, 31M, 70M, 160M, and 410M parameters. For each size, we release 9 additional training runs resulting in about 7k new checkpoints. In App. E, Table 5, we report the links to the model checkpoints. Each training run uses the same hyperparameters, codebase, and data as Biderman et al. (2023b) but varies the seeds for parameter initialisation and batch composition.[3] We use the standard (i.e., non-deduplicated) version of the Pile and release the tokenised and pre-shuffled datasets corresponding to the different seeds. More training details are in App. A.

A limitation of our suite is that it spans model sizes up to 410M parameters. This choice reflects computational constraints, prioritising seed exploration and checkpoint granularity over scaling up model size. Our aim is to provide an additional resource for researchers unable to train even 410M parameter models from scratch, thus enabling them to study training stability across model sizes.

Prior work that released multi-seed model suites includes Sellam et al. (2022) who introduced the MultiBERTs, a set of 25 BERT-base (final) checkpoints trained with similar hyper-parameters to the original encoder-only BERT architecture (Devlin et al., 2019) but with different random seeds. However, only 28 intermediate checkpoints are available for 5 of the runs, and the release is limited to encoder-only models in a single size. Karamcheti et al. (2021) introduced 10 GPT-2 (124M and 355M parameters; Radford et al., 2019) training runs, each with 600 intermediate checkpoints, but still limited to two model sizes. More recently, Madaan et al. (2024) trained a suite of 10 Llama-2-7B

[1] There exists a deduplicated version of the Pile dataset used to train a second version of the Pythia suite.

[2] Target tokens are the right-shifted input tokens; thus, an additional token is required to achieve the desired input and target sequence length of 2,048 tokens.

[3] We will use the term "batch composition" instead of "data order" because the GPT-NeoX codebase (Andonian et al., 2023), used to train the (Poly)Pythias, shuffles documents before packing them into sequences. This results in sequences that are not simply reshuffled across seeds; they are unique due to the different packing.

(Touvron et al., 2023) models initialised with different random seeds on 210B tokens, analysing 21 intermediate checkpoints which, however, remain publicly unavailable. PolyPythia compares favourably to these suites by spanning 5 model sizes with 154 checkpoints per run, using 10 seeds per model, for a total of almost 7k checkpoints trained on publicly available data.

3 Stability of Downstream Performance

We start by asking a key question, especially relevant for practitioners: *"How stable is model performance on downstream tasks to randomness factors?"* We first study how performance varies across seeds (controlling both batch composition and parameters' initialisation) for a given model size. Then, we analyse how models' predictions and learned gender biases vary throughout the training process and across seeds. We find that language modelling is largely stable and follows predictable scaling laws with respect to downstream performance.

Following Biderman et al. (2023b), we measure model performance as the average **accuracy** on a set of multiple-choice tasks: ARC (Easy) and ARC (Challenge) (Clark et al., 2018), LAMBADA (Paperno et al., 2016), Logiqa (Liu et al., 2020), Piqa (Bisk et al., 2020), SciQ (Welbl et al., 2017), WinoGrande (Sakaguchi et al., 2020), and WSC (Levesque et al., 2012). We measure how predictions agree across seeds and throughout training by computing the Cohen's κ (Cohen, 1960) on the individual multiple-choice answers, where $\kappa = 1$ means perfect agreement while $\kappa = 0$ denotes agreement at the chance level. Specifically, we define the **inter-seed agreement** as the Cohen's κ between the predictions of a model trained with a particular seed and the same model trained with seed 0. Additionally, we define **self-consistency** as the Cohen's κ between the predictions of a model at the last checkpoint and any previous one. Finally, we measure a model's gender bias as its accuracy on the BLiMP (Gender Agreement) (Warstadt et al., 2020), CrowS-Pairs (Gender) (Nangia et al., 2020), and Simple Co-occurrence Bias (Smith et al., 2022) benchmarks. More details about these benchmarks in App. B.

We perform the evaluation using the Language Model Evaluation Harness framework[4] (Gao et al., 2024; Biderman et al., 2024). To limit computational costs while being able to track model behaviour across training for each size and seed, we evaluate performance on a subset of the available checkpoints. Specifically, we use checkpoints at (log-spaced) steps $0, 1, 2, ..., 512$, 1k, and from step 3k onwards we choose every 10k-th step up to 143k, the final checkpoint. We report accuracy, inter-seed agreement, and self-consistency on ARC (Easy) and SciQ in Fig. 1 and the gender bias results in Fig. 2. We show the other benchmarks in App. D. For each metric, we show the median and interquartile range across seeds. We discuss the individual results below.

Downstream Performance. We find (unsurprisingly) that the larger models consistently outperform their smaller counterparts, as indicated by higher accuracy (Fig. 1, left column), except for the more challenging tasks for which all models perform as good as random (see ARC (Challenge), Logiqa, WSC, WinoGrande in App. D, Fig. 6), a result consistent with Biderman et al., 2023b. Performance improves most after step 10^3. However, it drops between step 10^4 and 10^5 for all but 410M, especially for the smallest models. This finding aligns with prior work showing that smaller LMs suffer from "saturation", i.e., a drop in performance at a later stage of the training process due to a mismatch between the dimension of the model representations and the high rank of the output embedding matrix (Michaelov and Bergen, 2023; Godey et al., 2024).

Inter-Seed Agreement and Self-Consistency. For a more fine-grained analysis of how much model behaviour changes across seeds, we compare inter-seed agreement and self-consistency in Fig. 1 (middle and right column, respectively). Inter-seed agreement peaks at $\kappa \approx 0.7$ around step 10^3 before converging towards a "moderate agreement"—i.e., $\kappa \approx 0.5$—around step 10^4 and remaining at this level until the end of training. Self-consistency steadily increases to "moderate agreement"—i.e., $\kappa \approx 0.5$—up to step 10^3, after which it plateaus up to step 10^4, before continuing to increase until the end of training as the models settle on their final answers. This finding aligns with the slow convergence observed in small models by Diehl Martinez et al. (2024).

What is special about step 10^3? A qualitative analysis of model predictions, across tasks and seeds, shows that around step 10^3 models begin generating non-random answers and accuracy

[4] github.com/EleutherAI/lm-evaluation-harness.

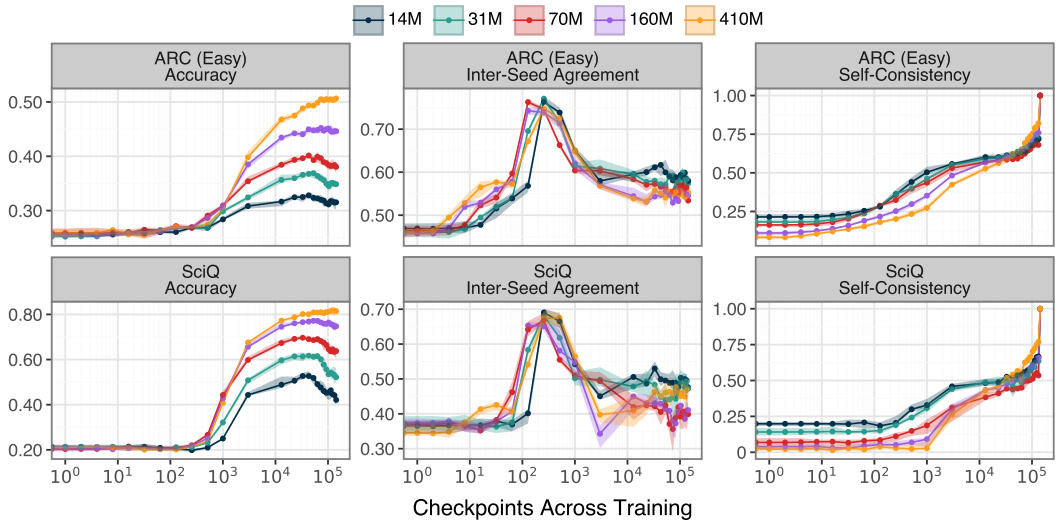


Figure 1: Accuracy, Inter-Seed Agreement, and Self-Consistency (median and interquartile range across seeds) on ARC (Easy) and SciQ.

