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