Forecasting Downstream Performance of LLMs With Proxy Metrics

Arkil Patel **Siva Reddy** **Marius Mosbach** **Dzmitry Bahdanau**

Mila – Quebec AI Institute & McGill University

Canada CIFAR AI Chair ServiceNow Research Periodic Labs

Correspondence to: arkil.patel@gmail.com McGill-NLP/proxy-metrics

Abstract

Progress in language model development is often driven by comparative decisions: which architecture to adopt, which pretraining corpus to use, or which training recipe to apply. Making these decisions well requires reliable performance forecasts, yet the two commonly used signals are fundamentally limited. Cross-entropy loss is poorly aligned with downstream capabilities, and direct downstream evaluation is expensive, sparse, and often uninformative at early training stages. Instead, we propose to construct proxy metrics by aggregating token-level statistics, such as entropy, top-k accuracy, and expert token rank, from a candidate model's next token distribution over expert-written solutions. Across three settings, our proxies consistently outperform loss- and compute-based baselines: 1) For cross-family model selection, they rank a heterogeneous population of reasoning models with mean Spearman $\rho = 0.81$ (vs. $\rho = 0.36$ for cross-entropy loss); 2) For pretraining data selection, they reliably rank 25 candidate corpora for a target model at roughly $10{,}000\times$ less compute than direct evaluation, pushing the Pareto frontier beyond existing methods; and 3) for training-time forecasting, they extrapolate downstream accuracy across an $18\times$ compute horizon with roughly half the error of existing alternatives. Together, these results suggest that expert trajectories are a broadly useful source of signal for assessing model capabilities, enabling reliable performance forecasting throughout the model development life cycle.

1 Introduction

Large language model (LLM) development requires making *comparative* decisions: which pretraining corpus is better, which post-training recipe increases performance on a target domain, and whether a new model architecture is better than the current frontier. A common signal for resolving such decisions has been cross-entropy loss, which scales smoothly with compute and extrapolates with remarkable fidelity (Hoffmann et al., 2022; Kaplan et al., 2020). However, the quantity we ultimately care about is downstream performance, not loss. Indeed, models with similar loss can exhibit sharply different downstream capabilities (Liu et al., 2023). Moreover, LLMs are increasingly judged on hard reasoning tasks where cross-entropy loss over generic text would offer little discriminative signal.

The natural response to resolve this discrepancy has been to fit scaling laws directly for downstream tasks, or to replace accuracy with smoother surrogates such as the likelihood of the correct answer (Bhagia et al., 2025; Brandfonbrener et al., 2025; Gadre et al., 2025; Hu et al., 2024; Ruan et al., 2024). These approaches have been shown to work well when we assume access to plentiful evaluations on a target task, often with a closed answer set, or candidate models that perform above chance. However,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

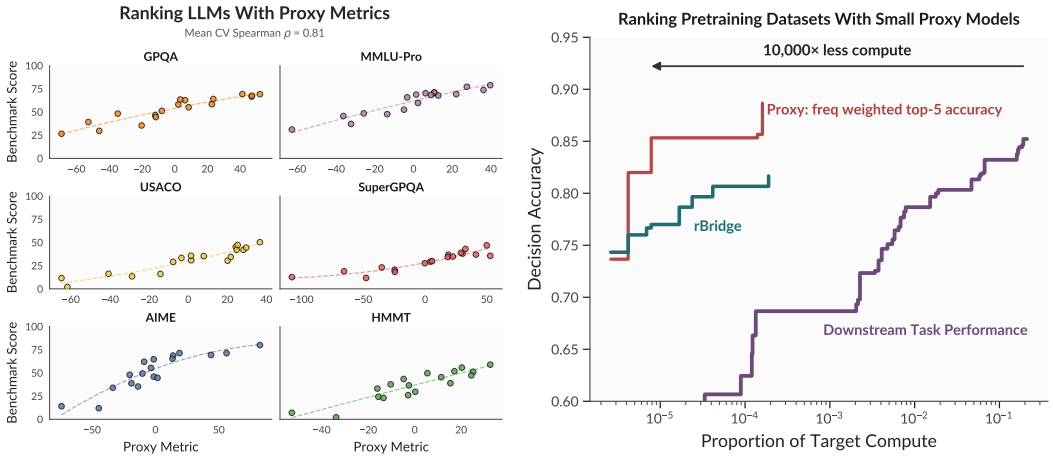


Figure 1: **Left.** Ranking models on held-out challenging reasoning tasks (as measured by mean CV Spearman ρ) using our linear RankSVM proxy. Our proxy uses features of the next-token prediction distributions of candidate models over expert reasoning traces. **Right:** Ranking 25 pretraining corpora for a target 1B LLM on the DataDecide testbed (Magnusson et al., 2025). Each method trains small proxy models (4M–90M) on each corpus and attempts to recover the ground-truth corpus ranking defined by the target 1B model's downstream accuracy. Decision accuracy is the fraction of corpus pairs ranked correctly. Our best proxy metric pushes the Pareto frontier, requiring roughly $10{,}000\times$ less compute to match the downstream evaluation baseline.

the regimes in which downstream forecasting is most valuable are precisely those in which these assumptions are not met. Evaluations at the frontier of LLMs are often expensive or inaccessible, e.g., requiring human experts, code execution, or an external experimental loop (Patwardhan et al., 2026; Wijk et al., 2025). Moreover, on hard reasoning tasks, small models or intermediate training checkpoints can all have indistinguishable accuracies (Phan et al., 2026), which leaves no ordinal signal to fit. Recent work has also cast doubt on the reliability of downstream scaling laws themselves, finding that many task-level fits break when asked to extrapolate (Lourie et al., 2025). The obstacle is not only that evaluation is expensive, but also that the quantities we can measure are often too sparse, too late, or too weakly tied to the reasoning process we hope to forecast.

In this paper, we propose a different approach for forecasting model performance: **compute proxy metrics based on the predictive distribution of the candidate model while it processes an expert solution.** Our intuition is the following. A final benchmark score records only whether the model succeeded or failed, but an expert trajectory contains a long sequence of local decisions, and a model that cannot yet solve a task may still assign high probability to the crucial steps once they appear in context. We build on this intuition by passing expert-written trajectories through the candidate in a single forward pass[1] and computing *token-level* statistics of its next-token distribution, e.g., entropy, top-k accuracy, rank of the expert token, etc. These statistics are aggregated with weights that emphasize important positions, such as rare tokens or tokens where the candidate is uncertain. Crucially, because the expert need only provide text, the same construction can use human solutions or traces from closed-weight frontier models.

We demonstrate our approach across three settings that mirror practical decisions in model development (Figure 1). ① In **cross-family model selection** (§4), where the goal is to rank heterogeneous models on a downstream task without direct evaluation, our best proxy ranks models on held-out reasoning benchmarks in close agreement with their true performance (mean Spearman $\rho = 0.81$, compared with 0.36 for cross-entropy loss). ② In **pretraining data selection** (§5), where the goal is to choose among candidate corpora before committing target-scale compute, our proxies reliably rank 25 diverse corpora using only small proxy models, achieving the same ranking quality as direct downstream evaluation at roughly $10{,}000\times$ less compute. ③ In **training-time forecasting** (§6), we show that proxy metrics follow smooth power laws along training trajectories, enabling extrapolation from early checkpoints, and that downstream accuracy is more predictable as a function of our proxy

[1] Our approach does not require generating from the candidate model, and hence is extremely efficient.

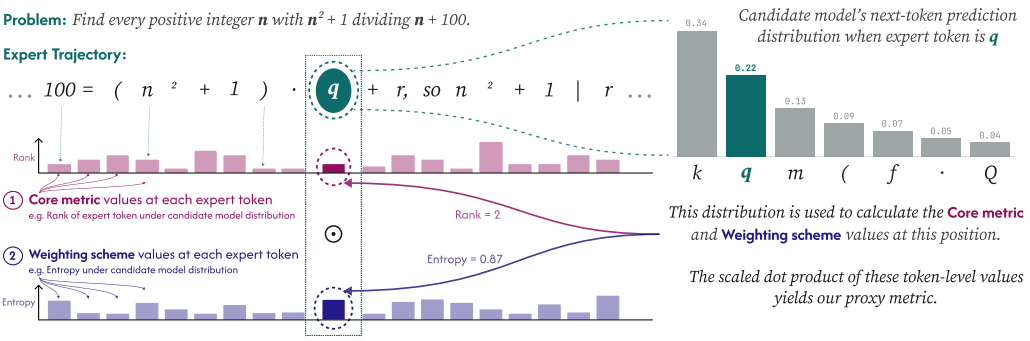


Figure 2: An illustration of our method. We use a candidate model's next-token prediction distribution at each token of an expert's trajectory to calculate our proxy metrics.

metric compared to cross-entropy loss or compute, roughly halving extrapolation error across an $18\times$ compute horizon. The pattern across all settings is the same: generic loss is smooth but task-agnostic, direct evaluation is task-specific but expensive and often uninformative at early training stages, and expert-trajectory proxies provide both smoothness and task-conditioning in a single forward pass.

2 Related Work

Scaling laws and downstream forecasting. Classical scaling laws predict pretraining loss as a function of compute, parameters, and data (Hoffmann et al., 2022; Kaplan et al., 2020). Subsequent work has attempted to extend this predictability to downstream task performance, whether by fitting accuracy directly against compute (Krajewski et al., 2026; Owen, 2024), mapping validation perplexity to downstream error (Gadre et al., 2025), decomposing the prediction into a compute-to-task-loss and task-loss-to-accuracy pipeline (Bhagia et al., 2025), fitting a latent capability axis over benchmark scores from public models (Ruan et al., 2024), or linking loss thresholds to capability emergence (Du et al., 2024). However, these approaches rest on assumptions that are often unmet in practice. Most require either a family of models trained at multiple scales or non-trivial benchmark scores across a broad population, neither of which is available when evaluating a new architecture or a single training run, or a task whose environment is inaccessible. Moreover, Lourie et al. (2025) find that only a minority of downstream scaling laws extrapolate reliably, and Liu et al. (2023) demonstrate that models with nearly identical loss can differ substantially in downstream performance. A separate line of work predicts smoother task-specific losses across distributions (Brandfonbrener et al., 2025; Mayilvahanan et al., 2025), avoiding the brittleness of accuracy, but this requires a closed answer set and does not resolve whether task loss tracks the downstream performance we ultimately care about. In this work, we focus on the problem of relative model ranking and show that proxy metrics derived from a single forward pass over expert trajectories can rank models on held-out tasks and across unseen models. Moreover, the benchmarks we consider such as graduate-level science (Rein et al., 2024) and olympiad programming (Shi et al., 2024), are precisely the reasoning tasks on which prior downstream scaling law approaches have not been tested.

Small-scale proxies for pretraining decisions. A separate line of work asks whether small proxy models can rank candidate pretraining corpora before committing target-scale compute. Prior approaches have selected domain weights (Xie et al., 2023) or data mixtures (Liu et al., 2025) using cheap small-scale runs. Magnusson et al. (2025) systematize this question with DataDecide, a controlled testbed of twenty-five pretraining corpora at fourteen proxy scales, and show that likelihood-style metrics predict the 1B target ranking at 0.01% of target compute. Koh et al. (2026) improve on this with rBridge, which reweights the proxy model's likelihood by the expert model's token-level probabilities, defining the previous state-of-the-art Pareto frontier on DataDecide. Our proxy metrics displace this frontier while requiring only the expert's tokens, not its probabilities, which opens the door to closed-weight models and human experts as sources of expert signal. An extended discussion of other related works is provided in Appendix D.

Table 1: Core metrics (left) and weighting schemes (right) constituting our proxy metric library. We write $p \coloneqq p_{\mathcal{M}}(\cdot \mid \mathbf{x}^{(i)}, \mathbf{y}^{(i)}_{<t})$, and y_t refers to the expert token. See Appendix B.1 for details.

(a) Core metrics

Core Metric (m_t)	**Definition**
Cross-entropy ($\mathcal{L}_t$)	$-\log p(y_t)$
Top-k accuracy	$\mathbf{1}[y_t \in \text{top-}k]$
	$k \in \{1, 2, 3, 5\}$
Entropy (H_t)	$-\sum_v p(v) \log p(v) / \log \lvert\mathcal{V}\rvert$
Rank	$\text{rank}(y_t)$
Reciprocal rank	$1/\text{rank}(y_t)$
Margin	$\max_v p(v) - p(y_t)$
Wrong-confidence	$\max_v p(v) \cdot \mathbf{1}[\text{rank}(y_t) > 1]$

(b) Weighting schemes

Scheme (w_t)	**Definition**
Uniform	1
Probability	$p(y_t)$
Expert-disagreement	$1 - p(y_t)$
Entropy	H_t
Inverse entropy	$1 - H_t$
Frequency	$\text{freq}(y_t)$
Inverse frequency	$1 - \text{freq}(y_t)$
Gaussian-NLL	$e^{-(\mathcal{L}_t - \bar{\mathcal{L}})^2 / 2\sigma^2_{\mathcal{L}}}$

3 Method

Our goal is to design a proxy signal that is both indicative of a candidate model's capability on a downstream task, and cheap to evaluate. We construct this signal from the candidate's predictive distribution over expert reasoning trajectories for the task instances as illustrated in Figure 2. The intuition is that a model whose distribution often matches the expert's reasoning at every step is one that has internalized how the task is solved, even when its own generation might have failed. We assume access to such expert trajectories, whether written by humans or by strong language models. Reference solutions are already standard for benchmarks of practical interest, and for frontier domains where current LLMs are not yet competent, e.g., drug discovery, protein design, or theorem proving, domain experts working in tandem with AI can provide high-quality reasoning traces. Obtaining such expert trajectories is a one-time cost per task that is amortized over every candidate model and checkpoint scored afterwards.

