Benchmark: MMLU Human Sexuality
Performance Metric: Target Greedily Decoded
Correlation Metric: Kendall