Why Has Predicting Downstream Capabilities of Frontier AI Models with Scale Remained Elusive?

Anonymous Authors[1]

Abstract

Predictable behavior from scaling advanced AI systems is an extremely desirable property. While a well-established literature exists on how pretraining performance scales, the literature on how particular downstream capabilities change with scale is significantly muddier. For instance, previous papers debated the origins of emergent abilities, and more recent work claimed that specific downstream capabilities become predictable only beyond a specific pretraining loss or if aggregated across dozens of benchmarks. In this work, we take a step back and ask: *what makes predicting specific downstream capabilities with scale difficult?* We identify a critical factor contributing to this difficulty on multiple-choice benchmarks. Using five model families and twelve widely-used benchmarks, we show that downstream performance is computed from negative log likelihoods via a sequence of transformations that progressively deteriorates the statistical relationship between performance and scale. We demonstrate that this deterioration is caused by metrics that require comparing the correct answer against a small number of specific incorrect answers, meaning that accurately predicting downstream capabilities requires predicting not just how probability mass concentrates on the correct behavior with scale, but also how probability mass changes on specific incorrect behaviors with scale. We empirically study how probability mass on the correct choice covaries with probability mass on incorrect choices with increasing compute, suggesting that scaling laws for *incorrect* choices might be achievable. Our work explains why pretraining scaling laws are regarded as more predictable and contribute towards establishing scaling-predictable evaluations of frontier AI models.

1. Introduction

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 their capabilities and informing key decisions around their development and deployment (Anthropic, 2023; OpenAI, 2023; Dragan et al., 2024). While scaling laws describing relationships between 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 concerning 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, with further work suggesting that such unpredictable changes might at times be artifacts of researchers' analyses, i.e., choices of metrics and lack of resolution (Srivastava et al., 2022; Schaeffer et al., 2023; Hu et al., 2024). More recently, Du et al. (2024) claimed that downstream capabilities *can* be predicted, but *only* after the pretraining cross-entropy loss falls below a certain threshold, and Gadre et al. (2024) claimed 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 take a step back and ask: *why has predicting specific downstream capabilities with scale remained elusive?*

While many factors are certainly responsible, we identify a new factor that makes modeling the scaling behavior on widely used multiple-choice question-answering benchmarks challenging. We demonstrate that common multiple-choice scores from raw model outputs (negative log probabilities) via a sequence of transformations that progressively degrade the statistical relationship between those outputs and scaling parameters. The cause is that these metrics rely

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the ICML NextGenAISafety 2024 Workshop. Do not distribute.

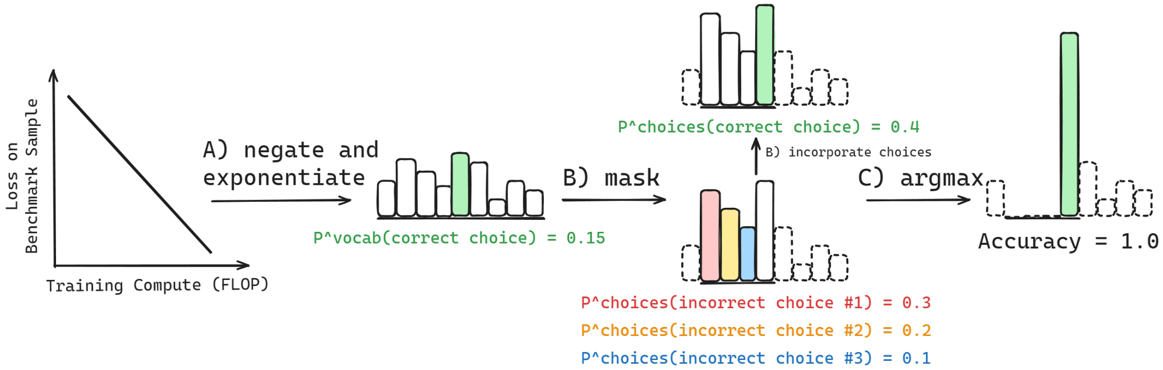
Loss on Benchmark Sample
Training Compute (FLOP)
A) negate and exponentiate
P^vocab(correct choice) = 0.15
B) mask
P^choices(correct choice) = 0.4
B) incorporate choices
P^choices(incorrect choice #1) = 0.3
P^choices(incorrect choice #2) = 0.2
P^choices(incorrect choice #3) = 0.1
C) argmax
Accuracy = 1.0

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**).

on a direct comparison between the ground truth output and a small set of specific incorrect outputs. As a result, *accurately predicting downstream performance requires modeling not only the concentration of probability mass on the correct output with increasing compute, but also modeling the fluctuations of probability mass on particular incorrect alternatives*, which (to-date) is a necessary but unaddressed step. We then empirically study how probability mass on incorrect choices fluctuates with increasing compute. Our findings help explain the apparent unpredictability of individual downstream metrics; more broadly, we argue that a precise understanding of the factors affecting downstream performance is essential for designing evaluations that can reliably track the progression of frontier AI capabilities.

2. Methodology: Data for Studying Scaling of Downstream Capabilities

To study how model families' downstream capabilities on specific tasks change with scale, 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 (LM) Evaluation Harness (Gao et al., 2023) rather than implementing our own evaluations. **Model Families** Because our goal was 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: Pythia (Biderman et al., 2023a), Cerebras-GPT (Dey et al., 2023), OLMo (Groeneveld et al., 2024), INCITE (AI, 2023) and LLM360 (Liu et al., 2023). For details, see App. D. **NLP Benchmarks** We evaluated the above model families on widely-used multiple-choice benchmarks: AI2 Reasoning Challenge (ARC) Easy and Hard (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MathQA (Amini et al., 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 common metrics for multiple choice benchmarks: Accuracy and probability mass on the correct choice relative to the available choices. **Compute Approximations** Following prior work, we approximated pretraining compute C (in terms of training FLOP) of a given model checkpoint as a function of the parameter count excluding the embedding layer N and the amount of training data seen in tokens D: $C = C(N, D) \approx 6\, N\, D$.

3. What Makes Predicting Downstream Performance Difficult?

Performance on multiple choice benchmarks is commonly presented as `Accuracy` or probability mass on the correct choice out of the available choices. These quantities are computed via a sequence of transformations that begins with the negative log-likelihood of the correct choice on this particular sample as some function $f(\cdot, \cdot)$ of compute:

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

Two details are critical. Firstly, this negative log-likelihood is specific to this particular sample in the benchmark. *All the scores we discuss are per-sample.* Secondly, this negative log-likelihood is computed over the vocabulary of the model. One can then compute the probability mass of the correct

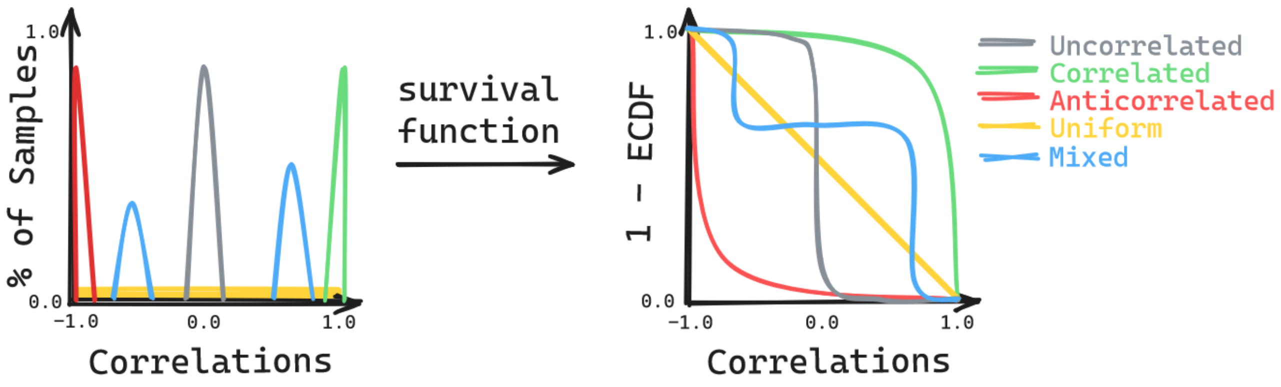

Figure 2. **Schematic distributions of compute-score correlations and their corresponding survival functions. Left:** For each benchmark, model family, performance metric and correlation metric, one can compute how correlated scores are with compute. This yields a distribution (over samples) of score-compute correlations. Note: the uniform (yellow) distribution is small but non-zero everywhere. **Right:** To easily extract what fraction of samples in a benchmark have score-compute correlations above any given threshold, we convert the probability distributions to *survival functions*, defined as 1 minus the empirical cumulative distribution function (ECDF).

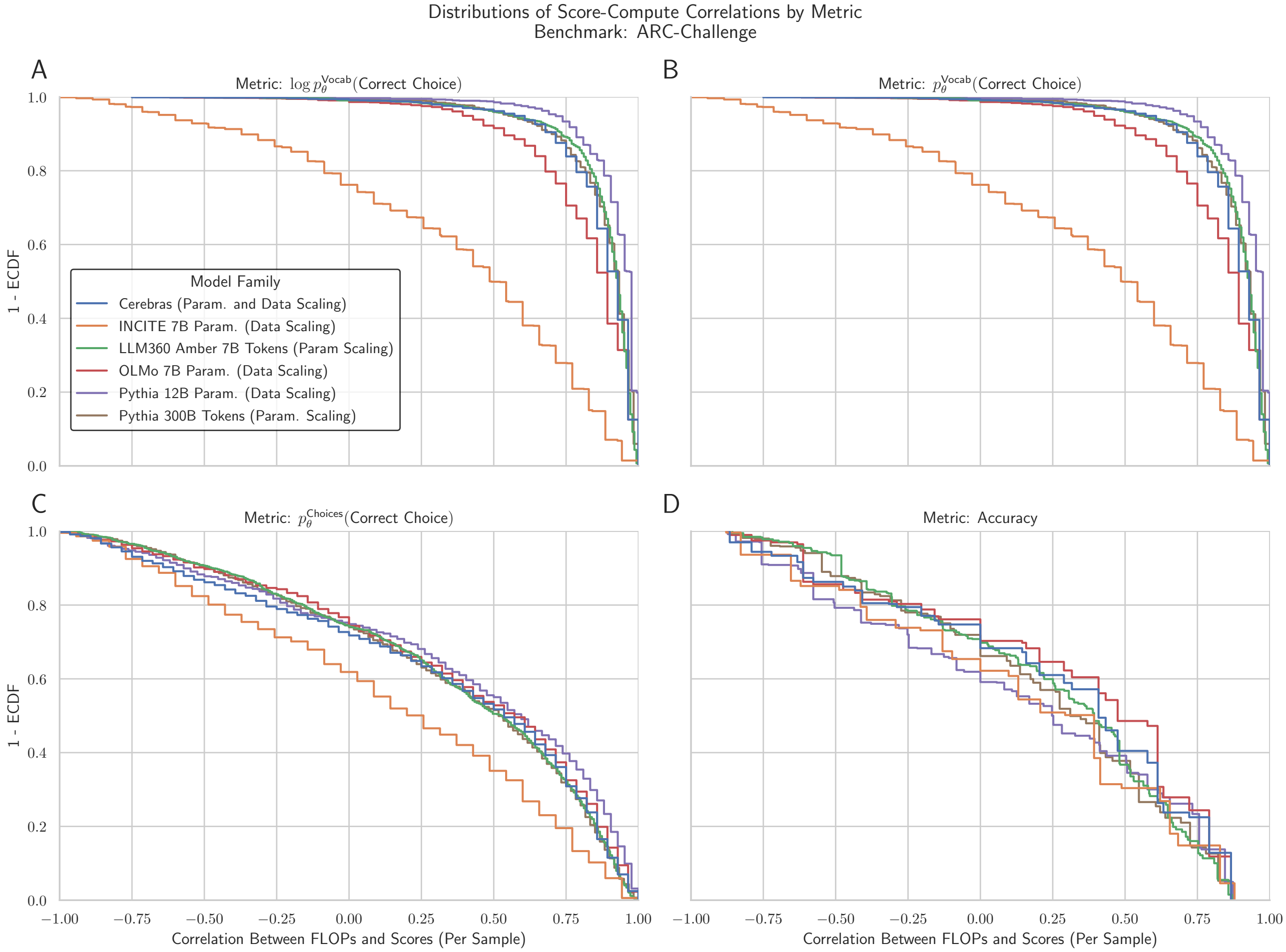

Figure 3. **Multiple-choice benchmark `Accuracy` is computed via a sequence of transformations that deteriorate correlations between performance scores and pretraining compute. (A)** Compute and scores $\log p_\theta^{\text{Vocab}}$(Correct Choice) begin highly correlated (shown: Spearman correlation). **(B)** Transforming $\log p_\theta^{\text{Vocab}}$(Correct Choice) into p_θ^{Vocab}(Correct Choice) preserves the high score-compute correlations. **(C)** Transforming p_θ^{Vocab}(Correct Choice) into $p_\theta^{\text{Choices}}$(Correct Choice) decorrelates scores from compute. **(D)** Transforming $p_\theta^{\text{Choices}}$(Correct Choice) into `Accuracy` further decorrelates scores from compute. Example from ARC Challenge (Clark et al., 2018). Results are consistent across NLP benchmarks and all three correlation metrics; for more, see App. I.

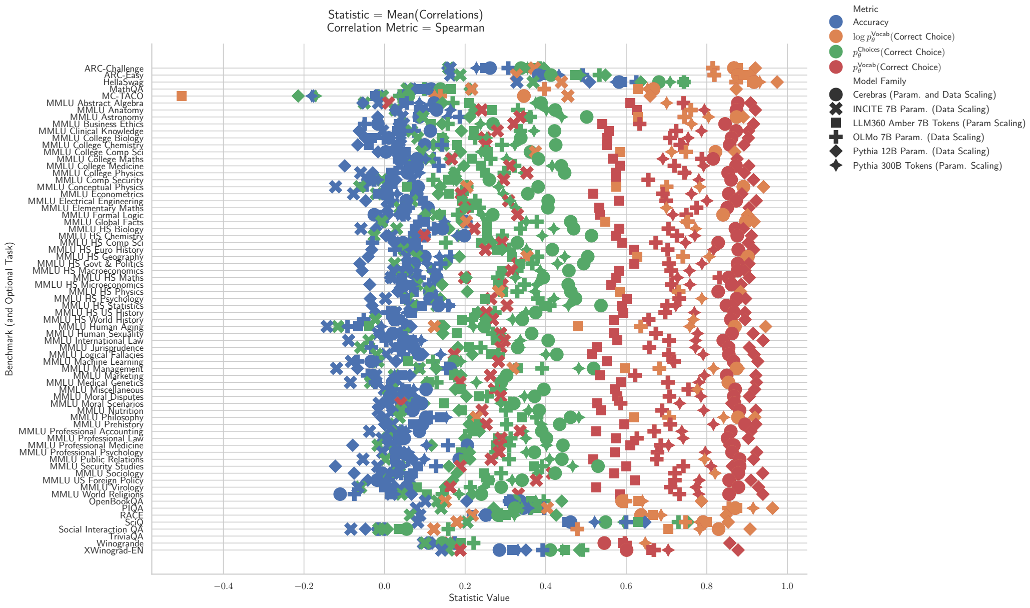

Figure 4. **All four statistics of score-compute correlation distributions demonstrate that transforming $\log p_\theta^{\text{Vocab}}$(Correct Choice) → p_θ^{Vocab}(Correct Choice) → $p_\theta^{\text{Choices}}$(Correct Choice) → `Accuracy` causes score-compute correlations to deteriorate.** We find a consistent trend 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) that the sequence of transformations degrades score-compute correlations, as shown by the right-to-left red-to-orange-to-green-to-blue vertical stripes. See App. Figs. 7,8,9 for other statistics and other correlation metrics.

choice, again with respect to the vocabulary:

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

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:

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

Finally, one uses the choices-normalized probability masses to compute accuracy:

$$\texttt{Acc}_\theta \overset{\text{def}}{=} \mathbb{1}\left(\text{Correct} = \arg\max_i \left\{p_\theta^{\text{Choices}}(\text{Available}_i)\right\}\right)$$

where $\mathbb{1}(\cdot)$ is an indicator variable.

To quantify how this sequence of transformations affects predictability, we measured how correlated per-sample scores are with pretraining compute, and then studied how the distribution (over samples) of correlation values shifted as one transitions from loglikelihoods to p_θ^{Vocab}(Correct Choice) to $p_\theta^{\text{Choices}}$(Correct Choice) to `Accuracy`. 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). Pearson, Kendall (1938) and Spearman (1961)) yielded consistent results. We present ARC Challenge (Clark et al., 2018) as an example, but note that all other benchmarks exhibited similar patterns (App. I). We visualized the distributions via their *complementary empirical cumulative distribution functions* (App. B; Fig. 2). For a given correlation value, e.g., 0.5, the survival function tells us what fraction of the benchmark's samples have score-compute correlations greater than the given value (Fig. 3A). Beginning with log likelihoods, approximately 90% of samples exhibit a score-compute correlation > 0.75, regardless of the model family (Fig. 3A). Transforming negative log-likelihoods into probability masses with respect to p_θ^{Vocab}(Correct Choice) does not affect the distribution of score-compute correlations for Spearman and Kendall correlations (Fig. 3B). However, transforming p_θ^{Vocab}(Correct Choice) into $p_\theta^{\text{Choices}}$(Correct Choice) causes a decrease in the distribution of score-compute correlations (Fig. 3C), with only 40% of samples having score-compute

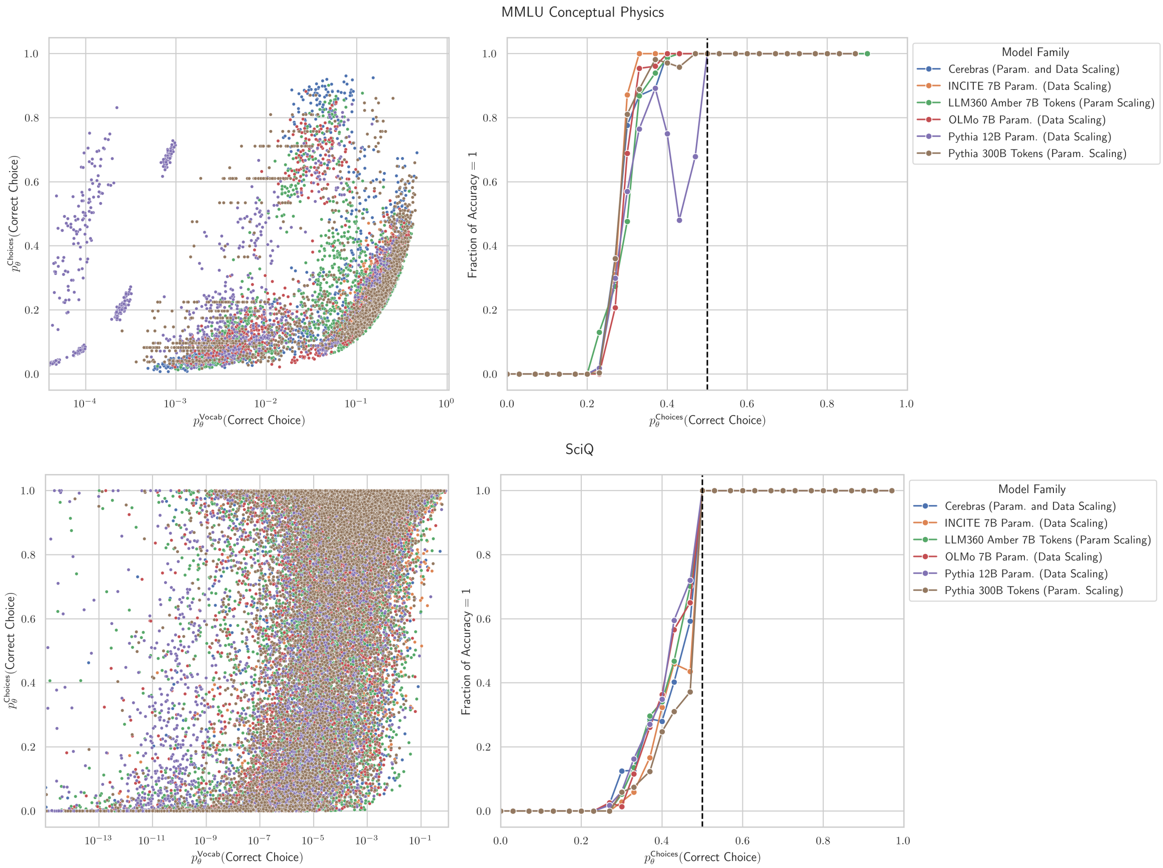

Figure 5. **Predictability deteriorates because of probability mass fluctuating on specific incorrect choices with scale. Left:** Transitioning from $p_\theta^{\text{Vocab}}(\text{Correct Choice})$ to $p_\theta^{\text{Choices}}(\text{Correct Choice})$ demonstrates that $p_\theta^{\text{Vocab}}(\text{Correct Choice})$ 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. **Right:** While $p_\theta^{\text{Choices}}(\text{Correct Choice}) > 0.5$ must yield 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. Two example benchmarks shown: MMLU Conceptual Physics (Hendrycks et al., 2020), SciQ (Welbl et al., 2017).

correlations > 0.75. Transforming $p_\theta^{\text{Choices}}(\text{Correct Choice})$ into `Accuracy` (Fig. 3D) furthers that decrease; only 20% of samples hold correlations > 0.75. To quantitatively test whether these transformations indeed decrease the correlation between scores and compute, we measured four statistics of these score-compute correlation distributions: the mean, the median, the area under the survival function and the negative of the minimum of two Wasserstein distances to ideal (anti) correlated distribtuions. Across all four summary statistics, for all benchmarks and for all model families, we discovered a consistent ordering among metrics of the score-compute correlation distributions (Fig. 4):

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

4. Unknown Incorrect-Choice Probabilities Produce Unpredictability

What is the mechanism that causes this deterioration of correlations between scores and compute? Metrics with degraded correlations - depend not just on how the model's probability mass concentrates on the correct choice as compute increases, but also depend on how the model's probability mass fluctuates on incorrect available choicesas compute increases. To demonstrate how drastically the probability mass placed on incorrect choices can alter performance, we visualized how $p_\theta^{Vocab}(\text{Correct Choice})$ relates to metrics that rely on probabilities assigned to each incorrect choice (Fig. 5). Once performance is evaluated using a metric which is a function of the incorrect choices, i.e. for $p_\theta^{\text{Choices}}(\text{Correct Choice})$ and for `Accuracy`), 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).

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

AI, T. 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.

Amini, A., Gabriel, S., Lin, S., Koncel-Kedziorski, R., Choi, Y., and Hajishirzi, H. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In Burstein, J., Doran, C., and Solorio, T. (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.

Beeching, E., Fourrier, C., Habib, N., Han, S., Lambert, N., Rajani, N., Sanseviero, O., Tunstall, L., and Wolf, T. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.

Besiroglu, T., Erdil, E., Barnett, M., and You, J. Chinchilla scaling: A replication attempt, 2024.

Biderman, S., Prashanth, U. S., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., and Raff, E. Emergent and predictable memorization in large language models, 2023a.

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and van der Wal, O. Pythia: A suite for analyzing large language models across training and scaling, 2023b.

Biderman, S., Schoelkopf, H., Sutawika, L., Gao, L., Tow, J., Abbasi, B., Aji, A. F., Ammanamanchi, P. S., Black, S., Clive, J., DiPofi, A., Etxaniz, J., Fattori, B., Forde, J. Z., Foster, C., Jaiswal, M., Lee, W. Y., Li, H., Lovering, C., Muennighoff, N., Pavlick, E., Phang, J., Skowron, A., Tan, S., Tang, X., Wang, K. A., Winata, G. I., Yvon, F., and Zou, A. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint*, 2024.

Bisk, Y., Zellers, R., Bras, R. L., Gao, J., and Choi, Y. Piqa: Reasoning about physical commonsense in natural language. 2020.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Clark, A., De Las Casas, D., Guy, A., Mensch, A., Paganini, M., Hoffmann, J., Damoc, B., Hechtman, B., Cai, T., Borgeaud, S., et al. Unified scaling laws for routed language models. In *International Conference on Machine Learning*, pp. 4057–4086. PMLR, 2022.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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

DeepSeek-AI, :, Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., Gao, H., Gao, K., Gao, W., Ge, R., Guan, K., Guo, D., Guo, J., Hao, G., Hao, Z., He, Y., Hu, W., Huang, P., Li, E., Li, G., Li, J., Li, Y., Li, Y. K., Liang, W., Lin, F., Liu, A. X., Liu, B., Liu, W., Liu, X., Liu, X., Liu, Y., Lu, H., Lu, S., Luo, F., Ma, S., Nie, X., Pei, T., Piao, Y., Qiu, J., Qu, H., Ren, T., Ren, Z., Ruan, C., Sha, Z., Shao, Z., Song, J., Su, X., Sun, J., Sun, Y., Tang, M., Wang, B., Wang, P., Wang, S., Wang, Y., Wang, Y., Wu, T., Wu, Y., Xie, X., Xie, Z., Xie, Z., Xiong, Y., Xu, H., Xu, R. X., Xu, Y., Yang, D., You, Y., Yu, S., Yu, X., Zhang, B., Zhang, H., Zhang, L., Zhang, L., Zhang, M., Zhang, M., Zhang, W., Zhang, Y., Zhao, C., Zhao, Y., Zhou, S., Zhou, S., Zhu, Q., and Zou, Y. Deepseek llm: Scaling open-source language models with longtermism, 2024.

Dey, N., Gosal, G., Zhiming, Chen, Khachane, H., Marshall, W., Pathria, R., Tom, M., and Hestness, J. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster, 2023.

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

Du, Z., Zeng, A., Dong, Y., and Tang, J. Understanding emergent abilities of language models from the loss perspective, 2024.

Gadre, S. Y., Smyrnis, G., Shankar, V., Gururangan, S., Wortsman, M., Shao, R., Mercat, J., Fang, A., Li, J., Keh, S., Xin, R., Nezhurina, M., Vasiljevic, I., Jitsev, J., Dimakis, A. G., Ilharco, G., Song, S., Kollar, T., Carmon, Y., Dave, A., Heckel, R., Muennighoff, N., and Schmidt, L. Language models scale reliably with over-training and on downstream tasks, 2024.

Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., Conerly, T., Dassarma, N., Drain, D., Elhage, N., et al. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1747–1764, 2022.

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization, 2022.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Gordon, M. A., Duh, K., and Kaplan, J. 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.

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K. R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M. E., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N. A., and Hajishirzi, H. Olmo: Accelerating the science of language models, 2024.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. 2021.

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Hernandez, D., Kaplan, J., Henighan, T., and McCandlish, S. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Hernandez, D., Brown, T., Conerly, T., DasSarma, N., Drain, D., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Henighan, T., Hume, T., et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Hu, S., Liu, X., Han, X., Zhang, X., He, C., Zhao, W., Lin, Y., Ding, N., Ou, Z., Zeng, G., Liu, Z., and Sun, M. Predicting emergent abilities with infinite resolution evaluation, 2024.

Huang, Y., Zhang, J., Shan, Z., and He, J. Compression represents intelligence linearly, 2024.

Isik, B., Ponomareva, N., Hazimeh, H., Paparas, D., Vassilvitskii, S., and Koyejo, S. Scaling laws for downstream task performance of large language models, 2024.

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

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints*, art. arXiv:1705.03551, 2017.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Keisuke, S., Ronan, L., Chandra, B., and Yejin, C. Winogrande: An adversarial winograd schema challenge at scale. 2019.

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

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

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., Lee, K., Toutanova, K. N., Jones, L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. 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.

Liu, Z., Qiao, A., Neiswanger, W., Wang, H., Tan, B., Tao, T., Li, J., Wang, Y., Sun, S., Pangarkar, O., Fan, R., Gu, Y., Miller, V., Zhuang, Y., He, G., Li, H., Koto, F., Tang, L., Ranjan, N., Shen, Z., Ren, X., Iriondo, R., Mu, C., Hu, Z., Schulze, M., Nakov, P., Baldwin, T., and Xing, E. P. Llm360: Towards fully transparent open-source llms, 2023.

Lyu, C., Wu, M., and Aji, A. F. Beyond probabilities: Unveiling the misalignment in evaluating large language models. *arXiv preprint arXiv:2402.13887*, 2024.

Maloney, A., Roberts, D. A., and Sully, J. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

McCandlish, S., Kaplan, J., Amodei, D., and Team, O. D. An empirical model of large-batch training, 2018.

McKenzie, I., Lyzhov, A., Parrish, A., Prabhu, A., Mueller, A., Kim, N., Bowman, S., and Perez, E. The inverse scaling prize, 2022. URL https://github.com/inverse-scaling/prize.

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

Muckatira, S., Deshpande, V., Lialin, V., and Rumshisky, A. Emergent abilities in reduced-scale generative language models, 2024.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., and Raffel, C. Crosslingual generalization through multitask finetuning, 2023.

Muennighoff, N., Rush, A., Barak, B., Le Scao, T., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., and Raffel, C. A. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Neumann, O. and Gros, C. 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, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024.

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

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y., and Shavit, N. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations*, 2019.

Ruan, Y., Maddison, C. J., and Hashimoto, T. Observational scaling laws and the predictability of language model performance, 2024.

Sap, M., Rashkin, H., Chen, D., Le Bras, R., and Choi, Y. 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.

Sap, M., Rashkin, H., Chen, D., LeBras, R., and Choi, Y. Socialiqa: Commonsense reasoning about social interactions, 2019b.

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

Schaeffer, R., Miranda, B., and Koyejo, S. Are emergent abilities of large language models a mirage? In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (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.

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

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Welbl, J., Liu, N. F., and Gardner, M. Crowdsourcing multiple choice science questions. In Derczynski, L., Xu, W., Ritter, A., and Baldwin, T. (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.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Zhou, B., Khashabi, D., Ning, Q., and Roth, D. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th In-*

ternational 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 **??**.

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; Isik 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):

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

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) $\hat{S}_X(x)$:

$$\hat{S}_X(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} 1\{x_i > x\} = 1 - \hat{F}_X(x) \tag{2}$$

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. Scaling Behavior of Probability Mass on Incorrect Choices

In order to accurately predict performance on multiple-choice question-answering benchmarks, one must predict not just how probability mass concentrates on correct choices with scale, but how probability mass also fluctuates on incorrect choices with scale. For metrics like `Accuracy`, these predictions *must* be made for each sample because knowing the average (across many samples) mass placed on incorrect choices says little about how much mass is placed on any single incorrect choice for a single sample. Achieving such a feat *might* be possible. We conduct a preliminary analysis how probability mass on correct choices and probability mass on incorrect choices for samples scale with compute (Fig. 6). Although preliminary, multiple benchmarks display strong positive relationships that suggest per-sample scaling laws predicting the probability mass assigned to each of the *incorrect* choices might be possible. We leave this challenge to future work.

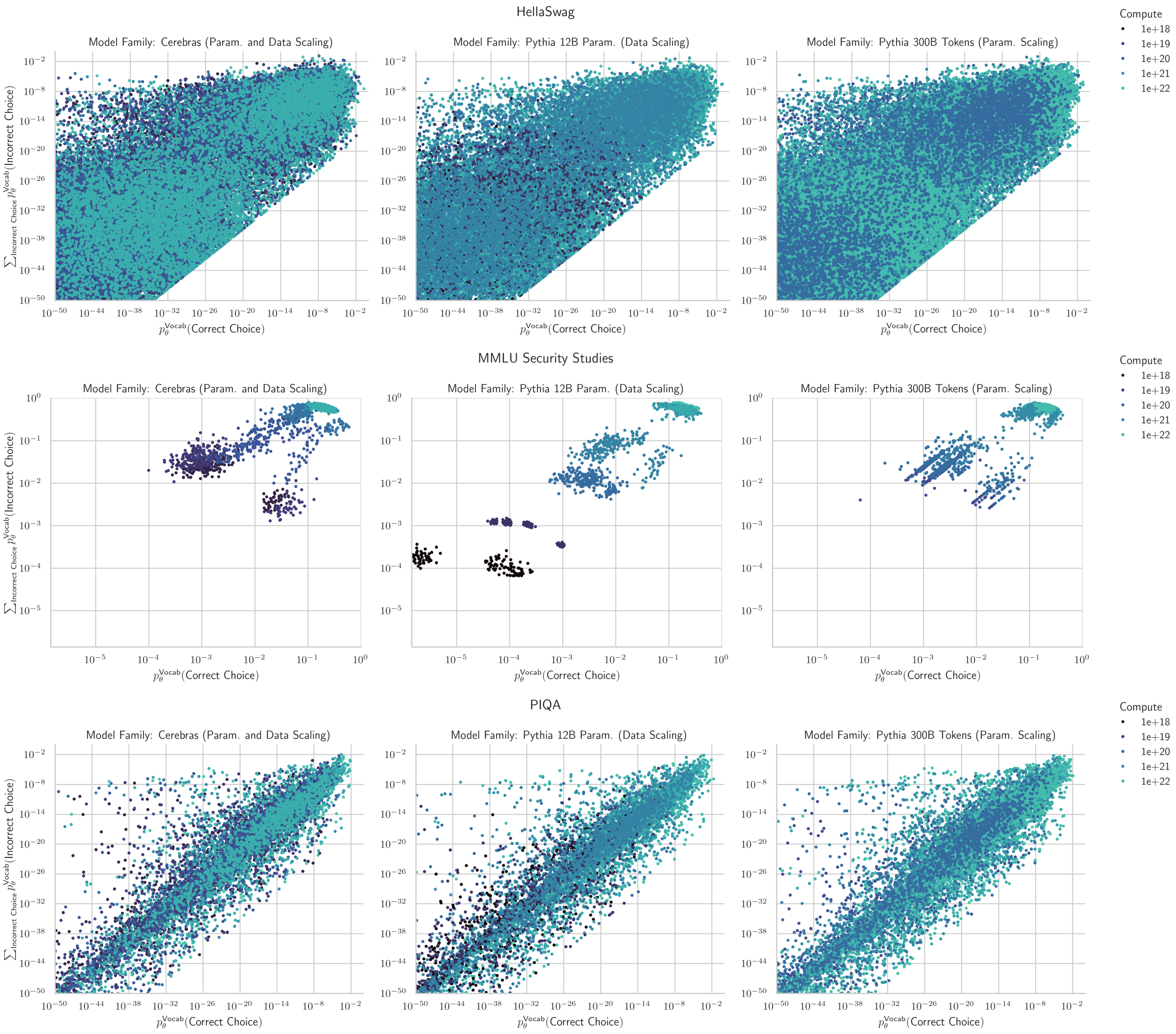

Figure 6. **Probability masses on correct versus incorrect are correlated but can fluctuate substantially.** We can observe clear linear correlations between the probability mass on correct choices and on incorrect choices. However, the spread is high: for any given value of p_θ^{Vocab}(Correct Choice), the probability mass across incorrect choices can vary by several orders of magnitude.

G. Discussion, Related Work and Future Directions

This work presents a mechanism that explains the (lack of) predictability of downstream task performance with compute. Our results have implications for the design of future evaluations for 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 For a treatise of related work, please see App. A.

Direction 1: Beyond Multiple Choice Benchmarks Our study is restricted to benchmarks evaluated via loglikelihood-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 indicates a reason for multiple-choice benchmark performance to not be 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 Owen (2024). A predictive model should be able to identify emergence points well-before scores rise on standard metrics like Accuracy.