Preliminaries. Given a downstream task $\mathcal{T}$ with instances $\{\mathbf{x}^{(i)}\}_{i=1}^N$ and expert trajectories $\{\mathbf{y}^{(i)}\}_{i=1}^N$, we pass each $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ pair through the candidate model $\mathcal{M}$. At each token position t we obtain the predictive distribution $p_{\mathcal{M}}(\cdot \mid \mathbf{x}^{(i)}, \mathbf{y}^{(i)}_{<t})$, from which we calculate a set of *core metrics* m_t and *weighting schemes* w_t. The 10 core metrics (Table 1a) span three aspects of model–expert alignment: how often the model agrees with the expert, how concentrated its distribution is, and how confidently it errs when it disagrees. Because not every token position is equally diagnostic, we aggregate each core metric as a weighted average under eight weighting schemes (Table 1b) that emphasize different notions of token importance such as model uncertainty, disagreement with the expert, or token rarity.

Proxy metrics. Each (metric, weighting) pair defines one *proxy metric*, indexed by j. Given an instance $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ with trajectory of length $T^{(i)}$, the proxy metric value is

$$\Phi_j(\mathcal{M}; \mathbf{x}^{(i)}, \mathbf{y}^{(i)}) = \frac{\sum_{t=1}^{T^{(i)}} s_j \cdot m^{(i)}_{j,t} \cdot w^{(i)}_{j,t}}{\sum_{t=1}^{T^{(i)}} w^{(i)}_{j,t}}, \tag{1}$$

where $m^{(i)}_{j,t}$ is the core metric value and $w^{(i)}_{j,t}$ is the weighting scheme value, both determined by j, at position t of instance i, and $s_j \in \{+1, -1\}$ is a sign convention so that higher values indicate a better model (e.g., $s_j = -1$ for cross-entropy loss). The task-level proxy metric is the mean over instances,

$$\Phi_j(\mathcal{M}, \mathcal{T}) = \frac{1}{N} \sum_{i=1}^{N} \Phi_j(\mathcal{M}; \mathbf{x}^{(i)}, \mathbf{y}^{(i)}). \tag{2}$$

With 10 core metrics and 8 weightings, we obtain a library of 80 proxy metrics $\Phi_1, \ldots, \Phi_{80}$, each assigning a scalar score $\Phi_j(\mathcal{M}, \mathcal{T})$ to a candidate model on a task. When needed, we write

$\Phi(\mathcal{M}, \mathcal{T}) \in \mathbb{R}^{80}$ for the full vector. The entire library is extracted from a single forward pass per instance, making computation extremely cheap while providing 80 complementary views of how closely the candidate's predictive distribution tracks the expert's reasoning.

Computing proxy metrics in practice. In all experiments we compute proxy metrics on the last 1,000 tokens of each expert trajectory, which empirically outperforms using the full trace. We do not filter out trajectories that yield incorrect answers, simulating the realistic setting of imperfect experts. When multiple experts are available, the 80 proxy metrics are averaged across experts and across instances, yielding $\Phi(\mathcal{M}, \mathcal{T}) \in \mathbb{R}^{80}$ per (model, task) pair.

4 Cross-Family Model Selection: Ranking LLMs on Unseen Tasks

A recurring decision in LLM development is choosing which of several candidate models will perform best on a downstream task of interest. The candidates may span different architectures, pretraining corpora, or post-training recipes, and the target evaluation is often inaccessible, requiring expert graders, code execution, or domain-specific infrastructure (Patwardhan et al., 2026; Wijk et al., 2025) that cannot be assembled at decision time. In this section we study whether the proxy metrics from §3 can be used to rank a heterogeneous model population on downstream tasks.

4.1 Experimental Setup

We evaluate 18 reasoning-capable language models spanning six model families and six post-training recipes, with sizes ranging from 0.6B to 70B parameters (full list in Appendix B.2), on six challenging reasoning benchmarks: **AIME** 2025 (Zhang and Math-AI, 2025), **HMMT** (Balunovic et al., 2025), **GPQA** (Rein et al., 2024), **USACO** (Shi et al., 2024), **MMLU-Pro** (Wang et al., 2024b), and **SuperGPQA** (Team et al., 2025) (details provided in Appendix B.3). Together these cover competition math, graduate-level science, broad professional knowledge, and competitive code. Expert trajectories are generated by three frontier open-weight reasoning models: Kimi-K2.5 (Kimi Team et al., 2026), MiniMax-M2.5 (MiniMax, 2026), and Qwen3-Next-80B (Yang et al., 2025a). We measure ranking quality by Spearman rank correlation (ρ) between proxy scores and downstream accuracy.

A natural first question is whether any single proxy metric is universally predictive across these tasks. To investigate, we select the best proxy using downstream scores from all six benchmarks and the full model population, an oracle setting that upper-bounds what any selection procedure can achieve (Tables 5 and 6 in the Appendix). The best proxy metric attains a mean ρ of 0.62, with per-task correlations ranging from 0.43 to 0.81. No single metric dominates universally. A linear combination of just three proxy metrics, however, reaches $\rho = 0.88$, indicating that the signal is present in the library but distributed across complementary metrics. A practitioner, however, will not have scores on the target task. We therefore ask: *given downstream accuracy on a subset of tasks and models, can we find a proxy that generalizes to held-out tasks and unseen models?*

Evaluation protocol. We use a two-level resampling scheme. At the *task* level, we perform leave-2-tasks-out cross validation over the six benchmarks, producing $\binom{6}{2} = 15$ folds. In each fold the proxy is selected on the four held-in tasks and scored by the mean ρ on the two held-out tasks. At the *model* level, for each fold we further sample 60% of the models at random for selection and evaluate ranking correlation on the full model set. We repeat the model sampling with 20 fixed seeds and report mean $\pm$ std across seeds.

Ranking models. The 80 proxy metrics from §3 reduce ranking to a low-dimensional learning problem: we seek a function $f : \mathbb{R}^{80} \to \mathbb{R}$ whose induced ordering over candidate models tracks their downstream ordering. We compare four model classes of increasing capacity: a **univariate proxy** $f(\Phi) = \Phi_j$; a **3-sparse proxy** $f(\Phi) = \sum_{k=1}^{3} \alpha_k \Phi_{j_k}$; a **linear RankSVM** $f(\Phi) = \mathbf{w}^\top \Phi$ trained under a pairwise hinge loss (Herbrich et al., 2000); and an **RBF RankSVM** $f(\Phi) = \sum_i \alpha_i \, k(\Phi, \Phi_i)$ with a Gaussian kernel, trained under the same objective.

Proxy selection. For the univariate proxy, we select the index j that maximizes the mean Spearman ρ between $\Phi_j(\cdot, \mathcal{T})$ and downstream accuracy, averaged over tasks $\mathcal{T} \in \mathcal{T}_{\text{in}}$. For the 3-sparse proxy, we enumerate all $\binom{80}{3}$ index triplets and sweep a signed log-spaced grid of coefficient ratios in

Table 2: Leave-2-tasks-out cross-validated Spearman rank correlation (ρ) between each proxy and downstream benchmark accuracy. The per-task columns report the mean test $|\rho|$ across all folds in which that task was held out, and the **Mean** ρ column averages over all folds. For learned proxies, values are additionally averaged over 20 random model-subsampling seeds and reported as mean $\pm$ std across seeds. Best per column is shown in **bold**.

Proxy	**Mean** ρ	AIME	GPQA	HMMT	MMLU-Pro	SuperGPQA	USACO
FineWeb CE loss	0.36	0.37	0.25	0.25	0.52	0.54	0.25
rBridge	0.33	0.22	0.37	0.13	0.33	0.55	0.39
Univariate proxy	0.54±0.08	0.44±0.09	0.57±0.05	0.42±0.04	0.62±0.12	0.64±0.17	0.53±0.13
3-sparse proxy	0.78±0.06	0.74±0.09	0.86±0.07	0.68±0.11	0.82±0.06	0.84±0.07	0.73±0.08
RankSVM (RBF)	**0.81**±0.05	**0.75**±0.11	0.84±0.09	**0.76**±0.09	0.87±0.07	0.84±0.07	**0.78**±0.09
RankSVM (linear)	**0.81**±0.04	0.73±0.11	**0.87**±0.07	0.73±0.10	**0.88**±0.07	**0.88**±0.05	**0.78**±0.10

$[-10^3, 10^3]$, selecting the triplet and ratios that maximize the same objective. For both RankSVM variants, the parameters are fit on $(\mathcal{M}_i \succ \mathcal{M}_j)$ preference pairs induced by the downstream scores.

Baselines. Cross-entropy loss on generic text has been widely used as a predictor of downstream capability (Brandfonbrener et al., 2025; Du et al., 2024; Mayilvahanan et al., 2025). We compute CE loss over 10M tokens from randomly sampled FineWeb (Penedo et al., 2024) documents. We also evaluate rBridge (Koh et al., 2026), which computes expert-probability-weighted CE loss over expert reasoning chains, requiring access to the expert model's logprobs.

4.2 Results and Discussion

Table 2 reports the leave-2-tasks-out Spearman ρ for all four proxy models and the two loss-based baselines, aggregated across 15 folds and 20 model-subsampling seeds.

Loss-based baselines fail to rank models. CE loss on FineWeb achieves only $\rho = 0.36$, confirming that a scalar summary of fit to generic text carries little information about relative performance on reasoning tasks. rBridge, which reweights the likelihood along a frontier-model reasoning trace and has access to expert logprobs, fares no better at $\rho = 0.33$. These results are further illustrated in Figure 6 (left) in the Appendix, where we visualize the loss-based baselines against MMLU-Pro accuracy and find no coherent pattern. Moreover, training our rankers on loss-based features alone reaches at most $\rho = 0.43$ with three loss signals and 0.58 with all ten loss-based signals we compute (Appendix Section C.2).

Proxy models show high correlation with performance. The univariate proxy reaches $\rho = 0.54$, which is higher than the best loss-based baseline. The 3-sparse proxy pushes this to $\rho = 0.78$, and the full linear RankSVM reaches $\rho = 0.81$, with the RBF variant tied. Figure 1 (left) plots downstream accuracy against the linear RankSVM proxy score for each of the six benchmarks in a randomly sampled held-out fold. Across all six tasks the relationship is monotonic. Figure 6 (right) in the Appendix zooms into MMLU-Pro, showing that this monotonic relationship holds across different base families and post-training recipes. Figures 7 and 8 in the Appendix show that similar trends hold even when we consider 3 held-out tasks and a lesser percentage of models used for selection.

Ranking signal concentrates on a few proxy metrics. Figure 5 in the Appendix shows how often each proxy metric is selected across folds and seeds. The distribution concentrates on a handful of cells. For the univariate proxy, inverse-frequency-weighted top-1 accuracy, a signal indicating model agreement with expert on rare tokens, dominates. For the 3-sparse proxy, entropy-weighted entropy and frequency-weighted top-5 accuracy are most frequently selected, capturing model uncertainty at positions where the candidate is least committed. These are precisely the "forking tokens" that Wang et al. (2025) identify as driving the majority of RL training signal in reasoning chains. Note that this analysis characterizes where the ranking signal concentrates rather than explaining why the proxy works, which does not depend on the selected features being interpretable.

5 Pretraining Data Selection: Ranking Datasets with Smaller LMs

Before committing to a target-scale pretraining run, a team must choose among various candidate pretraining corpora. The target run may cost millions of dollars, so the decision should ideally rest on evidence collected at a fraction of that budget. The standard approach is to train small proxy models on each candidate corpus and rank them by downstream benchmark accuracy or cross-entropy loss (Magnusson et al., 2025). But at small scale, benchmark accuracy is noisy or at chance, and cross-entropy loss on generic text correlates poorly with downstream performance. In this section, we ask: *can our proxy metrics, computed over small proxy models, rank pretraining corpora without ever evaluating on downstream tasks?*

5.1 Experimental Setup

We use the DataDecide testbed (Magnusson et al., 2025), which consists of twenty-five candidate pretraining corpora, each used to train proxy models at scales ranging from 4M to 90M parameters, together with realized 1B-parameter target models trained on the same corpora. The ground-truth ranking of the twenty-five corpora is defined by the mean downstream accuracy of the corresponding 1B target models on the **OLMES** suite of ten multiple-choice benchmarks (Gu et al., 2025).

Evaluation metric. Following Magnusson et al. (2025), we measure ranking quality by *decision accuracy*, which is the fraction of corpus pairs in which the proxy model agrees with the target-scale ranking. It can be formalized as follows. Let $\mathcal{P}$ be the set of all pretraining corpus pairs (A, B) with observed mean OLMES performance for the 1B target LLM as y_A, y_B respectively, and the predicted performance by the proxy model is denoted by $\hat{y}_A, \hat{y}_B$, respectively, then decision accuracy is:

$$\frac{1}{|\mathcal{P}|} \sum_{(A,B)\in\mathcal{P}} \mathbb{I}\big(\text{sign}(\hat{y}_A - \hat{y}_B) = \text{sign}(y_A - y_B)\big) \tag{3}$$

Estimating compute. We measure the cost of a ranking method by the fraction of the 1B target's training FLOPs consumed by the proxy model, following the standard approximation FLOPs $= 6ND$ (Kaplan et al., 2020). A method that ranks corpora using only 4M-parameter proxy models operates at roughly 10^{-5} of the target compute.

Method. We evaluate each univariate proxy metric on every (corpus, model-size) pair in DataDecide, producing a ranking of the 25 corpora at each compute budget. Due to the simpler nature of the OLMES tasks compared to the tasks in §4, here we use chain-of-thought (CoTs) produced by Llama-3.3-70B (Llama Team et al., 2024) as the expert trajectories for computing our proxy metrics. We compare against two baselines. The first is the DataDecide downstream task performance baseline, which scores each corpus by the mean OLMES accuracy of the corresponding proxy model. The second is rBridge (Koh et al., 2026), which reweighs the proxy model's negative log-likelihood at each token position of the expert's CoT by the expert model's token-level probabilities.

