LLM-Q: Hardware

Date: 2026-02-05
Author: Jan Quesker
Keywords: llm, quantization, hardware, benchmark, rtx 5090, rtx pro 6000, h100, gpu, inference, vllm, tensorrt-llm, gguf, awq, gptq, fp8, int8, int4, qwen, llama, mistral, deepseek, moe, mixtral, perplexity, kl divergence, throughput, latency, ttft, tpot, prefill, decode, batch size, kv cache, paged attention, continuous batching, speculative decoding, tensor parallelism, pipeline parallelism, multi-gpu, nvlink, pcie, blackwell, hopper, ada lowlevel, cuda, tensor core, memory bandwidth, vram, power consumption, tdp, tokens per watt, cost analysis, cloud gpu, runpod, benchmark methodology, reproducible, open source

Abstract: We benchmark 4 NVIDIA GPUs (RTX 5090, RTX PRO 6000 Blackwell, H100 PCIe, H100 SXM) across 3 quantization formats (GPTQ, AWQ, GGUF) and 2 inference engines (vLLM, TensorRT-LLM) on 5 models (Qwen 3 8B, Qwen 3 32B, Qwen 3 Coder 30B-A3B, Llama 3.1 8B, Llama 3.1 70B). We measure throughput, latency, TTFT, TPOT, power consumption, and quality (perplexity, KL divergence) under controlled conditions. We publish all raw data, configs, and methodology for reproducibility. Total cost: $1,624 on rented cloud GPUs.

Table of Contents

- Abstract
- Table of Contents
- 1. Introduction
- 2. Related Work
- 3. Hardware
- 4. Software
- 5. Models
- 6. Quantization Quality
- 7. Methodology
- 8. Results
- 9. Discussion
- 10. Conclusion
- 11. Appendix

Table of Contents

- 1. Introduction
  - 1.1 Why Quantization Matters
  - 1.2 Why Hardware Matters
  - 1.3 Why Open Source Matters
  - 1.4 Scope
  - 1.5 Contributions
- 2. Related Work
- 3. Hardware
  - 3.1 NVIDIA RTX 5090
  - 3.2 NVIDIA RTX PRO 6000 Blackwell
  - 3.3 NVIDIA H100 PCIe
  - 3.4 NVIDIA H100 SXM
  - 3.5 Cross-GPU Architecture Differences
  - 3.6 Cloud Pricing
- 4. Software
  - 4.1 Inference Engines
  - 4.2 Quantization Tools
  - 4.3 Why Pre-Quantized Models Matter
  - 4.4 Why Not Ollama
- 5. Models
  - 5.1 Selection Criteria
  - 5.2 Qwen 3 8B
  - 5.3 Qwen 3 32B
  - 5.4 Qwen 3 Coder 30B-A3B
  - 5.5 Llama 3.1 8B
  - 5.6 Llama 3.1 70B
- 6. Quantization Quality
  - 6.1 Perplexity
  - 6.2 KL Divergence
  - 6.3 Quantization Speed
- 7. Methodology
  - 7.1 Controlled Variables
  - 7.2 Metrics
  - 7.3 Benchmark Tools
  - 7.4 Reproducibility
- 8. Results
  - 8.1 Throughput
  - 8.2 Latency
  - 8.3 Power Consumption
  - 8.4 Tokens per Watt
  - 8.5 Cost Analysis
- 9. Discussion
  - 9.1 Hopper vs Blackwell
  - 9.2 Quantization Format Interaction
  - 9.3 Engine Maturity
  - 9.4 Multi-GPU Outlook
  - 9.5 Limitations
- 10. Conclusion
  - 10.1 Recommendations
  - 10.2 Future Work
- 11. Appendix
  - A. Raw Data
  - B. Configs
  - C. Hardware Setup
  - D. Software Versions

1. Introduction

1.1 Why Quantization Matters

Llama 3.1 8B
Llama 3.1 70B
Qwen 3 8B
Qwen 3 32B
Qwen 3 Coder 30B-A3B
FP16
FP8
INT8
INT4
Model size (GB)
Max context (tokens)