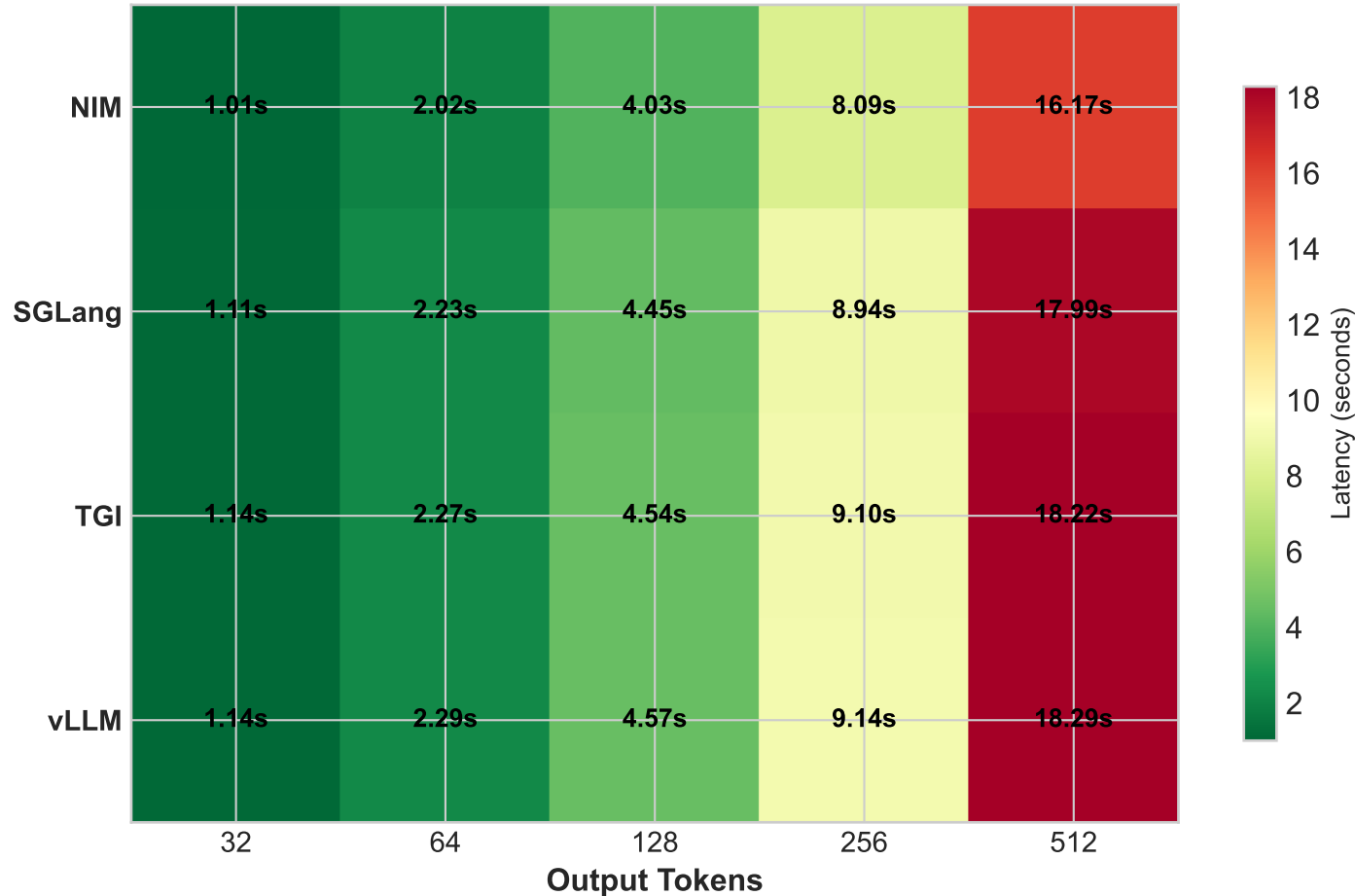


Inference Latency Heatmap (Lower is Better)

Llama-3-8B



Mistral-7B