H. Score-Compute Correlation Distributions' Statistics

H.1. Pearson Correlations

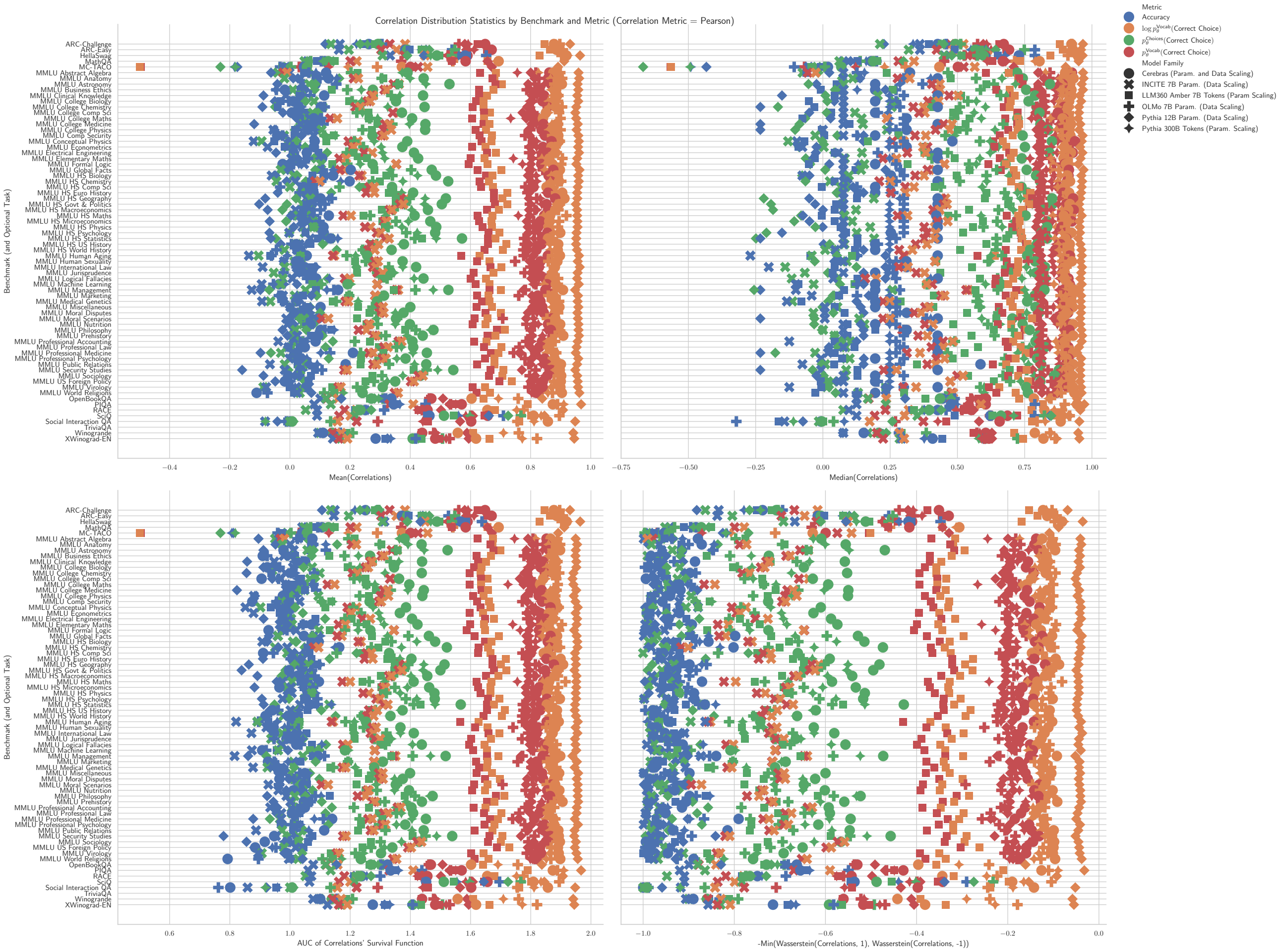

Figure 7. **Statistics for empirical distributions of correlations between scores and compute for all benchmarks and model families.** These correlation values were computed with Pearson correlation and are consistent with the main text's results computed with Spearman correlation (Fig. 4): The sequence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

H.2. Spearman Correlations

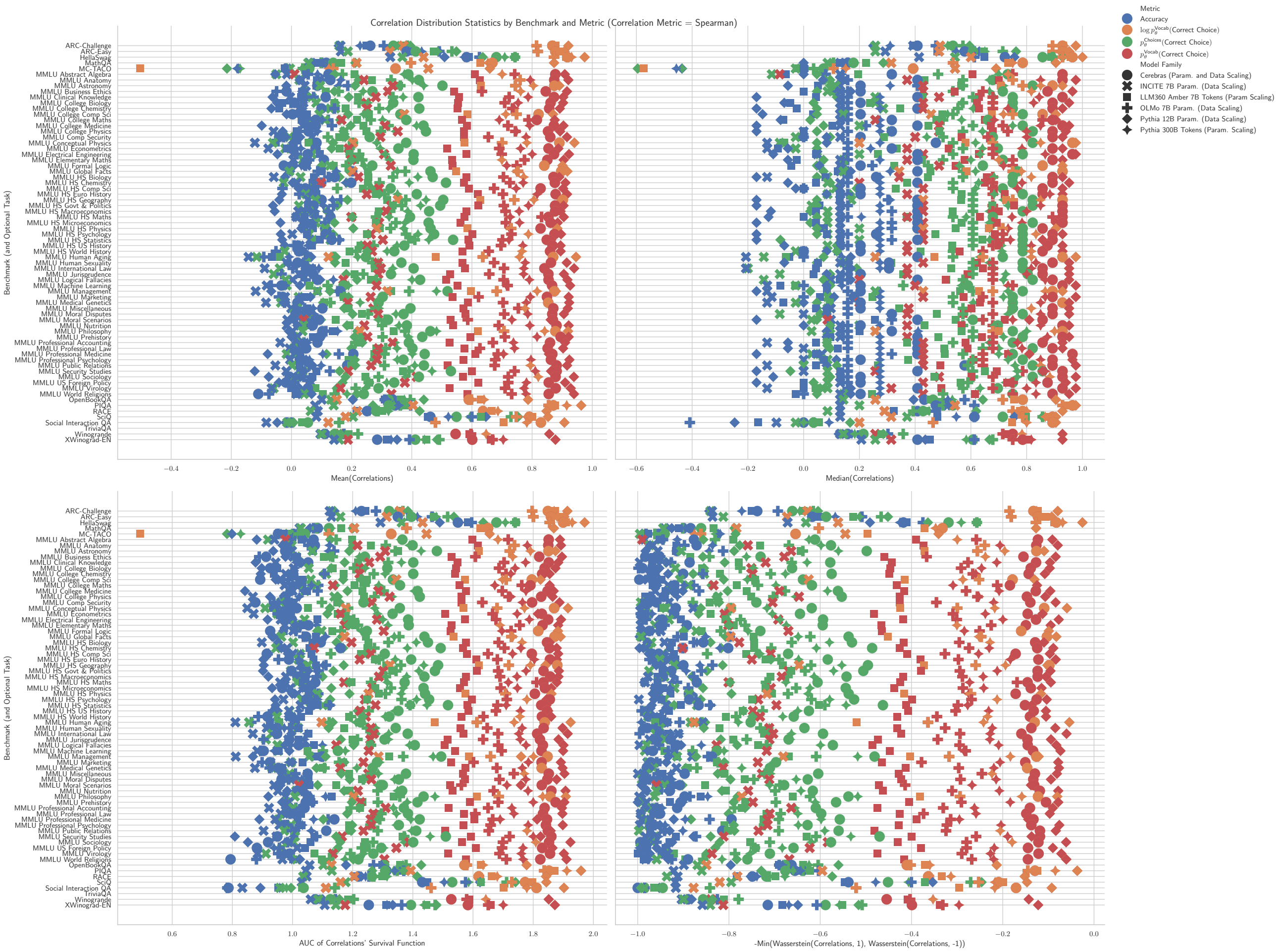

Figure 8. **Statistics for empirical distributions of correlations between scores and compute for all benchmarks and model families.** These correlation values were computed with Spearman correlation. The sequence of transformations from $\log p_\theta^{\text{Vocab}}$(Correct Choice) → p_θ^{Vocab}(Correct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

H.3. Kendall Correlations

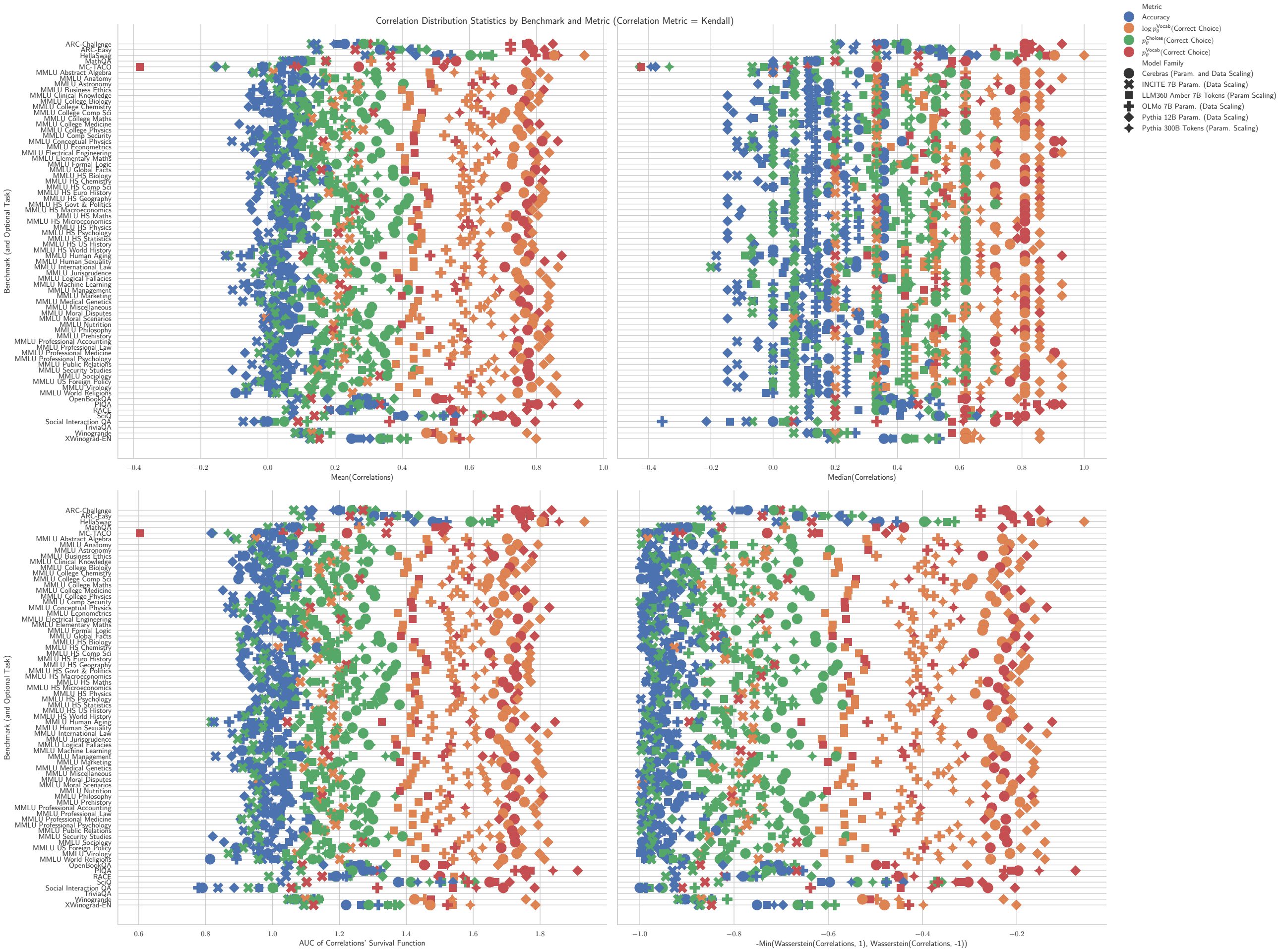

Figure 9. **Statistics for empirical distributions of correlations between scores and compute for all benchmarks and model families.** These correlation values were computed with Kendall correlation and are consistent with the main text's results computed with Spearman correlation (Fig. 4): The sequence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

I. Per-Benchmark Score-Compute Correlation Distributions

I.1. NLP Benchmark: ARC Challenge (Clark et al., 2018)

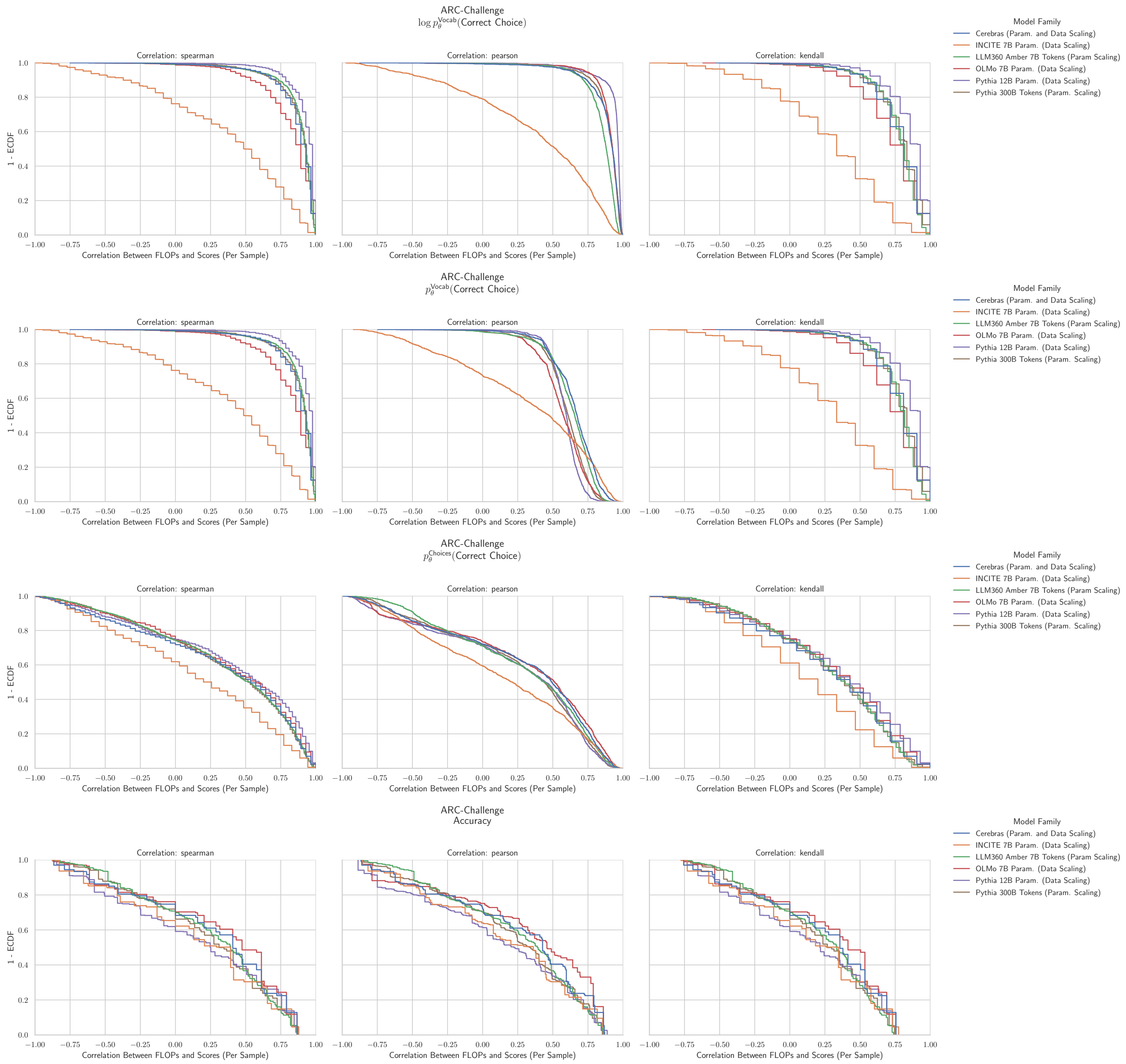

Figure 10. **ARC Challenge: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.2. NLP Benchmark: ARC Easy (Clark et al., 2018)

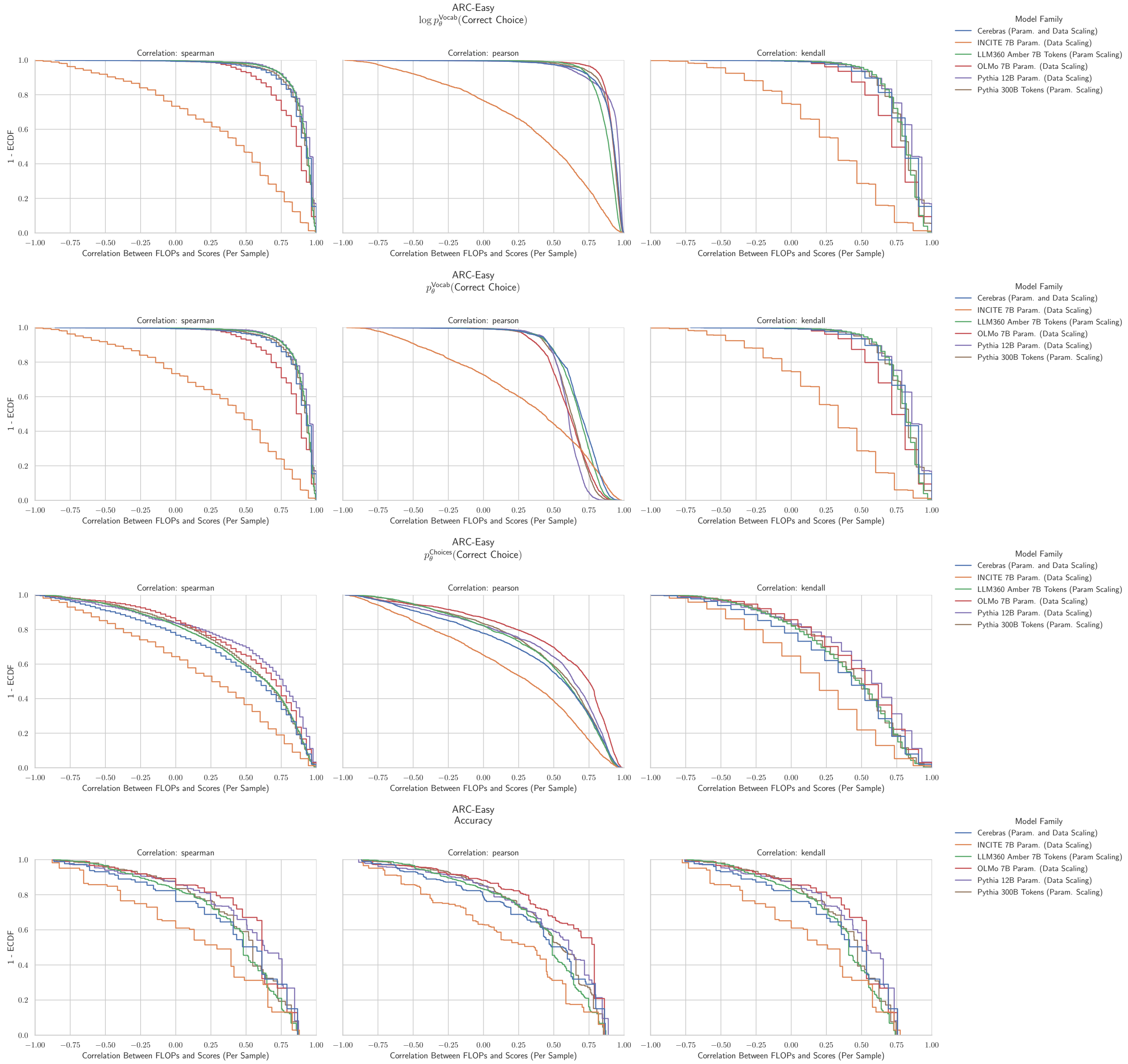

Figure 11. **ARC Easy: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.3. NLP Benchmark: HellaSwag (Zellers et al., 2019)

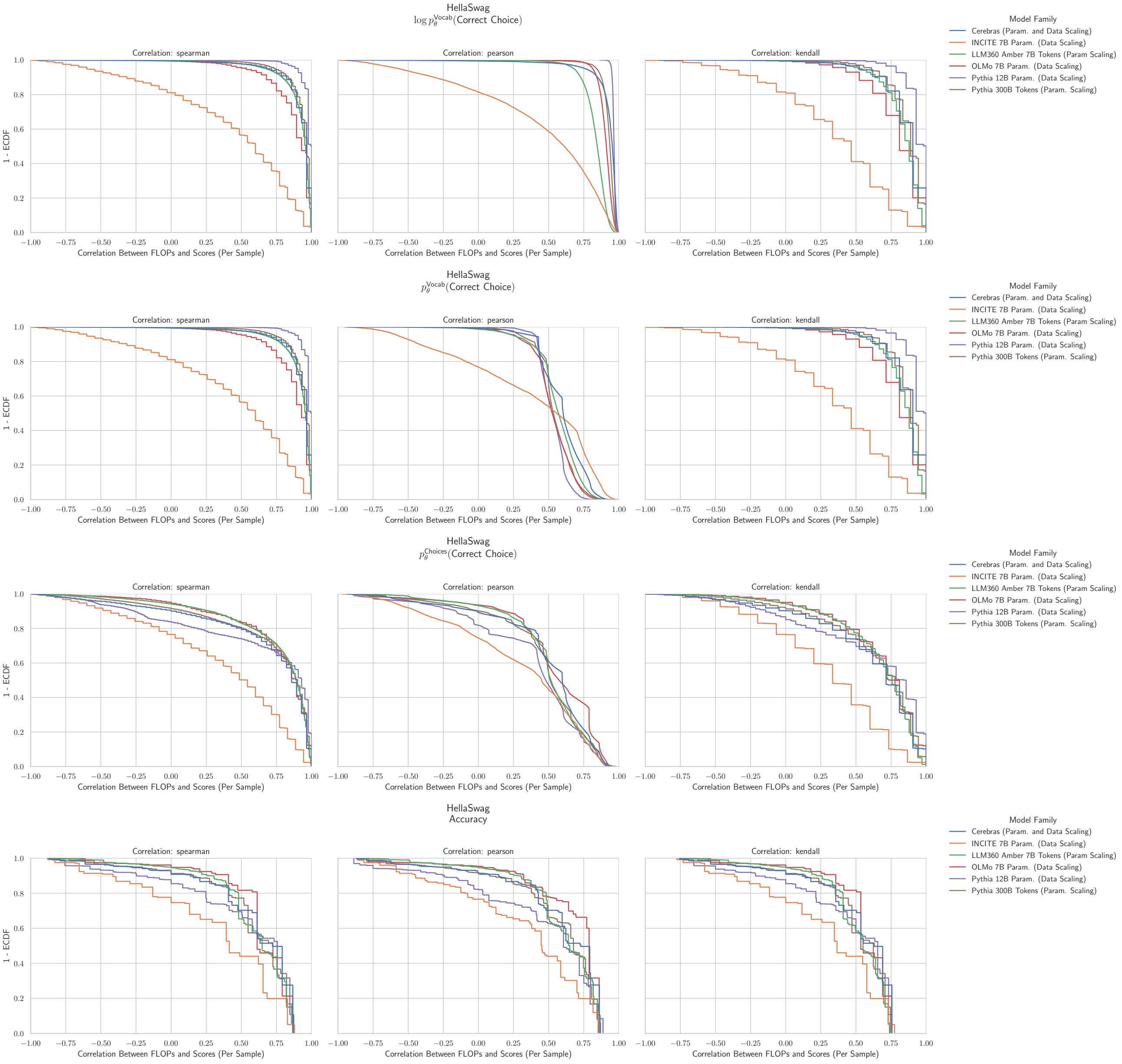

Figure 12. **HellaSwag: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.4. NLP Benchmark: MathQA (Amini et al., 2019)

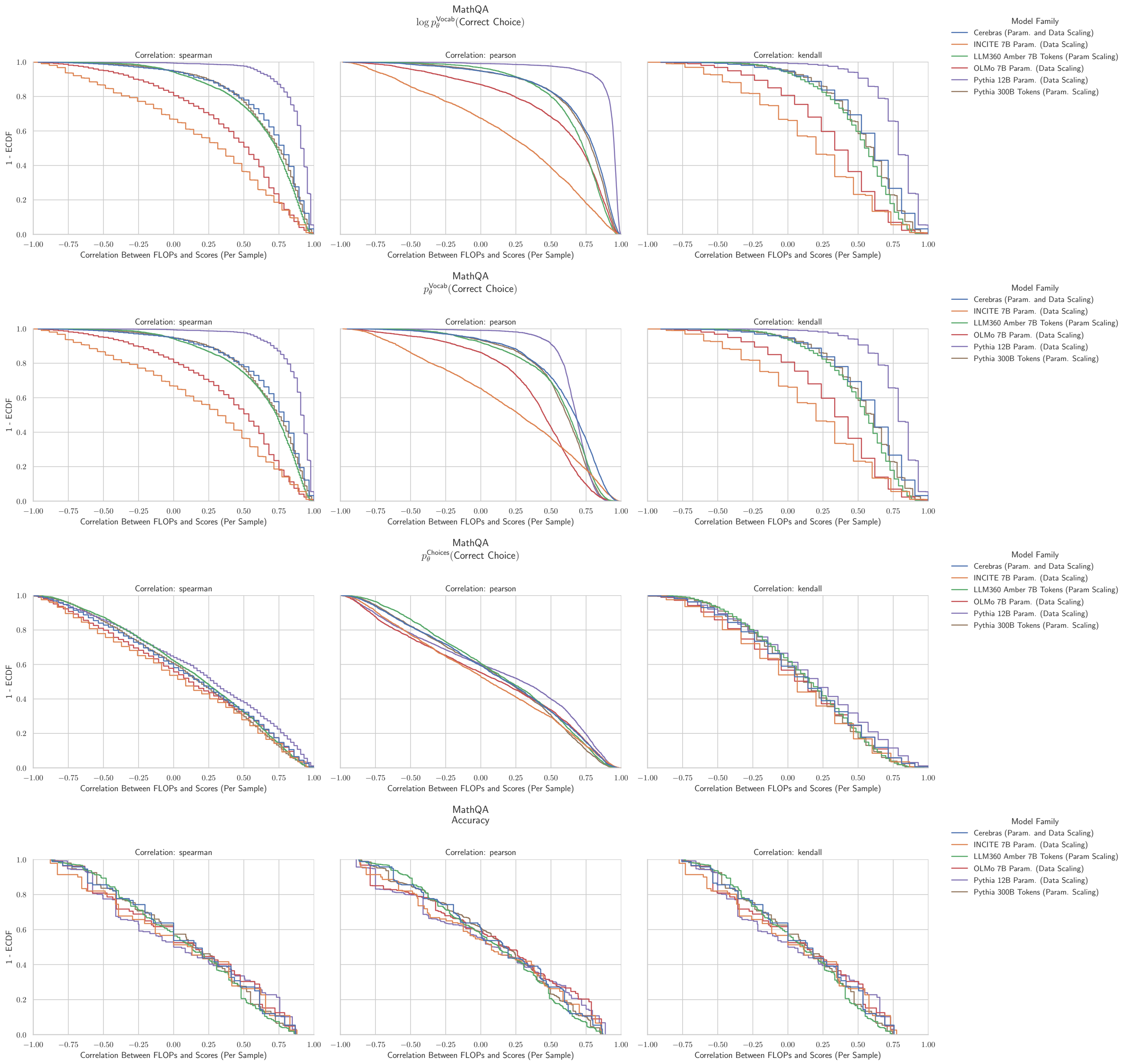

Figure 13. **HellaSwag: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.5. NLP Benchmark: MC TACO (Zhou et al., 2019)

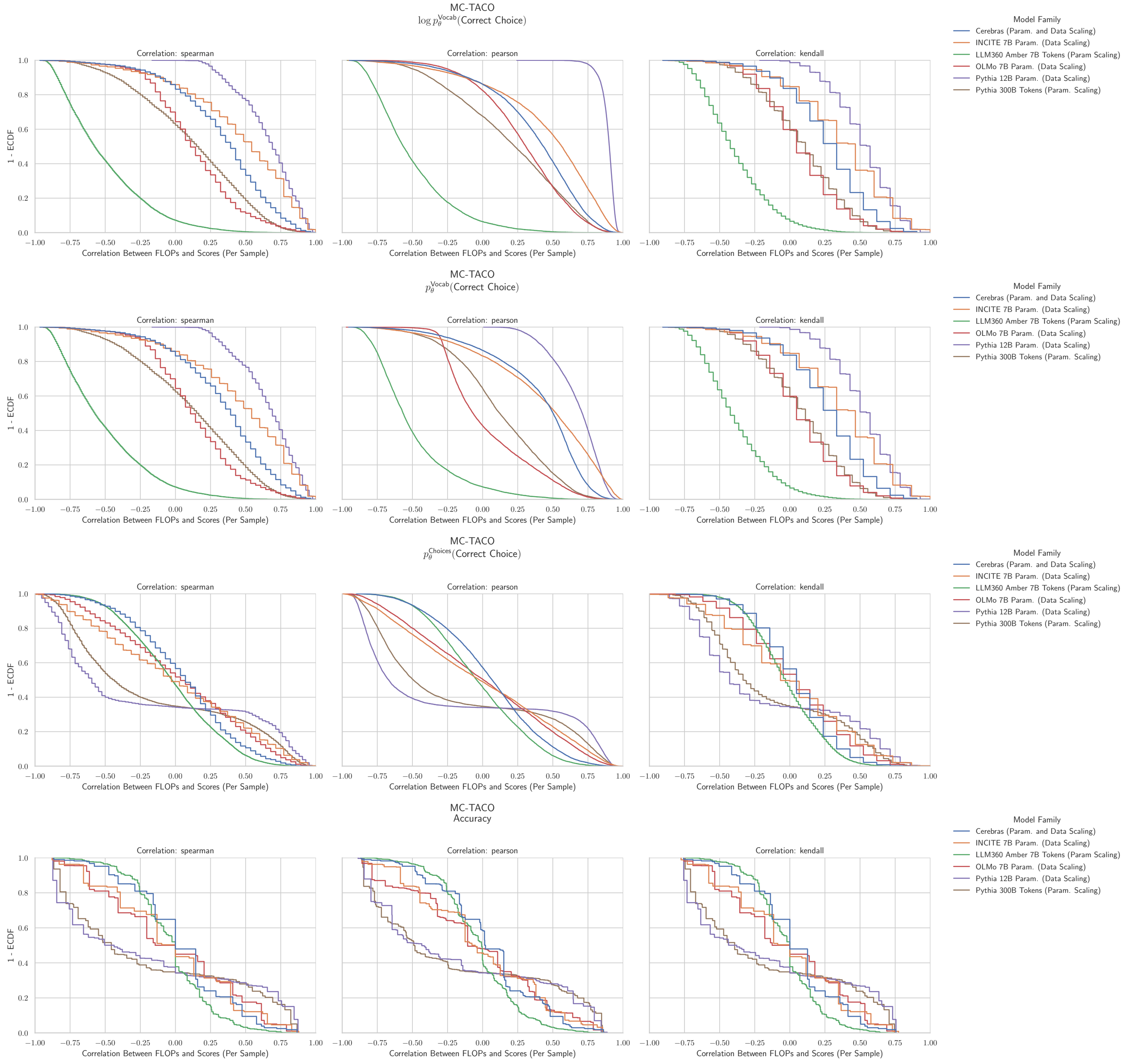

Figure 14. **MC TACO: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.6. NLP Benchmark: MMLU Abstract Algebra (Hendrycks et al., 2020)

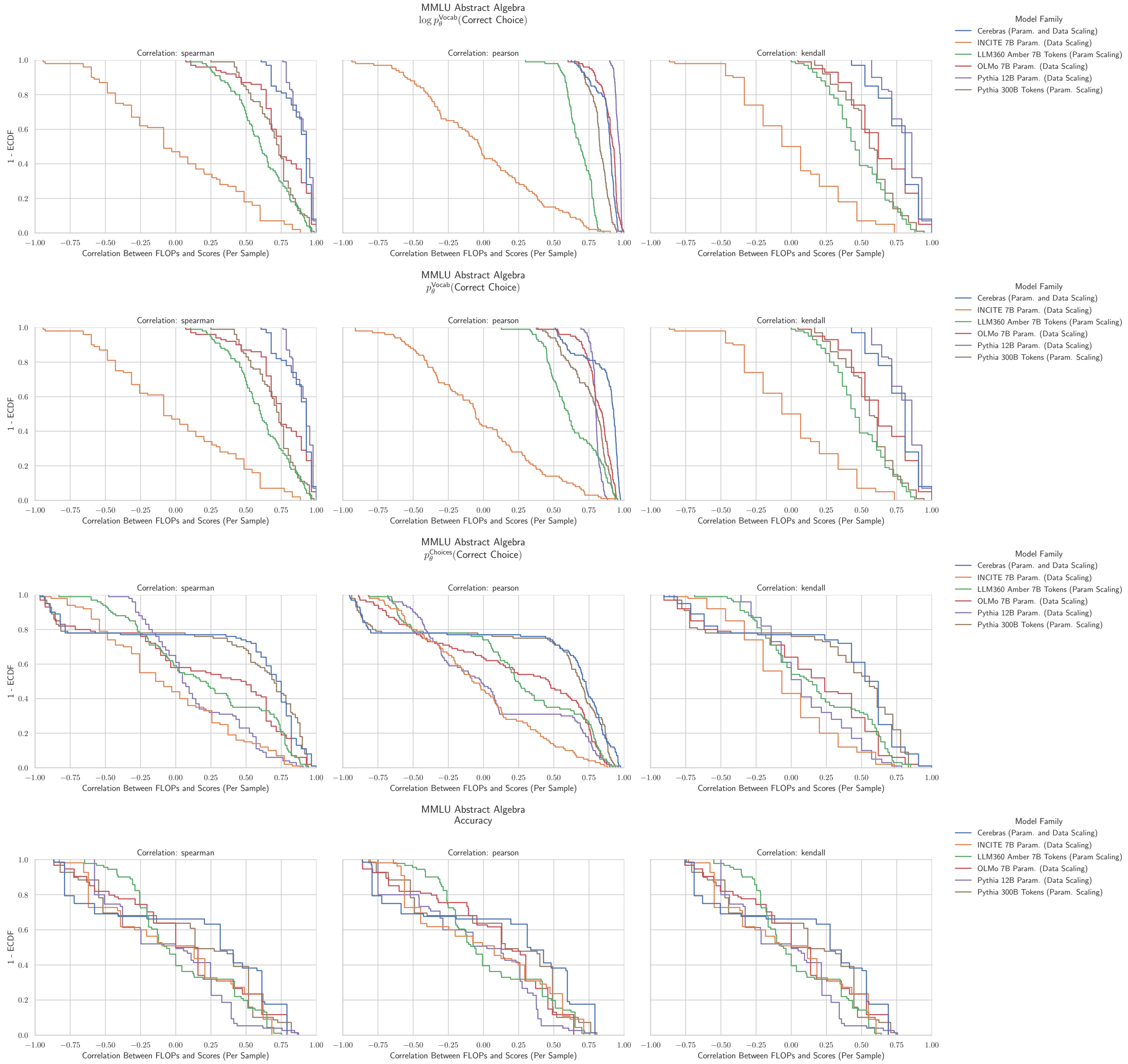

Figure 15. **MMLU Abstract Algebra: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.7. NLP Benchmark: MMLU Anatomy (Hendrycks et al., 2020)

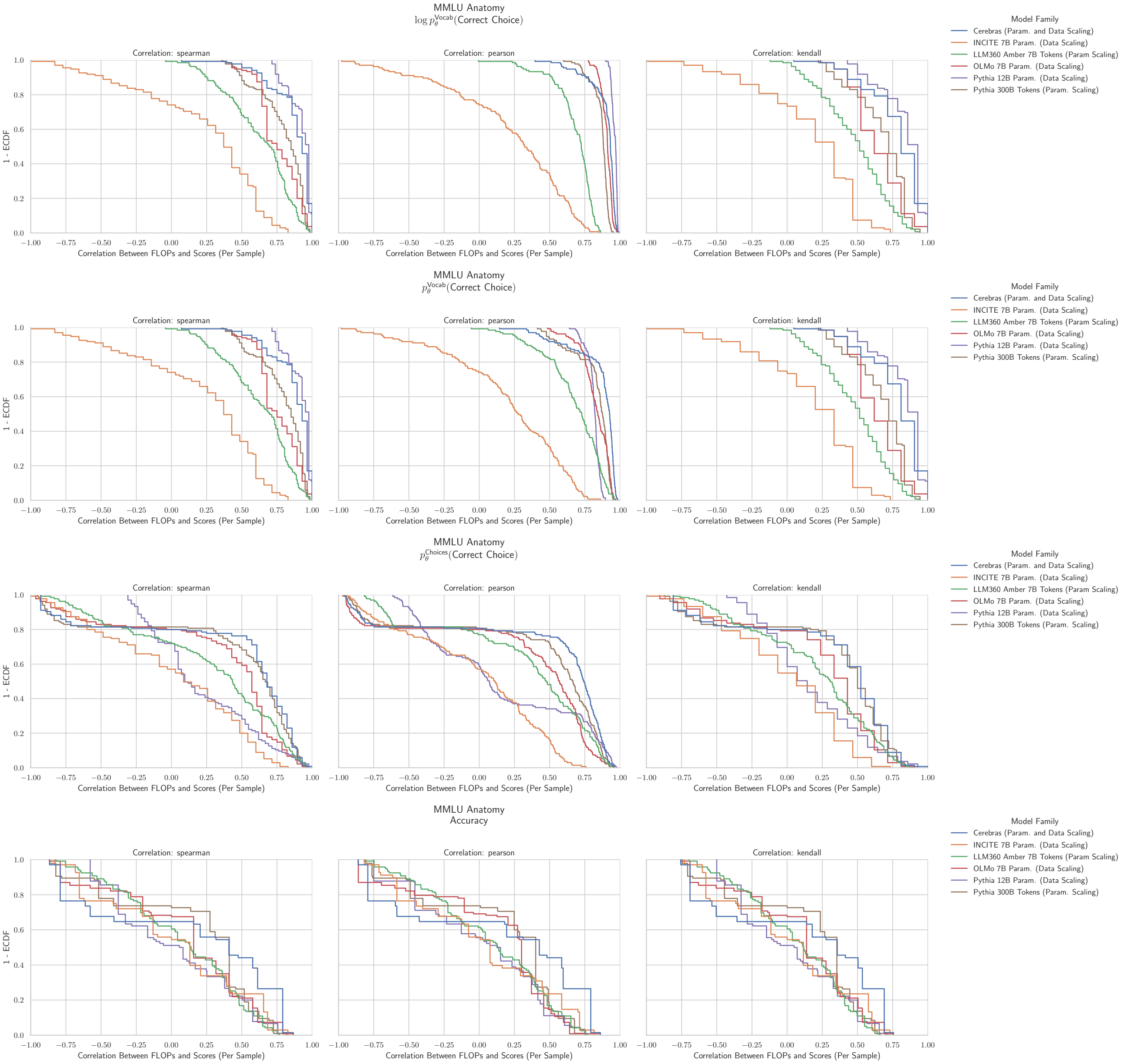

Figure 16. **MMLU Anatomy: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.8. NLP Benchmark: MMLU Astronomy (Hendrycks et al., 2020)

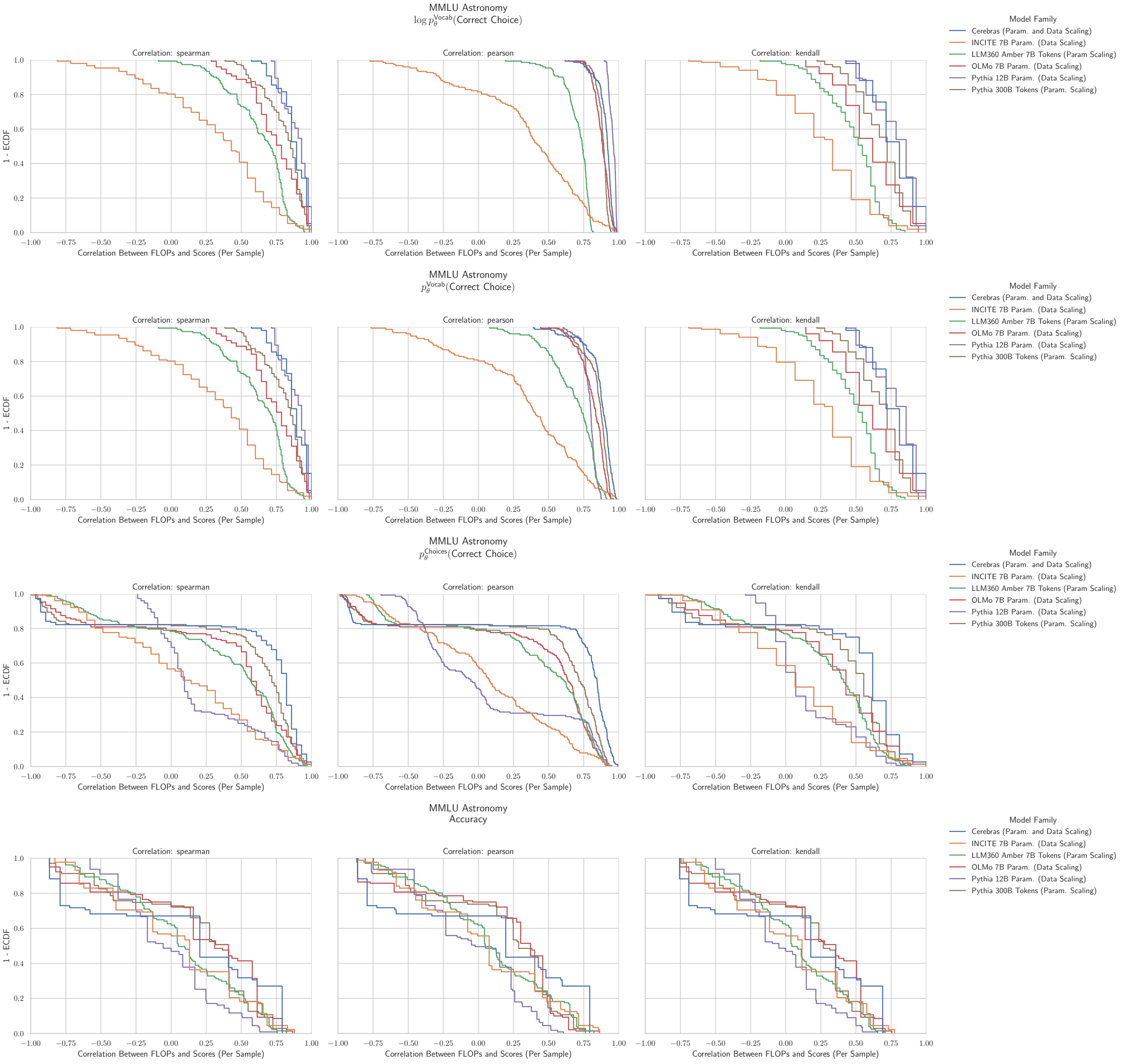

Figure 17. **MMLU Astronomy: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.9. NLP Benchmark: MMLU Business Ethics (Hendrycks et al., 2020)

