



■ Experiment Results 09.07.2025

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

Experiment Parameter	Value
NUM_WORKERS	20
NUM_TASKS	1000
OPENAI_URL	http://localhost:5000/v1/
MAX_TOKENS	1000
MODEL_NAME	Compressa-LLM
	0

Metric	Value
	0