Llama 3.1 8B Throughput tok/s Throughput tok/s 9000 9000 0000 0000 AMD MI250X AMD MI300X Intel XPU **NVIDIA A100 NVIDIA GH200 NVIDIA H100**