In a professional homelab environment like **Sentinel-Ops**, a GPU is **highly recommended but not strictly mandatory** for a RAG system, provided you make specific architectural choices.

Since you are running **RKE2 on Ubuntu 24.04** within **Proxmox**, you are currently relying on "CPU Inference." Here is the data-driven breakdown to help you decide if it's time to add a card.

1. The Reality of CPU vs. GPU Performance (2026 Benchmarks)

Using **Ollama** on a modern CPU (AVX-512 supported) vs. a mid-range consumer GPU (e.g., RTX 3060/4060):

Metric	CPU Only (16-core VM)	Consumer GPU (8GB+ VRAM)
Tokens per Second (8B Model)	3–8 tokens/sec (Slow Reading)	40–60 tokens/sec (Instant)
Time to First Token (TTFT)	2–5 seconds	< 0.5 seconds
System Impact	Near 100% CPU Load (Stalls other VMs)	< 10% CPU Load (Offloaded to GPU)
Concurrency	1 User (Sequential)	Multiple Users/Agents

> **Verdict:** If this RAG system is just for you to occasionally query your docs, **CPU is fine**. If you want this to act as a real-time agent or a "Sentinel" that constantly monitors and reacts to your RabbitMQ logs, the CPU latency will become a bottleneck.

---

2. How to succeed WITHOUT a GPU

If you choose not to buy a GPU, you must optimize your **Sentinel-Ops** stack for "Small Language Models" (SLMs) and efficient embeddings.

- **Model Selection:** Do not attempt to run Llama 3.1 8B or larger. Stick to the "Sub-4B" class.

    - **Recommendation:** `qwen2.5:3b` or `llama3.2:3b`. These are specifically optimized for CPU-heavy environments and still maintain high reasoning for RAG tasks.
- **Vector Database Choice:** Use **Qdrant** or **Milvus** with HNSW indexing. These are much faster at searching through chunks on a CPU than a basic ChromaDB instance.
- **Embeddings:** Use a lightweight embedding model like `bge-small-en-v1.5`. It produces smaller vectors (384 dimensions), making the retrieval phase significantly faster on your Ubuntu nodes.

---

3. Proxmox-Specific CPU Optimizations

To get the most out of your current hardware for AI:

1. **CPU Type:** In Proxmox, set the VM CPU type to **"host"** instead of "kvm64". This allows the VM to use modern instruction sets (AVX, AVX2, AVX-512) required for AI math.
2. **Pinning:** Use CPU pinning in Proxmox to ensure the RKE2 AI nodes have dedicated physical cores that aren't shared with your RabbitMQ or storage nodes.
3. **Memory:** Ensure you have at least **32GB of RAM** in the VM. CPU inference relies heavily on fast system RAM since there is no VRAM.

---

4. When should you definitely buy a GPU?

You should consider a low-cost used GPU (like an **NVIDIA Tesla T4** or an **RTX 3060 12GB**) if:

- **You want to use "Vision":** If you want to index screenshots of your Proxmox dashboards or network diagrams.
- **You want "Real-time" response:** If waiting 10 seconds for an answer feels like a friction point.
- **You want "Reasoning" models:** If you want to run `DeepSeek-R1` or larger models that require 14B+ parameters to be effective.

Would you like me to help you optimize your current RKE2 manifests to run a "CPU-Optimized" version of Ollama and Qdrant?

The **OptiPlex 7050 SFF** is a slight step up from the 7020, but the physical constraints for a GPU remain nearly identical. Because the PCIe x16 slot is still located at the bottom of the motherboard, the proximity to the Power Supply (PSU) is your primary hurdle.

For a 2026 RAG project on Ubuntu 24.04, here is the expert recommendation for the 7050 SFF:

1. The Physical Constraint: Single-Slot ONLY

In the 7050 SFF, the Blue PCIe x16 slot is extremely close to the PSU.

- **Do not buy a "Dual-Slot" or "2-Slot" card.** Even if it is low-profile (half-height), the thick heatsink/fan will physically hit the power supply or have zero air clearance.
- **The Workaround:** Some users put a dual-slot card in the **Black PCIe x4 slot** (the one further from the PSU). However, this drops your bandwidth significantly and is not recommended for LLM performance where data transfer between CPU/RAM and GPU VRAM is a frequent bottleneck.

---

2. Top GPU Picks for OptiPlex 7050 SFF

Option A: NVIDIA RTX A1000 (8GB GDDR6) — *The Best for AI*

This is the modern successor to the T1000 and is the absolute best card that fits this machine without modifications.

- **VRAM:** 8GB (Crucial for running Llama 3.1 8B or Mistral).
- **Form Factor:** Low-profile, **Single-slot**.
- **Power:** 50W (Safely powered by your stock 180W or 240W PSU).
- **Why for RAG:** It has 72 Tensor Cores specifically for AI acceleration. On Ubuntu 24.04, this card is natively supported by the standard `nvidia-driver-550` or higher.

Option B: NVIDIA RTX A400 (4GB GDDR6) — *The Budget Modern Choice*

If the A1000 is too expensive, this is a newer entry-level card (released 2024/2025).

- **VRAM:** 4GB (Tight for RAG, but works for "Small Language Models" like Phi-3 or Qwen-2.5 3B).
- **Benefit:** It is very power-efficient and ultra-slim.

Option C: NVIDIA T1000 (4GB or 8GB) — *The Reliable Used Market Pick*

The previous generation of the A1000. It is a solid single-slot card.

- **Note:** Try to find the **8GB version**. The 4GB version will frustrate you quickly when trying to load modern LLMs alongside your vector database.

---

3. Critical SRE Advice for your RKE2 Stack

Since you are running RKE2 on this machine, adding a GPU changes how you manage the cluster:

- **NVIDIA Operator:** Once the card is installed in the Ubuntu 24.04 host, you should deploy the **NVIDIA GPU Operator** via Helm. This automates the driver and toolkit injection into your RKE2 pods so `ollama` can see the hardware.
- **Power Supply Check:** Most 7050 SFFs come with a **180W PSU**. While a 50W GPU *should* work (the i7-7700 is 65W), if you find the system rebooting during heavy LLM inference, you may need to source the **240W PSU** (Dell Part: `J82T5` or similar) from an OptiPlex 7050 Tower or XE3 SFF.

Summary for your Build:

Part	Specification
Ideal GPU	NVIDIA RTX A1000 (8GB)
Max Width	Single Slot (approx. 18-20mm)
Max Height	Low Profile (Half-Height)
Power Needs	No external 6-pin/8-pin power

Would you like me to find the specific NVIDIA Helm chart values needed to enable GPU support in your RKE2 cluster once the card arrives?