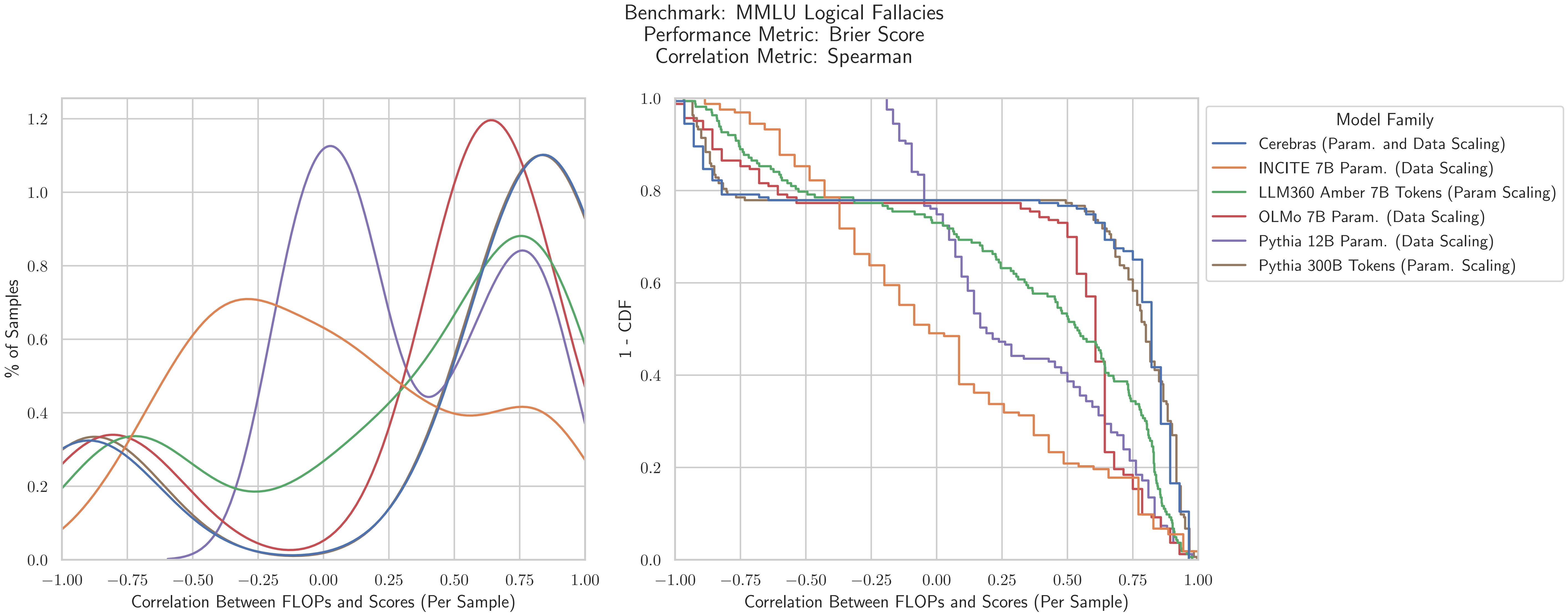
Benchmark: MMLU Logical Fallacies
Performance Metric: Brier Score
Correlation Metric: Spearman
% of Samples
0.0
0.2
0.4
0.6
0.8
1.0
1.2
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation Between FLOPs and Scores (Per Sample)
1 - CDF
0.0
0.2
0.4
0.6
0.8
1.0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Correlation Between FLOPs and Scores (Per Sample)
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)