WHY HAS PREDICTING DOWNSTREAM CAPABILITIES OF FRONTIER AI MODELS WITH SCALE REMAINED ELUSIVE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Predictable behavior from scaling advanced AI systems is an extremely desirable property for engineers, industry, economists and governments alike, and, while a well-established literature exists on how pretraining performance scales, predictable scaling behavior on downstream capabilities remains elusive. While many factors are certainly responsible, this paper identifies a significant factor that makes predicting scaling behavior on widely used multiple-choice question answering benchmarks challenging and illuminates a path towards making such downstream evaluations predictable with scale. Using five model families and twelve well-established multiple-choice benchmarks, we show that downstream performance is computed from negative log likelihoods via a sequence of transformations that progressively degrades the statistical relationship between performance and scale. We then pinpoint the mechanism causing this degradation: downstream metrics require comparing the correct choice against a small number of specific incorrect choices, meaning accurately predicting downstream capabilities requires predicting not just how probability mass concentrates on the correct choice with scale, but also how probability mass fluctuates on specific incorrect choices with scale. We empirically study how probability mass on the correct choice co-varies with probability mass on incorrect choices with increasing compute, suggesting that scaling laws for *incorrect* choices might be achievable. Our work also explains why pretraining scaling laws are commonly regarded as more predictable than downstream capabilities and contributes towards establishing scaling-predictable evaluations of frontier AI models.

1 THE IMPORTANCE OF PREDICTING CAPABILITIES WITH SCALE

Predictable scaling behavior of frontier AI systems such as GPT-4 (OpenAI, 2024; OpenAI et al., 2024), Claude (Anthropic, 2024) and Gemini (Team et al., 2023; Reid et al., 2024) is crucial for anticipating capabilities and informing key decisions regarding model development and deployment (Anthropic, 2023; OpenAI, 2023; Dragan et al., 2024). Predictable scaling behaviors enable engineers to make informed decisions about optimal model design choices and to de-risk investment in exceedingly expensive pretraining runs by determining the payoff from scaling up compute. For instance, OpenAI noted in the GPT-4 Technical Report (Achiam et al., 2023) that "A large focus of the GPT-4 project was building a deep learning stack that scales predictably" and that "[OpenAI] developed infrastructure and optimization methods that have very predictable behavior across multiple scales"; OpenAI noted that this ideally goes beyond predicting loss values, and that "Having a sense of the capabilities of a model before training can improve decisions around alignment, safety, and deployment". Meta's Llama Team similarly conducted experiments aimed at predicting the downstream performance of models, used to inform the design of their 405 billion parameter model (Dubey et al., 2024). Additionally, predicting capabilities is of interest beyond AI practitioners: economists and governments also have a significant interest in predicting the capabilities of current and future frontier AI systems for better decision-making (regulation, taxation, and safety) and forecasting of economic impacts (Council of Economic Advisers, 2024). Downstream capabilities are especially of interest because quantities like pretraining loss are difficult to translate into quantities more meaningful to society, such as the impact on economic labor or societal harms.

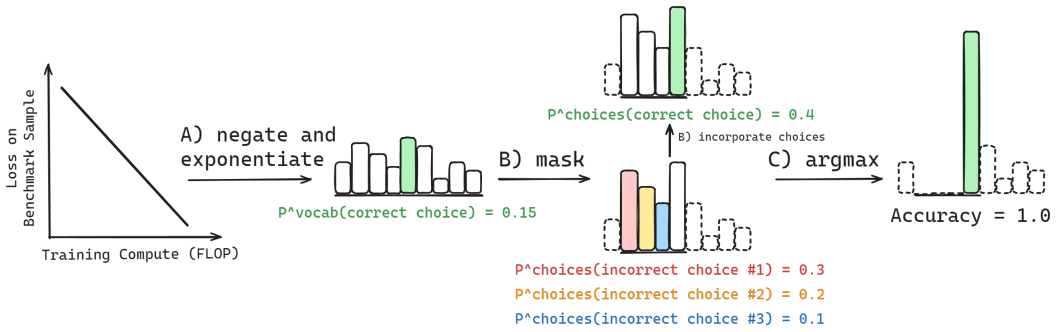

Figure 1: **Multiple-choice benchmark accuracy is computed from negative log likelihoods via a sequence of transformations that degrades predictability.** Computing `Accuracy` begins with computing the negative log likelihoods of each choice, then negating and exponentiating each to obtain the probability of each choice (**A**). Choices are then restricted to a set of available choices by *masking* invalid continuations, and renormalizing to obtain relative probability mass on each choice (**B**). Lastly, the model's choice is defined as $\arg\max_i \{p^{\text{Choices}}(\text{Available Choice}_i)\}$, and `Accuracy` is 1 if and only if the model's choice is the correct choice (**C**).

However, while scaling laws describing relationships amongst parameters, data, compute, and pretraining loss are well-established (Hestness et al., 2017; Rosenfeld et al., 2019; Henighan et al., 2020; Kaplan et al., 2020; Gordon et al., 2021; Hernandez et al., 2021; Jones, 2021; Zhai et al., 2022; Hoffmann et al., 2022; Clark et al., 2022; Neumann & Gros, 2022; Hernandez et al., 2022; Maloney et al., 2022; Sardana & Frankle, 2023; Muennighoff et al., 2024; Besiroglu et al., 2024), the literature is less conclusive regarding predicting specific downstream capabilities with scale. For instance, prior work has observed that performance on standard natural language processing (NLP) benchmarks can exhibit *emergent abilities* (Brown et al., 2020; Ganguli et al., 2022; Srivastava et al., 2022; Wei et al., 2022) where performance changes unpredictably with scale, but further work demonstrated that such unpredictable changes might at times be artifacts of researchers' analyses, i.e., choices of metrics and lack of sufficient resolution from too few samples (Srivastava et al., 2022; Schaeffer et al., 2023; Hu et al., 2024). More recently, Du et al. (2024) claim that downstream capabilities *can* be predicted, but *only* after the pretraining cross-entropy loss falls below a certain threshold, and Gadre et al. (2024) claim that while performance on individual tasks can be difficult to predict, aggregating results across dozens of diverse benchmarks yields clearer scaling trends. In this work, we ask: in contrast with strongly-predictable pretraining losses, *why has predicting specific downstream capabilities with scale remained elusive?*

2 Contribution: Explaining Why Predicting Downstream Capabilities With Scale Has Remained Elusive

Our goal is to understand what breaks down between the predictability of pretraining losses and the unpredictability of downstream evaluations. To do this, we investigated the relative predictability of different evaluation methodologies and setups, focusing on popular and comparatively simple (yet still highly difficult to predict) multiple-choice question answering benchmarks. We began with scaling-predictable pretraining log likelihoods and tracked how these log likelihoods are transformed in the process of calculating downstream evaluation metrics that are notoriously difficult to predict, such as `Accuracy` or `Brier Score` (See Fig. 1 and Sec. 4 for further detail):

$$\underbrace{\log p_\theta^{\text{Vocab}}(\text{Correct Choice})}_{\text{Scaling-Predictable}} \rightarrow p_\theta^{\text{Vocab}}(\text{Correct Choice}) \rightarrow p_\theta^{\text{Choices}}(\text{Correct Choice}) \rightarrow \underbrace{\texttt{Accuracy}}_{\text{Scaling-Unpredictable}}$$

This paper will demonstrate the following summary of our findings:

1. **Calculating downstream metrics requires a sequence of transformations applied to the original scaling-predictable quantities. These transformations progressively dete-**

riorate the statistical relationship between those metrics and the scaling parameters (parameters, data, and compute). This formalizes an intuition that "more complex" metrics might be less easily predictable.

2. **Accurately predicting downstream multiple-choice performance requires modeling not only the probability mass assigned to the correct choice with scale, but also the probability mass assigned to the incorrect alternatives.** This explains the cause of comparative unpredictability of multiple-choice benchmarks, and also suggests a potential path forward for successful predictive models of downstream performance in the area of multiple-choice question answering, and in general the need to model *external information* not related to scaling-predictable log likelihoods needed for downstream metric computation.
3. **Continuous metrics such as `Brier Score` are insufficient for recovering predictability.** We observe that, contrary to prior work showing that metrics such as `Brier Score` can hide *emergent behavior* at times, `Brier Score` is insufficient to improve the statistical relationship degraded by incorporating incorrect choices' probability mass.

Our findings explain that the apparent unpredictability of individual downstream evaluations is due to specific incorrect choices, which the strongly predictable pretraining losses do not depend upon. More broadly, we argue that a precise understanding of the factors affecting downstream performance is essential for designing evaluations to reliably track the progression of frontier AI models' capabilities.

3 Methodology: Data for Studying Scaling of Downstream Capabilities

To study how downstream capabilities on specific tasks change with scale for different model families, we generated per-sample scores from a large number of model families and multiple-choice NLP benchmarks. To ensure the computed scores were consistent with prior work, we used EleutherAI's Language Model Evaluation Harness (Gao et al., 2023).

Model Families Because our goal is to explore the scaling behavior of evaluations with increasing compute, we chose to evaluate model families with dense combinations of parameter counts and token counts. The following families were evaluated (additional details in App. D):

1. **Pythia** (Biderman et al., 2023b): The Pythia family contains 8 models from 70M to 12B parameters trained on the Pile (Gao et al., 2020) for 300B tokens. We used 8 checkpoints per size of the non-deduplicated variants.
2. **Cerebras-GPT** (Dey et al., 2023): The Cerebras-GPT family contains 7 models ranging from 111M to 13B parameters. The models were trained on the Pile (Gao et al., 2020) for different durations as part of a scaling study with a ratio of $\sim 20\times$ tokens to parameters in a "Chinchilla"-optimal manner (Hoffmann et al., 2022).
3. **OLMo** (Groeneveld et al., 2024): The OLMo family contains a 1B parameter model trained for 3T tokens and two 7B parameter models trained for 2T-2.5T tokens. We selected 7 checkpoints for 1B (spanning 84B[1] to 3T tokens) and 7 checkpoints for 7B (spanning 4B to 2.4T tokens).
4. **INCITE** (AI, 2023): The INCITE family contains 3B and 7B parameter models, trained on 0.8T and 1T tokens of RedPajama-v1(Computer, 2023). The 3B model has only a single checkpoint, so we excluded it. We found this family to be a slight outlier from other families, which we speculate is because its pretraining data were contaminated by benchmarks (Elazar et al., 2023).
5. **LLM360** (Liu et al., 2023): LLM360 includes two 7B parameter LLMs trained on 1.3T and 1.4T tokens. We selected 13 checkpoints of Amber spaced approximately logarithmically.

NLP Benchmarks We evaluated the above model families on widely-used multiple-choice benchmarks for assessing comprehension, reasoning, and world knowledge: AI2 Reasoning Challenge (ARC) Easy and Hard (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MathQA (Amini et al.,

[1]OLMo 1B checkpoints below 84B tokens were unfortunately accidentally lost by their creators.

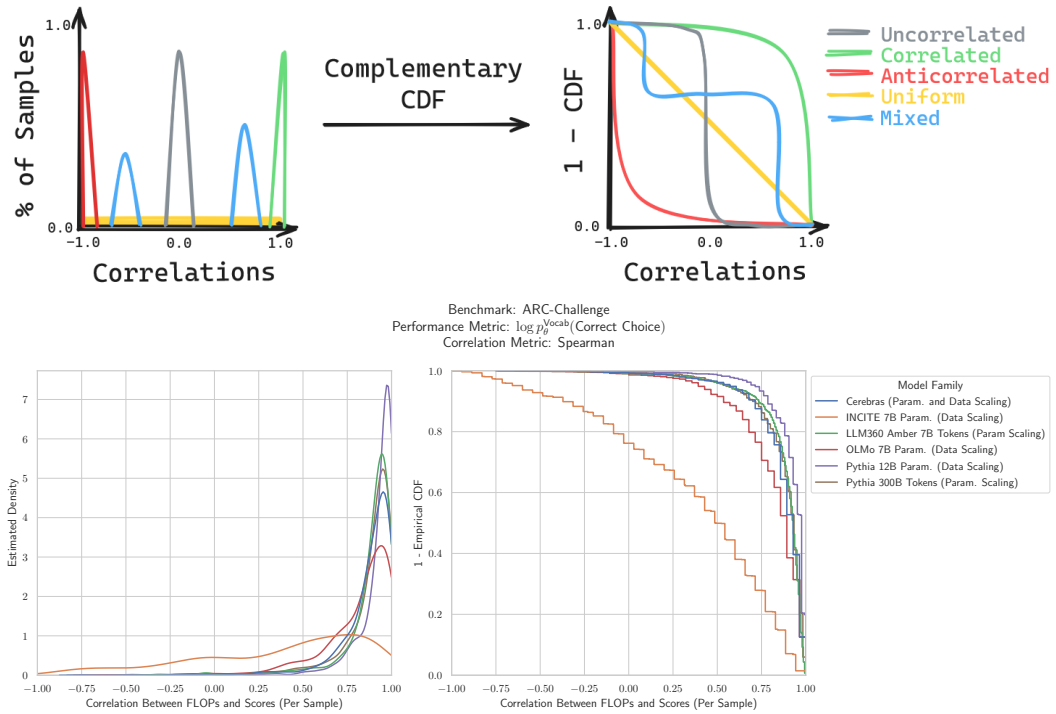

Figure 2: **Distributions of score-compute correlations and their corresponding complementary cumulative distribution functions. Left:** For each benchmark, model family, performance metric, and correlation metric, we computed how scores correlate with compute. This yields a distribution (over samples) of score-compute correlations. Note: the uniform distribution is small but non-zero everywhere. **Right:** To easily extract what fraction of samples in a benchmark has score-compute correlations above any given threshold, we converted the probability distributions to *complementary cumulative distribution functions*, i.e., 1 minus the empirical cumulative distribution function (CDF). **Top:** Idealized distributions. **Bottom:** Actual data on ARC Challenge.

2019), MCTACO (Zhou et al., 2019), MMLU (Hendrycks et al., 2020), OpenbookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), RACE (Lai et al., 2017), SciQ (Welbl et al., 2017), SIQA (Sap et al., 2019a), WinoGrande (Keisuke et al., 2019) and XWinoGrad En (Muennighoff et al., 2023). For MMLU, we analyzed each of the 57 subjects (e.g., Abstract Algebra) independently. For each benchmark, we used default evaluation settings from the LM Evaluation Harness (Gao et al., 2023).

Performance Metrics We used three common multiple-choice metrics (Srivastava et al., 2022; Schaeffer et al., 2023; Du et al., 2024): `Accuracy`, `Brier Score` (Brier, 1950), and probability mass on the correct choice relative to the available choices $p_\theta^{\text{Choices}}$(Correct Choice).

Compute Budget Calculations Following prior work (Kaplan et al., 2020), we approximated[2] the pretraining compute C (in terms of training FLOP) of a given model checkpoint as a function of the parameter count (excluding embeddings) N and the amount of training data seen in tokens D:

$$C = C(N, D) \approx 6\,N\,D$$

[2]This approximation neglects FLOP costs associated with attention calculations over sequence length; however, such operations are negligible so long as $d_{model} >> n_{ctx}/12$, and this approximation is therefore standard in most language model scaling law analyses.

4 WHAT MAKES PREDICTING DOWNSTREAM PERFORMANCE DIFFICULT?

Performance on multiple choice benchmarks is commonly published as `Accuracy`, `Brier Score`, or probability mass on the correct choice out of the available choices $p_\theta^{\text{Choices}}(\text{Correct Choice})$. These quantities are computed via a sequence of transformations that begins with the negative log likelihood of the correct choice on this particular benchmark sample as some function $f(\cdot,\cdot)$ of compute:

$$\mathcal{L}_\theta^{\text{Vocab}}(\text{Correct Choice}) = f(\text{Compute}, \text{Benchmark Datum}) \tag{1}$$

Two details are critical. Firstly, this negative log likelihood is not computed in expectation over a corpus; it is specific to this particular singular datum in the benchmark. *All the scores we discuss are per-datum.* Secondly, this negative log likelihood is computed over the vocabulary of the model. One can then compute the probability mass of the correct choice, again with respect to the vocabulary:

$$p_\theta^{\text{Vocab}}(\text{Correct Choice}) = \exp\big(-\mathcal{L}_\theta^{\text{Vocab}}(\text{Correct Choice})\big) \tag{2}$$

Next, probabilities are restricted to the set of available choices $\{\text{Available Choice}_i\}_i^{|\text{Available Choices}|}$ by masking invalid continuations and normalizing again with respect to this set:

$$p_\theta^{\text{Choices}}(\text{Correct Choice}) \stackrel{\text{def}}{=} \frac{p_\theta^{\text{Vocab}}(\text{Correct Choice})}{\sum_i p_\theta^{\text{Vocab}}(\text{Available Choice}_i)} \tag{3}$$

We distinguish the support over the token space of the model versus over the set of available choices in the benchmark's question because, as we will show, the support crucially affects predictability. Finally, the choices-normalized probability masses become standard downstream metrics:

$$\texttt{Accuracy}_\theta \stackrel{\text{def}}{=} \mathbb{1}\Big(\text{Correct Choice} == \arg\max_i \Big\{p_\theta^{\text{Choices}}(\text{Available Choice}_i)\Big\}\Big) \tag{4}$$

$$\texttt{Brier Score}_\theta \stackrel{\text{def}}{=} \sum_i \Big(\mathbb{1}(\text{Available Choice}_i == \text{Correct Choice}) - p_\theta^{\text{Choices}}(\text{Available Choice}_i)\Big)^2 \tag{5}$$

where $\mathbb{1}(\cdot)$ is an indicator variable. We next demonstrate that this sequence of transformations degrades how predictable performance is with scale before identifying the underlying mechanism.

To quantify how this sequence of transformations affects predictability of performance, we measured how per-sample scores correlate with pretraining compute, and then studied how the distribution (over samples) of correlation values shifted as one transitions from log likelihoods to $p_\theta^{\text{Vocab}}(\text{Correct Choice})$ to $p_\theta^{\text{Choices}}(\text{Correct Choice})$ to `Accuracy` or `Brier Score`. Specifically, for each combination of (*model family, benchmark, performance metric, correlation metric*), we computed a correlation value for each sample in the benchmark between pretraining compute and scores. This yielded a distribution (over samples) of correlation values for the combination (Fig. 2 Left). Visualizing the distribution of correlations for the combination told us what fraction of samples in the benchmark yielded scores that are correlated, uncorrelated or anticorrelated with compute (Fig. 2 Right). We used three standard correlation metrics – Pearson (1895), Kendall (1938) and Spearman (1961) – and found consistent results.

We demonstrate how the sequence of transformations affects the distribution of score-compute correlations using ARC Challenge (Clark et al., 2018) as an illustrative benchmark; we note that all other benchmarks exhibited similar patterns as well (App. G). We visualized the distributions via their complementary (empirical) cumulative distribution functions (complementary CDFs) (App. B):

$$\hat{S}(c) \stackrel{\text{def}}{=} \frac{1}{S}\sum_{s=1}^{S} \mathbb{1}\{C_s > c\}, \tag{6}$$

where S is the number of samples in the benchmark and C_s is the correlation (over the models in the model family) between compute and scores on the s-th sample in the benchmark. For a given threshold

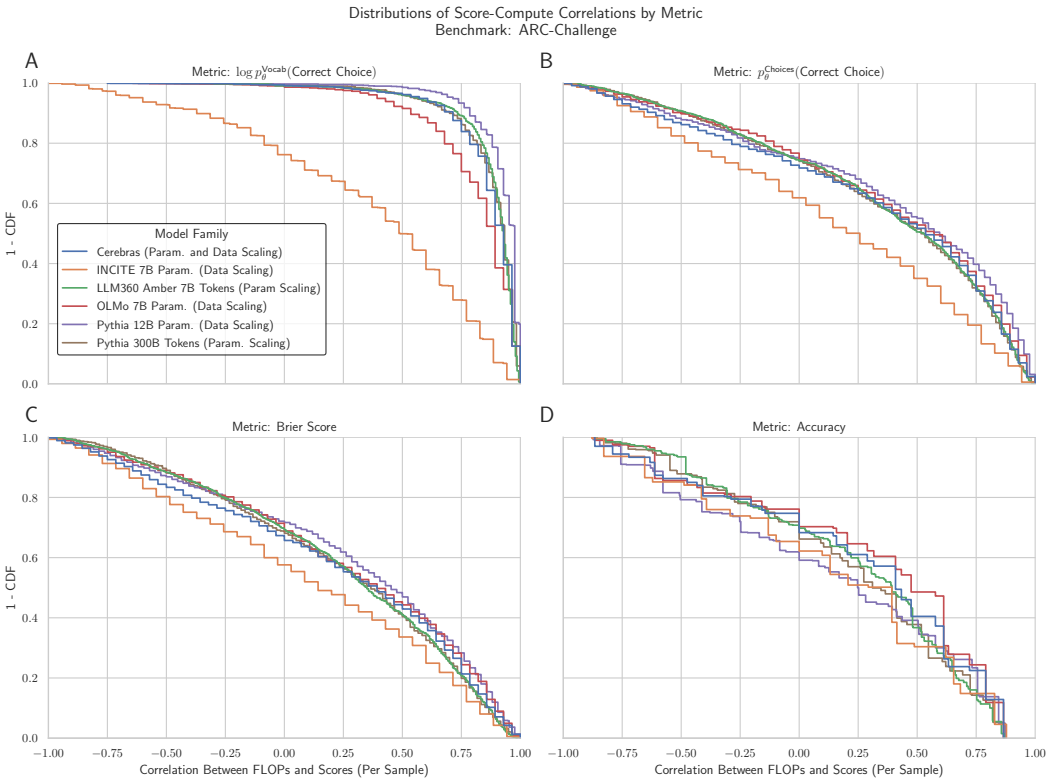

Figure 3: **Multiple-choice metrics like `Accuracy` and `Brier Score` are computed via a sequence of transformations that degrades correlations between performance scores and pre-training compute. (A)** Initially, scores under $\log p_\theta^{\text{Vocab}}(\text{Correct Choice})$ and compute are highly correlated. Transforming $\log p_\theta^{\text{Vocab}}(\text{Correct Choice})$ into $p_\theta^{\text{Vocab}}(\text{Correct Choice})$ has no effect for rank correlations. **(B)** Transforming $p_\theta^{\text{Vocab}}(\text{Correct Choice})$ into $p_\theta^{\text{Choices}}(\text{Correct Choice})$ decorrelates scores from compute. **(C)** Transforming $p_\theta^{\text{Choices}}(\text{Correct Choice})$ into `Brier Score` minorly decreases score-compute correlations. **(D)** Transforming $p_\theta^{\text{Choices}}(\text{Correct Choice})$ into `Accuracy` more substantially decorrelates scores from compute. Benchmark: ARC Challenge Clark et al. (2018). Correlation: Spearman. Results are consistent across benchmarks and all three correlation metrics; for more, see App. G.

c, the complementary CDF $\hat{S}(c)$ returns the fraction of the benchmark's samples with score-compute correlations greater than the threshold c (Fig. 3A). Beginning with log likelihoods, approximately 90% of samples exhibit score-compute correlations > 0.75, regardless of the model family (Fig. 3A). Transforming negative log likelihoods into probability masses $p_\theta^{\text{Vocab}}(\text{Correct Choice})$ does not affect the distribution of score-compute correlations for Spearman and Kendall. However, transforming $p_\theta^{\text{Vocab}}(\text{Correct Choice})$ into $p_\theta^{\text{Choices}}(\text{Correct Choice})$ decreases the distribution of score-compute correlations (Fig. 3B), with only 40% of samples having score-compute correlations > 0.75. Transforming $p_\theta^{\text{Choices}}(\text{Correct Choice})$ into `Brier Score` has little-to-no effect (Fig. 3C), but transforming into `Accuracy` (Fig. 3D) furthers decreases score-compute correlations. To quantitatively test whether these transformations indeed decrease the correlation between scores and compute, we measured four statistics of these score-compute correlation distributions: (1) the mean, (2) the median, (3) the area under the complementary CDF and (4) the negative[3] of the minimum of two Wasserstein distances: between the empirical correlation distribution and an ideal distribution of all correlations $= 1$, and between the empirical distribution and an ideal distribution of all correlations $= -1$. Across the four summary statistics, for most benchmarks and for most model families, we

[3]We chose the *negative* Wasserstein distance for consistency with the other statistics: higher values correspond to higher correlations between scores and compute.

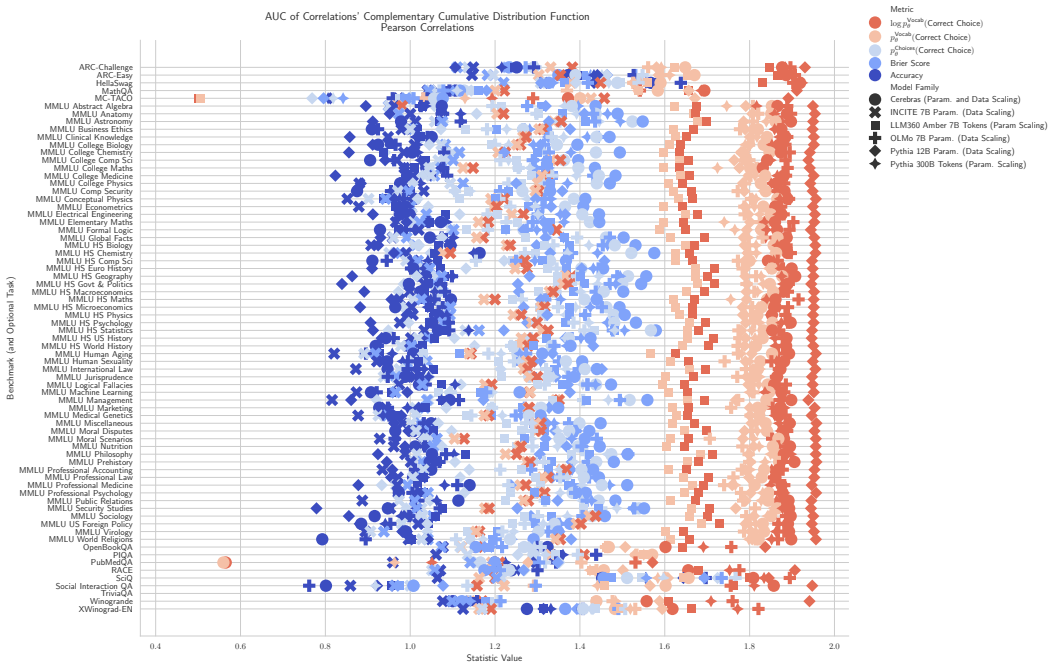

Figure 4: **All four statistics of score-compute correlation distributions demonstrate that transforming $\log p_\theta^{\text{Vocab}}$(Correct Choice) into `Accuracy` causes score-compute correlations to deteriorate.** We find a consistent trend that the sequence of transformations degrades score-compute correlations, as shown by the right-to-left $\log p_\theta^{\text{Vocab}}$(**Correct Choice**)-to-$p_\theta^{\text{Vocab}}$(**Correct Choice**)-to-$p_\theta^{\text{Choices}}$(**Correct Choice**)-or-**`Brier Score`**-to-**`Accuracy`** vertical stripes. This trend holds across benchmarks and model families for three correlation metrics (Spearman, Pearson and Kendall) and for four statistics of correlation distributions (mean, median, the area under the survival function, and negative Wasserstein distance from perfect correlation or perfect anti-correlation). See App. Figs. 7, 8, 9 for other correlation metrics and other score-compute correlation distribution statistics.

discovered a consistent ordering of metrics of the score-compute correlation distributions (Fig. 4):

$$\begin{aligned}
&\text{Corr}\big(\text{Compute}, \log p_\theta^{\text{Vocab}}(\text{Correct Choice})\big)\\
&\quad\geq \text{Corr}\big(\text{Compute}, p_\theta^{\text{Vocab}}(\text{Correct Choice})\big)\\
&\quad> \text{Corr}\big(\text{Compute}, p_\theta^{\text{Choices}}(\text{Correct Choice})\big)\\
&\quad\geq \text{Corr}\big(\text{Compute}, \texttt{Brier Score}\big)\\
&\quad> \text{Corr}\big(\text{Compute}, \texttt{Accuracy}\big)
\end{aligned}$$

5 Probability Mass on Incorrect Choices Causes Unpredictability

What is the mechanism that degrades how correlated scores are with compute? All three metrics with degraded correlations - $p_\theta^{\text{Choices}}$(Correct Choice), `Accuracy`, and `Brier Score` - depend not just on how the model's probability mass p_θ^{Vocab}(Correct Choice) concentrates on the correct choice as compute increases, but also depend on how the model's probability mass fluctuates on *incorrect* available choices $\{p_\theta^{\text{Vocab}}(\text{Incorrect Choice})\}_{\text{Incorrect Choices}}$ as compute increases. As an example, suppose $p_\theta^{Vocab}(\text{Correct Choice}) = 0.4$ on a 4-way multiple-choice question; what is the accuracy? Spreading the remaining mass uniformly on the incorrect choices will make `Accuracy` $= 1$, whereas concentrating mass on a single incorrect choice will make `Accuracy` $= 0$.

To demonstrate how drastically the probability mass placed on incorrect choices can alter performance, we visualized the relationships between pairs of metrics immediately preceding and following a given

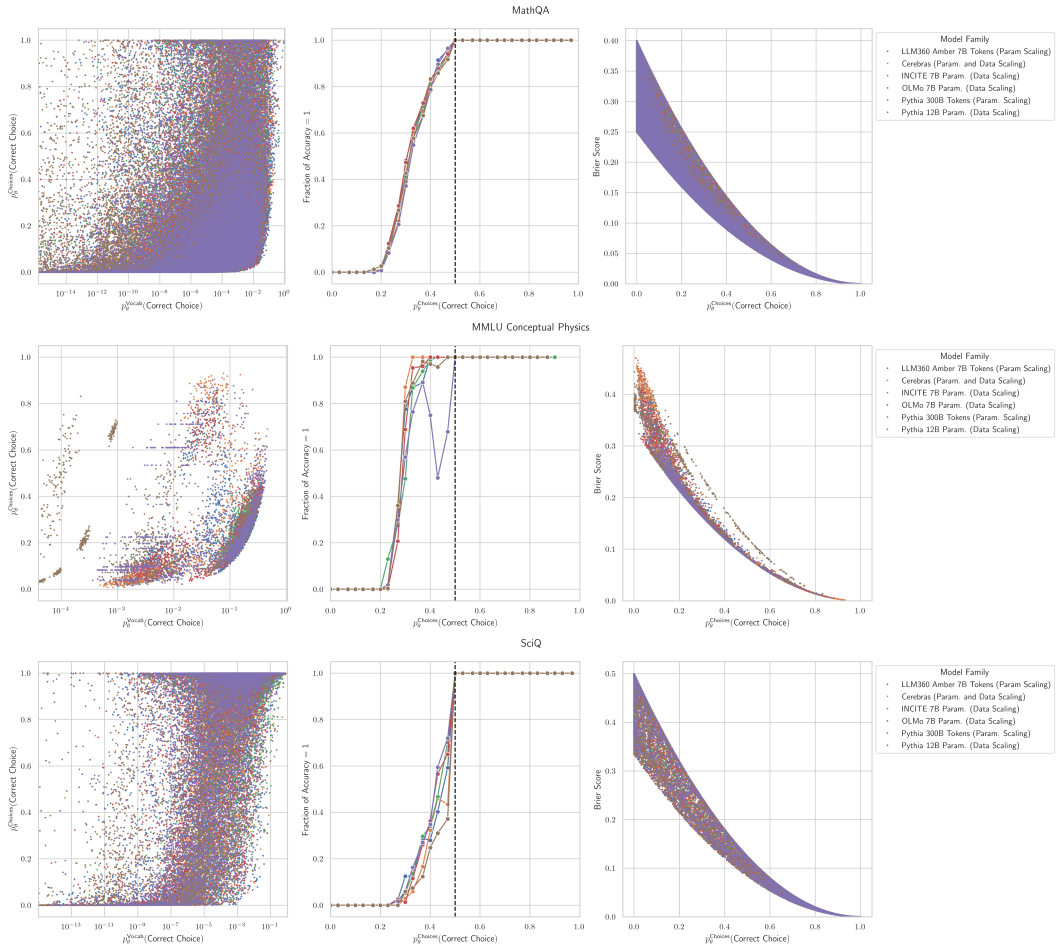

Figure 5: **Predictability deteriorates because of probability mass fluctuating on specific incorrect choices with scale. Left:** Transitioning from $p_\theta^{\text{Vocab}}(\text{Correct Choice})$ to $p_\theta^{\text{Choices}}(\text{Correct Choice})$ demonstrates that $p_\theta^{\text{Vocab}}(\text{Correct Choice})$ contains little information about $p_\theta^{\text{Choices}}(\text{Correct Choice})$ and vice versa; loosely speaking, any value of one can map to any value of the other. **Center:** While $p_\theta^{\text{Choices}}(\text{Correct Choice}) > 0.5$ must yield $\texttt{Accuracy} = 1$, for any $p_\theta^{\text{Choices}}(\text{Correct Choice}) < 0.5$, knowing $p_\theta^{\text{Choices}}(\text{Correct Choice})$ contains little information about `Accuracy` and vice versa. **Right**: `Brier Score` is more predictable from $p_\theta^{\text{Choices}}(\text{Correct Choice})$ than `Accuracy`, but still quite variable. Three benchmarks shown: MathQA Amini et al. (2019), MMLU Conceptual Physics Hendrycks et al. (2020), SciQ Welbl et al. (2017).

transformation (Fig. 5). For negative log likelihood of the correct choice and $p_\theta^{Vocab}(\text{Correct Choice})$ (not pictured), we observed a clean correspondence between performance under the metric and compute: one can reliably map a given value of these metrics to compute, and vice versa. In contrast, once performance is evaluated using a metric that is a function of the incorrect choices - $p_\theta^{\text{Choices}}(\text{Correct Choice})$, `Accuracy` or `Brier Score` - nearly any value of a score under one metric can map to any value of $p_\theta^{\text{Vocab}}(\text{Correct Choice})$ or $p_\theta^{\text{Choices}}(\text{Correct Choice})$ respectively (Fig. 5), thereby breaking the chain along which one can cleanly infer compute from an observed metric. We can see that `Brier Score`, a metric meant to produce more continuous scores (Schaeffer et al., 2023), is less variable than `Accuracy`, provided a known $p_\theta^{\text{Choices}}(\text{Correct Choice})$, but it cannot recover information about $p_\theta^{\text{Vocab}}(\text{Correct Choice})$ that is lost when shifting to $p_\theta^{\text{Choices}}(\text{Correct Choice})$. We next show that this is because of the additional information regarding the underdetermined values of $p_\theta^{\text{Choices}}(\text{Incorrect Choice})$ for each incorrect choice.

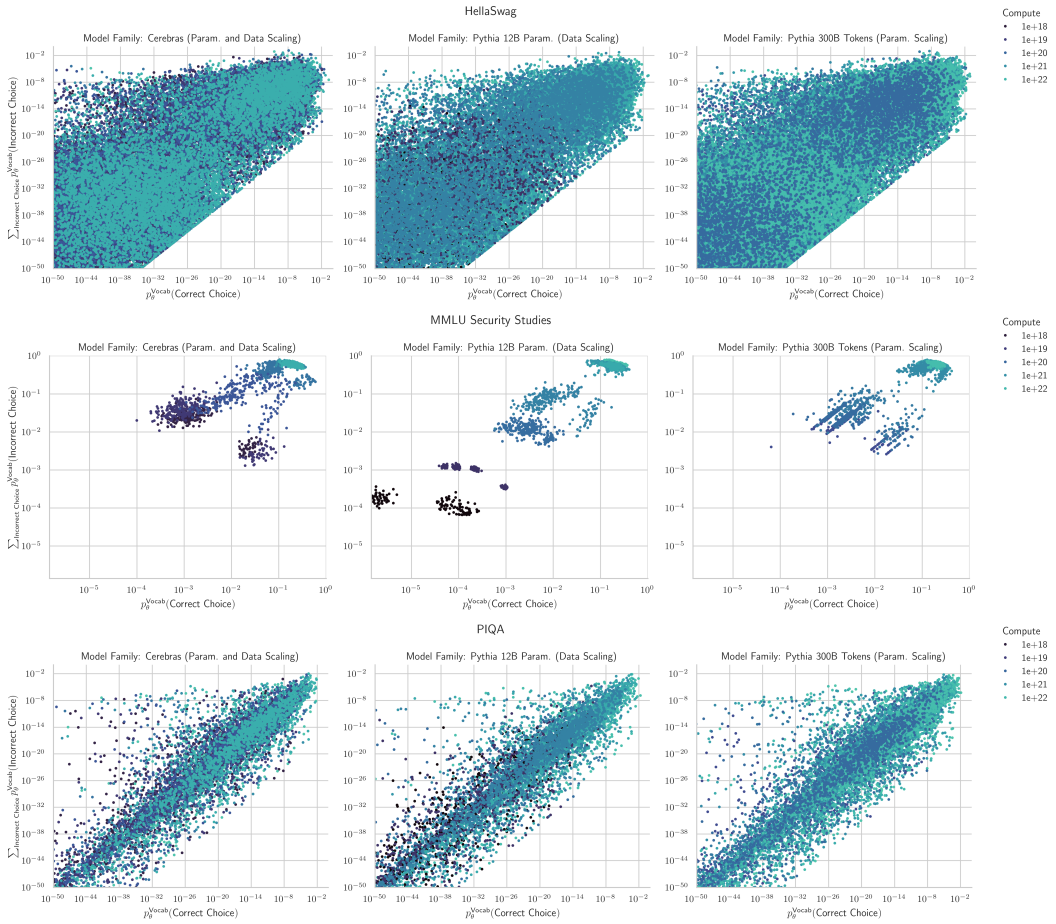

Figure 6: **Probability mass on the correct choices and the incorrect choices are correlated, but can fluctuate substantially.** Probability mass on correct choices and incorrect choices positively covaries and typically increases with compute. However, the spread is large: for any given value of p_θ^{Vocab}(Correct Choice), the mass on incorrect choices can vary by many orders of magnitude.

6 Scaling Behavior of Probability Mass on Incorrect Choices

In general, aggregate performance over a distribution is often of interest. Such a focus on aggregate performance leads to an important insight: in MCQA, probability mass fluctuations on incorrect choices do not "average out". Unlike estimating the mean of a random variable, where positive and negative deviations cancel, the nonlinear nature of metrics like Accuracy and Brier Score means that probability mass shifts between incorrect options affect scores in complex ways that persist under averaging. For example, if probability mass shifts from incorrect option A to incorrect option B, the impact on accuracy depends on whether either option had enough mass to compete with the correct answer - there's no natural cancellation. This perhaps counter-intuitive behavior partially explains why predicting aggregate performance using typical averaging has remained elusive.

