Benchmark: HellaSwag

Performance Metric: $p_\theta^{\text{Vocab}}(\text{Correct Choice})$

Correlation Metric: Pearson