■ Experiment Results 12.07.2025

Generated by compressa-perf library

Model Parameter	Value
MODEL	Compressa-LLM
ENGINE	vllm
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	20
NUM_TASKS	1000
MAX_TOKENS	1000
	0

Metric	Value
	0