This analysis suggests that modeling probability mass fluctuations on incorrect choices could improve predictions of metrics like Accuracy and Brier Score, though the magnitude of improvement remains an open question. For metrics like `Accuracy`, such predictions should be made for each sample because knowing the average mass (across many data) placed on incorrect choices says little about how much mass is placed on any single incorrect choice for a single sample. We conclude by providing preliminary evidence that achieving such a feat might be possible. Specifically, we test how probability masses on correct choices and probability masses on incorrect choices covary with increasing compute (Fig. 6). Multiple benchmarks display strong positive relationships between

Takeaway #1: Think through your metrics!
If one cares about scaling-predictable evaluations, then one needs to think through how their evaluations transform raw model outputs into useful signals to know what to expect.

Takeaway #2: Continuous metrics are insufficient to guarantee predictable changes.
As shown by $p_\theta^{\text{Choices}}(\text{Correct Choice})$ & `Brier Score`, even "continuous" metrics can be unpredictable, e.g., if the metric weighs correct behavior against specific incorrect behaviors.

Takeaway #3: Recommended scaling-predictable metrics for pretraining practitioners.
Pretraining practitioners seeking scaling-predictable signals for capabilities are advised to focus on $p_\theta^{Vocab}(\text{Correct Choice})$ on relevant benchmarks. Scores under this metric provide smoother scaling trends and are arguably more interpretable than the pretraining loss.

Takeaway #4: Evaluations should be reshaped based on intended desiderata.
Too often, we take evaluations as frozen static objects, but evaluations should be adapted to pertinent goals. For instance, if the goal is to predict capabilities with scale, evaluations should be designed or adapted to be scaling-predictable.

mass on correct choices and mass on incorrect choices, suggesting that fitting *per-sample scaling trends for each incorrect choice* might be possible; doing so would enable predicting changepoints in metrics like `Accuracy` or `Brier Score`. However, whether per-benchmark per-sample per-choice scaling trends can be fit and accurately extrapolated is unclear since the spread varies by several orders of magnitude. We leave this challenge to future work.

7 Discussion, Related Work and Future Directions

This work identifies a factor that induces unpredictability in multiple-choice assessments of frontier AI models, as well as the underlying mechanism: probability mass on incorrect choices. Our results have implications for the design of future evaluations of frontier AI models that are reliably predictable with scaling. We hope that our work will be extended to further the science of scaling-predictable evaluation of AI systems, especially for complex and important model capabilities. We note several future directions for extension of our work, and we hope that the community also adopts our framing to further improve scaling-predictable evaluations.

Related Work We intentionally wove key related work into our main text, with a particular emphasis in the Introduction. For a longer and more comprehensive exposition of related work, please see Appendix A.

Direction 1: Beyond Multiple Choice Benchmarks Our study is restricted to benchmarks evaluated via log likelihood-based multiple-choice formats. While we believe this is inherently valuable due to the usefulness and prevalence of such tasks, this limits the application of our findings. We hope that our discoveries and proposed mechanisms may be used to inform the study of predictable and reliable evaluation writ large, and that future work should explore the extent to which our findings can be generalized to more complex capabilities. Our findings corroborate those of Lyu et al. (2024), who find that multiple-choice answer scores often diverge from generative evaluations. Consequently, a particularly important direction for further study is to investigate generative evaluations, which may contain similar transformations distancing performance from the observed loss.

Direction 2: Predicting Benchmark Performance A Priori Our work provides an explanation why multiple-choice benchmark performance is not easily predictable for metrics such as `Accuracy` and `Brier Score`, as observed in the literature (Du et al., 2024). However, our analyses assume access to entire model families' scores across several orders of magnitude of pretraining FLOPs, and do not employ backtesting, as sensibly recommended by (Alabdulmohsin et al., 2022; Owen, 2024). A predictive model should be able to identify change points well in advance on standard metrics like `Accuracy` or `Brier Score`.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Together AI. Releasing 3b and 7b redpajama-incite family of models including base, instruction-tuned & chat models. https://www.together.ai/blog/redpajama-models-v1, 2023. Accessed: 2024-05-19.

Ibrahim Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision, 2022. URL https://arxiv.org/abs/2209.06640.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2357–2367, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1245. URL https://aclanthology.org/N19-1245.

Anthropic. Anthropic's responsible scaling policy. https://www.anthropic.com/news/anthropics-responsible-scaling-policy, 2023. Accessed: 2024-05-19.

Anthropic. Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family, 2024. Accessed: 2024-05-19.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt, 2024.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models, 2023a.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023b.

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

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. 2020.

Samuel R. Bowman. Eight things to know about large language models, 2023.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws, 2023.

Aidan Clark, Diego De Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. Unified scaling laws for routed language models. In *International Conference on Machine Learning*, pp. 4057–4086. PMLR, 2022.

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 preprint arXiv:1803.05457*, 2018.

Together Computer. Redpajama: An open source recipe to reproduce llama training dataset, 2023. URL `https://github.com/togethercomputer/RedPajama-Data`.

Wikipedia contributors. Survival function, 2023. URL `https://en.wikipedia.org/wiki/Survival_function`. [Online; accessed 22-May-2024].

Council of Economic Advisers. The 2024 economic report of the president, 03 2024. URL `https://www.whitehouse.gov/cea/written-materials/2024/03/21/the-2024-economic-report-of-the-president/`. Accessed on 09/29/2024.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. Deepseek llm: Scaling open-source language models with longtermism, 2024.

Nolan Dey, Gurpreet Gosal, Zhiming, Chen, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster, 2023.

Anca Dragan, Helen King, and Allan Dafoe. Introducing the frontier safety framework. `https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/`, 2024. Accessed: 2024-05-19.

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. Understanding emergent abilities of language models from the loss perspective, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz

Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh

Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What's in my big data? In *The Twelfth International Conference on Learning Representations*, 2023.

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

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1747–1764, 2022.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization, 2022.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Mitchell A Gordon, Kevin Duh, and Jared Kaplan. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5915–5922, 2021.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. Olmo: Accelerating the science of language models, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. 2021.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

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, 2024.

Yuzhen Huang, Jinghan Zhang, Zifei Shan, and Junxian He. Compression represents intelligence linearly, 2024.

Andy L Jones. Scaling scaling laws with board games. *arXiv preprint arXiv:2104.03113*, 2021.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints*, art. arXiv:1705.03551, 2017.

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. *arXiv preprint arXiv:2001.08361*, 2020.

Sakaguchi Keisuke, Le Ronan, Bhagavatula Chandra, and Choi Yejin. Winogrande: An adversarial winograd schema challenge at scale. 2019.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

David G. Kleinbaum and Mitchel Klein. *Survival Analysis: A Self-Learning Text*. Springer, 3 edition, 2012. ISBN 978-1441966452. doi: 10.1007/978-1-4419-6646-9.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL https://aclanthology.org/D17-1082.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P. Xing. Llm360: Towards fully transparent open-source llms, 2023.

Chenyang Lyu, Minghao Wu, and Alham Fikri Aji. Beyond probabilities: Unveiling the misalignment in evaluating large language models. *arXiv preprint arXiv:2402.13887*, 2024.

Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training, 2018.

Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. The inverse scaling prize, 2022. URL `https://github.com/inverse-scaling/prize`.

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 *EMNLP*, 2018.

Sherin Muckatira, Vijeta Deshpande, Vladislav Lialin, and Anna Rumshisky. Emergent abilities in reduced-scale generative language models, 2024.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning, 2023.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Oren Neumann and Claudius Gros. Scaling laws for a multi-agent reinforcement learning model. *arXiv preprint arXiv:2210.00849*, 2022.

OpenAI. Openai's approach to frontier risk. `https://openai.com/global-affairs/our-approach-to-frontier-risk/`, 2023. Accessed: 2024-05-19.

OpenAI. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`, 2024. Accessed: 2024-05-16.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo

Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

David Owen. How predictable is language model benchmark performance?, 2024.

Karl Pearson. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242, 1895.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations*, 2019.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance, 2024.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019a.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions, 2019b.

Nikhil Sardana and Jonathan Frankle. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws, 2023.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 55565–55581. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/adc98a266f45005c403b8311ca7e8bd7-Paper-Conference.pdf.

Charles Spearman. The proof and measurement of association between two things. 1961.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin (eds.), *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method, 2024. URL https://arxiv.org/abs/2402.17193.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3363–3369, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1332. URL https://aclanthology.org/D19-1332.

A Related Work

Language Model Evaluation The capabilities of AI models are typically evaluated using constructed datasets to assess performance on a specific task, acting as a proxy for some real-world usage scenario. However, performing robust and reliable evaluations is a challenge, with many potential pitfalls and unsolved problems (Biderman et al., 2024). For example, we might prefer to ask models open-ended questions and evaluate their answers in natural language, but it then often becomes difficult to robustly score the resulting model outputs, especially for partial correctness. For this reason, it is common practice for evaluation benchmarks to simplify their scoring via approximations, such as extracting a sub-string from free-form outputs heuristically (Joshi et al., 2017; Kwiatkowski et al., 2019; Hendrycks et al., 2021) and checking that it matches a specific gold target string, or casting a task to a *multiple-choice* format, in which a closed set of correct and incorrect answers is known, and the model's answer is determined by selecting the most likely option among these strings. For more details on the precise procedures typically used for multiple choice elsewhere in the literature, see Biderman et al. (2024). We believe that the multiple-choice format is valuable, due to its flexibility, popularity and relevance (Brown et al., 2020; Beeching et al., 2023; Biderman et al., 2024), but we discuss its limitations in Section 7.

Scaling Laws Many neural networks exhibit power-law scaling of the pretraining loss as a function of the amount of compute, data, or parameters used for training (Hestness et al., 2017; Brown et al., 2020; Hoffmann et al., 2022). These neural scaling laws demonstrate that the pretraining loss can be highly predictable as a function of these fundamental inputs, which has a number of practical applications: Scaling laws fit to smaller training runs can be used to predict the pretraining loss of a much larger training run, and can be used to determine effective hyperparameters (McCandlish et al., 2018; DeepSeek-AI et al., 2024), or the optimal allocation of dataset and model size for a given compute budget (Hoffmann et al., 2022; Muennighoff et al., 2024; Dey et al., 2023; Sardana & Frankle, 2023; Besiroglu et al., 2024). In some cases, such laws can be used to predict performance of a larger model in a particular domain, such as coding (Achiam et al., 2023). The existence of scaling laws turns deep learning into a predictable science at the macro level by providing a simple recipe for improving model quality and de-risking returns on increasing investment into scale (Ganguli et al., 2022; Bowman, 2023).

Emergent Abilities Language models have been observed to exhibit apparent *emergent abilities*—behaviors on downstream task performance that cannot be predicted from smaller scales (Wei et al., 2022; Srivastava et al., 2022). Emergence appears not to be simply a product of training compute or model size, but is also dependent on other factors such as dataset composition (Muckatira et al., 2024; Wei et al., 2022). Schaeffer et al. (2023) find that some emergent phenomena can be a "mirage" arising due to choices made by researchers such as the use of discontinuous metrics and insufficient resolution. However, Du et al. (2024) note that for many tasks, emergence remains despite the use of continuous metrics. Additionally, discontinuous metrics have been argued to often be the most reflective of real-world usefulness, so emergence in these hard metrics is important. Hu et al. (2024) found that for generative evaluations, infinite resolution can be achieved but requires significant compute and that generated answer be verifiable.

Predicting Downstream Task Performance Although predicting macroscopic pretraining loss is useful, a far more useful goal is to predict the scaling of model performance on particular downstream tasks or domains. If this was possible, then model developers could tune their datasets and training procedures in a more fine-grained way before launching computationally intensive training runs. Model performance on a particular downstream task is typically correlated with compute, albeit with a few exceptions (McKenzie et al., 2022; Huang et al., 2024). However, despite attempts to fit scaling laws to values other than loss, including benchmark scores (Gadre et al., 2024; Zhang et al., 2024), model memorization (Biderman et al., 2023a), or reward (Gao et al., 2022), these downstream performance metrics are usually more noisy or require more compute to fit accurately. Owen (2024) and Gadre et al. (2024) both find that while *aggregate* benchmark performance with more compute can be predicted, the scaling behaviour of individual tasks can be noisy. Additionally, Owen (2024), Du et al. (2024) and Gadre et al. (2024) claim that predicting scaling behavior on a task without access to models exhibiting better-than-random performance (i.e., "before emergence occurs") cannot be done reliably. Concurrently to our work, Ruan et al. (2024) propose Observational Scaling Laws

by mapping model capabilities from compute to a shared low-dimensional space of capabilities across model families before predicting performance on novel tasks. Our goal in this work is to investigate the comparative unpredictability of individual downstream performance scores, and advise how to create more scaling-predictable evaluations that are closely coupled with real-world use-cases.

B Definition of Survival Function

The survival function $S_X(x)$ – also known as the reliability function, the tail distribution, or the complementary cumulative distribution function – gives the probability that a random variable X exceeds a certain value x Kleinbaum & Klein (2012); contributors (2023):

$$S_X(x) \stackrel{\text{def}}{=} Pr[X > x] = \int_x^{\infty} f_X(x')\, dx' = 1 - F_X(x) \tag{7}$$

where $F_X(x) = Pr[X \leq x]$ is the cumulative distribution function (CDF) and $f_X(x)$ is the probability density function (pdf) or probability mass function (pmf) of the random variable X. The CDF $F_X(x)$ gives the probability that the random variable X is at most x, while the survival function $S_X(x)$ gives the probability that X exceeds x.

When the true distribution of X is unknown, we can use the empirical CDF (ECDF) $\hat{F}_X(x)$ and the empirical survival function (ESF) $\hat{S}_X(x)$:

$$\hat{S}_X(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} 1\{x_i > x\} = 1 - \hat{F}_X(x) \tag{8}$$

where n is the number of observations, x_i is the realized value of the random variable X for observation i, and $1\{x_i > x\}$ is the indicator function. The empirical survival function $\hat{S}_X(x)$ specifies the fraction of observations for which the sampled random variable X exceeds x.

C Compute Resources for Experiments

Experiments were done across a wide family of model families and sizes. The GPUs we used for medium-sized models (7B parameters and above) used a single A100s with 80GB of vRAM. For smaller models ($\leq$8B) we used A100s with 80GB of vRAM, Quadro RTX 8000 with 48GB of vRAM, or RTX A4000 with 16GB of vRAM. For 70B parameter models, we used at least 2 A100 GPUs with 80GB of vRAM.

D Additional Model Family Details

Here we provide further experimental details regarding our selection of model families.

1. **Pythia** (Biderman et al., 2023b): We consider two "families" for Pythia in our experiments. **Pythia (Parameter Scaling)** refers to the use of fully-trained checkpoints from 9 different model sizes (all model sizes documented in Biderman et al. (2023), as well as a 14M parameter model trained later by the authors). **Pythia-12B (Data Scaling)** refers to the use of 8 checkpoints across training for the Pythia-12B model, namely having seen 2M, 64M, 2B, 6B, 20B, 60B, 200B, and 300B tokens in training.
2. **Cerebras-GPT** (Dey et al., 2023): **Cerebras (Parameter and Data Scaling)** refers to our use of 1 checkpoint per model in the Cerebras-GPT family, each fully trained for differing quantities of data as documented by the model creators, for 7 checkpoints in total.
3. **OLMo** (Groeneveld et al., 2024): **OLMo (7B Data Scaling)** refers to the use of 7 checkpoints for OLMo-7B across training, namely, checkpoints having seen 4B, 44B, 133B, 442B, 885B, 1.5T, and 2.4T tokens.
4. **INCITE** (AI, 2023): **INCITE-7B (Data Scaling)** considers 6 checkpoints over training for the 7B parameter model, having seen 240B, 280B, 400B, 500B, 700B, and 1T tokens.
5. **LLM360** (Liu et al., 2023): **LLM360 Amber (Data Scaling)** considers 13 checkpoints of the Amber model, having seen 0B, 3.5B, 7B, 10.5B, 17.5B, 31.5B, 49B, 87.5B, 147B, 252B, 430B, 738B, and 1.26T tokens.

E BROADER IMPACT

This paper contributes to a better understanding of the predictability of large language models (LLMs), which can have both positive and negative societal impacts. On the positive side, by making LLM benchmarks more predictable, this research can help society anticipate and plan for potential challenges associated with their development and deployment. This increased predictability can facilitate proactive measures to mitigate risks and ensure the responsible use of AI technologies.

However, the increased predictability of LLMs could theoretically be exploited by malicious actors to accelerate the development of AI systems designed for malicious purposes. We also stress the importance of proactive risk assessment and the implementation of safeguards to prevent the misuse of AI technologies.

F SCORE-COMPUTE CORRELATION DISTRIBUTIONS' STATISTICS

F.1 PEARSON CORRELATIONS

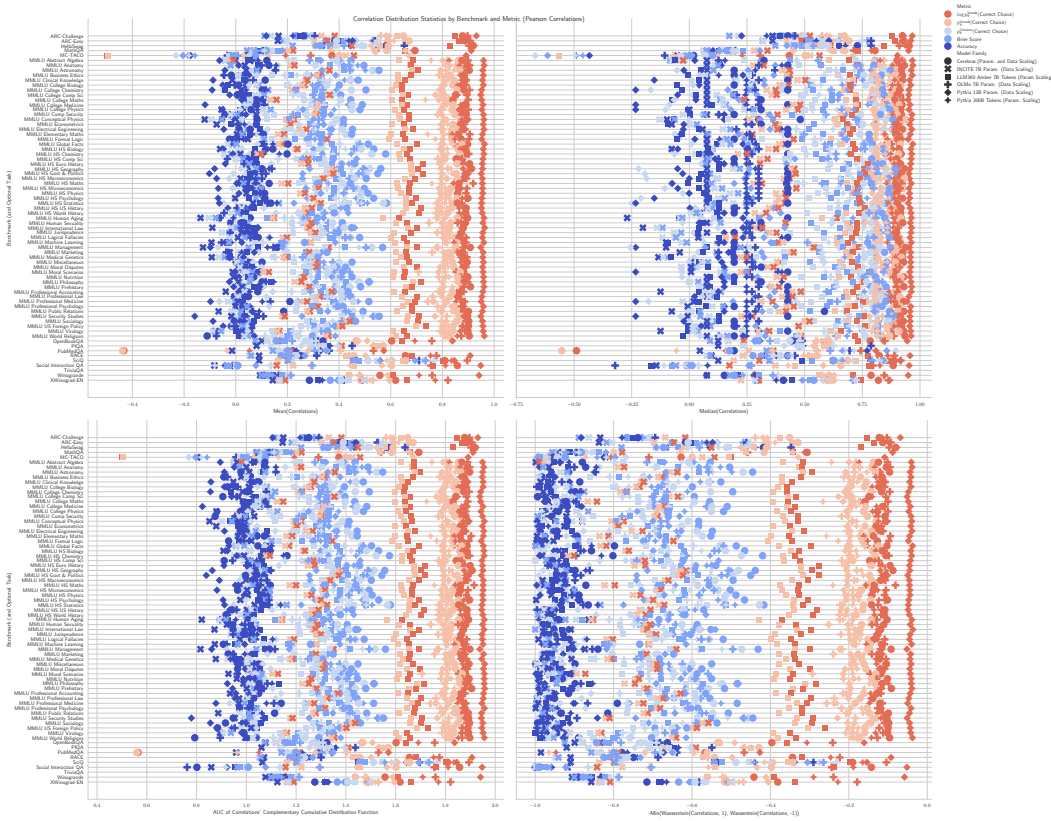

Figure 7: **Statistics for empirical distributions of correlations between scores and compute for all benchmarks and model families.** These correlation values were computed with Pearson correlation and are consistent with the main text's results computed with Spearman correlation (Fig. 4): The sequence of transformations from $\log p_\theta^{\text{Vocab}}(\text{Correct Choice}) \rightarrow p_\theta^{\text{Vocab}}(\text{Correct Choice}) \rightarrow p_\theta^{\text{Choices}}(\text{Correct Choice}) \rightarrow$ Accuracy degrades predictability.

F.2 SPEARMAN CORRELATIONS

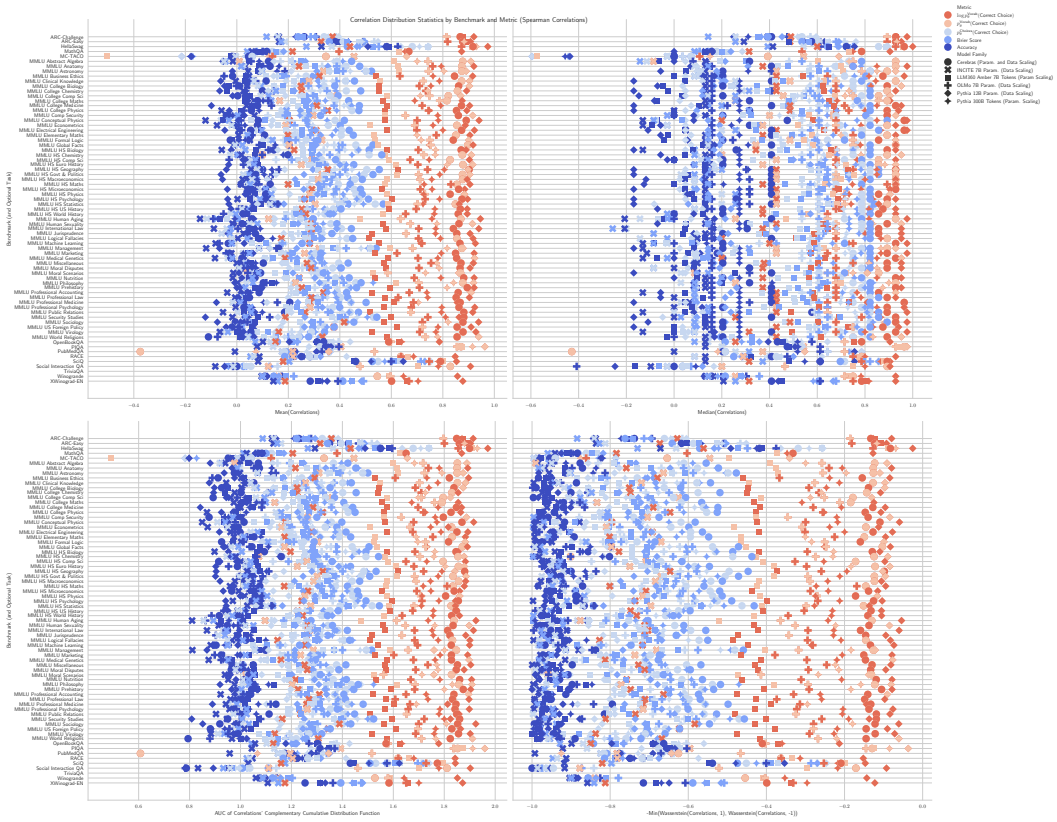

Figure 8: **Statistics for empirical distributions of correlations between scores and compute for all benchmarks and model families.** These correlation values were computed with Spearman correlation. The sequence of transformations from $\log p_\theta^{\text{Vocab}}(\text{Correct Choice}) \rightarrow p_\theta^{\text{Vocab}}(\text{Correct Choice}) \rightarrow p_\theta^{\text{Choices}}(\text{Correct Choice}) \rightarrow$ Accuracy degrades predictability.

F.3 KENDALL CORRELATIONS

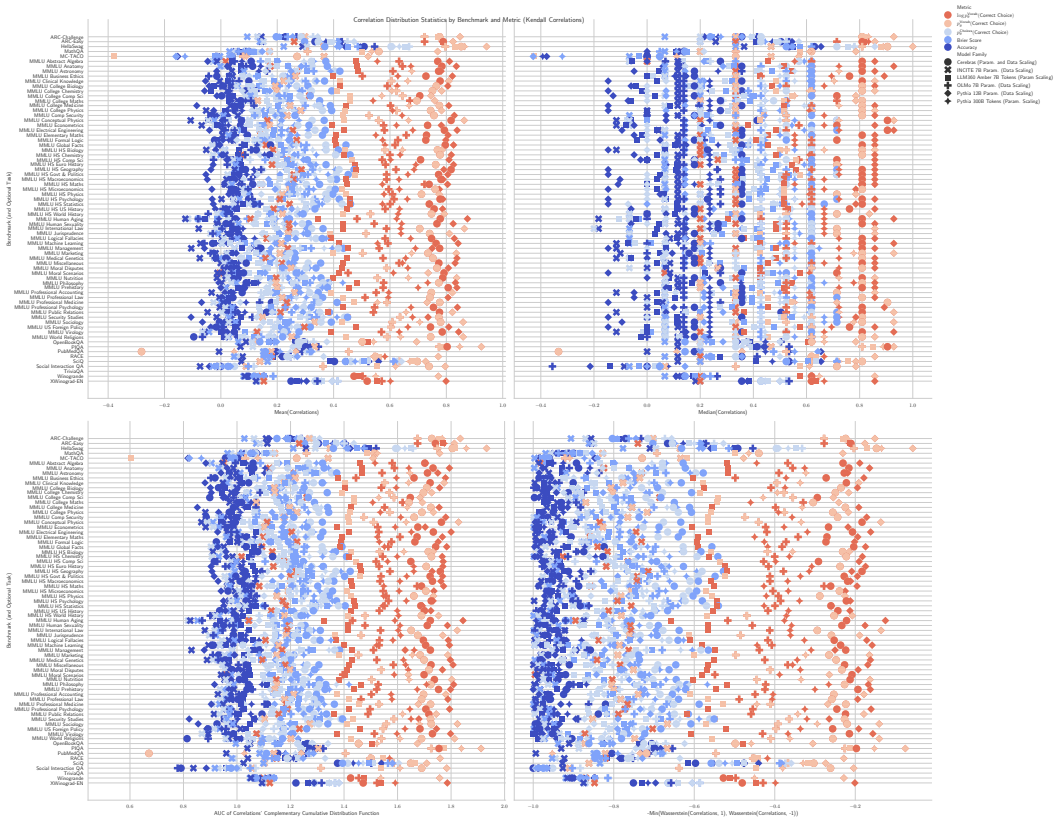

Figure 9: **Statistics for empirical distributions of correlations between scores and compute for all benchmarks and model families.** These correlation values were computed with Kendall correlation and are consistent with the main text's results computed with Spearman correlation (Fig. 4): The sequence of transformations from $\log p_\theta^{\text{Vocab}}(\text{Correct Choice}) \to p_\theta^{\text{Vocab}}(\text{Correct Choice}) \to p_\theta^{\text{Choices}}(\text{Correct Choice}) \to$ Accuracy degrades predictability.

G PER-BENCHMARK SCORE-COMPUTE CORRELATION DISTRIBUTIONS

G.1 NLP BENCHMARK: ARC CHALLENGE CLARK ET AL. (2018)

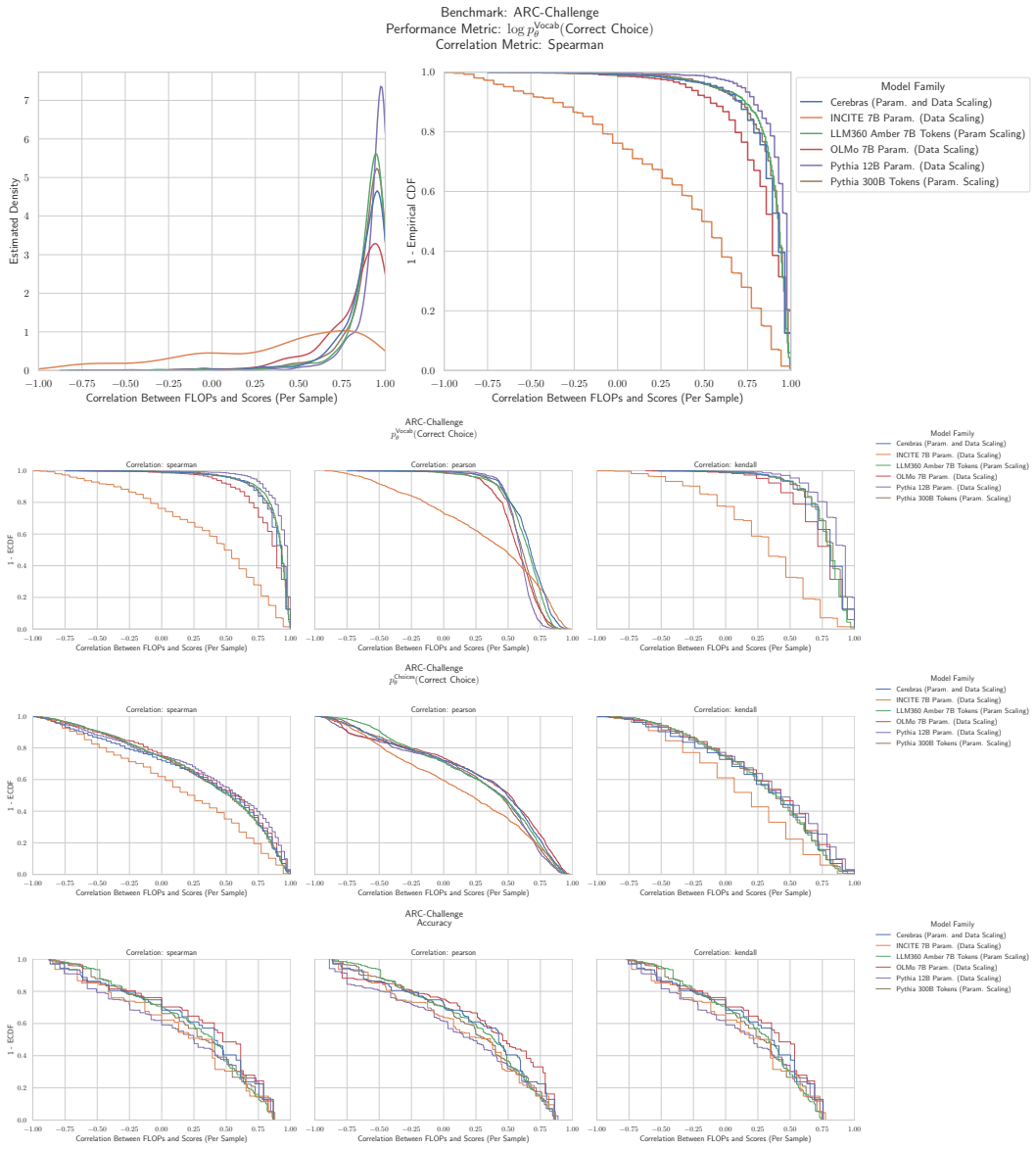

Figure 10: **ARC Challenge: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.2 NLP BENCHMARK: ARC EASY CLARK ET AL. (2018)

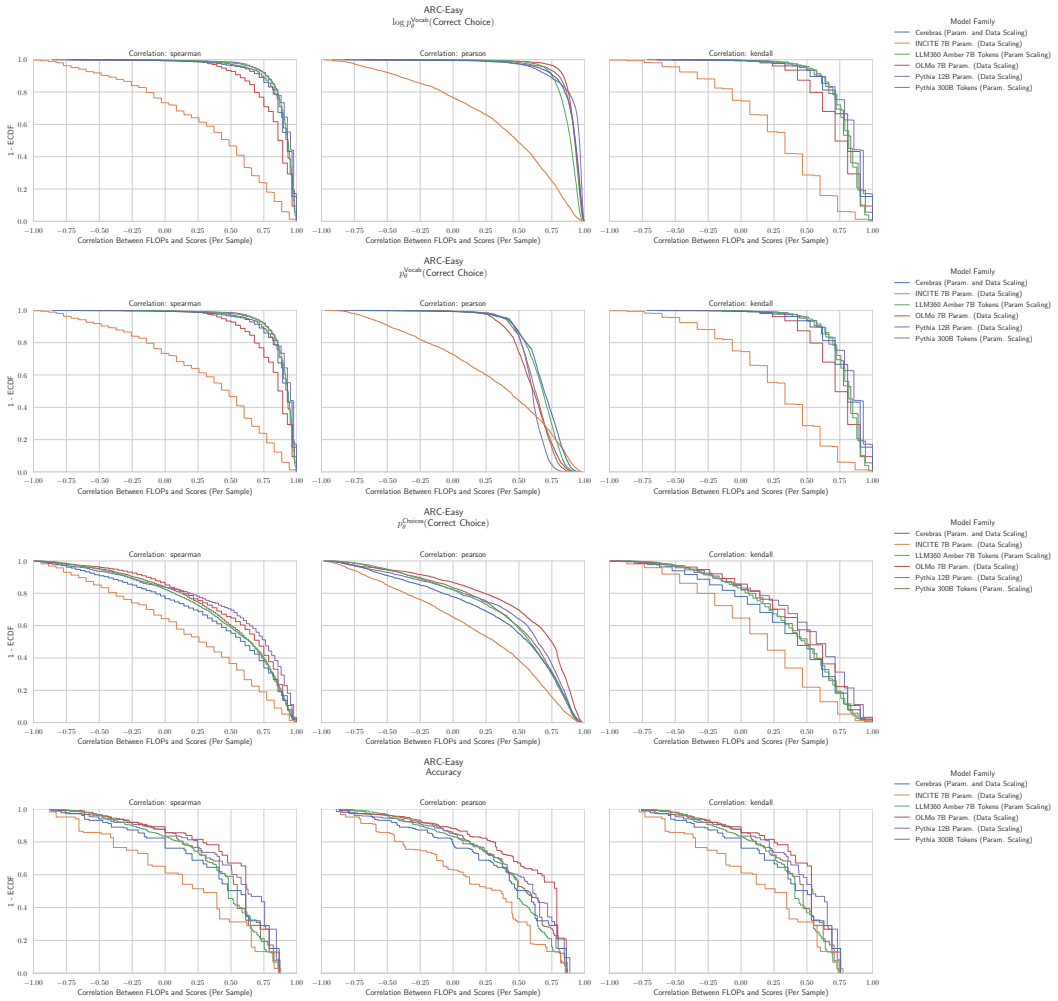

Figure 11: **ARC Easy: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.3 NLP BENCHMARK: HELLASWAG ZELLERS ET AL. (2019)

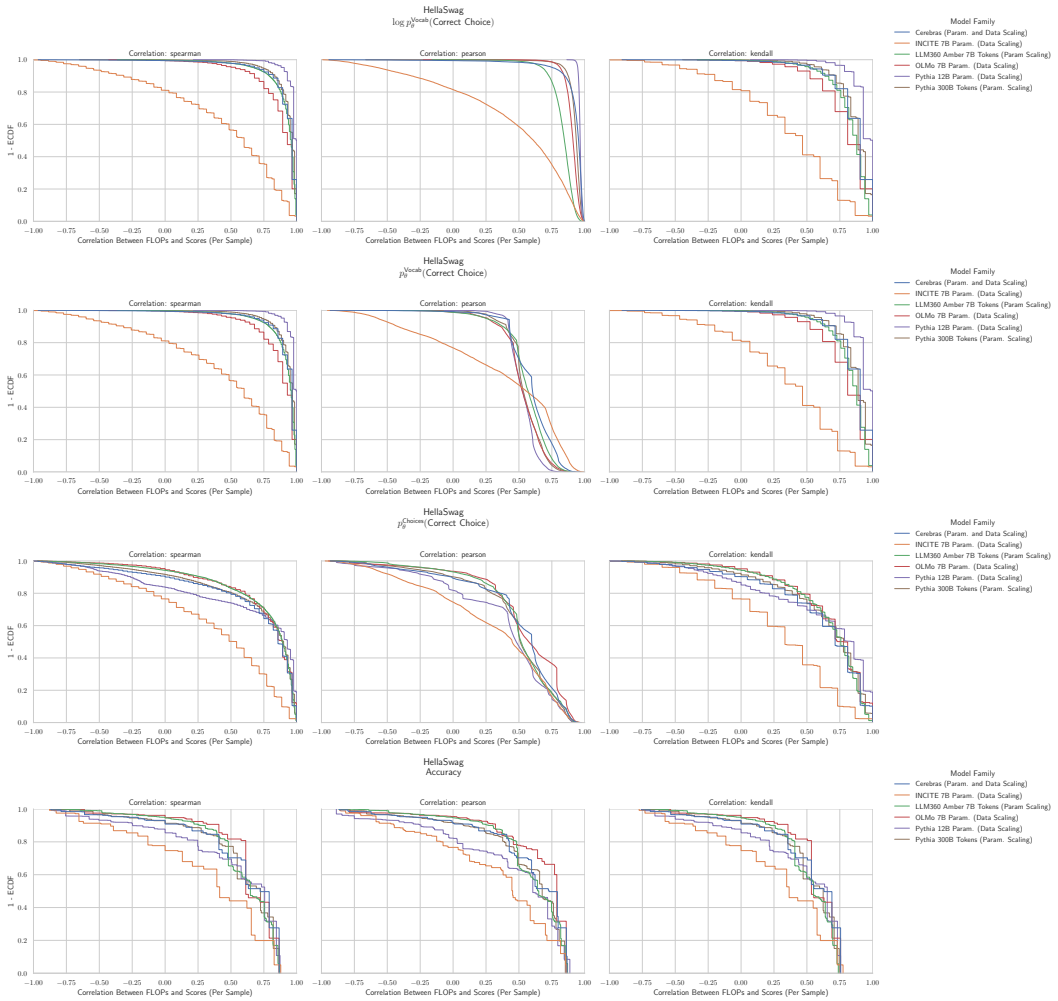

Figure 12: **HellaSwag: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.4 NLP BENCHMARK: MATHQA AMINI ET AL. (2019)