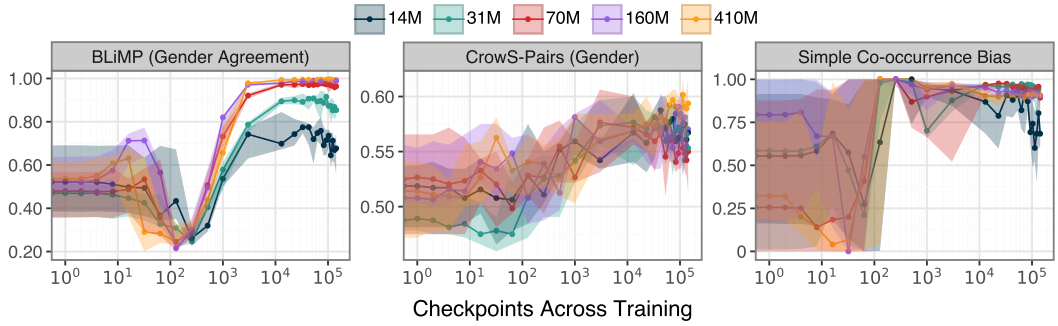


Figure 2: Accuracy, proportions of times the stereotypical answer is chosen, and proportion of times the male option is preferred for, respectively, BLiMP (Gender Agreement), CrowS-Pairs (Gender), Simple Co-occurrence Bias (median and interquartile range across seeds).

improves. Since PolyPythias use a batch size of slightly over 2M tokens, this step aligns with the 2.5B-token mark identified by Olsson et al. (2022) as the point where induction heads form and in-context learning capabilities emerge. Also, this aligns with Tigges et al. (2024) who observe a variety of circuits emerge in Pythia models, regardless of model size, between 2B and 10B tokens (i.e., steps 1k –5k)—e.g., induction heads, successor heads (Gould et al., 2024), copy-suppression heads (McDougall et al., 2023), and name-mover heads (Wang et al., 2023). Further, Chang and Bergen (2022) find that early in training LMs primarily rely on unigram token frequencies before gradually shifting to more contextual predictions—a finding further corroborated by Meister et al. (2023). Similarly, Jumelet et al. (2024) show that Pythia models develop adjective order preferences within this same training range. Collectively, these findings suggest that core semantic functions emerge at a consistent stage in training, regardless of model size or randomness factors.

Finding outlier seeds based on Accuracy and Validation Loss. Zooming in on Fig. 1, we see that per-size accuracy is generally consistent across seeds, with two exceptions. For a given model size, we can identify outlier seeds as follows.[5] First, we consider the accuracy of the last checkpoints on the ARC (Easy), LAMBADA, Piqa, and SciQ tasks, for which all models perform better than random. Second, for each task and model size, we standardise accuracy to have mean zero and a standard deviation of one by standardising across seeds. Finally, we define a region of 2 standard deviations from the mean of a model on that task and consider "outliers" those model-seed combinations that fall outside this region. We (remarkably) find only two such combinations: 410M seed 3 and 4; only

[5]A formal statistical test (e.g., ANOVA and Tukey's test) would have required a larger sample size (i.e., more tasks). Nonetheless, this simple heuristic allows us to discover the same outliers we find using other approaches.

for these very seeds we observe "loss spikes" (see Fig. 5 in App. A). We will further explore these outlier seeds using other metrics and approaches in §5.

Gender Bias. We find distinct phases in the development of gender, both grammatical and bias, in Fig. 2. Specifically, the models start to learn gender agreement between step 10^2 and 10^3 (see BLiMP (Gender Agreement)), which coincides with a sudden shift to a strong bias for using "male identifier words" (see Simple Co-occurrence Bias). Around this step, but less sharply, we also see an increase in gender bias measured by the CrowS-Pairs (Gender) benchmark, which measures more semantically diverse stereotypes than simple co-occurrence statistics. We posit that the large variance observed for the bias measures reflects the poor reliability (e.g., due to small benchmark size, poor quality test items, etc) rather than actual bias differences (see Van der Wal et al., 2024; Delobelle et al., 2024).

4 Representational Stability of Linguistic Information

We now focus on the step before output generation and analyse token representations and find that they remain similar across seeds. Also, representational stability follows consistent trajectories across model sizes, suggesting that trends in smaller models reliably predict those in larger ones.

We study representational (in)stability by applying the information-theoretic probing approach proposed in Müller-Eberstein et al. (2023).[6] We consider seven linguistically motivated token-classification tasks: coreference resolution (Coref; Pradhan et al., 2013), dependency parsing (Dep; Silveira et al., 2014), named entity recognition (NER; Pradhan et al., 2013), part-of-speech tagging (PoS; Pradhan et al., 2013), semantic tagging (SemTag; Abzianidze et al., 2017), sentiment analysis (Senti; Socher et al., 2013), and topic classification (Topic; Lang, 1995); more details in App. B. For each task, we train a **probe** classifier $\boldsymbol{\theta} \in \mathbb{R}^{d \times |\mathcal{Y}|}$ as follows. First, given an input-output pair, $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}$, we collect the model representations, $\mathbf{h}_l(x) \in \mathbb{R}^d$, for each token x in the input sequence $\boldsymbol{x}$ at each layer l. Then, we aggregate per-layer representations into a global representation for that token using a learned weighting scheme $\boldsymbol{\alpha} \in \mathbb{R}^L$, obtaining $\mathbf{h}(x) = \sum_{l=1}^{L} \alpha_l \, \mathbf{h}_l(x)$. This global representation is finally passed as input to the probe, which outputs class probabilities for that token. Both $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are jointly learned by optimising the minimum description length (MDL) loss:[7]

$$\mathcal{L}_{\text{MDL}} = -\mathop{\mathbb{E}}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta})} \Bigg[\sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}} \underbrace{\Bigg(\sum_{x \in \boldsymbol{x}, y \in \boldsymbol{y}} \overbrace{\mathcal{L}_{\text{CE}}\left(\boldsymbol{\theta}^\top \mathbf{h}(x), y\right)}^{\text{Token-level loss}} \Bigg)}_{\text{Sequence-level loss}} \Bigg] + \text{KL}\big(p(\boldsymbol{\theta}) \,\|\, q(\boldsymbol{\theta})\big) \tag{1}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss. Probes trained with only cross-entropy loss can correctly map random representations to labels (Voita and Titov, 2020; Pimentel et al., 2020). Thus, eq. (1) includes a KL-divergence term to keep the probe's distribution $p(\boldsymbol{\theta})$ close to a sparsity-inducing prior $q(\boldsymbol{\theta})$.

We study representational stability through three metrics. First, we measure the **information content** of representations using the probe's macro-F1 score. Second, we note that eq. (1) corresponds to the probe's codelength; thus, we measure **representational efficiency** by defining the **codelength ratio** between random[8] *vs.* probe's representations where values close to 0 indicate high representational efficiency. Finally, we measure the **representational shift** between two consecutive checkpoints using the principal **subspace angles** (**SSA**s; Knyazev and Argentati, 2002): given two probes, SSAs return an angle between $0°$ and $90°$ where lower values represent more similar representations.

Results are summarised in Fig. 3, which shows the macro-F1, Codelength Ratio, and SSAs (rows) for each task (colour) averaged across seeds (line) and the respective standard deviation (shaded area) for each model size (columns) and checkpoint (x-axis). We discuss these results and explore their correlation across different model sizes below.

Information Content. The macro-F1 scores mirror (unsurprisingly) the scaling behaviour observed for the ARC (Easy) and SciQ tasks (§3), where scores consistently improve with model size across

[6] For space reasons, we only sketch the method here and refer to the original paper for a detailed introduction.
[7] We refer to Voita and Titov (2020) for a formal derivation.
[8] We use the probes obtained from the randomly initialized models at step 0.

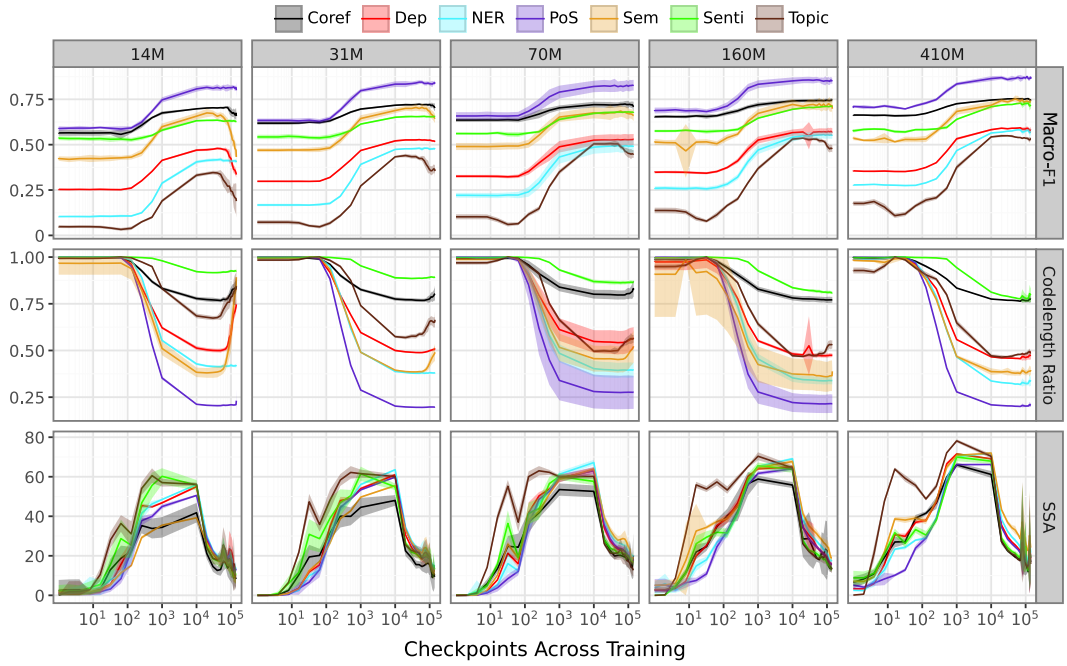


Figure 3: Changes in latent representations of linguistic phenomena (coreference, syntactic dependencies, named entities, parts-of-speech, semantic tags, sentiment, topic). macro-F1, Codelength Ratio, and SSA (rows) for each linguistic task (colour) averaged across seeds (line) and the respective standard deviation (shaded area) for each model size (columns) and checkpoint (x-axis).

