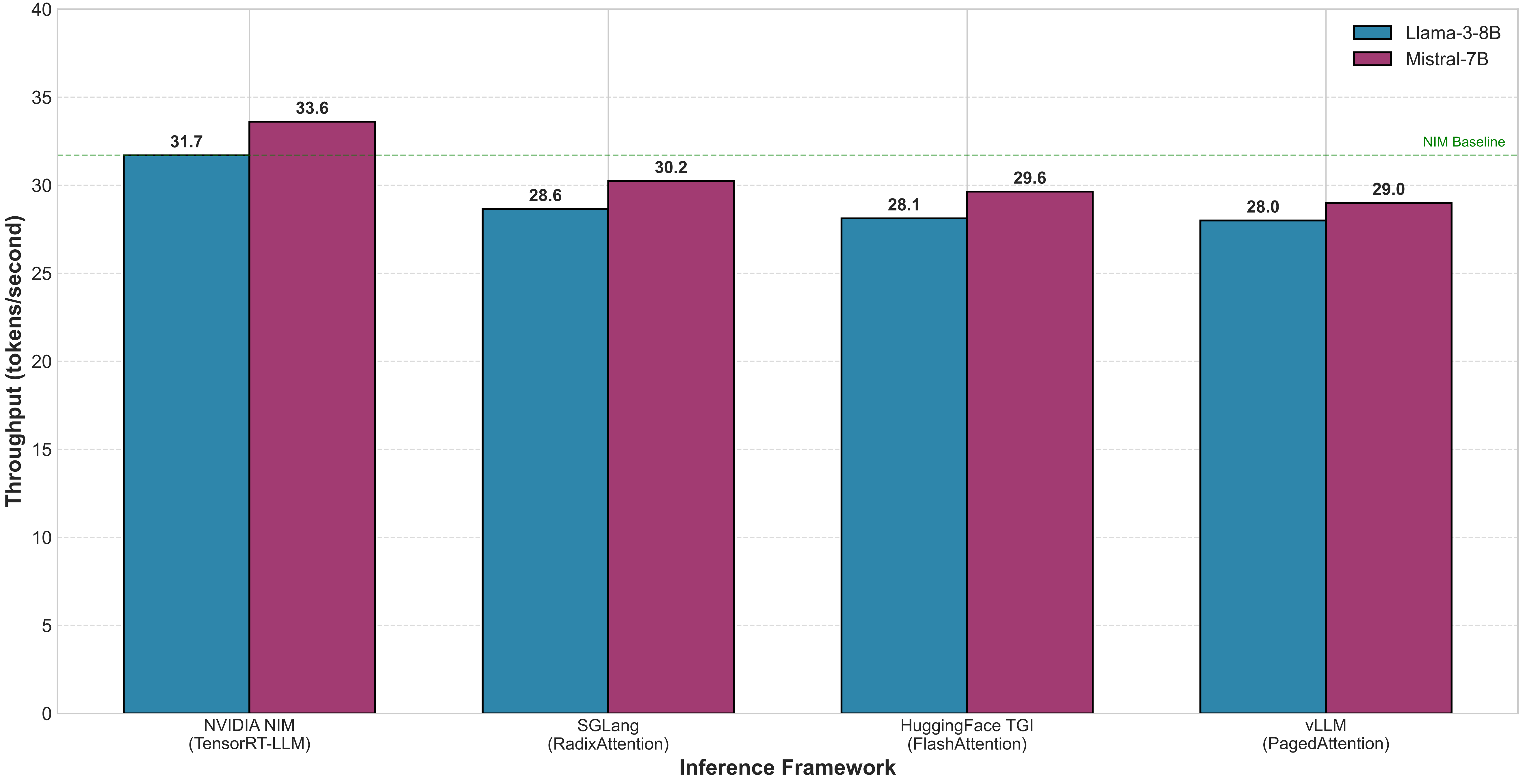
LLM Inference Throughput Comparison
NVIDIA A10 GPU (24GB)


Throughput (tokens/second)
Llama-3-8B
Mistral-7B
NIM Baseline
NVIDIA NIM
(TensorRT-LLM)
SGLang
(RadixAttention)
HuggingFace TGI
(FlashAttention)
vLLM
(PagedAttention)
Inference Framework