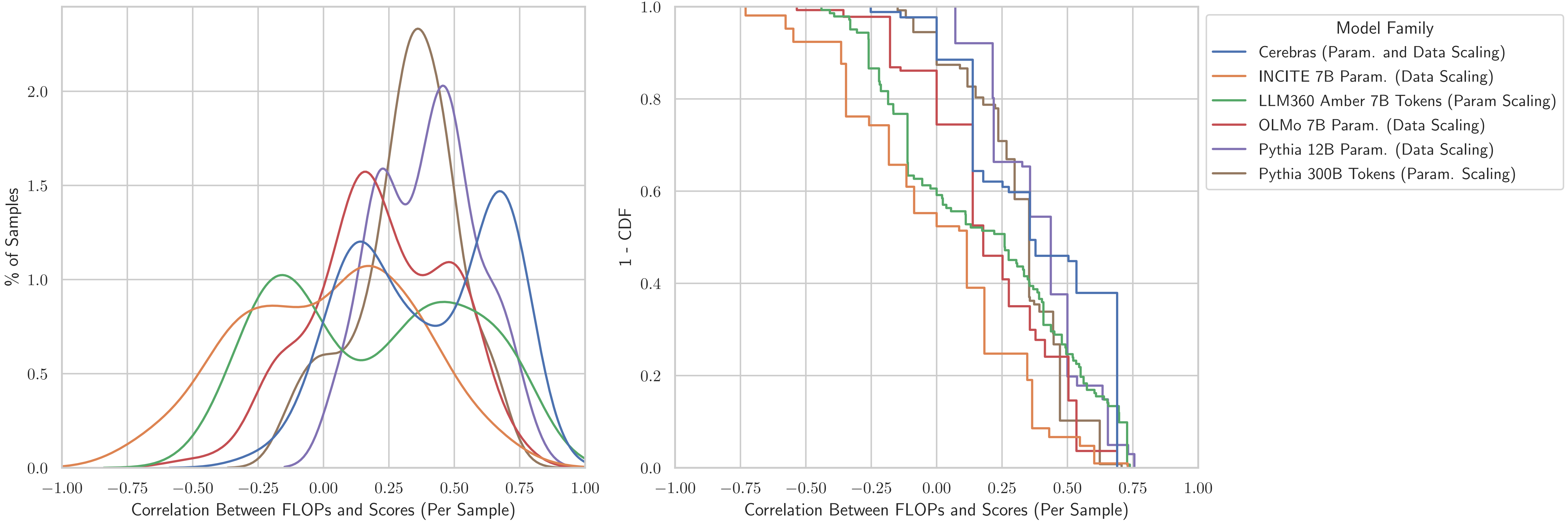
Benchmark: MMLU Logical Fallacies
Performance Metric: Target Greedily Decoded
Correlation Metric: Kendall


% of Samples
Correlation Between FLOPs and Scores (Per Sample)
1 - CDF
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)