both lower-level syntactic tasks (e.g., PoS, Dep) and higher-level semantic ones (e.g., NER, Senti). In other words, larger models return representations with higher information content. While final values differ across model sizes, their trajectory is consistent: mirroring accuracy scores in Fig. 1 (top row), we observe initial improvements around step 10^3 and a rapid increase in macro-F1 up to step 10^4, after which performance for most tasks remains stable, except for the smaller model sizes.

Representational Efficiency. Like macro-F1, the Codelength Ratio starts improving around step 10^3 and reaches its stable value around step 10^4. The final Codelength Ratio for each task depends on the model size, with a distinction between models below and above 100M parameters. Specifically, we observe similar final values for syntactic tasks (e.g., PoS, Dep, Coref) across model sizes, while larger models achieve a 5-10% abs. lower Codelength Ratio on semantic tasks (e.g., NER, Senti).

Representational Shift. The SSAs follow (remarkably) similar trajectories across model sizes. For each task and model size, SSA increases up to step 10^3, then plateaus until step 10^4, and finally decreases until the end of training to a value around $20°$. In other words, model representations after step 10^4 tend to become more stable, matching the reduced rate of change in macro-F1 and Codelength Ratio. The SSA does not directly correlate with macro-F1 or Codelength Ratio: larger changes in SSA early on have positive effects, while smaller changes later on can negatively affect these metrics (e.g., 14M). However, we note that SSA roughly follows the learning rate schedule: peaking around step 2k (end of the warm-up phase) and then decreasing until the end of training.

Correlation across model sizes. We investigate the similarity of the trajectories of representational stability metrics across model sizes. A metric's trajectory for a model size is based on the concatenation of its values for all tasks across all checkpoints. We first compute the Pearson correlation $r_{i,j}$ between a metric's trajectory for each pair of model sizes $\langle i, j\rangle \in \mathcal{M} \times \mathcal{M}$ where $\mathcal{M} = \{\text{14M}, \text{31M}, \text{70M}, \text{160M}, \text{410M}\}$. To obtain the average $\overline{r}$ across all pairs of model sizes, we use the Fisher transformation over correlation coefficients (Fisher, 1970): $\overline{r} = \tanh\left(\frac{1}{|\mathcal{M}|^2}\sum_{i,j\in\mathcal{M}\times\mathcal{M}} \operatorname{arctanh}(r_{i,j})\right)$. Finally, we average the resulting $\overline{r}$'s across seeds in the same way to obtain per-metric average correlations. We find that macro-F1, Codelength Ratio, and SSA are highly correlated across model sizes with an average $\overline{r}$ of 0.99 for macro-F1, 0.98 for

Codelength Ratio, and 0.94 for SSA (all with a p-value smaller than 0.001). This indicates that the representational stability of smaller models is strongly indicative of that of larger models.

5 Training Phases and Outlier Seeds

> After analysing the stability of performance and intermediate representations, we now examine the dynamics of model parameters using training maps. We find mostly consistent training dynamics across model sizes and seeds, with some exceptions. Additionally, training maps from smaller models can predict those of larger models and their final performance.

We investigate the dynamics of model parameters using training maps. A **training map** (Hu et al., 2023) associates each checkpoint with a latent state by fitting a Hidden Markov Model (HMM; Baum and Petrie, 1966) to a vector of statistics derived from the model parameters (e.g., L_2-norm). To fit the HMM, we first gather statistics—listed in App. C, Table 4—from all checkpoints. For each checkpoint within each model size, we standardise the statistics across seeds so that they have a mean of zero and a standard deviation of one, as HMMs are sensitive to the scale of the inputs. We use the standardised sequence to train the HMM with the Baum-Welch algorithm (Baum et al., 1970). Typically, the number of latent states, the primary hyperparameter in HMMs, is determined by minimising some information criterion (e.g., Schwarz, 1978; Akaike, 1998) computed on a validation set. To enable comparisons across model sizes, we use 5 states, as this value is near optimal for the Bayesian Information Criterion across all sizes. Consequently, for each training run, we obtain a sequence of latent states (one per each checkpoint) representing its training map.

First, we use training maps to compare training runs and find outliers (Fig. 4), and we report which properties of the parameters drive state transitions (Table 1). Second, we study the relationship between training maps and final model performance (Tables 2 and 3), and investigate whether it is possible to zero-shot predict final performance from the training map alone. Specifically, we represent each training map as a **bag-of-states**, i.e., a vector counting the number of times a model visits each state. We then use this as input for a linear regression model to predict final performance. As our performance metric, we use the average accuracy of the final model checkpoint on ARC (Easy), LAMBADA, Piqa, and SciQ; we choose these tasks as all models perform better than random. To average accuracy across tasks consistently, for each model size, we standardise it across seeds to have a mean of zero and a standard deviation of one. We refer to the resulting scores as **z-scores**.

Characterising the training maps of outlier seeds. Across model sizes and seeds, we find that training maps are linear graphs with a few exceptions for specific seeds, i.e., seed 3 and 4 of model size 410M. In Fig. 4, we visualise the HMM (left) and training map (right) for 410M and show outlier (top) and stable (bottom) seeds separately. In line with Hu et al. (2023), we find that linear training maps (Fig. 4, bottom-left) describe stable dynamics and performance (Fig. 4, bottom-right), while maps with "forks"—i.e., regressions to an earlier state—are associated with instability. Specifically, forks are only present in the training maps of the outlier seeds (Fig. 4, top-left); these seeds showcase sudden drops in performance (Fig. 4, top-right) and loss spikes (see Fig. 5 in App. A).

Drivers of state transition. In Table 1, we report the three main drivers of state transition, focusing on the transitions that appear in the stable maps but not in the outlier maps and those that are only present in the outlier maps. First, outlier maps fail to perform the transition 2 $\rightarrow$ 3 in which the parameters L_2, median bias, and the average weight variance decrease. Abnormal state transitions for the outlier maps (2 $\rightarrow$ 4) are driven by an increase in the variance of the weights' singular values (σ_λ). The subsequent state transitions (4 $\rightarrow$ 0 and 4 $\rightarrow$ 1) coincide with a strong performance drop and are driven by a sharp decrease in σ_λ. This phenomenon is also described by Godey et al. (2024) as "representation degeneration": the distribution of singular values first becomes increasingly uniform and then abruptly degenerates around a point. We leave further exploration as future work.

Consistency in state transitions. In Table 2 (right), we report the steps at which state transitions occur. We find that state transitions mostly happen around the same step for each model size and across different seeds. The only exception is 410M, for which we observe variability caused by outlier seeds. Additionally, transitions tend to occur at similar steps: the first transition happens around step 10^3, followed by two more transitions between 10^3 and 10^4, and the final transition

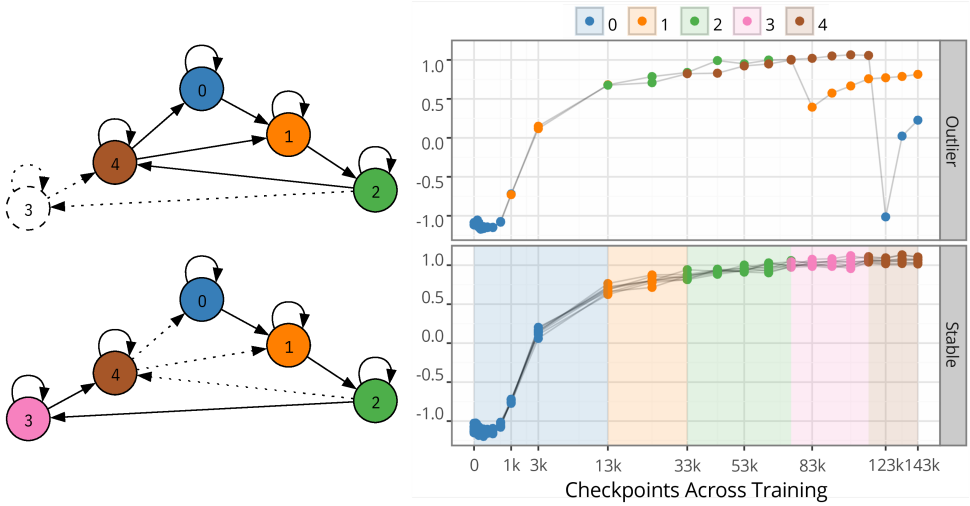


Figure 4: Average standardised accuracy (z-score) across the datasets ARC (Easy), LAMBADA, Piqa, and SciQ (right) for each seed of the model size 410M, along with their corresponding training maps (left). The HMM transitions are colour-coded according to the training map. The results are divided into outlier runs (top; seeds 3 and 4) and stable runs (bottom).

