Benchmark: Social Interaction QA
Performance Metric: Target Greedily Decoded
Correlation Metric: Spearman