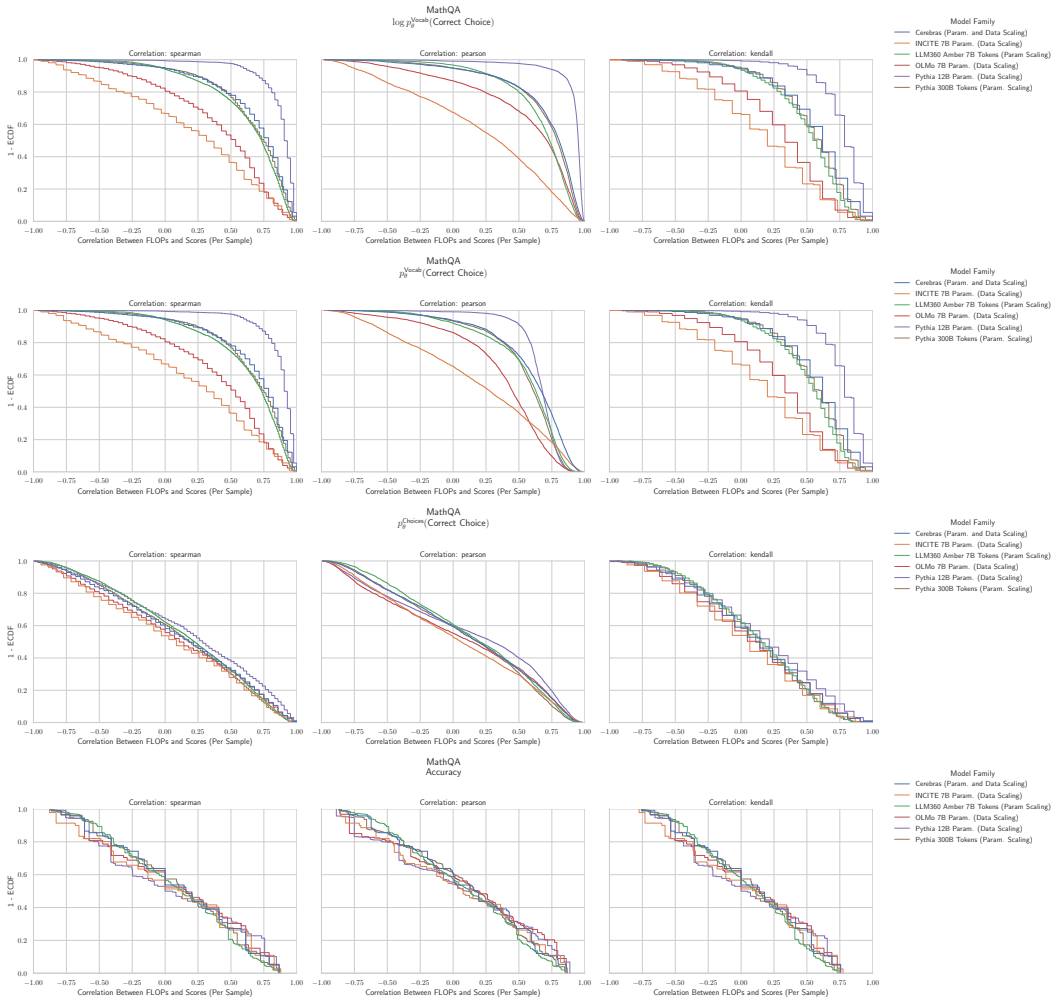

Figure 13: **HellaSwag: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.5 NLP BENCHMARK: MC TACO ZHOU ET AL. (2019)

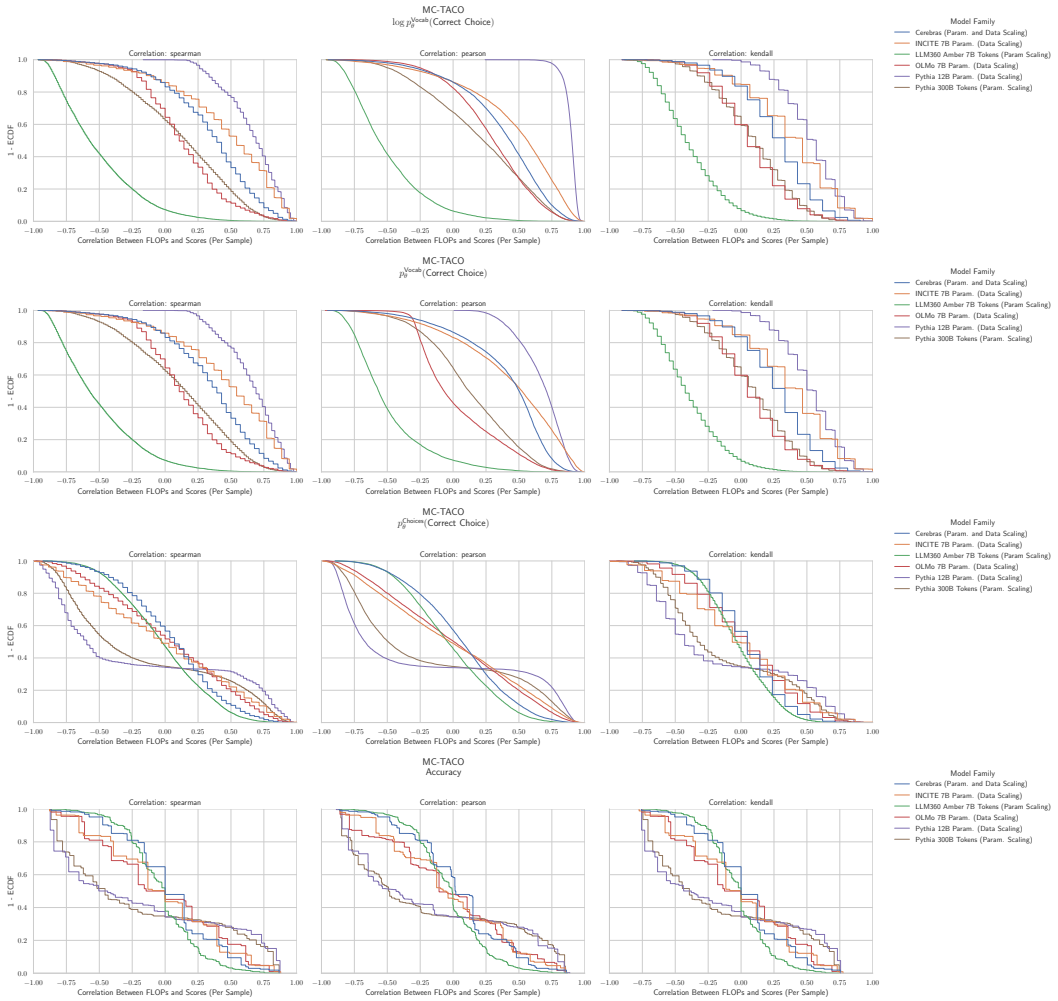

Figure 14: **MC TACO: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.6 NLP BENCHMARK: MMLU ABSTRACT ALGEBRA HENDRYCKS ET AL. (2020)

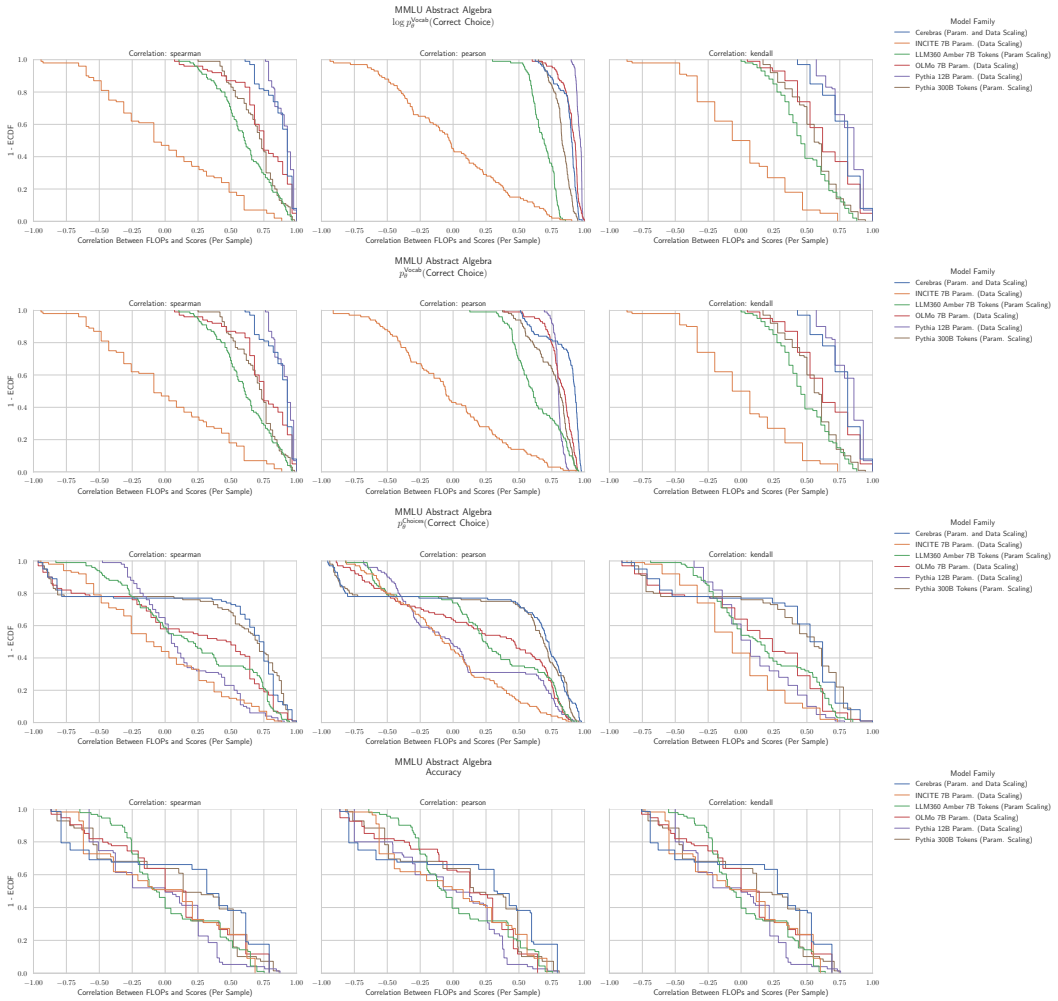

Figure 15: **MMLU Abstract Algebra: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.7 NLP BENCHMARK: MMLU ANATOMY HENDRYCKS ET AL. (2020)

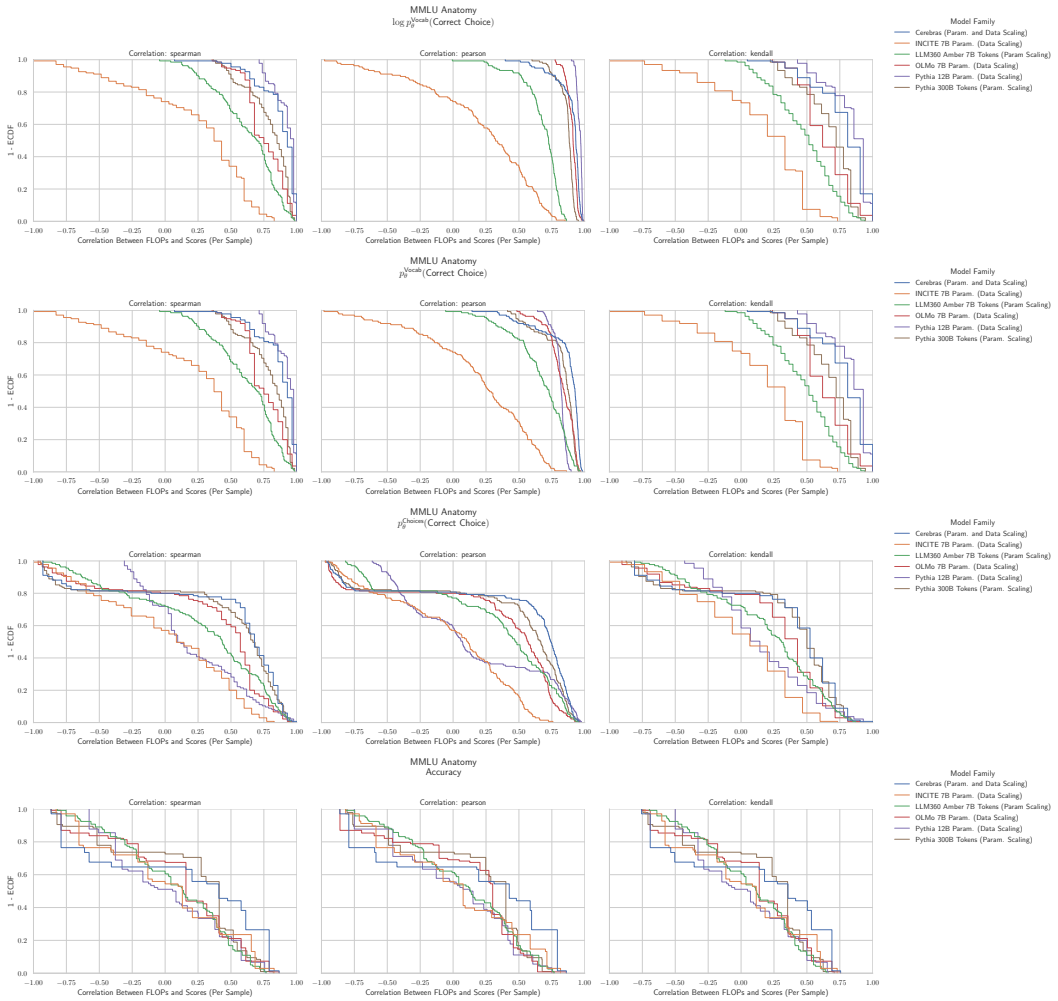

Figure 16: **MMLU Anatomy: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.8 NLP BENCHMARK: MMLU ASTRONOMY HENDRYCKS ET AL. (2020)

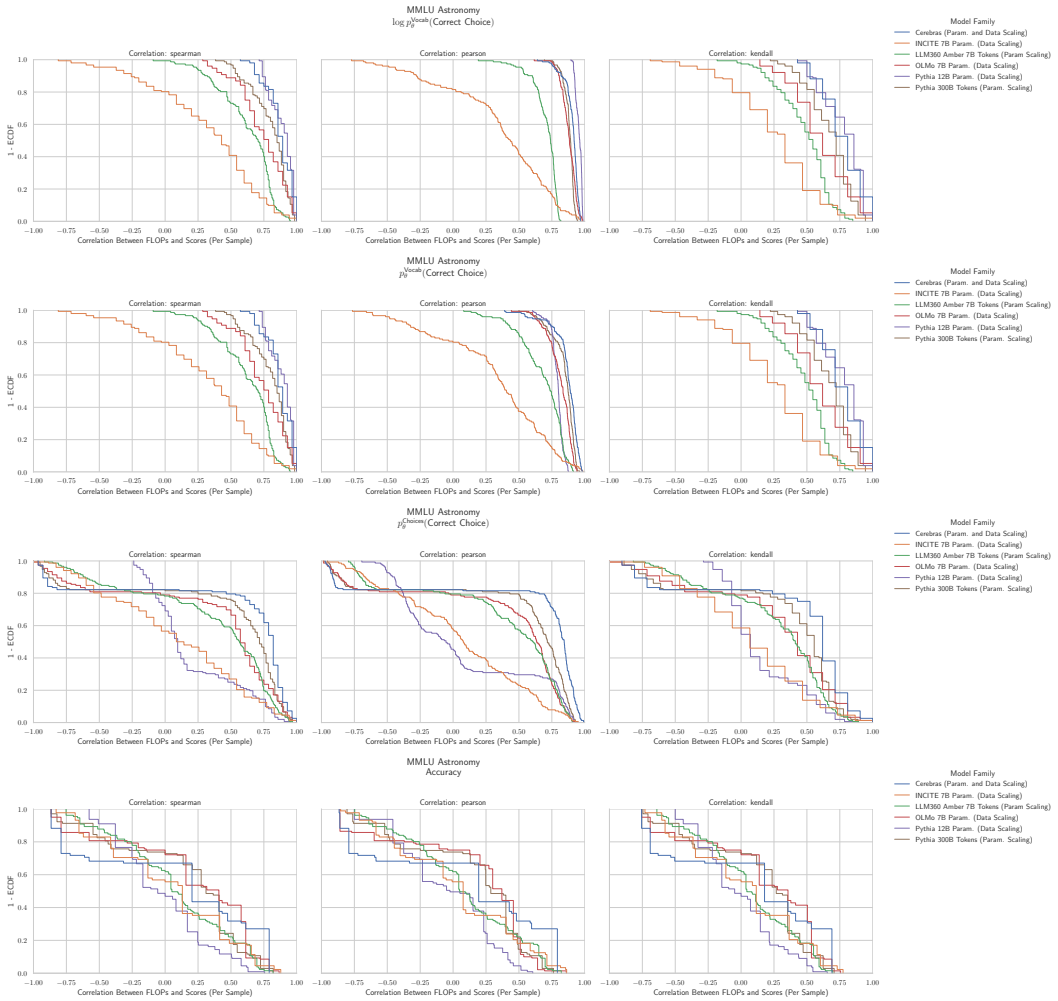

Figure 17: **MMLU Astronomy: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.9 NLP BENCHMARK: MMLU BUSINESS ETHICS HENDRYCKS ET AL. (2020)

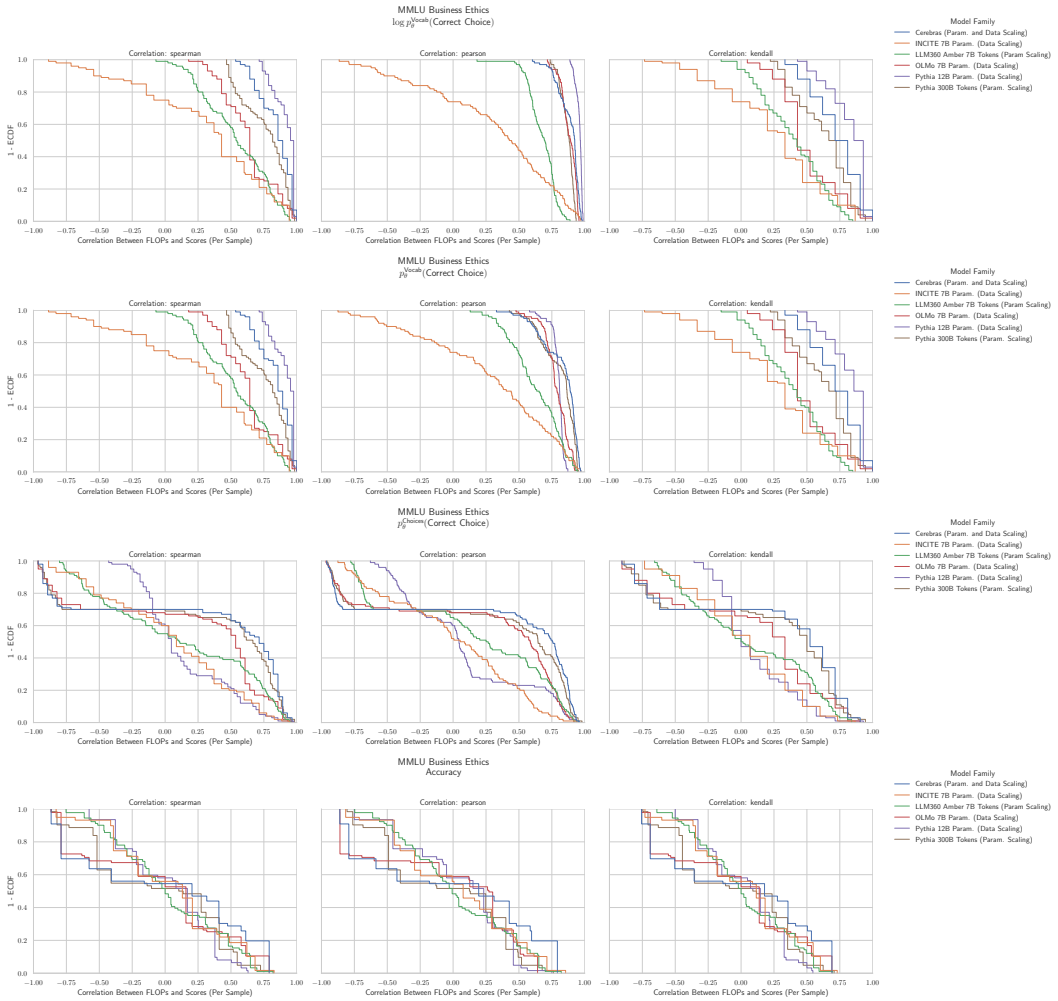

Figure 18: **MMLU Business Ethics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.10 NLP BENCHMARK: MMLU CLINICAL KNOWLEDGE HENDRYCKS ET AL. (2020)

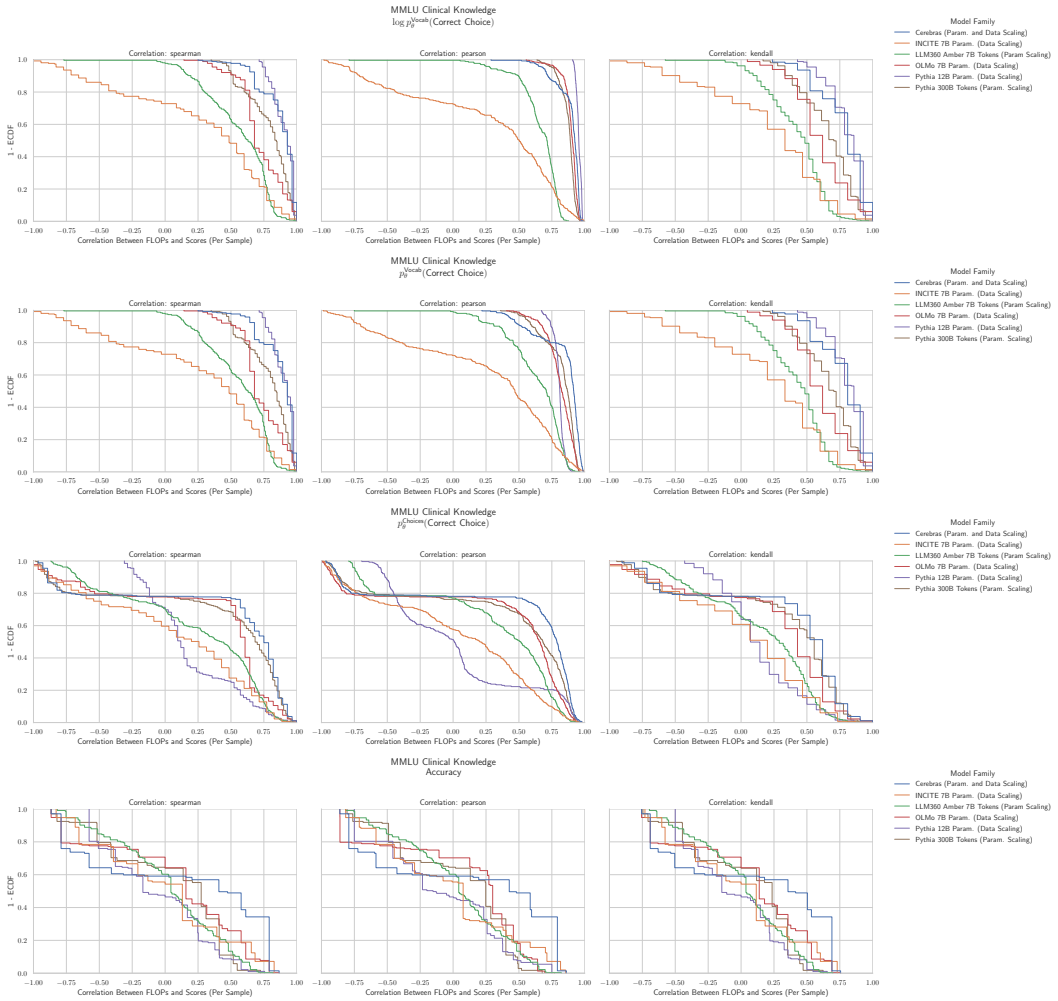

Figure 19: **MMLU Clinical Knowledge: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.11 NLP BENCHMARK: MMLU COLLEGE BIOLOGY HENDRYCKS ET AL. (2020)

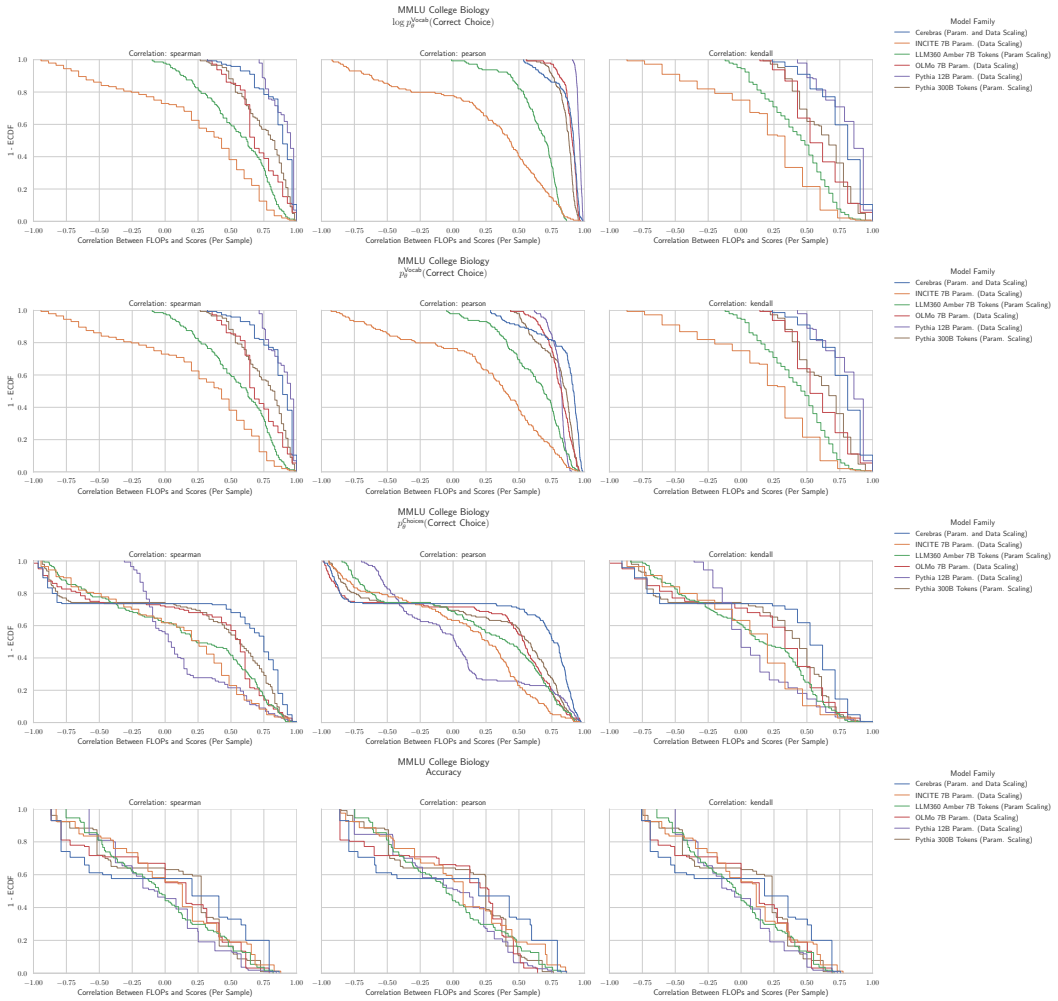

Figure 20: **MMLU College Biology: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.12 NLP BENCHMARK: MMLU COLLEGE CHEMISTRY HENDRYCKS ET AL. (2020)

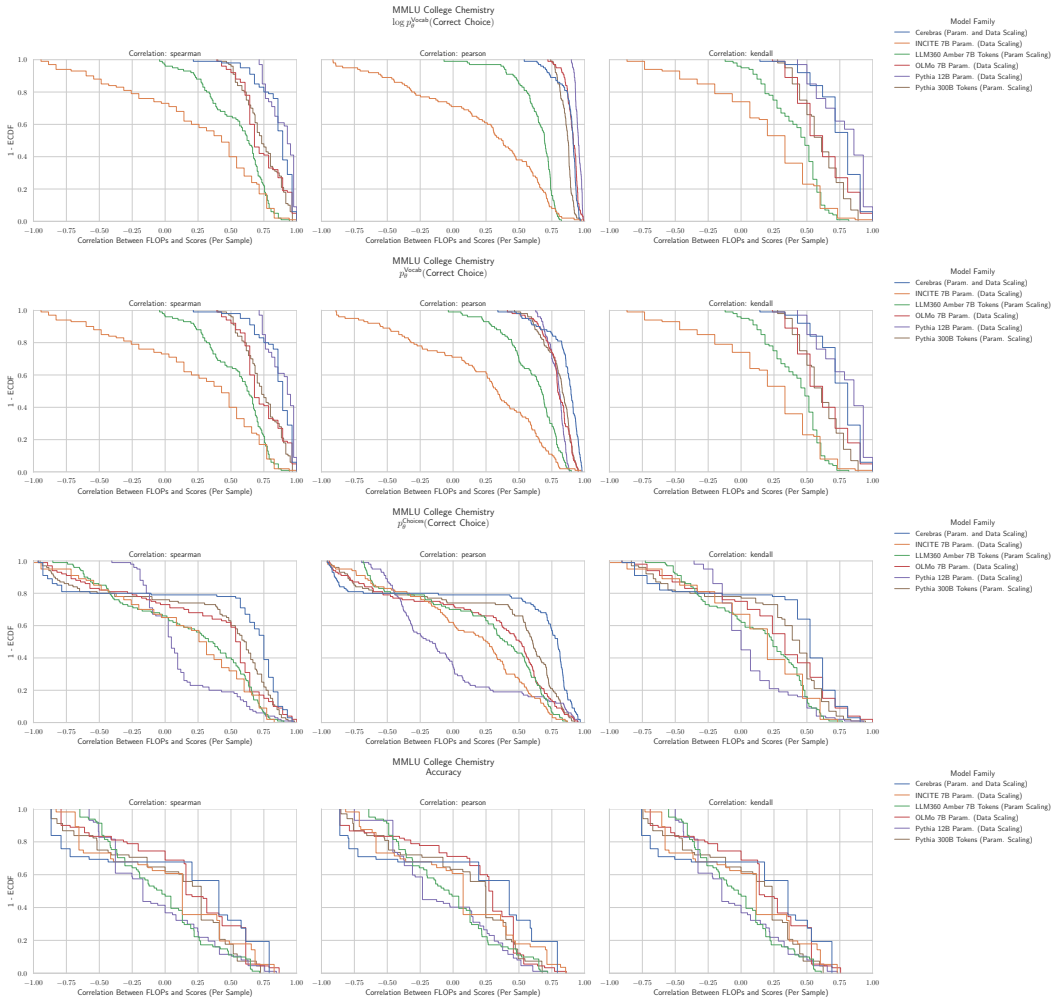

Figure 21: **MMLU College Chemistry: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.13 NLP BENCHMARK: MMLU COLLEGE COMPUTER SCIENCE HENDRYCKS ET AL. (2020)

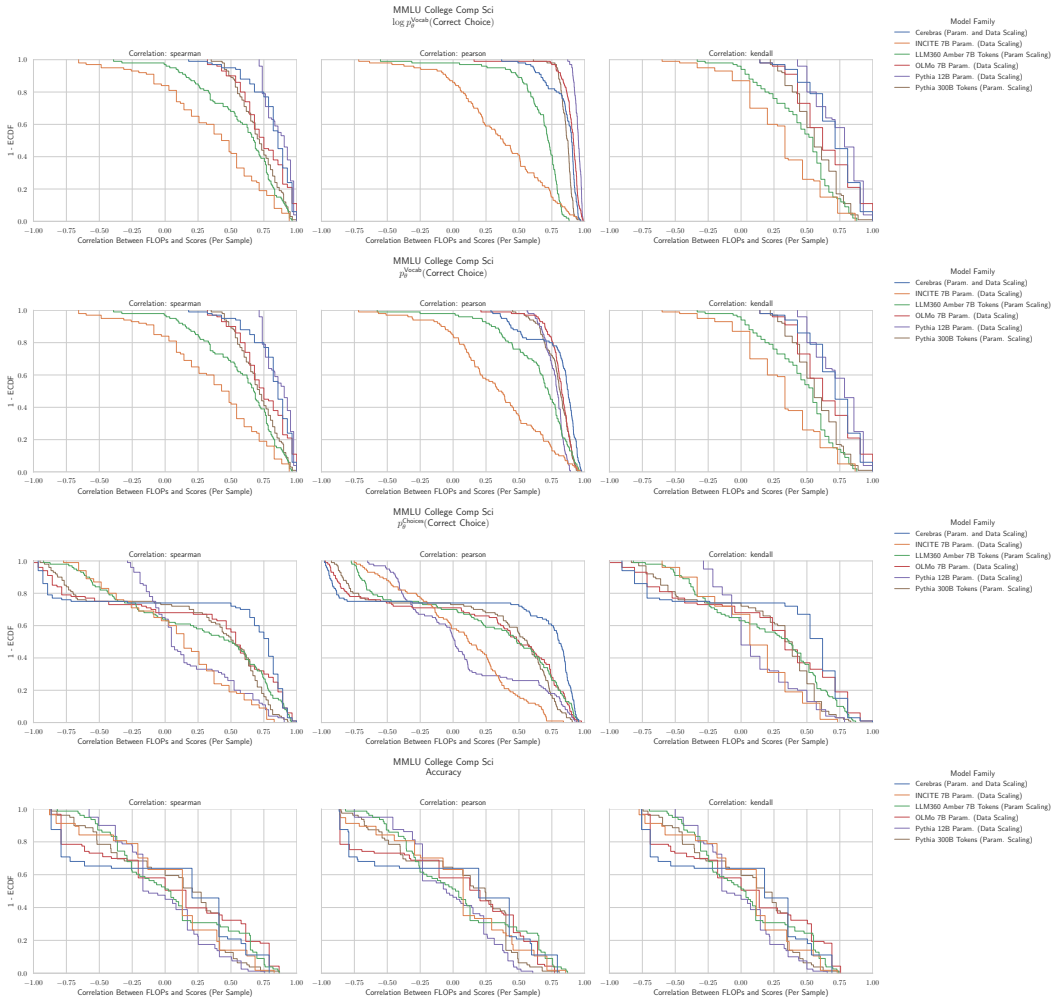

Figure 22: **MMLU College Computer Science: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.14 NLP BENCHMARK: MMLU COLLEGE MATHEMATICS HENDRYCKS ET AL. (2020)

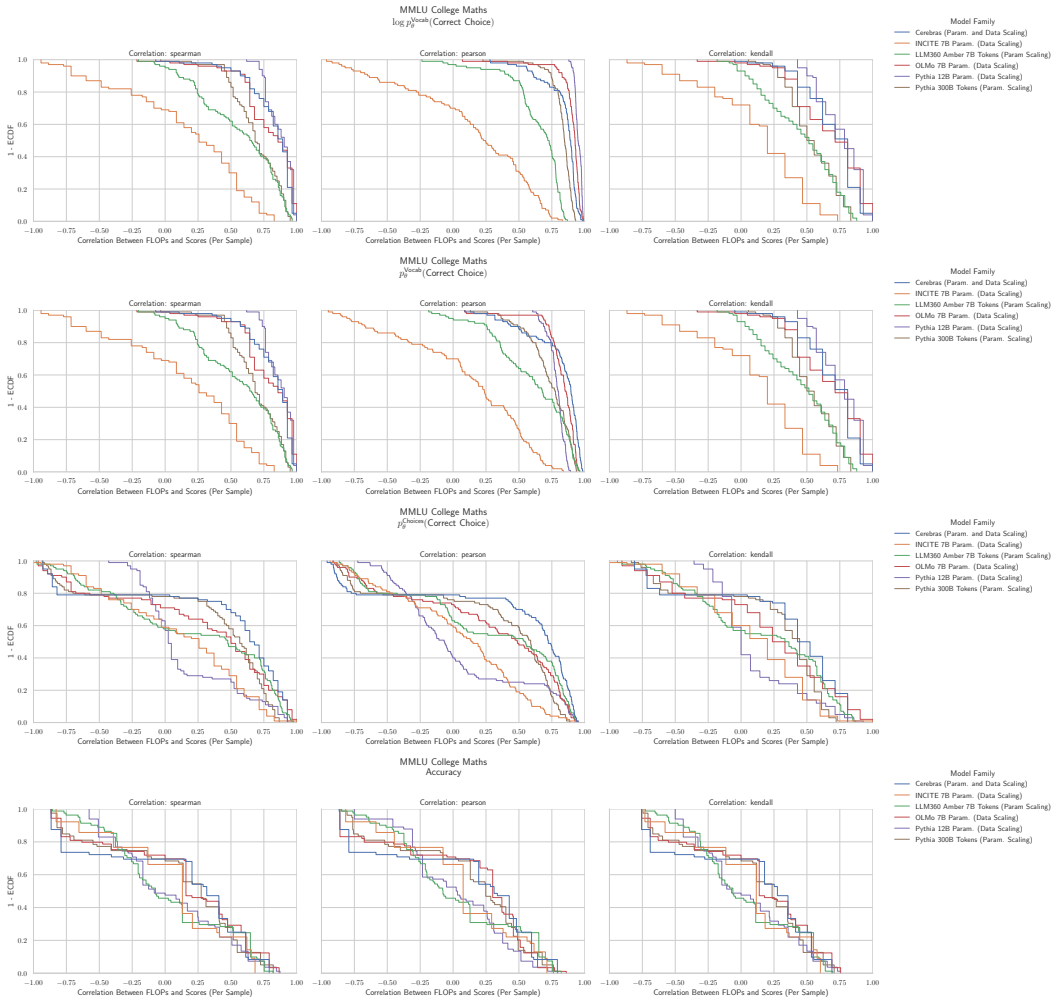

Figure 23: **MMLU College Mathematics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.15 NLP BENCHMARK: MMLU COLLEGE MEDICINE HENDRYCKS ET AL. (2020)

