Benchmark: HellaSwag
Performance Metric: Target Greedily Decoded
Correlation Metric: Kendall