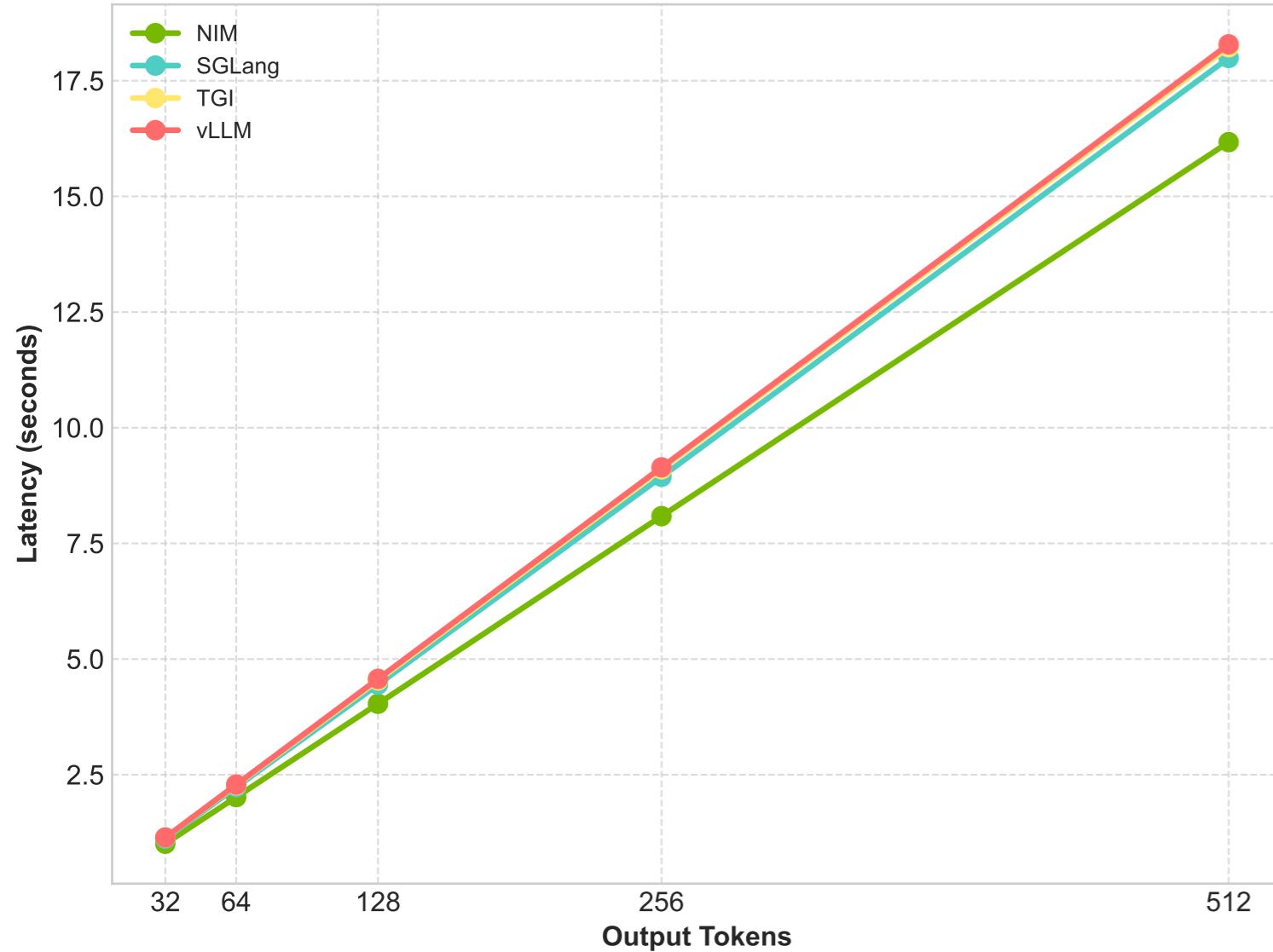


Inference Latency vs Output Token Count

NVIDIA A10 GPU

Llama-3-8B Latency Scaling



Mistral-7B Latency Scaling