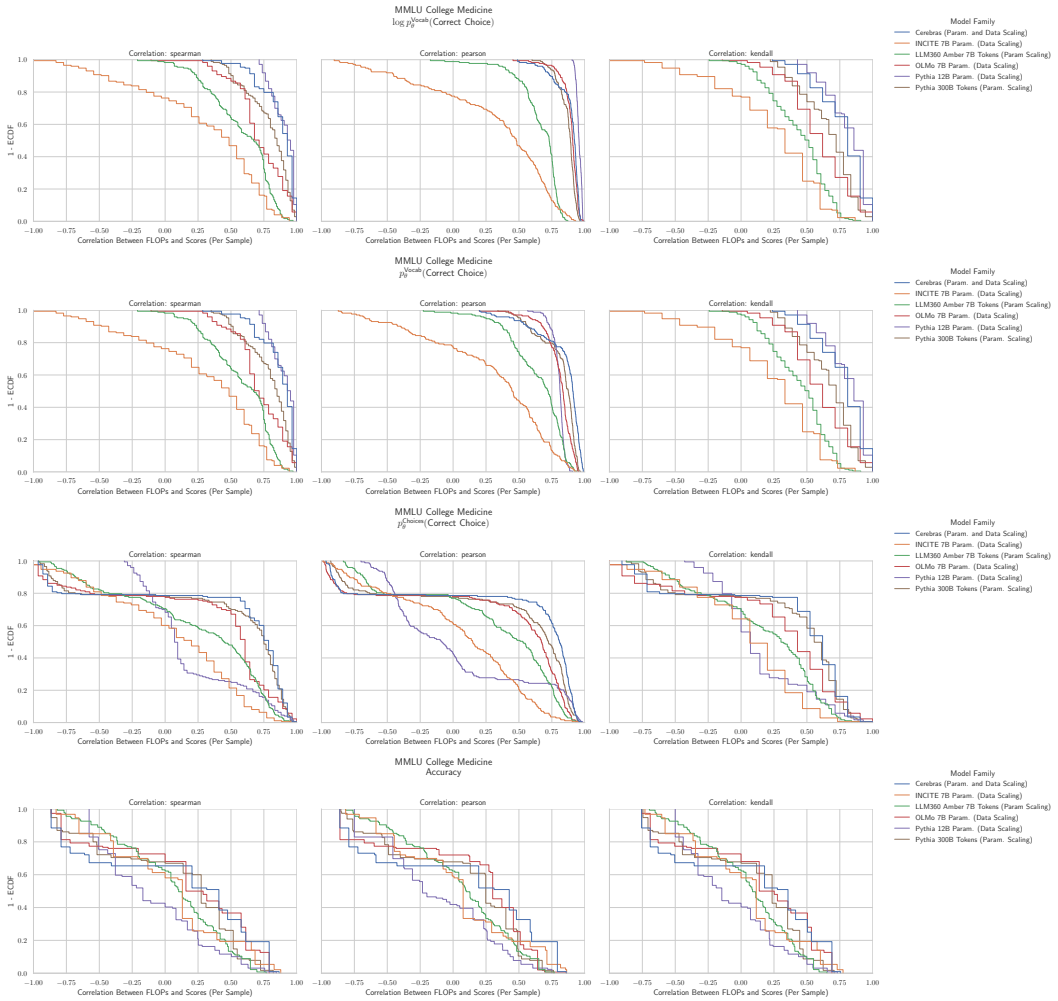

Figure 24: **MMLU College Medicine: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.16 NLP BENCHMARK: MMLU COLLEGE PHYSICS HENDRYCKS ET AL. (2020)

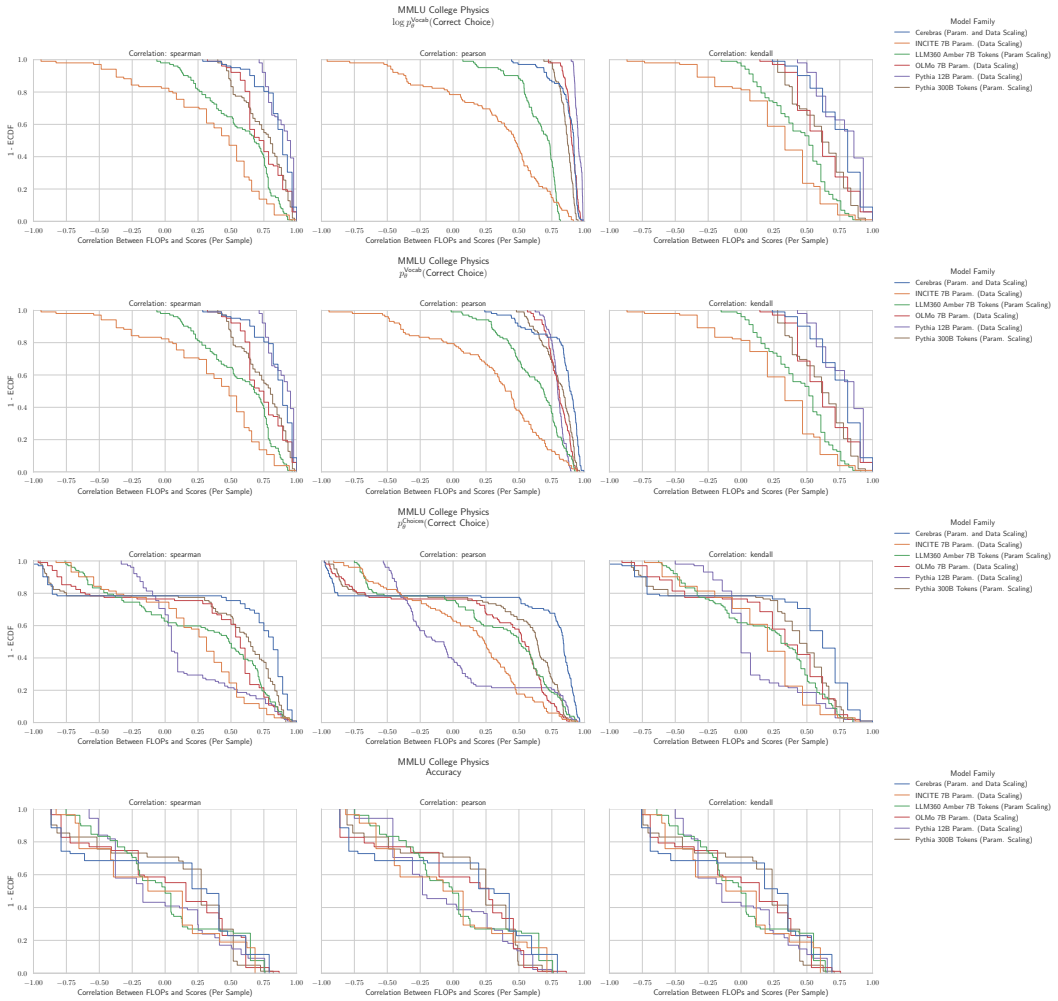

Figure 25: **MMLU College Physics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.17 NLP BENCHMARK: MMLU COMPUTER SECURITY HENDRYCKS ET AL. (2020)

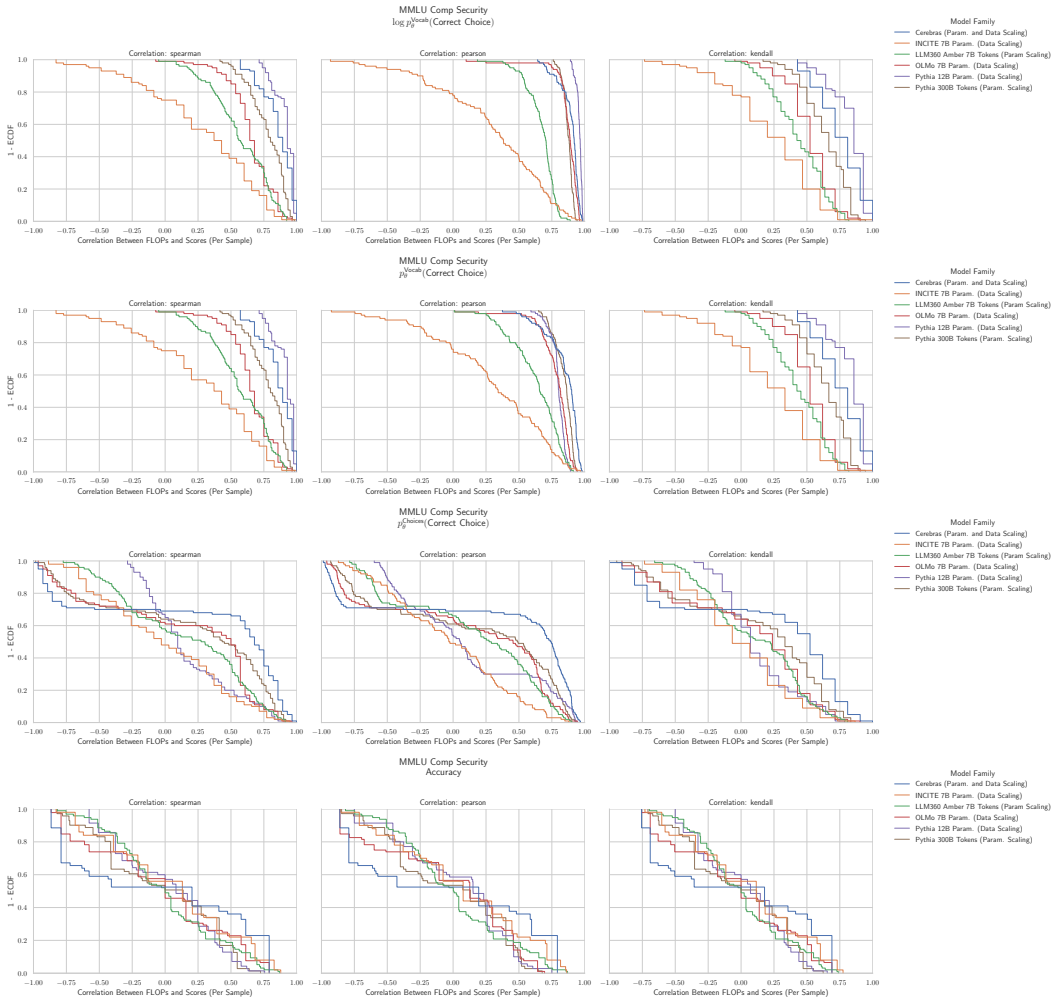

Figure 26: **MMLU Computer Security: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.18 NLP BENCHMARK: MMLU CONCEPTUAL PHYSICS HENDRYCKS ET AL. (2020)

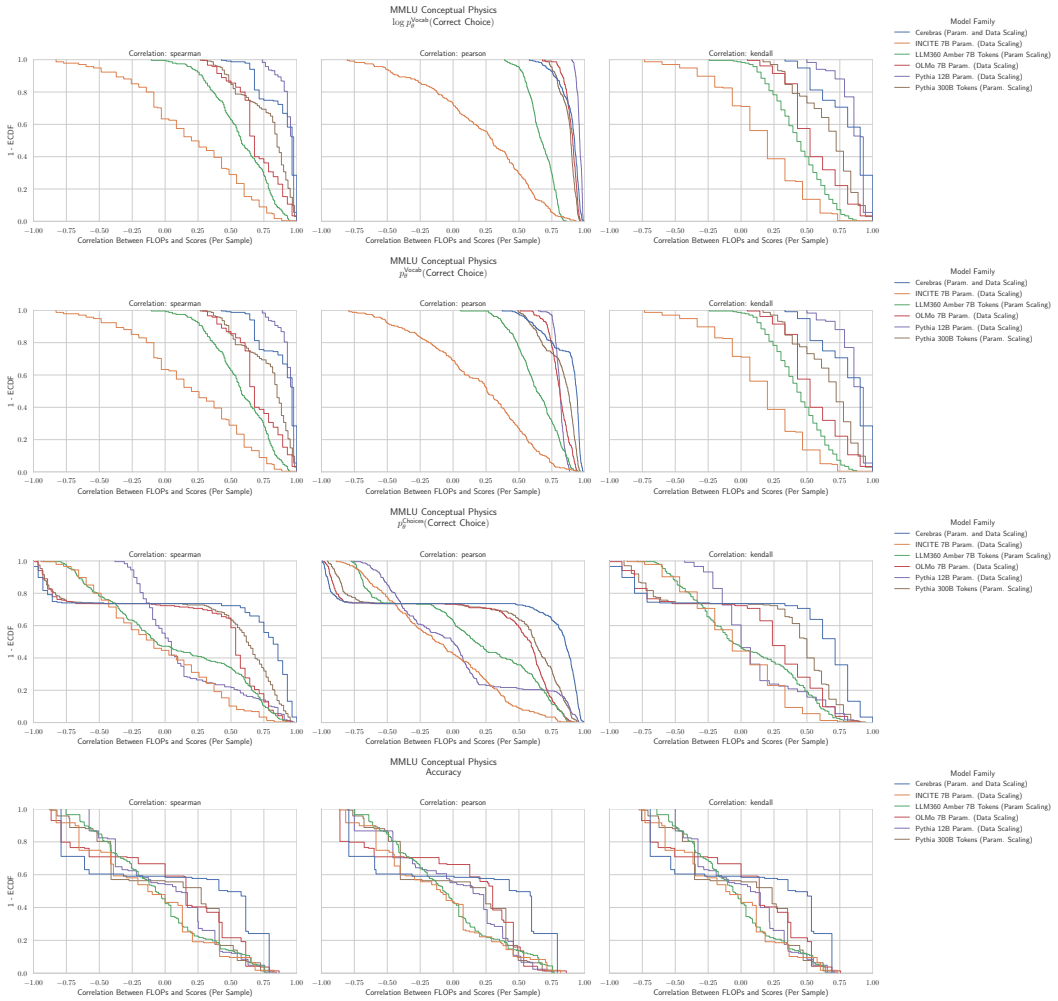

Figure 27: **MMLU Conceptual Physics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.19 NLP BENCHMARK: MMLU ECONOMETRICS HENDRYCKS ET AL. (2020)

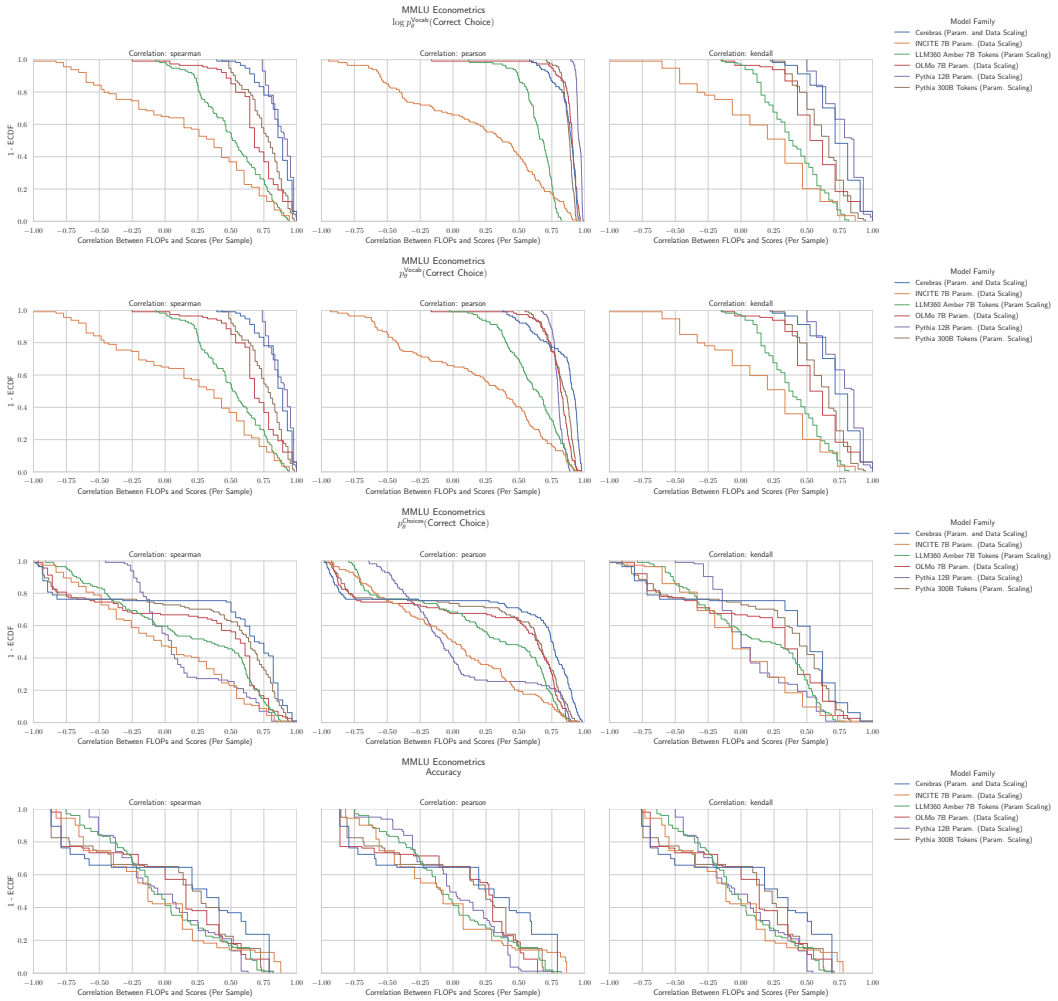

Figure 28: **MMLU Econometrics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.20 NLP BENCHMARK: MMLU ELECTRICAL ENGINEERING HENDRYCKS ET AL. (2020)

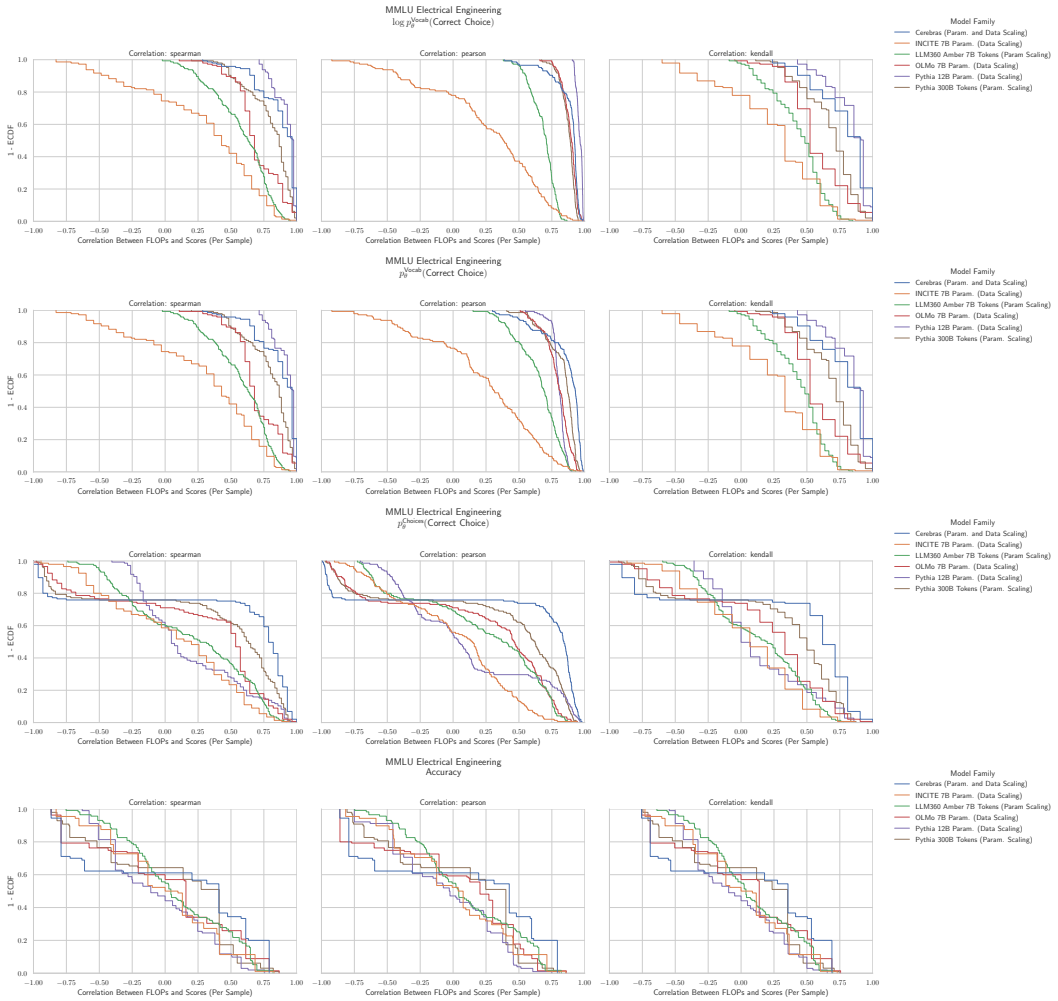

Figure 29: **MMLU Electrical Engineering: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.21 NLP BENCHMARK: MMLU ELEMENTARY MATHEMATICS HENDRYCKS ET AL. (2020)

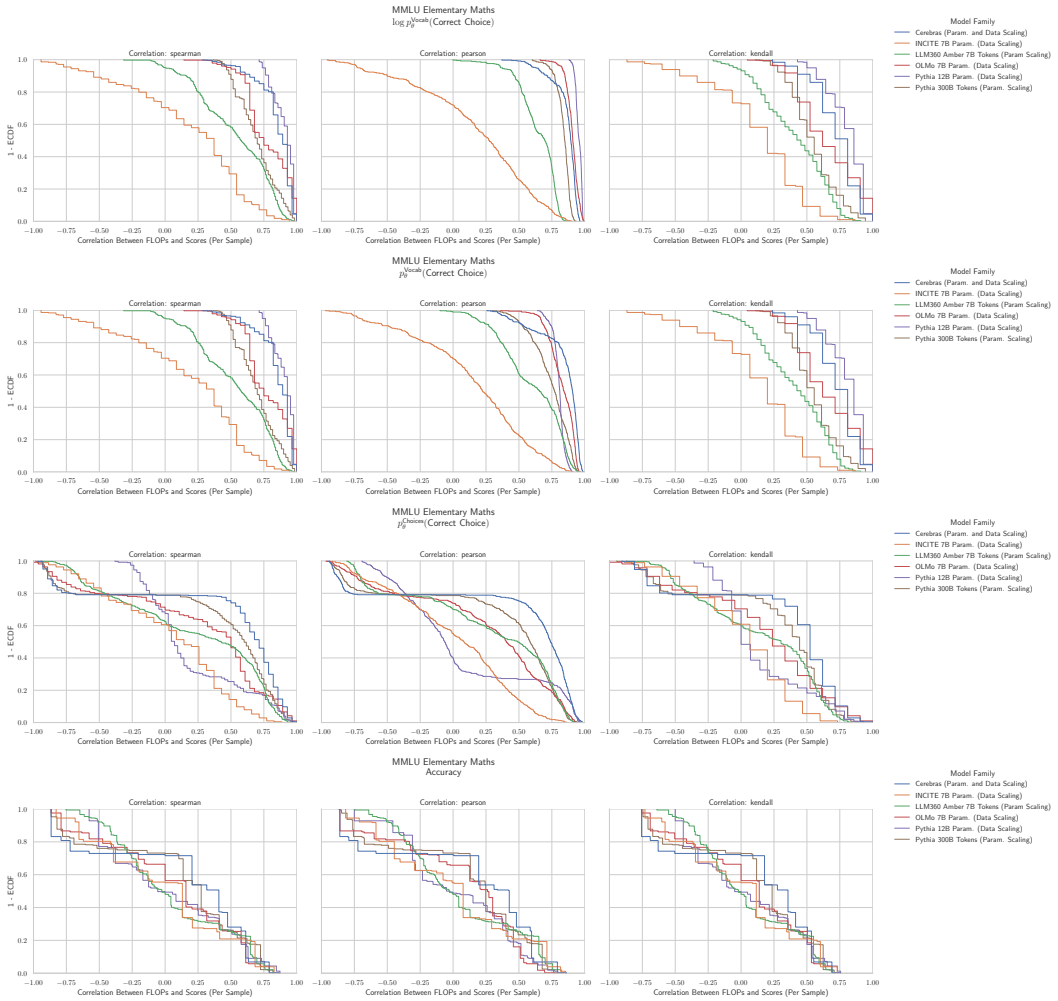

Figure 30: **MMLU Elementary Mathematics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.22 NLP BENCHMARK: MMLU FORMAL LOGIC HENDRYCKS ET AL. (2020)

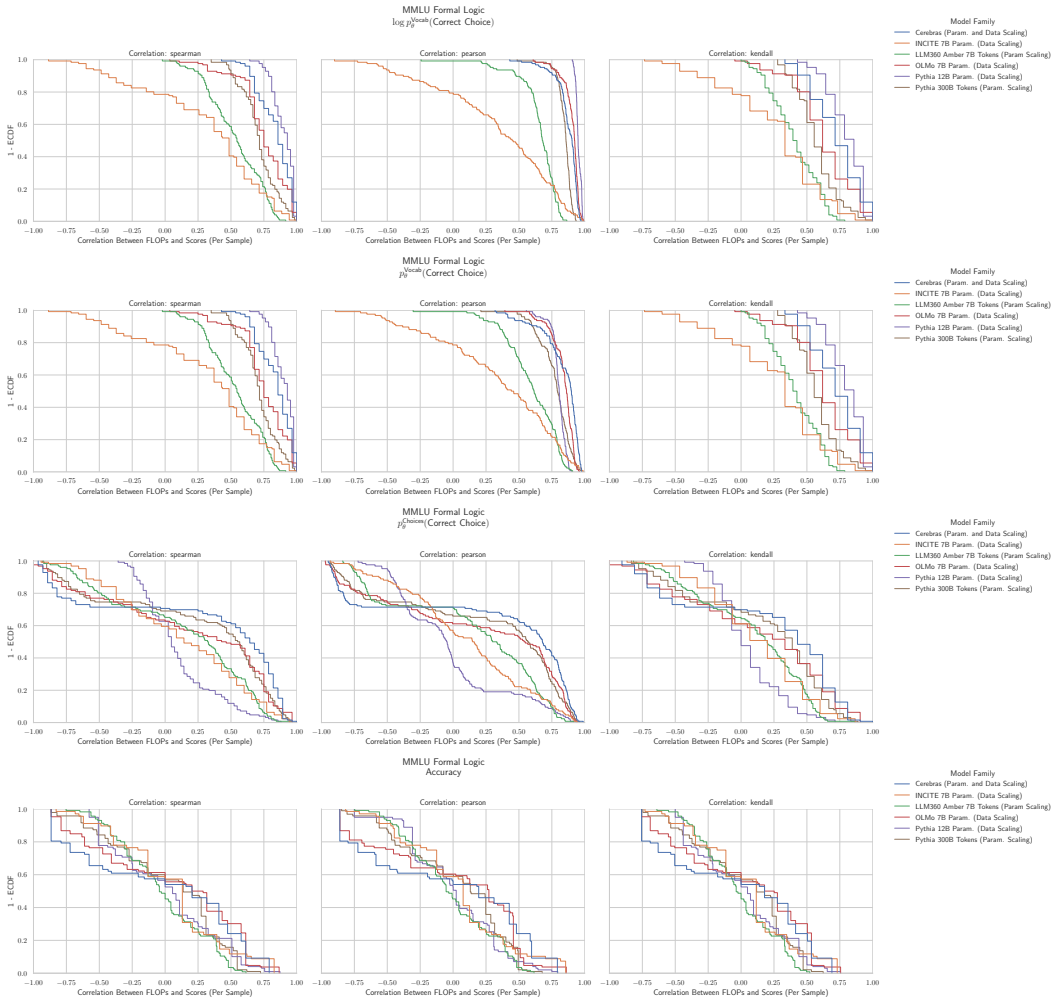

Figure 31: **MMLU Formal Logic: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.23 NLP BENCHMARK: MMLU GLOBAL FACTS HENDRYCKS ET AL. (2020)

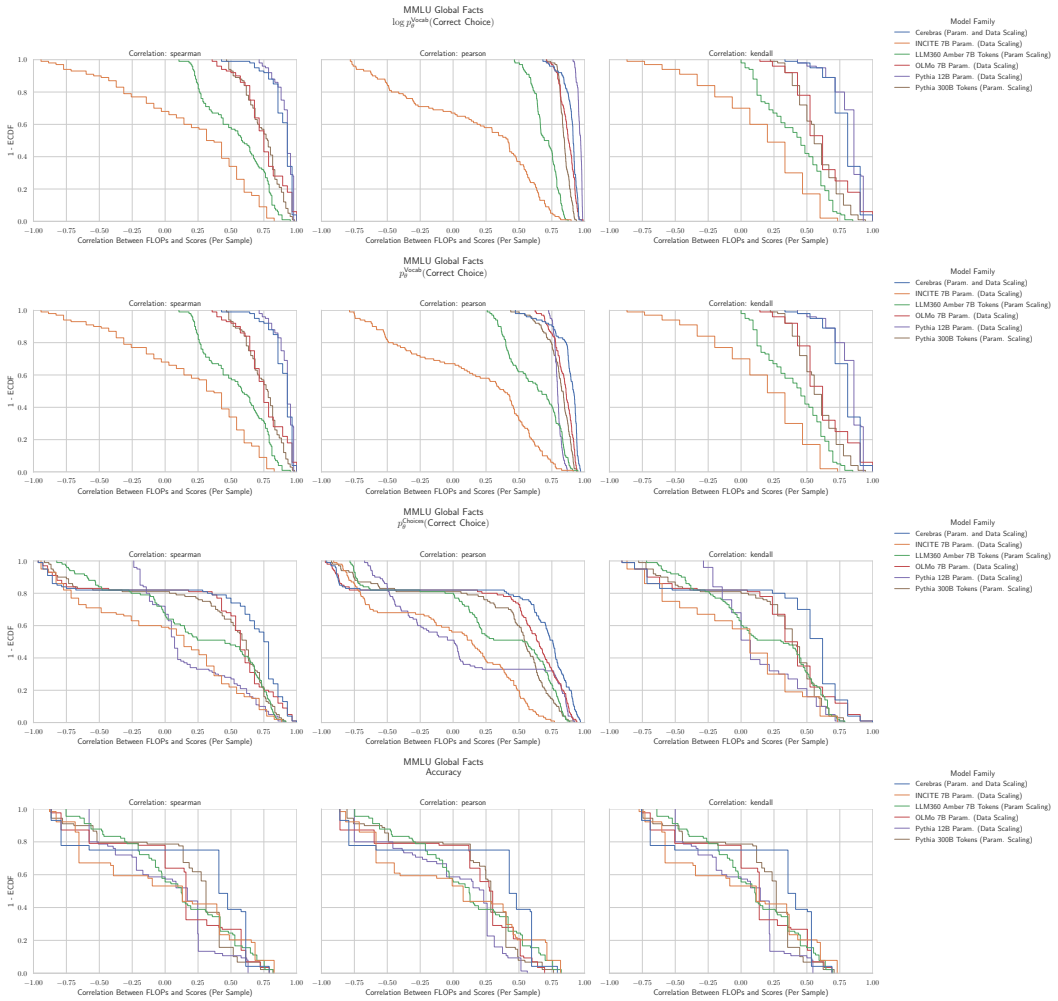

Figure 32: **MMLU Global Facts: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.24 NLP BENCHMARK: MMLU HIGH SCHOOL BIOLOGY HENDRYCKS ET AL. (2020)

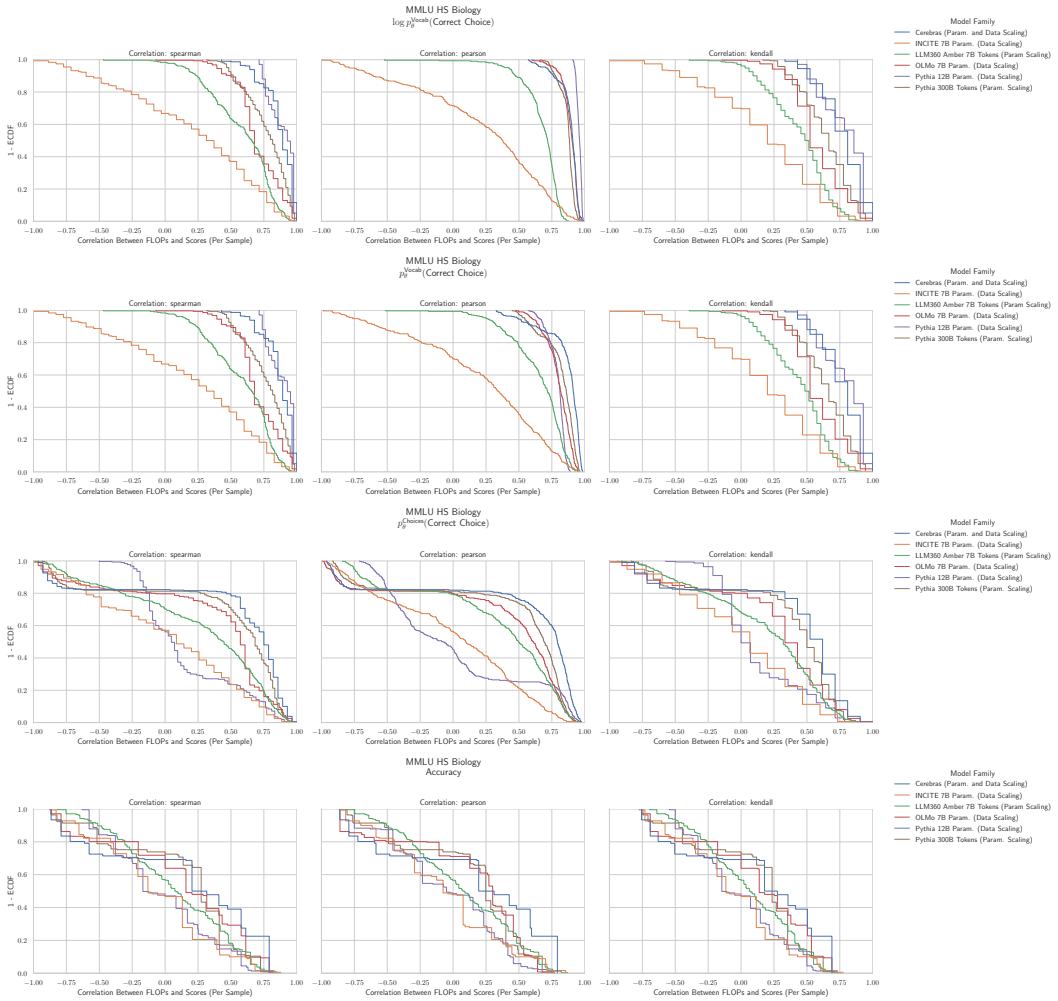

Figure 33: **MMLU High School Biology: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.25 NLP BENCHMARK: MMLU HIGH SCHOOL CHEMISTRY HENDRYCKS ET AL. (2020)

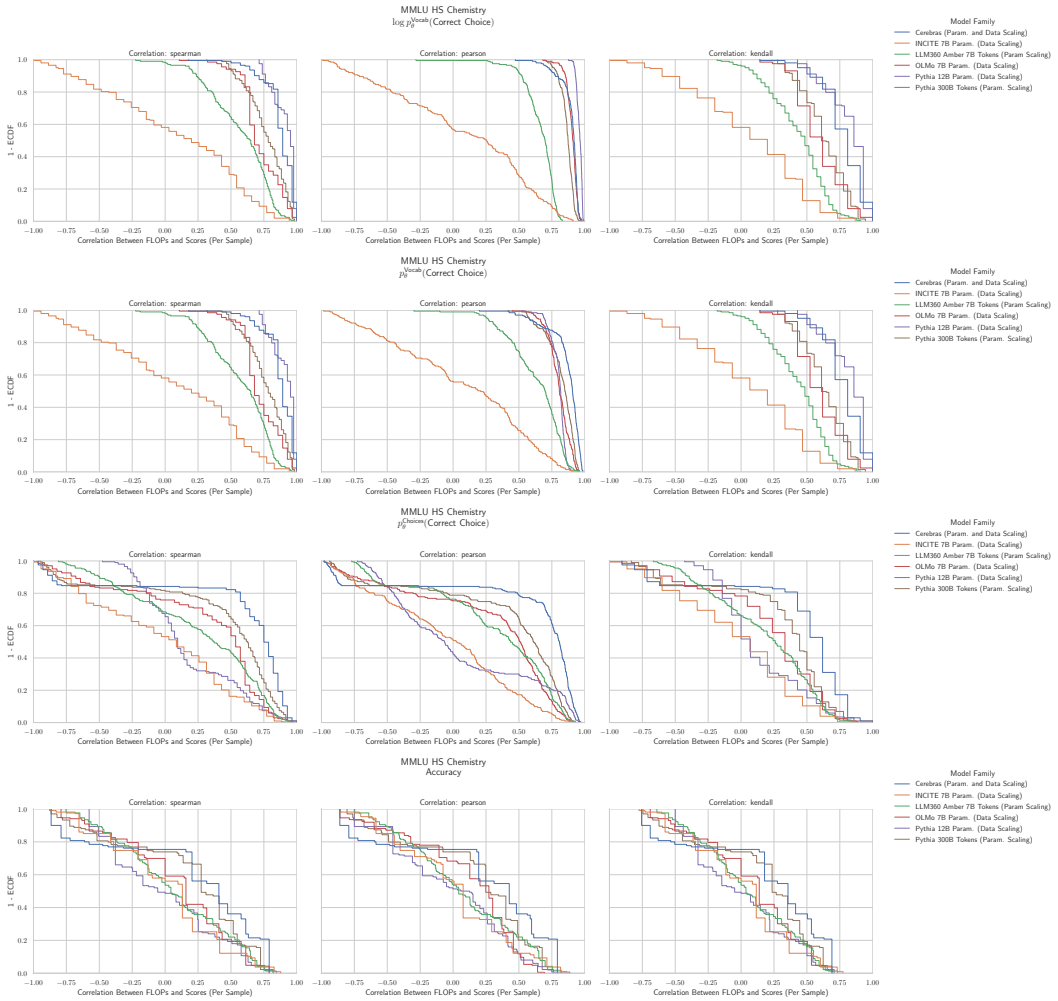

Figure 34: **MMLU High School Chemistry: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.26 NLP BENCHMARK: MMLU HIGH SCHOOL COMPUTER SCIENCE HENDRYCKS ET AL. (2020)