5.2 Results and Discussion

Figure 1 (right) plots decision accuracy against compute budget for all methods. Prior to this work, rBridge defined the state-of-the-art Pareto frontier on DataDecide.

New state-of-the-art on DataDecide. Our best proxy metric, frequency-weighted top-5 accuracy, pushes this frontier. At roughly 10^{-5} of the target compute, it reaches decision accuracy above 0.85, and at matched compute budgets it dominates rBridge at almost every point where both methods are defined. To reach comparable decision accuracy using the downstream performance baseline, one must scale proxy models to a budget exceeding 10^{-1} of the target, which is roughly $10{,}000$ times more compute. This advantage comes under *strictly weaker assumptions*: our proxy metrics require only the expert's discrete tokens, not its logprobs, enabling closed-weight frontier models as well as human experts to be valid sources.

Why proxies are suitable for small models. We believe that the reason proxy metrics discriminate among corpora at these scales is that a model which cannot solve a problem can still track the CoT written by an expert. Benchmark accuracy requires the model to produce a correct answer, which

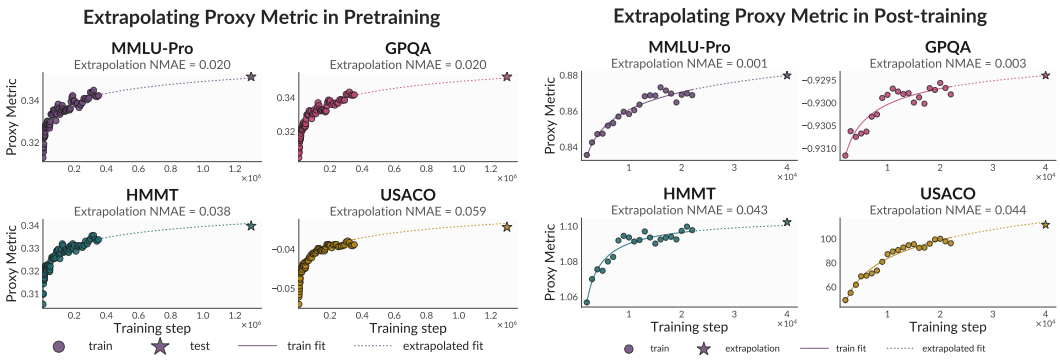


Figure 3: **Extrapolating proxy metrics along the training trajectory.** *Left:* pretraining checkpoints of OLMo-3-7B on four reasoning benchmarks. *Right:* post-training checkpoints of OLMo-3-7B-Think on four reasoning benchmarks. Filled markers are the training window, stars are held-out checkpoints, solid curves are power-law fits from the training window, and dashed curves are extrapolations. The plots for other benchmarks are provided in Figures 10 and 11 in the Appendix.

may not be possible at such small scales. But the model's token-level compatibility with an expert trajectory varies measurably across corpora long before any benchmark score exits the noise floor.

6 Training-time Forecasting: Proxy Metrics Facilitate Extrapolation

The previous sections showed that proxy metrics can rank models and pretraining corpora for downstream task performance. But a practitioner often needs to forecast end-of-training capability with a fraction of the training compute expended. This requires that the proxy evolve predictably with training compute, so that a fit from early checkpoints can be extrapolated to later ones.

6.1 Extrapolating Proxy Metric During Training

First, we test whether *we can find proxy metrics that follow a simple functional form along the training trajectory*.

Experimental setup. We study two trajectories: the pretraining checkpoints of OLMo-3-7B (Olmo et al., 2026), evaluated on all six reasoning benchmarks from §4, and the post-training checkpoints of OLMo-3-7B-Think, evaluated on five (excluding SuperGPQA, on which the model shows negligible improvement). At each checkpoint we compute univariate proxy metrics and fit a power law $f(t) = a - bt^{-c}$. We select the proxy via an inner split: fit on the first $k = 50\%$ of checkpoints, choose the proxy whose extrapolation to the remainder has the lowest normalized mean absolute error (NMAE), defined as mean absolute error divided by the proxy's range over training. The selected proxy is then refit on the full window and evaluated on the held-out checkpoint.

Results and discussion. The results are shown in Figure 3. In pretraining, we fit on checkpoints up to 350K steps and extrapolate to $\sim 4\times$ the compute. On all six benchmarks the best selected proxy follows a smooth power law, with mean NMAE of 0.03 across tasks. In post-training, the proxy extrapolates to nearly $2\times$ the training compute with mean NMAE of 0.038. As a sanity check, the best selected proxies also correlate strongly with downstream accuracy at the post-training checkpoints (mean Spearman $\rho = 0.84$ illustrated in Appendix Figure 12), confirming that the extrapolated quantity tracks the ranking we care about.

6.2 Downstream Accuracy Is More Predictable from Proxy Metrics

Existing approaches for predicting downstream task performance fit exponential functions over validation loss (Gadre et al., 2025) or sigmoids over log-compute (Owen, 2024), but Lourie et al. (2025) show that the majority of such fits fail to extrapolate reliably. Hence, we next ask: *is downstream accuracy more predictable as a function of our proposed proxy metrics compared to cross-entropy loss or training compute?*

Table 3: Extrapolation RMSE for predicting downstream accuracy at 1.4M steps from checkpoints up to 80K steps ($\sim 18\times$). **Bold** marks the lowest RMSE per task.

Task	Proxy	CE Loss	Compute
HellaSwag	**0.003**	0.094	0.105
BoolQ	0.037	0.111	**0.034**
SocialIQA	0.016	**0.011**	0.025
Winogrande	**0.010**	0.078	0.018
ARC Challenge	**0.022**	0.131	0.071
ARC Easy	0.049	**0.001**	0.036
CommonsenseQA	**0.004**	0.023	0.129
MMLU	**0.001**	0.036	0.058
OpenBookQA	0.069	0.086	**0.000**
PIQA	0.029	**0.025**	0.068
Mean	**0.024**	0.059	0.055

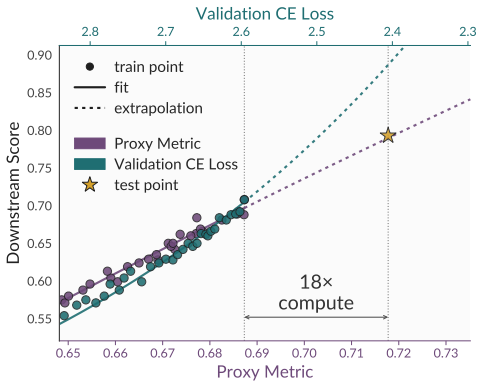


Figure 4: Extrapolating HellaSwag accuracy along the OLMo-3-7B pretraining trajectory. The proxy power-law fit (RMSE $= 0.003$) tracks the target far more closely than the CE loss exponential (RMSE $= 0.09$).

Experimental setup. We use pretraining checkpoints of OLMo-3-7B (Olmo et al., 2026) across ten OLMES benchmarks (Gu et al., 2025). For each benchmark we fit a predictor→accuracy curve on checkpoints up to 80,000 steps and evaluate at the final checkpoint (1.4M steps, roughly $18\times$ the fitting horizon). We compare three predictors: (1) CE loss, via an exponential from FineWeb validation loss to accuracy (Gadre et al., 2025); (2) compute, via a sigmoid against $\log_{10}(\text{steps})$ (Owen, 2024); and (3) proxy metric, via a power law, selected by the inner-split protocol of §6.1.

Results and discussion. Table 3 summarizes the extrapolation RMSE across all ten OLMES benchmarks. The proxy-based fit achieves a mean RMSE of 0.024, roughly half that of the CE loss predictor (0.059) and the compute-based predictor (0.055). Even on the tasks where a baseline achieves the lowest RMSE, the proxy-based fit is comparable. Figure 4 illustrates the contrast on HellaSwag: within the training window all fits are strong ($R^2 > 0.9$), but at extrapolation the CE loss fit overshoots (RMSE $= 0.09$) while the proxy fit tracks the target closely (RMSE $= 0.003$). We attribute this to the fact that CE loss and compute are both task-agnostic predictors, whereas the proxy is conditioned on the target task through the expert trajectories, and so its relationship with downstream accuracy is less likely to drift as training progresses.

7 Conclusion

We proposed proxy metrics computed from token-level statistics of a candidate model's forward pass over expert reasoning trajectories and showed that they carry substantial information about downstream task capability. ① For **cross-family model selection** under leave-two-tasks-out cross-validation across 6 reasoning benchmarks and 18 models spanning 6 base families, a linear ranker over these metrics achieves Spearman $\rho = 0.81$, compared to $\rho = 0.36$ for cross-entropy loss. ② For **pretraining data selection** on the DataDecide testbed, a single univariate proxy ranks 25 pretraining corpora with decision accuracy above 0.85 at roughly 10^{-5} of target compute, displacing the prior Pareto frontier while requiring only the expert's tokens, not its logprobs. ③ For **training-time forecasting** along the pretraining trajectory of OLMo-3-7B, proxy-to-accuracy fits extrapolate downstream performance across an $18\times$ compute horizon with roughly half the error of loss-based or compute-based alternatives. Across all three settings, the mechanism is the same: the candidate model's token-level distribution over an expert trajectory provides a dense, task-conditioned signal that inherits the smoothness of loss while remaining grounded in the reasoning process that the downstream evaluation is designed to measure.

Several important questions remain. Our experiments establish that proxy metrics work across model families and training stages, but the boundaries of this finding are not yet fully mapped: whether the same construction generalizes to mixture-of-experts architectures, to task types beyond reasoning such as long-context retrieval and agentic evaluation, and whether the power-law fits that enable

extrapolation hold across model scales and not only along a single training trajectory. A related question concerns the expert itself. We have assumed access to high-quality reasoning traces, but in practice expert quality will vary, and understanding how proxy signal degrades with weaker or noisier experts is important for deployment. Finally, we have treated the eighty proxy metrics as a fixed library and selected among them post hoc. Learning the aggregation function end-to-end, or designing metrics that target specific failure modes of downstream evaluation, could close the remaining gap between the oracle upper bound and the cross-validated performance we report.

8 Limitations

Our experimental scope has boundaries that should inform how the results are interpreted.

Model and checkpoint coverage. The cross-family experiment evaluates 18 models spanning six base families and six post-training recipes. This covers most of the major open-weight reasoning-capable models available while this work was done, but the population is still modest. The training-time forecasting experiments (§6.1, §6.2) rely entirely on checkpoints from OLMo-3-7B and OLMo-3-7B-Think. We chose OLMo because it publicly releases both pretraining and post-training checkpoints at the granularity needed for power-law fitting. This means the extrapolation results are established on a single architecture and scale, and whether the same power-law regularity holds across model sizes, architectures, or mixture-of-experts models remains an open empirical question.

No universal proxy metric. No single proxy metric dominates across all settings. The best univariate proxy for cross-family ranking is not the same as the best proxy for data selection or for extrapolating any particular benchmark. In practice, the specific proxy metric that works best will depend on the available held-in benchmarks, the model population, and hyperparameters of the selection procedure such as the inner-split fraction and the training-window size used for fitting. We have not conducted a systematic sensitivity analysis over these choices. The oracle gap (Table 5) shows that the 80-dimensional library contains enough signal to reach $\rho = 0.88$ with just three features, but extracting that signal reliably under realistic selection constraints is challenging.

Task coverage. The cross-family ranking experiment focuses on challenging reasoning benchmarks, while the data selection and downstream extrapolation experiments use OLMES, a suite of non-reasoning multiple-choice tasks. This split is driven by resource constraints: DataDecide provides a controlled testbed only for OLMES, and, to the best of our knowledge, no comparable infrastructure exists for reasoning benchmarks at the time of writing. As a result, we have not demonstrated pretraining data selection for reasoning tasks or downstream extrapolation on hard reasoning benchmarks. More broadly, we have also not tested the method on generative tasks (e.g., open-ended writing, translation), long-context benchmarks, or agentic evaluations.

Acknowledgments

Arkil is supported by the 2026 Google PhD Fellowship, and the Canada Graduate Scholarships (Doctoral) funded by the Natural Sciences and Engineering Research Council (NSERC) [funding reference no. 601601]. We thank our colleagues at Mila and McGill University for helpful discussions and for providing valuable feedback.