Transition	Description	Top 3 feature determining each transition		
2 → 3	Missing in Outlier	$\text{median}_b \downarrow 0.44$	$L_2 \downarrow 0.42$	$\sigma_w \downarrow 0.50$
2 → 4	Only in Outlier	$\lambda_{\max} \uparrow 1.41$	$\sigma_b \uparrow 1.79$	$\sigma_\lambda \uparrow 1.71$
4 → 0	Only in Outlier	$L_1/L_2 \uparrow 2.76$	$\sigma_b \downarrow 2.43$	$\sigma_\lambda \downarrow 2.99$
4 → 1	Only in Outlier	$L_1/L_2 \uparrow 2.20$	$\sigma_\lambda \downarrow 2.21$	$\lambda_{\max} \downarrow 1.94$

Table 1: Top three drivers of state transitions for 410M. Focus on transitions unique to stable maps and those only present in outlier maps.

around 10^5. Notably, 410M is less consistent, even when we do not account for the outlier seeds. We leave the exploration of this aspect for future work.

Training maps and final performance. The training maps (interestingly) reveal that the outlier runs start to deviate from the other runs long before they show worse scores on the performance metrics. In Fig. 4 (top), we see that both runs enter state 2 prematurely and fail to make the transition to state 3. To investigate whether a model's training maps are informative of that model's final performance, we perform linear regressions using the bag-of-states to predict the final average z-score. We report the regression R^2 (Table 2, R^2 column). Only the training map of 410M is predictive of performance ($R^2 = 0.99$). Also, this model, due to the outlier seeds, is associated with a high performance variance (Table 2, σ^2 column). We report the regression coefficients for 410M in the last row of Table 3. We observe that the outlier seeds (3 and 4) are the only ones receiving a negative coefficient. In other words, the bag-of-states obtained from the training map of 410M has enough information to predict this model will underperform.

Predicting performance across model sizes using training maps. We investigate whether training maps from smaller models can help predict the performance of a larger model. Specifically, we attempt to predict the performance of 410M using training maps from smaller models. Our approach consists of two steps. First, for a given small model (e.g., 14M), we train an HMM on checkpoints from all seeds. We then use this HMM to assign each checkpoint of 410M to a latent state, effectively performing a "zero-shot" prediction of its training map. Second, we aggregate the training maps of the large model across all seeds into a bag-of-states representation, which we use to predict the performance of each seed. Our hypothesis is that if this regression successfully predicts model performance, then the training map of the smaller model carries meaningful information that can be used to predict the performance of the larger model. In Table 3, we report the R^2 and coefficients of the regression that uses each size-specific training map to predict the performance of 410M. We find

		z-scores		Step at which transition happens*			
Size	**Fork**	R^2	σ^2	0 → 1	1 → 2	2 → 3	3 → 4
14M		0.75	0.65	$5\text{k}_{\pm 0.7\text{k}}$	$38\text{k}_{\pm 3.5\text{k}}$	$72\text{k}_{\pm 2.2\text{k}}$	$108\text{k}_{\pm 2.7\text{k}}$
31M		0.64	0.84	$7\text{k}_{\pm 0.7\text{k}}$	$42\text{k}_{\pm 2.6\text{k}}$	$69\text{k}_{\pm 2.1\text{k}}$	$104\text{k}_{\pm 2.0\text{k}}$
70M		0.41	0.71	$6\text{k}_{\pm 0.6\text{k}}$	$28\text{k}_{\pm 3.1\text{k}}$	$58\text{k}_{\pm 3.3\text{k}}$	$94\text{k}_{\pm 2.5\text{k}}$
160M		0.03	0.58	$2\text{k}_{\pm 0}$	$18\text{k}_{\pm 0.8\text{k}}$	$61\text{k}_{\pm 1.5\text{k}}$	$100\text{k}_{\pm 1.6\text{k}}$
410M	✓	0.99	0.98	$18\text{k}_{\pm 22.6\text{k}}$	$35\text{k}_{\pm 24.1\text{k}}$	$73\text{k}_{\pm 20.9\text{k}}$	$114\text{k}_{\pm 4.9\text{k}}$

Table 2: Overview of statistics for the training maps found for the different sizes of Pythia models. R^2: the goodness of fit for the linear regression for the bag-of-states and the z-scores. σ^2: the variance of the average performance z-scores. *Averaged across seeds, except for outlier seeds 3 and 4 for 410M to remove non-linearities due to forks in the training maps.

		Seed used to train 410M									
HMM	R^2	0	1	2	3	4	5	6	7	8	9
14M	0.18	0.01	0.06	0.39	−1.12	−0.21	0.03	0.35	0.01	0.08	0.39
31M	0.97	0.42	0.52	0.48	−0.56	−2.78	0.42	0.31	0.32	0.44	0.43
70M	0.97	0.34	0.43	0.48	−0.60	−2.75	0.23	0.19	0.51	0.67	0.49
160M	0.96	0.46	0.46	0.69	−0.58	−2.74	0.31	0.28	0.37	0.25	0.50
410M	0.99	0.39	0.45	0.38	−0.57	−2.80	0.47	0.48	0.42	0.49	0.29

Table 3: Linear regression coefficients returned by regressing the average z-score of 410M on its bag-of-states obtained from the zero-shot training map constructed using the HMM trained on the model listed in column "HMM". Outlier seeds are highlighted in grey.

that forming the bag-of-states using predict the average z-score successfully with $R^2 > 0.9$ for all models, except 14M, the smallest size. Also, the negative coefficients for the outlier seeds indicate that it is possible to predict underperforming runs across sizes. When computing the bag-of-states for 410M, we pass all its checkpoints through the HMM to generate the full training map. Ideally, we would like to use only the initial checkpoints to predict the final performance of a partially trained model. This would allow us to decide early on whether to stop a specific run. However, when we construct the bag-of-states using only a partial training run, we fail to predict the average z-score accurately. Empirically, we find that at least 120k steps (out of the total 143k) are required for a reliable prediction. We leave the investigation of which properties can be predicted from the bag-of-states of early checkpoint metrics as future work.

6 Discussion

Our experiments on downstream performance (§3), intermediate representations (§4), and model parameters (§5) allow us to examine the stability of training and find outlier runs using different methods across the model training pipeline. In this section, we analyse commonalities across the resulting metrics to identify broader characteristics of LM pre-training dynamics.

Language modelling is largely stable. Generally, we observe LM pre-training dynamics to follow consistent trajectories. Across seeds and model sizes, downstream performance and representational efficiency consistently increase during pre-training, and training maps are linear (except for the outlier seeds). Furthermore, model scaling laws seem to hold across seeds, not only for downstream performance but also for information content and representational efficiency of model representations. Similarly, both at the performance and representational level, we observe the effect of "saturation" (Michaelov and Bergen, 2023; Godey et al., 2024) in smaller models.

Linguistic information is encoded in the initial learning phase (10^3–10^4 steps). Across all experiments, metrics start moving away from the initial random baseline around step 10^3 (2B tokens circa) and reach their convergence level around step 10^4 (20B tokens circa). In this initial phase, representational shift peaks and linguistic information begins to be encoded into the models' latent representations. Through the lens of multiple metrics, we can analyse model behaviour in this phase

in detail. Specifically, while the amount and the efficiency with which linguistic information is encoded in model representations have already increased substantially at step 10^3, the model does not yet generate coherent outputs, as indicated by low performance on linguistic acceptability benchmarks like BLiMP (Gender Agreement). Simultaneously, self-consistency is low while inter-seed agreement is high, which we hypothesise to be an artefact of all models initially choosing an incorrect baseline answer. In terms of training maps, this phase corresponds to the 0 → 1 transition, which occurs consistently in this initial training phase for all model sizes (except for the 410M outliers).