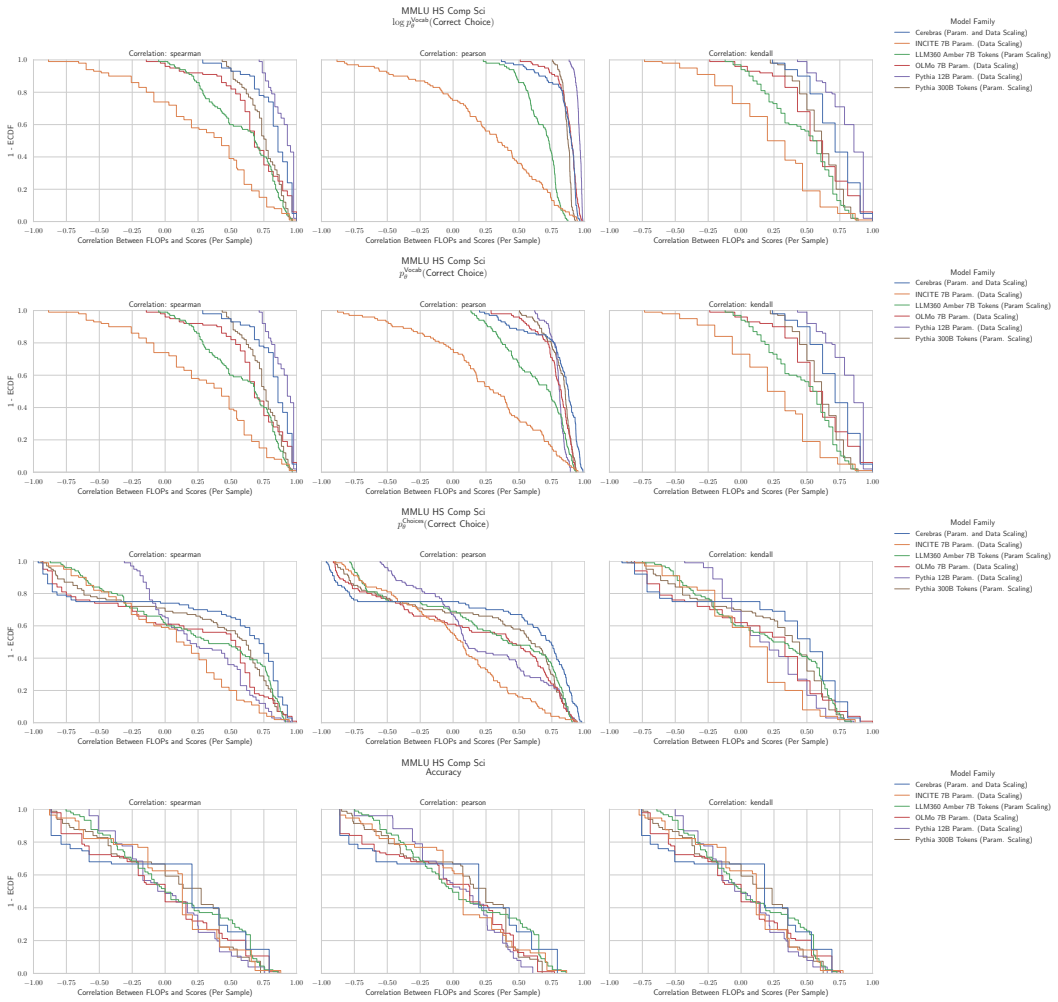

Figure 35: **MMLU High School Computer Science: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.27 NLP BENCHMARK: MMLU HIGH SCHOOL CHEMISTRY HENDRYCKS ET AL. (2020)

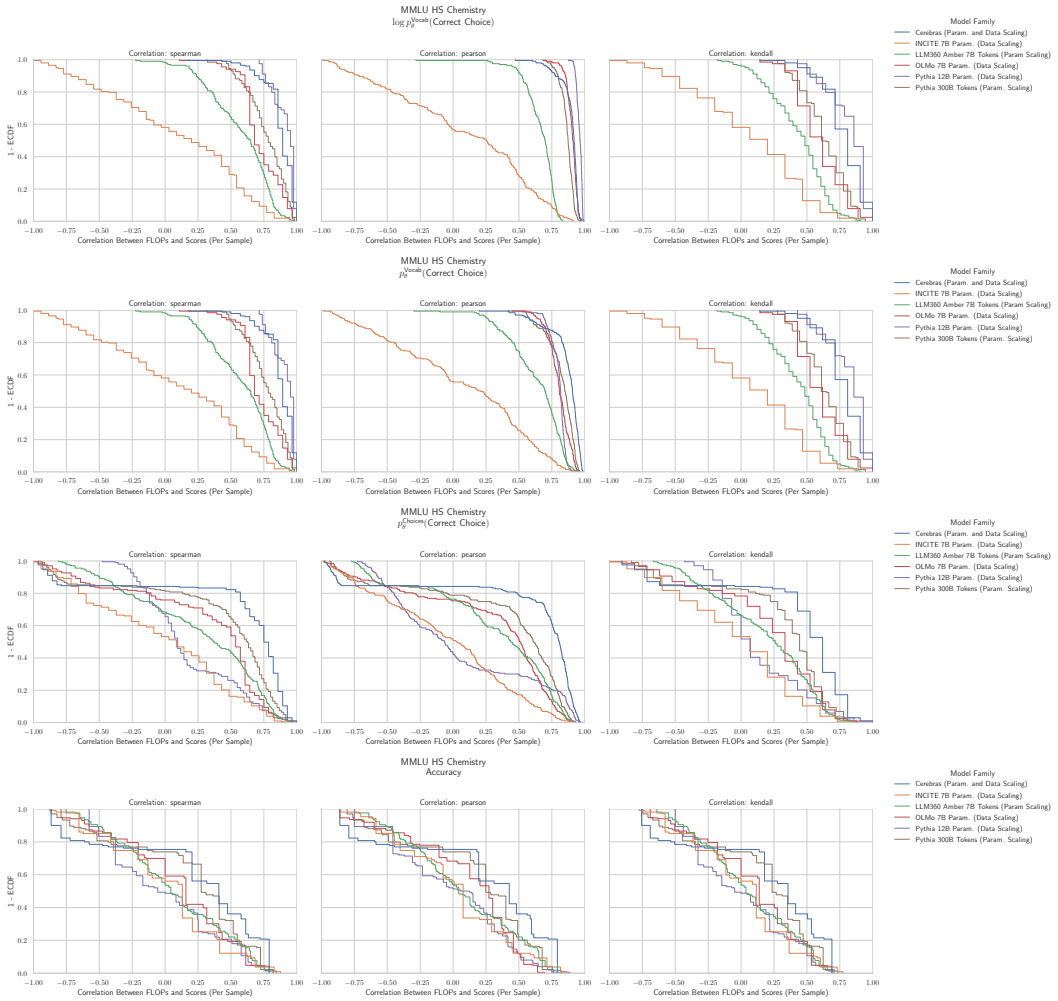

Figure 36: **MMLU High School Chemistry: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.28 NLP BENCHMARK: MMLU HIGH SCHOOL EUROPEAN HISTORY HENDRYCKS ET AL. (2020)

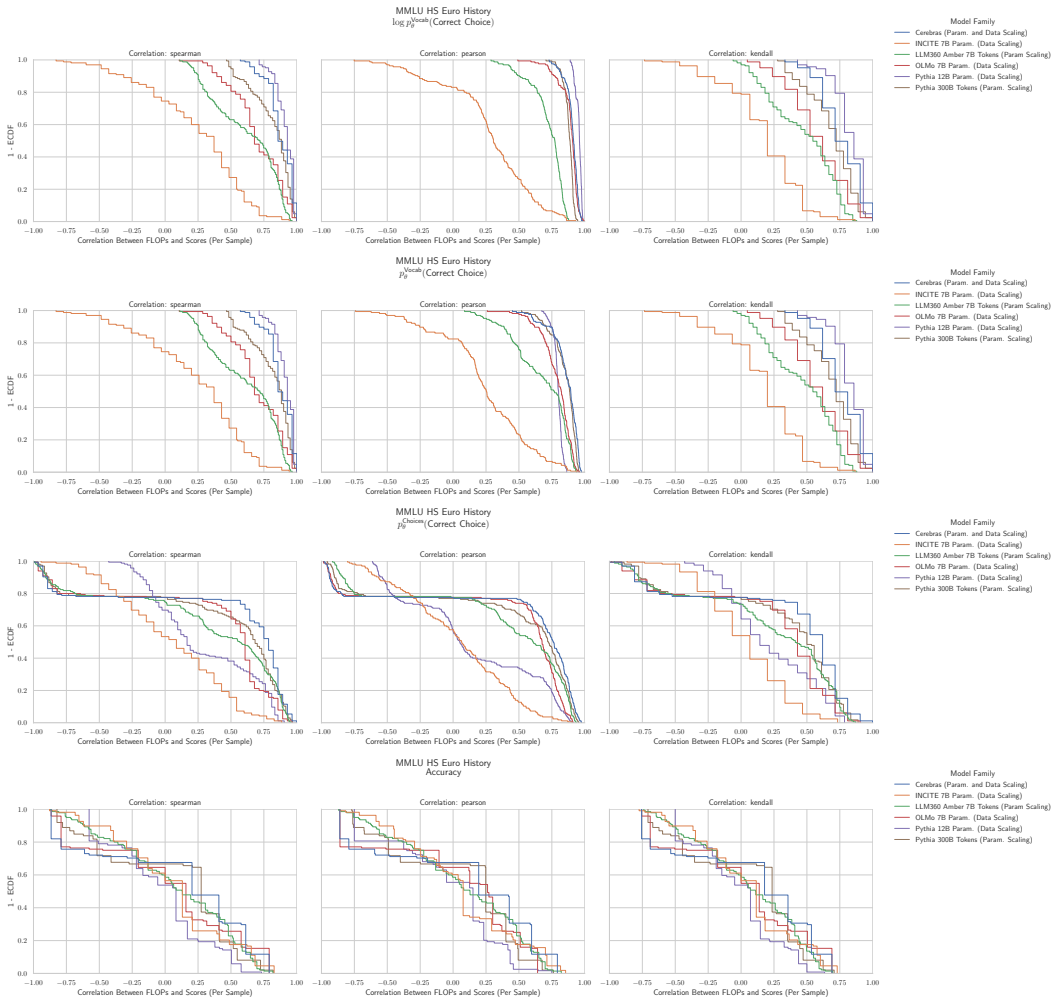

Figure 37: **MMLU High School European History: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.29 NLP BENCHMARK: MMLU HIGH SCHOOL GEOGRAPHY HENDRYCKS ET AL. (2020)

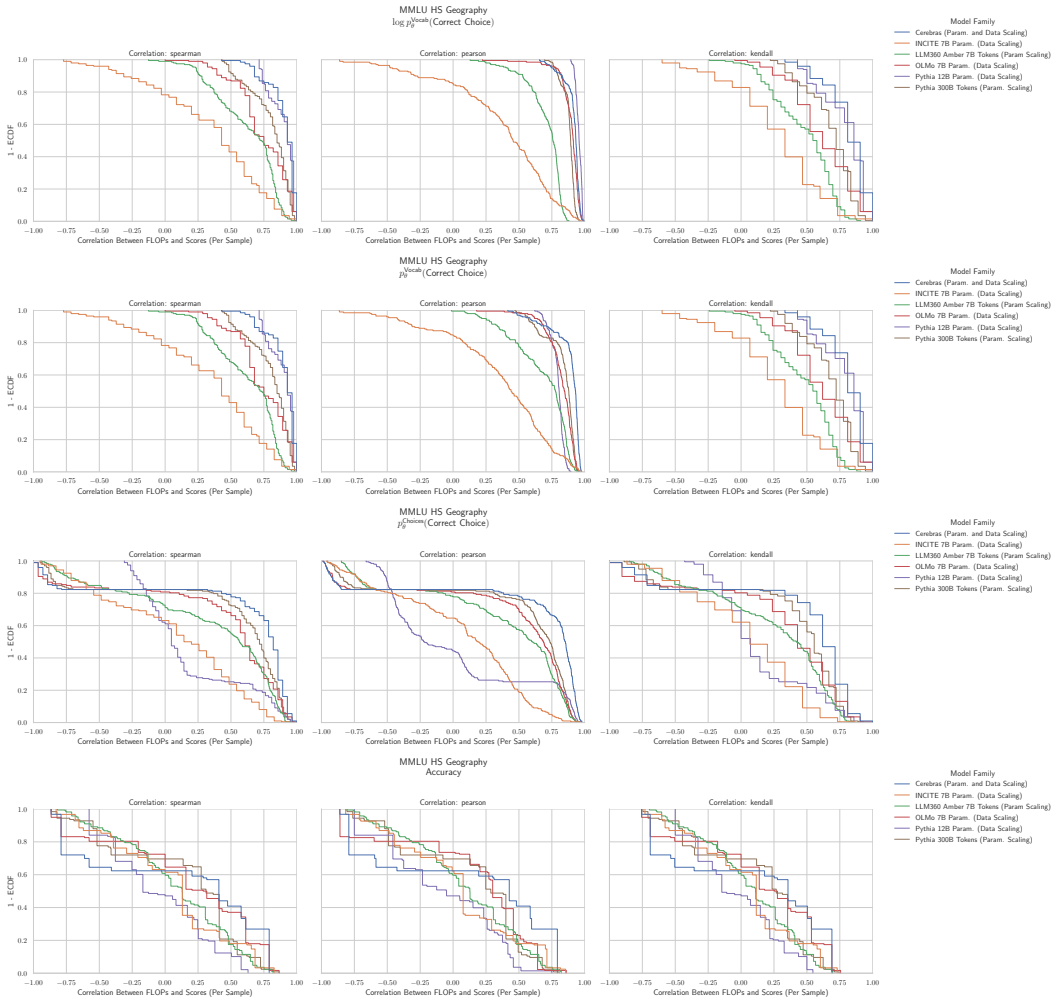

Figure 38: **MMLU High School Geography: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.30 NLP BENCHMARK: MMLU HIGH SCHOOL GOVERNMENT & POLITICS HENDRYCKS ET AL. (2020)

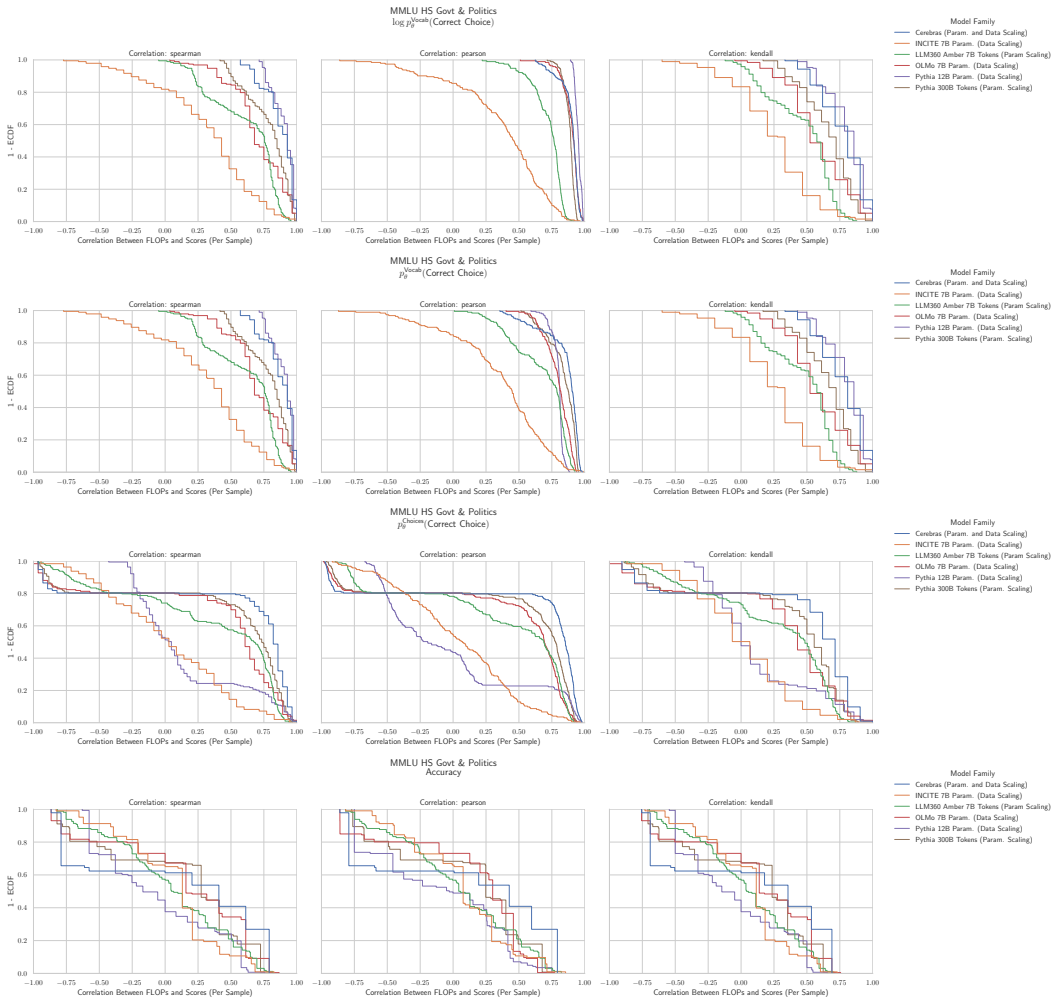

Figure 39: **MMLU High School Government & Politics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.31 NLP BENCHMARK: MMLU HIGH SCHOOL MACROECONOMICS HENDRYCKS ET AL. (2020)

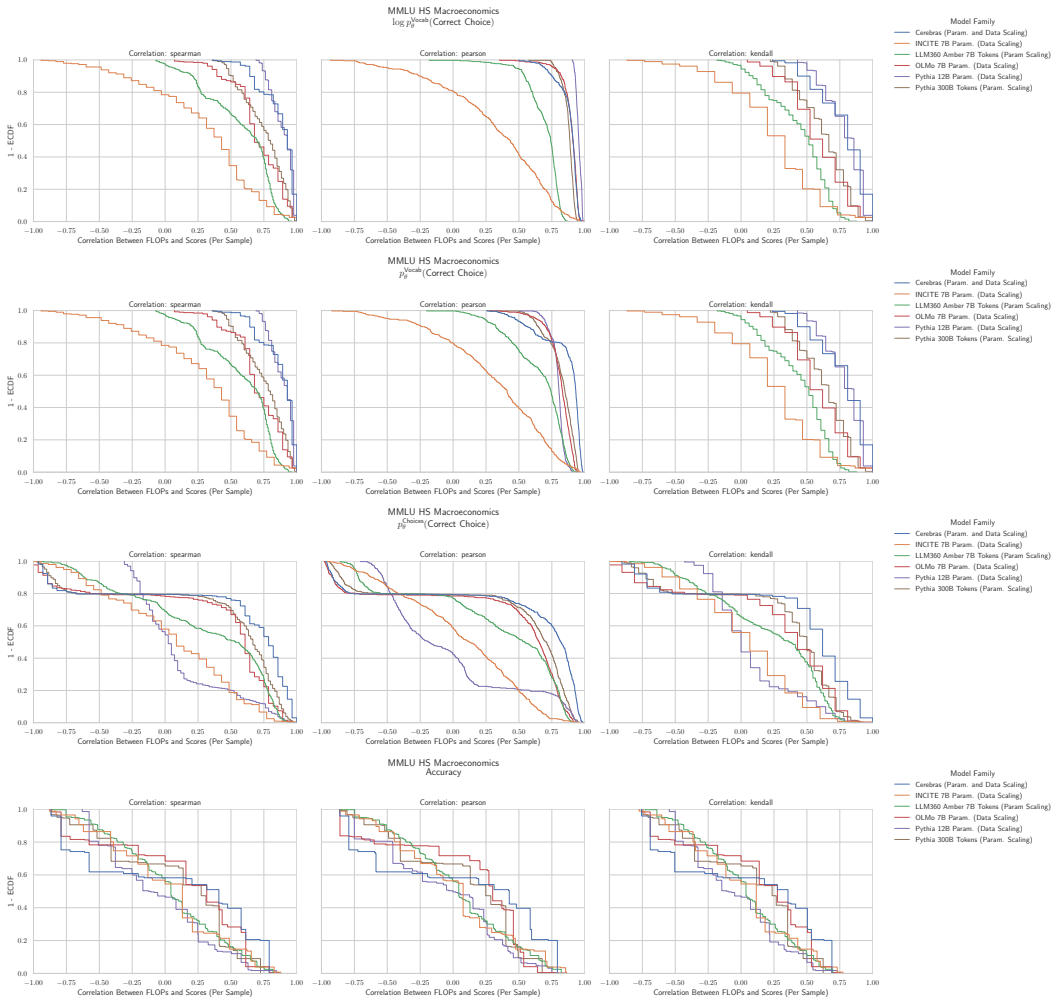

Figure 40: **MMLU High School Macroeconomics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.32 NLP BENCHMARK: MMLU HIGH SCHOOL MATHEMATICS HENDRYCKS ET AL. (2020)

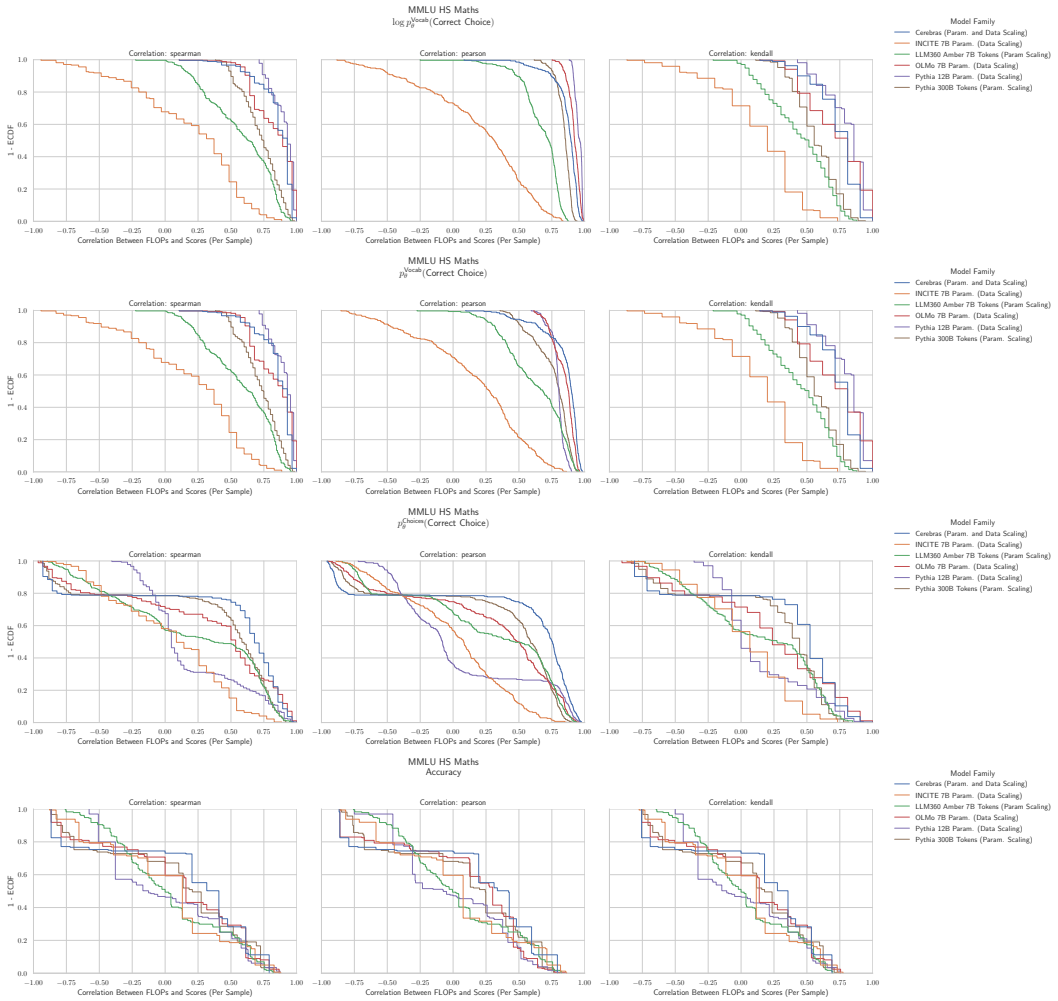

Figure 41: **MMLU High School Mathematics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.33 NLP BENCHMARK: MMLU HIGH SCHOOL MICROECONOMICS HENDRYCKS ET AL. (2020)

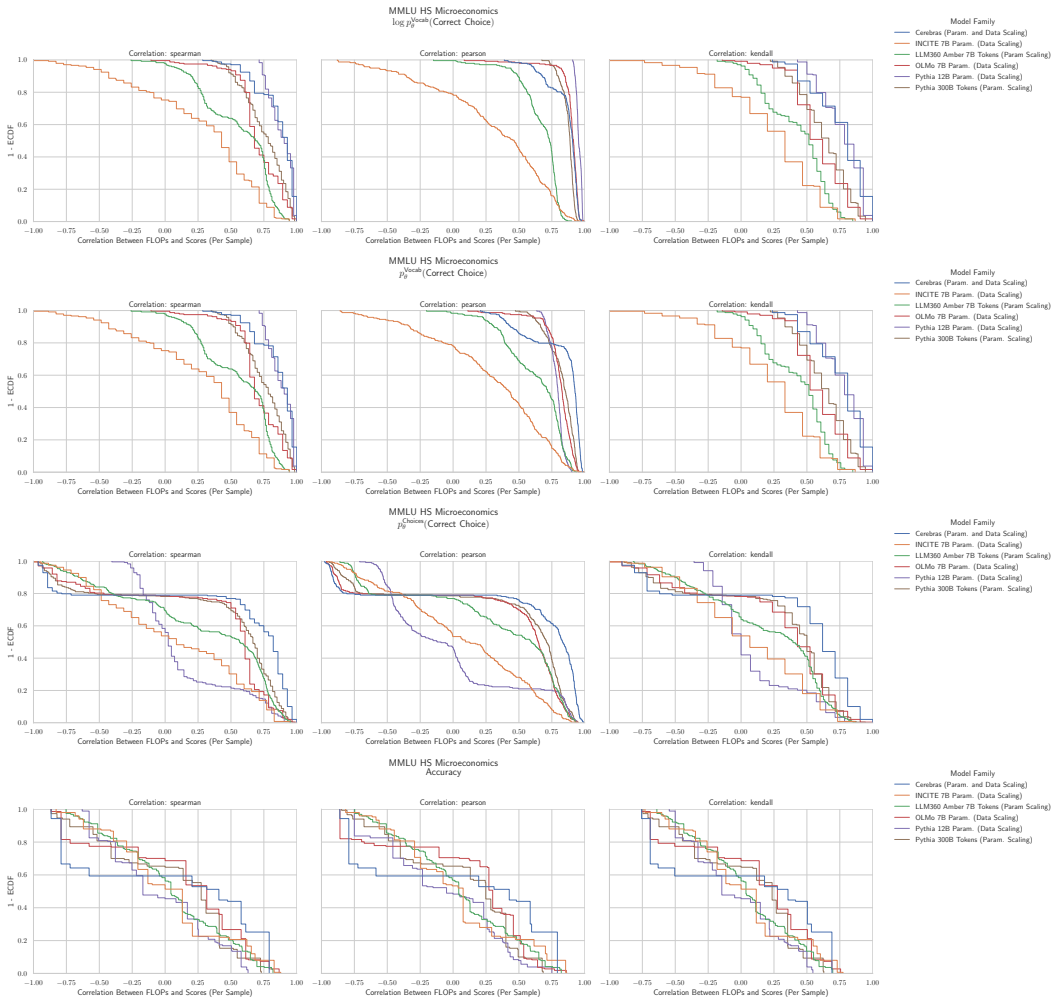

Figure 42: **MMLU High School Microeconomics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.34 NLP BENCHMARK: MMLU HIGH SCHOOL PHYSICS HENDRYCKS ET AL. (2020)

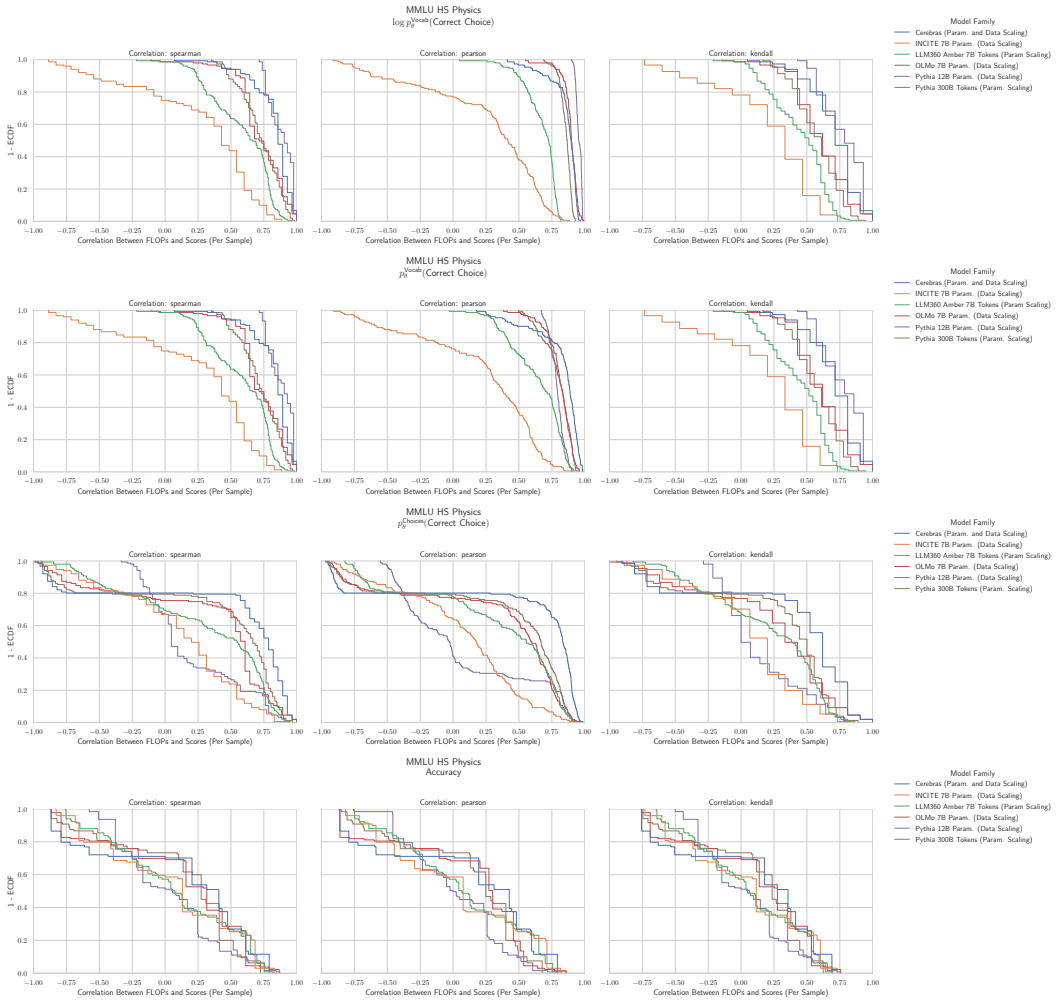

Figure 43: **MMLU High School Physics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.35 NLP BENCHMARK: MMLU HIGH SCHOOL PSYCHOLOGY HENDRYCKS ET AL. (2020)

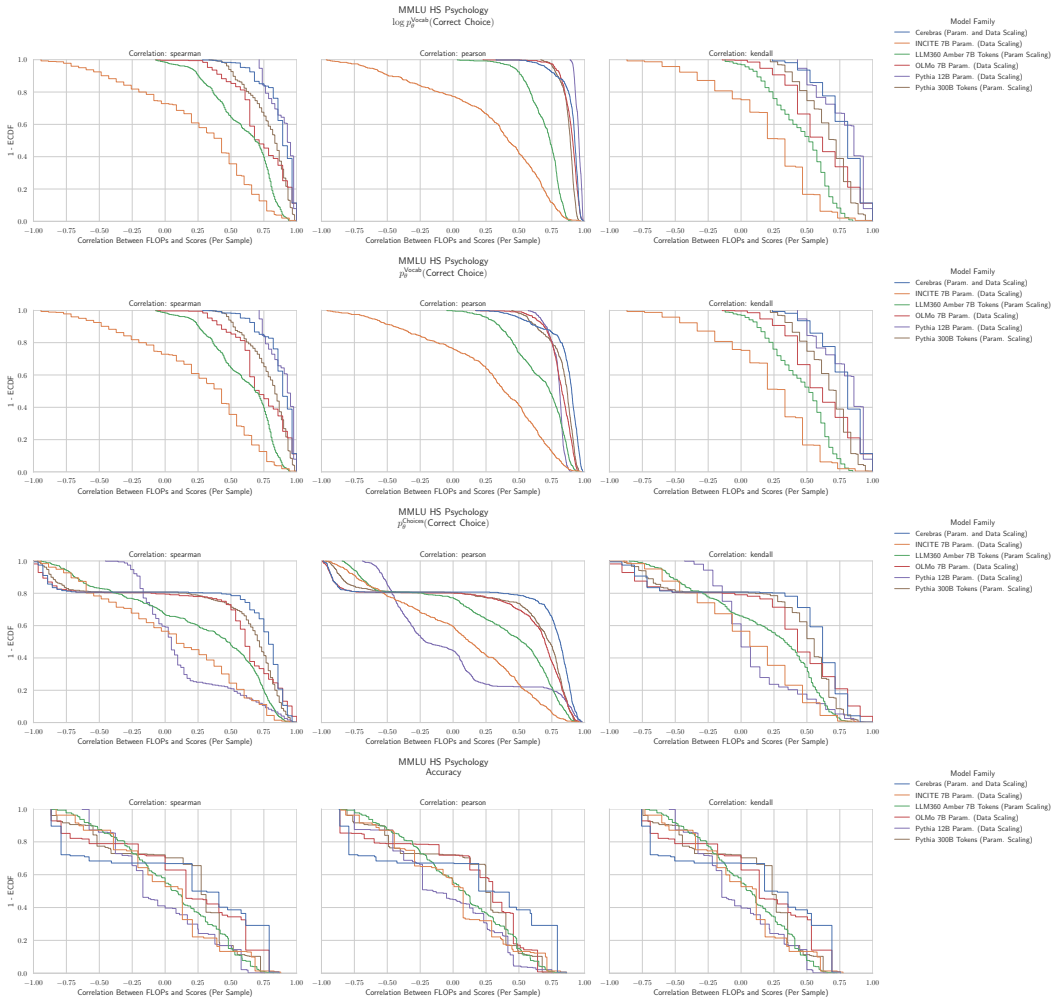

Figure 44: **MMLU High School Psychology: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.36 NLP BENCHMARK: MMLU HIGH SCHOOL STATISTICS HENDRYCKS ET AL. (2020)

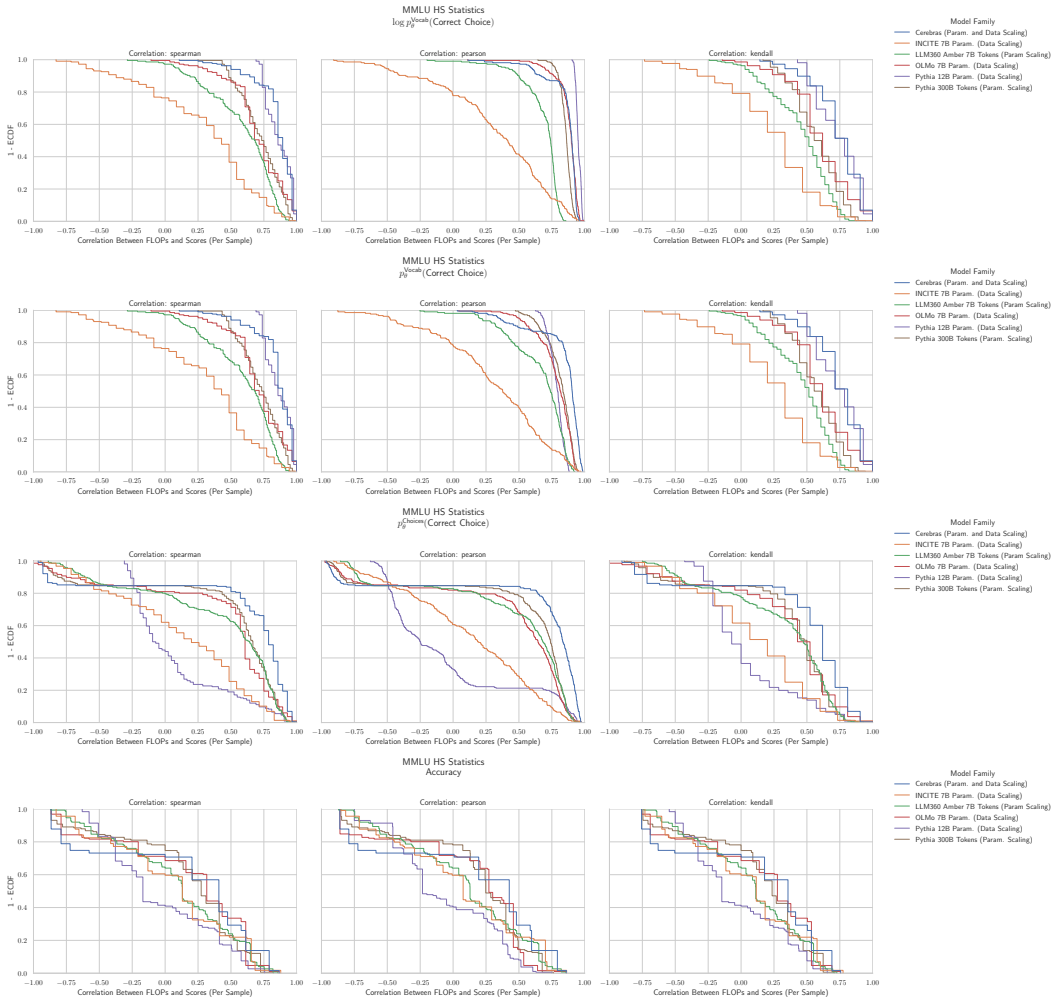

Figure 45: **MMLU High School Statistics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.37 NLP BENCHMARK: MMLU HIGH SCHOOL US HISTORY HENDRYCKS ET AL. (2020)

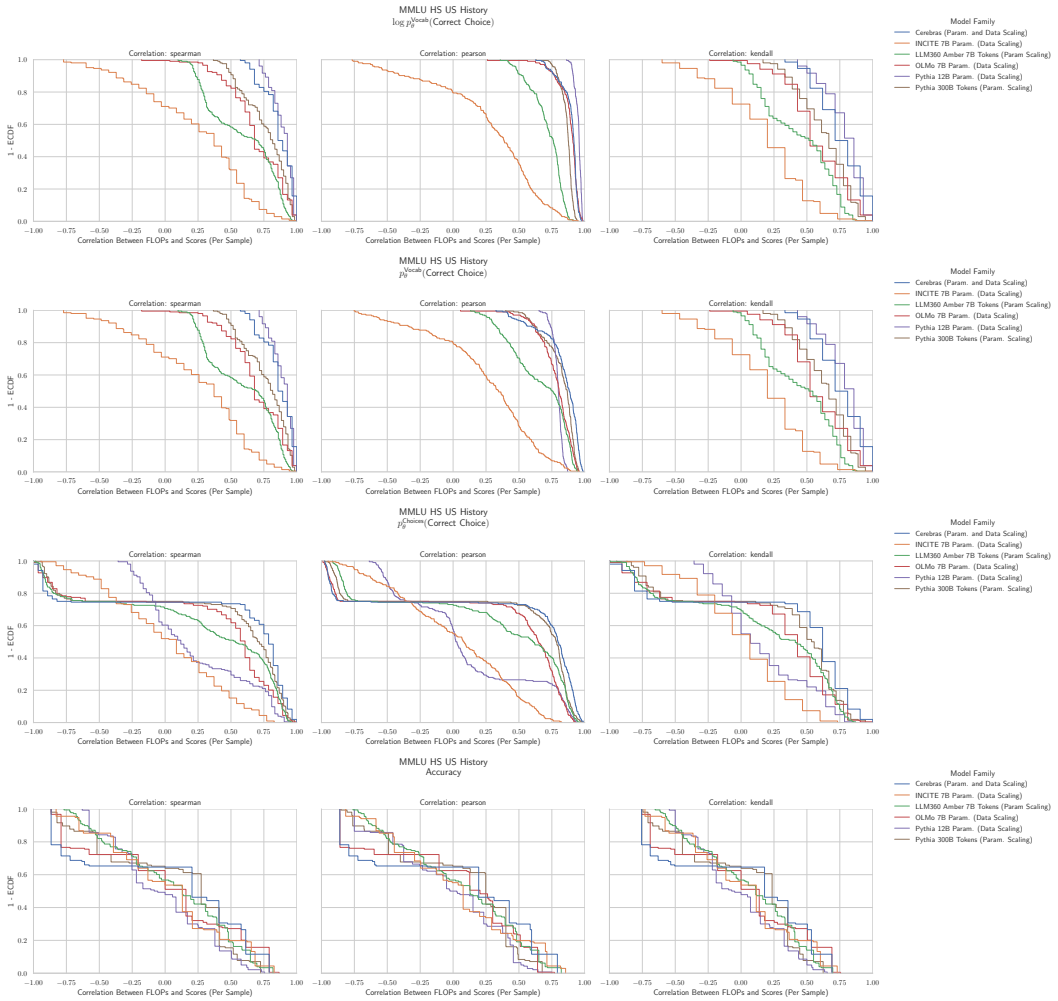

Figure 46: **MMLU High School US History: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.38 NLP BENCHMARK: MMLU HIGH SCHOOL WORLD HISTORY HENDRYCKS ET AL. (2020)