References

Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max Marion, Matthew L Leavitt, and Mansheej Paul. Perplexed by perplexity: Perplexity-based data pruning with small reference models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1GTARJhxtq. (Cited on page 33)

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025. (Cited on page 21)

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating LLMs on uncontaminated math competitions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=y0zL9IZxZ7. (Cited on pages 5 and 22)

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt, 2024. URL https://arxiv.org/abs/2404.10102. (Cited on page 31)

Akshita Bhagia, Jiacheng Liu, Alexander Wettig, David Heineman, Oyvind Tafjord, Ananya Harsh Jha, Luca Soldaini, Noah A. Smith, Dirk Groeneveld, Pang Wei Koh, Jesse Dodge, and Hannaneh Hajishirzi. Establishing task scaling laws via compute-efficient model ladders. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=FeAM2RVO8l. (Cited on pages 1, 3, and 32)

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439, Apr. 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239. (Cited on page 22)

David Brandfonbrener, Nikhil Anand, Nikhil Vyas, Eran Malach, and Sham M. Kakade. Loss-to-loss prediction: Scaling laws for all datasets. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=1Avb4jYjLb. Featured Certification. (Cited on pages 1, 3, 6, and 32)

Yangyi Chen, Binxuan Huang, Yifan Gao, Zhengyang Wang, Jingfeng Yang, and Heng Ji. Scaling laws for predicting downstream performance in LLMs. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=PJUbMDkQVY. (Cited on page 32)

Leshem Choshen, Yang Zhang, and Jacob Andreas. A hitchhiker's guide to scaling law estimation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=KugSHTH0c8. (Cited on page 31)

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *NAACL*, pages 2924–2936, Minneapolis, Minnesota, June 2019. doi: 10.18653/v1/N19-1300. (Cited on page 22)

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, 2018. URL http://arxiv.org/abs/1803.05457. (Cited on page 22)

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards, 2025. URL https://arxiv.org/abs/2502.01456. (Cited on page 33)

DeepSeek-AI et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z. (Cited on page 21)

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. Understanding emergent abilities of language models from the loss perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=35DAviqMFo. (Cited on pages 3, 6, and 32)

Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. What is wrong with perplexity for long-context language modeling? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fL4qWkSmtM. (Cited on page 33)

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Luca Soldaini, Jenia Jitsev, Alex Dimakis, Gabriel Ilharco, Pang Wei Koh, Shuran Song, Thomas Kollar, Yair Carmon, Achal Dave, Reinhard Heckel, Niklas Muennighoff, and Ludwig Schmidt. Language models scale reliably with over-training and on downstream tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iZeQBqJamf. (Cited on pages 1, 3, 8, 9, 30, 31, and 32)

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. Demystifying prompts in language models via perplexity estimation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.679. URL https://aclanthology.org/2023.findings-emnlp.679/. (Cited on page 32)

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. OLMES: A standard for language model evaluations. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5020–5048, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.282. URL https://aclanthology.org/2025.findings-naacl.282/. (Cited on pages 7 and 9)

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.507. URL https://aclanthology.org/2023.emnlp-main.507/. (Cited on page 33)

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. (Cited on page 22)

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large-Margin Classifiers*. The MIT Press, 09 2000. ISBN 9780262283977. doi: 10.7551/mitpress/1113.003.0010. URL https://doi.org/10.7551/mitpress/1113.003.0010. (Cited on page 5)

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=iBBcRUlOAPR. (Cited on pages 1, 3, and 31)

Shengding Hu, Xin Liu, Xu Han, Xinrong Zhang, Chaoqun He, Weilin Zhao, Yankai Lin, Ning Ding, Zebin Ou, Guoyang Zeng, Zhiyuan Liu, and Maosong Sun. Predicting emergent abilities with infinite resolution evaluation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lDbjooxLkD. (Cited on pages 1 and 32)

Yuzhen Huang, Jinghan Zhang, Zifei Shan, and Junxian He. Compression represents intelligence linearly. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=SHMj84U5SH. (Cited on page 32)

Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Koyejo. Scaling laws for downstream task performance in machine translation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vPOMTkmSiu. (Cited on page 32)

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361. (Cited on pages 1, 3, 7, and 31)

Kimi Team et al. Kimi k2.5: Visual agentic intelligence, 2026. URL https://arxiv.org/abs/2602.02276. (Cited on page 5)

Woosung Koh, Juyoung Suk, Sungjun Han, Se-Young Yun, and Jay Shin. Predicting LLM reasoning performance with small proxy model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=JSE40ljyKm. (Cited on pages 3, 6, 7, 22, and 33)

Jakub Krajewski, Amitis Shidani, Dan Busbridge, Sam Wiseman, and Jason Ramapuram. Revisiting the scaling properties of downstream metrics in large language model training. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YnJ2s4WeNF. (Cited on pages 3 and 32)

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023. (Cited on page 22)

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi. (Cited on page 33)

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang,

Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584. (Cited on page 21)

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22188–22214. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/liu23ao.html. (Cited on pages 1, 3, and 32)

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=5BjQOUXq7i. (Cited on pages 3 and 33)

Llama Team et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783. (Cited on pages 7 and 21)

Nicholas Lourie, Michael Y. Hu, and Kyunghyun Cho. Scaling laws are unreliable for downstream tasks: A reality check. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16167–16180, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.877. URL https://aclanthology.org/2025.findings-emnlp.877/. (Cited on pages 2, 3, 8, and 32)

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. Improve mathematical reasoning in language models by automated process supervision, 2024. URL https://arxiv.org/abs/2406.06592. (Cited on page 33)

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Pete Walsh, Yanai Elazar, Kyle Lo, Dirk Groeneveld, Iz Beltagy, Hannaneh Hajishirzi, Noah A. Smith, Kyle Richardson, and Jesse Dodge. Paloma: A benchmark for evaluating language model fit. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=yjj8ele147. (Cited on page 32)

Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. Datadecide: How to predict best pretraining data with small experiments. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=p9YlQPF8fE. (Cited on pages 2, 3, 7, and 33)

Prasanna Mayilvahanan, Thaddäus Wiedemer, Sayak Mallick, Matthias Bethge, and Wieland Brendel. LLMs on the line: Data determines loss-to-loss scaling laws. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=IVUjRWnU6c. (Cited on pages 3, 6, and 32)

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *EMNLP*, pages 2381–2391, Brussels, Belgium, October-November 2018. doi: 10.18653/v1/D18-1260. (Cited on page 22)

MiniMax. Minimax m2.5: Built for real-world productivity, 2026. URL https://www.minimax.io/news/minimax-m25. (Cited on page 5)

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=j5BuTrEj35. (Cited on page 31)

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2026. URL https://arxiv.org/abs/2512.13961. (Cited on pages 8, 9, and 21)

OpenAI et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774. (Cited on page 22)

David Owen. How predictable is language model benchmark performance?, 2024. URL https://arxiv.org/abs/2401.04757. (Cited on pages 3, 8, 9, 28, 30, 31, and 32)

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf. (Cited on page 22)

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simon Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Samuel Miserendino, Gildas Chabot, David Li, Patrick Chao, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. GDPval: Evaluating AI model performance on real-world economically valuable tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hcuEdq6eKD. (Cited on pages 2 and 5)

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=n6SCkn2QaG. (Cited on page 6)

Long Phan, Alice Gatti, Nathaniel Li, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dan Hendrycks, Ziwen Han, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Aakaash Nattanmai, Gordon McKellips, Anish Cheraku, Asim Suhail, Ethan Luo, Marvin Deng, Jason Luo, Ashley Zhang, Kavin Jindel, Jay Paek, Kasper Halevy, Allen Baranov, Michael Liu, Advaith Avadhanam, David Zhang, Vincent Cheng, Brad Ma, Evan Fu, Liam Do, Joshua Lass, Hubert Yang, Surya Sunkari, Vishruth Bharath, Violet Ai, James Leung, Rishit Agrawal, Alan Zhou, Kevin Chen, Tejas Kalpathi, Ziqi Xu, Gavin Wang, Tyler Xiao, Erik Maung, Sam Lee, Ryan Yang, Roy Yue, Ben Zhao, Julia Yoon, Xiangwan Sun, Aryan Singh, Clark Peng, Tyler Osbey, Taozhi Wang, Daryl Echeazu, Timothy Wu, Spandan Patel, Vidhi Kulkarni, Vijaykaarti Sundarapandiyan, Andrew Le, Zafir Nasim, Srikar Yalam, Ritesh Kasamsetty, Soham Samal, David Sun, Nihar Shah, Abhijeet Saha, Alex Zhang, Leon Nguyen, Laasya Nagumalli, Kaixin Wang, Aidan Wu, Anwith Telluri, Summer Yue, Alexandr Wang, Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala,

Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C. Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G. Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow, Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod, Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang, Paolo Rissone, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad Hogg, Carlo Bosio, Brian P. Coppola, Julian Salazar, Jaehyeok Jin, Rafael Sayous, Stefan Ivanov, Philippe Schwaller, Shaipranesh Senthilkumar, Andres M. Bran, Andres Algaba, Kelsey Van den Houte, Lynn Van Der Sypt, Brecht Verbeken, David Noever, Alexei Kopylov, Benjamin Myklebust, Bikun Li, Lisa Schut, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Tong Yang, John Maar, Julian Wykowski, Mart Oller, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Ariel Ghislain Kemogne Kamdoum, Alvin Jin, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Gongbo Sun, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Joseph M. Cavanagh, Daofeng Li, Jiawei Shen, Donato Crisostomi, Wenjin Zhang, Ali Dehghan, Sergey Ivanov, David Perrella, Nurdin Kaparov, Allen Zang, Ilia Sucholutsky, Arina Kharlamova, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Shankar Sivarajan, Dan Bar Hava, Aleksey Kuchkin, David Holmes, Alexandra Rodriguez-Romero, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Zakayo Kazibwe, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Sara Fish, Veit Elser, Tobias Kreiman, Victor Efren Guadarrama Vilchis, Immo Klose, Ujjwala Anantheswaran, Adam Zweiger, Kaivalya Rawal, Jeffery Li, Jeremy Nguyen, Nicolas Daans, Haline Heidinger, Maksim Radionov, Václav Rozhoň, Vincent Ginis, Christian Stump, Niv Cohen, Rafał Poświata, Josef Tkadlec, Alan Goldfarb, Chenguang Wang, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Ryan Stendall, Jamie Tucker-Foltz, Jack Stade, T. Ryan Rogers, Tom Goertzen, Declan Grabb, Abhishek Shukla, Alan Givré, John Arnold Ambay, Archan Sen, Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess ai capabilities. *Nature*, 649(8099):1139–1146, 2026. doi: 10.1038/s41586-025-09962-4. URL https://doi.org/10.1038/s41586-025-09962-4. (Cited on page 2)

Felipe Maia Polo, Seamus Somerstep, Leshem Choshen, Yuekai Sun, and Mikhail Yurochkin. Sloth: scaling laws for LLM skills to predict multi-benchmark performance across families. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=9GN5Jsa3lv. (Cited on page 32)

Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving discrepancies in compute-optimal scaling of language models. In *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML 2024)*, 2024. URL https://openreview.net/forum?id=zhCBrgaQZ0. (Cited on page 31)

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98. (Cited on pages 3, 5, and 22)

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of langauge model performance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=On5WIN7xyD. (Cited on pages 1, 3, and 32)

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740, Apr. 2020. doi: 10.1609/aaai.v34i05.6399. URL https://ojs.aaai.org/index.php/AAAI/article/view/6399. (Cited on page 22)

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *EMNLP*, pages 4463–4473, Hong Kong, China, November 2019. doi: 10.18653/v1/D19-1454. (Cited on page 22)

Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=0bmXrtTDUu. (Cited on page 31)

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ITw9edRDlD. (Cited on page 32)

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. Why has predicting downstream capabilities of frontier AI models with scale remained elusive? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=I1NtlLvJal. (Cited on page 32)

Ben Shi, Michael Tang, Karthik R Narasimhan, and Shunyu Yao. Can language models solve olympiad programming? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kGa4fMtP9l. (Cited on pages 3, 5, and 22)

Charlie Victor Snell, Eric Wallace, Dan Klein, and Sergey Levine. Predicting emergent capabilities by finetuning. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=vL8BIGuFTF. (Cited on page 32)

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=UmvSlP-PyV. (Cited on page 33)

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *NAACL*, pages 4149–4158, Minneapolis, Minnesota, June 2019. doi: 10.18653/v1/N19-1421. (Cited on page 22)

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Tran, Dani Yogatama, and Donald Metzler. Scaling laws vs model architectures: How does inductive bias influence scaling? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 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/. (Cited on page 32)

P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shawn Gavin, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, Yizhi Li, Yunwen Li, David Ma, Yuansheng Ni, Haoran Que, Qiyao Wang, Zhoufutu Wen, Siwei Wu, Tyshawn Hsing, Ming Xu, Zhenzhu Yang, Zekun Moore Wang, Junting Zhou, Yuelin Bai, Xingyuan Bu, Chenglin Cai, Liang Chen, Yifan Chen, Chengtuo Cheng, Tianhao Cheng, Keyi Ding, Siming Huang, Yun Huang, Yaoru Li, Yizhe Li, Zhaoqun Li, Tianhao Liang, Chengdong Lin, Hongquan Lin, Yinghao Ma, Tianyang Pang, Zhongyuan Peng, Zifan Peng, Qige Qi, Shi Qiu, Xingwei Qu, Shanghaoran Quan, Yizhou Tan, Zili Wang, Chenqing Wang, Hao Wang, Yiya Wang, Yubo Wang, Jiajun Xu, Kexin Yang, Ruibin Yuan, Yuanhao Yue, Tianyang Zhan, Chun Zhang, Jinyang Zhang, Xiyue Zhang, Xingjian Zhang, Yue Zhang, Yongchi Zhao, Xiangyu Zheng, Chenghua Zhong, Yang Gao, Zhoujun Li, Dayiheng Liu, Qian Liu, Tianyu Liu, Shiwen Ni, Junran Peng, Yujia Qin, Wenbo Su, Guoyin Wang, Shi Wang, Jian Yang, Min Yang, Meng Cao, Xiang Yue, Zhaoxiang Zhang, Wangchunshu Zhou, Jiaheng Liu, Qunshu Lin, Wenhao Huang, and Ge Zhang. Supergpqa: Scaling llm evaluation across 285 graduate disciplines, 2025. URL https://arxiv.org/abs/2502.14739. (Cited on pages 5 and 22)

Tristan Thrush, Christopher Potts, and Tatsunori Hashimoto. Improving pretraining data using perplexity correlations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=huuKoVQnB0. (Cited on page 33)

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback, 2022. URL https://arxiv.org/abs/2211.14275. (Cited on page 33)

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Validad Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 furious (COLM's version). In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=2ezugTT9kU. (Cited on page 29)

Boxin Wang, Chankyu Lee, Nayeon Lee, Sheng-Chieh Lin, Wenliang Dai, Yang Chen, Yangyi Chen, Zhuolin Yang, Zihan Liu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nemotron-cascade: Scaling cascaded reinforcement learning for general-purpose reasoning models, 2026. URL https://arxiv.org/abs/2512.13607. (Cited on page 21)

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/. (Cited on page 33)

