Benchmark: MMLU Logical Fallacies
Performance Metric: Brier Score
Correlation Metric: Kendall