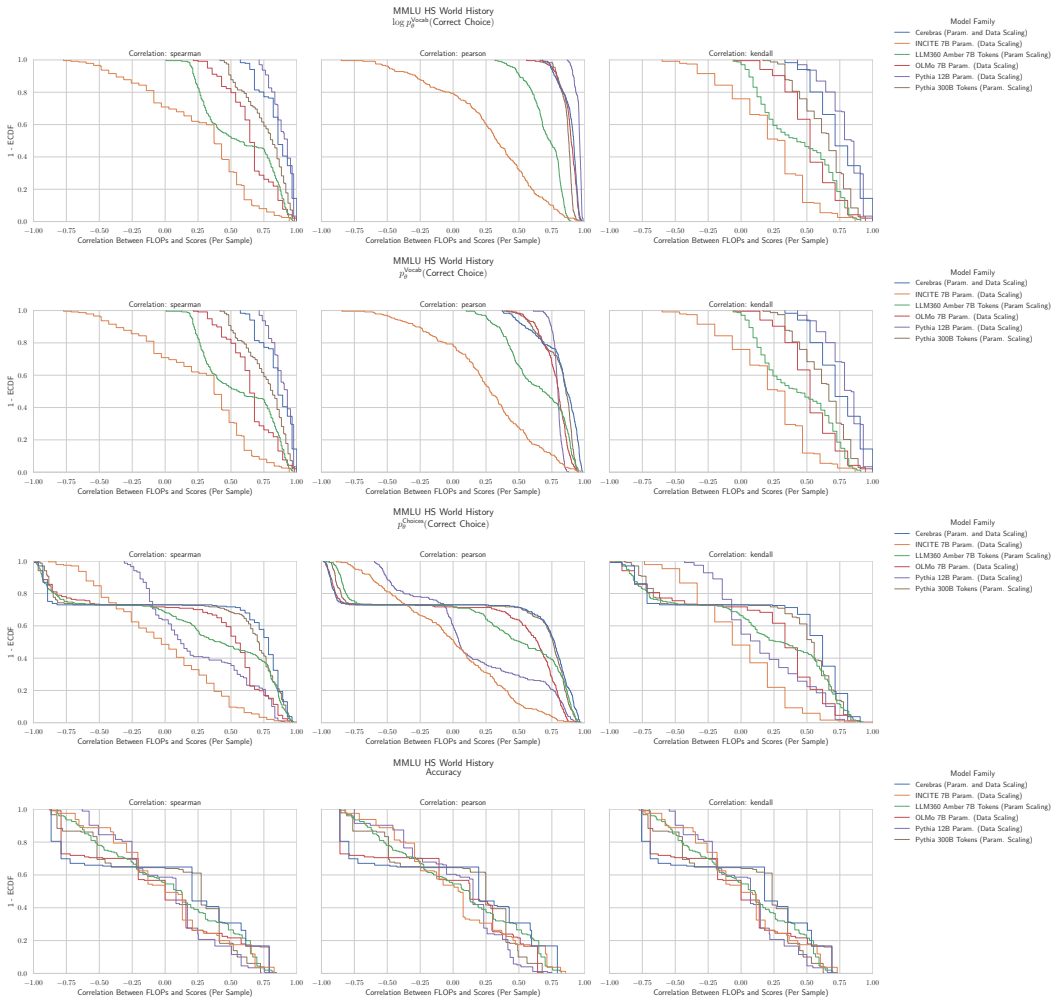

Figure 47: **MMLU High School World History: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.39 NLP BENCHMARK: MMLU HUMAN AGING HENDRYCKS ET AL. (2020)

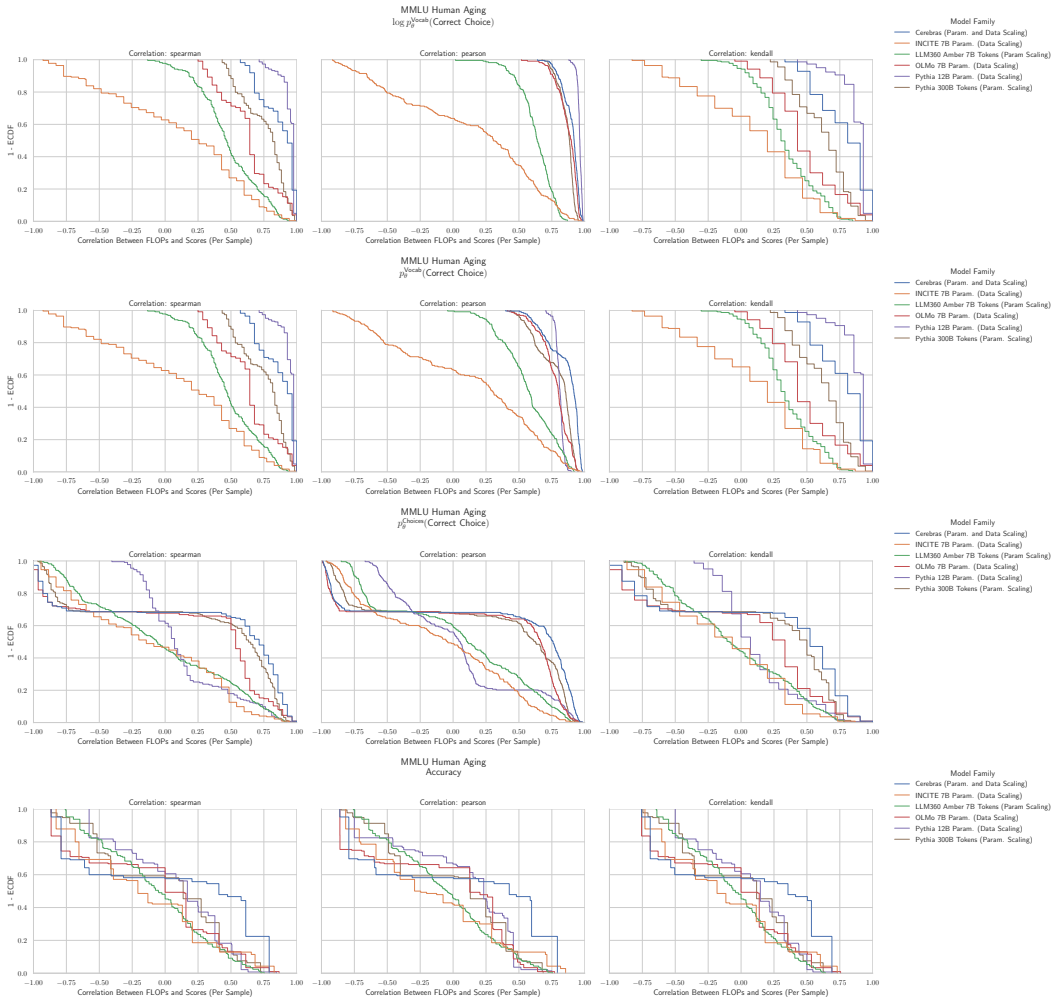

Figure 48: **MMLU Human Aging: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.40 NLP BENCHMARK: MMLU HUMAN SEXUALITY HENDRYCKS ET AL. (2020)

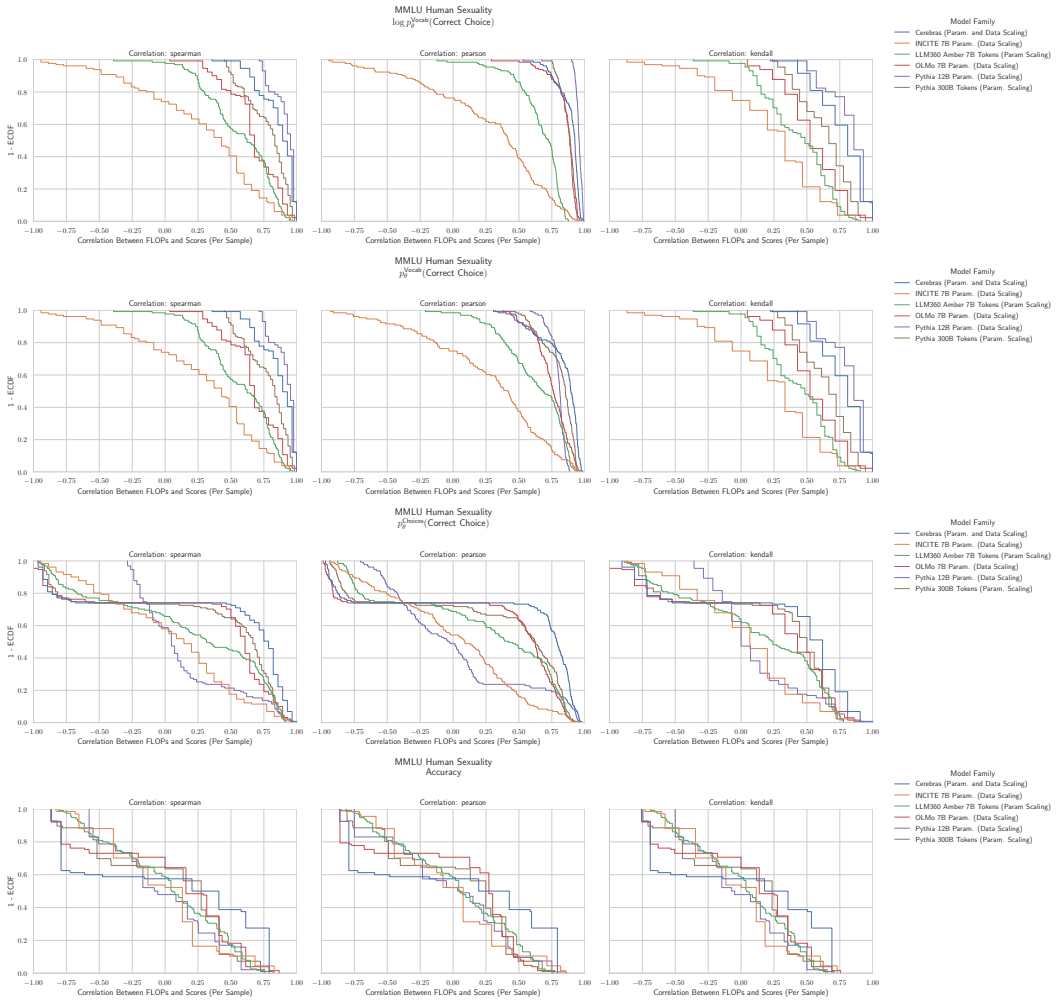

Figure 49: **MMLU Human Sexuality: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.41 NLP BENCHMARK: MMLU INTERNATIONAL LAW HENDRYCKS ET AL. (2020)

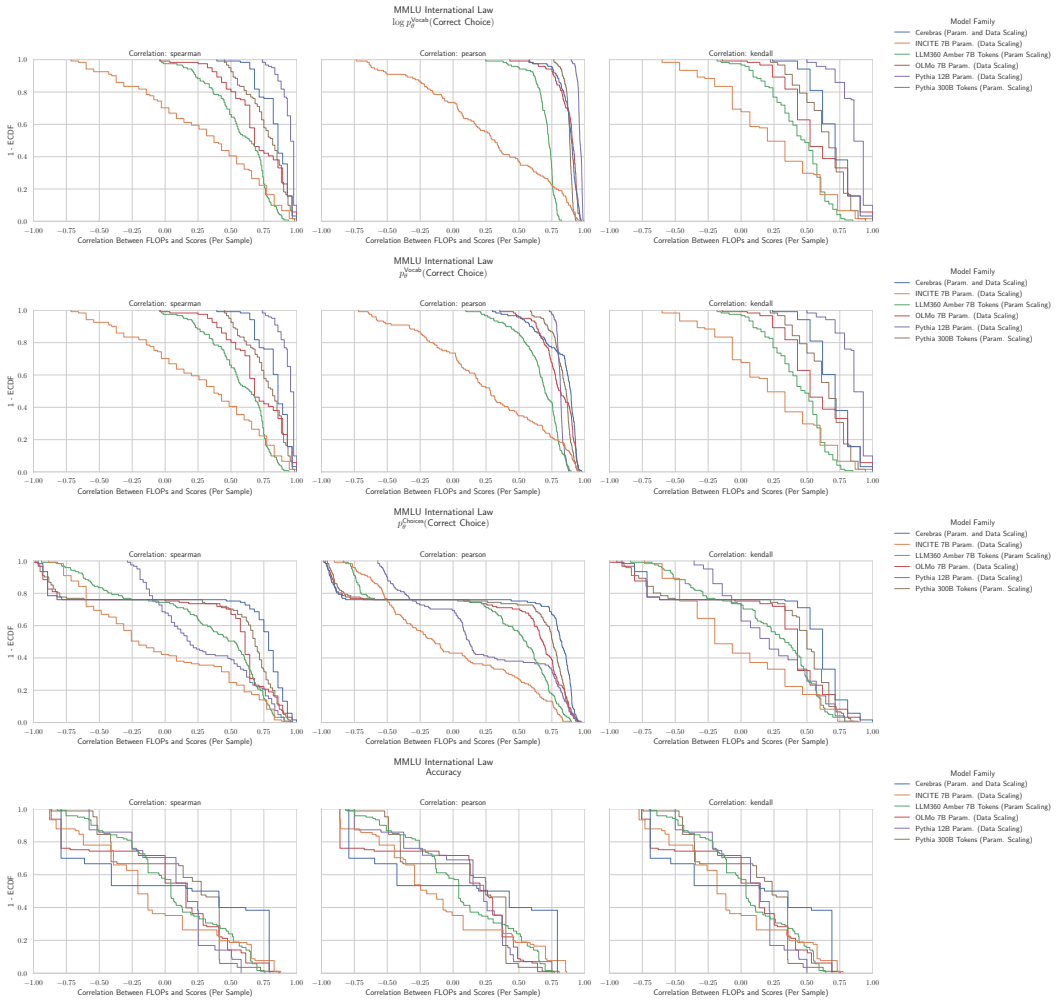

Figure 50: **MMLU International Law: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.42 NLP BENCHMARK: MMLU JURISPRUDENCE HENDRYCKS ET AL. (2020)

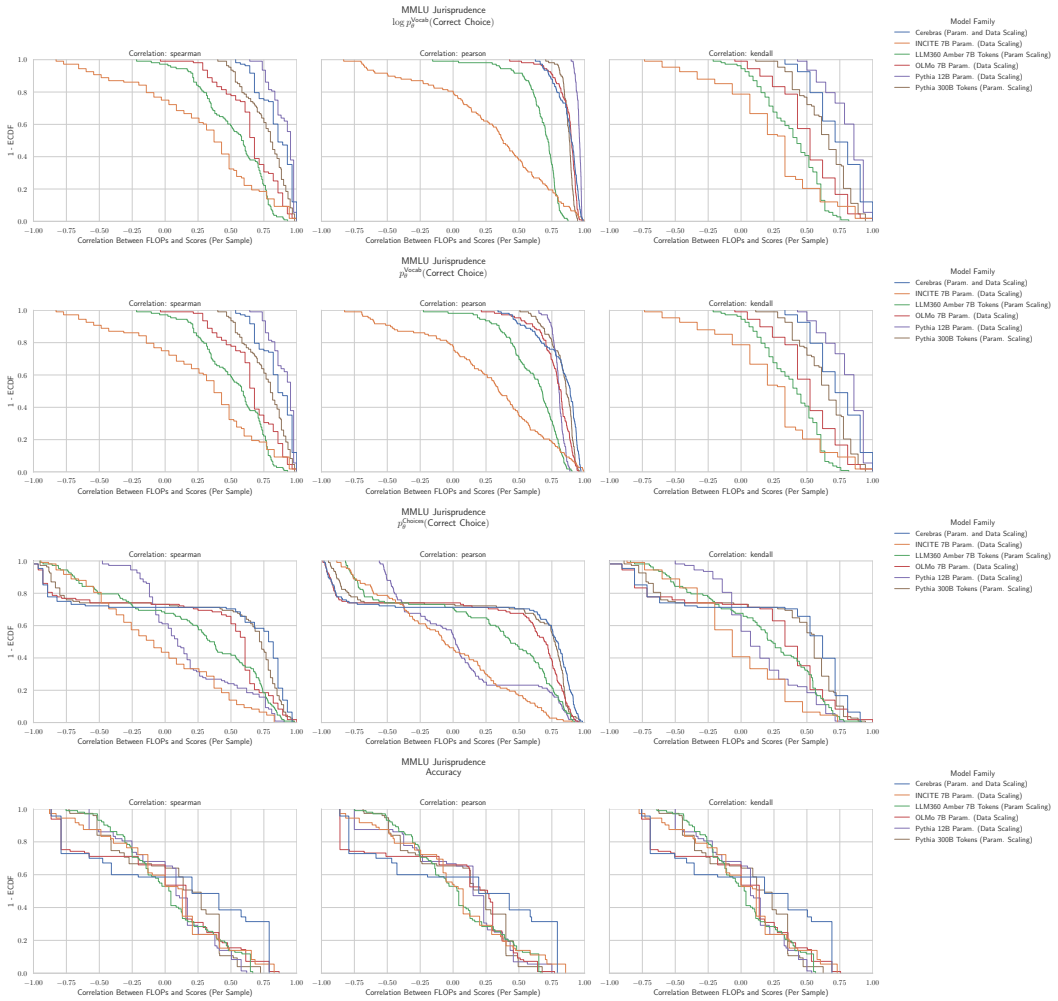

Figure 51: **MMLU Jurisprudence: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.43 NLP BENCHMARK: MMLU LOGICAL FALLACIES HENDRYCKS ET AL. (2020)

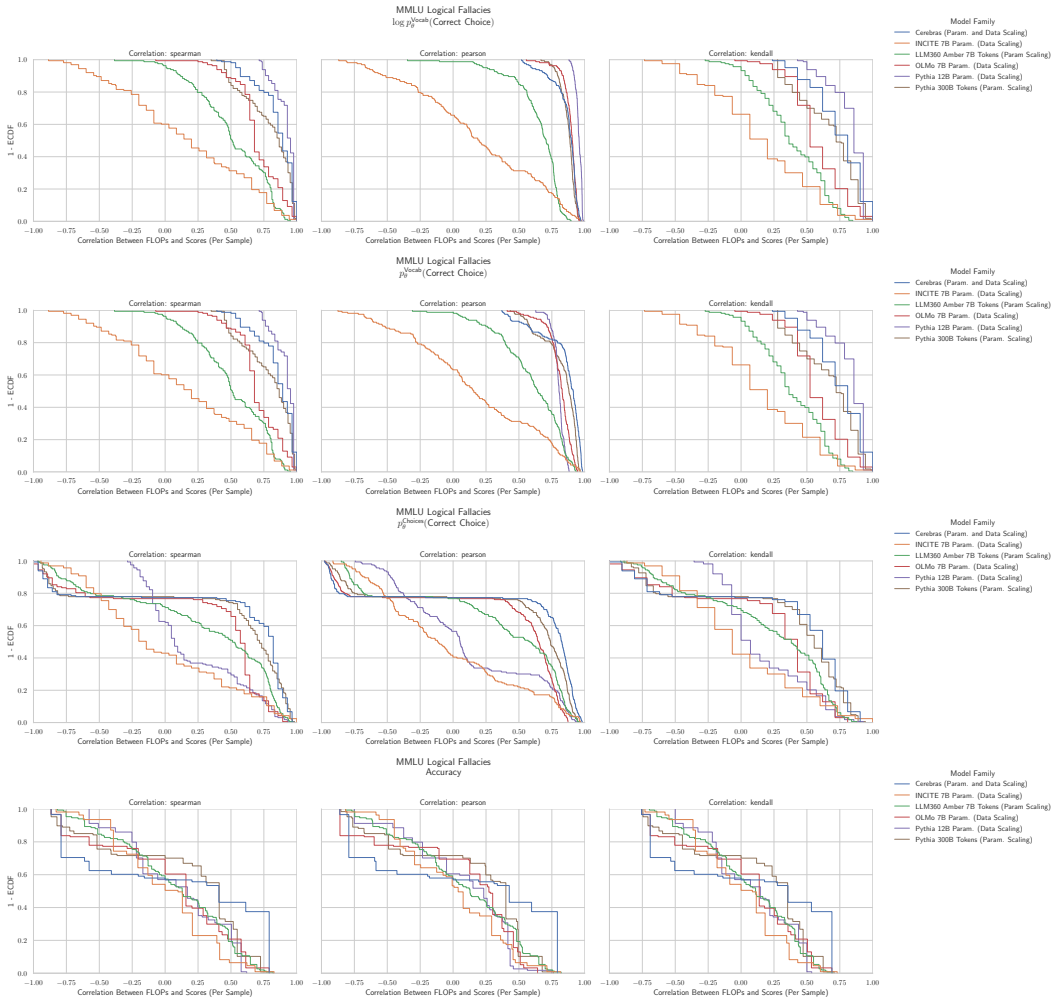

Figure 52: **MMLU Logical Fallacies: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.44 NLP BENCHMARK: MMLU MACHINE LEARNING HENDRYCKS ET AL. (2020)

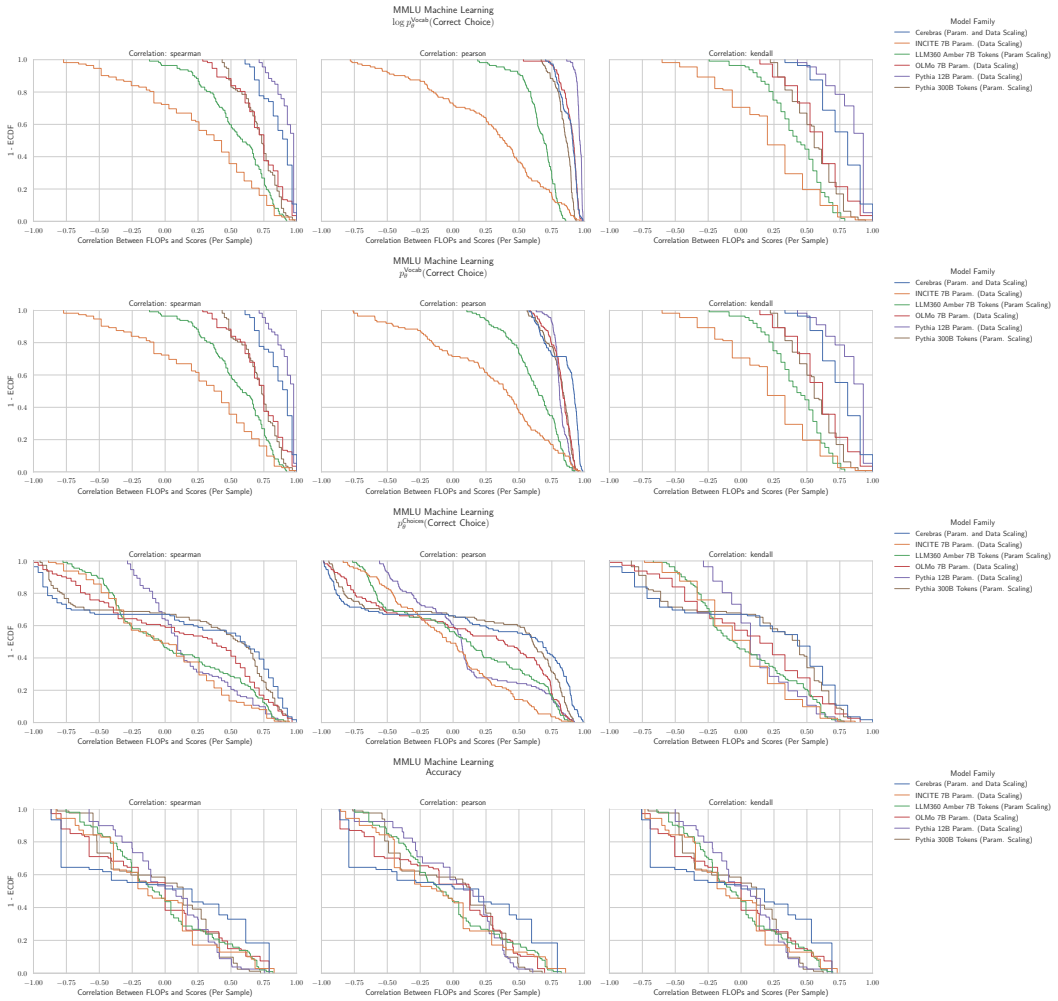

Figure 53: **MMLU Machine Learning: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.45 NLP BENCHMARK: MMLU MANAGEMENT HENDRYCKS ET AL. (2020)

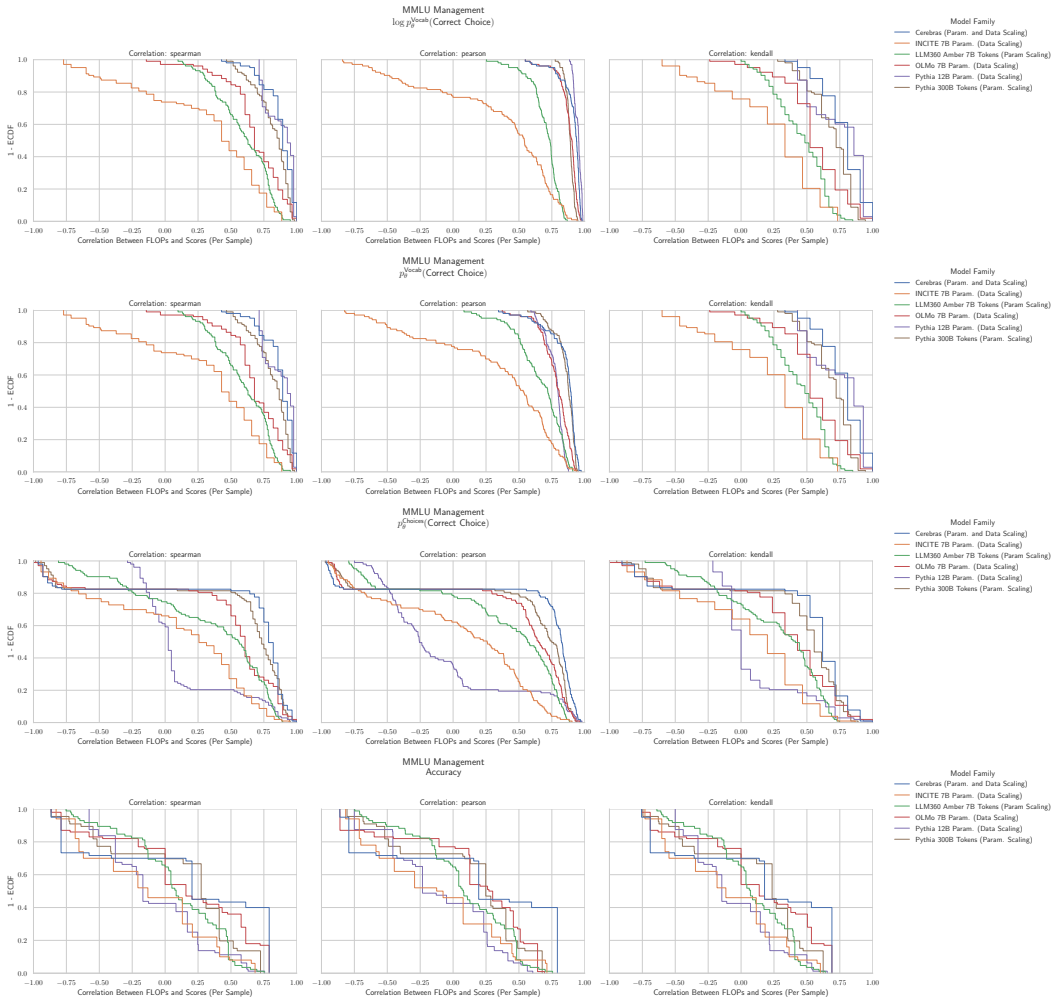

Figure 54: **MMLU Management: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.46 NLP BENCHMARK: MMLU MARKETING HENDRYCKS ET AL. (2020)

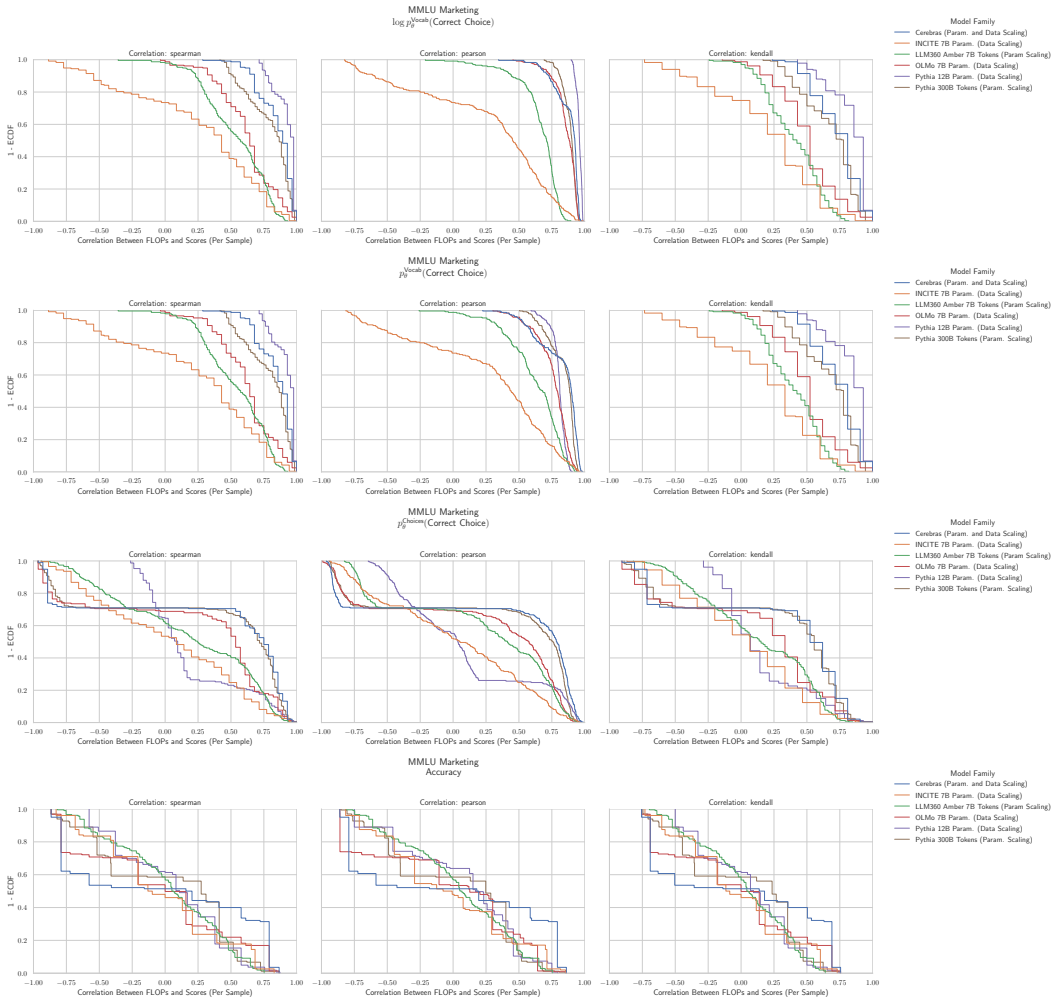

Figure 55: **MMLU Marketing: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.47 NLP BENCHMARK: MMLU MEDICAL GENETICS HENDRYCKS ET AL. (2020)

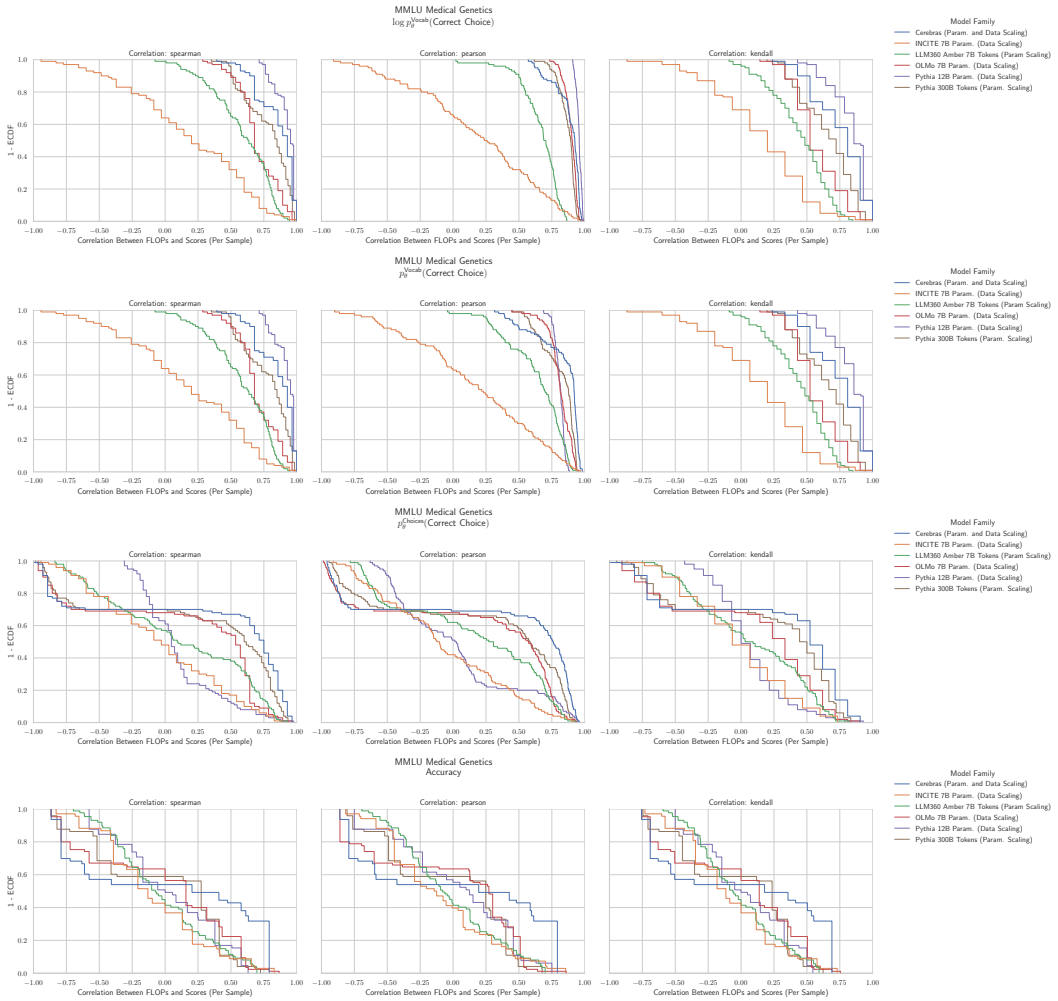

Figure 56: **MMLU Medical Genetics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.48 NLP BENCHMARK: MMLU MISCELLANEOUS HENDRYCKS ET AL. (2020)

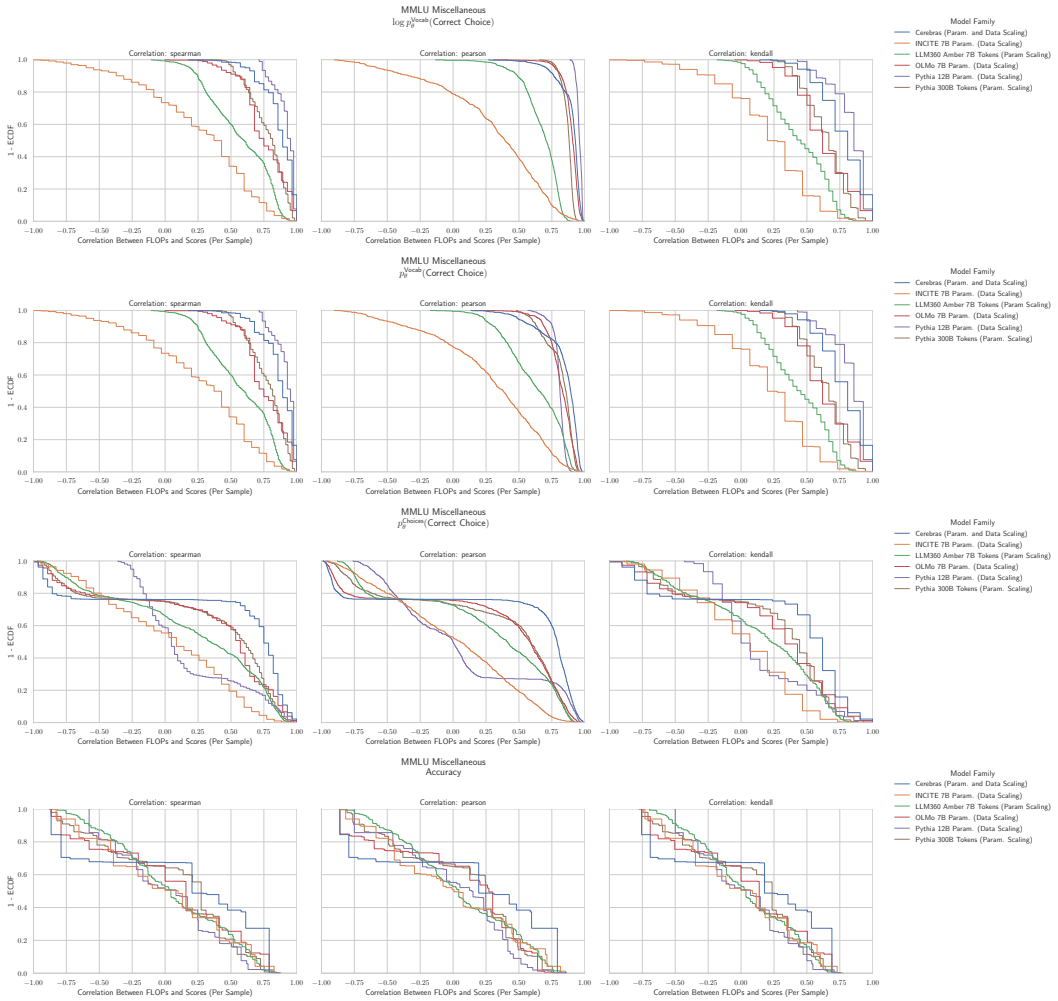

Figure 57: **MMLU Miscellaneous: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.49 NLP BENCHMARK: MMLU MORAL DISPUTES HENDRYCKS ET AL. (2020)

