MC-TACO Accuracy — Cerebras (Param. and Data Scaling) — INCITE 7B Param. (Data Scaling) Correlation: pearson Correlation: kendall LLM360 Amber 7B Tokens (Param Scaling) — OLMo 7B Param. (Data Scaling) — Pythia 12B Param. (Data Scaling) —— Pythia 300B Tokens (Param. Scaling)

Correlation Between FLOPs and Scores (Per Sample)

1.00

Correlation Between FLOPs and Scores (Per Sample)

-1.00

Correlation: spearman

Correlation Between FLOPs and Scores (Per Sample)

1.00 -1.00

2.0

0.5

Model Family

1.00