

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