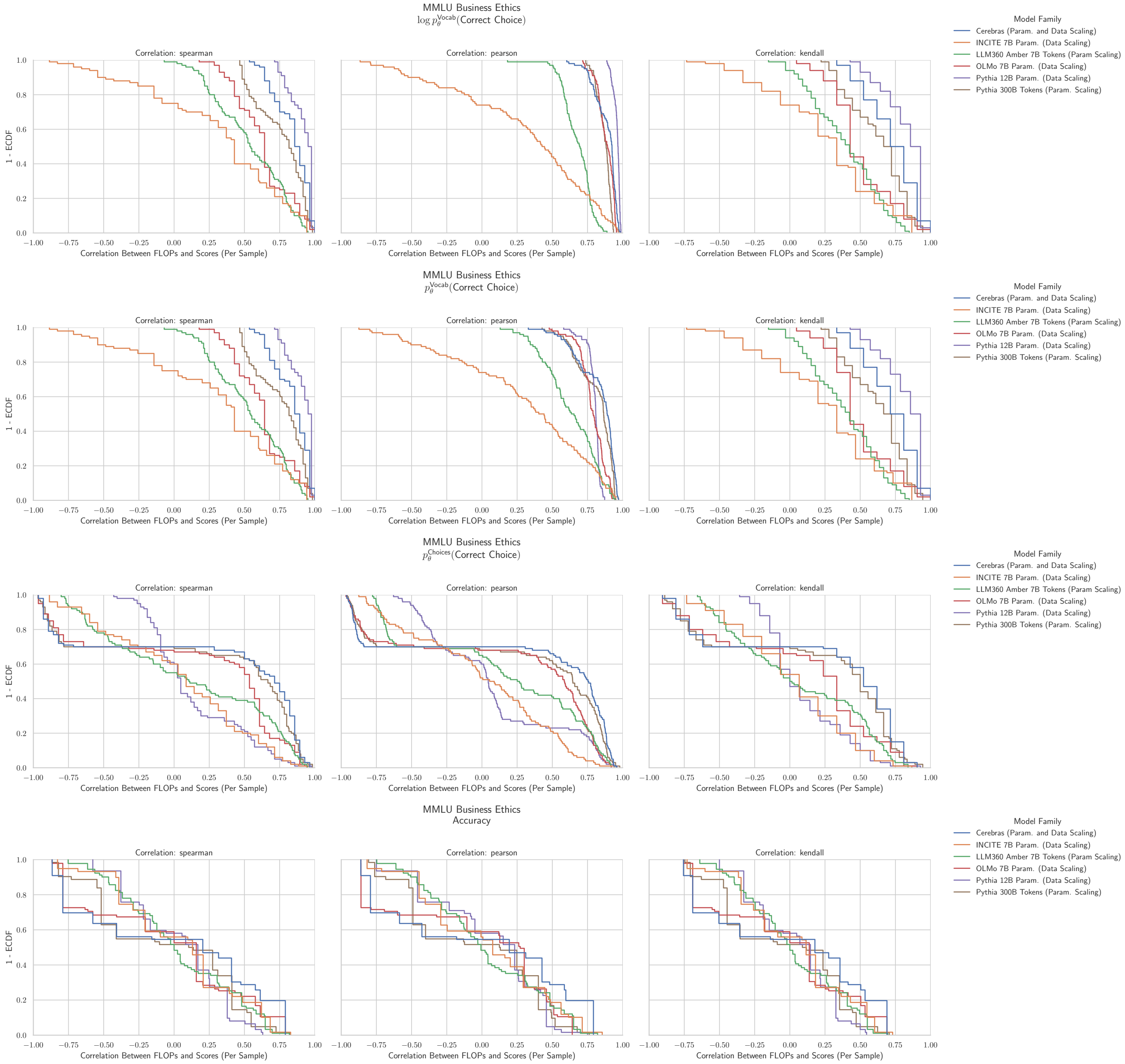

Figure 18. **MMLU Business Ethics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.10. NLP Benchmark: MMLU Clinical Knowledge (Hendrycks et al., 2020)

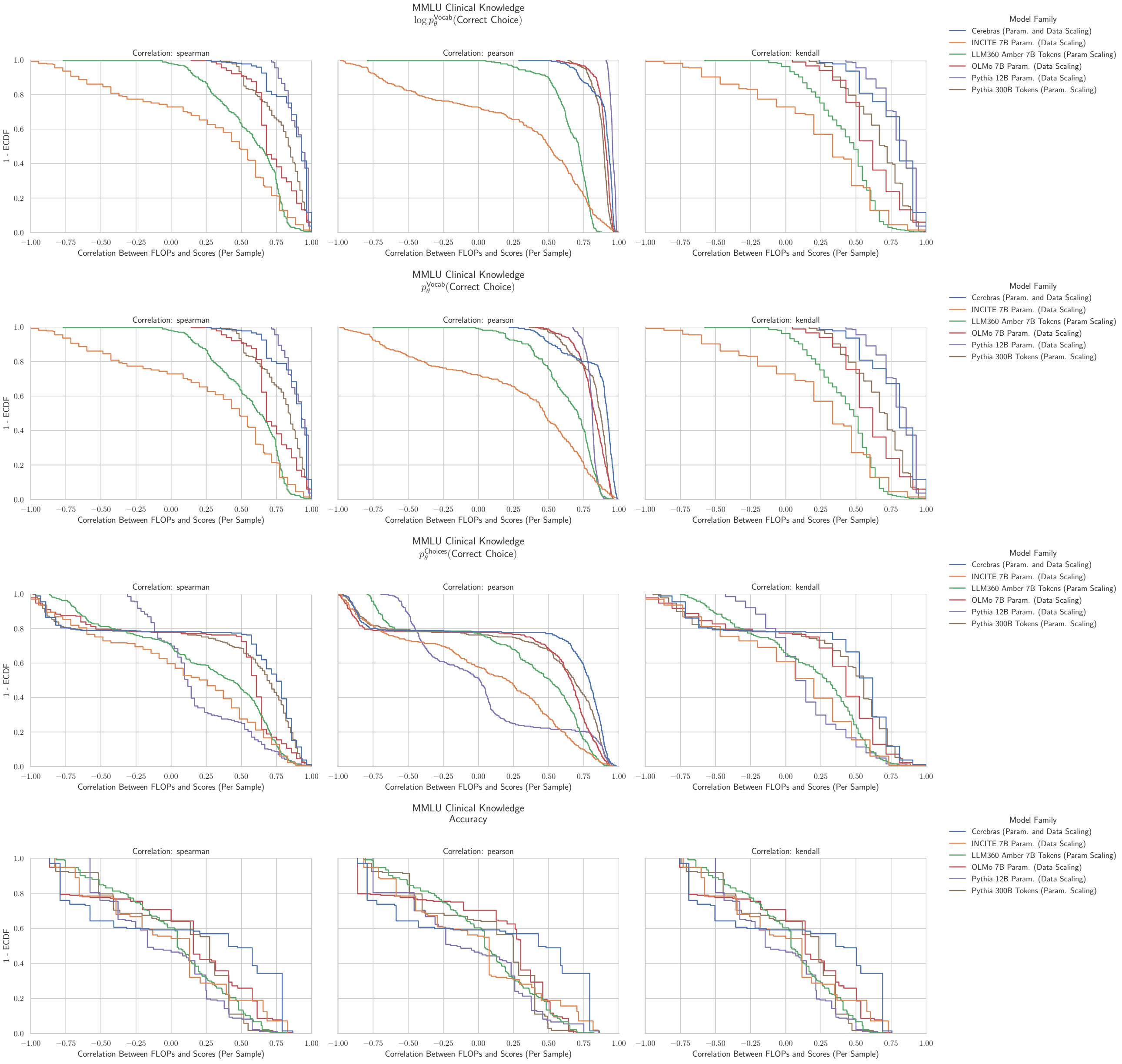

Figure 19. **MMLU Clinical Knowledge: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.11. NLP Benchmark: MMLU College Biology (Hendrycks et al., 2020)

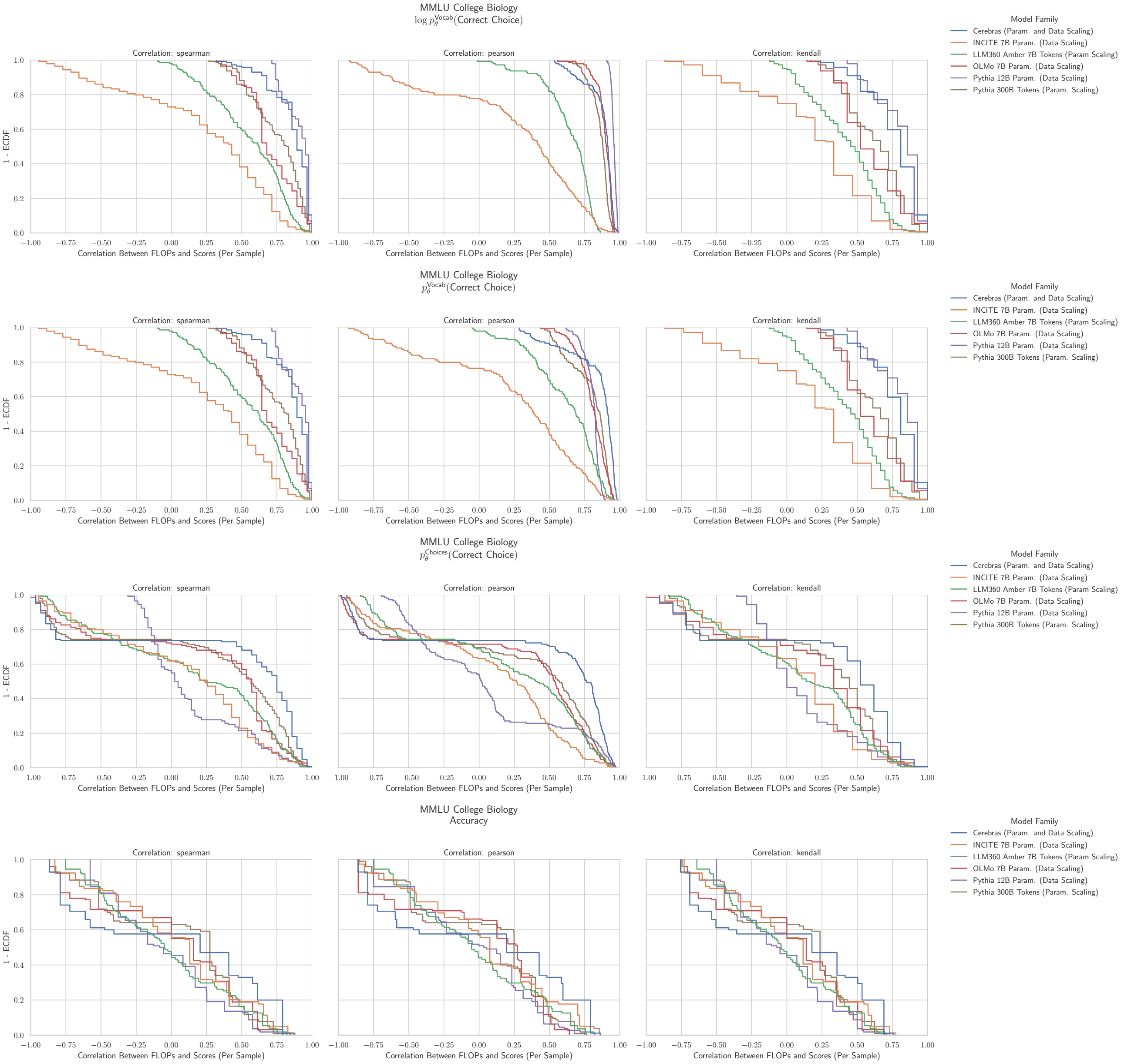

Figure 20. **MMLU College Biology: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.12. NLP Benchmark: MMLU College Chemistry (Hendrycks et al., 2020)

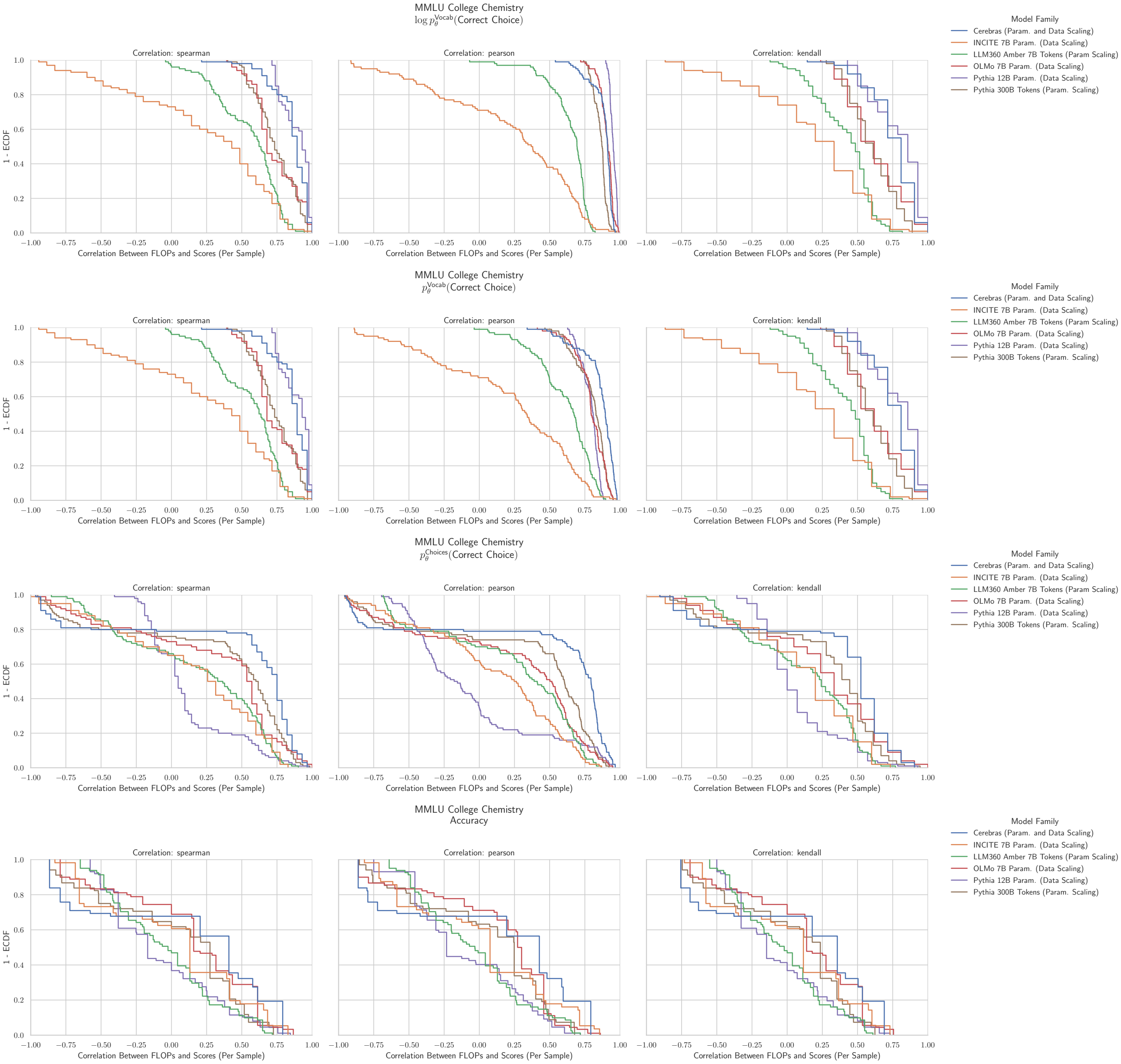

Figure 21. **MMLU College Chemistry: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.13. NLP Benchmark: MMLU College Computer Science (Hendrycks et al., 2020)

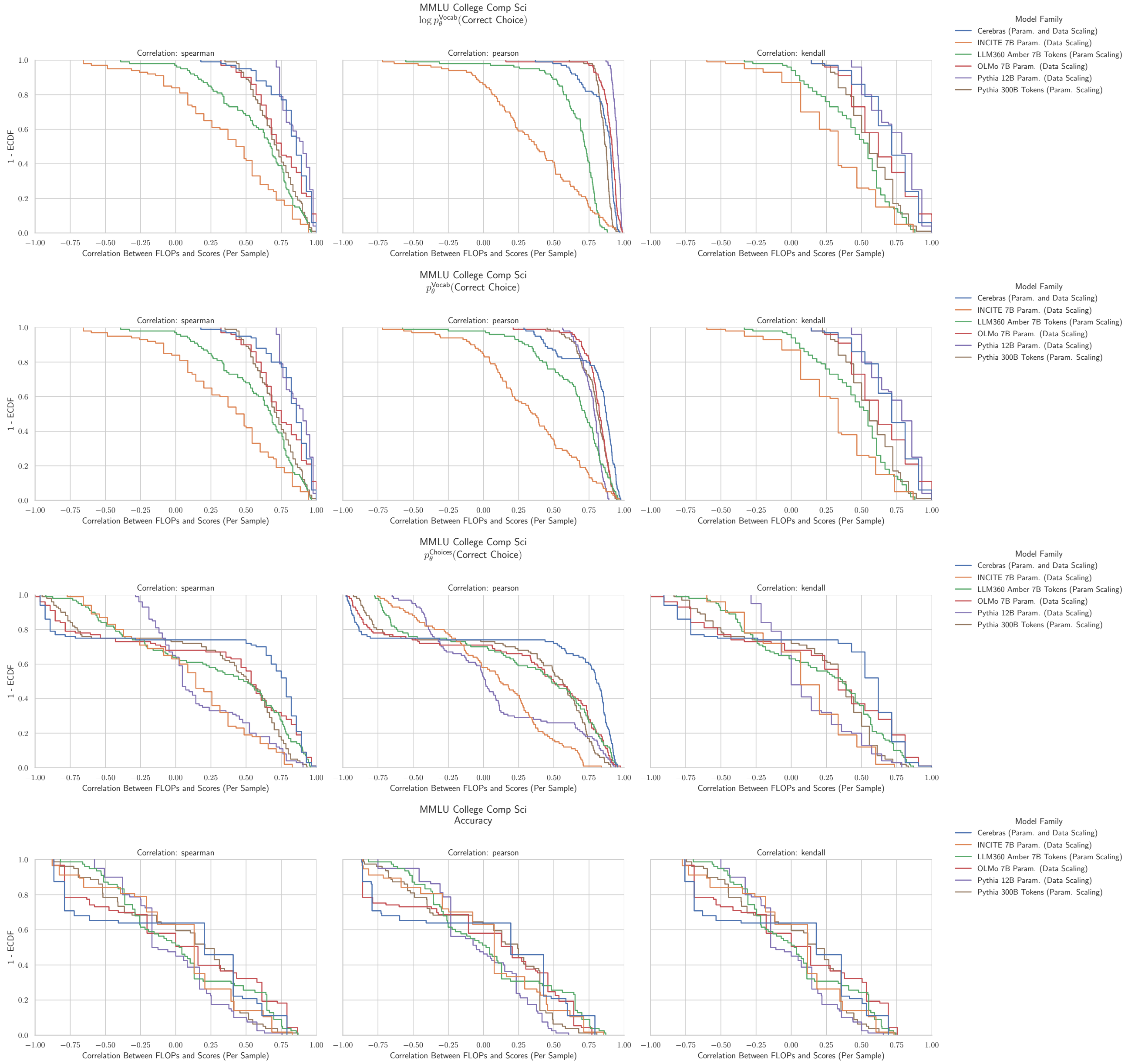

Figure 22. **MMLU College Computer Science: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.14. NLP Benchmark: MMLU College Mathematics (Hendrycks et al., 2020)

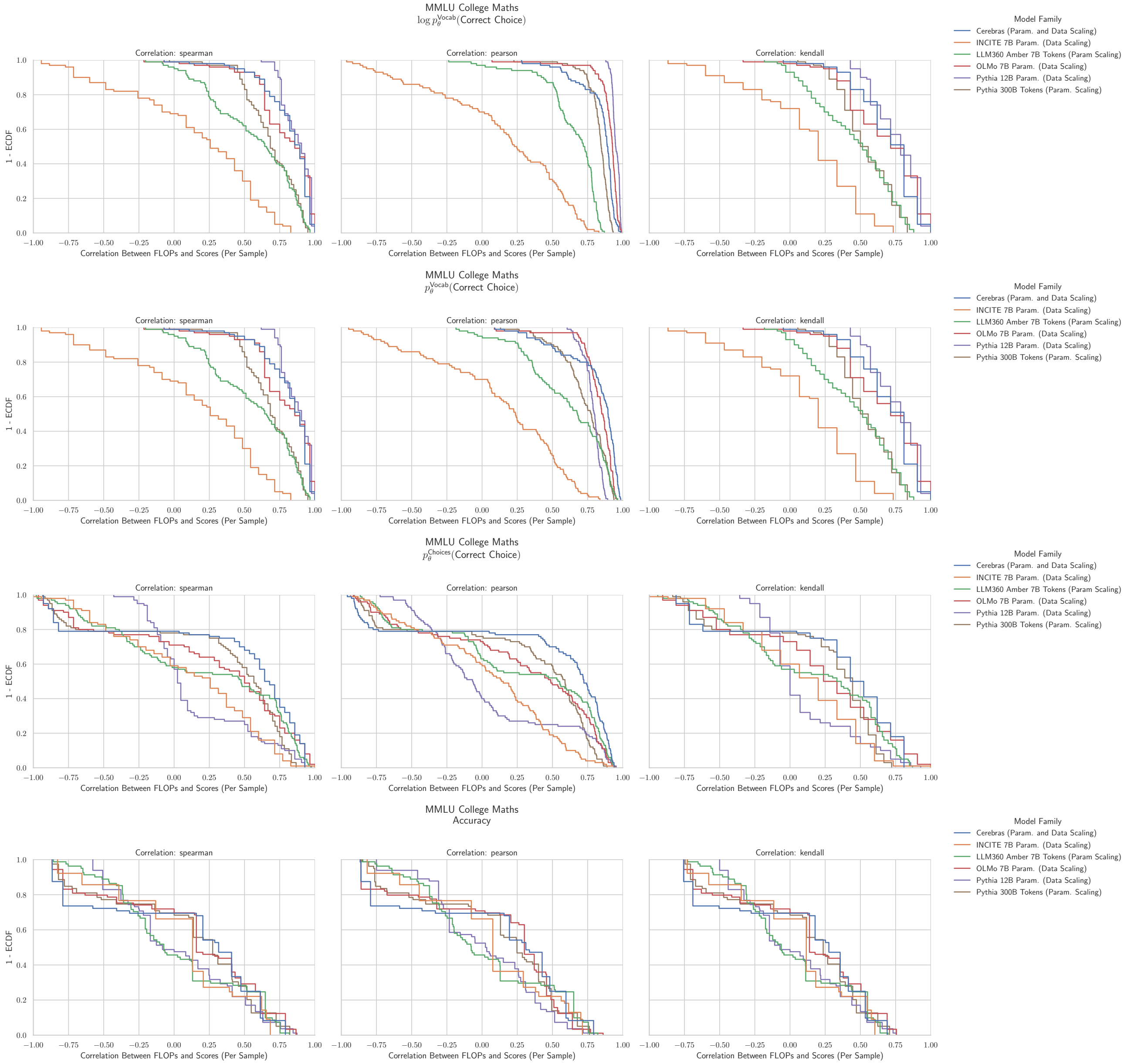

Figure 23. **MMLU College Mathematics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.15. NLP Benchmark: MMLU College Medicine (Hendrycks et al., 2020)

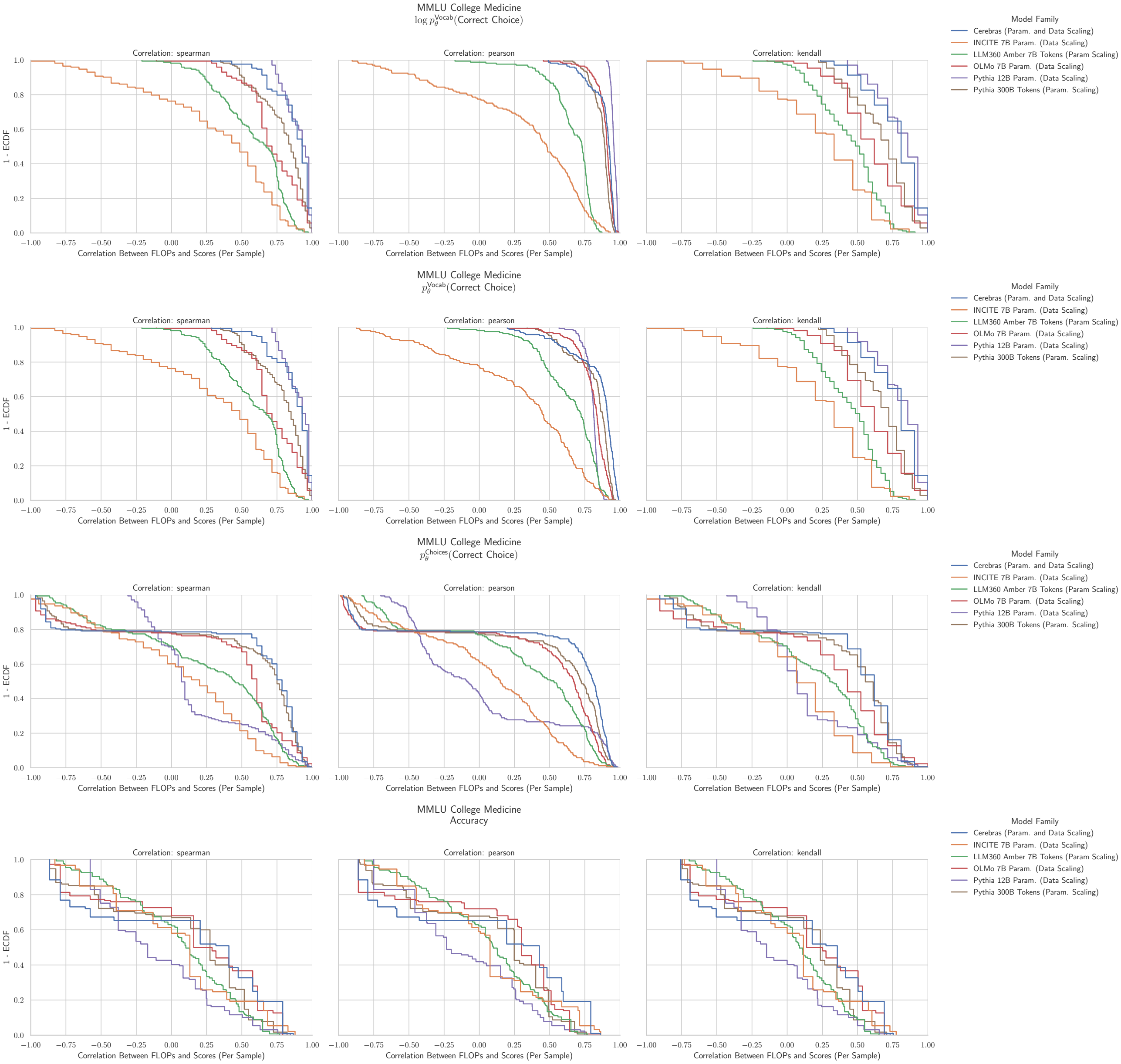

Figure 24. **MMLU College Medicine: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.16. NLP Benchmark: MMLU College Physics (Hendrycks et al., 2020)

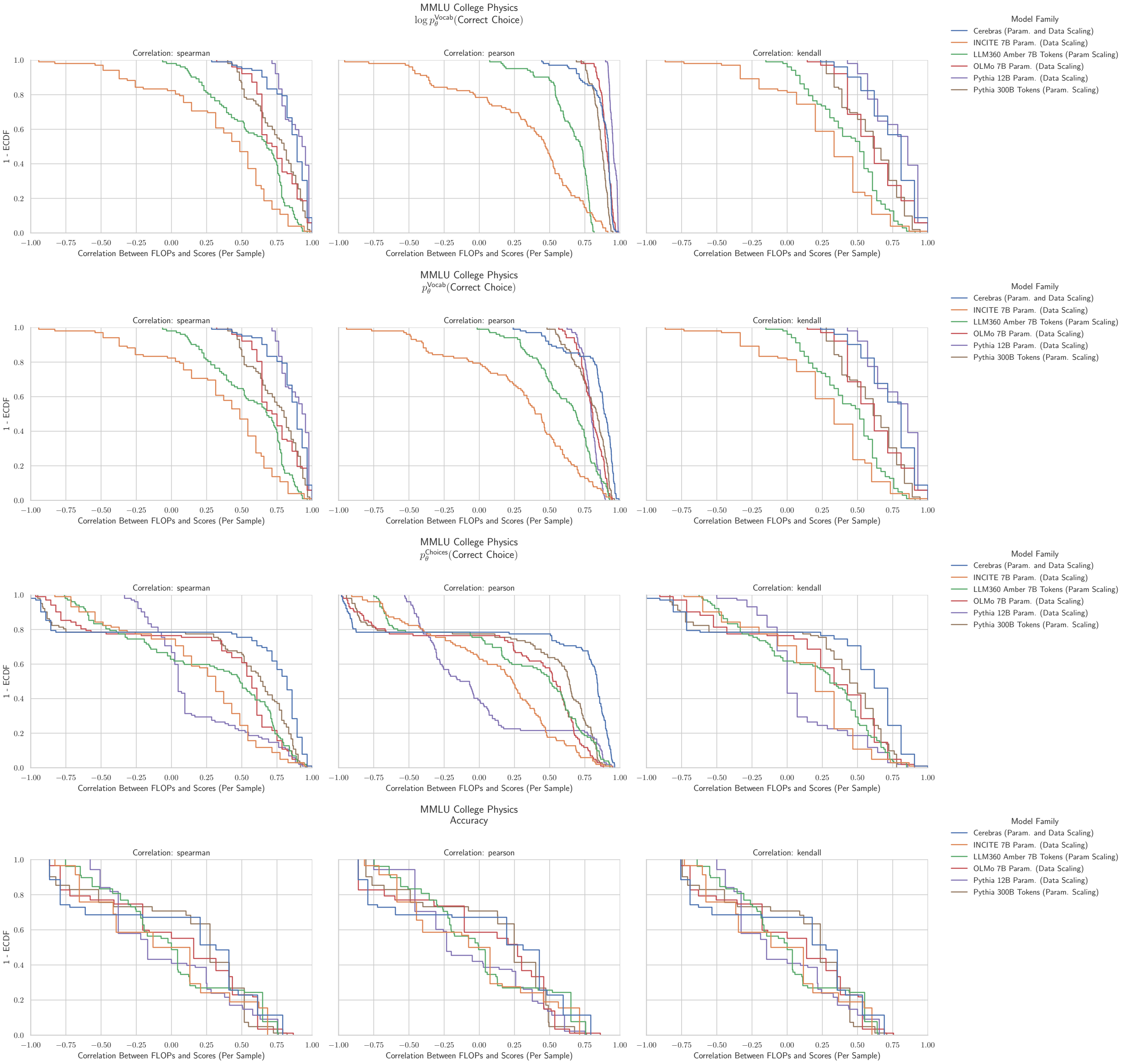

Figure 25. **MMLU College Physics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.17. NLP Benchmark: MMLU Computer Security (Hendrycks et al., 2020)

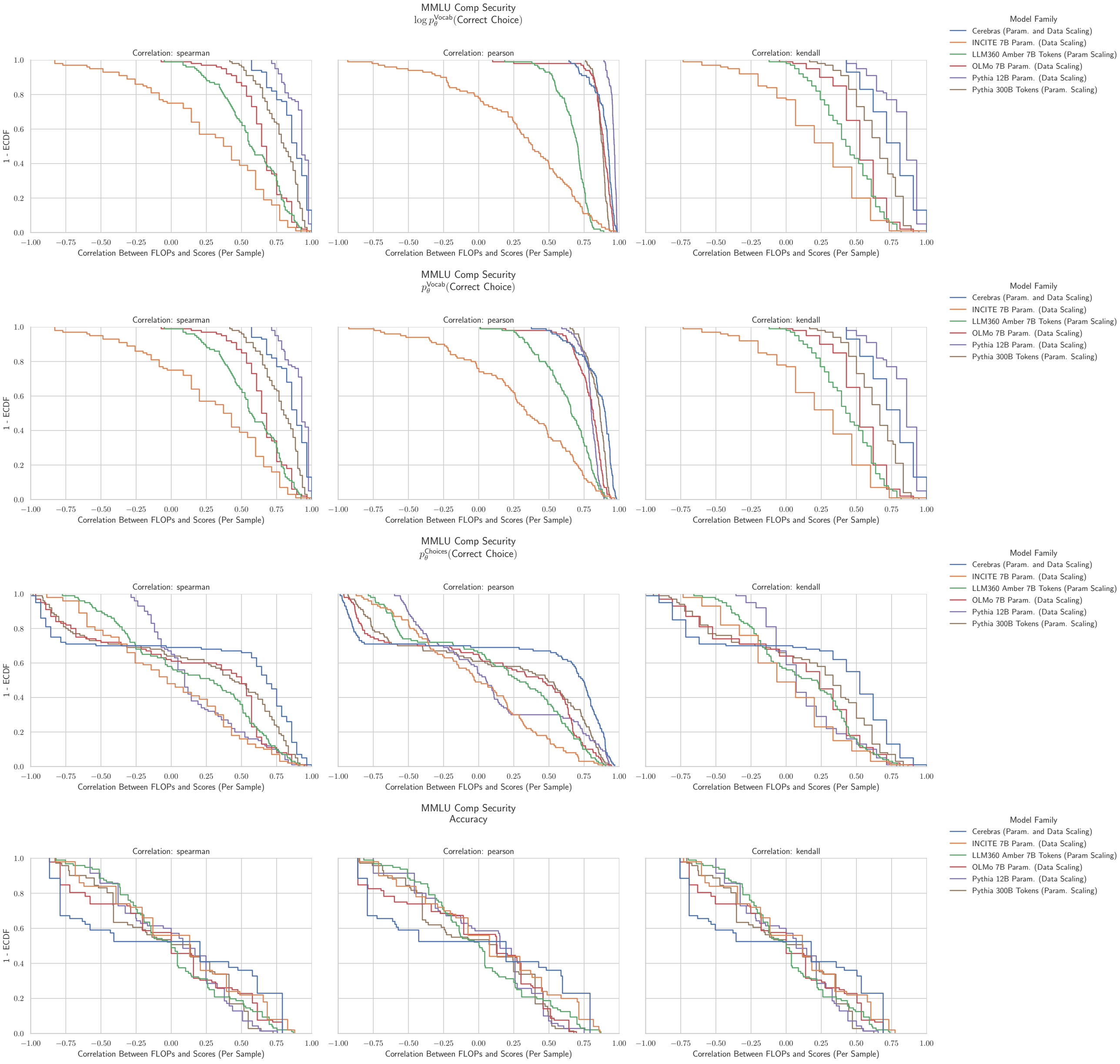

Figure 26. **MMLU Computer Security: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.18. NLP Benchmark: MMLU Conceptual Physics (Hendrycks et al., 2020)

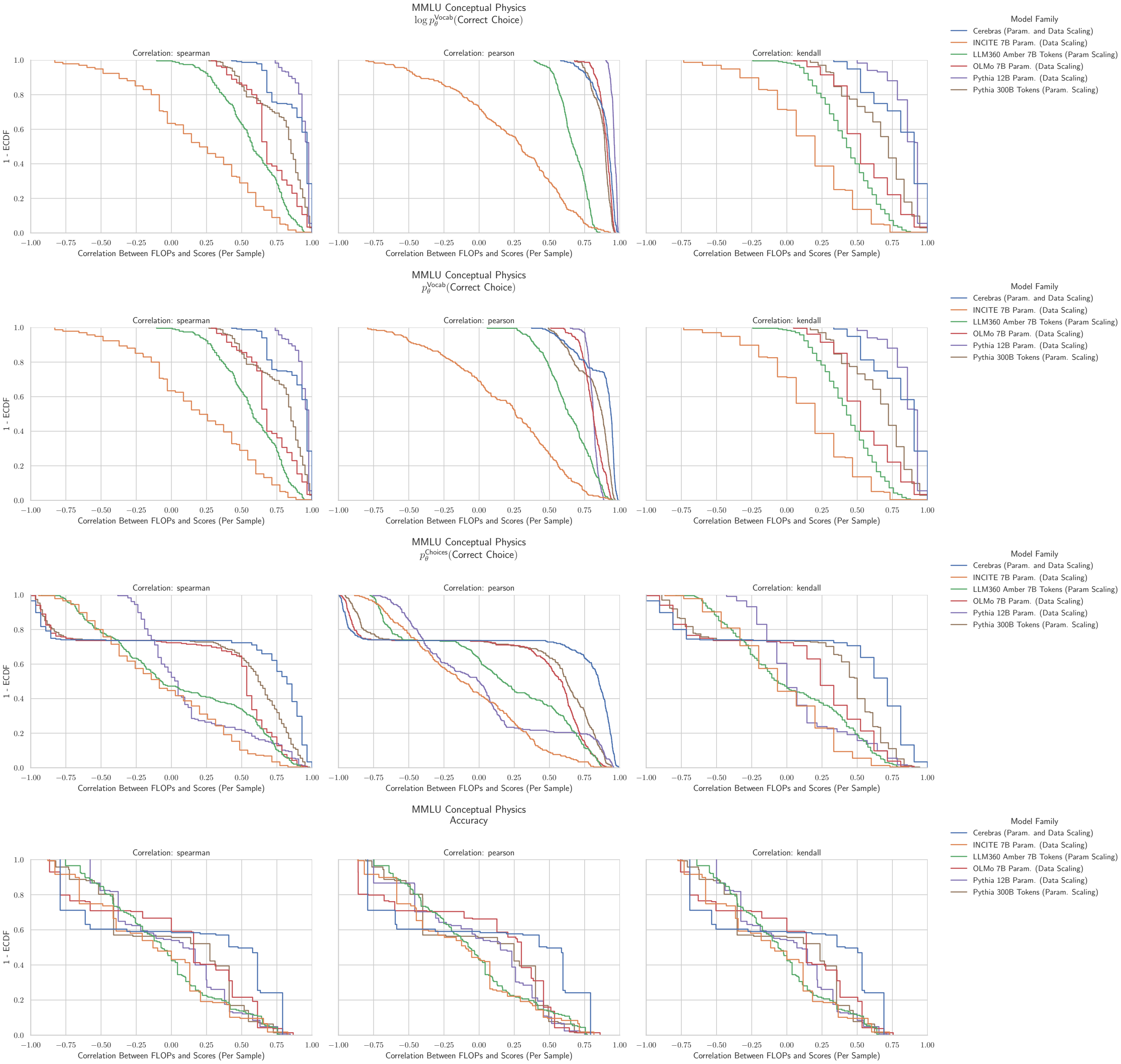

Figure 27. **MMLU Conceptual Physics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.19. NLP Benchmark: MMLU Econometrics (Hendrycks et al., 2020)

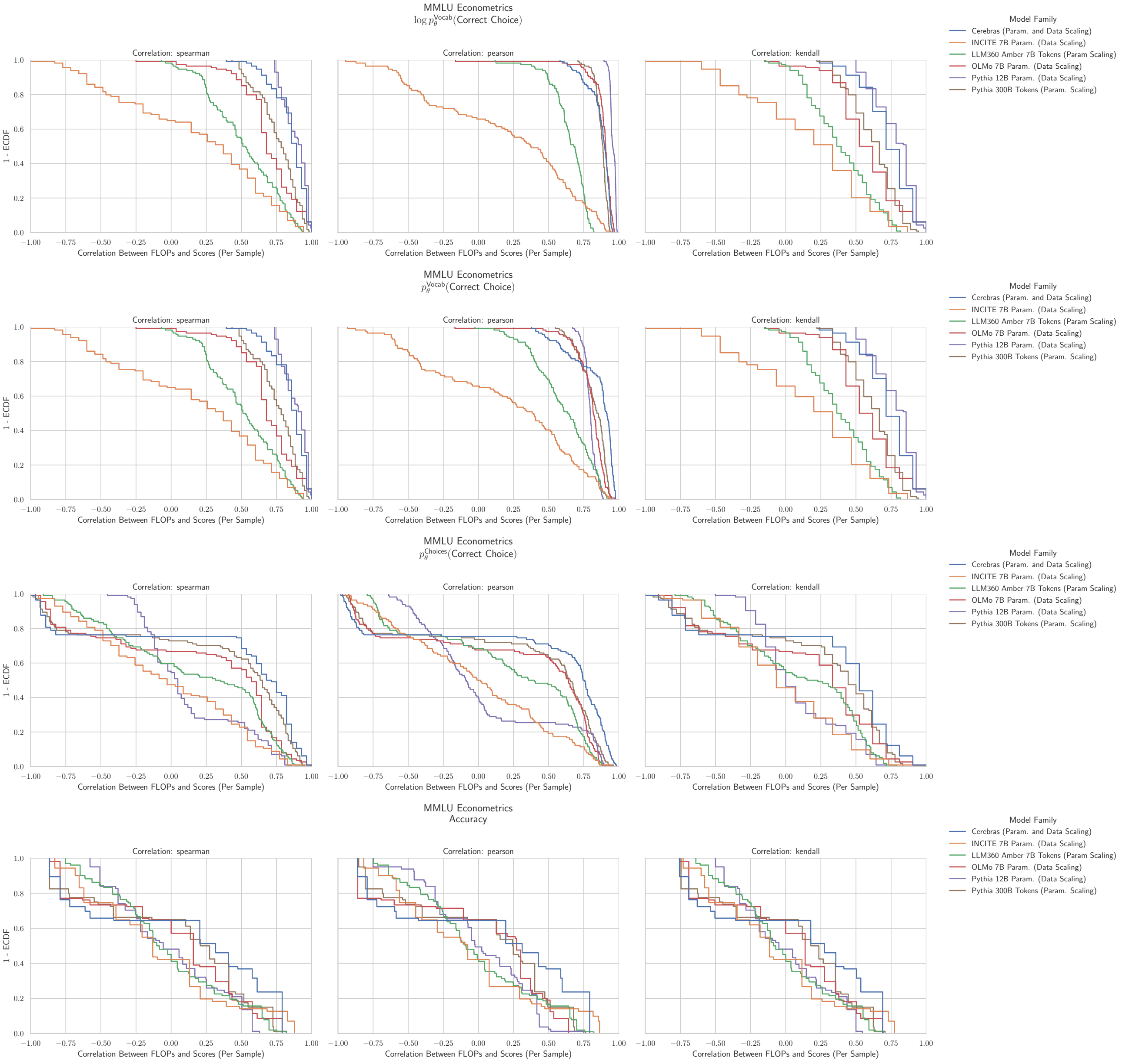

Figure 28. **MMLU Econometrics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.20. NLP Benchmark: MMLU Electrical Engineering (Hendrycks et al., 2020)