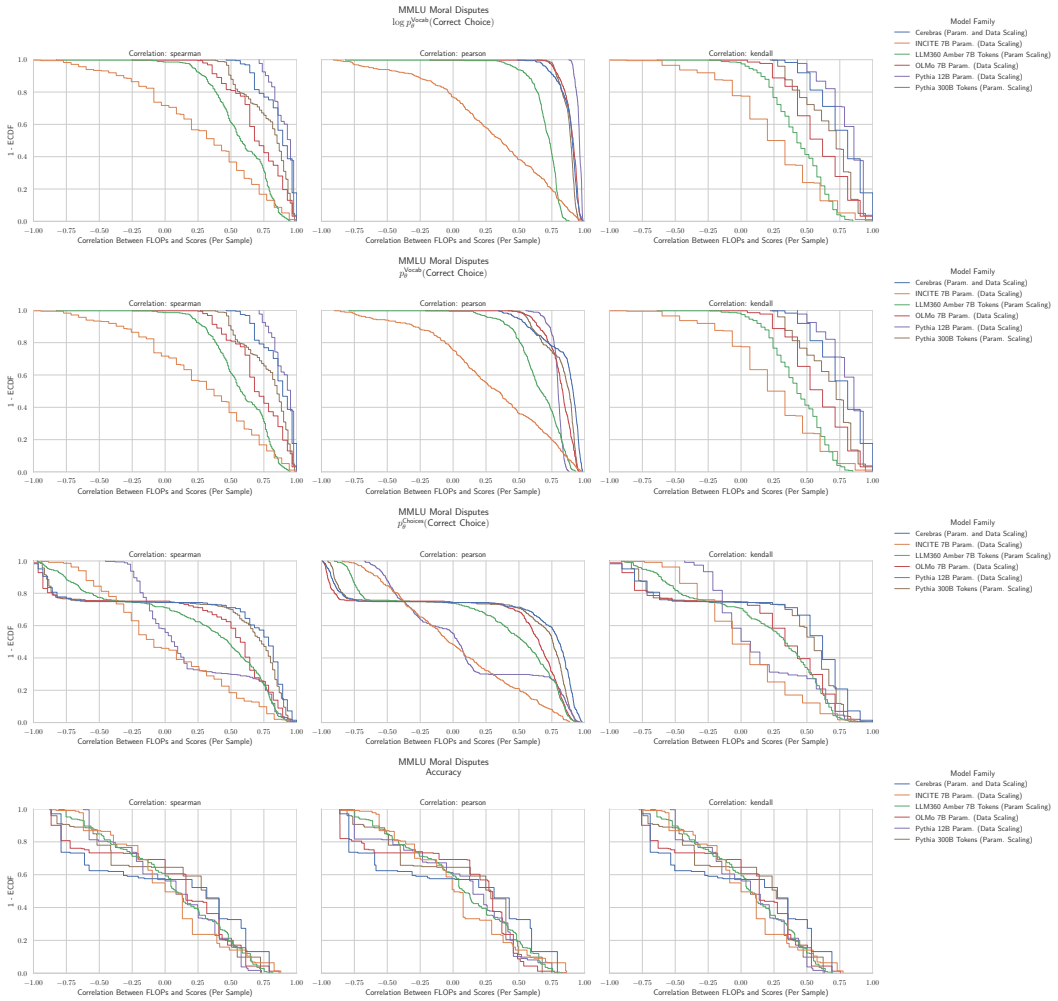

Figure 58: **MMLU Moral Disputes: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.50 NLP BENCHMARK: MMLU MORAL SCENARIOS HENDRYCKS ET AL. (2020)

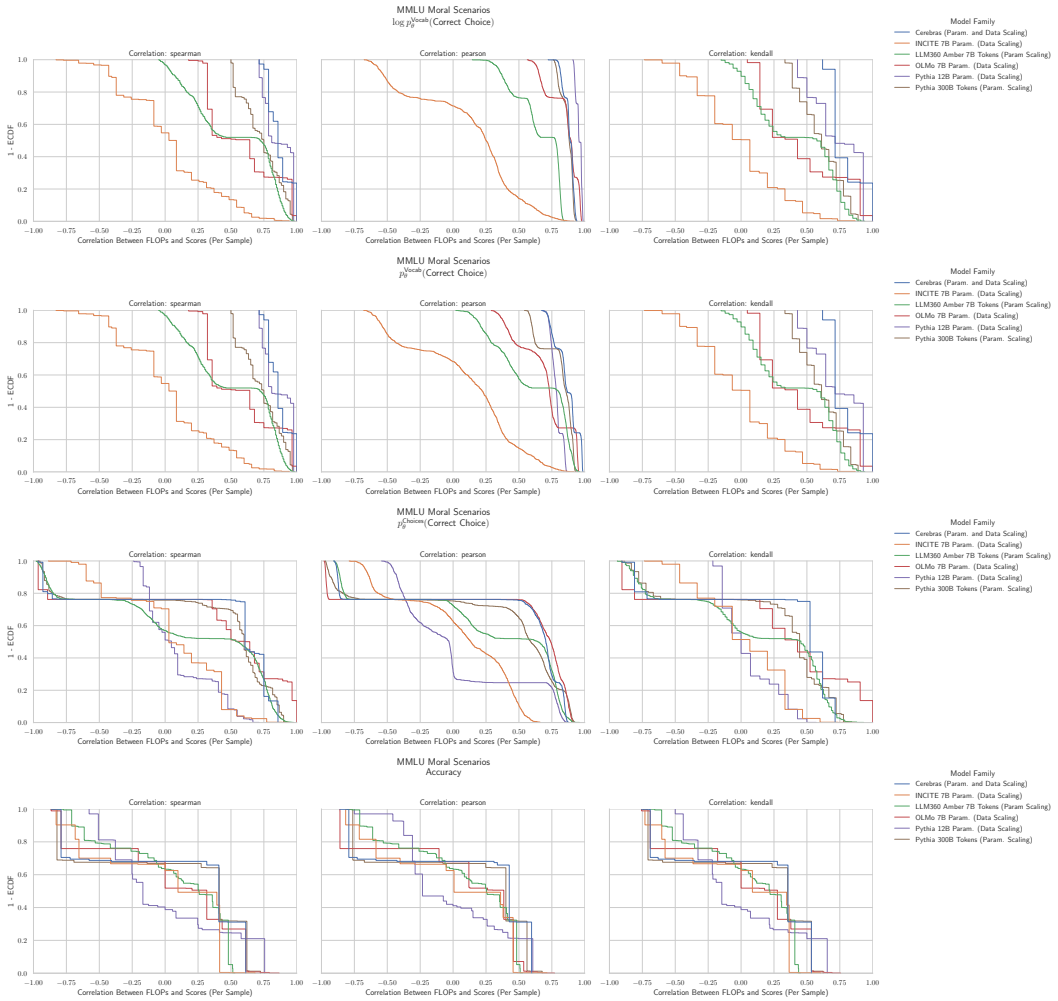

Figure 59: **MMLU Moral Scenarios: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.51 NLP BENCHMARK: MMLU NUTRITION HENDRYCKS ET AL. (2020)

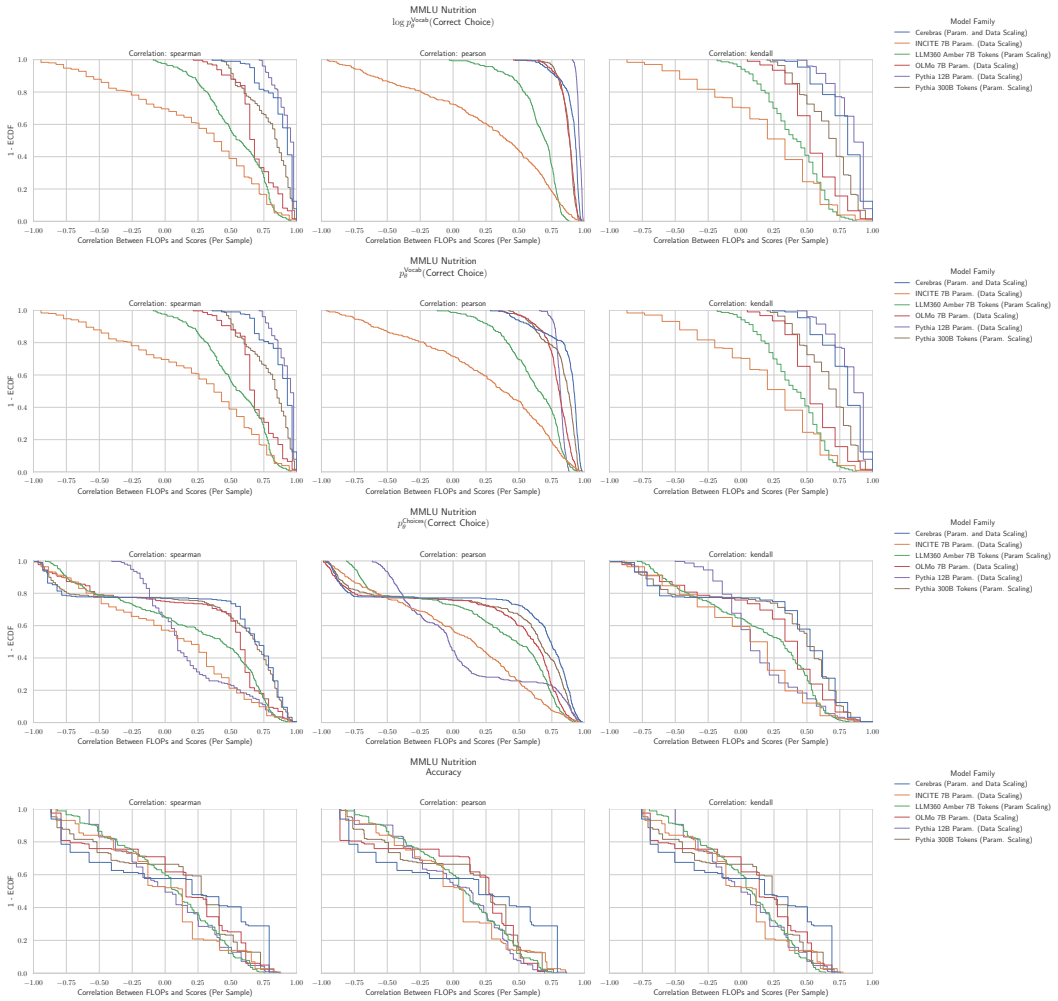

Figure 60: **MMLU Nutrition: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.52 NLP BENCHMARK: MMLU PHILOSOPHY HENDRYCKS ET AL. (2020)

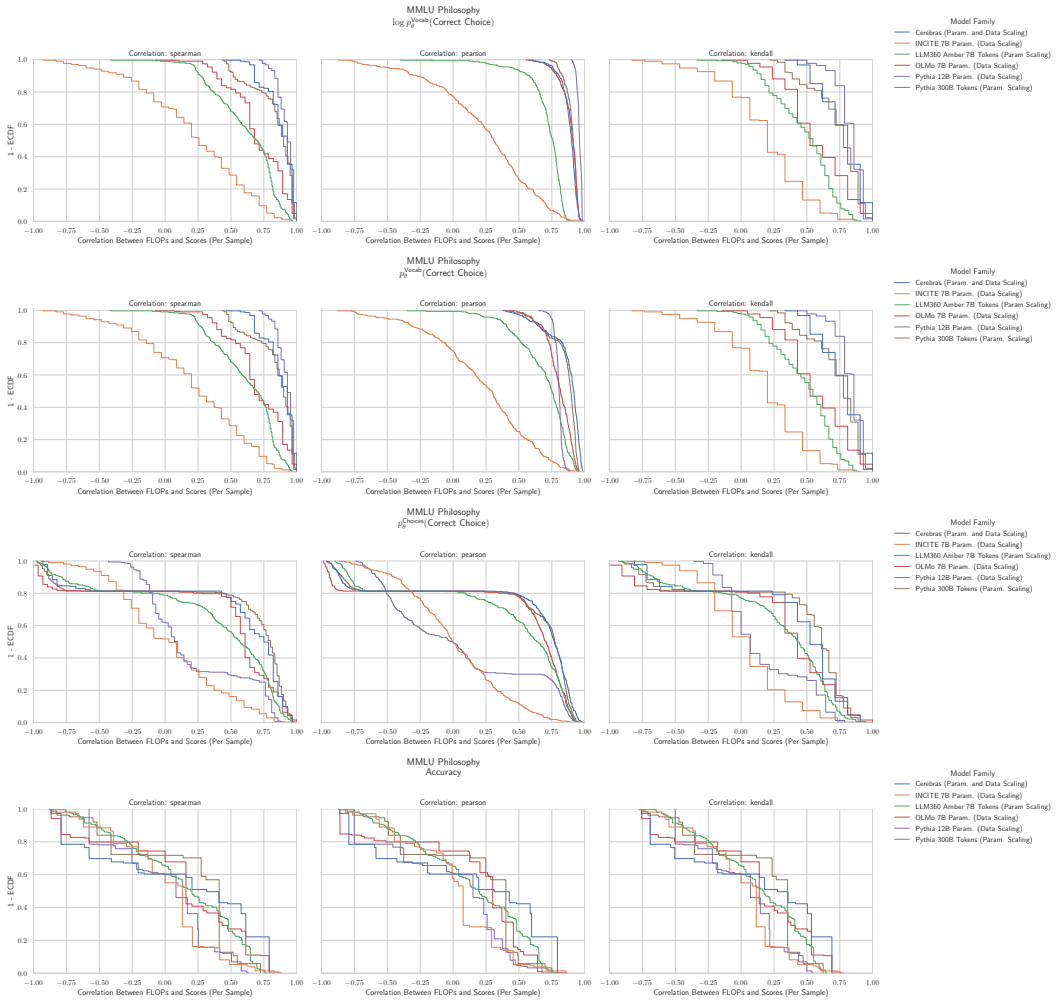

Figure 61: **MMLU Philosophy: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.53 NLP BENCHMARK: MMLU PREHISTORY HENDRYCKS ET AL. (2020)

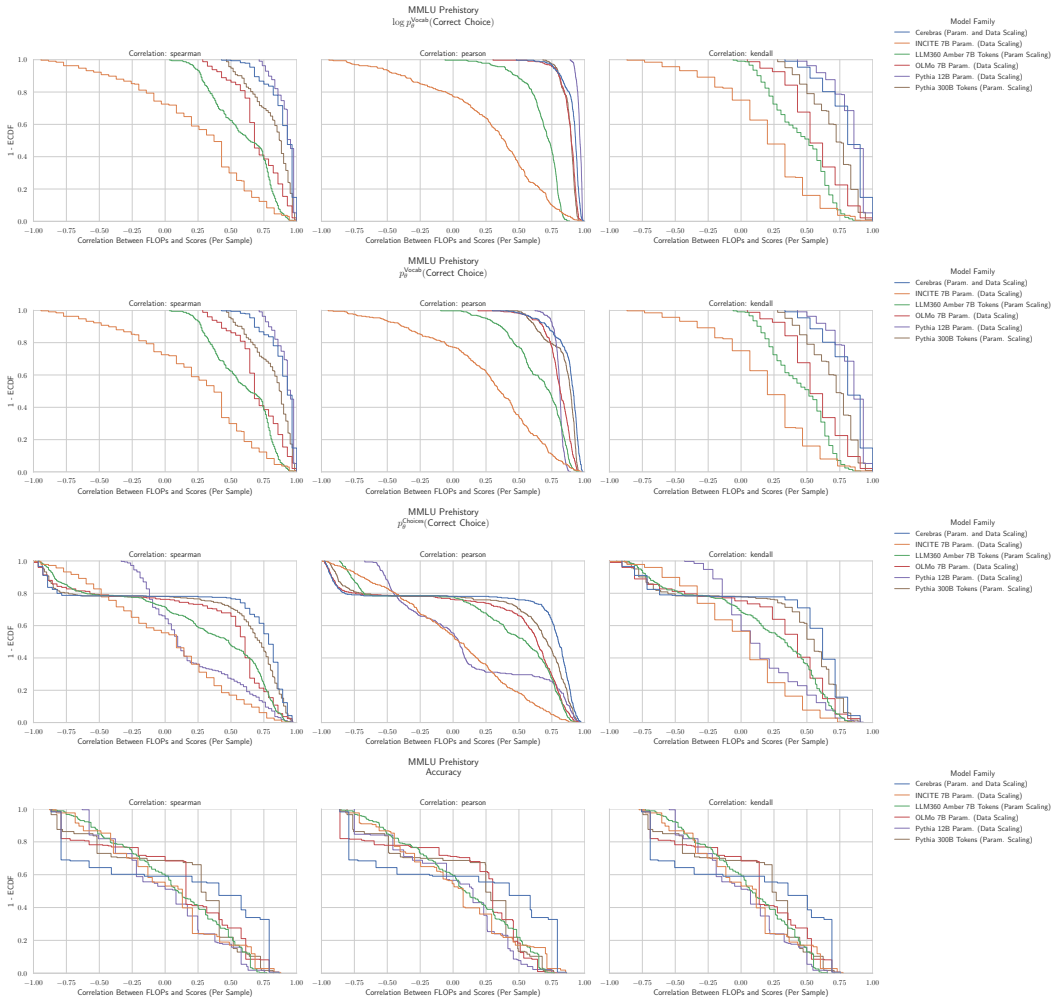

Figure 62: **MMLU Prehistory: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.54 NLP BENCHMARK: MMLU PROFESSIONAL ACCOUNTING HENDRYCKS ET AL. (2020)

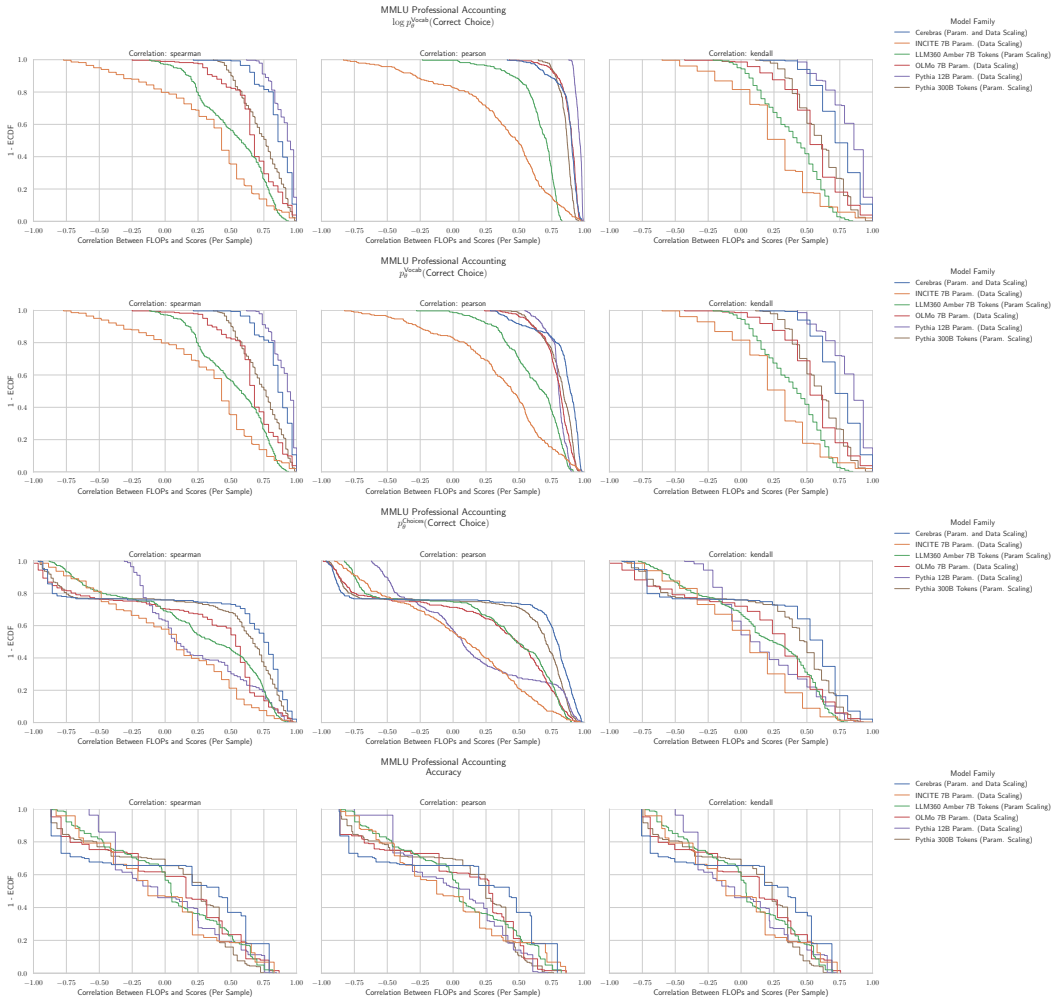

Figure 63: **MMLU Professional Accounting: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.55 NLP BENCHMARK: MMLU PROFESSIONAL LAW HENDRYCKS ET AL. (2020)

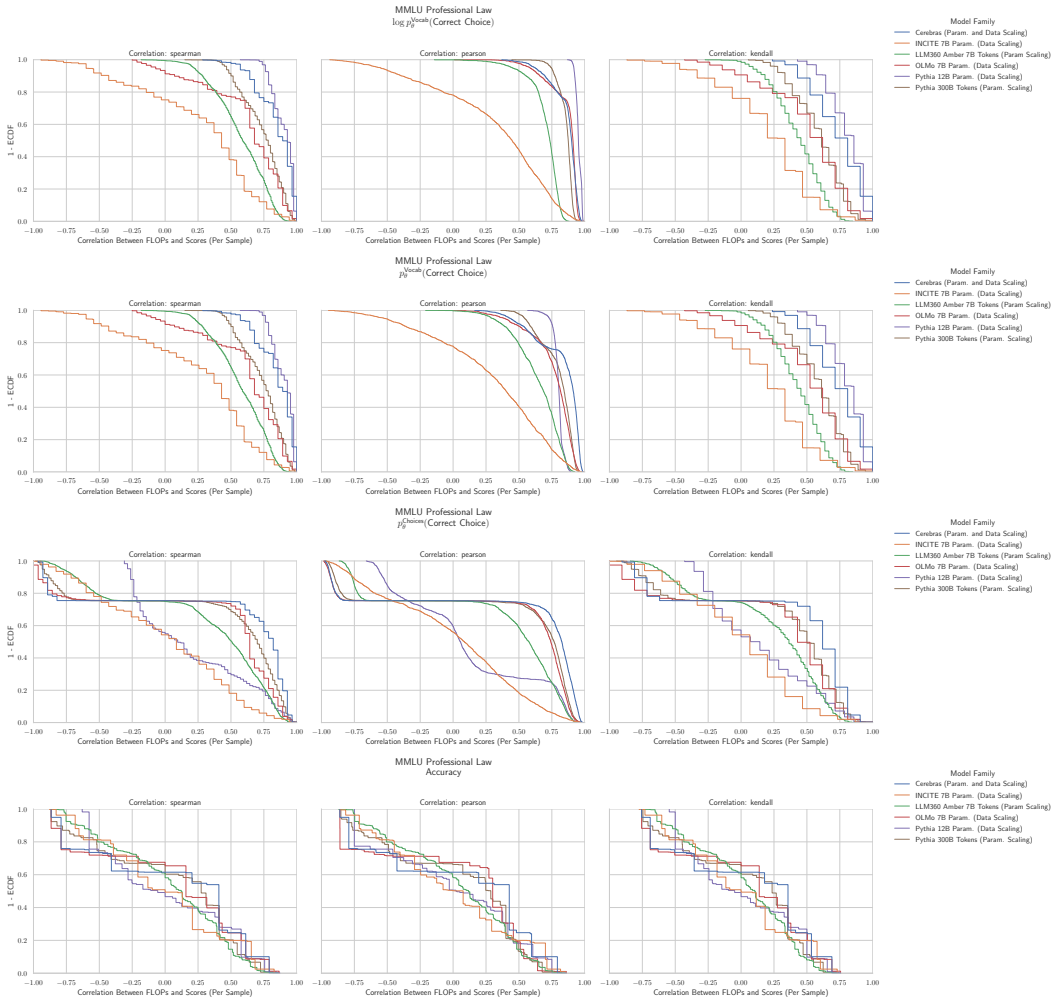

Figure 64: **MMLU Professional Law: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.56 NLP BENCHMARK: MMLU PROFESSIONAL MEDICINE HENDRYCKS ET AL. (2020)

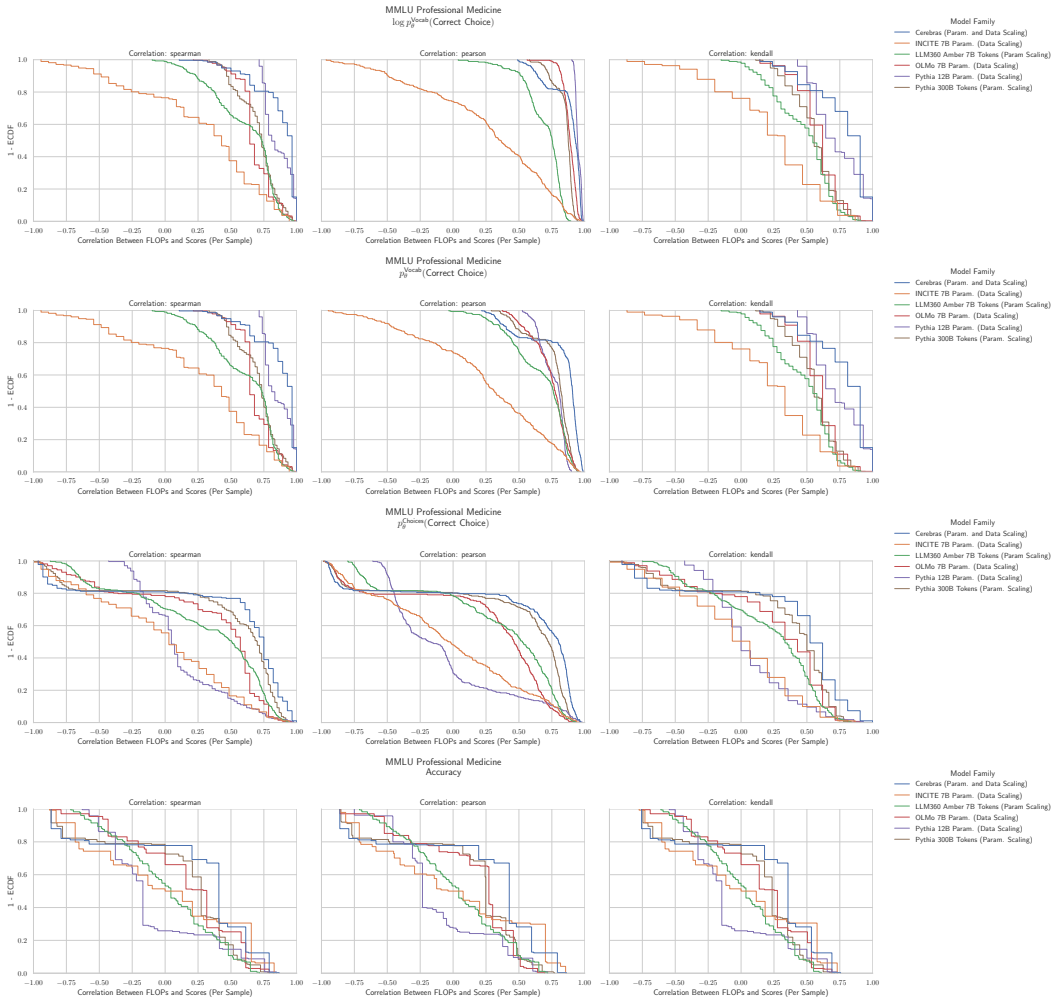

Figure 65: **MMLU Professional Medicine: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.57 NLP BENCHMARK: MMLU PROFESSIONAL PSYCHOLOGY HENDRYCKS ET AL. (2020)

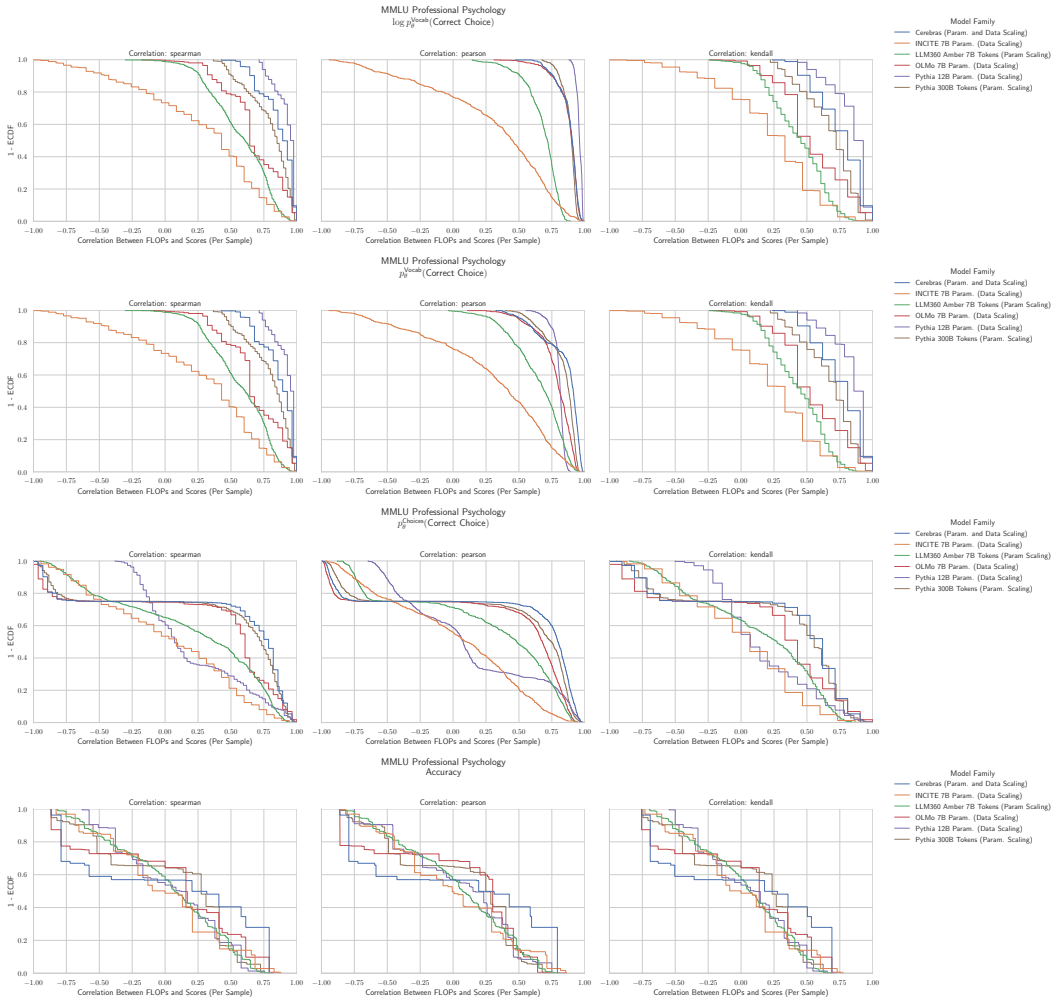

Figure 66: **MMLU Professional Psychology: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.58 NLP BENCHMARK: MMLU PUBLIC RELATIONS HENDRYCKS ET AL. (2020)

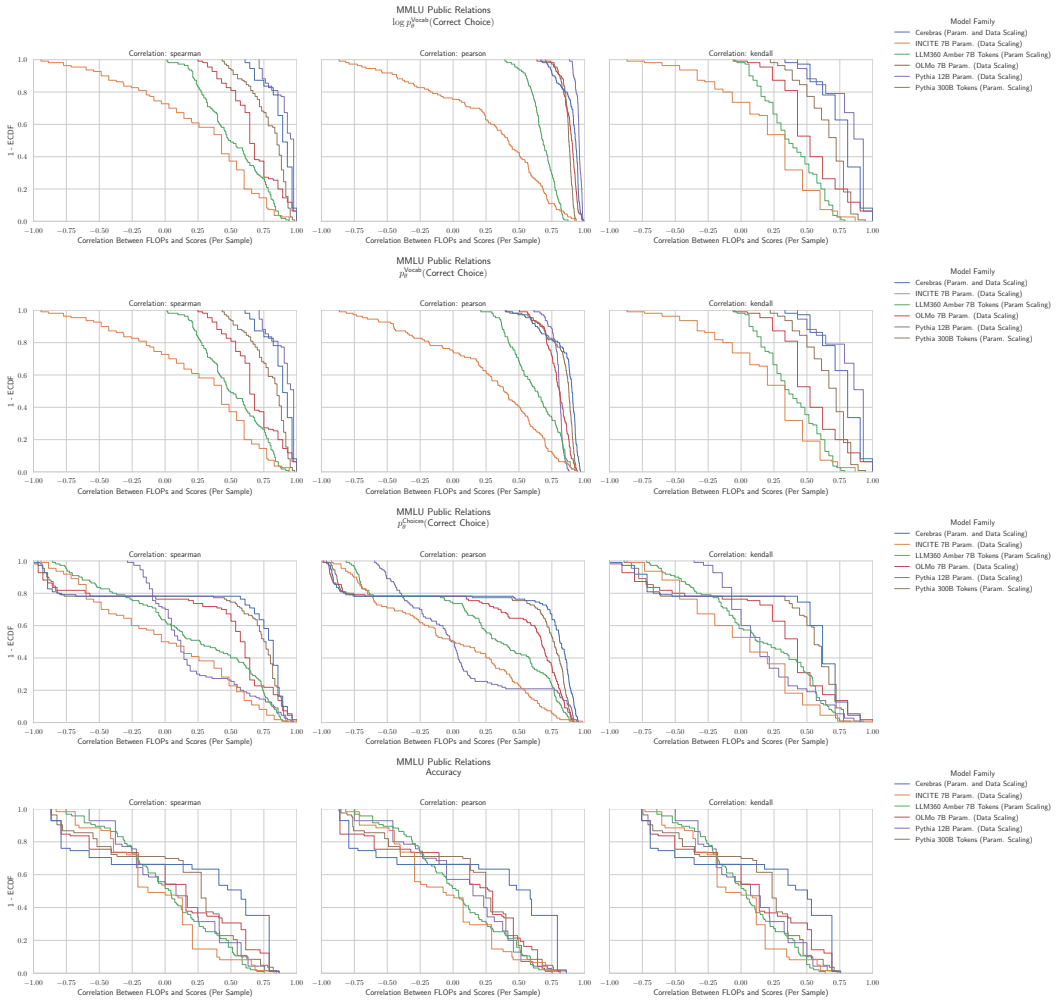

Figure 67: **MMLU Public Relations: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.59 NLP BENCHMARK: MMLU SECURITY STUDIES HENDRYCKS ET AL. (2020)

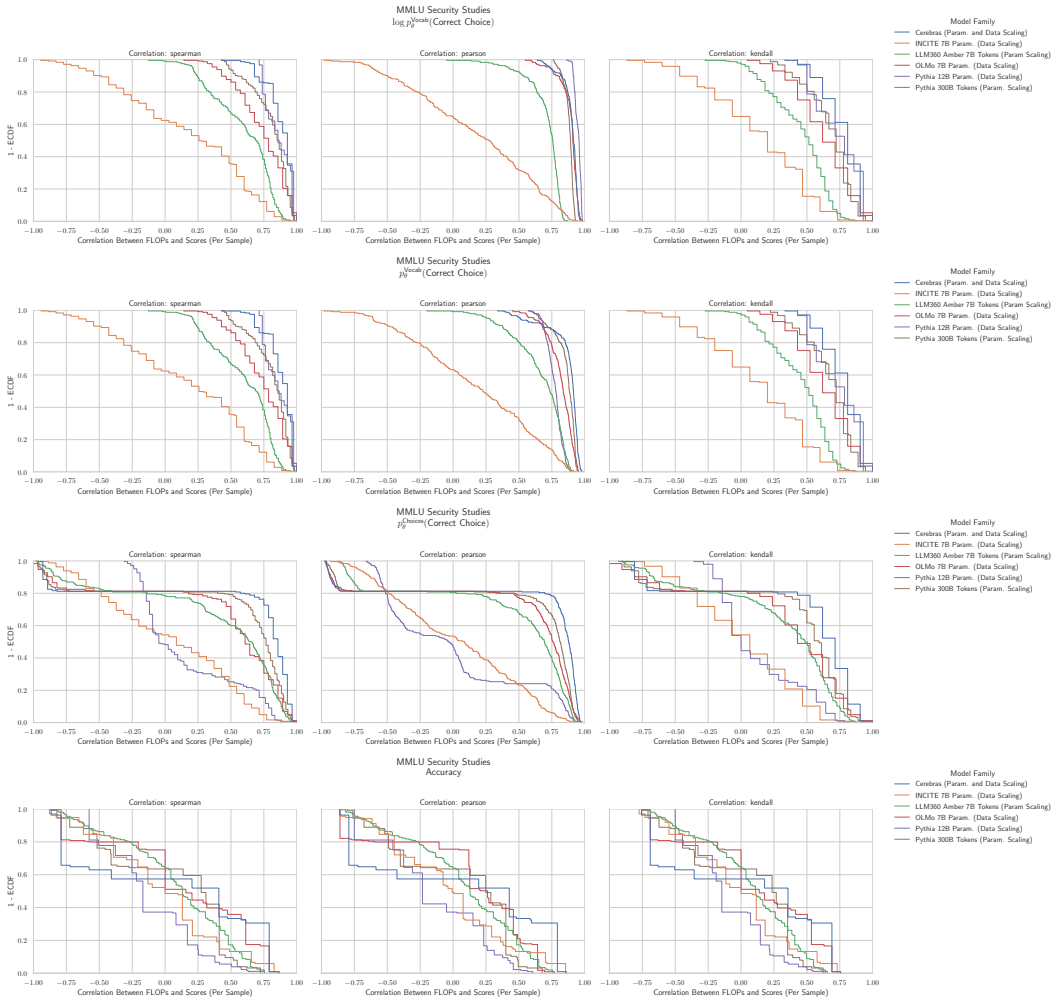

Figure 68: **MMLU Security Studies: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.60 NLP BENCHMARK: MMLU SOCIOLOGY HENDRYCKS ET AL. (2020)

