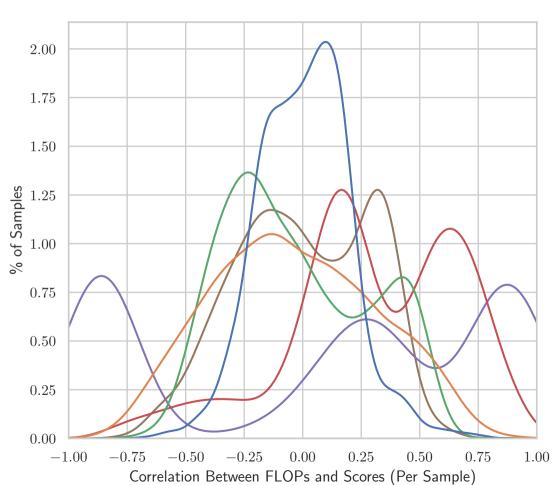
Benchmark: Social Interaction QA Performance Metric: Brier Score Correlation Metric: Pearson Model Family Cerebras (Param. and Data Scaling) —— INCITE 7B Param. (Data Scaling) 0.8 LLM360 Amber 7B Tokens (Param Scaling) — OLMo 7B Param. (Data Scaling) — Pythia 12B Param. (Data Scaling) —— Pythia 300B Tokens (Param. Scaling) 0.6 CDF 0.40.2

0.25

0.50

0.75

1.00



0.0

-1.00

-0.75

-0.50

-0.25

0.00

Correlation Between FLOPs and Scores (Per Sample)