

Benchmark: MMLU Logical Fallacies
Performance Metric: $p_{\theta}^{\text{Vocab}}$ (Correct Choice)
Correlation Metric: Kendall

