Report: GPU Upgrade for Local LLM Deployment

Current Problem Setup:
- Existing Hardware: AMD Ryzen Threadripper 3960X (24C/48T), 256 GB RAM, 1 × NVIDIA RTX 3090 (24 GB VRAM)
- Models: LLaMA 3.1 (11B) and DeepSeek Coder v2 (16B), AWQ 4-bit quantized
- Serving Framework: vLLM
- Requirement: Long context inference (≥ 15k–20k tokens), at least 300 concurrent users
- Current Limitation: RTX 3090 (24 GB) cannot hold both models with 15k+ tokens in parallel. Relies on paged KV cache → high latency, poor concurrency for long contexts.

Upgrade Requirement:
- Need GPUs with ≥ 48 GB VRAM to support long-context inference efficiently.
- Target setup: One GPU for LLaMA 11B, one GPU for DeepSeek 16B.
- Must ensure stability and throughput for ~300 concurrent users, while supporting 15k–20k tokens.

GPU Model	VRAM	Strengths	Limitations	Approx Price (INR)
NVIDIA RTX 3090 (existing)	24 GB GDDR6X	Good single-model inference up to ~8k tokens	Cannot handle 15k–20k ctx efficiently; paging required	Already Owned
NVIDIA RTX A6000 (Ampere)	48 GB GDDR6	Reliable for 15k–20k ctx; strong parallelism with 3090	Slower than Ada/H100 but affordable	■3.5–4.0 L (used/new)
NVIDIA RTX 6000 Ada	48 GB GDDR6	High throughput; modern Ada arch; 15k–20k ctx fits comfortably	Expensive vs A6000	■4.5–5.0 L
NVIDIA H100 PCIe	80 GB HBM2e	Enterprise-grade; MIG for multi-tenant workloads; top performance	Very expensive	■27–30 L
NVIDIA A100 (80 GB)	80 GB HBM2e	Proven LLM inference GPU; supports long ctx	Older, slower vs H100	■9–11 L
NVIDIA RTX 4090	24 GB GDDR6X	High raw TFLOPs; great for shorter ctx workloads	VRAM too small for 20k ctx per model	■2.0–2.5 L

Recommendation:
- Best Value Upgrade: NVIDIA RTX A6000 (48 GB) → Pair with existing RTX 3090 for true parallel model serving.
- Stretch Goal: NVIDIA RTX 6000 Ada (48 GB) if budget allows.
- Long-Term Enterprise Option: NVIDIA H100 (80 GB) for maximum concurrency, but outside current budget.