Most improvements happen in the "critical" learning phase (10^4–10^5 steps). In the range of 10^3 to 10^4 steps, most learning occurs, as measured by all of our metrics. Performance increases the most, and linguistic information content and representational efficiency converge to close-to-final values. At the same time, the representational shift begins to decrease from its peak, showing how the information encoded in the model begins to stabilise. This is reflected in terms of performance by the simultaneous increase in self-consistency. The fact that inter-seed agreement decreases before remaining relatively constant until the end of training indicates that models settle on their final answers after this stage. In the training maps, this phase corresponds to the 1 → 2 and 2 → 3 transitions that occur before step 10^5. The fact that our outlier runs (410M, seeds 3 and 4) exit state 2 prematurely indicates that this phase is important for the LM's downstream performance. Furthermore, we note that these learning phases are (remarkably) similar and occur at the same time during pre-training as for encoder-only models (Müller-Eberstein et al., 2023). Also, they follow similar trajectories as recurrent LSTM architectures (Saphra and Lopez, 2019). We leave the exploration of which modelling or data decisions may result in these learning phases for future work. Finally, we observe that all benchmarks still improve past the optimal token count (e.g., 8.2B tokens, or around 4k–5k steps) predicted by the Chinchilla scaling law (Hoffmann et al., 2022).

Training maps describe outlier seeds. Using the downstream performance results and training loss (§3), and training maps (§5), we identify outlier seeds and explain how the model parameters change in the unstable pre-training regime. Moreover, we find that it is possible to predict training maps in a "zero-shot" fashion (e.g., using HMM trained on smaller models to predict the training map of larger ones), suggesting that statistics of the parameters of smaller models are informative of their larger counterparts. We propose using PolyPythias to investigate potential connections between the state transitions observed in training maps and the emergence of components or behaviours in circuit analyses, as discussed in §3. However, a limitation is that early-training state transitions may occur too consistently across seeds for each model size to provide meaningful insights into this relationship. To better understand whether connections exist between circuit formation (e.g., the development of induction heads or other circuit components in each seed) and state transitions, interventional studies of training dynamics may be necessary. We leave the exploration of this aspect for future work.

7 Conclusions

In this work, we introduce PolyPythia, a multi-seed extension of the Pythia model suite (Biderman et al., 2023b), adding 45 extra pre-training runs for a total of 10 seeds across 5 model sizes. This expanded resource is designed to facilitate research on training dynamics and model stability. Through our experiments, we demonstrate the usefulness of PolyPythias by analysing the stability of the language modelling pipeline. We examine downstream performance, intermediate representations, and parameter training dynamics across seeds and model scales. Our findings suggest that, with some exceptions, language modelling remains largely stable.

We hope these additional seeds will support further research into the impact of randomness in model pre-training. This includes studying the robustness of different evaluation metrics (e.g., Sellam et al., 2022; Van der Wal et al., 2024; Madaan et al., 2024), identifying factors contributing to suboptimal training (e.g., Zoph et al., 2022; Zhai et al., 2023; Chowdhery et al., 2023; Zeng et al., 2023; Chung et al., 2024), and improving the predictability of model performance across scales (e.g., Kaplan et al., 2020; Srivastava et al., 2023). Further, the diverse training runs with different data orders enable further studies on memorisation (e.g., Biderman et al., 2023a; Lesci et al., 2024) or on the relationship of the data to the emergence of learned behaviours (Van der Wal et al., 2022; Biderman et al., 2023b; Jumelet et al., 2024; Belrose et al., 2024). Lastly, PolyPythia provides a valuable testbed for assessing benchmark evaluations' reliability and model interventions' effectiveness.

ACKNOWLEDGEMENTS

This research was made possible through computational resources generously provided by StabilityAI. We express our gratitude to Michael Hu for his assistance with the implementation of the HMM training maps presented in §5. We thank Andreas Vlachos, Tiago Pimentel, and Clara Meister for their insightful feedback on earlier drafts. Finally, we extend our thanks to Davide Lesci and Marco Lesci for proofreading the final version of the manuscript.

Oskar van der Wal initiated this project during his research internship at EleutherAI and gratefully acknowledges their support. His work is partially funded by the Dutch Research Council (NWO) through the project "The biased reality of online media" (406.DI.19.059).

Pietro Lesci received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation programme grant AVeriTeC (Grant agreement No. 865958).

Max Müller-Eberstein thanks the IT University of Copenhagen's High-Performance Computing Cluster for supporting the representational stability experiments in §4.

Naomi Saphra's work was enabled in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence.

Hailey Schoelkopf thanks EleutherAI for the opportunity to conduct this work.

Willem Zuidema is funded by the Institute for Logic, Language and Computation of the University of Amsterdam.

REFERENCES

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Mayank Agarwal, Mikhail Yurochkin, and Yuekai Sun. 2021. On sensitivity of meta-learning to support data. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 20447–20460.

Hirotogu Akaike. 1998. Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand. Association for Computational Linguistics.

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Jason Phang, Shivanshu Purohit, Hailey Schoelkopf, Dashiell Stander, Tri Songz, Curt Tigges, Benjamin Thérien, Phil Wang, and Samuel Weinbach. 2023. GPT-NeoX: Large scale autoregressive language modeling in PyTorch.

Julian Arnold, Flemming Holtorf, Frank Schäfer, and Niels Lörch. 2024. Phase transitions in the output distribution of large language models. *ArXiv preprint 2405.17088*.

Leonard E. Baum and Ted Petrie. 1966. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563.

Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171.

Nora Belrose, Quintin Pope, Lucia Quirke, Alex Mallen, and Xiaoli Fern. 2024. Neural networks learn statistics of increasing complexity. *ArXiv preprint*, abs/2402.04362.

Stella Biderman, Kieran Bicheno, and Leo Gao. 2022. Datasheet for the Pile. *ArXiv preprint 2201.07311*.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023a. Emergent and predictable memorization in large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023b. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. 2024. Lessons from the trenches on reproducible evaluation of language models. *ArXiv preprint 2405.14782*.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tyler A. Chang and Benjamin K. Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Woojin Chung, Jiwoo Hong, Na Min An, James Thorne, and Se-Young Yun. 2024. Stable language model pre-training by reducing embedding variability. In *Proceedings of the 2024 Conference on*

Empirical Methods in Natural Language Processing, pages 10852–10863, Miami, Florida, USA. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *ArXiv preprint 1803.05457*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2022. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61.

Pieter Delobelle, Giuseppe Attanasio, Debora Nozza, Su Lin Blodgett, and Zeerak Talat. 2024. Metrics for what, metrics for whom: Assessing actionability of bias evaluation metrics in NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21669–21691, Miami, Florida, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Richard Diehl Martinez, Pietro Lesci, and Paula Buttery. 2024. Tending towards stability: Convergence challenges in small language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3275–3286, Miami, Florida, USA. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *ArXiv preprint 2002.06305*.

Yupei Du and Dong Nguyen. 2023. Measuring the instability of fine-tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6209–6230, Toronto, Canada. Association for Computational Linguistics.

Ronald A. Fisher. 1970. *Statistical methods for research workers.*, 14. ed. edition. Oliver and Boyd, Edinburgh.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The Pile: An 800GB dataset of diverse text for language modeling. *ArXiv preprint 2101.00027*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. 2024. Why do small language models underperform? Studying language model saturation via the softmax bottleneck. *ArXiv preprint 2404.07647*.

Seraphina Goldfarb-Tarrant, Pedro Rodriguez, Jane Dwivedi-Yu, and Patrick Lewis. 2024. Multi-Contrievers: Analysis of dense retrieval representations. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 118–139, Miami, Florida, US. Association for Computational Linguistics.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. 2024. Successor heads: Recurring, interpretable attention heads in the wild. In *The Twelfth International Conference on Learning Representations*.

Odd Erik Gundersen, Kevin Coakley, Christine Kirkpatrick, and Yolanda Gil. 2022. Sources of irreproducibility in machine learning: A review. *ArXiv preprint 2204.07610*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Michael Y. Hu, Angelica Chen, Naomi Saphra, and Kyunghyun Cho. 2023. Latent state models of training dynamics. *Transactions on Machine Learning Research*.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2024. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75.

Jaap Jumelet, Lisa Bylinina, Willem Zuidema, and Jakub Szymanik. 2024. Black big boxes: Do language models hide a theory of adjective order? *ArXiv preprint 2407.02136*.

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

Siddharth Karamcheti, Laurel Orr, Jason Bolton, Tianyi Zhang, Karan Goel, Avanika Narayan, Rishi Bommasani, Deepak Narayanan, Tatsunori Hashimoto, Dan Jurafsky, Christopher D. Manning, Christopher Potts, Christopher Ré, and Percy Liang. 2021. Mistral–A journey towards reproducible language model training.

