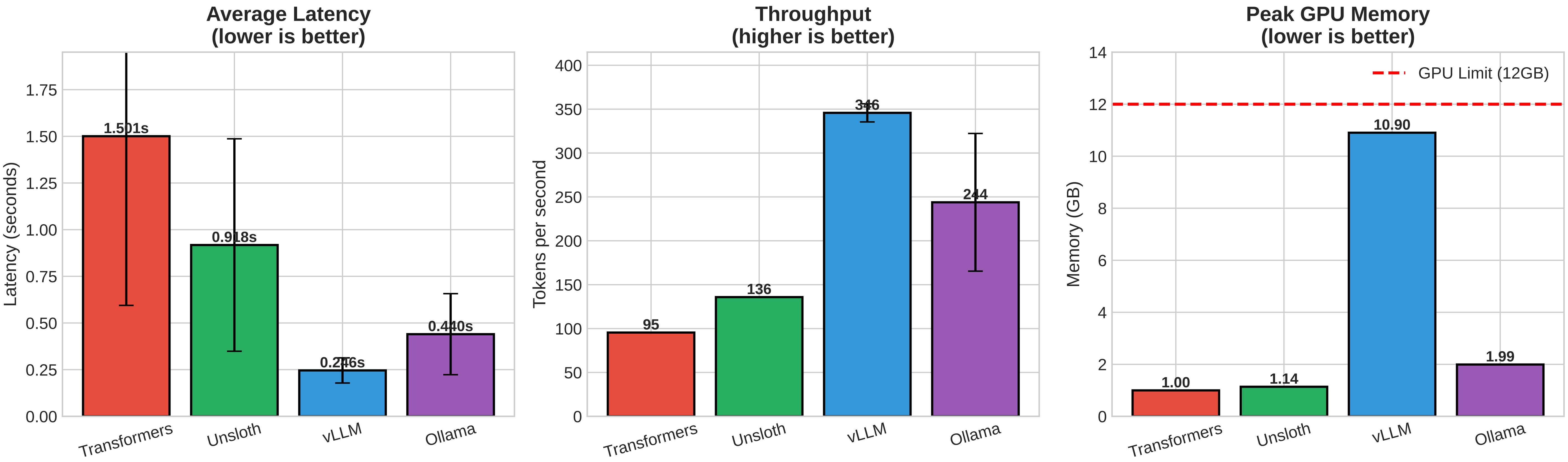
# Inference Benchmark: Framework Comparison


Average Latency
(lower is better)
Latency (seconds)
1.501s
0.918s
0.246s
0.440s
Transformers
Unsloth
vLLM
Ollama
Throughput
(higher is better)
Tokens per second
Transformers
Unsloth
vLLM
Ollama
Peak GPU Memory
(lower is better)
Memory (GB)
GPU Limit (12GB)
Transformers
Unsloth
vLLM
Ollama