Benchmark: MMLU HS US History
Performance Metric: Target Greedily Decoded
Correlation Metric: Pearson