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