${\sf OpenBookQA}$  $p_{\theta}^{\mathsf{Vocab}}(\mathsf{Correct\ Choice})$ Correlation: spearman Correlation: pearson Correlation: kendall — OLMo 7B Param. (Data Scaling) 2.0 0.50.750.751.00 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)

LLM360 Amber 7B Tokens (Param Scaling)

— Pythia 12B Param. (Data Scaling)

— Pythia 300B Tokens (Param. Scaling)