Andrew V. Knyazev and Merico E. Argentati. 2002. Principal angles between subspaces in an A-based scalar product: Algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2008–2040.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, ICML'95, page 331–339, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. 2024. Causal estimation of memorisation profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15616–15635, Bangkok, Thailand. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3622–3628. ijcai.org.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. 2024. Quantifying variance in evaluation benchmarks. *ArXiv preprint 2406.10229*.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. 2023. Copy suppression: Comprehensively understanding an attention head. *ArXiv preprint 2310.04625*.

Clara Meister, Wojciech Stokowiec, Tiago Pimentel, Lei Yu, Laura Rimell, and Adhiguna Kuncoro. 2023. A natural bias for language generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 243–255, Toronto, Canada. Association for Computational Linguistics.

James Michaelov and Ben Bergen. 2023. Emergent inabilities? Inverse scaling over the course of pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14607–14615, Singapore. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Max Müller-Eberstein, Rob van der Goot, Barbara Plank, and Ivan Titov. 2023. Subspace chronicles: How linguistic information emerges, shifts and interacts during language model training. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13190–13208, Singapore. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *ArXiv preprint 2209.11895*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Branislav Pecher, Ivan Srba, and Maria Bielikova. 2024. A survey on stability of learning with limited labelled data and its sensitivity to the effects of randomness. *ACM Comput. Surv.*

Hung Viet Pham, Shangshu Qian, Jiannan Wang, Thibaud Lutellier, Jonathan Rosenthal, Lin Tan, Yaoliang Yu, and Nachiappan Nagappan. 2020. Problems and opportunities in training deep learning software systems: An analysis of variance. In *Proceedings of the 35th IEEE/ACM international conference on automated software engineering*, pages 771–783.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial*

Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 8732–8740. AAAI Press.

Naomi Saphra and Adam Lopez. 2019. Understanding learning dynamics of language models with SVCCA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3257–3267, Minneapolis, Minnesota. Association for Computational Linguistics.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. 2022. The MultiBERTs: BERT reproductions for robustness analysis. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, a large-scale generative language model. *Arxiv preprint 2201.11990*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germàn Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar,

Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Sophie Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. 2024. LLM circuit analyses are consistent across training and scale. In *Advances in Neural Information Processing Systems*, volume 37, pages 40699–40731. Curran Associates, Inc.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu,

Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint 2307.09288*.

Oskar Van der Wal, Dominik Bachmann, Alina Leidinger, Leendert van Maanen, Willem Zuidema, and Katrin Schulz. 2024. Undesirable biases in NLP: Addressing challenges of measurement. *Journal of Artificial Intelligence Research*, 79:1–40.

Oskar Van der Wal, Jaap Jumelet, Katrin Schulz, and Willem Zuidema. 2022. The birth of bias: A case study on the evolution of gender bias in an English language model. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 75–75, Seattle, Washington. Association for Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

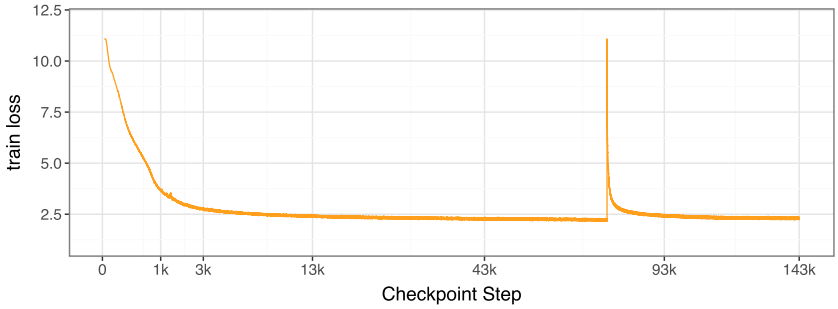
Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. 2023. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 40770–40803. PMLR.

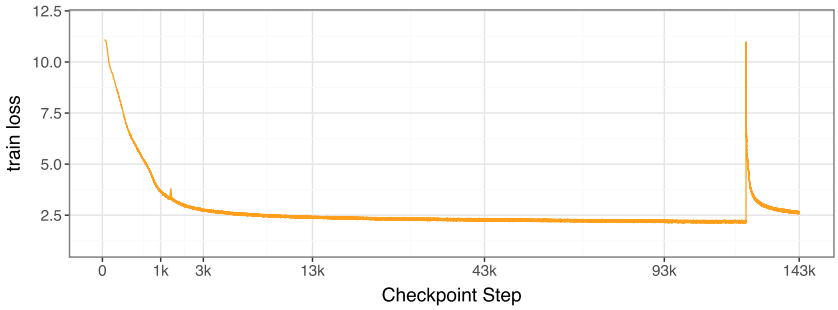
Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. ST-MoE: Designing stable and transferable sparse expert models. *ArXiv preprint 2202.08906*.

A Training Details

We used the v1.0 version of the GPT-Neox codebase[9] for model training. We report the training loss for the two outlier pre-training runs, 410M seed 3 and 4, below. We refer to the Weights & Biases space for the training losses and logs of all runs.



(a) Seed 3



(b) Seed 4

Figure 5: Training loss for 410M outlier seeds 3 and 4 showing "loss spikes".

B Evaluation Datasets Details

In this section, we describe the datasets used in §3 and §4.

ARC (Easy) and ARC (Challenge) (Clark et al., 2018). The AI2 Reasoning Challenge (ARC) is a benchmark for evaluating natural science question answering. It consists of two subsets, Easy and Challenge, where questions are drawn from standardized science exams. Following Brown et al. (2020), evaluation is performed using log-likelihood-based scoring and reported in terms of accuracy.

LAMBADA (Paperno et al., 2016). LAMBADA is a language modelling benchmark that measures a model's ability to predict the final word of a sentence given its full context. The dataset consists of narrative texts, and successful prediction requires long-range context comprehension.

Logiqa (Liu et al., 2020). Logiqa is a reading comprehension dataset designed to test logical reasoning in language models. The questions are derived from LSAT exam questions and require deductive reasoning and critical thinking skills. Evaluation is performed using multiple-choice accuracy.

Piqa (Bisk et al., 2020). Piqa is a dataset designed to evaluate physical commonsense reasoning. It consists of multiple-choice questions requiring an understanding of everyday physical interactions. Evaluation is based on accuracy.

[9] github.com/EleutherAI/gpt-neox tag v1.0.

SciQ (Welbl et al., 2017). SciQ is a question-answering dataset focused on scientific topics. It includes multiple-choice, direct-answer, and "cloze"-style questions sourced from educational materials. Evaluation involves measuring accuracy on multiple-choice and direct-answer formats.

WSC (Levesque et al., 2012). The Winograd Schema Challenge (WSC) is a coreference resolution task designed to test commonsense reasoning. It consists of sentence pairs that differ by a single word, requiring the model to correctly resolve ambiguous pronouns based on contextual cues. Accuracy is used as the primary evaluation metric.

WinoGrande (Sakaguchi et al., 2020). WinoGrande is an expanded version of the Winograd Schema Challenge, containing sentence pairs with minor lexical variations. The task evaluates a model's ability to resolve ambiguous pronouns by comparing the probabilities of different completions. Accuracy is reported as the primary metric.

BLiMP (Gender Agreement) (Warstadt et al., 2020). BLiMP (Gender Agreement) is a subset of the BLiMP benchmark designed to evaluate gender bias in language models. It consists of minimal sentence pairs that differ only in gender-marked words, allowing the assessment of whether a model exhibits gender preference in syntactic and morphological structures.

CrowS-Pairs (Nangia et al., 2020). CrowS-Pairs assesses biases in language models by presenting minimal sentence pairs that contrast stereotypical and non-stereotypical perspectives on US-protected demographic groups. We use the adaptation by Névéol et al. (2022), which removes instances identified as problematic for validity following Blodgett et al. (2021).

Simple Co-occurrence Bias (Smith et al., 2022). Simple Co-occurrence Bias evaluates gender associations in language models by analysing the likelihood of gendered identifiers (e.g., "man", "woman") appearing in simple template-based prompts. Following Brown et al. (2020); Smith et al. (2022), we report directional bias in model predictions. The dataset is available at huggingface.co/datasets/oskarvanderwal/simple-cooccurrence-bias.

Coref (Pradhan et al., 2013). Coref is a dataset for coreference resolution, which involves linking mentions of the same entity in a text. We use the coreference annotation layer from OntoNotes 5.0, which is commonly used to evaluate models' ability to resolve pronouns and named entity references in complex passages.

Dep (Silveira et al., 2014). Dep is a dataset for dependency parsing, providing annotations of relative syntactic functions based on the English Web Treebank. It is used to evaluate a model's ability to predict grammatical relations between words in a sentence.

NER (Pradhan et al., 2013). NER (Named Entity Recognition) is a dataset used to identify and classify proper nouns into 18 predefined categories, such as persons, organizations, and locations. We use the NER annotation layer from OntoNotes 5.0, which serves as a standard benchmark for evaluating entity recognition performance.