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning, 2025. URL https://arxiv.org/abs/2506.01939. (Cited on page 6)

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw. (Cited on page 33)

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b. URL https://openreview.net/forum?id=y10DM6R2r3. (Cited on pages 5 and 22)

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022a. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification. (Cited on page 32)

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=_VjQlMeSB_J. (Cited on page 33)

Hjalmar Wijk, Tao Roa Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Joshua M Clymer, Jai Dhyani, Elena Ericheva, Katharyn Garcia, Brian Goodrich, Nikola Jurkovic, Megan Kinniment, Aron Lajko, Seraphina Nix, Lucas Jun Koba Sato, William Saunders, Maksym Taran, Ben West, and Elizabeth Barnes. RE-bench: Evaluating frontier AI r&d capabilities of language model agents against human experts. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=3rB0bVU6z6. (Cited on pages 2 and 5)

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6. (Cited on page 22)

Mitchell Wortsman, Peter J Liu, Lechao Xiao, Katie E Everett, Alexander A Alemi, Ben Adlam, John D Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=d8w0pmvXbZ. (Cited on page 33)

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. Training trajectories of language models across scales. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 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/. (Cited on page 32)

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=lXuByUeHhd. (Cited on pages 3 and 33)

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388. (Cited on pages 5 and 21)

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. Qwen2.5-1m technical report. *arXiv preprint arXiv:2501.15383*, 2025b. (Cited on page 21)

Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Bx6qKuBM2AD. (Cited on page 33)

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=8ThnPFhGm8. (Cited on page 33)

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STar: Bootstrapping reasoning with reasoning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_3ELRdg2sgI. (Cited on page 33)

Eric Zelikman, Georges Raif Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah Goodman. Quiet-STar: Language models can teach themselves to think before speaking. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=oRXPiSOGH9. (Cited on page 33)

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *ACL*, pages 4791–4800, Florence, Italy, July 2019. doi: 10.18653/v1/P19-1472. (Cited on page 22)

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2025, 2025. (Cited on pages 5 and 21)

A Broader Impact Statement

This paper introduces a methodology for forecasting the relative performance of language models on downstream tasks. The most direct consequence of this work, if adopted, would be a reduction in the computational cost of model development decisions. Pretraining runs, post-training recipes, and data mixture choices currently require expensive evaluations or full-scale training to compare. By enabling such comparisons at a fraction of the cost, our method could lower the resource barrier for teams and organizations that cannot afford extensive evaluation infrastructure, and reduce the energy and hardware costs associated with training models that turn out to underperform.

B Details of Experiments

B.1 Proxy metric definitions

Below, we describe the exact formulation of each of our core metrics and weighting schemes.

Core metrics. At each position t we compute the **(1) cross-entropy loss** $\mathcal{L}_t = -\log p_{\mathcal{M}}(y_t \mid \mathbf{x}, y_{<t})$, **(2) top-$k$ accuracy** $A_{k,t} = \mathbf{1}[y_t \in \text{top-}k(\mathcal{M}, t)]$ for $k \in \{1, 2, 3, 5\}$, **(3) entropy** $H_t = -\sum_v p_{\mathcal{M}}(v \mid \mathbf{x}, y_{<t}) \log p_{\mathcal{M}}(v \mid \mathbf{x}, y_{<t}) / \log |\mathcal{V}|$ (vocabulary-normalized), **(4) rank** of y_t, **(5) reciprocal rank** $1/\text{rank}(y_t)$, **(6) margin** $\max_v p_{\mathcal{M}}(v \mid \mathbf{x}, y_{<t}) - p_{\mathcal{M}}(y_t \mid \mathbf{x}, y_{<t})$, and **(7) wrong-confidence mass** $\max_v p_{\mathcal{M}}(v \mid \mathbf{x}, y_{<t}) \cdot \mathbf{1}[\text{rank}(y_t) > 1]$ that the model places on the wrong token when it misses. Together these metrics span three regimes of model-expert alignment: how often the model agrees with the expert, how peaked the model is, and how badly the model errs when it errs.

Weighting schemes. Not every token position is equally important, e.g., function words and punctuation are predicted well by most models and drown out signal. We therefore introduce per-token weights $w_t \geq 0$ and aggregate each metric as a weighted average. The eight schemes we consider are: **(1) uniform** ($w_t = 1$), **(2) probability** ($w_t = p_{\mathcal{M}}(y_t)$), **(3) expert-disagreement** ($w_t = 1 - p_{\mathcal{M}}(y_t)$), **(4) entropy** ($w_t = H_t$, where H_t is vocabulary-normalized so $H_t \in [0, 1]$), **(5) inverse entropy** ($w_t = 1 - H_t$), **(6) frequency** ($w_t = \text{freq}(y_t)$, where $\text{freq}(y_t)$ is the frequency of the token in the entire expert trajectories corpus $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$), **(7) inverse frequency** ($w_t = 1 - \text{freq}(y_t)$), and **(8) Gaussian-NLL** kernel that isolates positions near the typical loss level ($w_t = \exp(-(\mathcal{L}_t - \bar{\mathcal{L}})^2 / 2\sigma_{\mathcal{L}}^2)$, where $\bar{\mathcal{L}}$ and $\sigma_{\mathcal{L}}$ are the mean and standard deviation of $\mathcal{L}_t$ along the trajectory).

B.2 Model details

We evaluate 18 reasoning-capable language models spanning six base families (Qwen3 (Yang et al., 2025a), Qwen2.5 (Yang et al., 2025b), Llama3 (Llama Team et al., 2024), Ministral (Liu et al., 2026), OLMo3 (Olmo et al., 2026), SmolLM3 (Bakouch et al., 2025)) and six post-training recipes (Qwen3, R1-distillation (DeepSeek-AI et al., 2025), Nemotron (Wang et al., 2026), Ministral-Reasoning, OLMo3-Think, and SmolLM), with sizes varying from 0.6B to 70B parameters. Table 4 lists these 18 models. We treat Qwen3-A3B as a part of the Qwen3 family. We additionally experimented with Ministral-3-14B-Reasoning, but could not obtain consistent performance evaluations on benchmarks (for instance, on some benchmarks it would perform even worse than the 3B model). We consider this a failure of our implementation or lack of reproducibility and hence do not include those results in our analysis.

We limit generations for all models to a maximum of 28000 tokens. This is to ensure fair comparison across models since some support shorter context lengths than others. In general, we generate with temperature $T = 0.6$, top-$p = 0.95$, and top-k disabled. Where needed, we followed model-specific overrides based on developer recommendations, for e.g., Ministral $\leq$ 8B uses $T = 0.7$, along with a special system prompt.[2] All benchmark performance evaluations are averaged over 5 random seeds.

B.3 Details of Benchmarks

In our experiments in Section 4 and Section 6.1, we use six challenging reasoning benchmarks: **AIME** 2025 (Zhang and Math-AI, 2025) consisting of 30 problems, **HMMT** Feb-25, Nov-25, and

[2] https://huggingface.co/mistralai/Ministral-3-8B-Reasoning-2512

Table 4: Models evaluated in the cross-family rank forecasting experiment.

Base Family	Model	Post-training	Short Name
Qwen2.5	R1-Distill-Qwen-7B	R1	R1-Q-7B
Qwen2.5	R1-Distill-Qwen-14B	R1	R1-Q-14B
Qwen2.5	R1-Distill-Qwen-32B	R1	R1-Q-32B
Llama3	R1-Distill-Llama-8B	R1	R1-L-8B
Llama3	R1-Distill-Llama-70B	R1	R1-L-70B
Ministral	Ministral-3-3B-Reasoning	Ministral	Ministral-3B
Ministral	Ministral-3-8B-Reasoning	Ministral	Ministral-8B
OLMo3	OLMo-3-32B-Think	OLMo3	OLMo3-32B
SmolLM3	SmolLM3-3B	SmolLM	SmolLM3-3B
Qwen3	Qwen3-0.6B	Qwen3	Qwen3-0.6B
Qwen3	Qwen3-1.7B	Qwen3	Qwen3-1.7B
Qwen3	Qwen3-4B	Qwen3	Qwen3-4B
Qwen3	Qwen3-8B	Qwen3	Qwen3-8B
Qwen3	Qwen3-14B	Qwen3	Qwen3-14B
Qwen3	Qwen3-32B	Qwen3	Qwen3-32B
Qwen3-A3B	Qwen3-30B-A3B-Thinking	Qwen3-A3B	Qwen3-A3B-30B
Qwen3	Nemotron-Cascade-8B	Nemotron	Nemo-8B
Qwen3	Nemotron-Cascade-14B	Nemotron	Nemo-14B

Feb-26 (Balunovic et al., 2025) consisting a total of 93 problems, **GPQA** main set (Rein et al., 2024) consisting a total of 448 problems, **USACO** competitive programming (Shi et al., 2024) consisting of 307 problems, the computer-science and engineering subsets of **MMLU-Pro** (Wang et al., 2024b) consisting a total of 1379 problems, and the hard non-science/engineering subset of **SuperGPQA** (Team et al., 2025) consisting of 382 problems. Together these cover competition math, graduate-level science, broad professional knowledge, and competitive code.

In our experiments in Section 5 and Section 6.2, we use OLMES which consists of MMLU (Hendrycks et al., 2021), HellaSwag (Zellers et al., 2019), ARC Challenge (Clark et al., 2018), ARC Easy (Clark et al., 2018), PIQA (Bisk et al., 2020), CommonsenseQA (Talmor et al., 2019), SocialIQA (Sap et al., 2019), OpenBookQA (Mihaylov et al., 2018), BoolQ (Clark et al., 2019), and WinoGrande (Sakaguchi et al., 2020).

B.4 Other Experimental Details

The experiments in Section 5 use DataDecide evaluation results with checkpoints of the default seed. The rBridge (Koh et al., 2026) results are based on our reproduction of their approach as described in their paper. While Koh et al. (2026) experimented with GPT-4o (OpenAI et al., 2024) as the teacher, we instead work with an open-weights expert (Llama-3.3-70B) for more reliable access to logprobs. The results for the downstream task performance baseline are taken directly from the DataDecide huggingface repository.[3]

The experiments in Section 6.1 use MiniMax-M2.5 and Qwen3Next-80B as the experts. For pre-training extrapolation, we set the inner-split-fraction as 0.5, fit on checkpoints up to step 350000 and test the extrapolation at step 1300000. To calculate the proxy metrics, since these are pretraining checkpoints, we use a standard chain-of-thought prompt which appends "Let's think how to answer this question step by step." to the problem before the expert trajectory. For post-training extrapolation, we fit on checkpoints up to step 22000 and test the extrapolation at step 40000.

B.5 Software and Compute Requirements

Our code is implemented in PyTorch (Paszke et al., 2019) and makes use of the HuggingFace Transformers library (Wolf et al., 2020) and the vLLM library (Kwon et al., 2023) for running efficient inference locally on LLMs. All benchmark evaluation and proxy metric calculation experiments were done on our cluster with 4 NVIDIA H100 GPUs with 80 GB memory. Experiments with Kimi-K2.5

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

Table 5: **Oracle proxy selection (upper bound).** Spearman ρ when the proxy is selected using downstream scores from all six benchmarks and the full model population. Because the proxy is fit and evaluated on the same tasks, these numbers upper-bound the cross-validated results in Table 2.

Proxy	**Mean ρ**	AIME	GPQA	HMMT	MMLU-Pro	SuperGPQA	USACO
Univariate proxy	0.62	0.51	0.64	0.43	0.73	0.81	0.62
3-sparse proxy	0.88	0.89	0.97	0.89	0.88	0.89	0.78

Table 6: **Proxy metrics selected under oracle access.** The specific proxy metrics and coefficients chosen when fitting on all six benchmarks and the full model population simultaneously. For the univariate proxy, the selected metric is inverse-frequency-weighted top-1 accuracy, the same metric that dominates the cross-validated selection (Figure 5). For the 3-sparse proxy, the negative coefficient on frequency-weighted top-5 accuracy penalizes models whose accuracy concentrates on frequent (and therefore easy) tokens, complementing the two positive components that reward uncertainty-aware agreement with the expert.

Model	**Selected proxy metric**	**Coefficient**
Univariate	`inverse_frequency` / `top_1_accuracy`	$+1$
3-sparse	`expert_disagreement` / `entropy`	$+1$
	`inverse_entropy` / `margin`	$+1$
	`frequency` / `top_5_accuracy`	-1

were carried out using Together API.[4] Obtaining model generations for benchmark evaluations takes around 2 hours on average for a single seed on a benchmark of 100 problems. Computing proxy metrics is extremely fast and takes less than 5 minutes per evaluation.

C Additional Results and Discussion

C.1 Cross-Family Model Selection

This section provides additional analysis of the cross-family model selection experiment (§4). We examine which proxy metrics are selected under cross-validation and oracle access, how the ranking signal compares to loss-based baselines, and how robust the results are to the number of held-out tasks and the fraction of models available for selection.

Which proxy metrics carry the ranking signal? Figure 5 visualizes the frequency with which each of the 80 proxy metrics is selected across all leave-2-tasks-out folds and model-subsampling seeds. In both the univariate and 3-sparse settings, the selection mass concentrates on a small number of cells, with the majority of the 80 proxy metrics never selected.

For the univariate proxy (Figure 5, left), inverse-frequency-weighted top-1 accuracy accounts for 32% of all selections, more than twice the frequency of any other cell. This metric measures whether the candidate model's top prediction matches the expert token, but only at positions where the expert token is rare. Frequent tokens, such as punctuation, articles, and common function words, are predicted well by nearly all models and therefore carry little discriminative signal. Upweighting rare tokens isolates the positions where models are most likely to differ: variable names in code, technical terms in science, or key numerical quantities in mathematics. The remaining selection mass spreads across other top-k accuracy variants (uniform top-1 at 0.16, uniform top-5 at 0.07) and expert-disagreement-weighted rank (0.12), all of which measure the same broad phenomenon, whether the model agrees with the expert, through slightly different lenses.