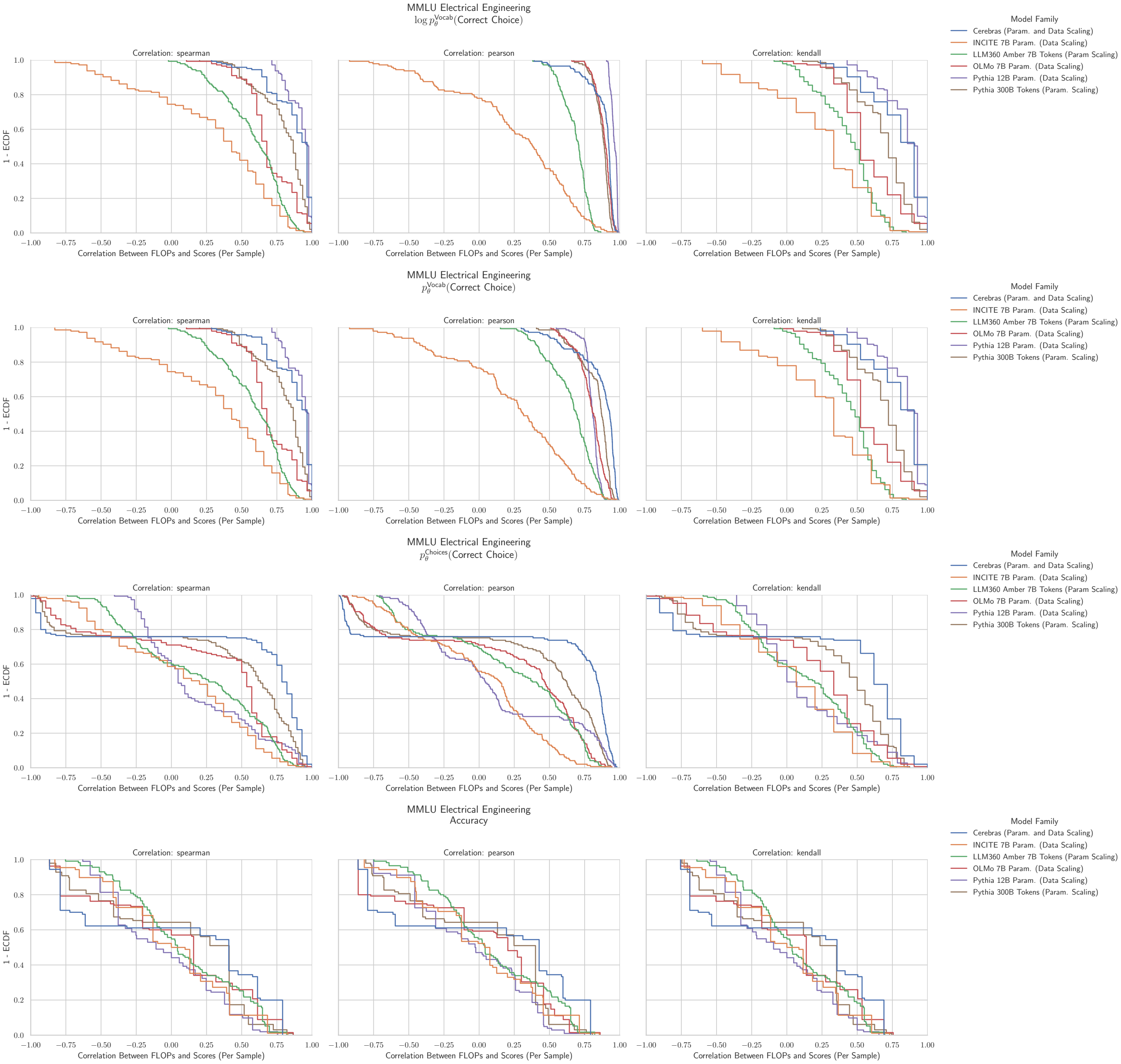

Figure 29. **MMLU Electrical Engineering: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.21. NLP Benchmark: MMLU Elementary Mathematics (Hendrycks et al., 2020)

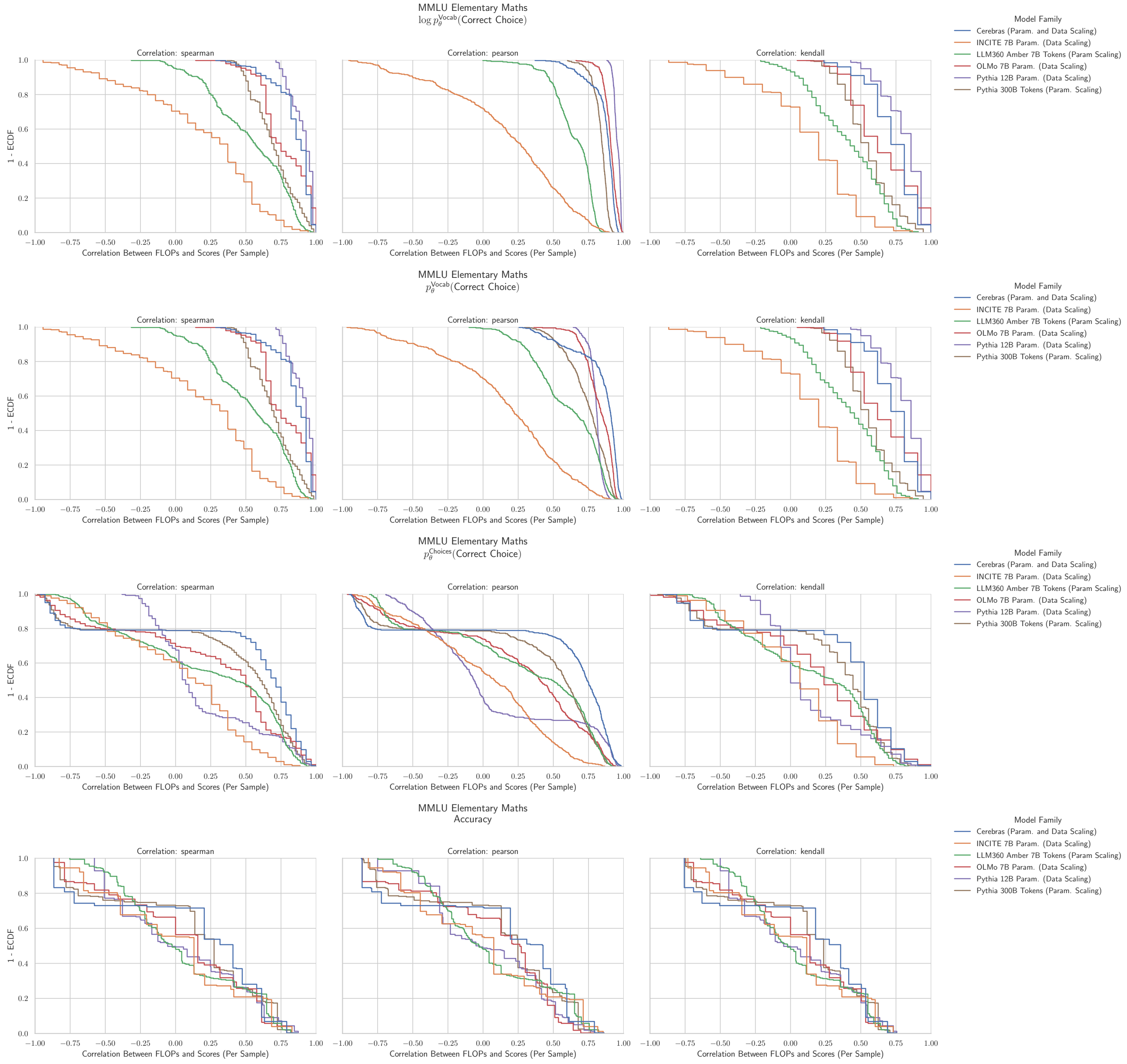

Figure 30. **MMLU Elementary Mathematics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.22. NLP Benchmark: MMLU Formal Logic (Hendrycks et al., 2020)

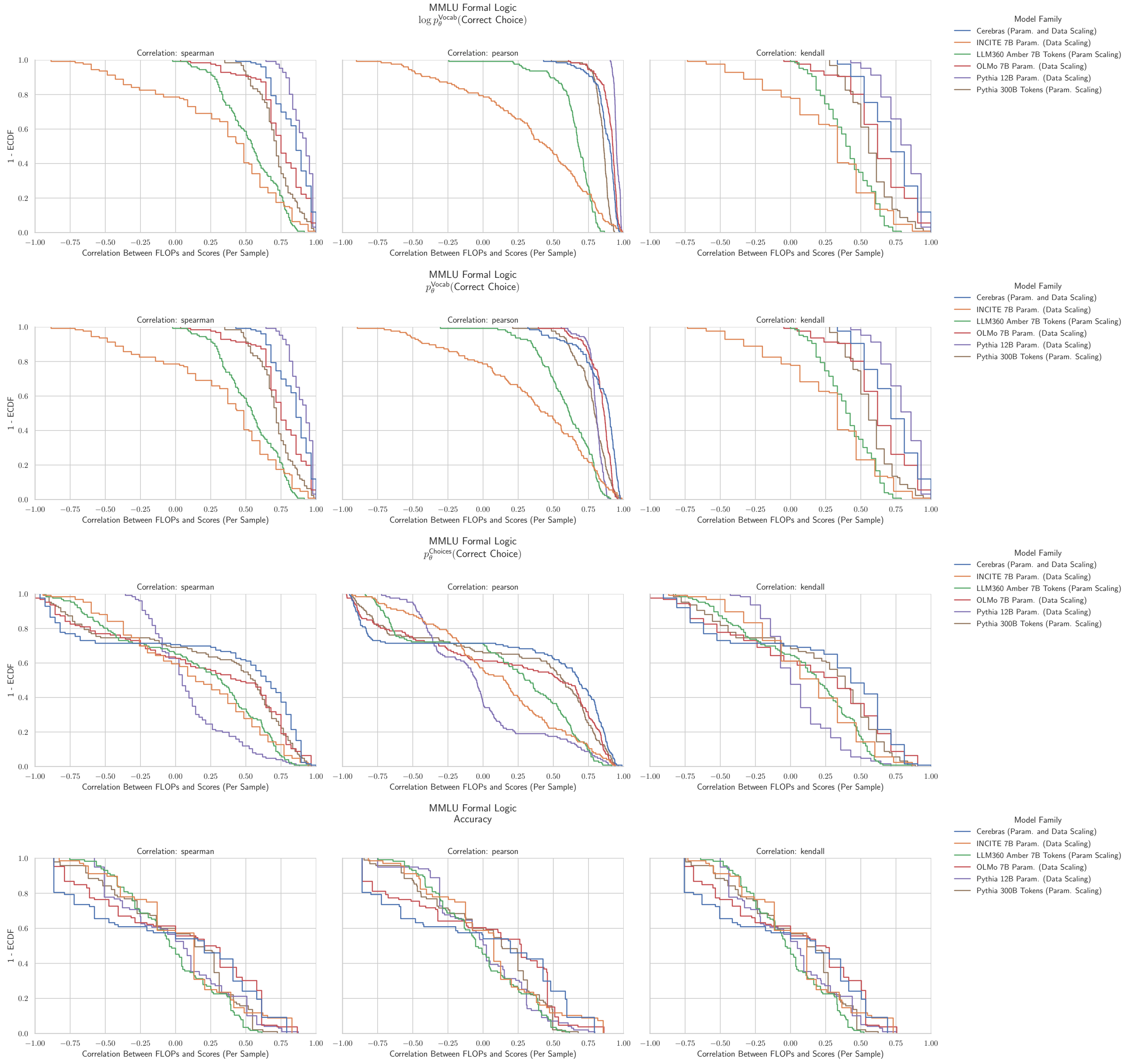

Figure 31. **MMLU Formal Logic: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.23. NLP Benchmark: MMLU Global Facts (Hendrycks et al., 2020)

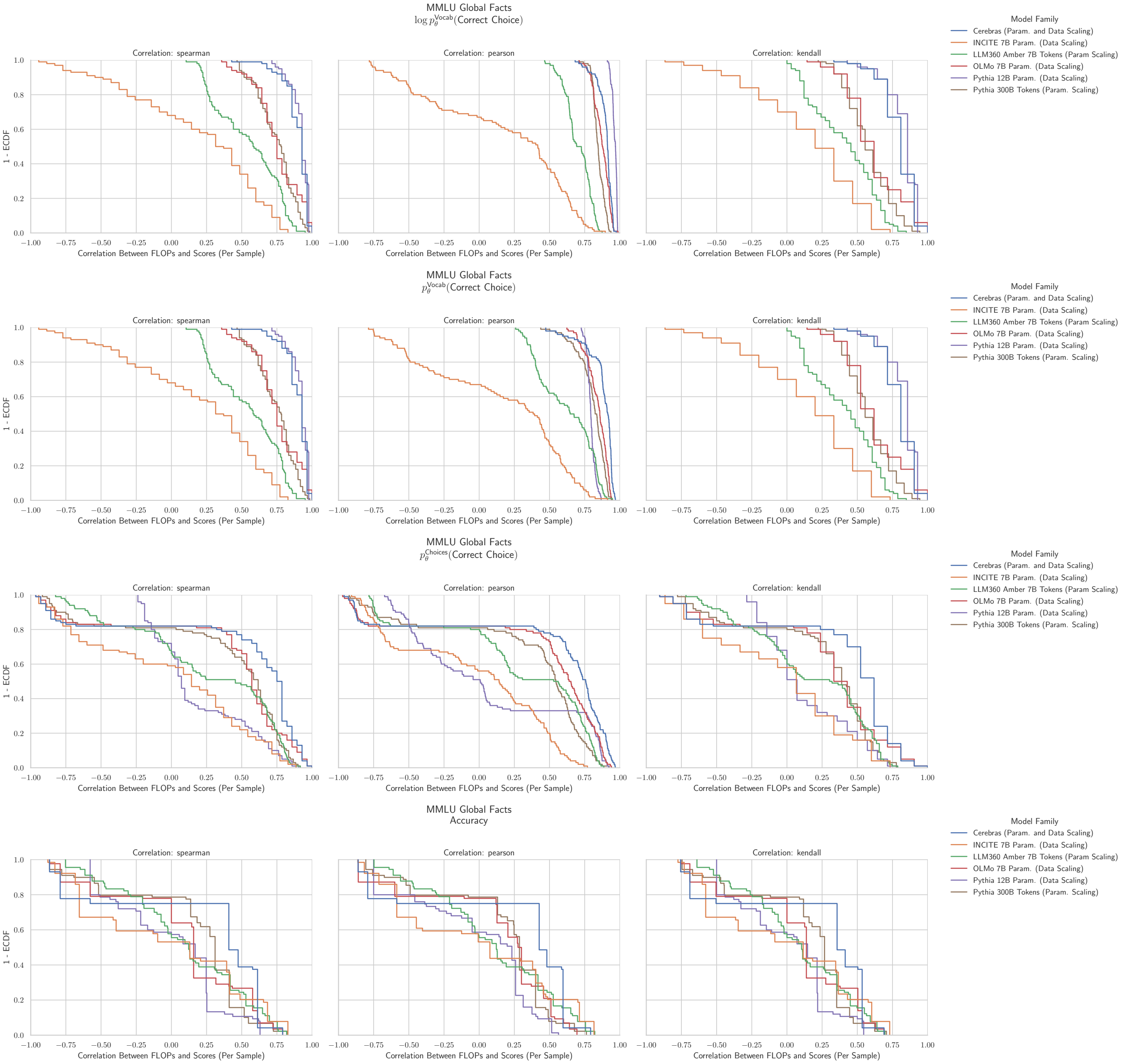

Figure 32. **MMLU Global Facts: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.24. NLP Benchmark: MMLU High School Biology (Hendrycks et al., 2020)

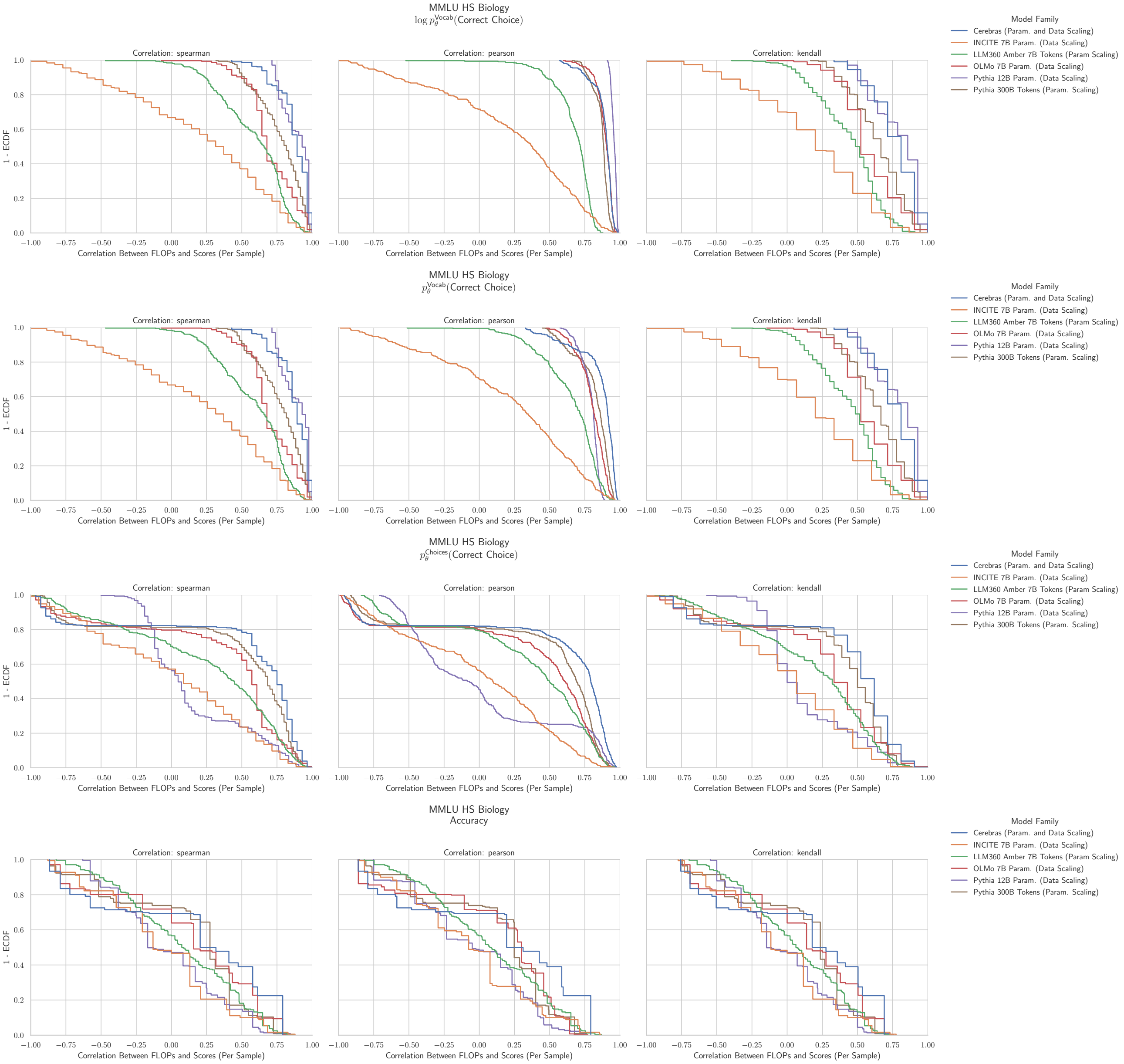

Figure 33. **MMLU High School Biology: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.25. NLP Benchmark: MMLU High School Chemistry (Hendrycks et al., 2020)

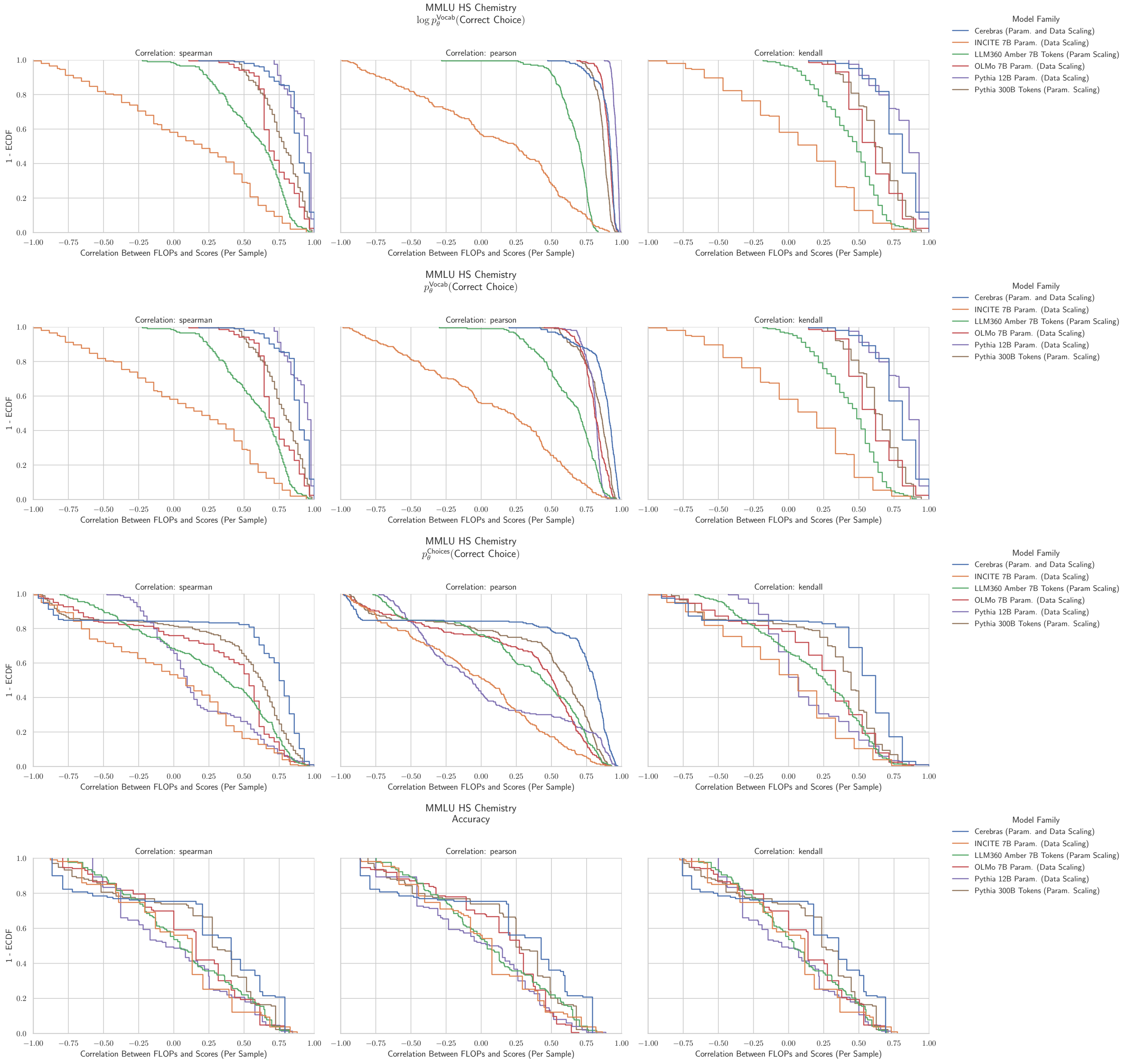

Figure 34. **MMLU High School Chemistry: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.26. NLP Benchmark: MMLU High School Computer Science (Hendrycks et al., 2020)

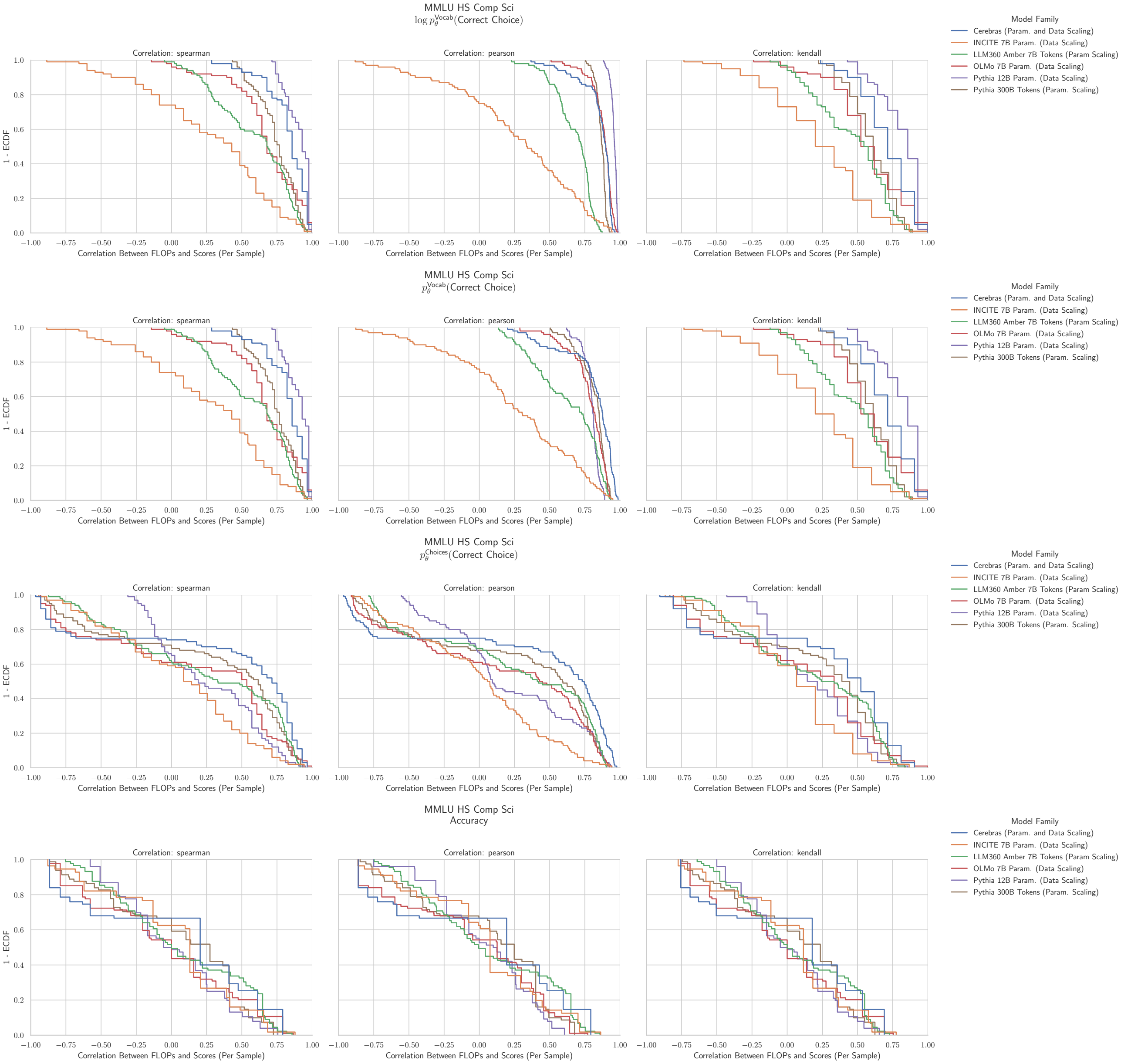

Figure 35. **MMLU High School Computer Science: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.27. NLP Benchmark: MMLU High School Chemistry (Hendrycks et al., 2020)

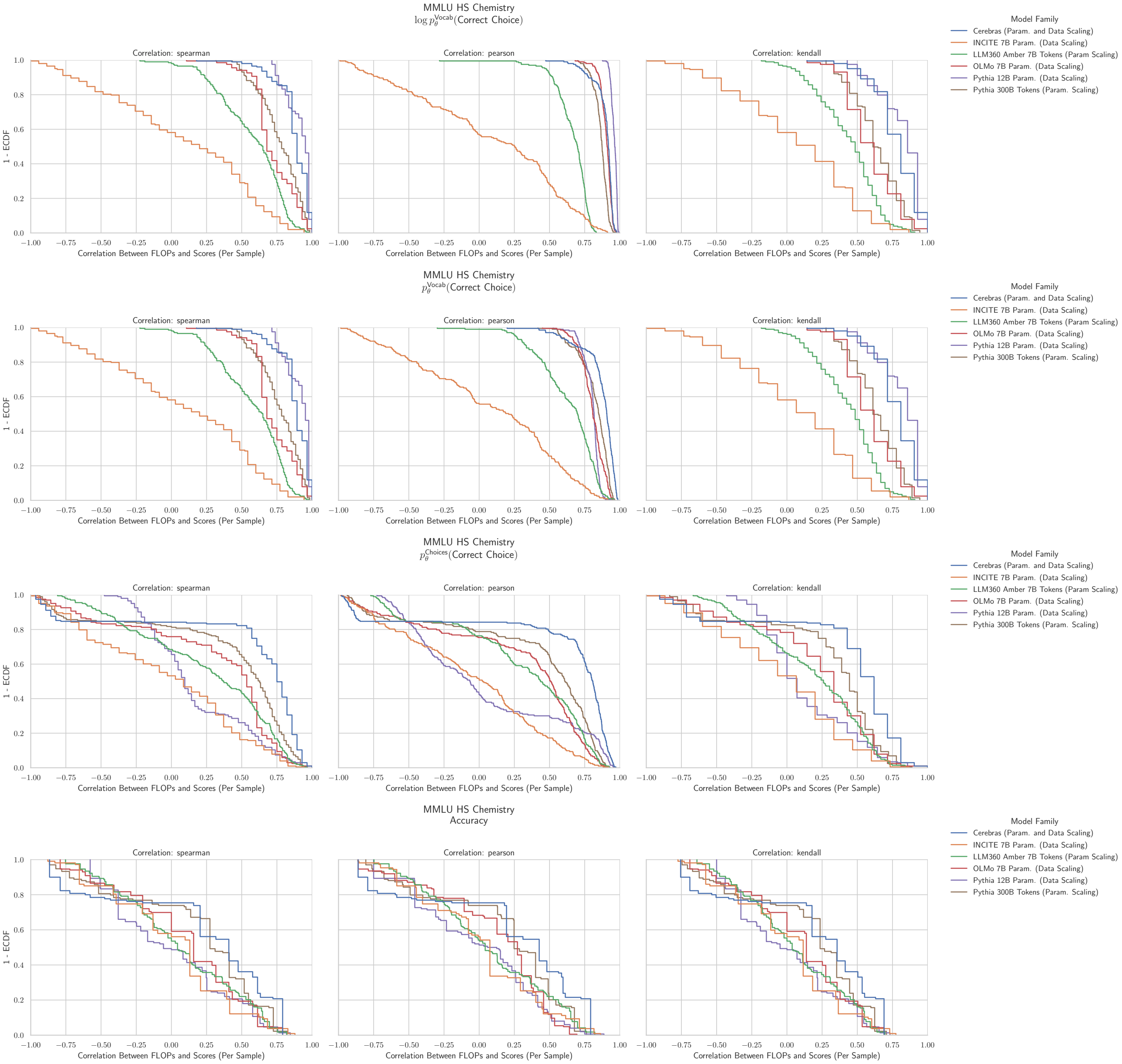

Figure 36. **MMLU High School Chemistry: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.28. NLP Benchmark: MMLU High School European History (Hendrycks et al., 2020)

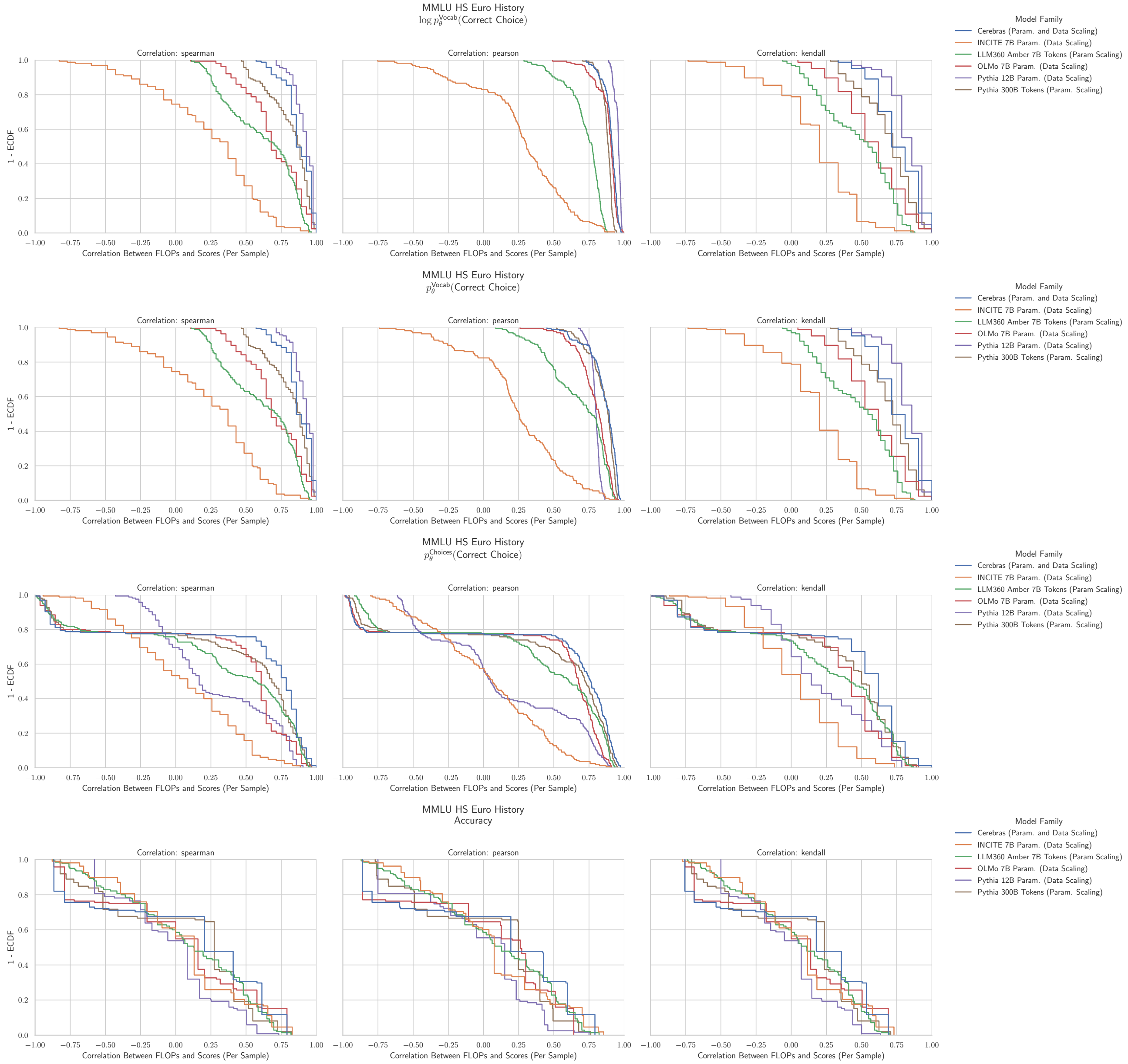

Figure 37. **MMLU High School European History: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.29. NLP Benchmark: MMLU High School Geography (Hendrycks et al., 2020)

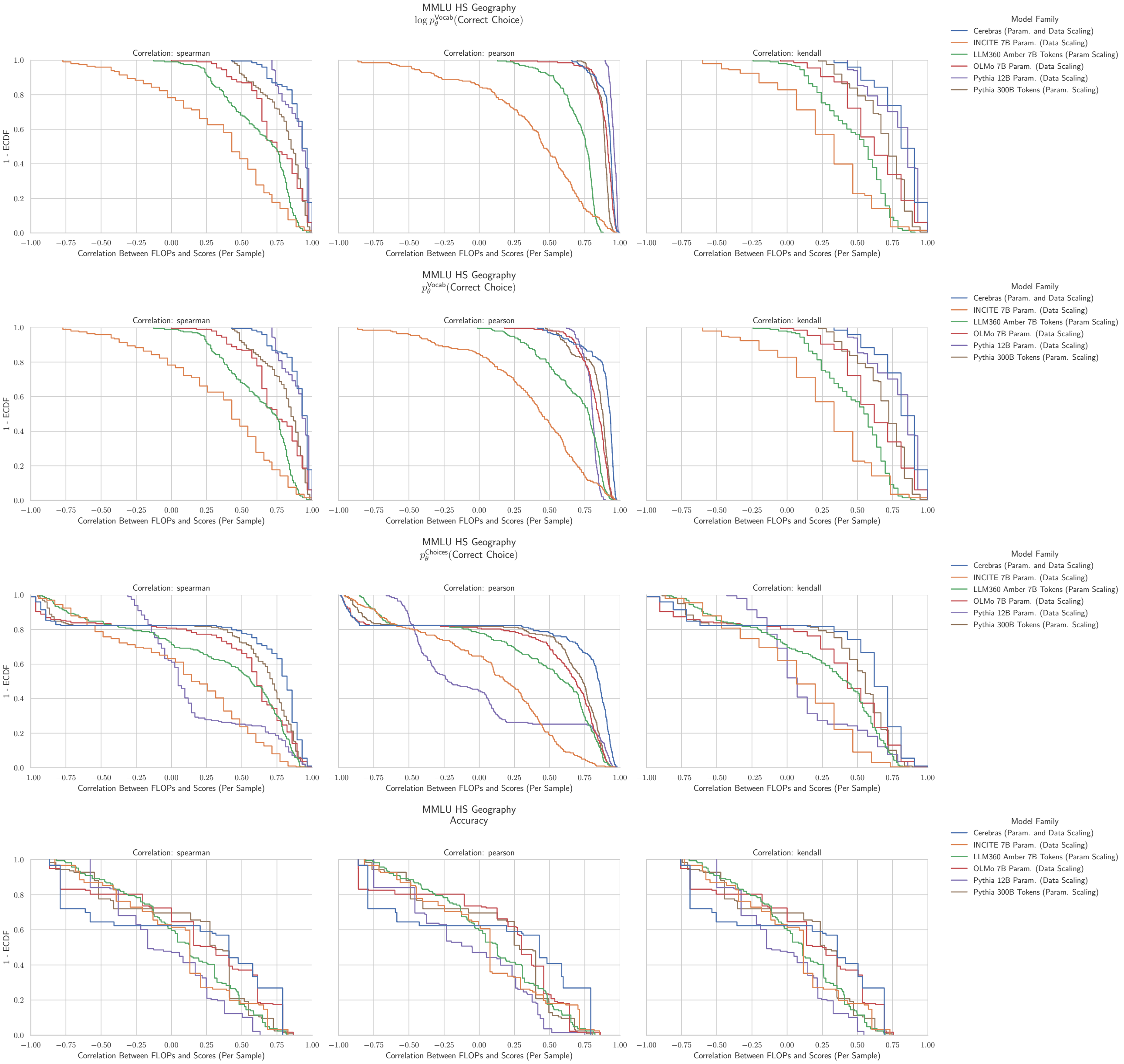

Figure 38. **MMLU High School Geography: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.30. NLP Benchmark: MMLU High School Government & Politics (Hendrycks et al., 2020)

