Benchmark: Social Interaction QA
Performance Metric: Target Greedily Decoded
Correlation Metric: Pearson

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)

Left plot: Y-axis: % of Samples; X-axis: Correlation Between FLOPs and Scores (Per Sample)

Right plot: Y-axis: 1 - CDF; X-axis: Correlation Between FLOPs and Scores (Per Sample)