

Evaluation Pipeline

lm-evaluation-harness

Gao et al., 2023

CoQA Benchmark

Conversational QA | Zero-shot

Metrics

F1 Score

EM Score

Memory

Latency