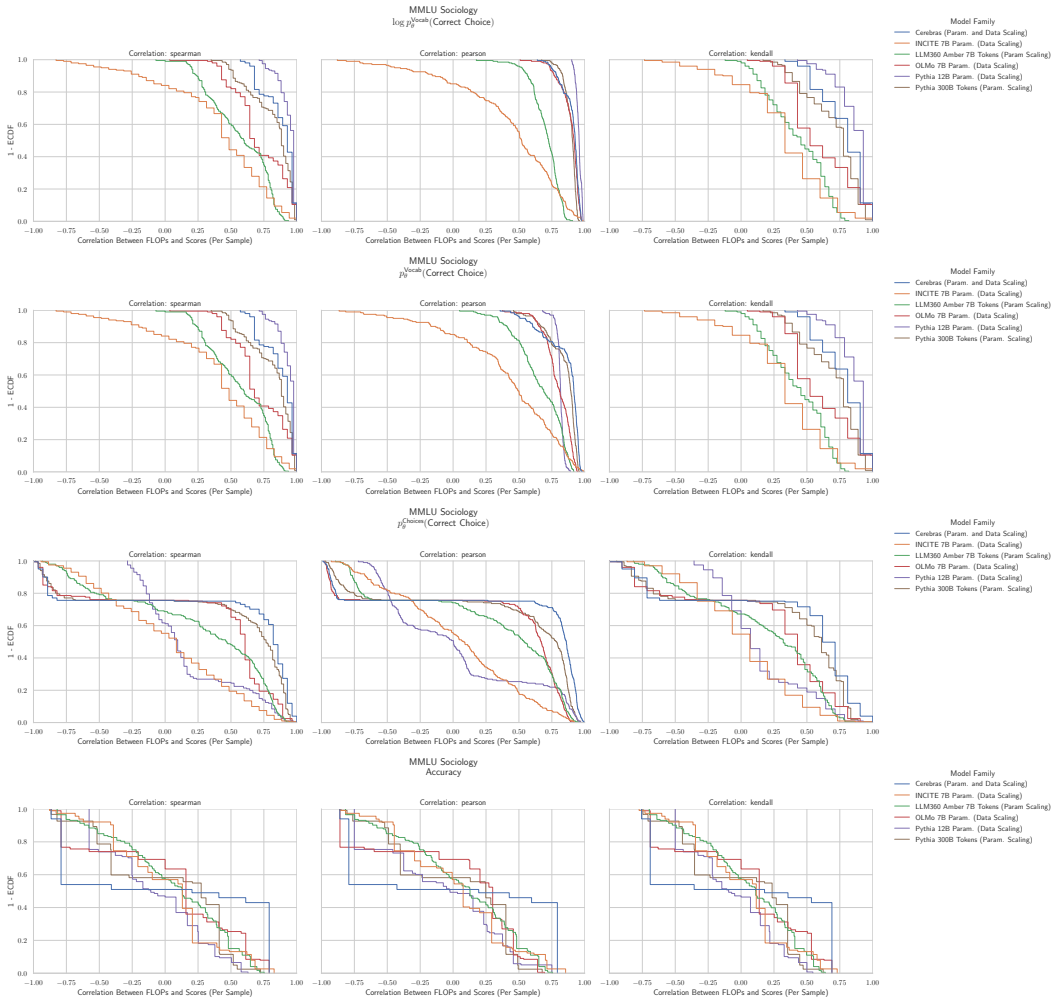

Figure 69: **MMLU Sociology: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.61 NLP BENCHMARK: MMLU US FOREIGN POLICY HENDRYCKS ET AL. (2020)

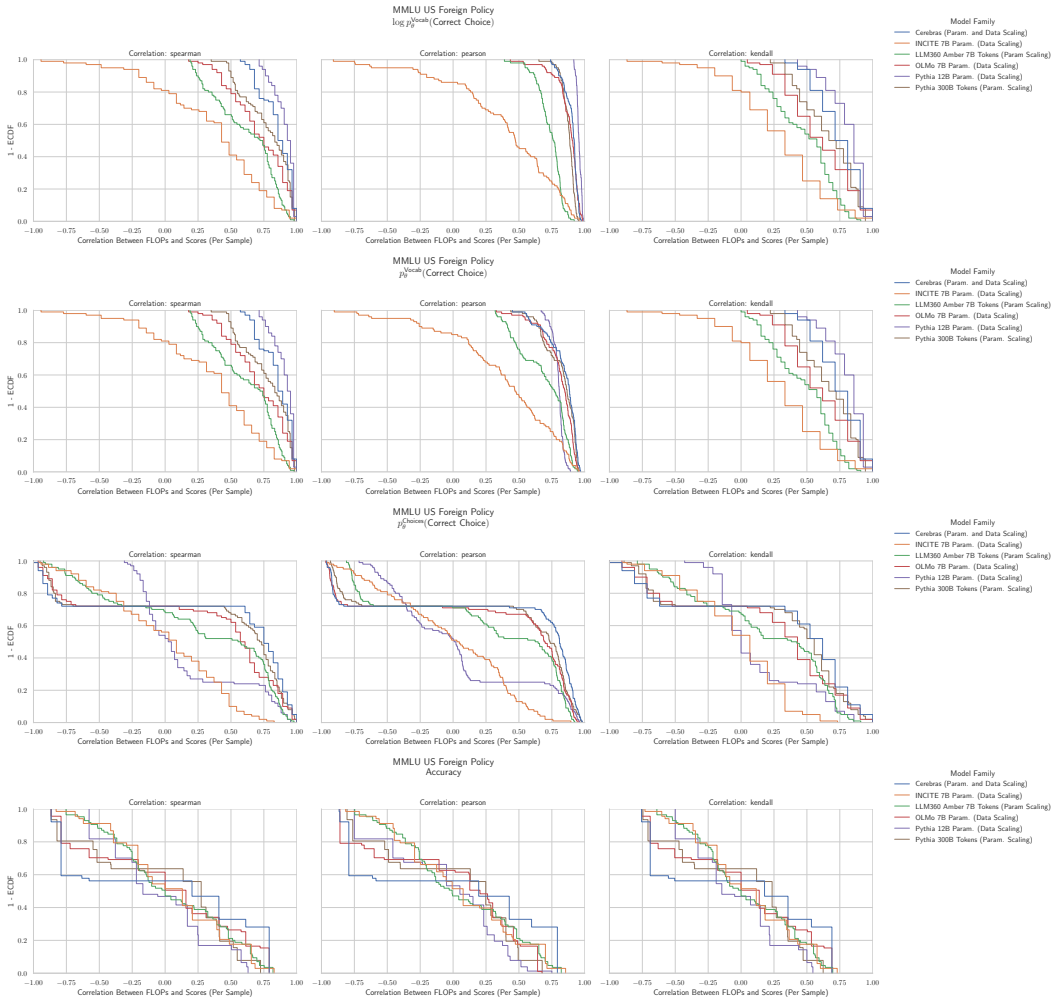

Figure 70: **MMLU US Foreign Policy: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.62 NLP BENCHMARK: MMLU VIROLOGY HENDRYCKS ET AL. (2020)

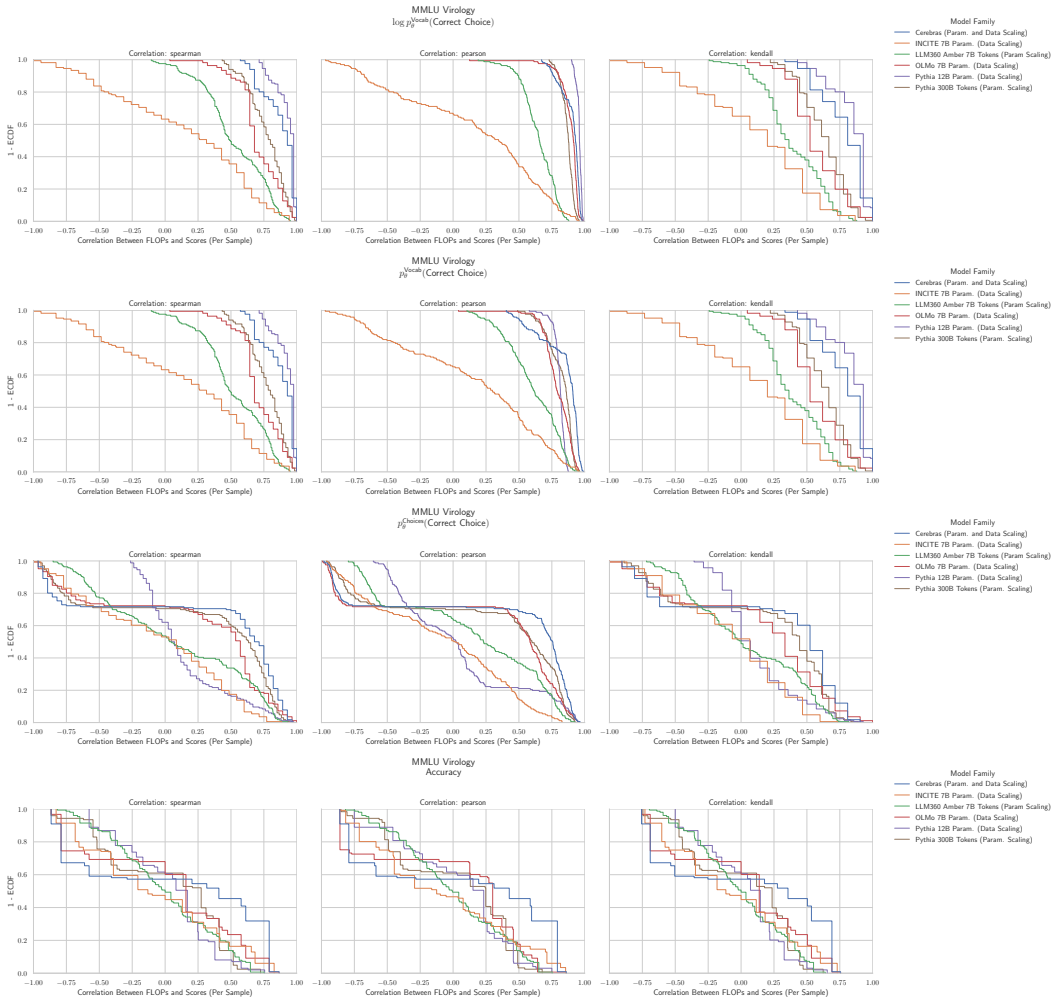

Figure 71: **MMLU Virology: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.63 NLP BENCHMARK: MMLU WORLD RELIGIONS HENDRYCKS ET AL. (2020)

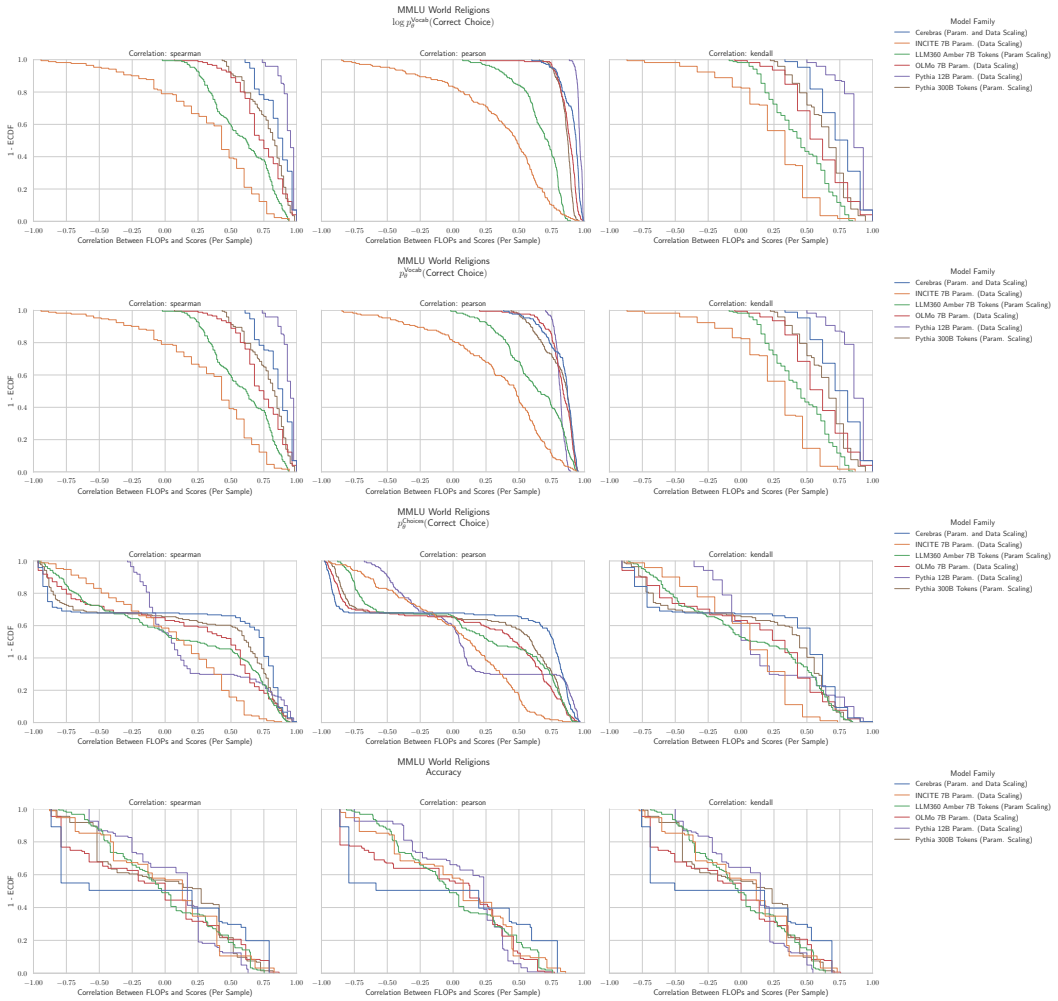

Figure 72: **MMLU World Religions: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.64 NLP BENCHMARK: OPENBOOKQA MIHAYLOV ET AL. (2018)

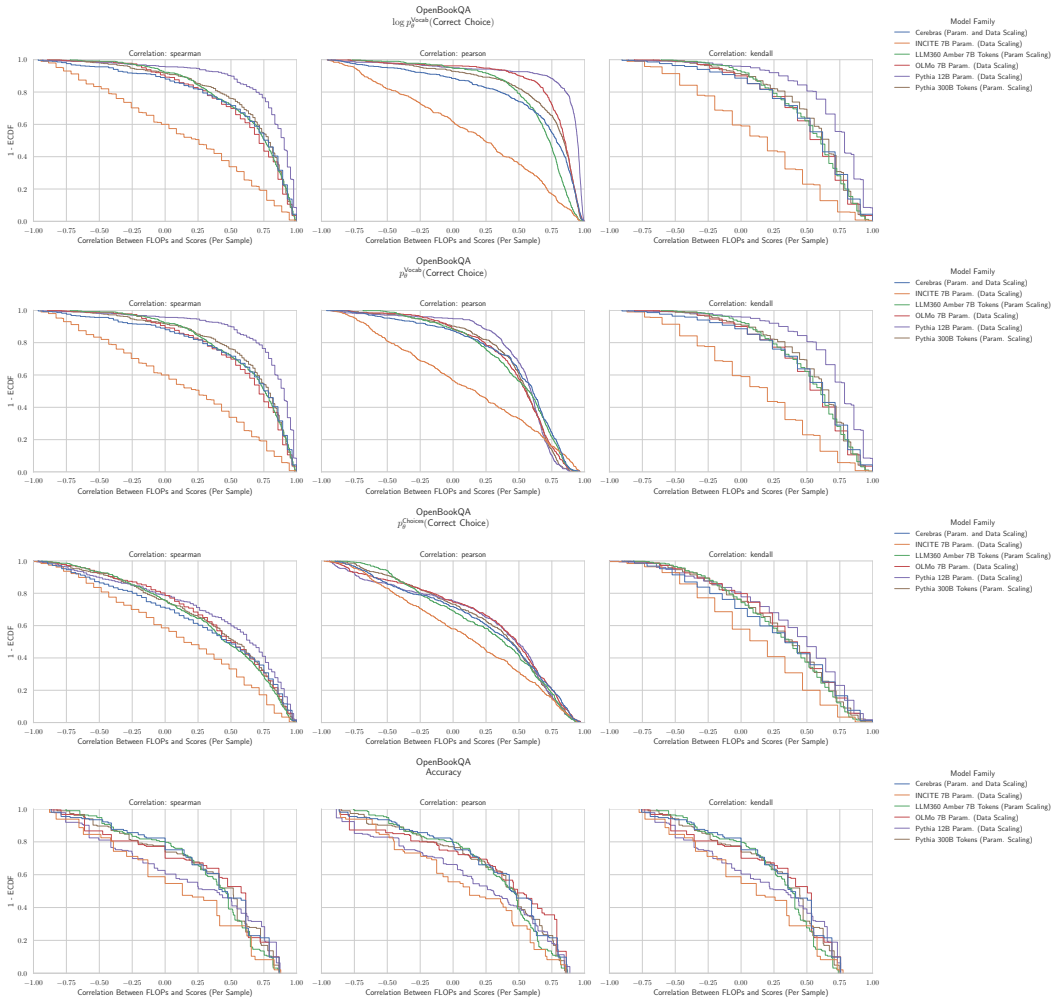

Figure 73: **OpenBookQA: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.65 NLP BENCHMARK: PIQA BISK ET AL. (2020)

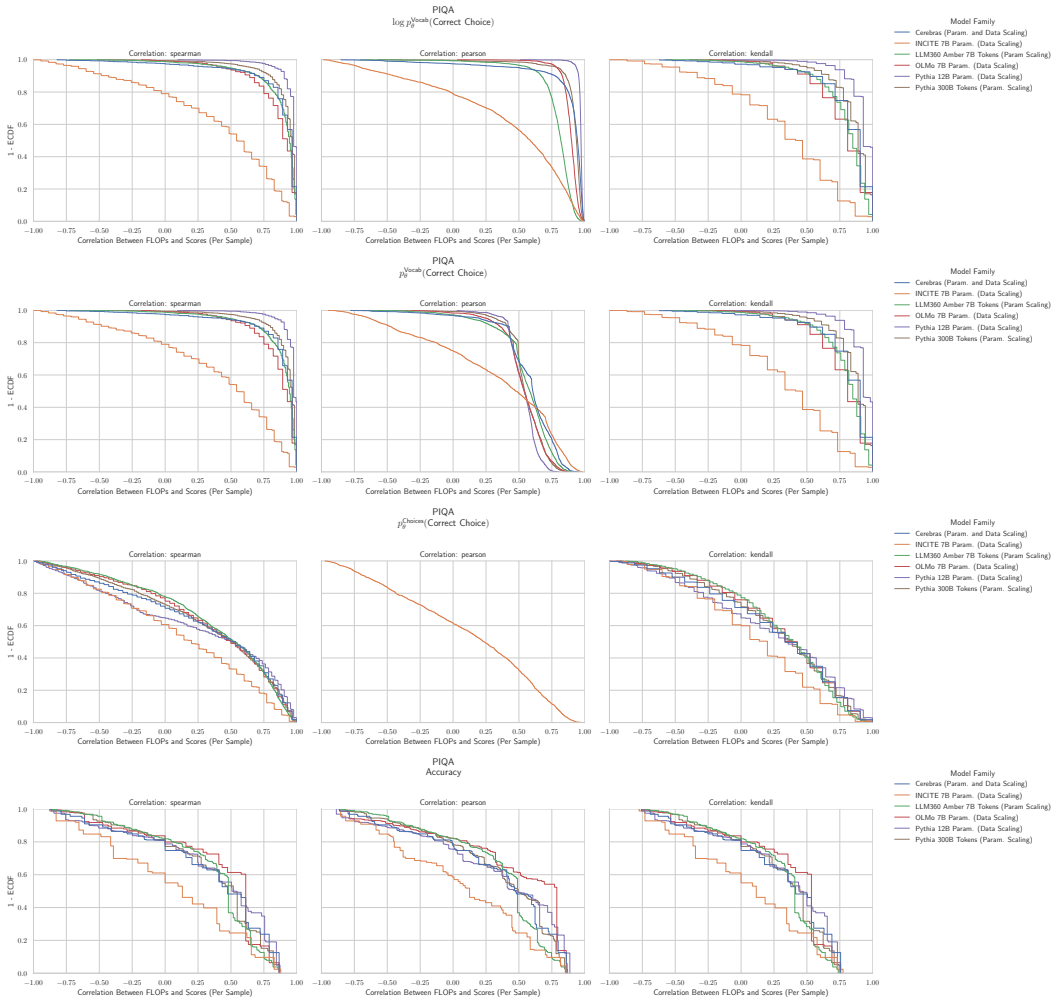

Figure 74: **PIQA: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.66 NLP BENCHMARK: RACE LAI ET AL. (2017)

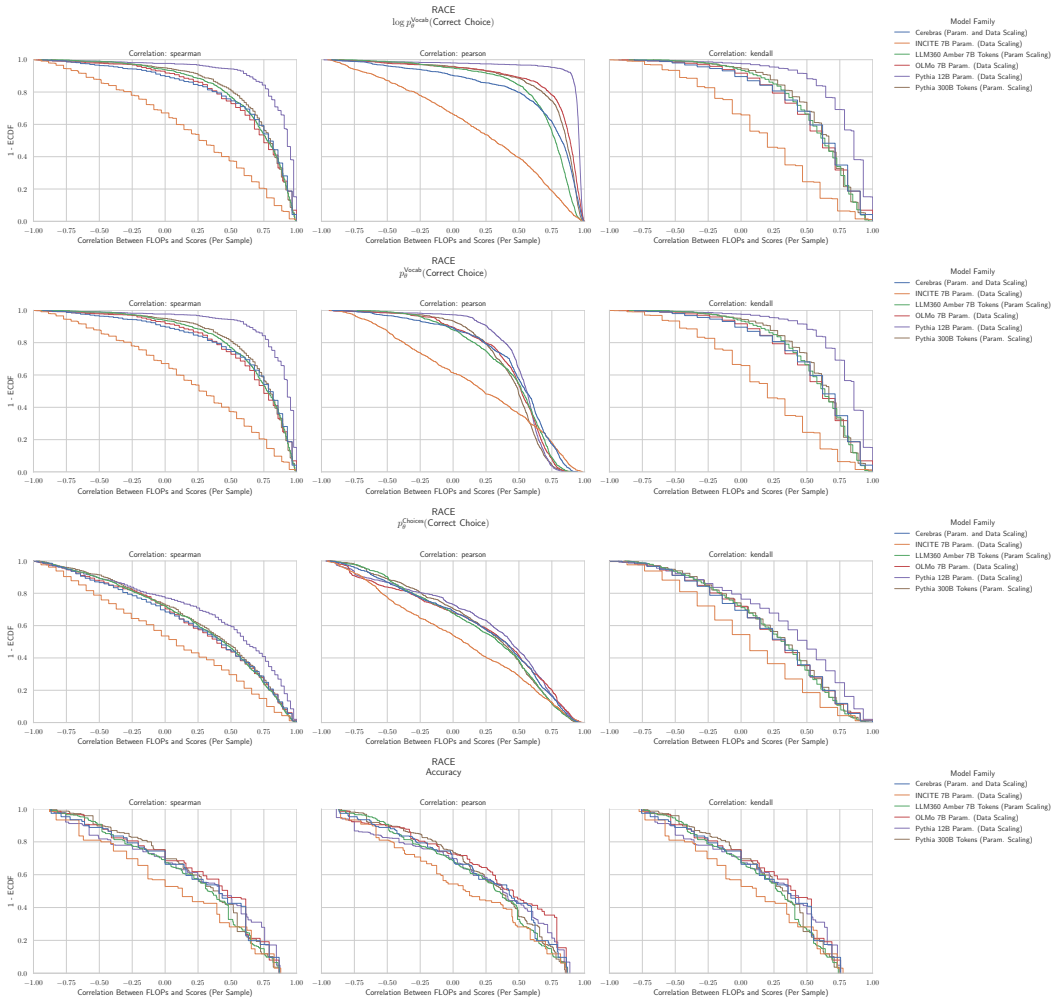

Figure 75: **RACE: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.67 NLP BENCHMARK: SCIQ WELBL ET AL. (2017)

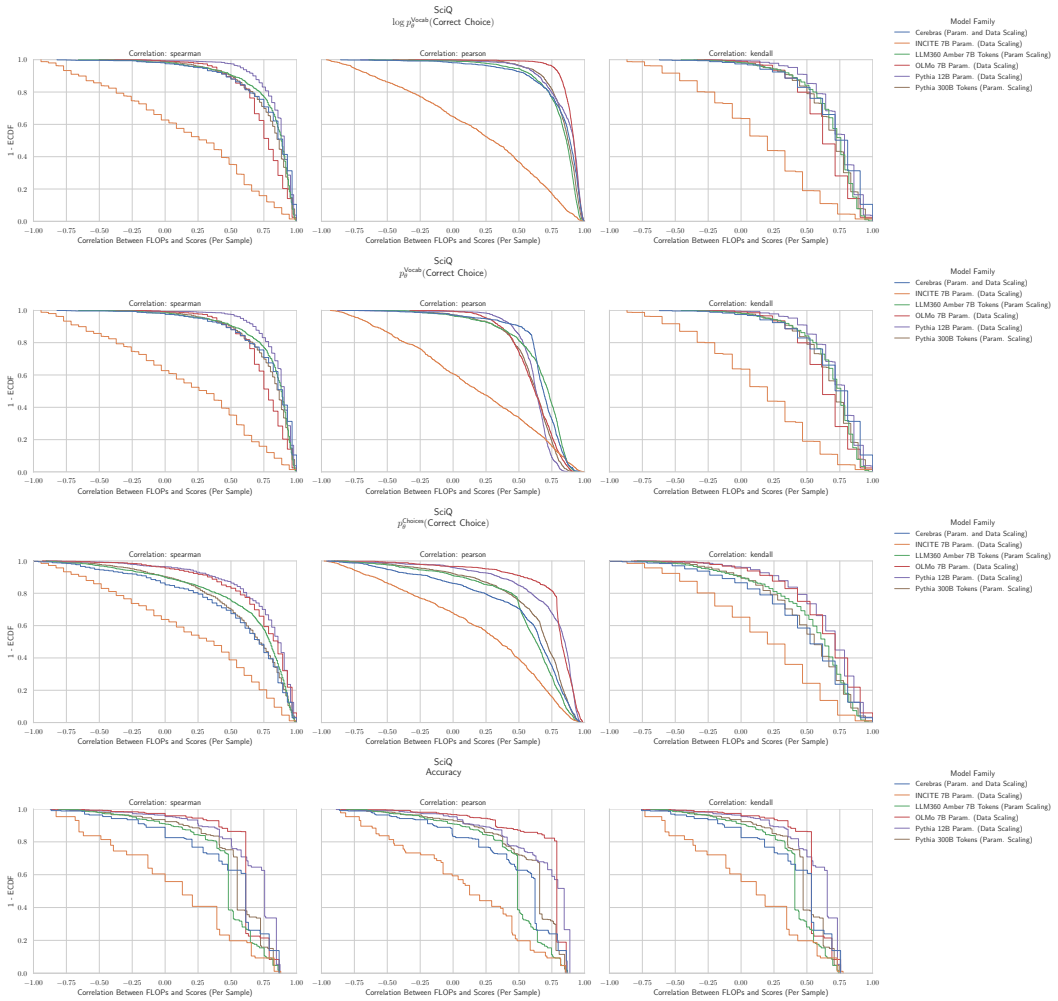

Figure 76: **SciQ: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.68 NLP BENCHMARK: SOCIAL IQA SAP ET AL. (2019B)

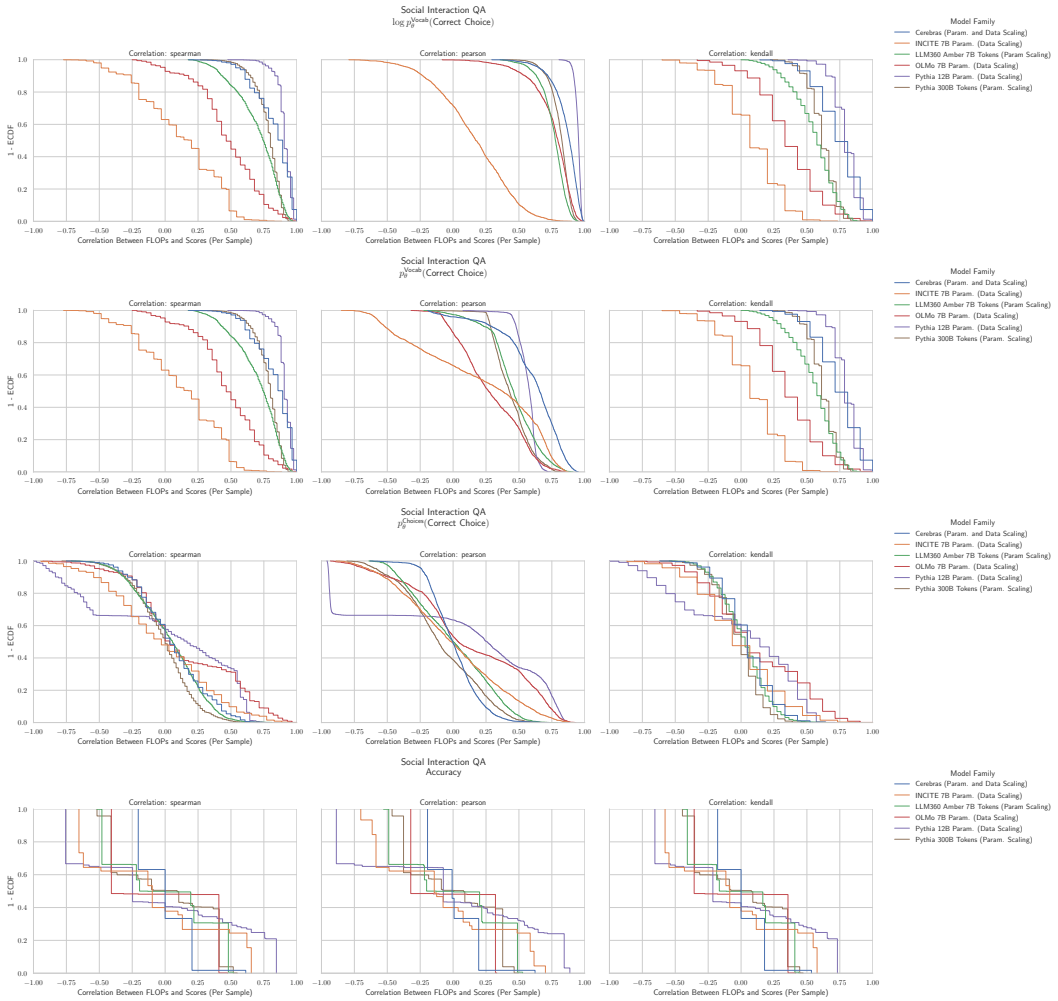

Figure 77: **Social IQA: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.69 NLP BENCHMARK: WINOGRANDE KEISUKE ET AL. (2019)

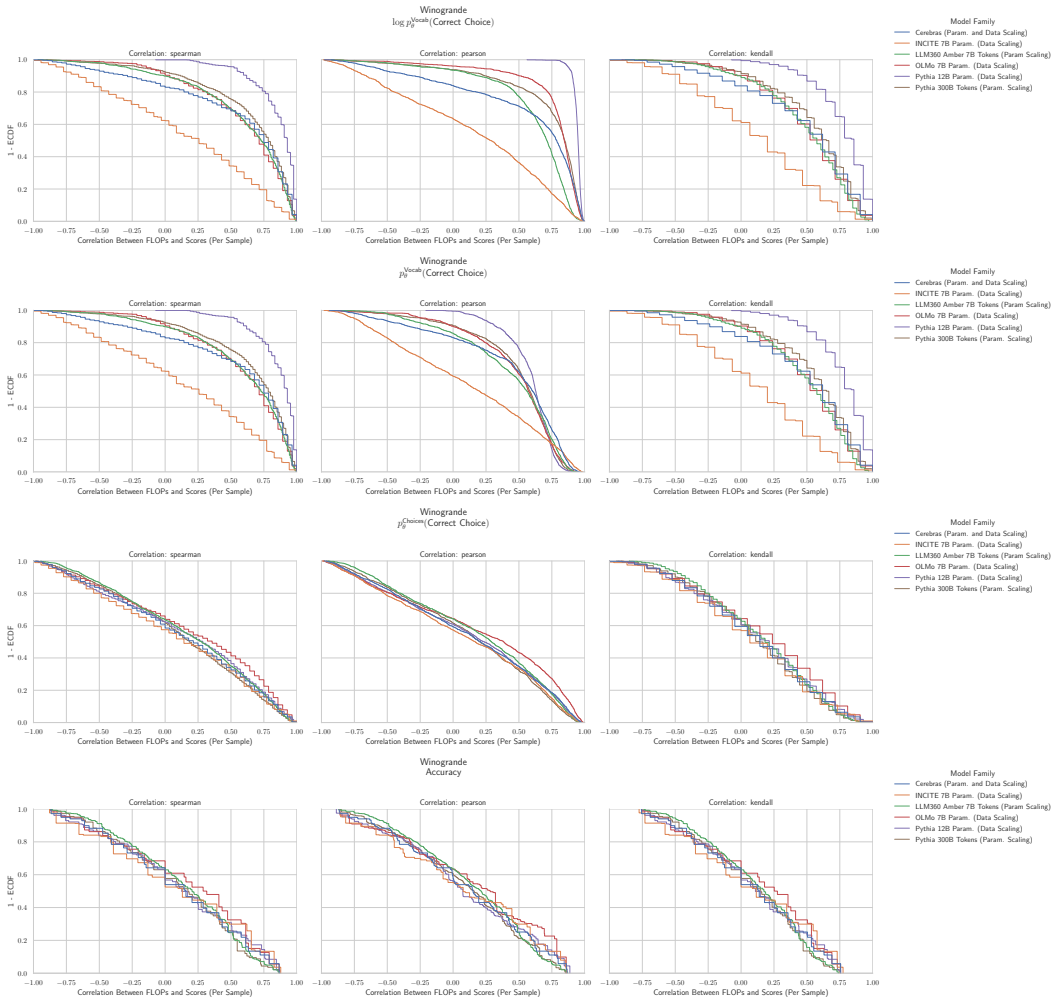

Figure 78: **Social IQA: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

G.70 NLP BENCHMARK: XWINOGRAD ENGLISH MUENNIGHOFF ET AL. (2023)

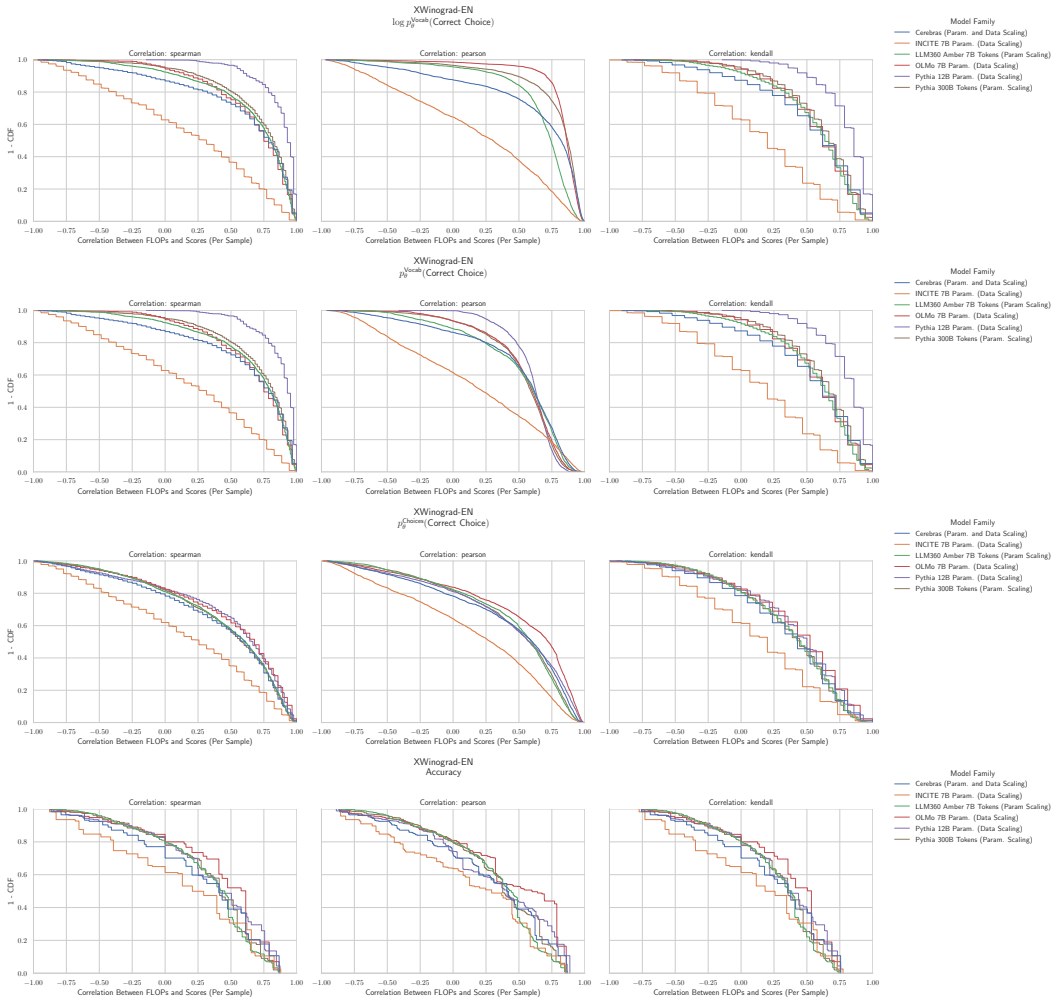

Figure 79: **XWinograd English: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**