Distributions of Score-Compute Correlations by Metric
Benchmark: TriviaQA

A Metric: $\log p_\theta^{\text{Vocab}}(\text{Correct Choice})$

B Metric: $p_\theta^{\text{Choices}}(\text{Correct Choice})$

C Metric: Brier Score

D Metric: Accuracy

Model Family
- Cerebras (Param. and Data Scaling)
- INCITE 7B Param. (Data Scaling)
- LLM360 Amber 7B Tokens (Param Scaling)
- OLMo 7B Param. (Data Scaling)
- Pythia 12B Param. (Data Scaling)
- Pythia 300B Tokens (Param. Scaling)

1 - CDF

Correlation Between FLOPs and Scores (Per Sample)