Benchmark: MMLU Prehistory
Performance Metric: Target Greedily Decoded
Correlation Metric: Spearman