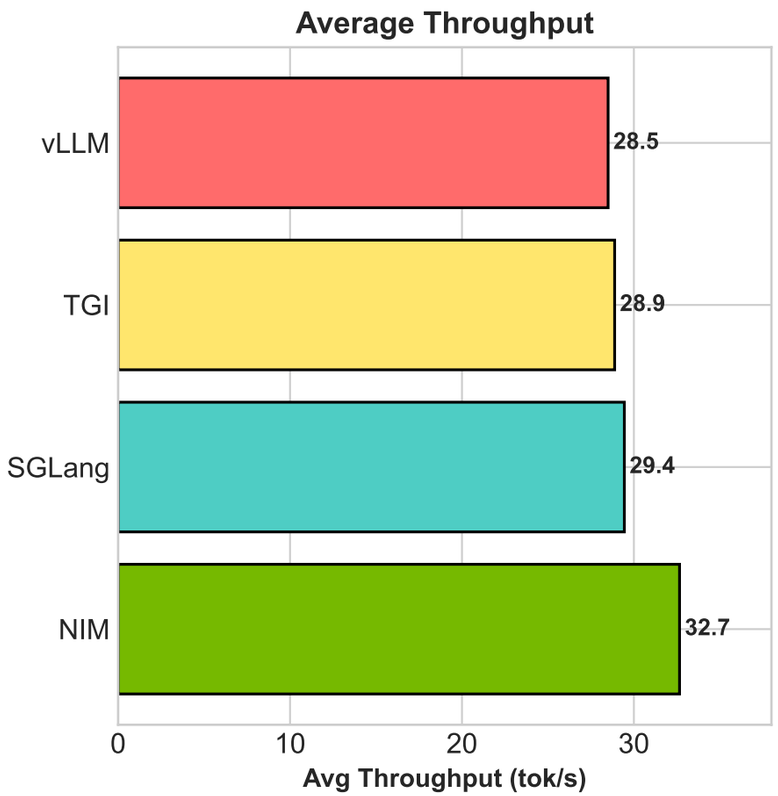
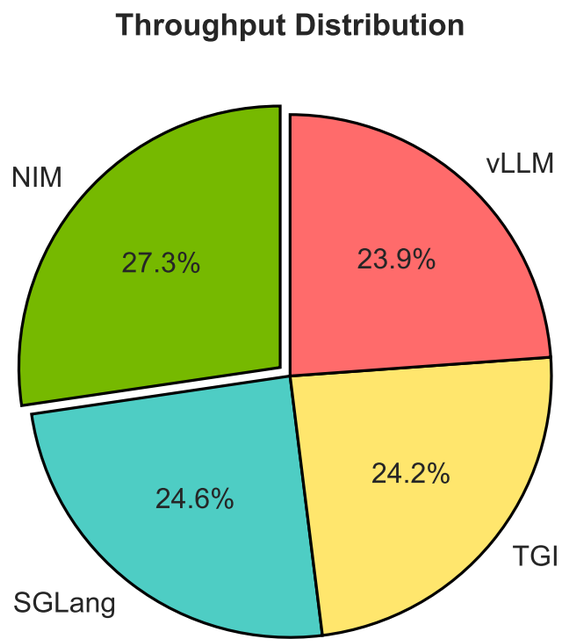
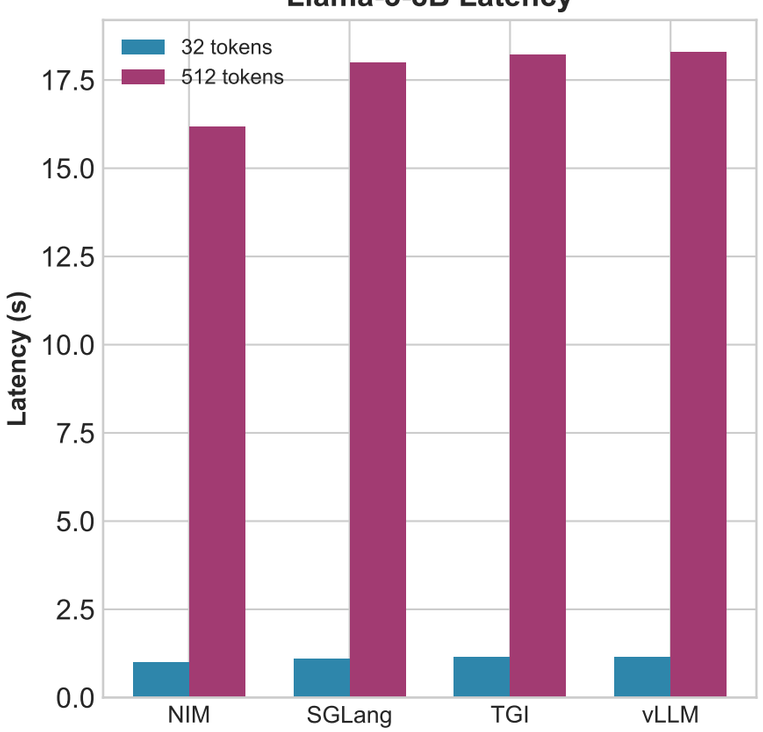
LLM Inference Benchmark Executive Summary
NVIDIA A10 GPU Performance Analysis