For the 3-sparse proxy (Figure 5, right), the pattern shifts. The two most frequently selected cells are entropy-weighted entropy (0.16) and frequency-weighted top-5 accuracy (0.14). Entropy-

[4] https://api.together.xyz/

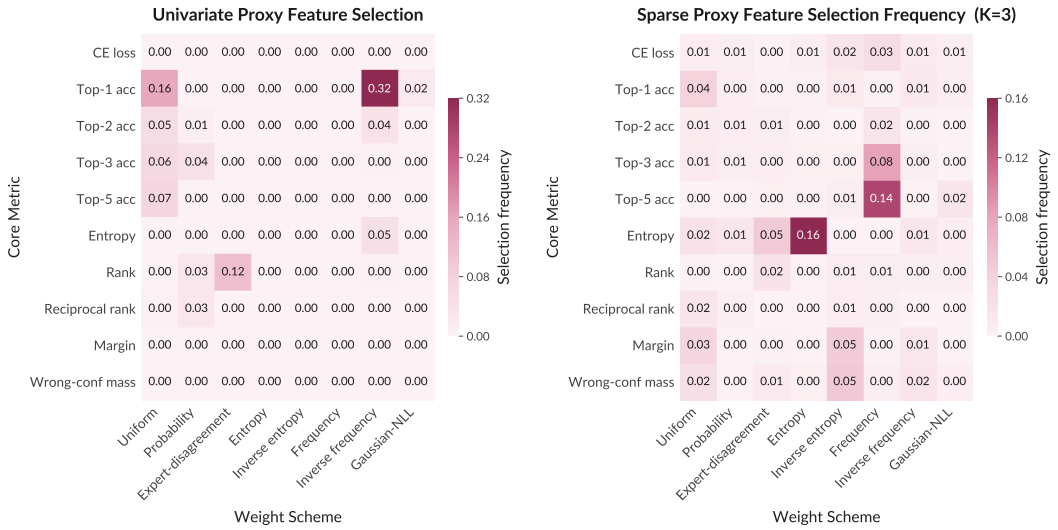


Figure 5: Proxy metric selection frequency (normalized) for univariate (*left*) and 3-sparse proxy (*right*). Darker cells are chosen more often. In both cases, a small cluster of proxy metrics accounts for almost all selections. The bulk of the 80 proxy metrics are unused.

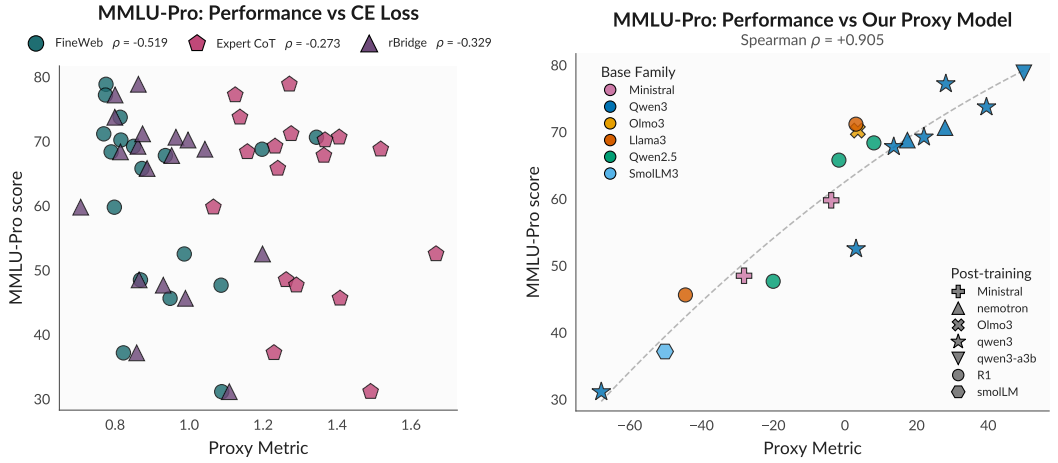


Figure 6: **Likelihood-based baselines against a learned proxy.** *Left:* the three loss-based baselines (FineWeb cross-entropy loss, uniform expert-trajectory cross-entropy loss, and rBridge) plotted against MMLU-Pro accuracy across 18 language models. Low loss is a weak and non-monotonic indicator of downstream ranking across model families and post-training recipes. *Right:* the learned RankSVM (linear) proxy, evaluated on MMLU-Pro in a fold where MMLU-Pro was held out, produces a nearly monotonic relationship with downstream accuracy. The same proxy was not exposed to MMLU-Pro during fitting.

weighted entropy upweights positions where the candidate model is uncertain and asks how diffuse its distribution is at those positions. A model that is uncertain at the right places, the critical reasoning steps, and concentrated elsewhere, is one that has learned the structure of the task without necessarily being able to solve it. Frequency-weighted top-5 accuracy, by contrast, measures agreement with the expert on common tokens. These two signals are complementary: the first captures behavior at hard, high-entropy positions, while the second captures baseline competence on the easy positions. The oracle 3-sparse proxy (Table 6) confirms this complementarity by including both an uncertainty-aware component (expert-disagreement-weighted entropy, coefficient $+1$) and a negative coefficient on frequency-weighted top-5 accuracy (-1), which penalizes models whose agreement with the expert concentrates on frequent tokens where all models perform well.

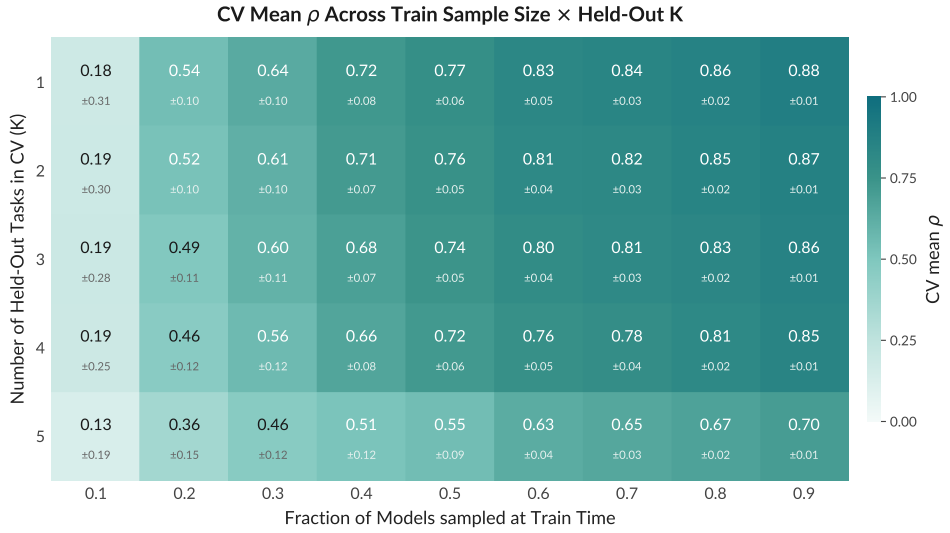


Figure 7: Performance of the linear RankSVM proxy as we vary the number of held-out tasks and the fraction of models used for selection. Each cell reports the mean Spearman ρ on held-out tasks.

Oracle upper bounds. Table 5 reports the Spearman ρ when the proxy is selected using downstream scores from all six benchmarks and the full model population. Because the proxy is fit and evaluated on the same data, these numbers upper-bound the cross-validated results in Table 2. The univariate oracle reaches $\rho = 0.62$, with substantial variation across tasks (HMMT at 0.43, SuperGPQA at 0.81). The 3-sparse oracle reaches $\rho = 0.88$ with per-task correlations between 0.78 and 0.97, indicating that three complementary proxy metrics contain nearly all the ranking information present in the 80-dimensional library. The gap between the oracle (0.88) and the cross-validated 3-sparse result (0.78) reflects the difficulty of selecting the right triplet without access to the target task, and suggests that better proxy selection procedures could close part of this gap.

Table 6 lists the specific proxy metrics selected under oracle access. The univariate oracle selects inverse-frequency-weighted top-1 accuracy, the same metric that dominates the cross-validated heatmap, providing evidence that the cross-validated selection procedure converges to a genuinely informative signal rather than overfitting to the held-in tasks. The 3-sparse oracle selects three metrics from three distinct regimes of model–expert alignment: expert-disagreement-weighted entropy (an uncertainty signal at positions where the model disagrees with the expert), inverse-entropy-weighted margin (a confidence signal at positions where the model is peaked), and frequency-weighted top-5 accuracy with a negative sign (a correction that discounts agreement on easy tokens). This combination is consistent with the view that ranking heterogeneous models requires measuring not just how often a model agrees with the expert, but how its uncertainty and confidence are distributed across token positions of varying difficulty.

Loss-based baselines fail across model families. Figure 6 contrasts the loss-based baselines with the learned proxy on MMLU-Pro. The left panel plots three loss-based signals, FineWeb cross-entropy ($\rho = -0.52$), uniform expert-trajectory cross-entropy ($\rho = -0.27$), and rBridge ($\rho = -0.33$), against MMLU-Pro accuracy. None of these produces a coherent ranking. Models with very different downstream scores overlap at similar loss values, and the overall relationship is weak and non-monotonic. The right panel shows the linear RankSVM proxy evaluated on a fold where MMLU-Pro was held out. Despite never seeing MMLU-Pro scores during fitting, the proxy produces a nearly monotonic relationship with downstream accuracy (Spearman $\rho = 0.91$), and this relationship holds across all six base families and seven post-training recipes. The contrast illustrates the central claim of §4: a scalar summary of fit to generic text is a poor predictor of relative downstream performance across a heterogeneous model population, but a small number of task-conditioned token-level statistics can recover the ranking reliably.

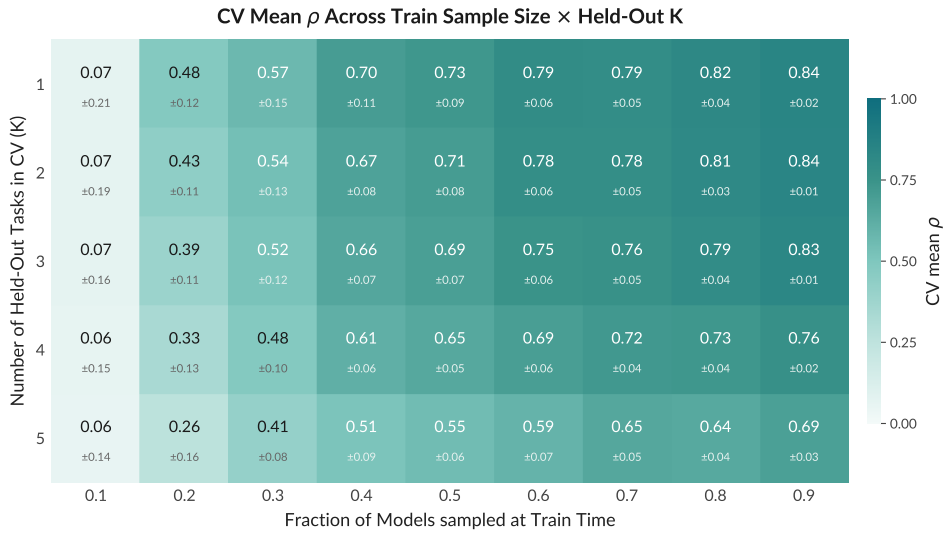


Figure 8: Performance of the 3-sparse proxy as we vary the number of held-out tasks and the fraction of models used for selection. Each cell reports the mean Spearman ρ on held-out tasks.

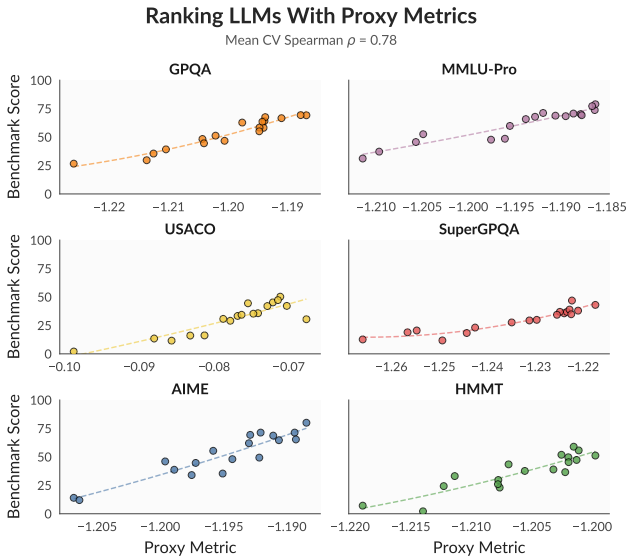


Figure 9: **Ranking LLMs with the 3-sparse proxy.** Downstream accuracy vs. proxy score for each of the six benchmarks on a randomly sampled held-out fold. Same format as Figure 1 (left) but using the 3-sparse proxy instead of the linear RankSVM.

Robustness to held-out tasks and model fraction. Figures 7 and 8 show how the linear RankSVM and 3-sparse proxies degrade as we increase the number of held-out tasks (rows) and decrease the fraction of models available for selection (columns). Performance degrades gracefully in both dimensions. For the linear RankSVM, holding out $K = 2$ tasks with 60% of models yields $\rho = 0.81$, and even holding out $K = 3$ tasks with 50% of models still achieves $\rho = 0.74$. Variance decreases steadily with more models, from ± 0.10 at 20% to ± 0.01 at 90%. The 3-sparse proxy shows a similar pattern. The one regime where performance collapses is $K = 5$ held-out tasks (i.e., selecting from a single held-in task), which drops to $\rho \approx 0.36$–0.70 depending on model fraction. This is expected: with only one task for selection, the procedure cannot distinguish proxy metrics that generalize from those that happen to work on that particular task.

