MC-TACO Brier Score Correlation: pearson Correlation: kendall 1.00 1.00 -1.00-0.50

Correlation Between FLOPs and Scores (Per Sample)

Correlation: spearman

Correlation Between FLOPs and Scores (Per Sample)

1.00 -1.00

Correlation Between FLOPs and Scores (Per Sample)

1.0

0.6

0.2

Model Family

LLM360 Amber 7B Tokens (Param Scaling)

Cerebras (Param. and Data Scaling)INCITE 7B Param. (Data Scaling)

OLMo 7B Param. (Data Scaling)
Pythia 12B Param. (Data Scaling)
Pythia 300B Tokens (Param. Scaling)