

Metric  $\log p_{\theta}^{\mathsf{Vocab}}(\mathsf{Correct\ Choice})$   $p_{\theta}^{\mathsf{Vocab}}(\mathsf{Correct\ Choice})$   $p_{\theta}^{\mathsf{Choices}}(\mathsf{Correct\ Choice})$ Brier Score  $\mathsf{Accuracy}$   $\mathsf{Model\ Family}$   $\mathsf{Cerebras\ (Param.\ and\ Data\ Scaling)}$   $\mathsf{INCITE\ 7B\ Param.\ (Data\ Scaling)}$   $\mathsf{LLM360\ Amber\ 7B\ Tokens\ (Param\ Scaling)}$ 

OLMo 7B Param. (Data Scaling)

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