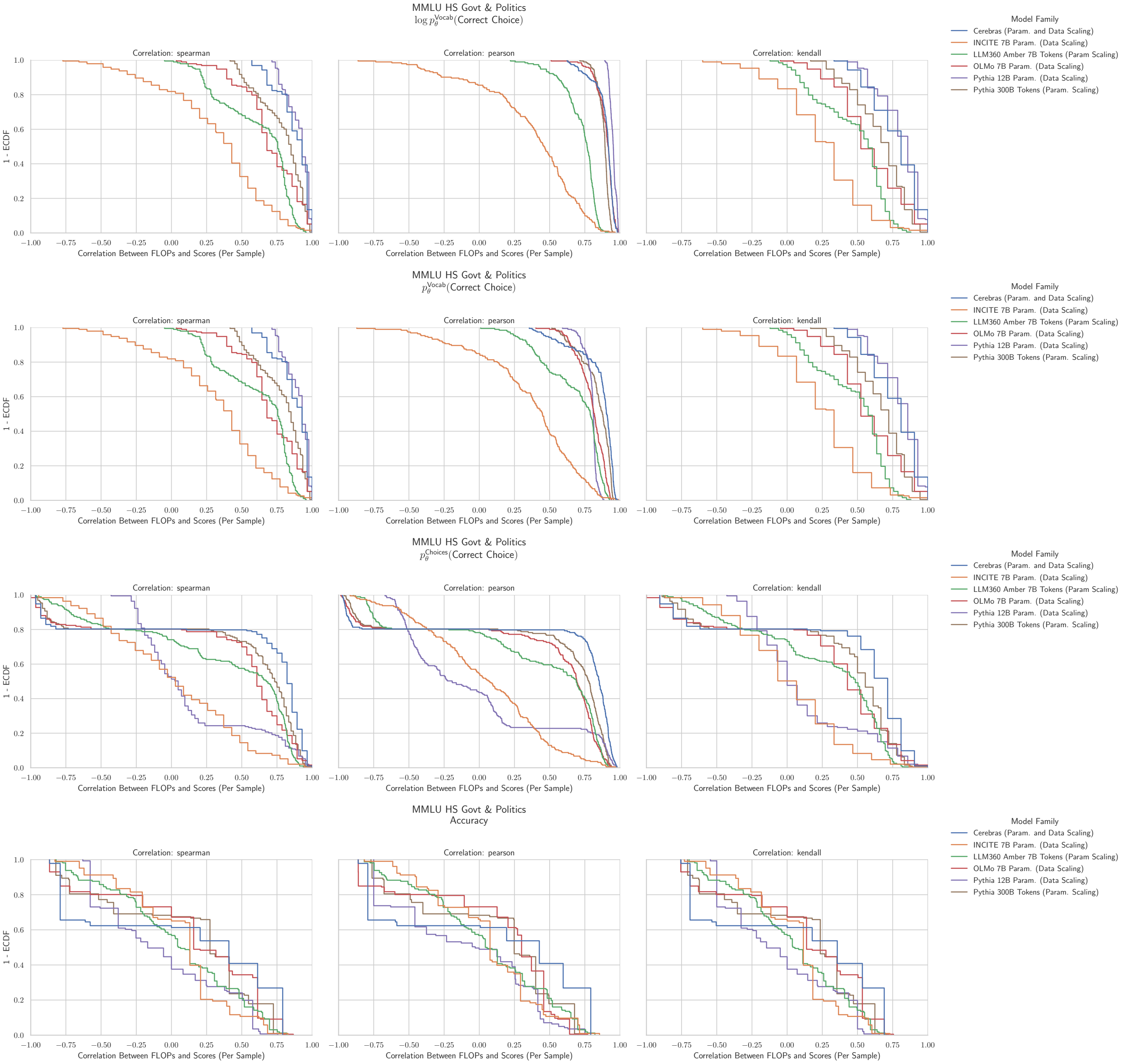

Figure 39. **MMLU High School Government & Politics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.31. NLP Benchmark: MMLU High School Macroeconomics (Hendrycks et al., 2020)

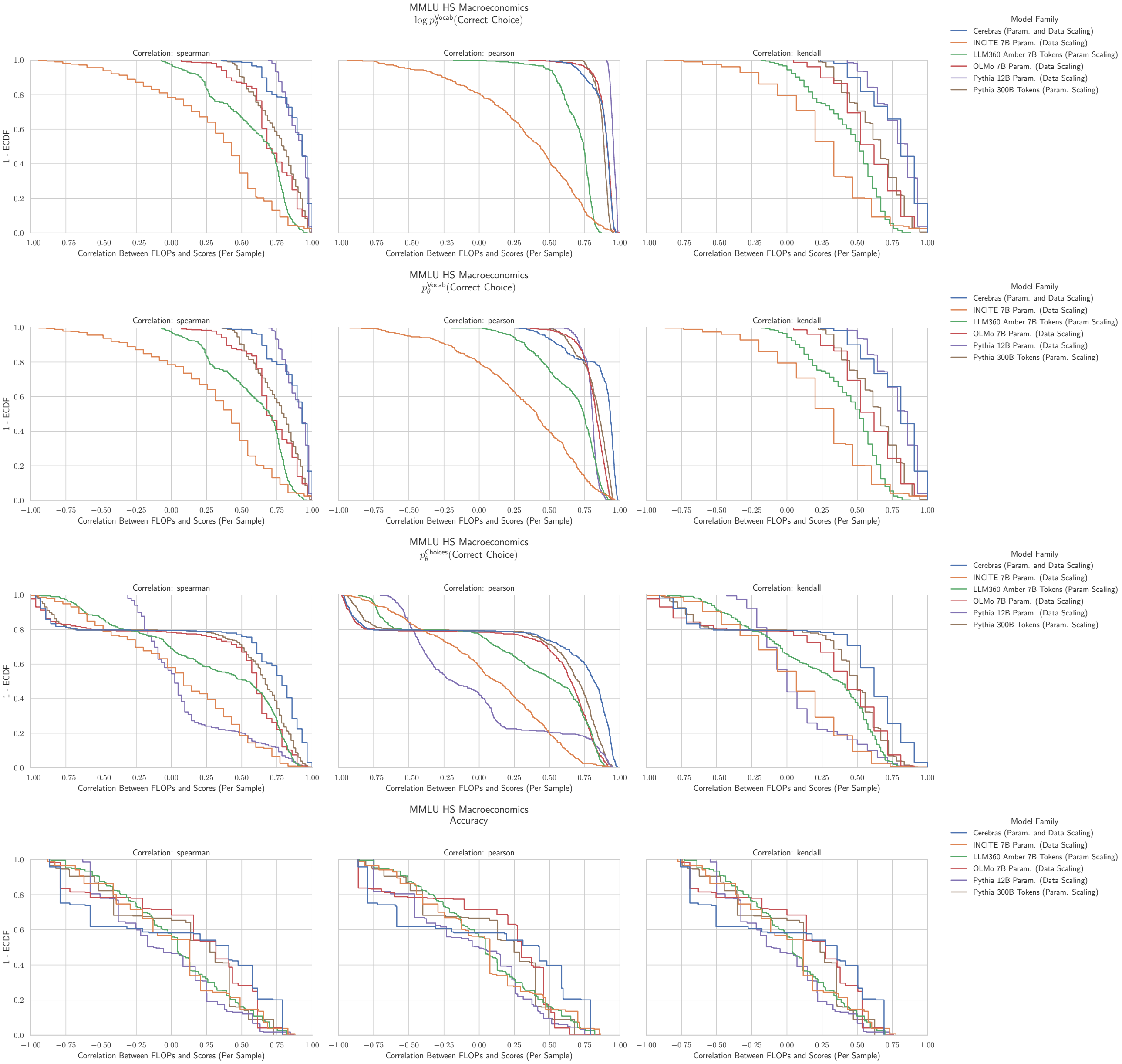

Figure 40. **MMLU High School Macroeconomics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.32. NLP Benchmark: MMLU High School Mathematics (Hendrycks et al., 2020)

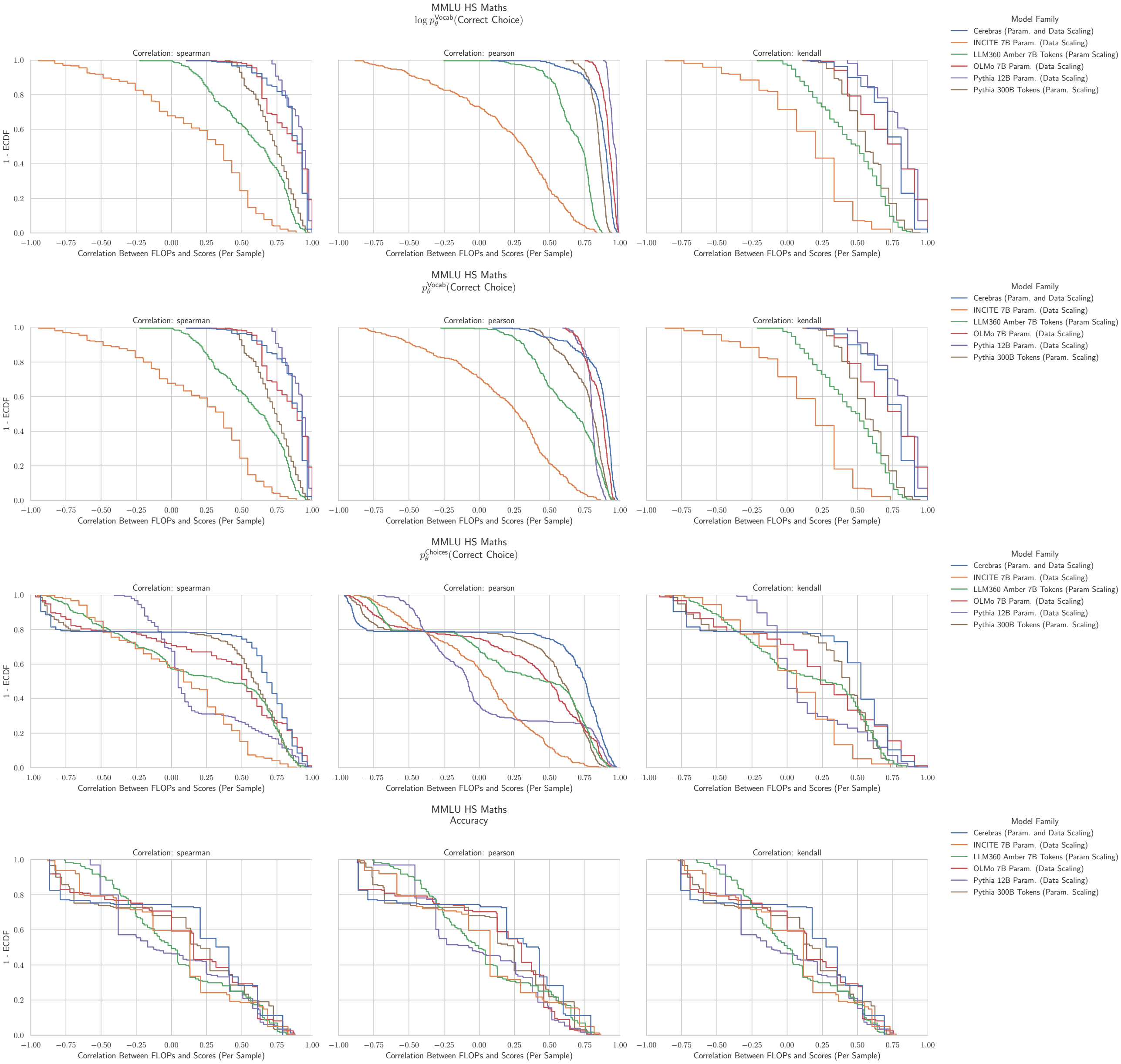

Figure 41. **MMLU High School Mathematics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.33. NLP Benchmark: MMLU High School Microeconomics (Hendrycks et al., 2020)

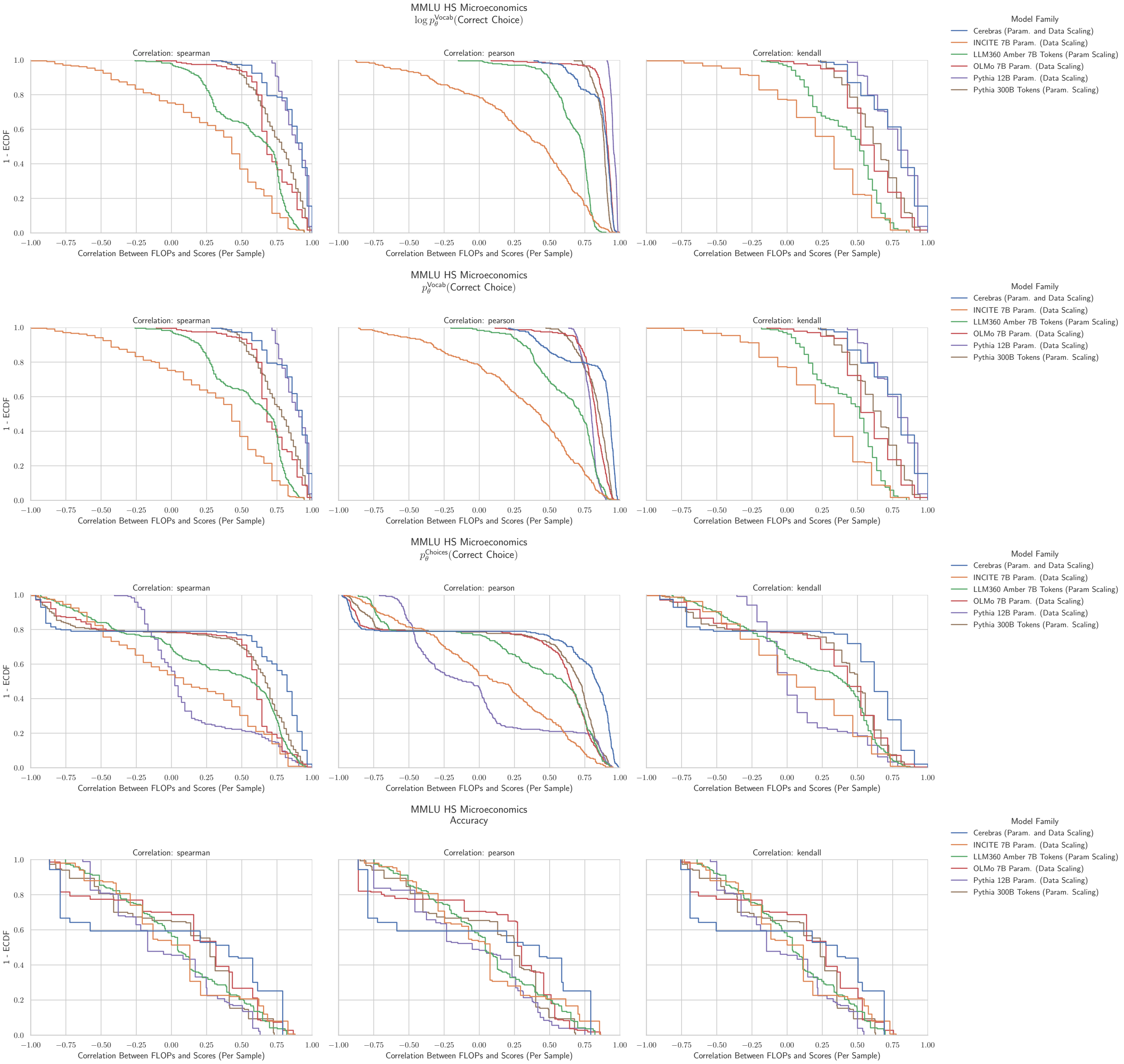

Figure 42. **MMLU High School Microeconomics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.34. NLP Benchmark: MMLU High School Physics (Hendrycks et al., 2020)

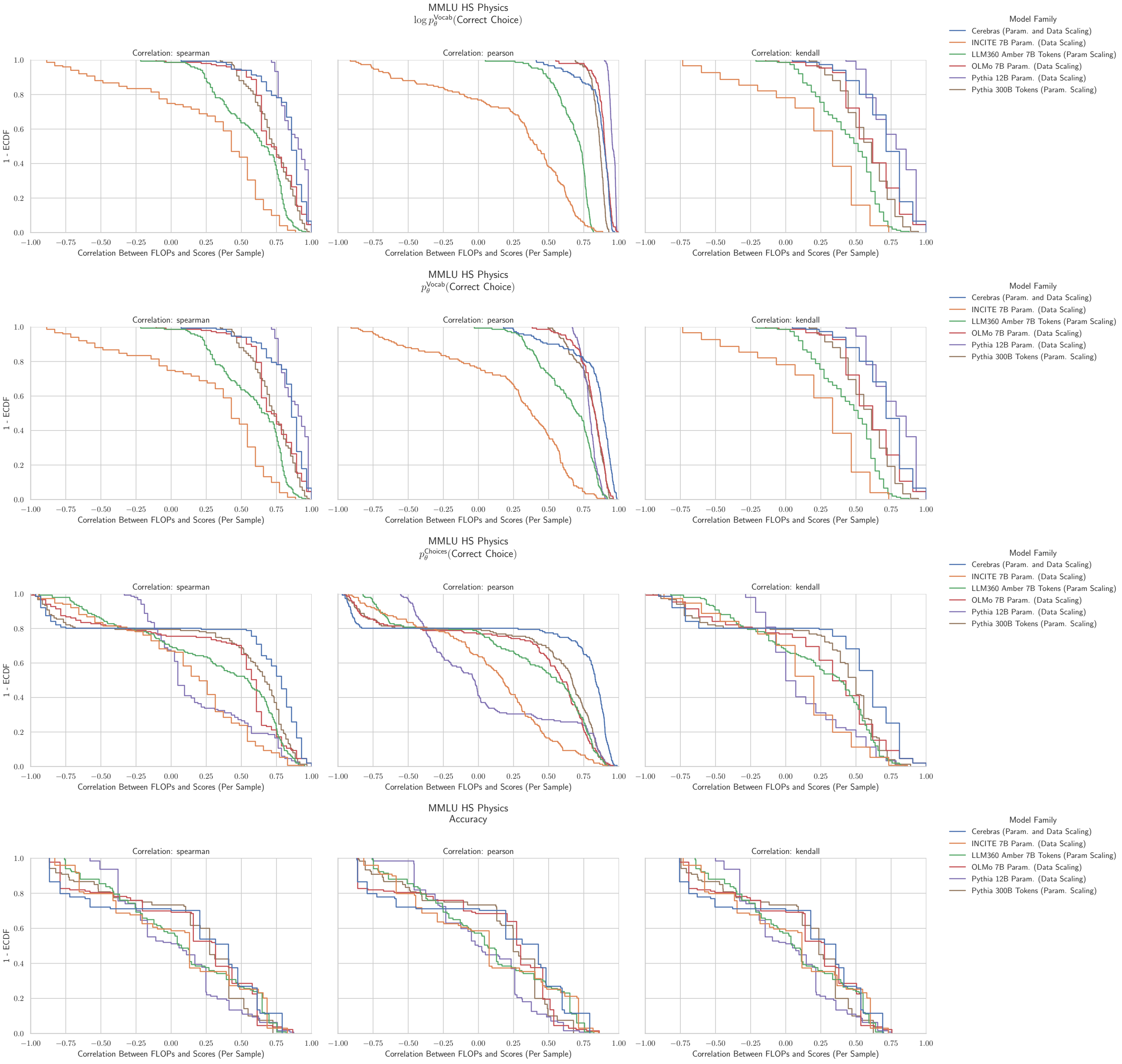

Figure 43. **MMLU High School Physics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.35. NLP Benchmark: MMLU High School Psychology (Hendrycks et al., 2020)

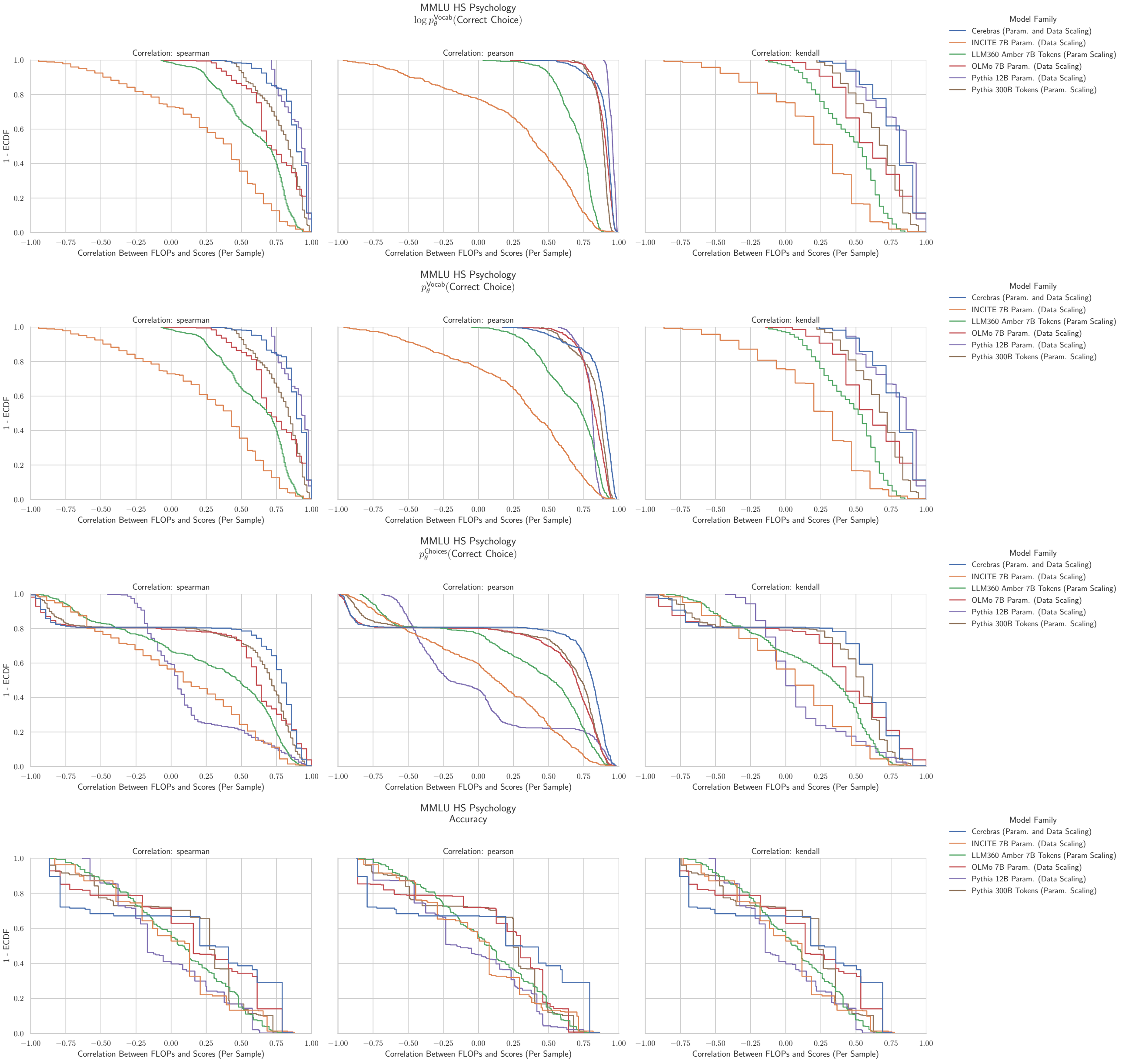

Figure 44. **MMLU High School Psychology: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.36. NLP Benchmark: MMLU High School Statistics (Hendrycks et al., 2020)

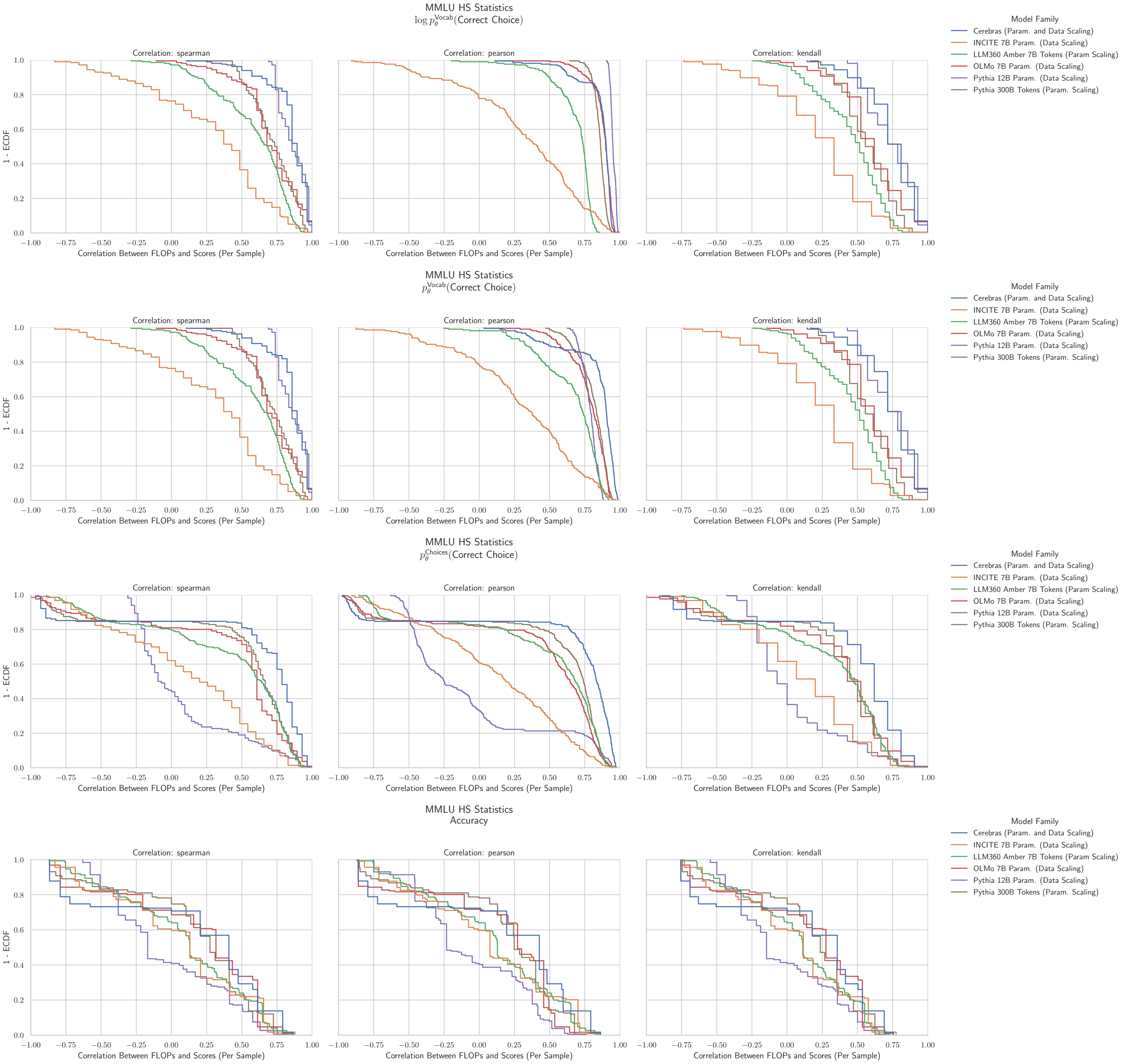

Figure 45. **MMLU High School Statistics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.37. NLP Benchmark: MMLU High School US History (Hendrycks et al., 2020)

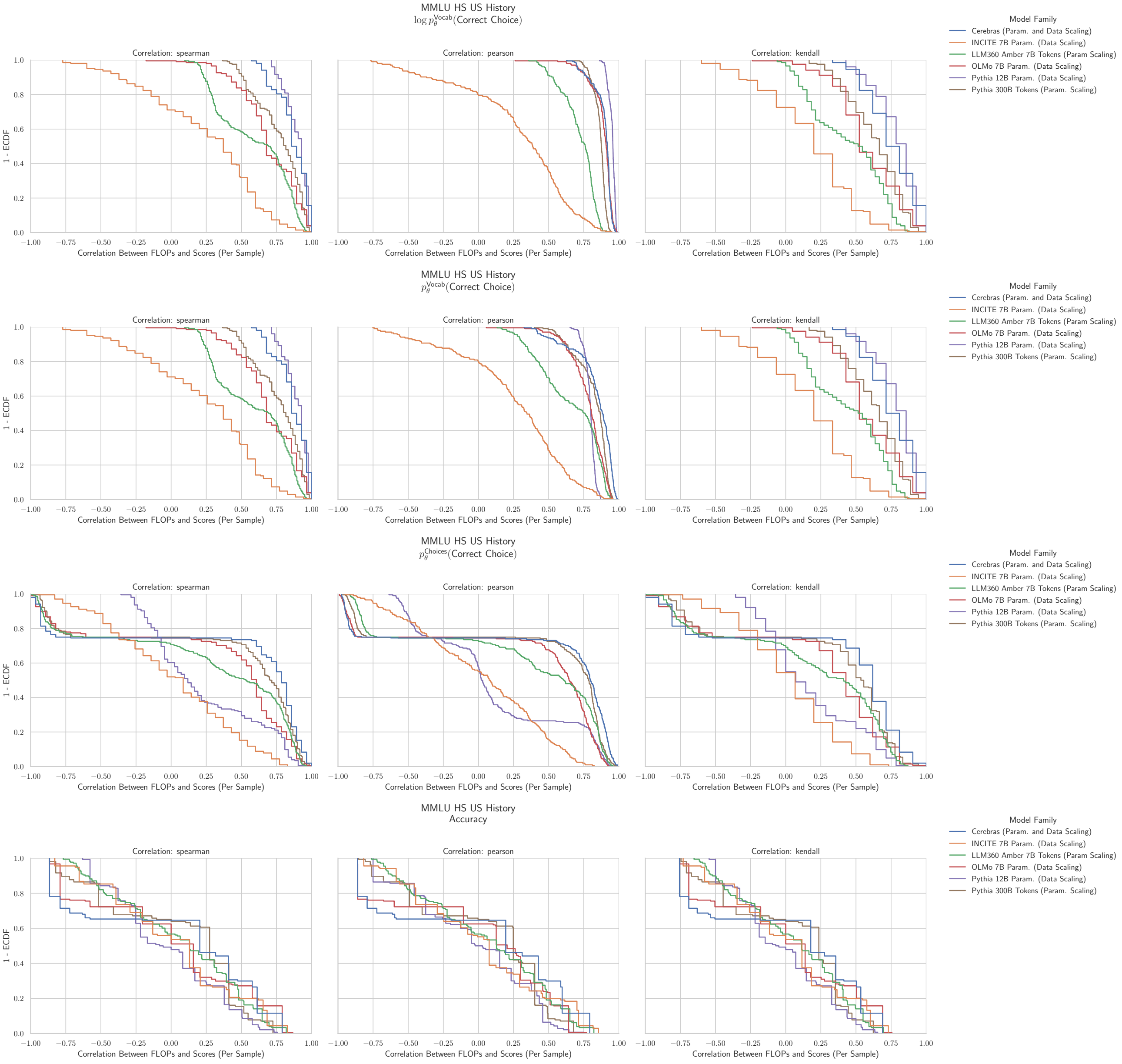

Figure 46. **MMLU High School US History: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.38. NLP Benchmark: MMLU High School World History (Hendrycks et al., 2020)

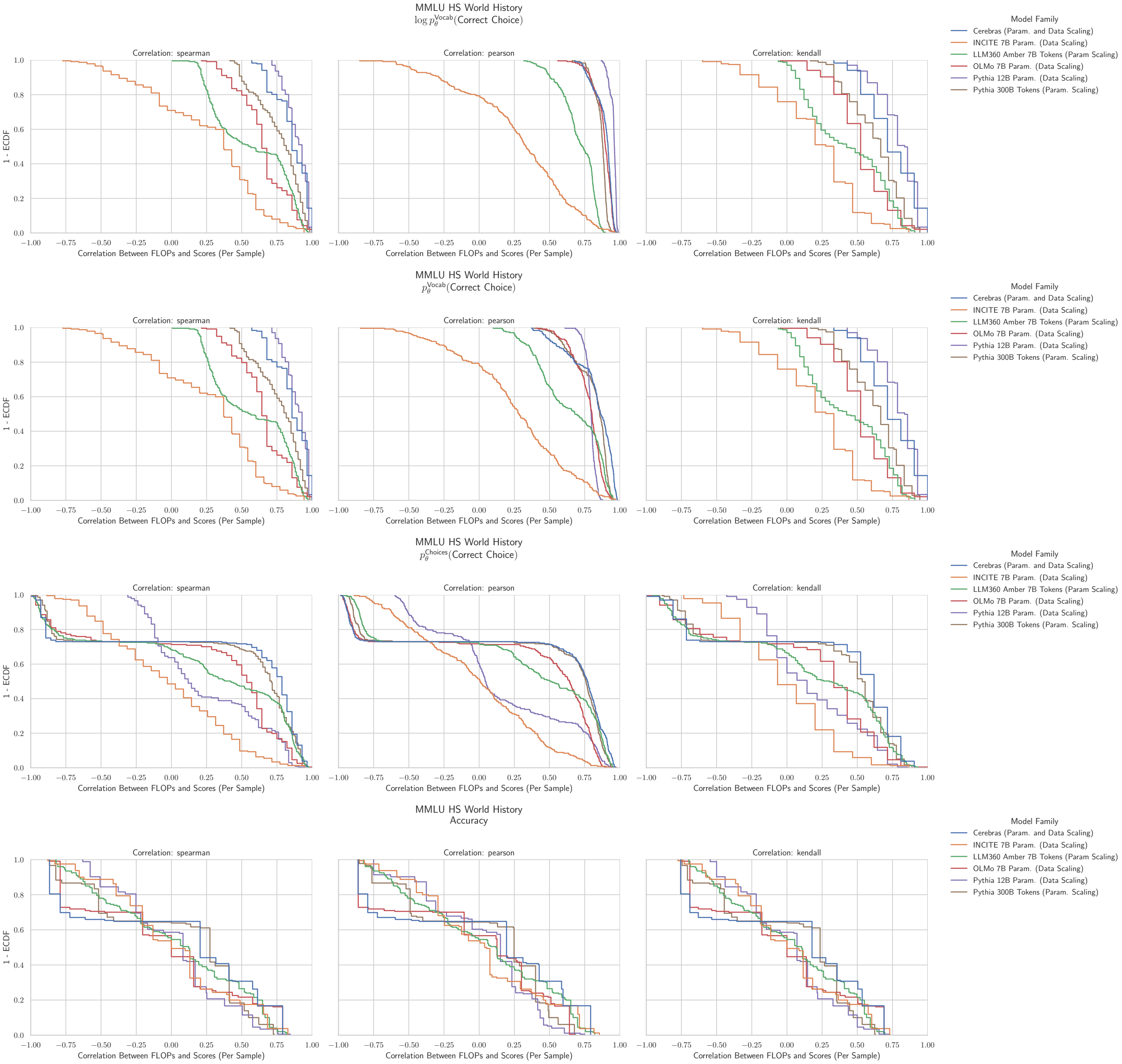

Figure 47. **MMLU High School World History: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.39. NLP Benchmark: MMLU Human Aging (Hendrycks et al., 2020)

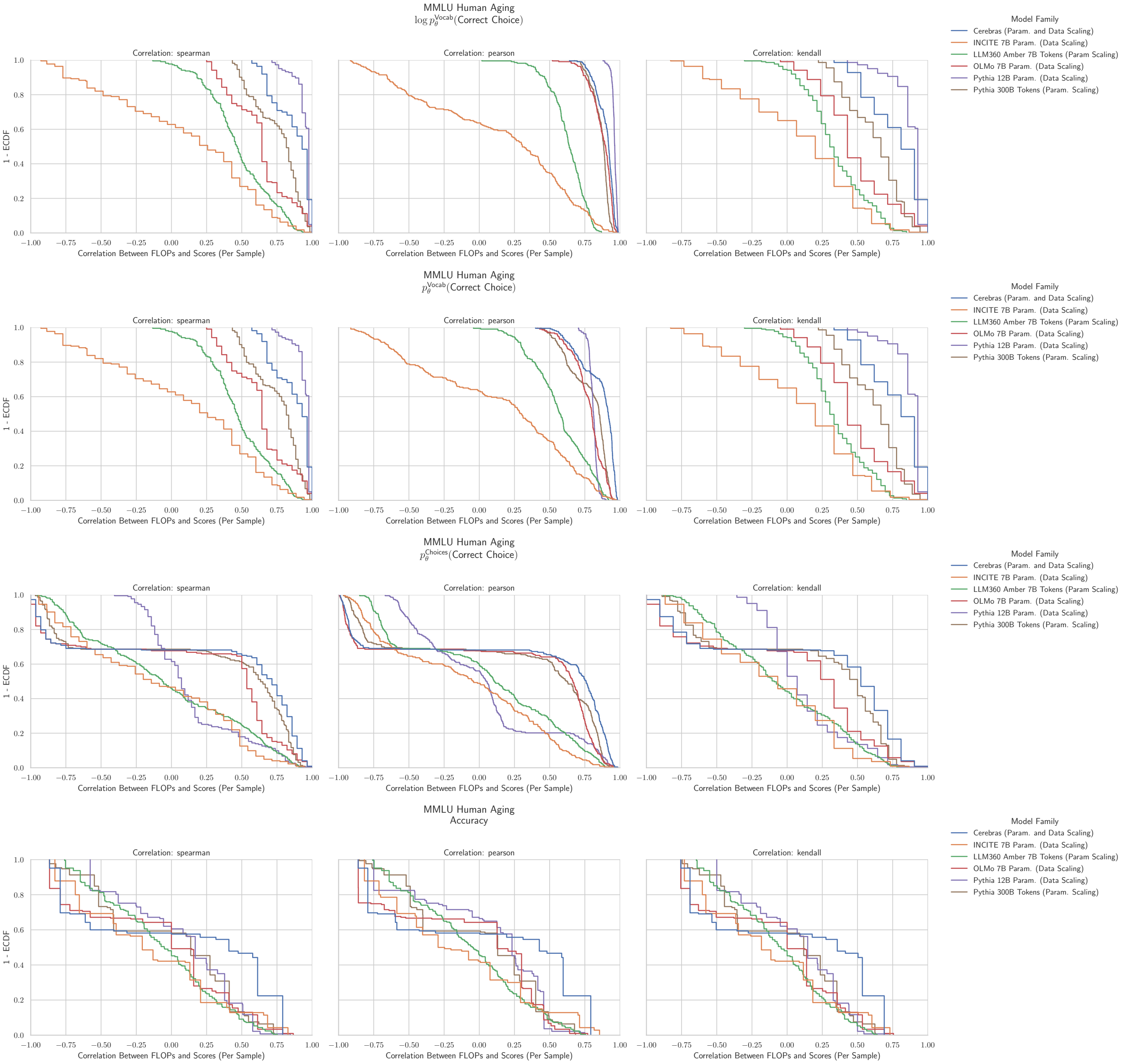

Figure 48. **MMLU Human Aging: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.40. NLP Benchmark: MMLU Human Sexuality (Hendrycks et al., 2020)