PoS (Pradhan et al., 2013). PoS (Part-of-Speech Tagging) is a dataset that annotates words in a sentence with one of 51 syntactic categories (e.g., noun, verb, adjective), as taken from OntoNotes 5.0. It evaluates the models' ability to perform syntactic parsing at the word level.

SemTag (Abzianidze et al., 2017). SemTag is a dataset for semantic tagging, which classifies words or phrases based on 69 semantic roles and meanings. It is used to assess the models' ability to understand and differentiate word meanings in context.

Senti (Socher et al., 2013). Senti is a sentiment analysis dataset containing sentences labelled with sentiment polarity (positive, negative, neutral). It is used to evaluate models' ability to infer sentiment from textual data.

Topic (Lang, 1995). Topic is a dataset for topic classification, consisting of documents labelled with 20 predefined topic categories. It serves as a benchmark for evaluating models' ability to perform document-level topic classification.

C Training Maps

In Table 4, we report the metrics used to fit the HMMs to create the training maps in §5. These metrics are (partially) taken from Hu et al. (2023) Appendix B, to which we refer for further details.

Metric	Description
L_1	The L_1-norm, averaged over the weight matrices
L_2	The L_2-norm, averaged over the weight matrices
L_1/L_2	Weight sparsity (ratio of their L_1 and L_2 norms), averaged over the weight matrices
μ_w	Sample mean of the weights
median_w	Median of the weights
σ_w	Sample variance of weights
μ_b	Sample mean of the biases
median_b	Median of the biases
σ_b	Sample variance of biases
trace	The average trace over the weight matrices
λ_{max}	The average spectral norm of the weights
$\text{trace}/\lambda_{max}$	The average trace over spectral norm
μ_λ	The average singular value over the weights
σ_λ	Sample variance of singular values over the weights

Table 4: Statistics of model parameters (both weights and biases) used to fit the HMMs in §5.

D Additional Figures

We report the figures for all the benchmarks analysed in §3. The figures follow in the next pages.

E Release Links

Besides the link to the Hugging Face collection listing all checkpoints reported in the abstract,[10] in Table 5 we report the links to the individual checkpoints on the Hugging Face Hub. Also, the indices used to recreate the pre-shuffled datasets are available at huggingface.co/datasets/EleutherAI/pile-preshuffled-seeds. Links follow on the next page (after the figures).

[10]Explicitly, huggingface.co/collections/EleutherAI/polypythias-67bed6916110c8933e1ea561.

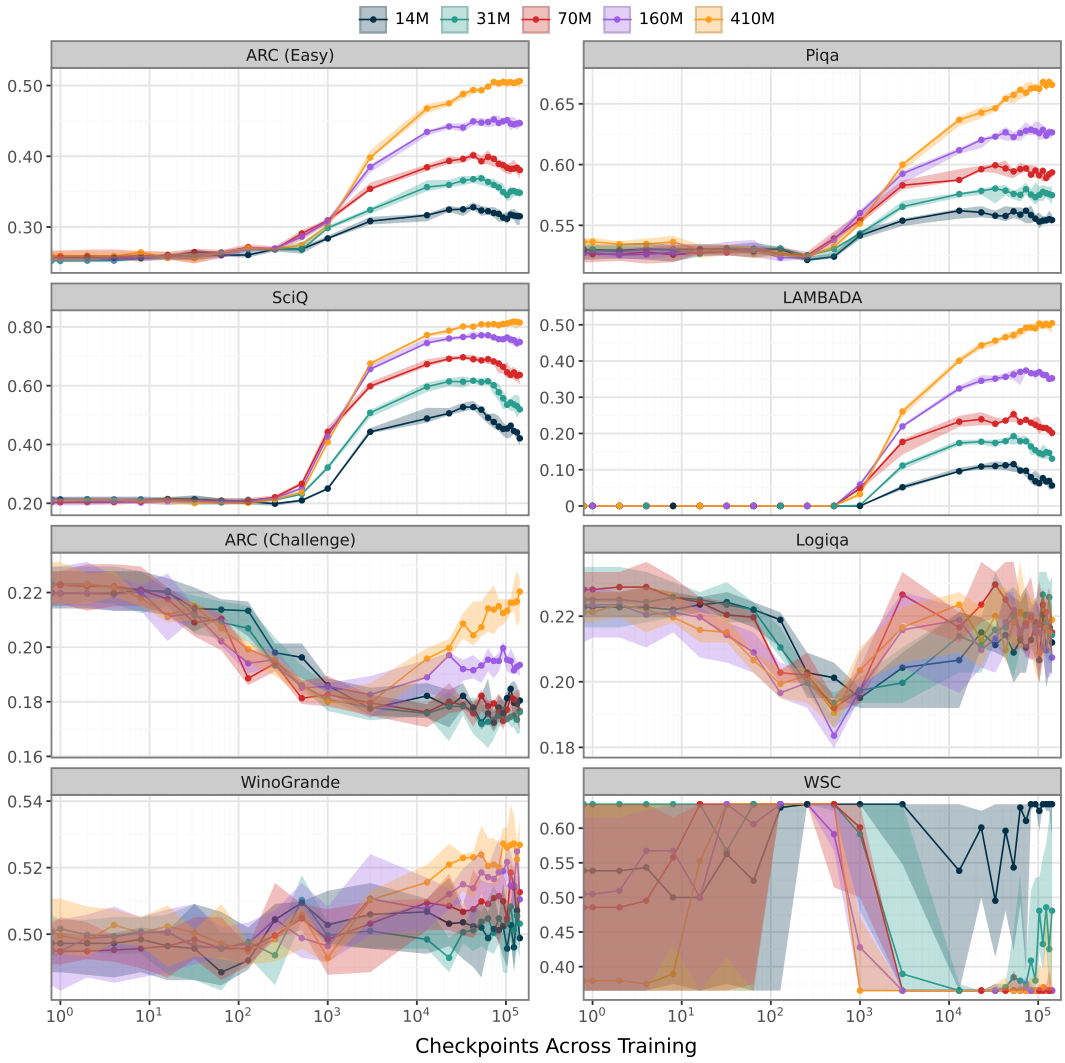


Figure 6: Accuracy (median and interquartile range across seeds) on ARC (Easy), Piqa, SciQ, LAMBADA, ARC (Challenge), Logiqa, WinoGrande, and WSC tasks.

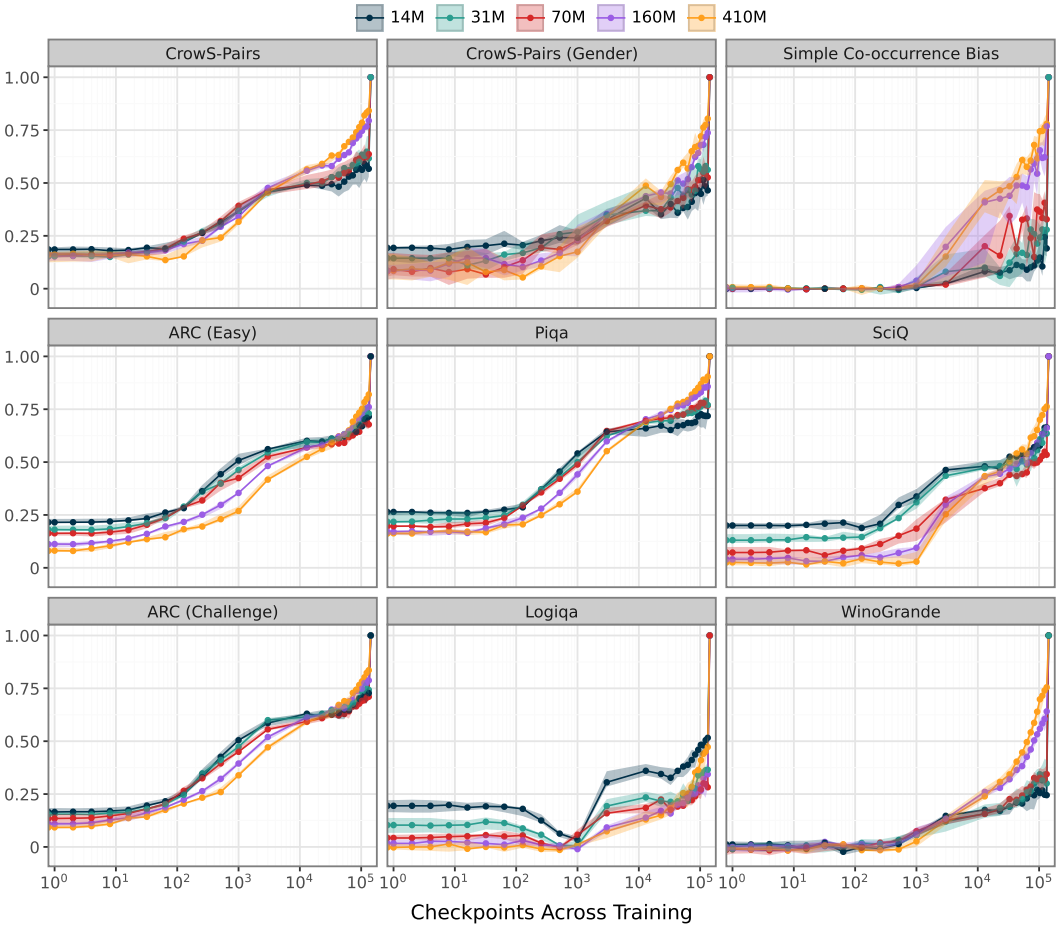


Figure 7: Self-consistency (median and interquartile range across seeds) on CrowS-Pairs, CrowS-Pairs (Gender), Simple Co-occurrence Bias, ARC (Easy), ARC (Challenge), Piqa, SciQ, ARC (Challenge), Logiqa, and WinoGrande.

