Benchmark: MMLU HS US History
Performance Metric: $\log p_\theta^{\mathsf{Vocab}}(\text{Correct Choice})$
Correlation Metric: Spearman

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)

% of Samples

Correlation Between FLOPs and Scores (Per Sample)

1 - CDF

Correlation Between FLOPs and Scores (Per Sample)