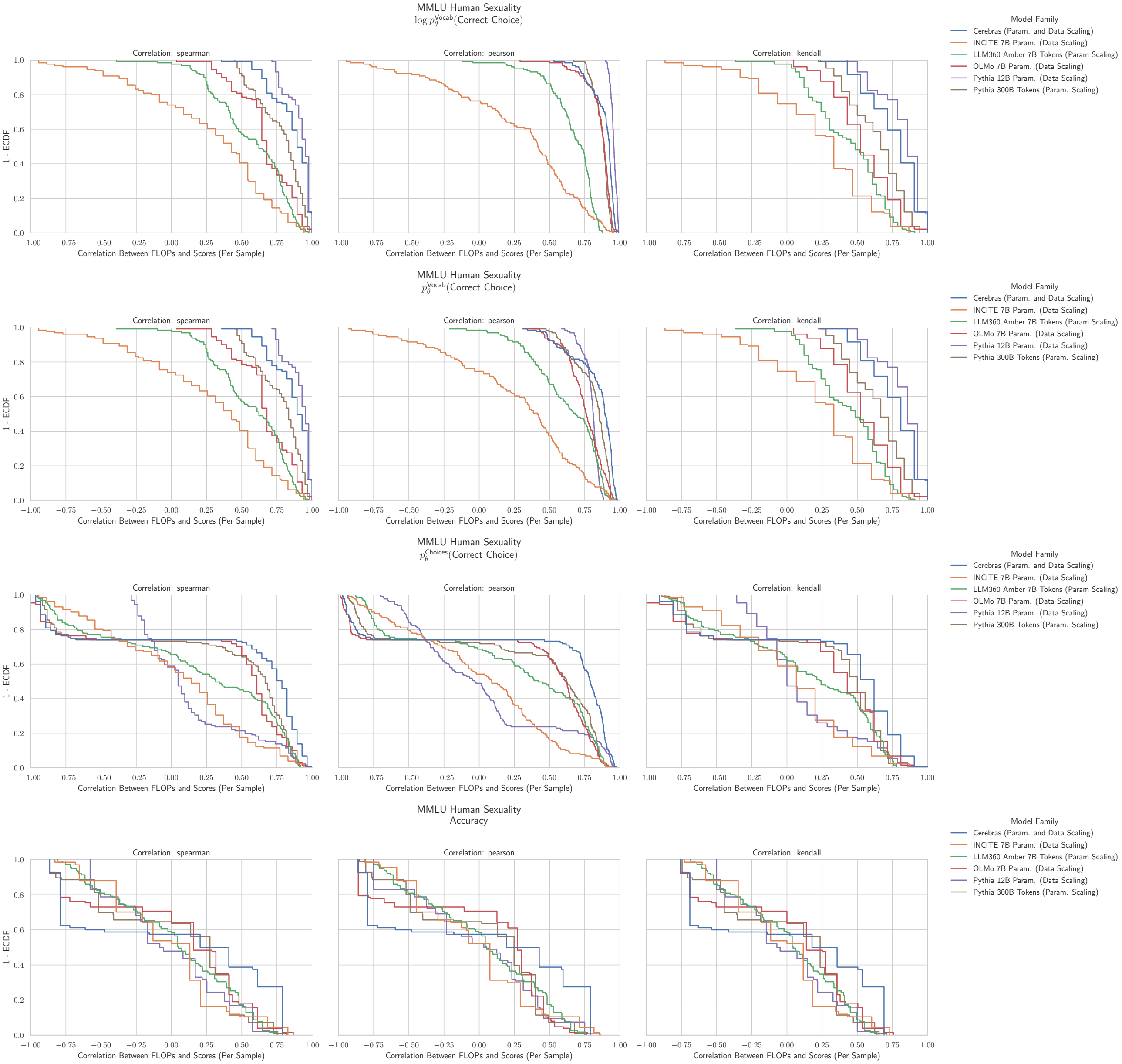

Figure 49. **MMLU Human Sexuality: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.41. NLP Benchmark: MMLU International Law (Hendrycks et al., 2020)

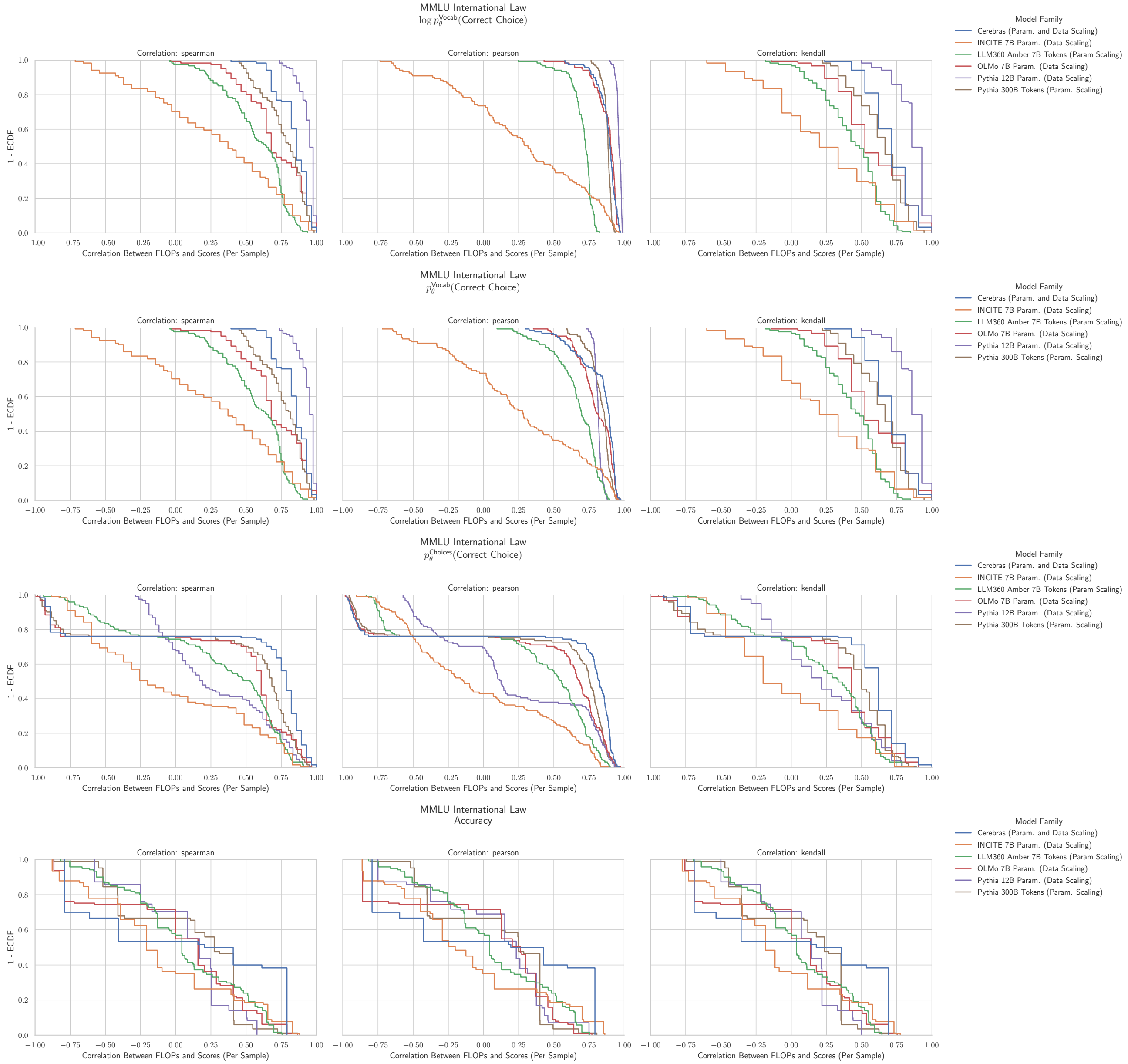

Figure 50. **MMLU International Law: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.42. NLP Benchmark: MMLU Jurisprudence (Hendrycks et al., 2020)

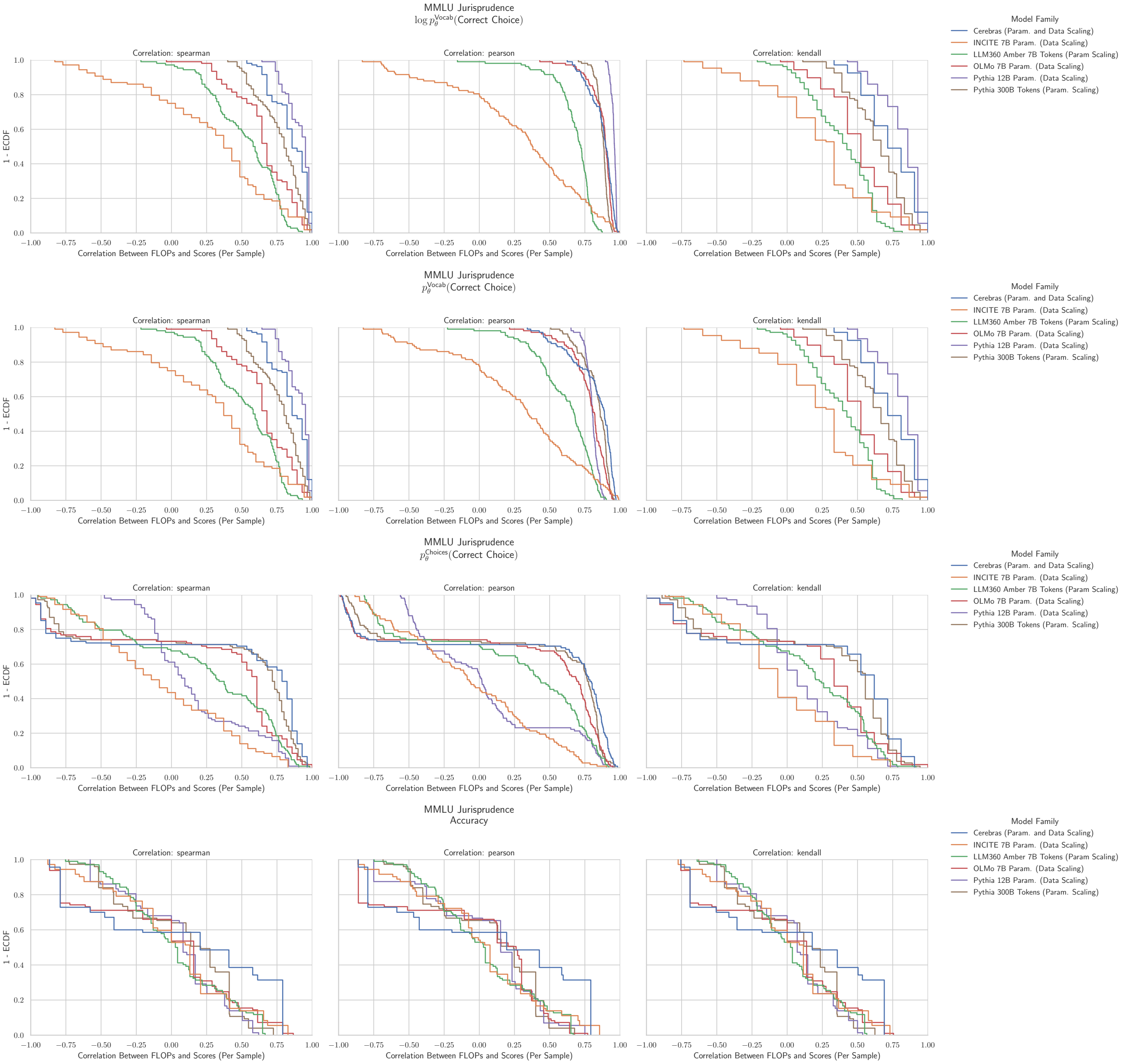

Figure 51. **MMLU Jurisprudence: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.43. NLP Benchmark: MMLU Logical Fallacies (Hendrycks et al., 2020)

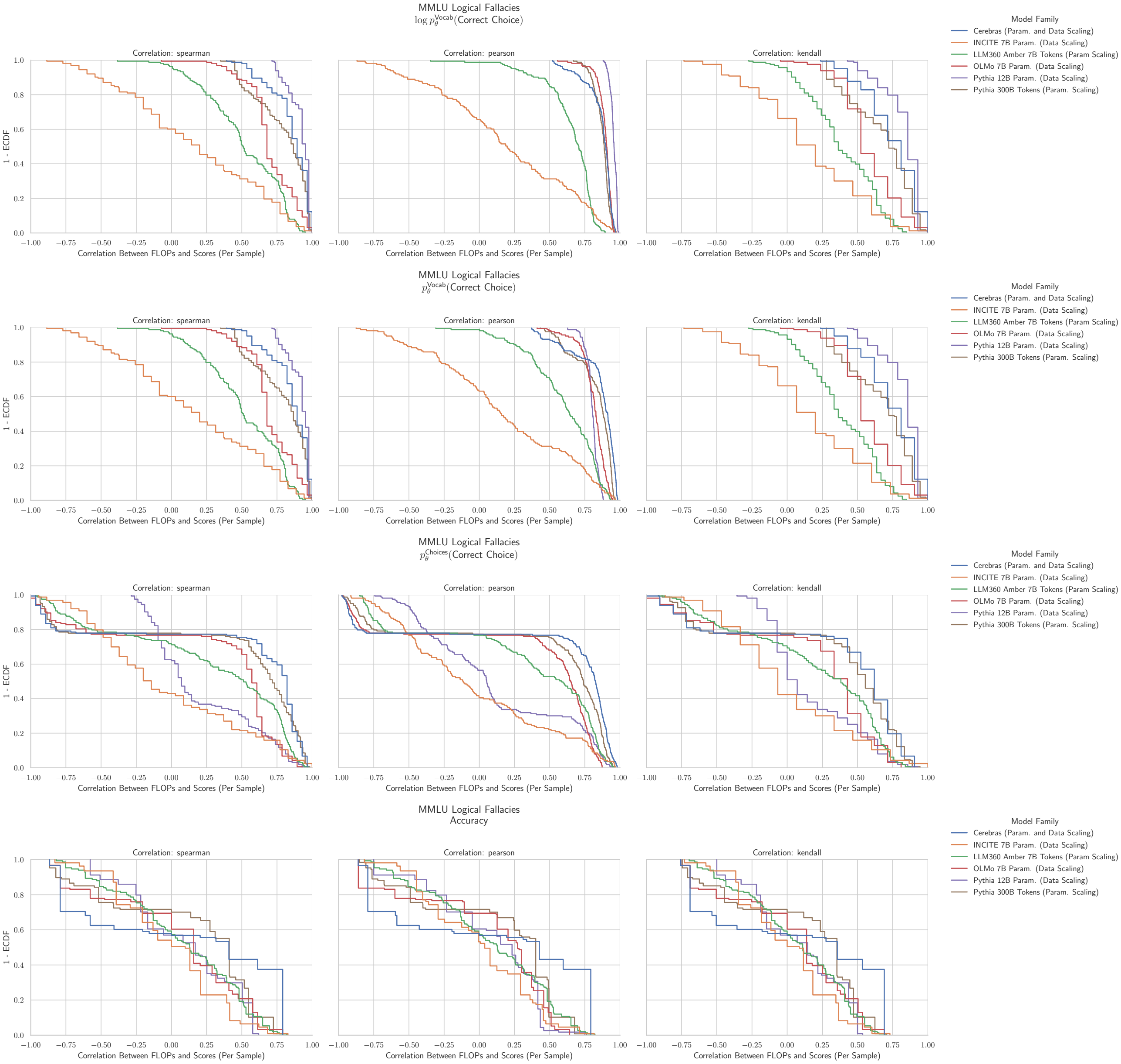

Figure 52. **MMLU Logical Fallacies: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.44. NLP Benchmark: MMLU Machine Learning (Hendrycks et al., 2020)

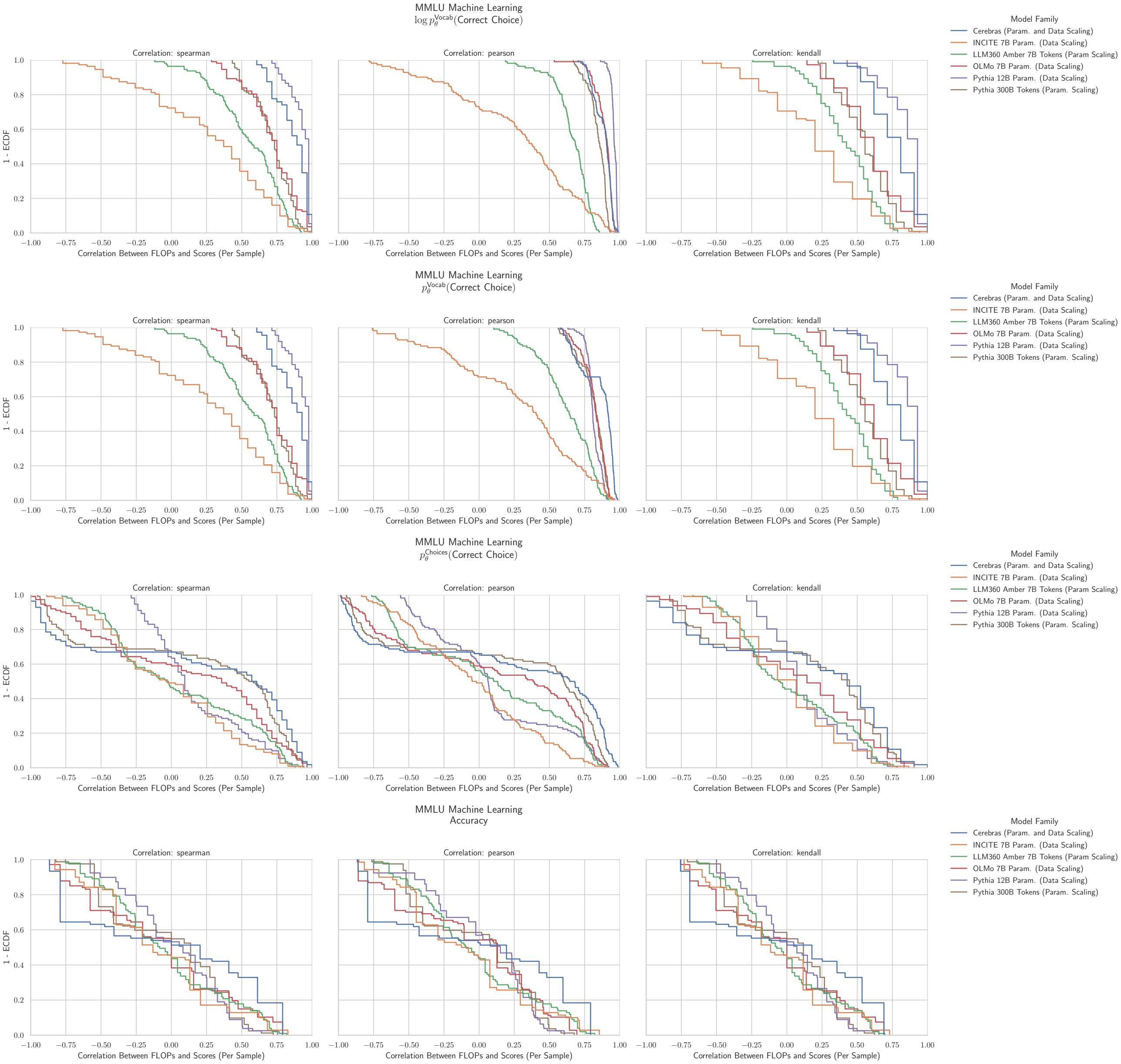

Figure 53. **MMLU Machine Learning: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.45. NLP Benchmark: MMLU Management (Hendrycks et al., 2020)

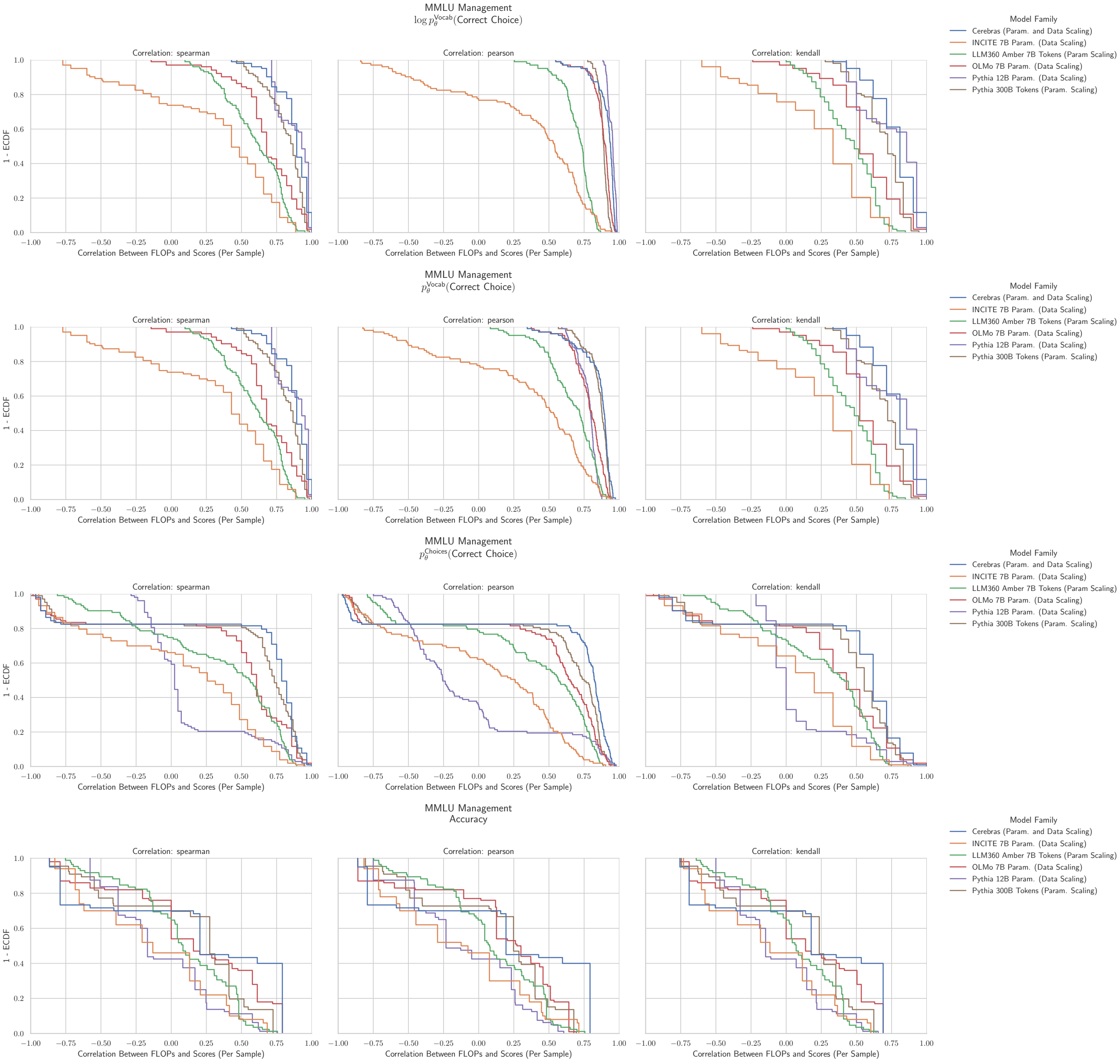

Figure 54. **MMLU Management: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.46. NLP Benchmark: MMLU Marketing (Hendrycks et al., 2020)

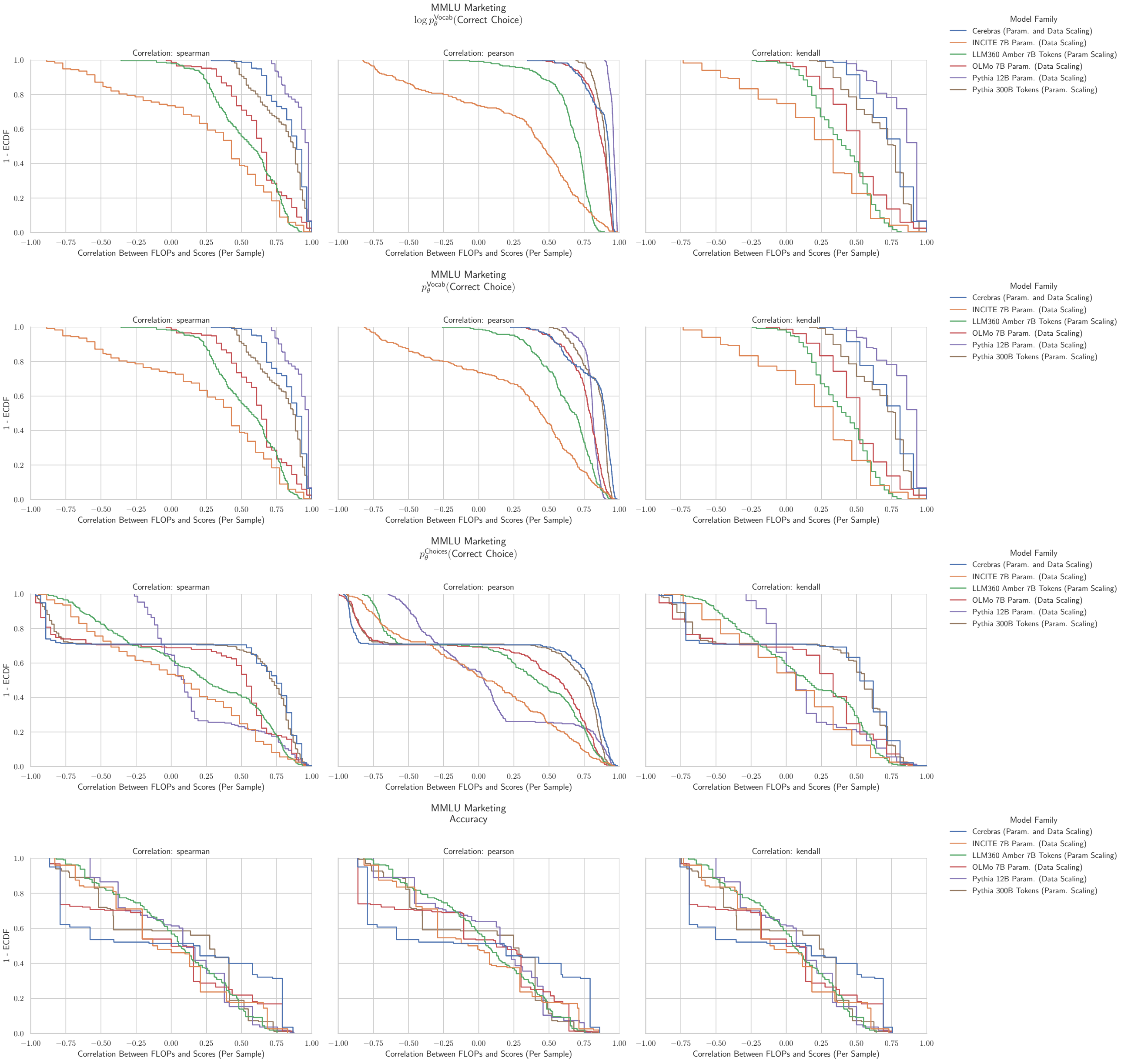

Figure 55. **MMLU Marketing: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.47. NLP Benchmark: MMLU Medical Genetics (Hendrycks et al., 2020)

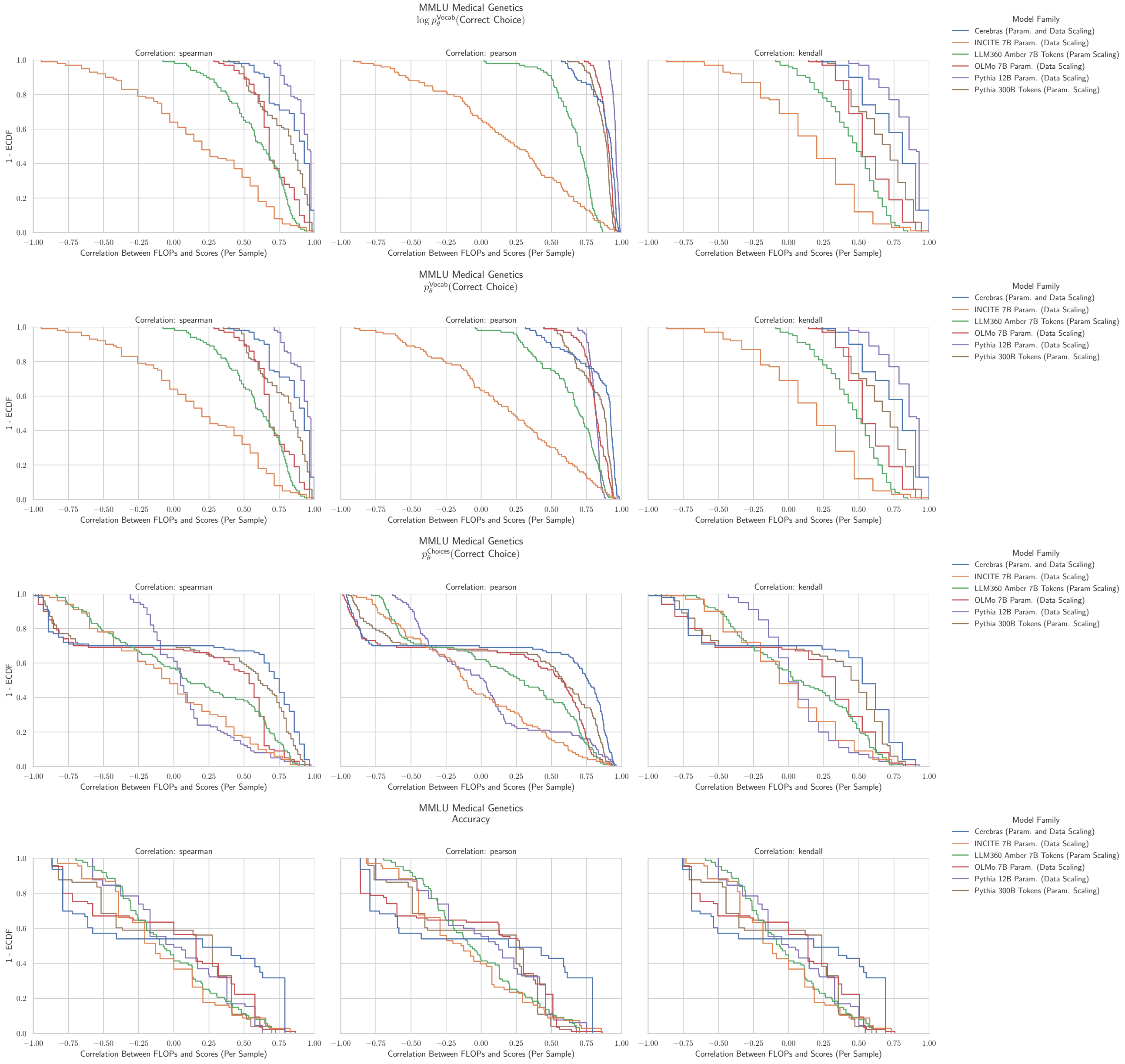

Figure 56. **MMLU Medical Genetics: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.48. NLP Benchmark: MMLU Miscellaneous (Hendrycks et al., 2020)

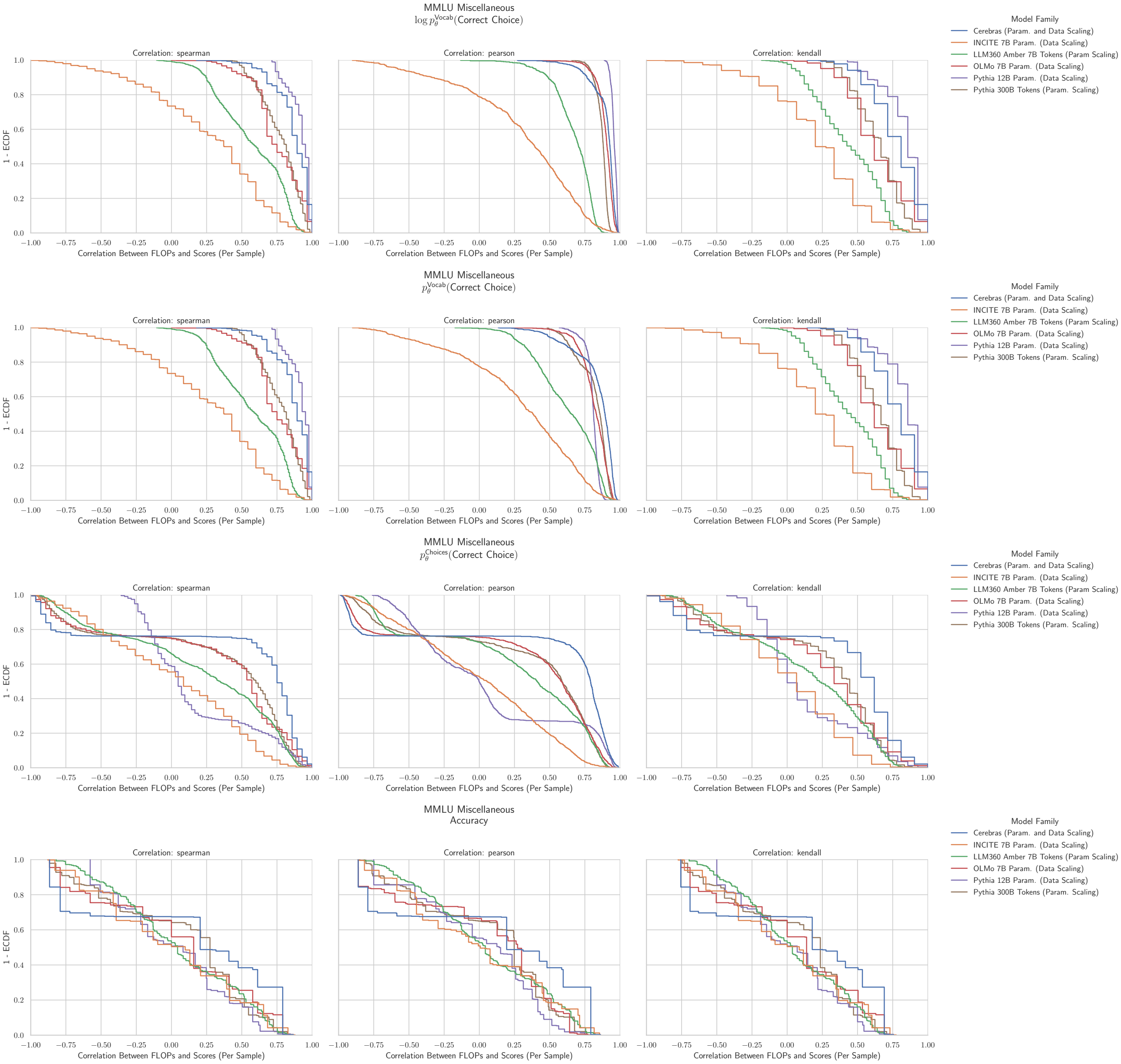

Figure 57. **MMLU Miscellaneous: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.49. NLP Benchmark: MMLU Moral Disputes (Hendrycks et al., 2020)

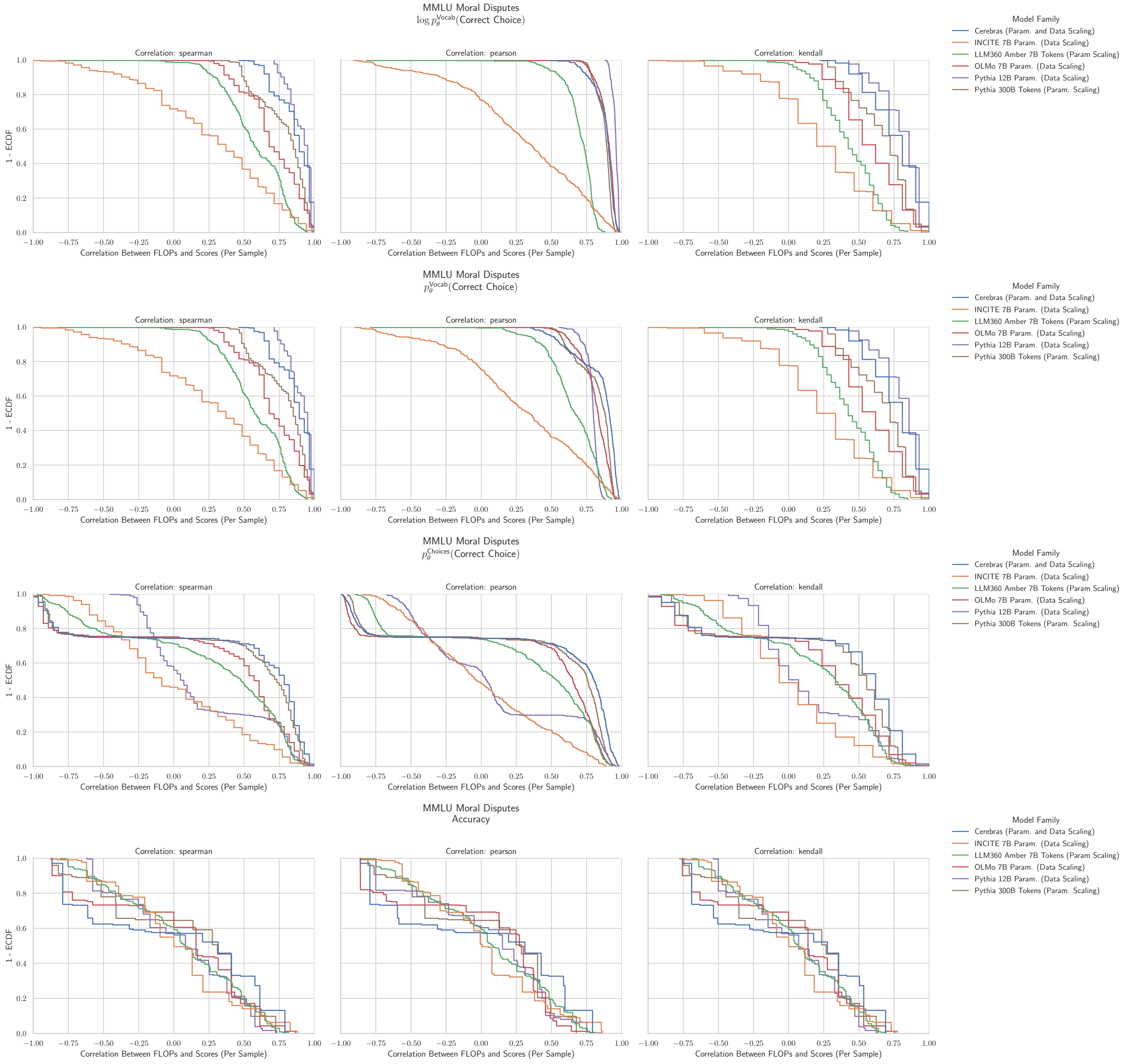

Figure 58. **MMLU Moral Disputes: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.50. NLP Benchmark: MMLU Moral Scenarios (Hendrycks et al., 2020)

