■ Experiment Results 07.07.2025

Generated by compressa-perf library

Parameter	Value
MODEL	Compressa-LLM
ENGINE	sglang
MAX_MODEL_LENGTH	27000
TTFT	0.059
TTFT_95	0.061
TOP_5_TTFT	0.062
LATENCY	0.462
LATENCY_95	0.52
TOP_5_LATENCY	0.521
TPOT	0.026
THROUGHPUT	100.805
THROUGHPUT_INPUT_TOKENS	62.733
THROUGHPUT_OUTPUT_TOKENS	38.072
RPS	2.163
LONGER_THAN_60_LATENCY	0
LONGER_THAN_120_LATENCY	0
LONGER_THAN_180_LATENCY	0
FAILED_REQUESTS	0
FAILED_REQUESTS_PER_HOUR	0.0