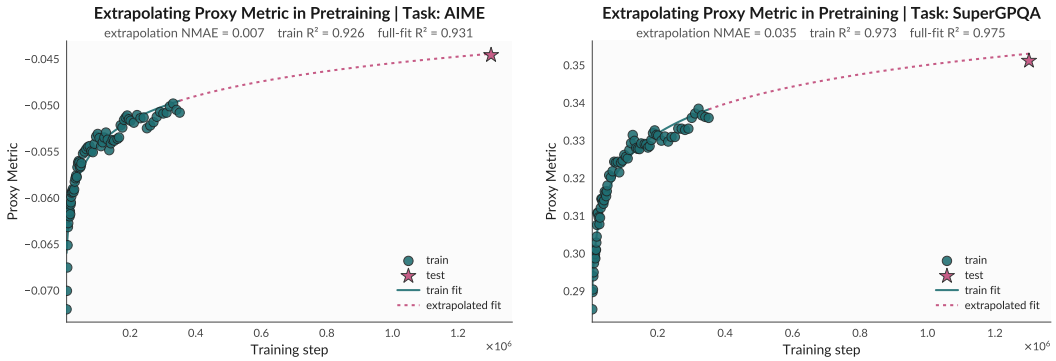


Figure 10: **Pretraining extrapolation on AIME and SuperGPQA.** Same protocol as Figure 3 (left) for the two benchmarks not shown in the main figure. The power-law fits from the training window (filled markers) extrapolate to the held-out checkpoint (star) on both tasks.

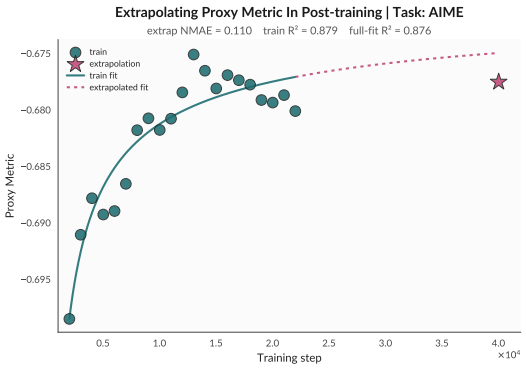


Figure 11: **Post-training extrapolation on AIME.** Same protocol as Figure 3 (right). The proxy metric evolves monotonically with post-training step and the power-law fit extrapolates to the held-out checkpoint.

Visual comparison of proxy models. Figure 9 plots downstream accuracy against the 3-sparse proxy score across all six benchmarks in a randomly sampled held-out fold, providing a counterpart to Figure 1 (left). The 3-sparse proxy produces monotonic trends on all six tasks, with somewhat tighter scatter than the linear RankSVM on GPQA and MMLU-Pro but slightly more variance on HMMT. This is consistent with the quantitative comparison in Table 2, where the two models achieve similar mean ρ (0.78 vs. 0.81) but differ on individual tasks.

C.2 Learned Rankers and Capacity-Matched Baselines

We test whether the gains in §4 come from merely fitting a higher-capacity ranker compared to scalar loss. Both experiments below use the evaluation protocol of §4 (leave-2-tasks-out cross-validation with 20 model-subsampling seeds), and the results are summarized in Table 8.

End-to-end learned ranker. A function of the raw (sequence length $\times$ vocabulary) next-token distributions is not well defined across a heterogeneous model population, since they might use different tokenizers and vocabularies. We therefore build the closest well-defined alternative, which removes all hand-designed features. At each position of the expert trajectory, the network receives 32 vocabulary-invariant inputs: the candidate's 31 largest next-token log-probabilities in sorted order, and its log-probability of the expert token. Each position is encoded by an MLP of hidden size 64. The positions are aggregated with a learned token weighting implemented as 4-head attention pooling. An MLP head outputs the candidate's score. The network has 21K parameters and is trained with a pairwise ranking loss on the same preference pairs used by the RankSVMs. It fits the held-in tasks almost perfectly (mean train correlation 0.96) but reaches only $\rho = 0.51 \pm 0.19$ on held-out tasks, compared with 0.81 ± 0.04 for the linear RankSVM over our library, with more than four times

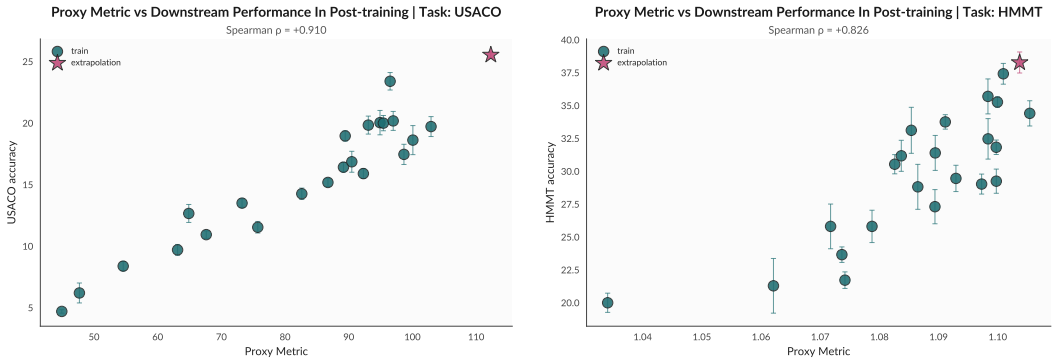


Figure 12: **Proxy metric vs. downstream accuracy at post-training checkpoints.** The best selected univariate proxy is plotted against downstream accuracy on USACO (left) and HMMT (right) across post-training checkpoints of OLMo-3-7B-Think. The strong monotonic relationship confirms that the extrapolated proxy tracks the ranking of interest.

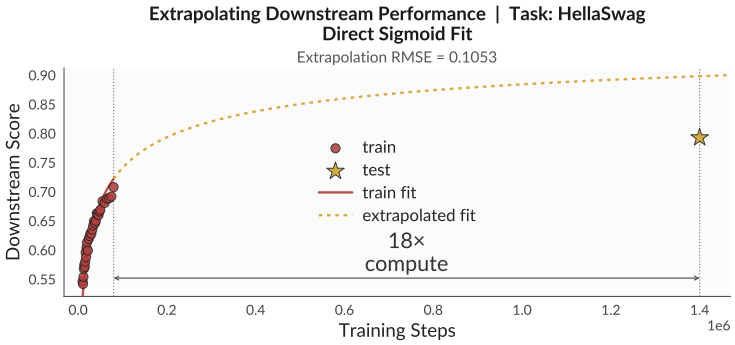


Figure 13: **Direct sigmoid extrapolation of HellaSwag accuracy.** Accuracy is fit as a sigmoid of $\log_{10}(\text{steps})$ following Owen (2024). Circles are the training window (up to 80K steps), the star is the held-out checkpoint at 1.4M steps. The fit overshoots the held-out accuracy (RMSE $= 0.11$).

the variance. Each fold provides about 11 models on 4 tasks, i.e., roughly 44 labeled outcomes and a few hundred preference pairs. Since the network fits these almost perfectly, its failure is one of generalization to unseen tasks and model families, and closing it would require more labeled (model, task) pairs, which are scarce by construction in this setting.

Capacity-matched loss baselines. Conversely, we give loss-based signals the same supervision and rankers as our proxies, changing only the input features. With three loss-based signals (FineWeb CE loss, uniform CE loss over expert trajectories, and rBridge), the best ranker (linear RankSVM) reaches $\rho = 0.43$. Adding the seven remaining weighted-CE variants from our library, so that the input contains every loss-based signal we compute (10 features), raises this to 0.58. In contrast, a 10-feature library of non-loss proxy metrics reaches $\rho = 0.85$ with the same linear RankSVM. We form this library from the metrics selected most frequently in Figure 5, excluding cross-entropy. Loss-based features thus plateau well below our proxies even when supervised and given matched capacity.

C.3 Robustness to Expert Quality

Our proxies are computed from expert trajectories, so their reliability may depend on the quality of the expert. We quantify this dependence below. Unless noted otherwise, results use the cross-validated ranking protocol of §4 with the linear RankSVM.

Expert errors. We never filter trajectories by whether they reach a correct final answer, so all results in the main paper already include substantial expert noise: between 7–13% (AIME) and

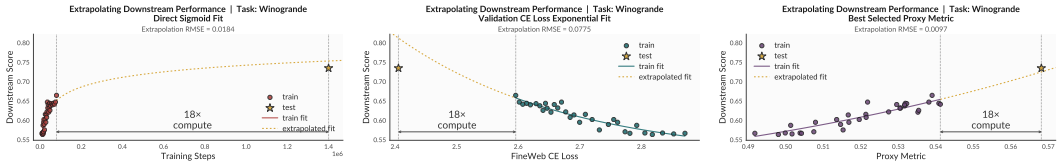


Figure 14: **Extrapolating Winogrande accuracy along the pretraining trajectory of OLMo-3-7B.** Circles are the training window (up to 80K steps), the star is the held-out checkpoint at 1.4M steps ($\sim$18$\times$ the training compute). *Left:* accuracy vs. $\log_{10}$(steps), fit with a sigmoid (RMSE $= 0.02$). *Centre:* accuracy vs. CE loss on FineWeb, fit with an exponential (RMSE $= 0.08$). *Right:* accuracy vs. the best univariate proxy, fit with a power law (RMSE $= 0.01$).

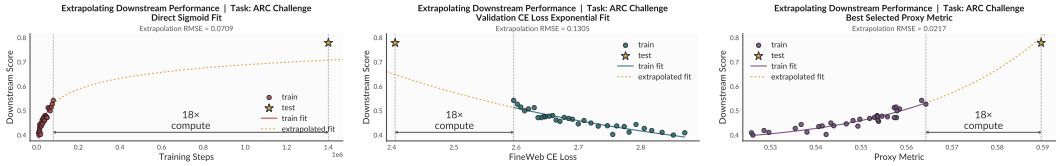


Figure 15: **Extrapolating ARC Challenge accuracy along the pretraining trajectory of OLMo-3-7B.** Circles are the training window (up to 80K steps), the star is the held-out checkpoint at 1.4M steps ($\sim$18$\times$ the training compute). *Left:* accuracy vs. $\log_{10}$(steps), fit with a sigmoid (RMSE $= 0.07$). *Centre:* accuracy vs. CE loss on FineWeb, fit with an exponential (RMSE $= 0.13$). *Right:* accuracy vs. the best univariate proxy, fit with a power law (RMSE $= 0.02$).

45–48% (SuperGPQA) of trajectories end in a wrong answer (Table 9). To test whether this noise hurts ranking, we recompute all proxy metrics using only trajectories whose final answer is correct, in an instance-matched comparison. Ranking quality is essentially unchanged (0.80 ± 0.05 versus 0.81 ± 0.04). We attribute this robustness to what the proxies measure: token-level agreement between the candidate's predictive distribution and the expert's text at every position. A trajectory that ends in a wrong answer still consists mostly of meaningful reasoning steps, and agreement with those steps separates strong candidates from weak ones just as well. In practice, our method therefore requires no answer verification or trajectory filtering before use.

Weaker experts. The data selection experiment in §5 already uses a much weaker expert than §4: chains of thought from Llama-3.3-70B, a non-reasoning instruction-tuned model. Replacing it with OLMo-2-32B-Instruct (Walsh et al., 2025) preserves the trend in Figure 1 (right): our proxy metric still outperforms rBridge and reaches a decision accuracy of about 0.83 at orders of magnitude less compute than the downstream-evaluation baseline.

C.4 Truncation Length of Expert Trajectories

All main results compute proxy metrics on the last 1,000 tokens of each expert trajectory. Table 10 varies which tokens are scored, re-running the full cross-validated ranking experiment of §4 for each choice. Every last-K variant outperforms every first-K variant as well as the full trajectory, so the informative tokens sit late in the trajectory.

C.5 Training-time Forecasting

This section provides some additional results for the training-time forecasting experiments of §6.

Pretraining extrapolation for AIME and SuperGPQA. Figure 10 extends the pretraining extrapolation of Figure 3 (left) to the two benchmarks not shown in the main figure: AIME and SuperGPQA. On AIME the selected proxy follows a clean power law with NMAE $= 0.007$ and train-window $R^2 = 0.93$, and the extrapolated value lands almost exactly on the held-out checkpoint. SuperGPQA is similarly well-behaved (NMAE $= 0.035$, $R^2 = 0.97$), with a slightly larger extrapolation gap. Together with the four benchmarks in the main figure (mean NMAE $= 0.034$), these results confirm that the power-law regularity of the proxy is not confined to a subset of tasks: across all six reasoning benchmarks the mean pretraining NMAE is 0.030.

Table 7: Extrapolation RMSE for predicting downstream accuracy at 1.4M steps from checkpoints up to 80K steps ($\sim 18\times$ compute extrapolation). Three predictors are compared: the best univariate proxy (selected by inner split), an exponential fit of accuracy against FineWeb validation CE loss (Gadre et al., 2025), and a direct sigmoid fit of accuracy against $\log_{10}(\text{steps})$ (Owen, 2024). **Bold** marks the lowest RMSE per task.

Task	Best Proxy	Proxy	CE Loss	Compute
HellaSwag	`uniform / top_5_acc`	**0.003**	0.094	0.105
BoolQ	`entropy / top_5_acc`	0.037	0.111	**0.034**
SocialIQA	`inverse_freq / top_3_acc`	0.016	**0.011**	0.025
Winogrande	`inverse_freq / top_2_acc`	**0.010**	0.078	0.018
ARC Challenge	`expert_disagreement / top_3_acc`	**0.022**	0.131	0.071
ARC Easy	`probability / top_3_acc`	0.049	**0.001**	0.036
CommonsenseQA	`inverse_freq / top_1_acc`	**0.004**	0.023	0.129
MMLU	`freq / top_2_acc`	**0.001**	0.036	0.058
OpenBookQA	`entropy / top_1_acc`	0.069	0.086	**0.000**
PIQA	`probability / reciprocal_rank`	0.029	**0.025**	0.068
Mean	—	**0.024**	0.059	0.055

Table 8: **Learned rankers and capacity-matched baselines.** Mean leave-2-tasks-out Spearman ρ of the best of our four rankers (§4) for each input feature set, and of an end-to-end neural ranker over raw log-probabilities.

