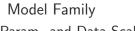
HellaSwag  $p_{\theta}^{\mathsf{Choices}}(\mathsf{Correct\ Choice})$ — Cerebras (Param. and Data Scaling) Correlation: spearman Correlation: pearson Correlation: kendall LLM360 Amber 7B Tokens (Param Scaling) 3.5— Pythia 12B Param. (Data Scaling) 3.0 2.51.0 0.51.00 0.751.00 1.00 -1.00-1.00Correlation Between FLOPs and Scores (Per Sample) Correlation Between FLOPs and Scores (Per Sample) Correlation Between FLOPs and Scores (Per Sample)



— INCITE 7B Param. (Data Scaling)

— OLMo 7B Param. (Data Scaling)

— Pythia 300B Tokens (Param. Scaling)