

Benchmark: MMLUHS US History
Performance Metric: Target Greedily Decoded
Correlation Metric: Spearman