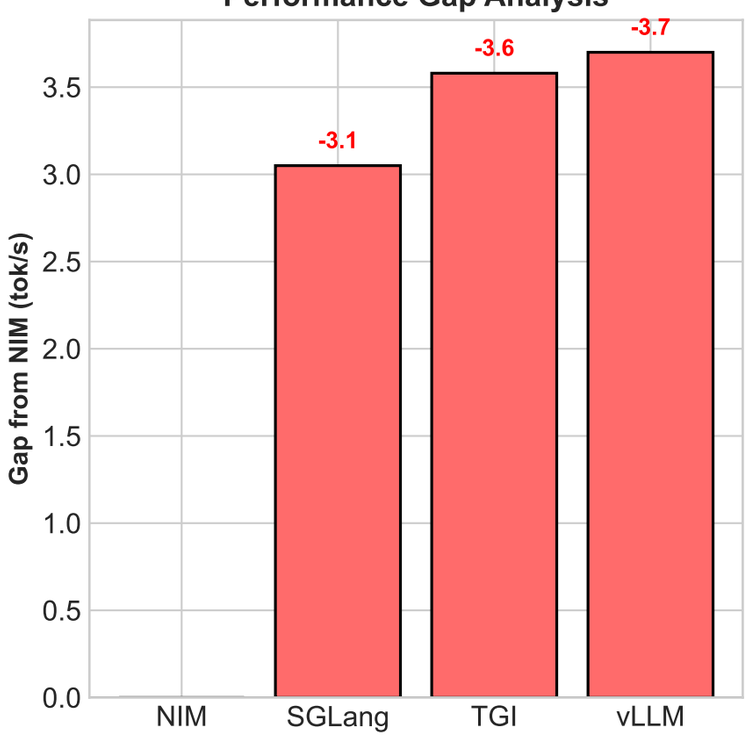




Throughput (tokens/second)

Framework	Llama-3-8B	Mistral-7B	Avg
NIM	**31.7**	**33.6**	**32.7**
SGLang	28.7	30.2	29.5
TGI	28.1	29.6	28.9
vLLM	28.0	29.0	28.5






```
KEY FINDINGS
═══════════════════════════════════════

[1] NVIDIA NIM leads with 10-18%
    higher throughput than alternatives

[2] SGLang achieves ~90% of NIM
    performance with open-source stack

[3] TGI & vLLM offer comparable
    performance (~88% of NIM)

[4] All frameworks show linear
    latency scaling with token count

[5] Mistral-7B shows ~6% higher
    throughput than Llama-3-8B

═══════════════════════════════════════
GPU: NVIDIA A10 (24GB) | FP16
```