

Benchmark: Social Interaction QA  
Performance Metric:  $\log p_{\theta}^{\text{Vocab}}(\text{Correct Choice})$   
Correlation Metric: Pearson