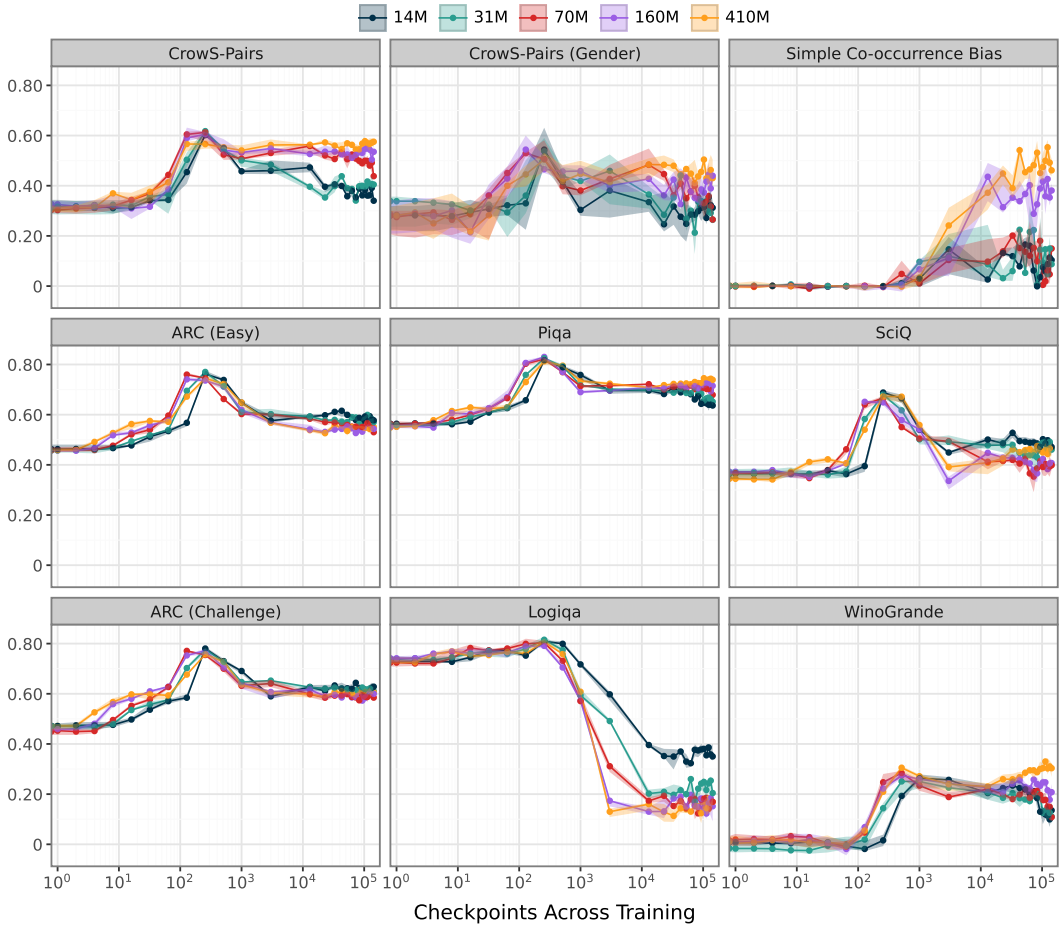


Figure 8: Inter-seed agreement (median and interquartile range across seeds) on CrowS-Pairs, CrowS-Pairs (Gender), Simple Co-occurrence Bias, ARC (Easy), ARC (Challenge), Piqa, SciQ, ARC (Challenge), Logiqa, and WinoGrande.

Model Size	Seed	Links
14M	0 (default)	huggingface.co/EleutherAI/pythia-14m
	1	huggingface.co/EleutherAI/pythia-14m-seed1
	2	huggingface.co/EleutherAI/pythia-14m-seed2
	3	huggingface.co/EleutherAI/pythia-14m-seed3
	4	huggingface.co/EleutherAI/pythia-14m-seed4
	5	huggingface.co/EleutherAI/pythia-14m-seed5
	6	huggingface.co/EleutherAI/pythia-14m-seed6
	7	huggingface.co/EleutherAI/pythia-14m-seed7
	8	huggingface.co/EleutherAI/pythia-14m-seed8
	9	huggingface.co/EleutherAI/pythia-14m-seed9
31M	0 (default)	huggingface.co/EleutherAI/pythia-31m
	1	huggingface.co/EleutherAI/pythia-31m-seed1
	2	huggingface.co/EleutherAI/pythia-31m-seed2
	3	huggingface.co/EleutherAI/pythia-31m-seed3
	4	huggingface.co/EleutherAI/pythia-31m-seed4
	5	huggingface.co/EleutherAI/pythia-31m-seed5
	6	huggingface.co/EleutherAI/pythia-31m-seed6
	7	huggingface.co/EleutherAI/pythia-31m-seed7
	8	huggingface.co/EleutherAI/pythia-31m-seed8
	9	huggingface.co/EleutherAI/pythia-31m-seed9
70M	0 (default)	huggingface.co/EleutherAI/pythia-70m
	1	huggingface.co/EleutherAI/pythia-70m-seed1
	2	huggingface.co/EleutherAI/pythia-70m-seed2
	3	huggingface.co/EleutherAI/pythia-70m-seed3
	4	huggingface.co/EleutherAI/pythia-70m-seed4
	5	huggingface.co/EleutherAI/pythia-70m-seed5
	6	huggingface.co/EleutherAI/pythia-70m-seed6
	7	huggingface.co/EleutherAI/pythia-70m-seed7
	8	huggingface.co/EleutherAI/pythia-70m-seed8
	9	huggingface.co/EleutherAI/pythia-70m-seed9
160M	0 (default)	huggingface.co/EleutherAI/pythia-160m
	1	huggingface.co/EleutherAI/pythia-160m-seed1
	2	huggingface.co/EleutherAI/pythia-160m-seed2
	3	huggingface.co/EleutherAI/pythia-160m-seed3
	4	huggingface.co/EleutherAI/pythia-160m-seed4
	5	huggingface.co/EleutherAI/pythia-160m-seed5
	6	huggingface.co/EleutherAI/pythia-160m-seed6
	7	huggingface.co/EleutherAI/pythia-160m-seed7
	8	huggingface.co/EleutherAI/pythia-160m-seed8
	9	huggingface.co/EleutherAI/pythia-160m-seed9
	1 (only data)	huggingface.co/EleutherAI/pythia-160m-data-seed1
	2 (only data)	huggingface.co/EleutherAI/pythia-160m-data-seed2
	3 (only data)	huggingface.co/EleutherAI/pythia-160m-data-seed3
	1 (only parameters)	huggingface.co/EleutherAI/pythia-160m-weight-seed1
	2 (only parameters)	huggingface.co/EleutherAI/pythia-160m-weight-seed2
	3 (only parameters)	huggingface.co/EleutherAI/pythia-160m-weight-seed3
410M	0 (default)	huggingface.co/EleutherAI/pythia-410m
	1	huggingface.co/EleutherAI/pythia-410m-seed1
	2	huggingface.co/EleutherAI/pythia-410m-seed2
	3	huggingface.co/EleutherAI/pythia-410m-seed3
	4	huggingface.co/EleutherAI/pythia-410m-seed4
	5	huggingface.co/EleutherAI/pythia-410m-seed5
	6	huggingface.co/EleutherAI/pythia-410m-seed6
	7	huggingface.co/EleutherAI/pythia-410m-seed7
	8	huggingface.co/EleutherAI/pythia-410m-seed8
	9	huggingface.co/EleutherAI/pythia-410m-seed9

Table 5: Links to the individual checkpoints in the PolyPythias release.