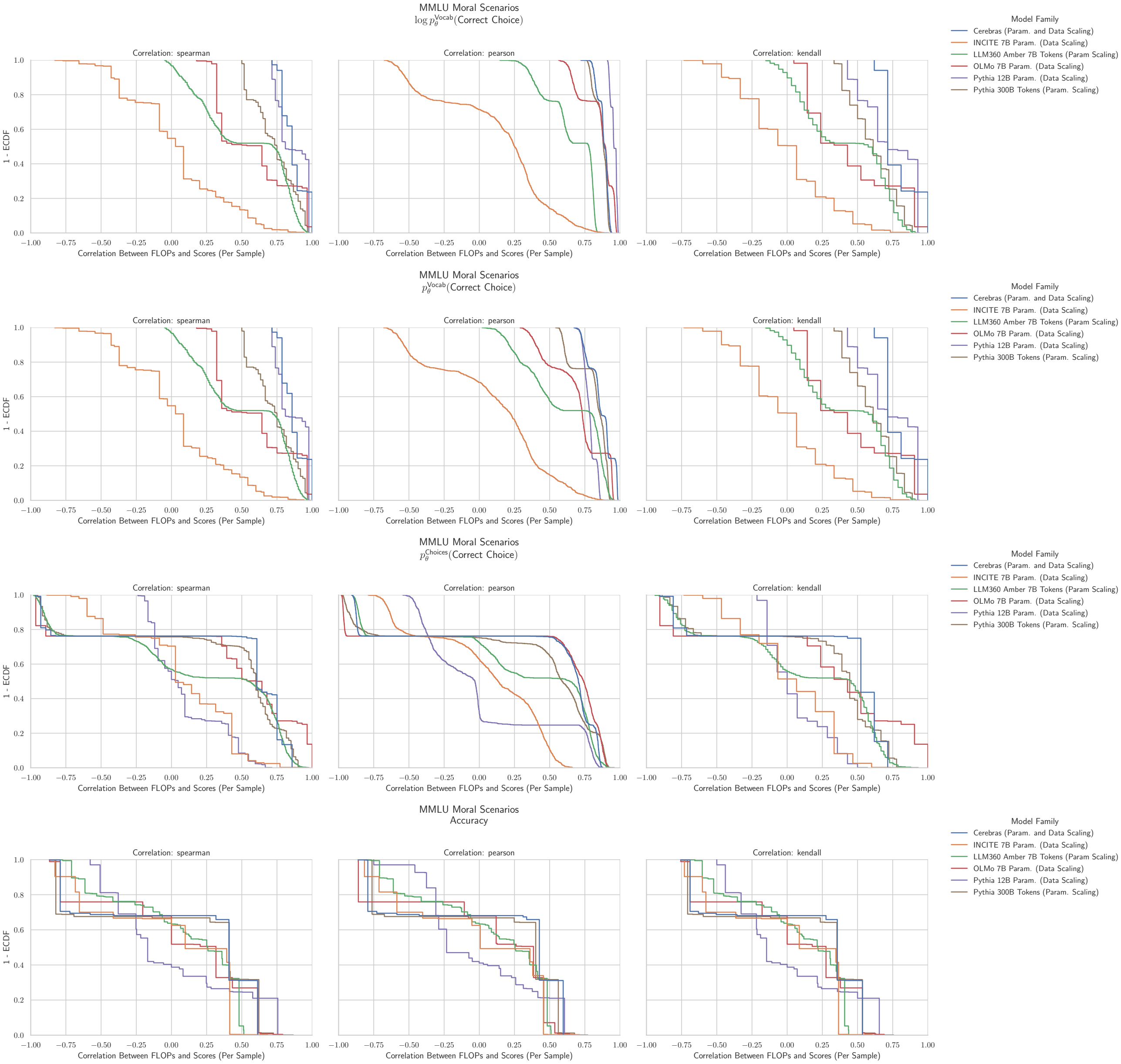

Figure 59. **MMLU Moral Scenarios: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.51. NLP Benchmark: MMLU Nutrition (Hendrycks et al., 2020)

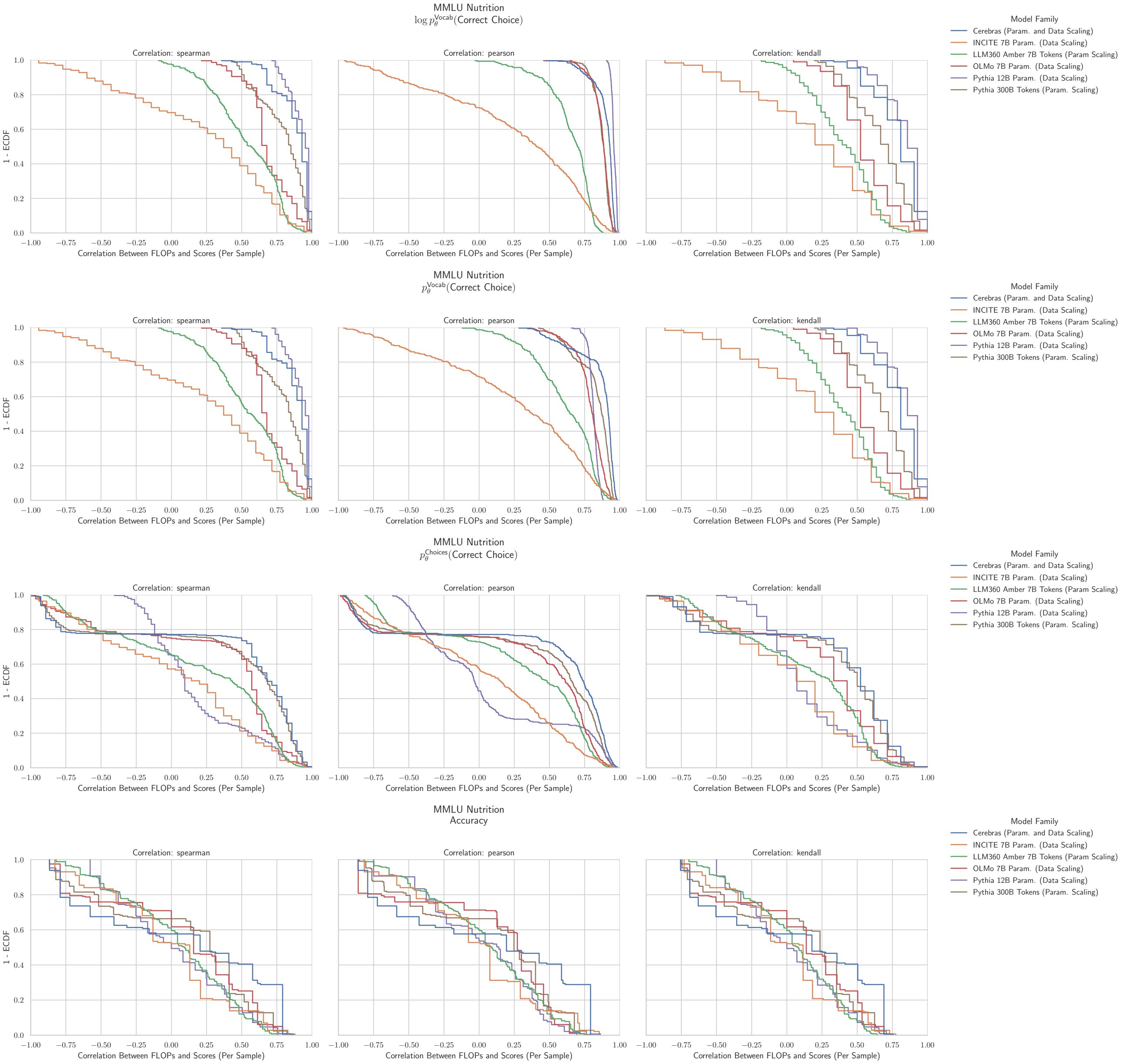

Figure 60. **MMLU Nutrition: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.52. NLP Benchmark: MMLU Philosophy (Hendrycks et al., 2020)

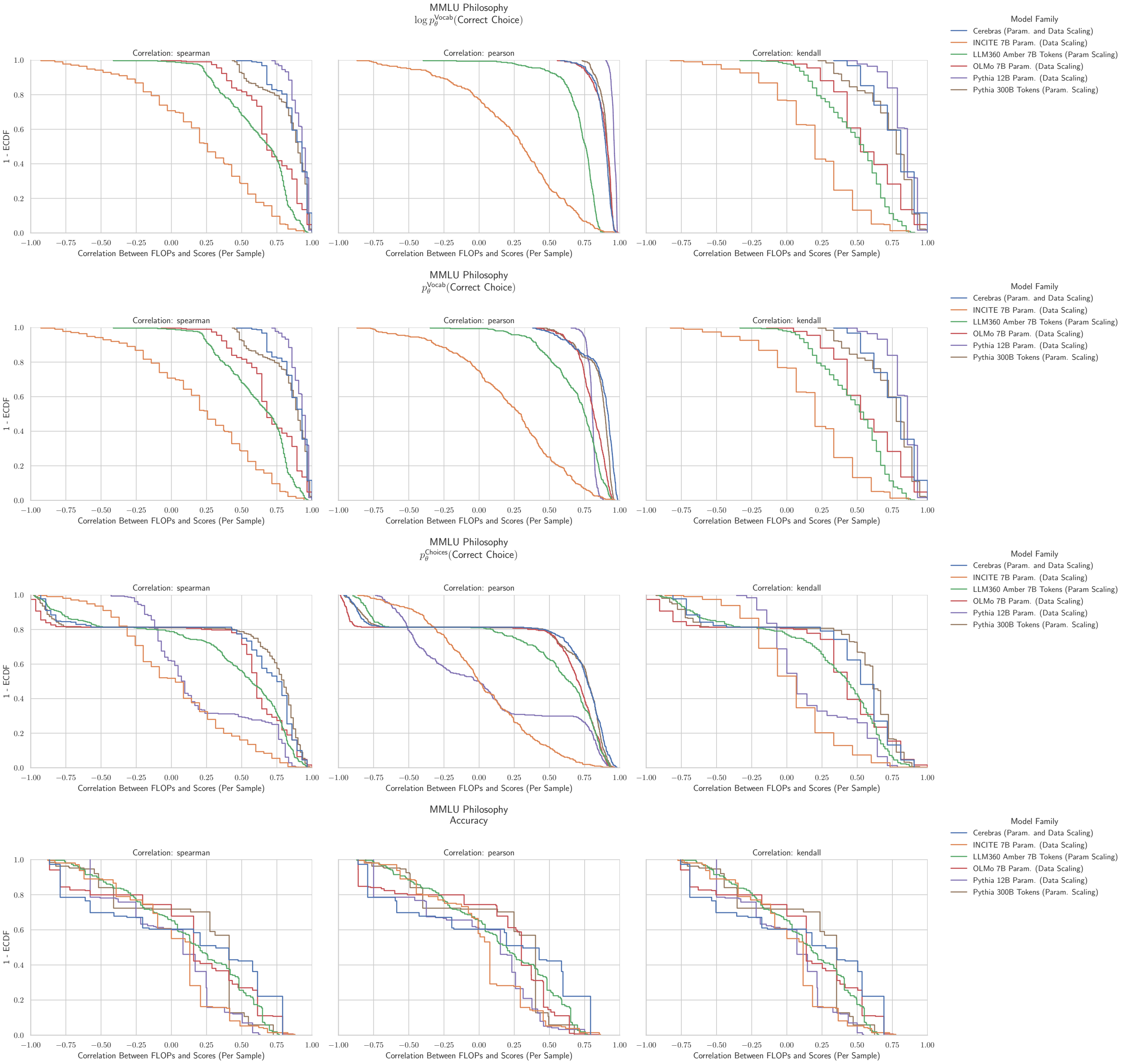

Figure 61. **MMLU Philosophy: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.53. NLP Benchmark: MMLU Prehistory (Hendrycks et al., 2020)

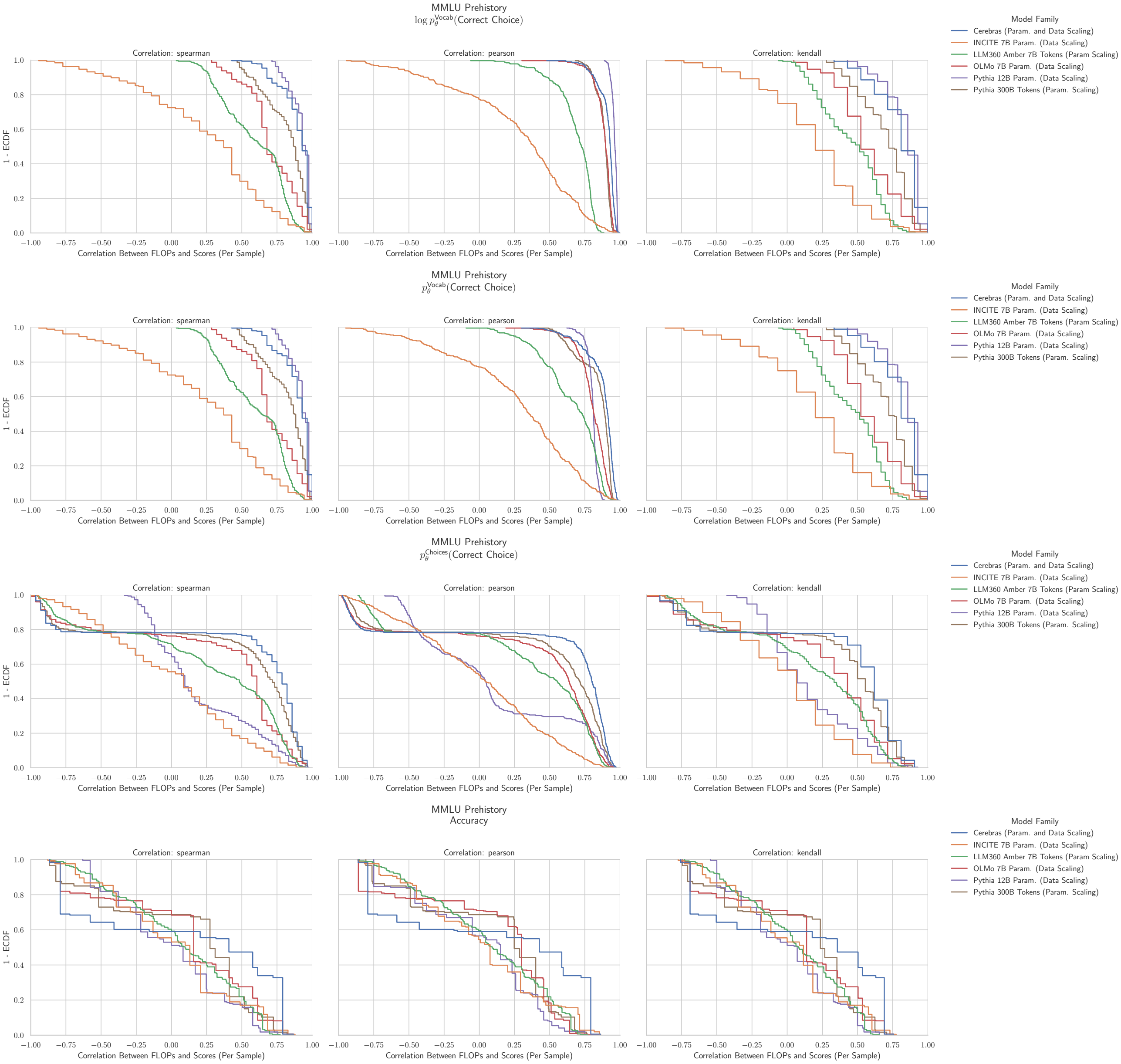

Figure 62. **MMLU Prehistory: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.54. NLP Benchmark: MMLU Professional Accounting (Hendrycks et al., 2020)

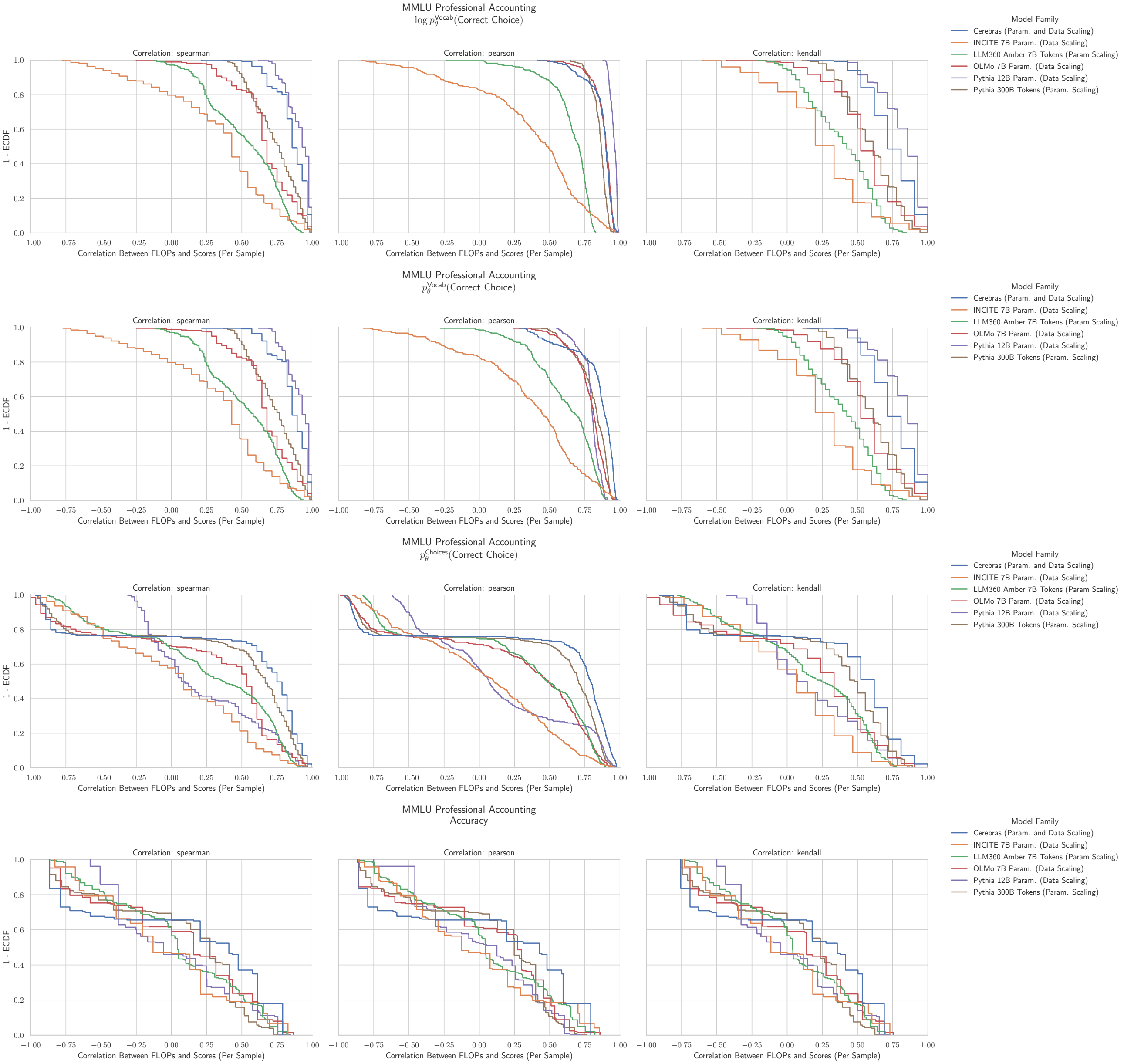

Figure 63. **MMLU Professional Accounting: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.55. NLP Benchmark: MMLU Professional Law (Hendrycks et al., 2020)

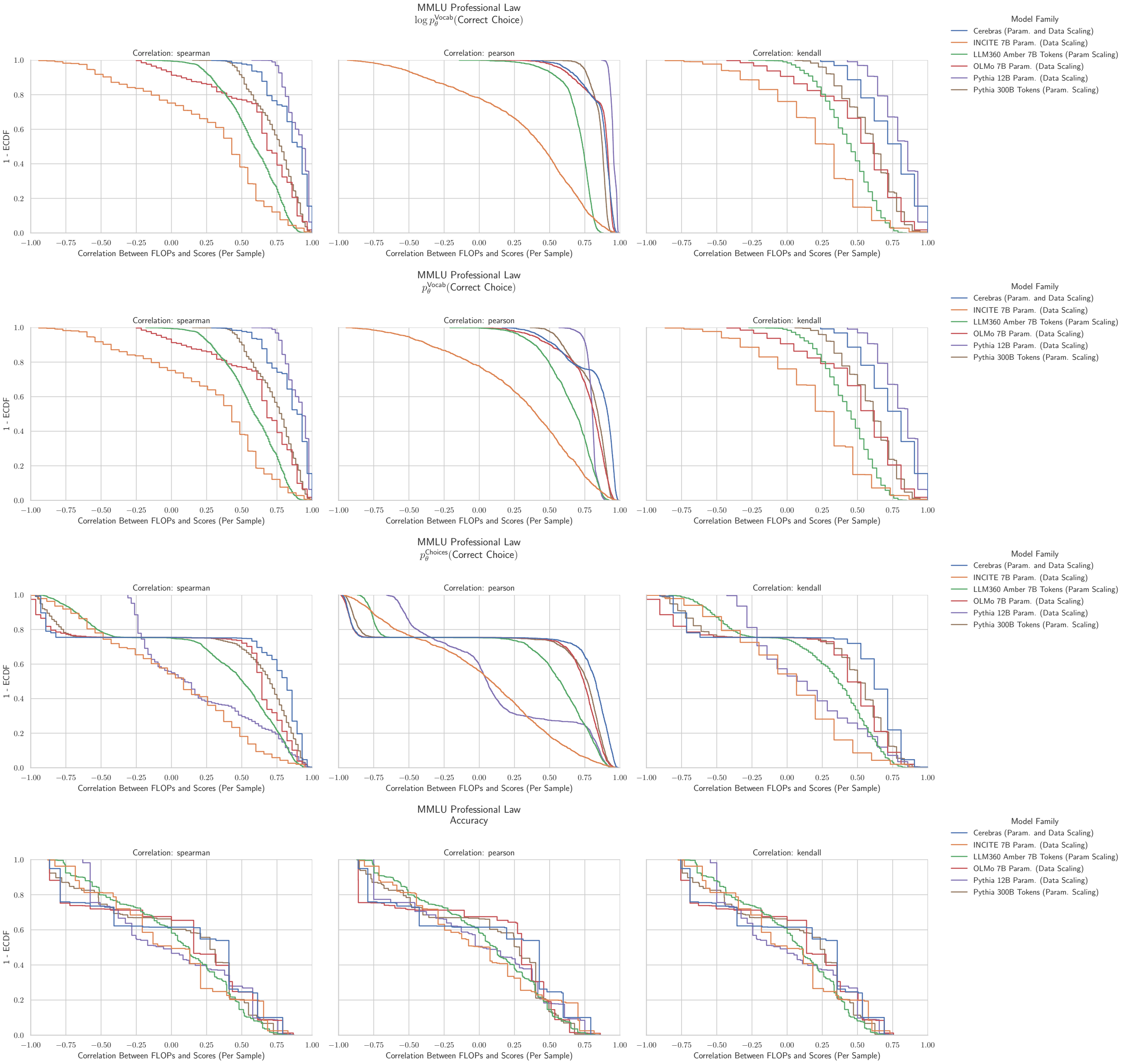

Figure 64. **MMLU Professional Law: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.56. NLP Benchmark: MMLU Professional Medicine (Hendrycks et al., 2020)

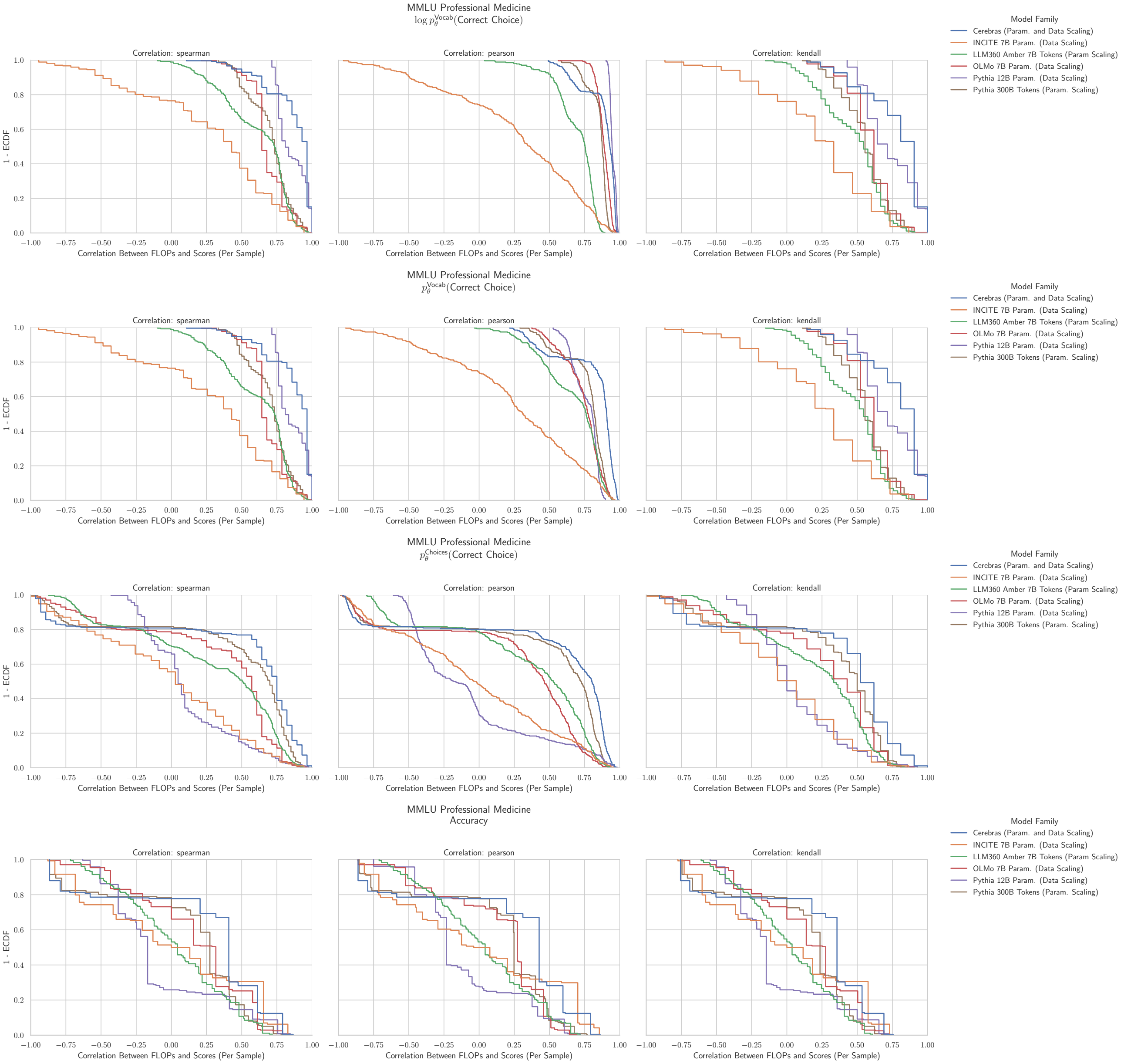

Figure 65. **MMLU Professional Medicine: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.57. NLP Benchmark: MMLU Professional Psychology (Hendrycks et al., 2020)

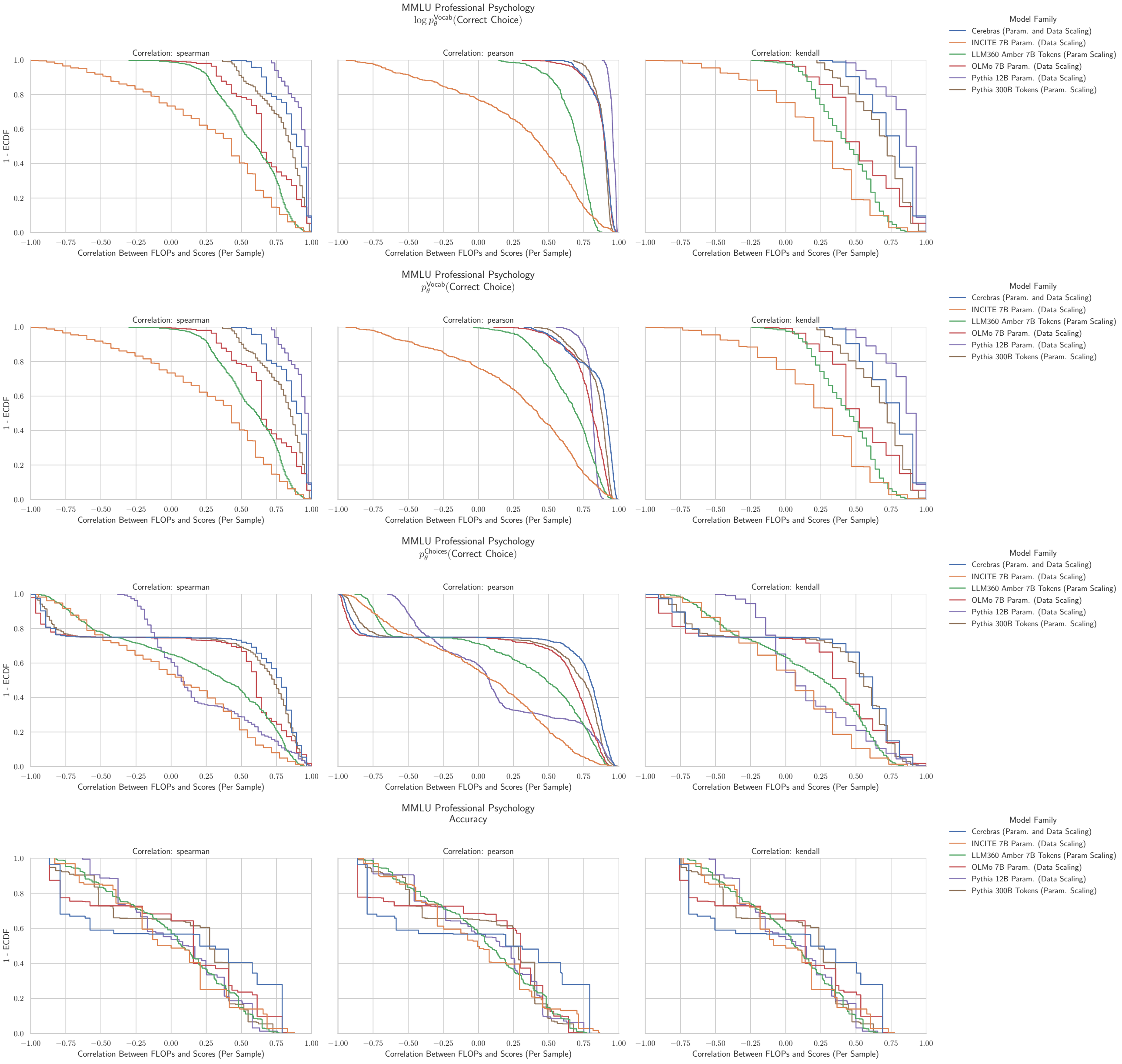

Figure 66. **MMLU Professional Psychology: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.58. NLP Benchmark: MMLU Public Relations (Hendrycks et al., 2020)

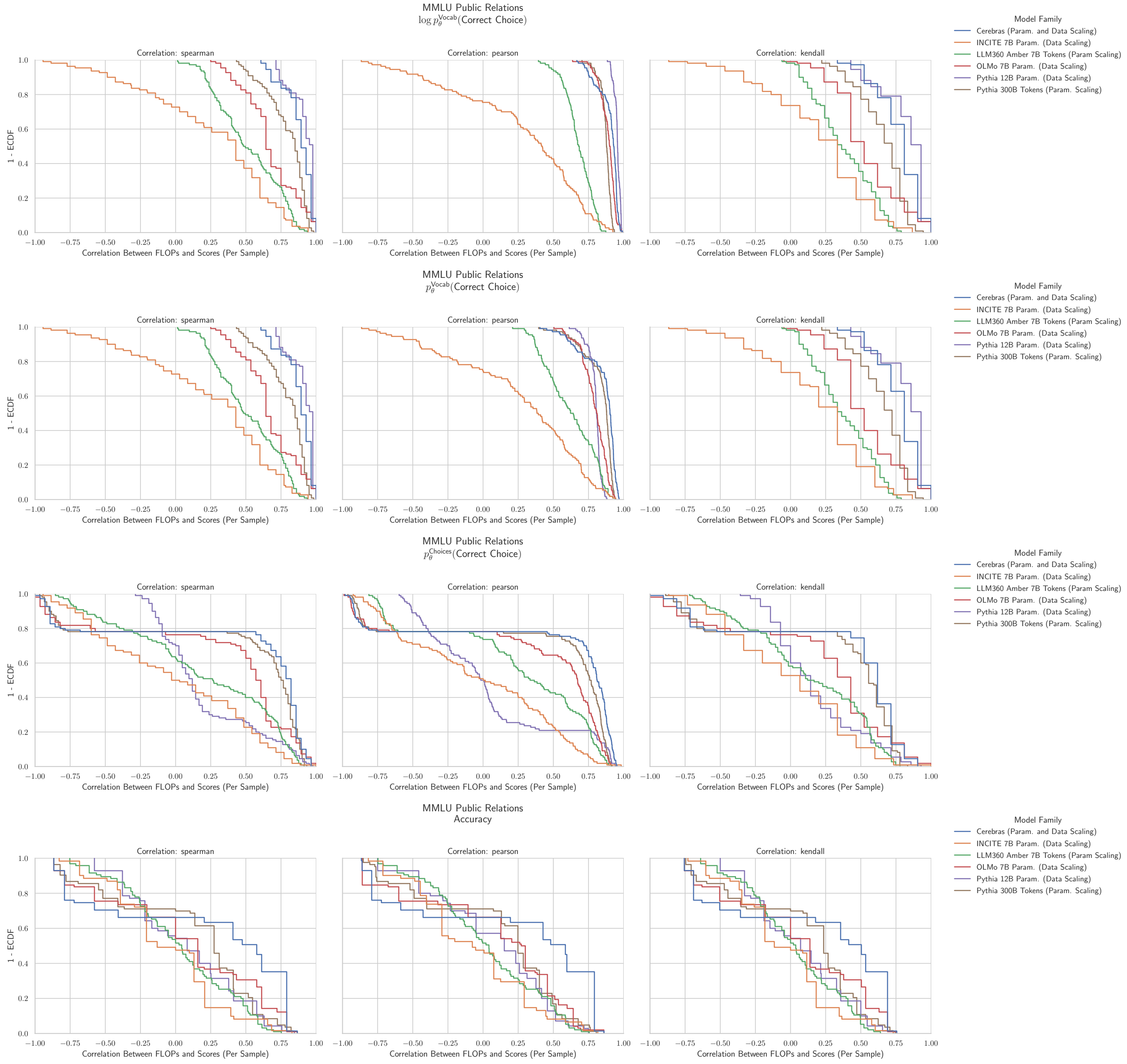

Figure 67. **MMLU Public Relations: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.59. NLP Benchmark: MMLU Security Studies (Hendrycks et al., 2020)

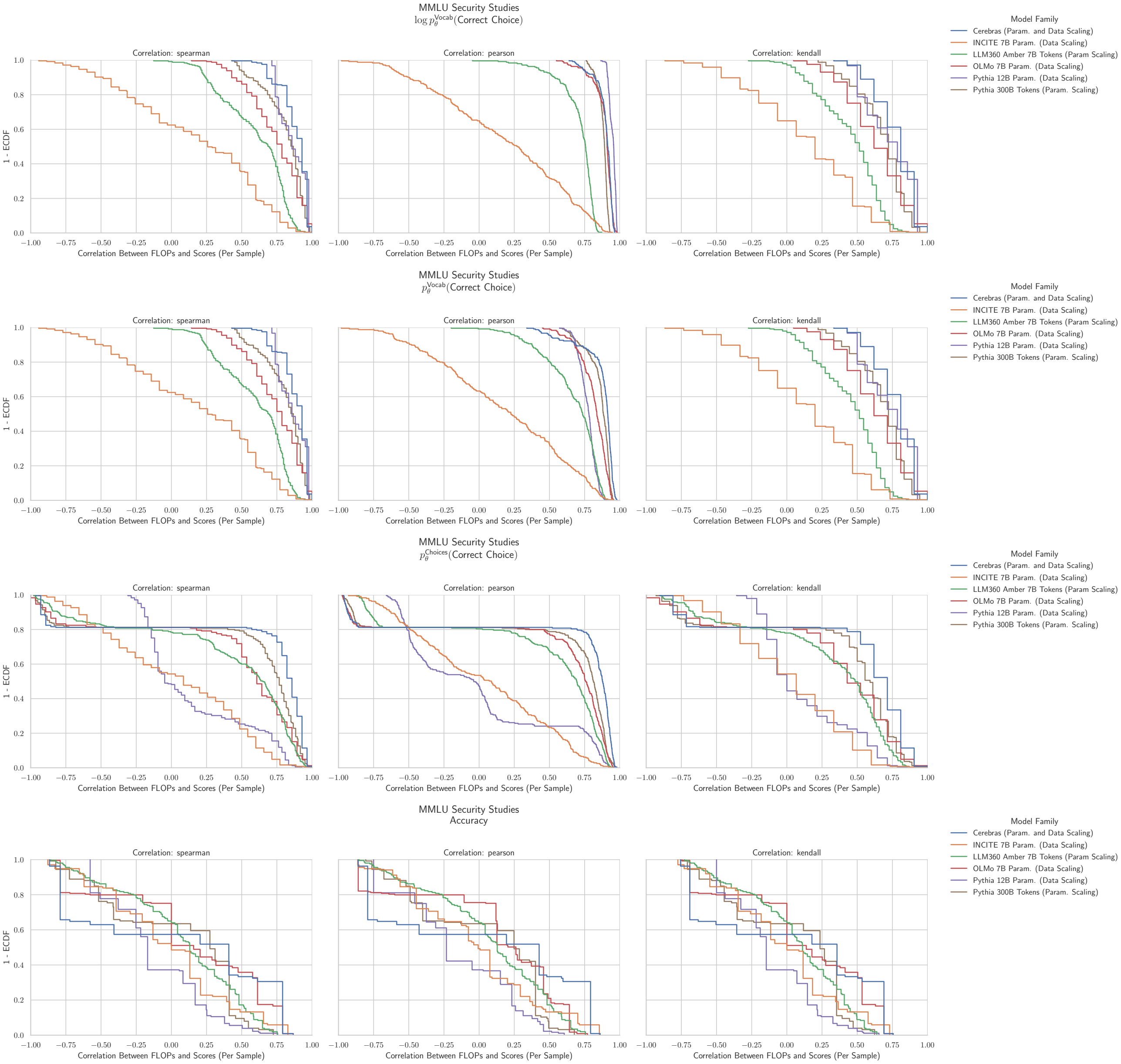

Figure 68. **MMLU Security Studies: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.60. NLP Benchmark: MMLU Sociology (Hendrycks et al., 2020)

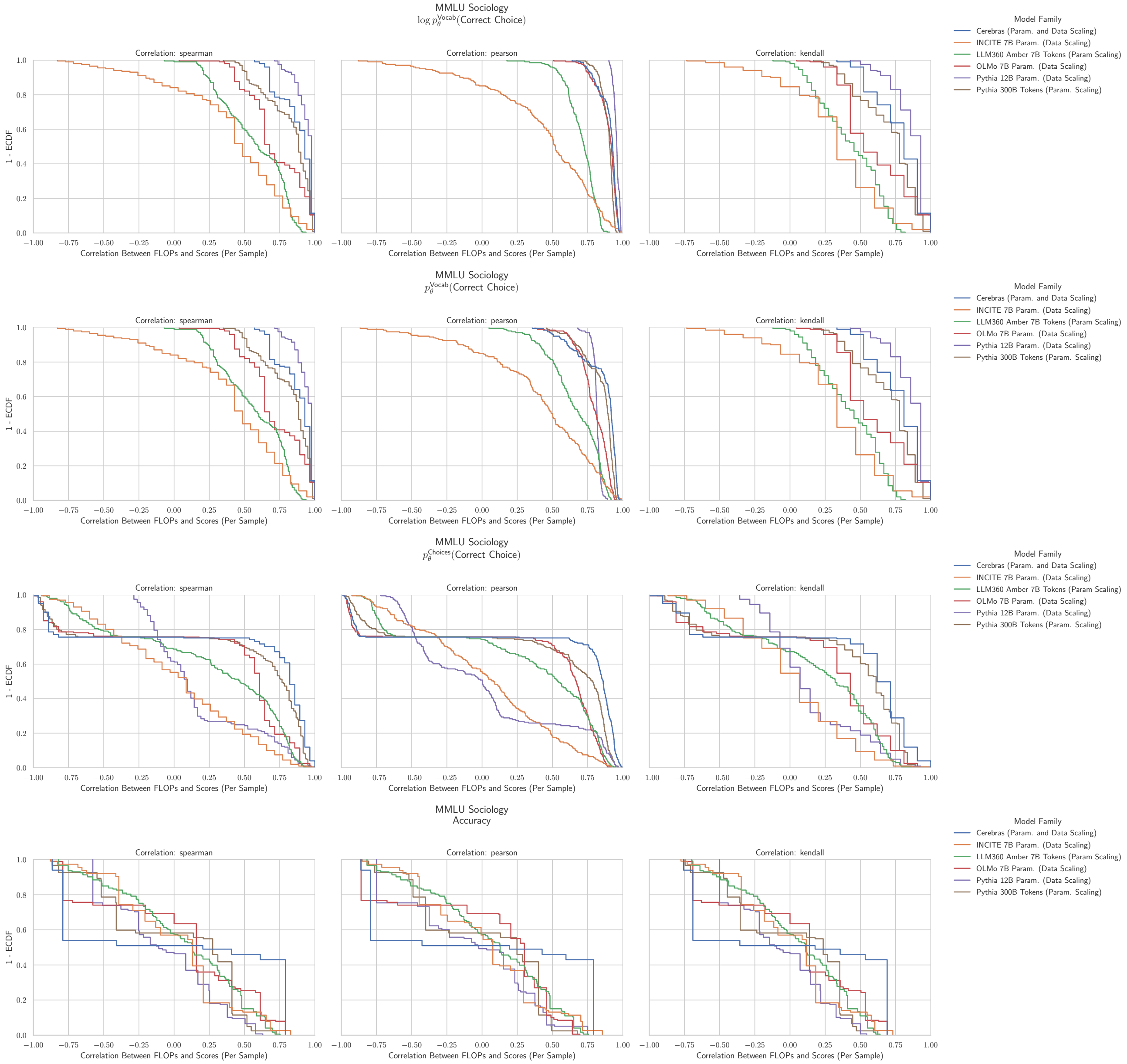

Figure 69. **MMLU Sociology: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.61. NLP Benchmark: MMLU US Foreign Policy (Hendrycks et al., 2020)

