■ Experiment Results 11.07.2025

Generated by compressa-perf library

Model Parameter	Value
MODEL	Compressa-LLM
ENGINE	sglang
MAX_MODEL_LENGTH	27000
DRIVER VERSION	535.129.03
CUDA VERSION	12.4
HARDWARE	A100-PCIE-40GB
OPENAI_URL	http://localhost:5000/v1/

Experiment Parameter	Value
NUM_WORKERS	2.0
NUM_TASKS	100.0
MAX_TOKENS	1000.0
AVG_N_INPUT	29.0
STD_N_INPUT	0.0
AVG_N_OUTPUT	18.52
STD_N_OUTPUT	4.92

Metric	Value
TTFT	0.078
TTFT_95	0.091
TOP_5_TTFT	0.107
LATENCY	0.524
LATENCY_95	0.603
TOP_5_LATENCY	0.77
TPOT	0.028
THROUGHPUT	180.544
THROUGHPUT_INPUT_TOKENS	110.192
THROUGHPUT_OUTPUT_TOKENS	70.352
RPS	3.8
LONGER_THAN_60_LATENCY	0.0
LONGER_THAN_120_LATENCY	0.0
LONGER_THAN_180_LATENCY	0.0
FAILED_REQUESTS	0.0
FAILED_REQUESTS_PER_HOUR	0.0