Input features	**# Features**	**Mean ρ**
Loss-based (FineWeb CE, uniform expert CE, rBridge)	3	0.43
All loss-based signals	10	0.58
Non-loss proxy metrics	10	0.85
Full library (Table 2)	80	0.81 ± 0.04
Raw log-probabilities (end-to-end network)	–	0.51 ± 0.19

Post-training extrapolation for AIME. Figure 11 shows the post-training extrapolation on AIME, complementing the set of results in Figure 3 (right). This is the noisiest setting we observe: the training-window points exhibit visible scatter around the fitted power law ($R^2 = 0.88$), and the NMAE is 0.110, roughly three times larger than the post-training average on the other four benchmarks. We attribute this to the difficulty and noisiness of the task itself. AIME consists of only 30 problems, all of which are competition-math level difficulty.

Proxy-accuracy correspondence at post-training checkpoints. The extrapolation experiments predict the *proxy metric* at a future checkpoint. For this prediction to be useful, the proxy must actually track the downstream quantity of interest. Figure 12 provides this sanity check by plotting the selected univariate proxy against downstream accuracy across post-training checkpoints of OLMo-3-7B-Think on USACO and HMMT. On USACO the relationship is strongly monotonic (Spearman $\rho = 0.91$). On HMMT the trend is noisier ($\rho = 0.83$) with wider confidence intervals in general, reflecting the smaller size of the test set and the high difficulty of competition mathematics. Both correlations are well above what would be needed for the extrapolated proxy to serve as a reliable early indicator of downstream ranking, and the mean across all five post-training benchmarks ($\rho = 0.84$, as reported in §6.1) confirms that this correspondence holds across the other benchmarks.

Proxy metrics selected for per-task downstream accuracy extrapolation. Table 7 expands Table 3 by reporting the identity of the best proxy selected by the inner-split procedure for each of the ten OLMES benchmarks. Two patterns stand out. First, the selected proxies are overwhelmingly top-k accuracy variants: nine of the ten tasks select a top-k accuracy metric under some weighting scheme,

Table 9: **Expert error rates.** Percentage of expert trajectories whose final answer is incorrect, reported as the range across the three experts (Kimi-K2.5, MiniMax-M2.5, and Qwen3-Next-80B). All results in the main paper use these trajectories without filtering.

AIME	GPQA	HMMT	MMLU-Pro	SuperGPQA	USACO
7–13%	23–28%	18–25%	20–24%	45–48%	23–37%

Table 10: **Truncation length ablation.** Mean leave-2-tasks-out Spearman ρ of the linear RankSVM when proxy metrics are computed on different segments of each expert trajectory.

Tokens scored	**Mean ρ**
First 500	0.72 ± 0.07
First 1,000	0.72 ± 0.05
First 3,000	0.71 ± 0.05
Full trajectory	0.74 ± 0.03
Last 500	0.79 ± 0.06
Last 1,000 (main setting)	0.81 ± 0.04
Last 3,000	0.76 ± 0.04

with k ranging from 1 to 5. The one exception is PIQA, which selects probability-weighted reciprocal rank. This convergence on top-k agreement is consistent with the cross-family heatmap in Figure 5, where top-k accuracy under various weightings dominates the selected features, and suggests that how often a model places the expert's token in its top predictions is a broadly informative signal across diverse evaluation settings. Second, the weighting schemes are heterogeneous: inverse-frequency weighting appears on three tasks (Winogrande, CommonsenseQA, SocialIQA), frequency weighting on one (MMLU), entropy weighting on one (BoolQ), and other schemes on the remainder. No single weighting dominates, which is why the full library of eighty proxy metrics is necessary for broad coverage.

Baseline comparisons for per-task downstream accuracy extrapolation. Figure 13 shows the direct sigmoid baseline of Owen (2024) on HellaSwag. Within the training window (up to 80K steps) the sigmoid fits the data well, but at the held-out checkpoint (1.4M steps) it overshoots the true accuracy substantially (RMSE $= 0.11$). Figures 14 and 15 provide side-by-side comparisons of all three predictors on Winogrande and ARC Challenge. On Winogrande the sigmoid fit undershoots slightly (RMSE $= 0.02$), the CE loss exponential overshoots markedly (RMSE $= 0.08$), and the proxy power law tracks the target closely (RMSE $= 0.01$). ARC Challenge exhibits the same pattern more dramatically: the CE loss exponential diverges (RMSE $= 0.13$), the sigmoid is closer but still off (RMSE $= 0.07$), while the proxy fit remains tight (RMSE $= 0.02$).

D Extended Related Work

In this section, we mention other related works.

Compute-based scaling laws for pretraining loss. Classical scaling laws connect training compute, parameters, and tokens to a single pretraining cross-entropy loss under controlled architectures (Hoffmann et al., 2022; Kaplan et al., 2020). These laws are extraordinarily powerful as planning tools when the quantity of interest is loss itself, and the functional form, a power law with additive terms in N and D, is the starting point for virtually every downstream forecasting pipeline that follows, including ours when we extrapolate the proxy metric along a training trajectory in Section 6. They are silent, however, on the quantity practitioners actually care about, which is the score of the trained model on a specific downstream task. Follow-up work extends these laws to over-trained regimes (Gadre et al., 2025), data-constrained regimes (Muennighoff et al., 2023), inference-aware compute budgets (Sardana et al., 2024), and has sharpened their fitting protocols (Besiroglu et al., 2024; Choshen et al., 2025; Porian et al., 2024). The proxy metrics we study complement this line

by producing a dense, task-specific, and well-behaved scalar that is defined at every training step and obeys its own smooth law, even in regimes where pretraining loss has long since saturated its informativeness about task capability.

Forecasting downstream task performance. A growing line of work converts the pretraining loss into a forecast of downstream accuracy. Gadre et al. (2025) propose the exponential map $y = \epsilon - k \cdot \exp(-\gamma x)$ from validation perplexity to average downstream top-1 error across many tasks. Bhagia et al. (2025) decompose the problem into two stages, $(N, D) \to$ task-specific loss on the correct answer $\to$ ranked-classification accuracy, using compute-efficient ladders at roughly 1% of target compute. Chen et al. (2025) extend the two-stage approach to domain-mixed pretraining with FLP and FLP-M. Ruan et al. (2024) take a different route with observational scaling laws, fitting a low-dimensional latent capability axis over roughly one hundred public models via PCA on benchmark scores, and showing that agentic and reasoning behaviors become predictable in this latent space. Owen (2024) fits sigmoidal forms from scaling-estimated loss to BIG-Bench and MMLU across eleven families, and Polo et al. (2026) show that benchmark performance is well explained by a handful of latent skills that transfer across families. Krajewski et al. (2026) push the single-recipe direct law further by showing that a two-parameter law on training compute can fit downstream accuracy directly when the task is above chance.

The unreliability of downstream scaling laws under realistic constraints. Several recent papers document concrete failure modes of the scaling-law-to-downstream pipeline. Lourie et al. (2025) find that only about 39% of the tasks studied by Gadre et al. (2025) exhibit predictably linear scaling of downstream accuracy with compute, and Schaeffer et al. (2025) argue that predicting downstream capabilities of frontier models from scale has remained elusive because argmax-based accuracy degrades the cross-entropy-to-score relationship. Accurate prediction requires modeling probability mass on both correct and incorrect completions. This observation is a direct motivation for the margin and wrong-confidence features in our library, which measure exactly the two sides of this error surface. Isik et al. (2025) provide a complementary warning. On machine translation, pretraining cross-entropy can continue to fall while BLEU or COMET degrade, so a scalar loss is not even a monotone predictor of the downstream metric. Liu et al. (2023) show that models with identical pretraining loss can differ substantially downstream, a direct counterexample to the simplest loss-as-rank hypothesis, and Tay et al. (2023) show that the upstream-downstream ordering can invert across architectures.

Emergence and the role of continuous metrics. Wei et al. (2022a) coined the term emergent abilities for capabilities that appear absent in smaller models and abruptly present at larger scales, a framing that was interpreted by many as evidence against forecastability. Schaeffer et al. (2023) push back by showing that apparent emergence largely evaporates under continuous and smooth metrics such as log-likelihood, Brier score, or edit distance, with abrupt jumps appearing only when the evaluation collapses the predictive distribution to a discrete score. Du et al. (2024) strengthen this picture from the opposite direction by arguing that downstream abilities align more cleanly with pretraining loss than with parameter count, and that emergence often reflects a loss threshold. Hu et al. (2024) develop PassUntil, a continuous evaluation with effectively infinite resolution that makes per-instance task-solve probability predictable from small-scale models, and Snell et al. (2024) fit emergence laws by finetuning smaller models to shift the emergence point.

Loss-to-loss and cross-distribution prediction. Brandfonbrener et al. (2025) introduce loss-to-loss prediction, fitting shifted power laws that map training loss on one dataset to training or test loss on another and extrapolating up to roughly twenty times the training compute. Mayilvahanan et al. (2025) extend this to thousands of checkpoints, showing that the loss-to-loss trend is robust across architectures and tokenizers. Huang et al. (2024) provide cross-family evidence in the same spirit, reporting a Pearson correlation of approximately -0.95 between bits-per-character and downstream scores across thirty-one LLMs and twelve benchmarks. Magnusson et al. (2024) propose Paloma as a tokenizer-invariant bits-per-byte benchmark over 546 domains, and Xia et al. (2023) show that across the OPT family perplexity predicts in-context learning on seventy-four BIG-Bench tasks better than model size or compute.

Token-level likelihood signals and fine-grained weighting. Several papers have recognized that not every token position carries the same signal. Gonen et al. (2023) show that prompt perplexity is a

label-free predictor of task performance, Ankner et al. (2025) use small-reference-model perplexity to prune pretraining data and improve downstream accuracy at 3B scale, and Thrush et al. (2025) select pretraining documents whose per-document perplexity correlates most strongly with downstream benchmarks. Closest in spirit to our weighting schemes, Fang et al. (2025) introduce LongPPL, which weights perplexity only on key tokens identified by a reference model and achieves a -0.96 correlation with long-context benchmarks where uniform perplexity shows almost none. Sorscher et al. (2022) use the teacher's margin (top minus second-top logit) as a per-example difficulty signal for data pruning, anticipating our use of margin as a token-level feature.

Small-scale proxies for pretraining decisions. Another line of works asks whether small models can be used to decide between candidate pretraining datasets or recipes before committing target-scale compute. Wortsman et al. (2024) formalize this as small-scale proxies for large-scale transformer behavior, and μP and μTransfer (Yang et al., 2021) provide the parameterization that makes such extrapolation work for hyperparameters. Xie et al. (2023) introduce DoReMi, which trains a 280M proxy with Group DRO on excess NLL and transfers domain weights to 8B. Liu et al. (2025) introduce RegMix, which fits a regression over 512 1M-parameter proxies to select optimal mixtures at 1B and 7B with roughly 2% of full compute. Magnusson et al. (2025) construct DataDecide, a large controlled testbed of twenty-five pretraining corpora and fourteen proxy scales, and show that continuous likelihood-style metrics make MMLU, ARC, HellaSwag, MBPP, and HumanEval over 80% predictable at 0.01% of target compute. Koh et al. (2026) propose rBridge, which aligns the small-proxy NLL with the target task by weighting by a frontier-model reasoning trace.

Reasoning traces as supervision and as signal. Reasoning traces have been used most extensively as a training signal. Chain-of-thought prompting (Wei et al., 2022b) and self-consistency (Wang et al., 2023) exploit reasoning trajectories at inference, and STaR (Zelikman et al., 2022) and Quiet-STaR (Zelikman et al., 2024) bootstrap models from generated rationales. Process reward models make the token-level view explicit. (Uesato et al., 2022) compare process and outcome supervision, Lightman et al. (2024) release PRM800K and show that process supervision dominates outcome supervision, and Wang et al. (2024a) and (Luo et al., 2024) develop automatic step-level process rewards from Monte Carlo rollouts. Two recent methods are particularly close in spirit to ours at the scoring step. Hao et al. (2023) use per-step LM token likelihood together with self-evaluation as an MCTS reward, and Cui et al. (2025) and Yuan et al. (2025) derive implicit token-level process rewards from log-likelihood ratios of outcome-supervised models. All of these works treat trajectories as signals for training or search. Our paper uses trajectories as an evaluation substrate instead.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the claims, and each is supported by a dedicated experimental section with quantitative results: cross-family ranking (Section 4, Table 2), pretraining data selection (Section 5, Figure 1), and training-time forecasting (Section 6, Table 3).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Appendix 8.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results; all contributions are empirical.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All information needed for reproduction are provided either in the main paper or in the Appendix B. Note that we only work with publicly available models and datasets.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide code in the supplementary materials.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: All experimental setting details are provided in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: For the cross-family model selection experiment, Table 2 reports mean $\pm$ std across 20 random model-subsampling seeds over 15 cross-validation folds. The pretraining data selection experiment is done over the default seed and training-time forecasting experiments use deterministic evaluation where variability does not arise from random splits or random seeds.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute requirements are specified in Appendix B.5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We comply with the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts of our work in Appendix A.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper proposes evaluation metrics for existing models and does not release new models, datasets, or artifacts that pose risks of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We respect the license and terms of use for all datasets and models used in this work, and properly credit them where required.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release any new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: All research work in this paper was carried out by humans.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.