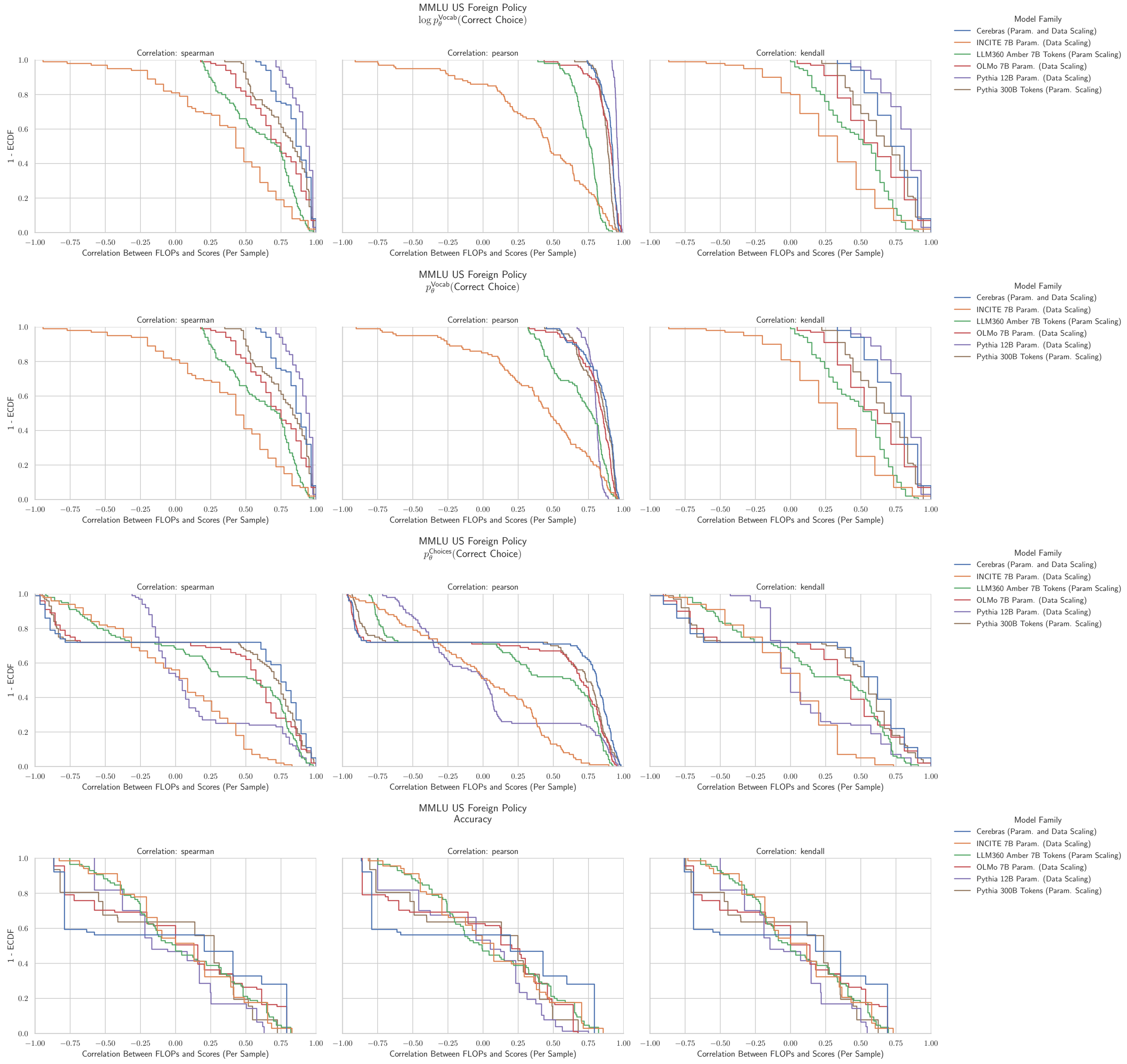

Figure 70. **MMLU US Foreign Policy: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.62. NLP Benchmark: MMLU Virology (Hendrycks et al., 2020)

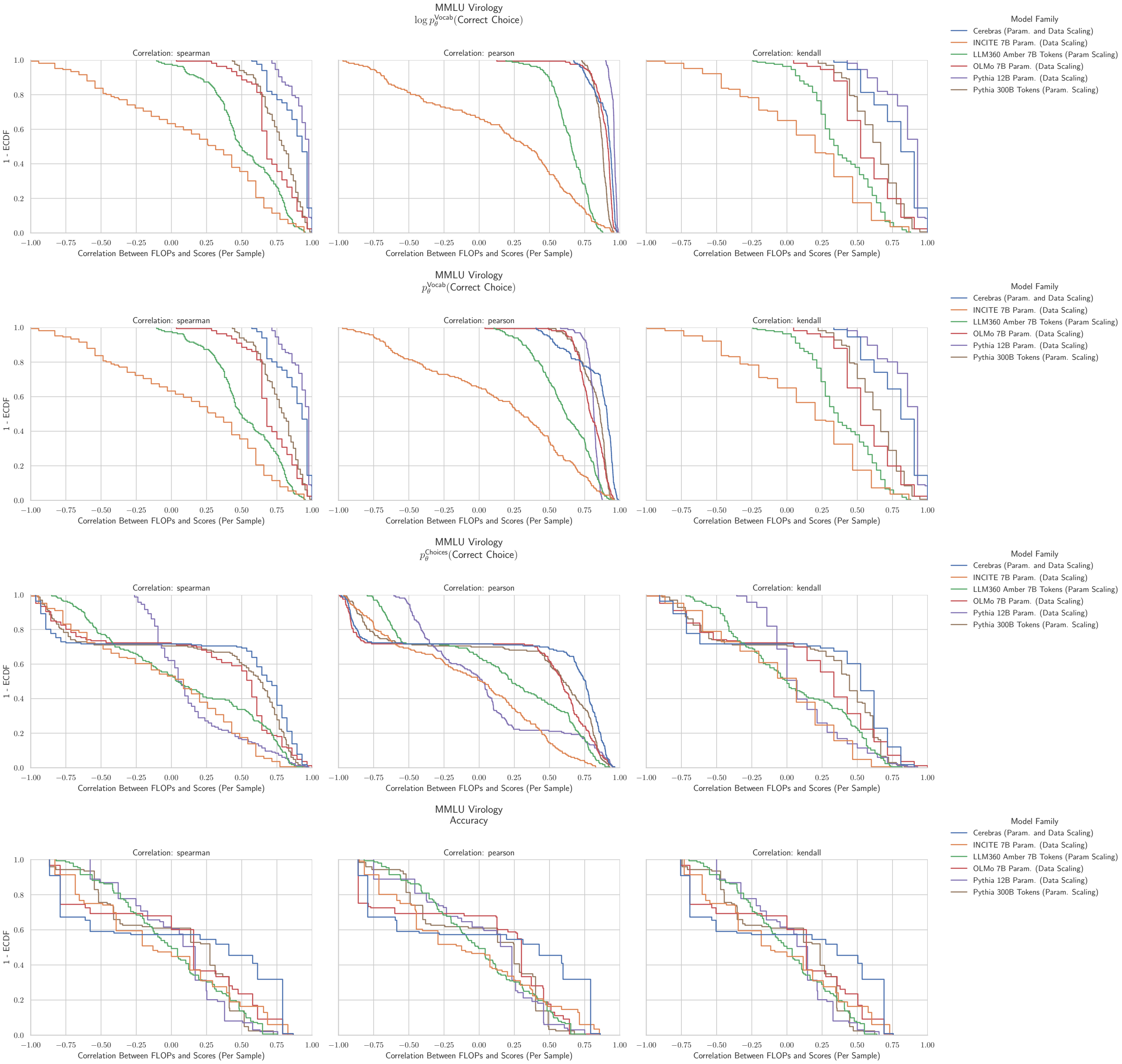

Figure 71. **MMLU Virology: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.63. NLP Benchmark: MMLU World Religions (Hendrycks et al., 2020)

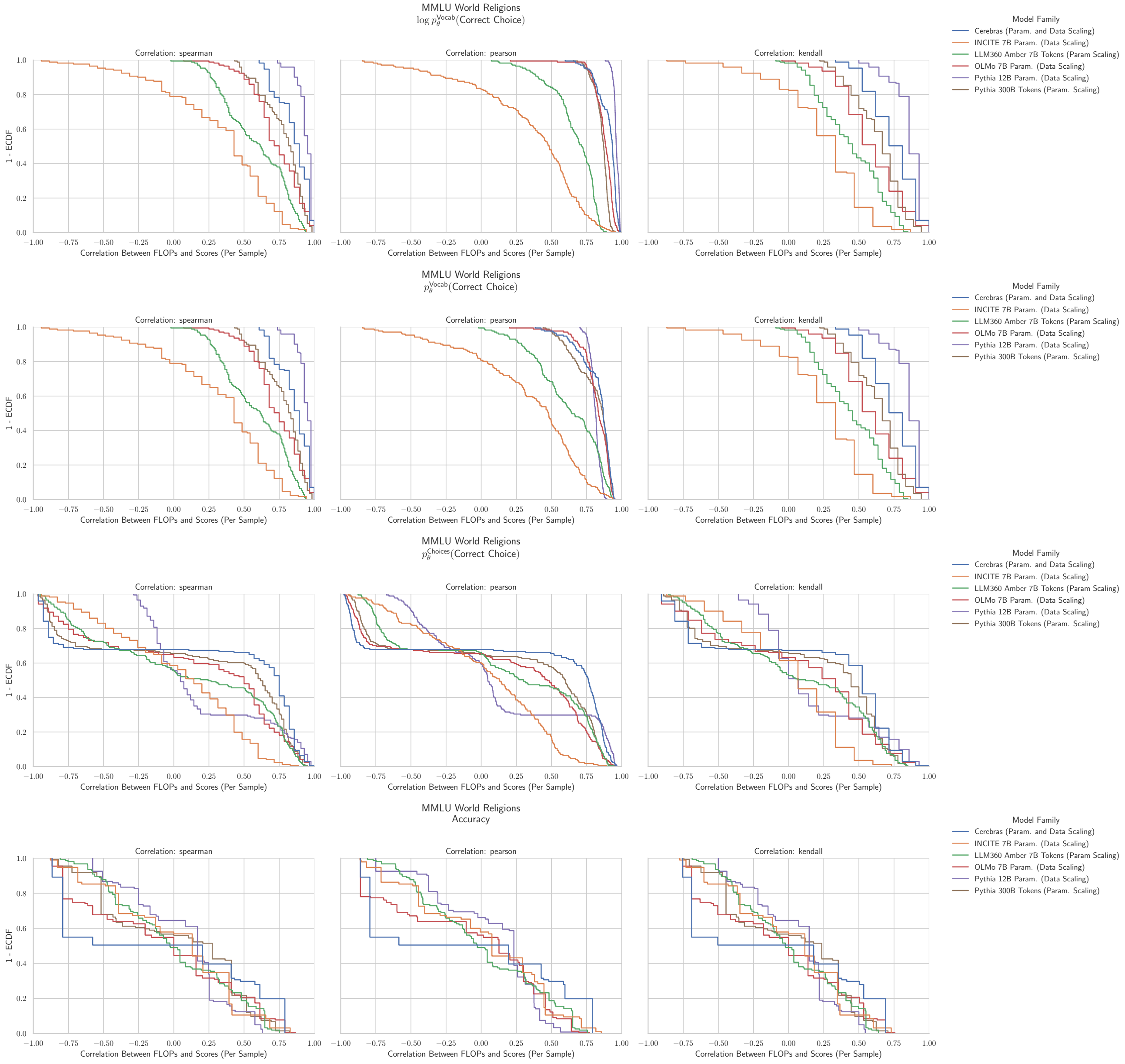

Figure 72. **MMLU World Religions: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.64. NLP Benchmark: OpenBookQA (Mihaylov et al., 2018)

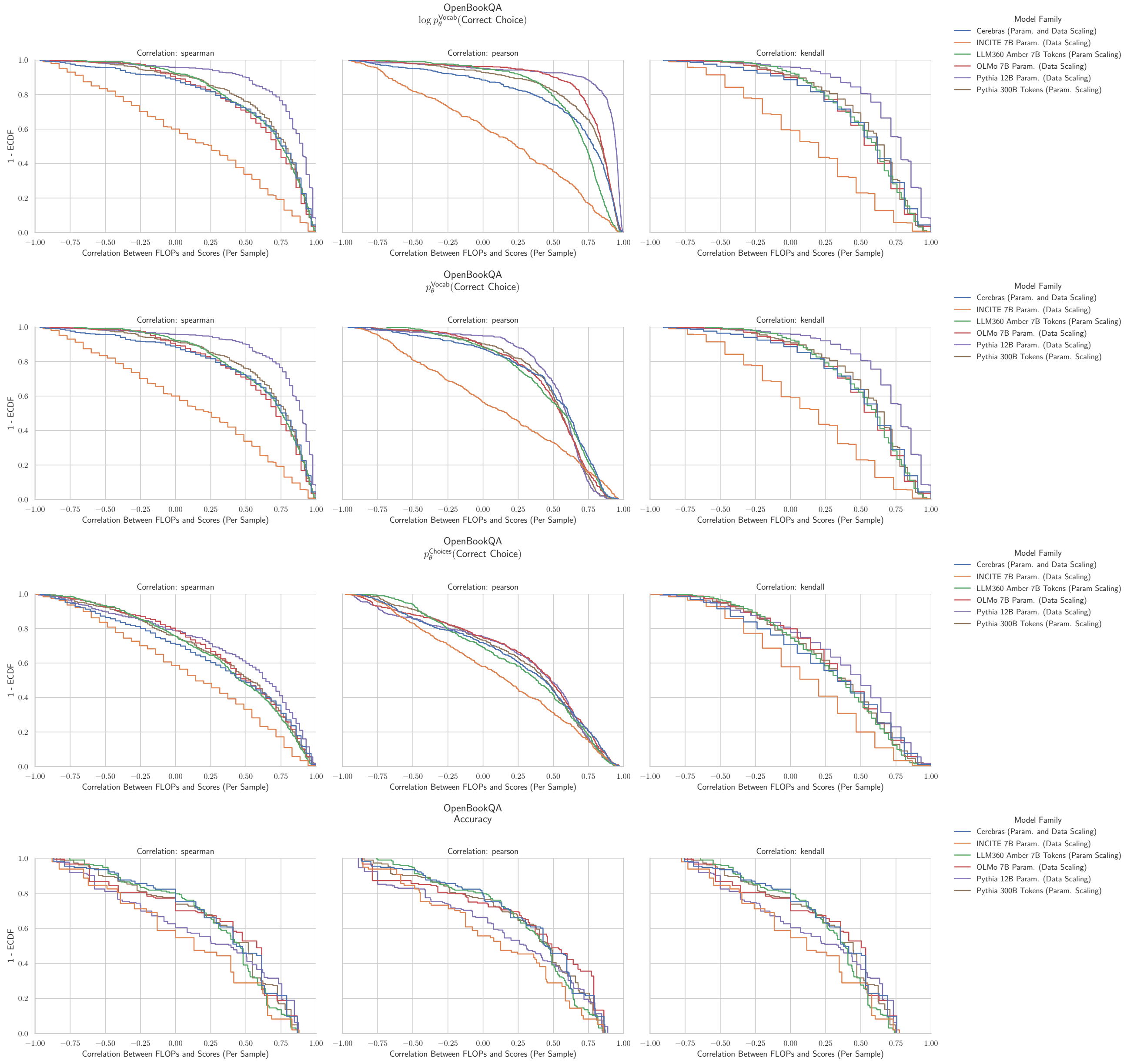

Figure 73. **OpenBookQA: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.65. NLP Benchmark: PIQA (Bisk et al., 2020)

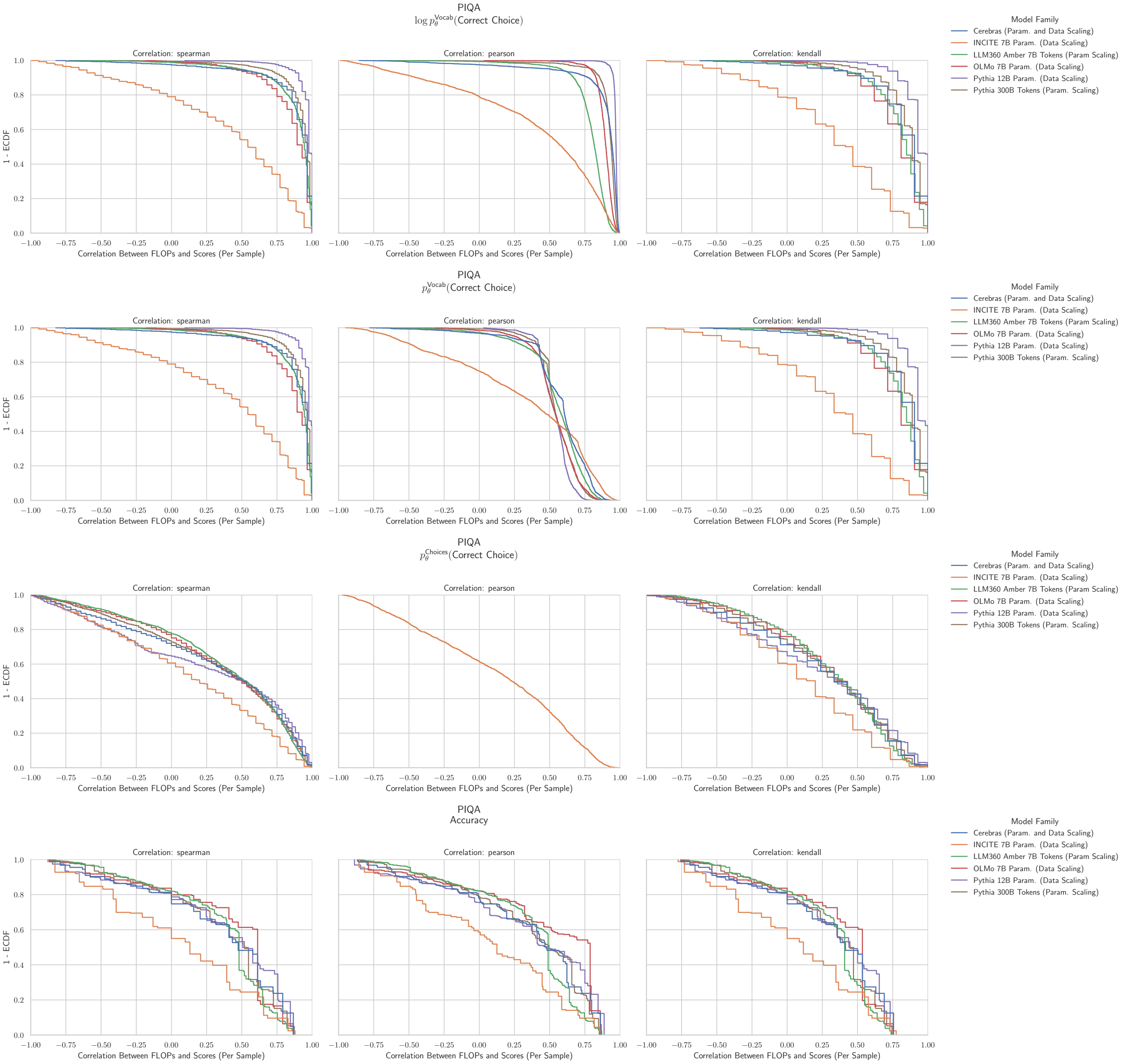

Figure 74. **PIQA: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.66. NLP Benchmark: RACE (Lai et al., 2017)

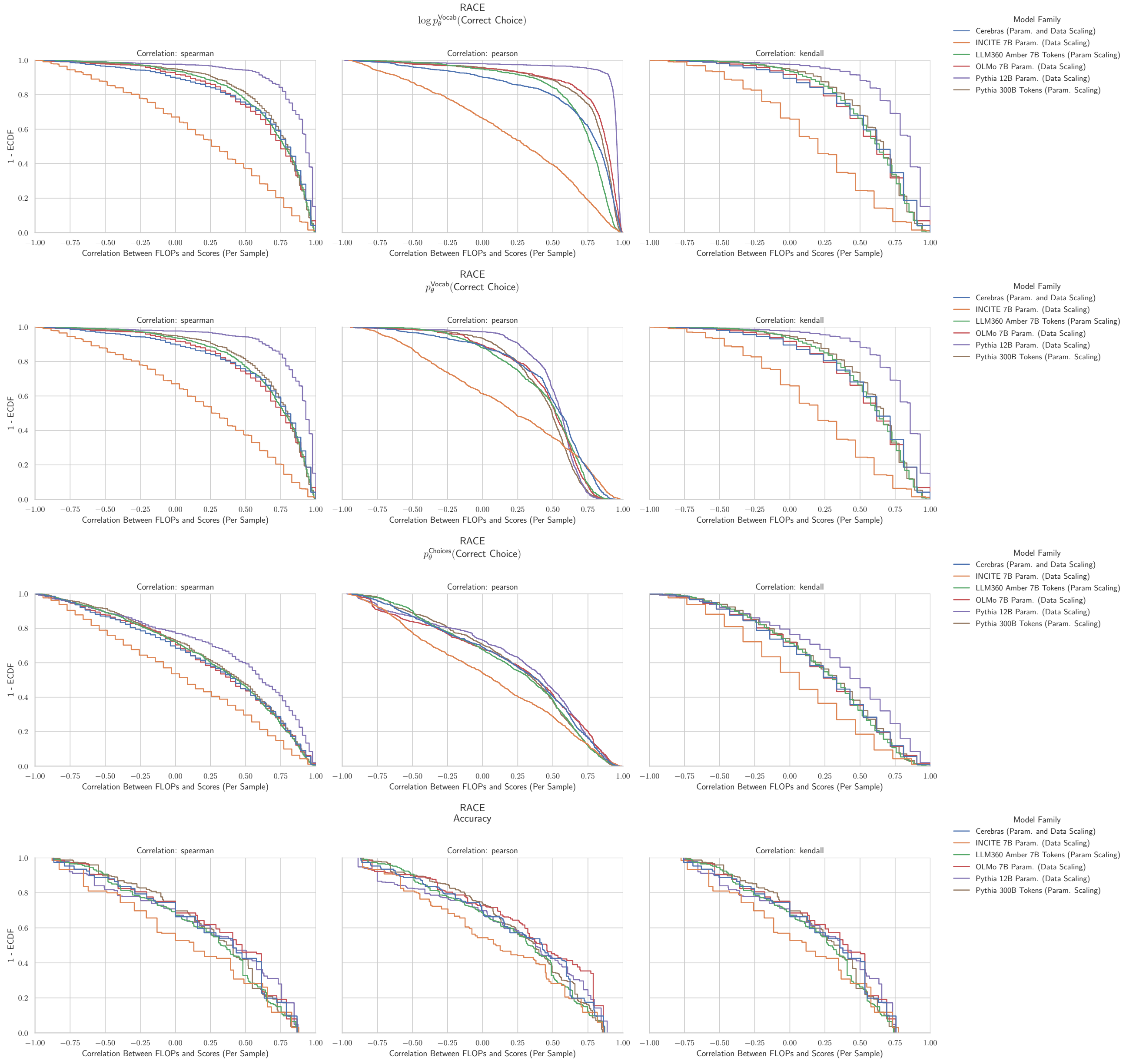

Figure 75. **RACE: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.67. NLP Benchmark: SciQ (Welbl et al., 2017)

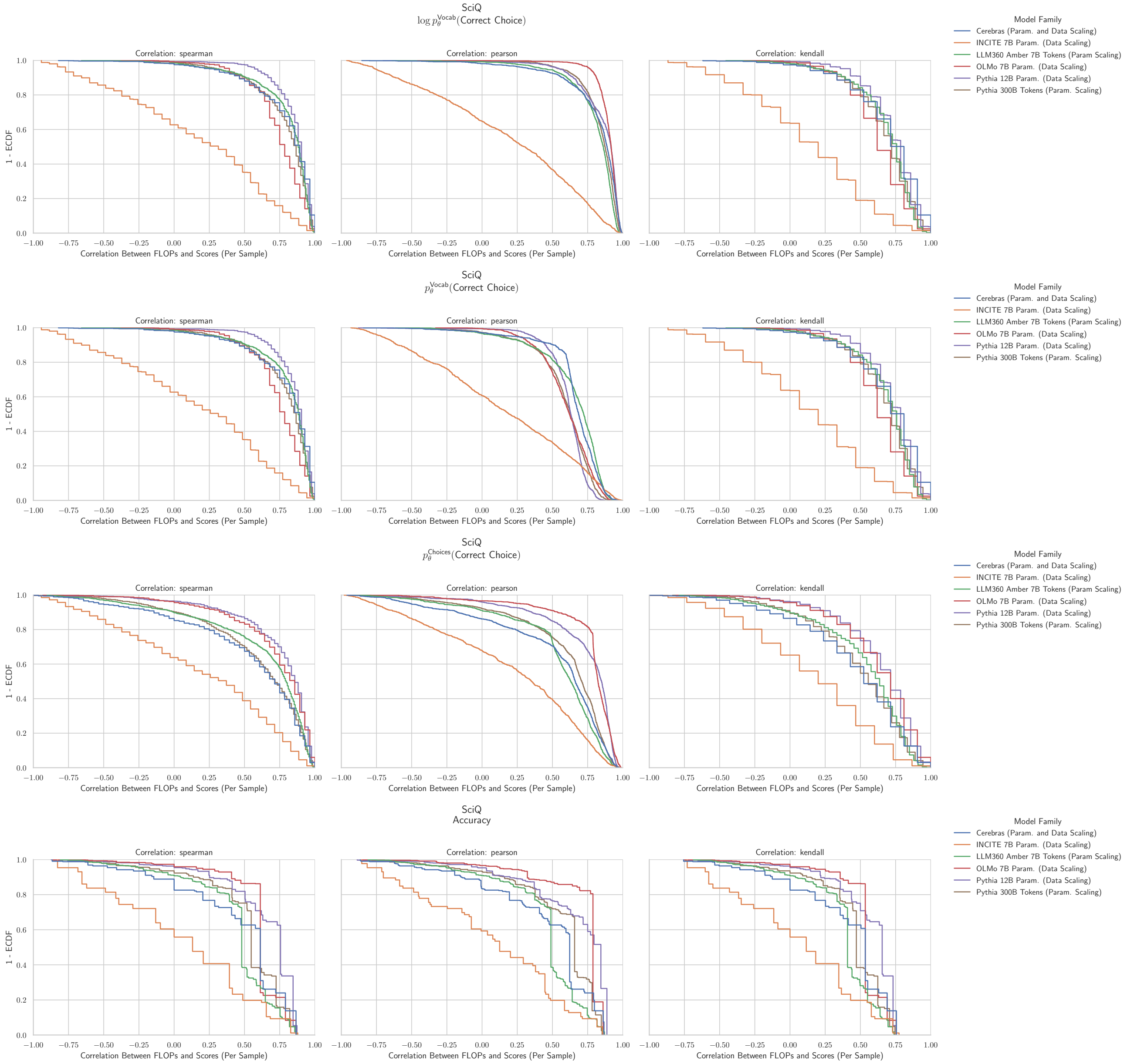

Figure 76. **SciQ: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.68. NLP Benchmark: Social IQA (Sap et al., 2019b)

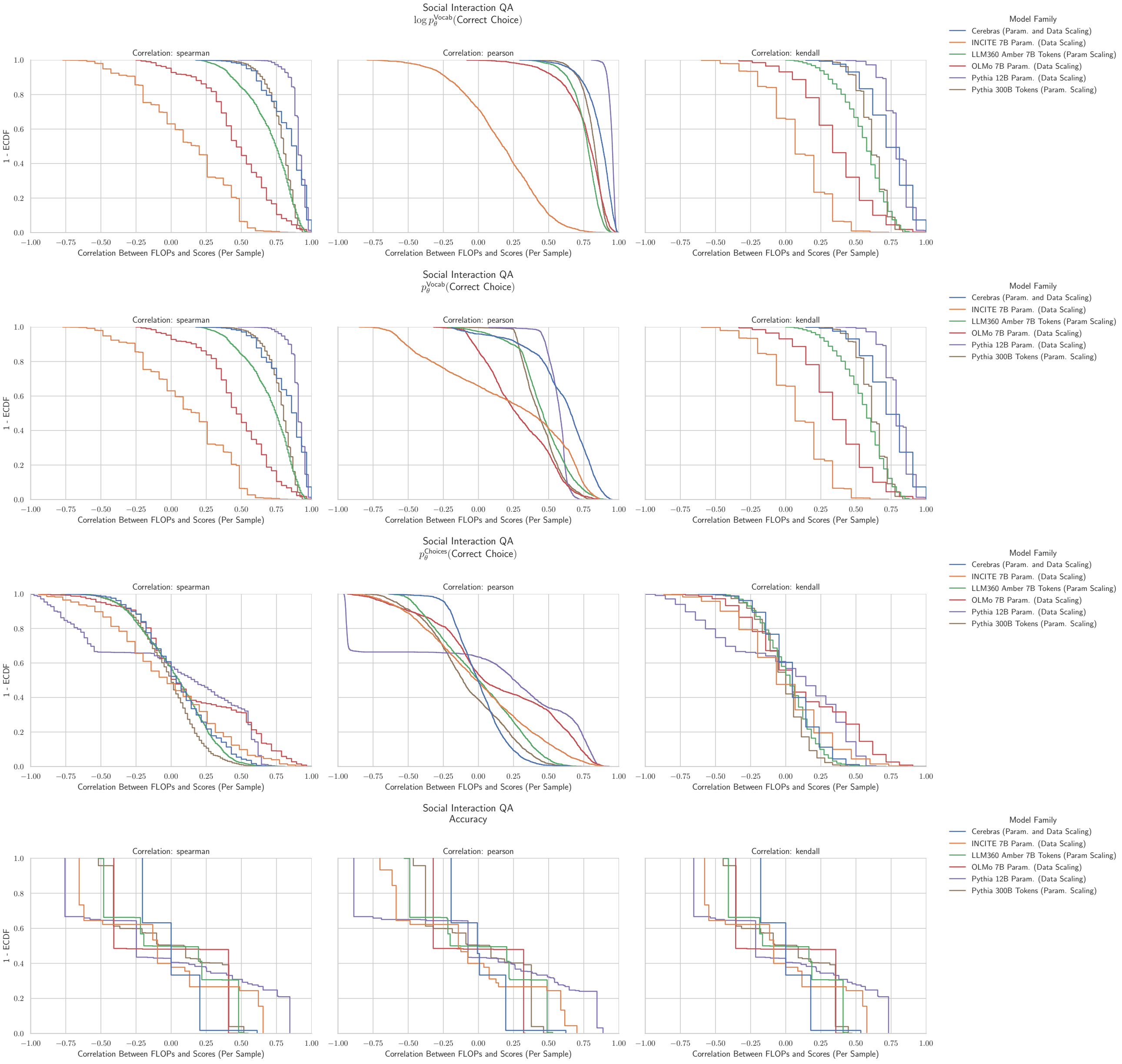

Figure 77. **Social IQA: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.69. NLP Benchmark: Winogrande (Keisuke et al., 2019)

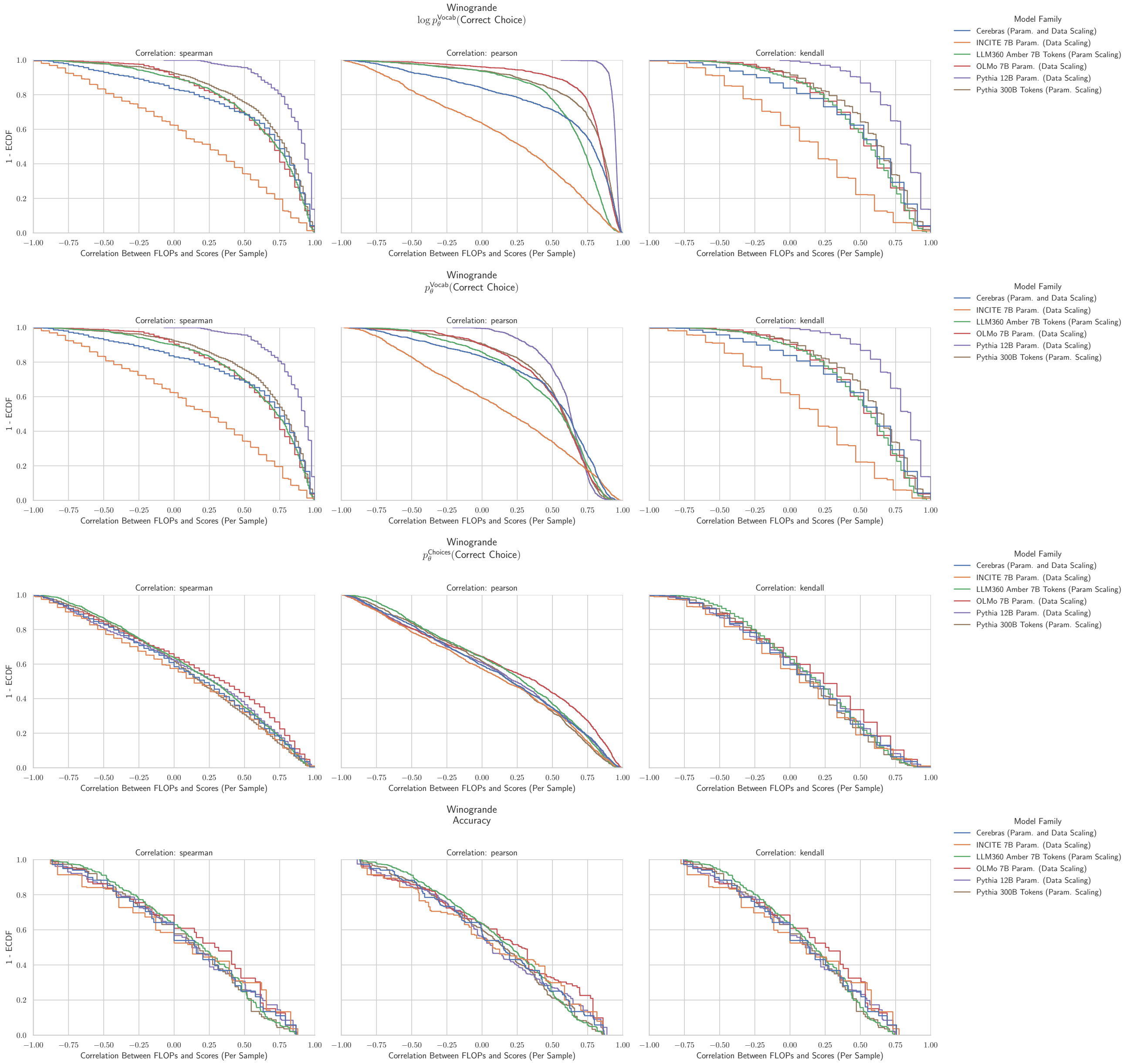

Figure 78. **Social IQA: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**

I.70. NLP Benchmark: XWinograd English (Muennighoff et al., 2023)

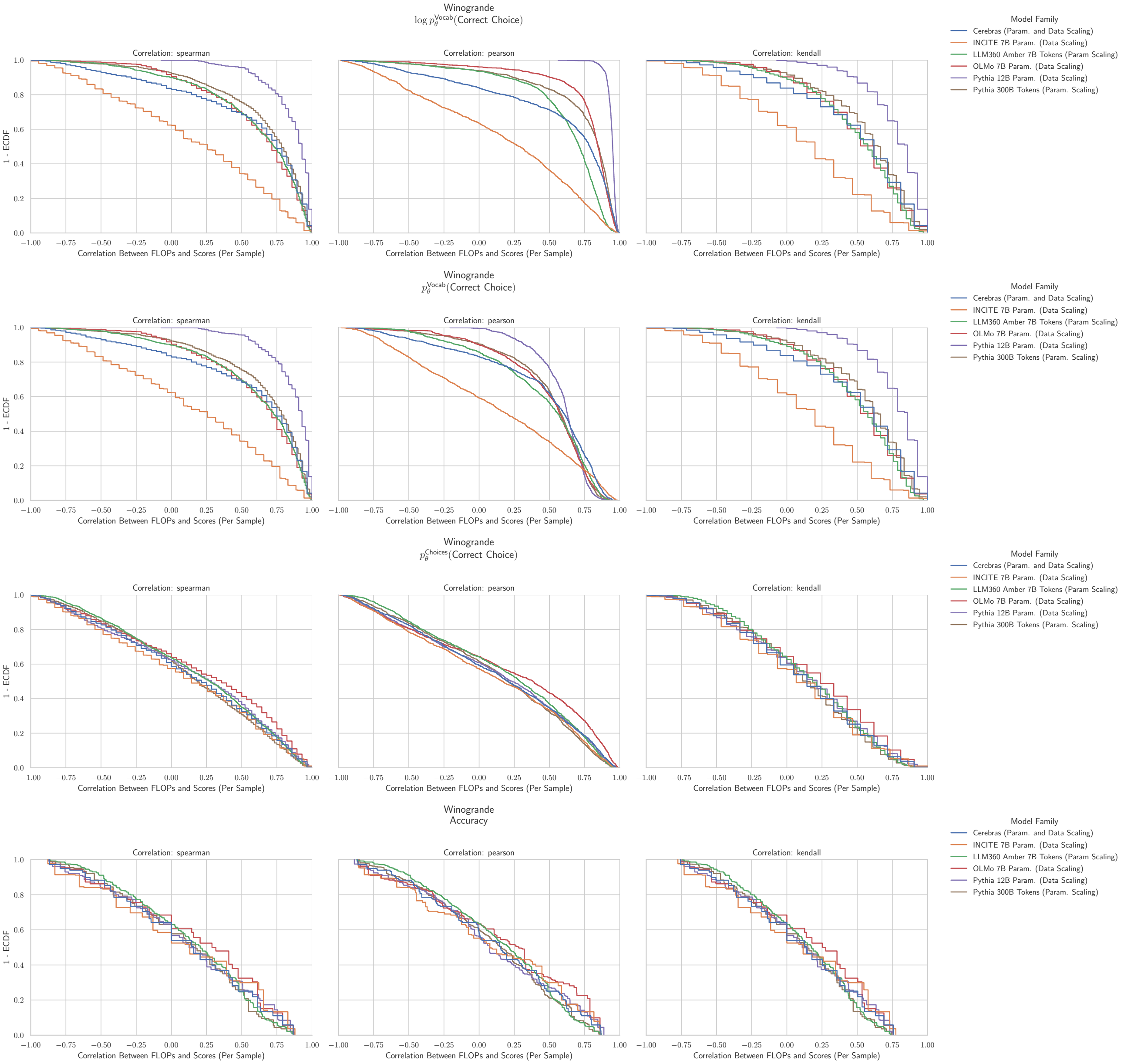

Figure 79. **XWinograd English: Downstream performance is computed via a sequence of transformations that deteriorate correlations between scores and pretraining compute.**