RTX 5060 Ti 16GB Killed? Local AI Alternatives

January 26, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer**: The RTX 5060 Ti 16GB isn't officially discontinued, but production is being quietly strangled by GDDR7 shortages. Street prices have jumped from $429 MSRP to ~$500—a 17% markup in under a year. If you need affordable 16GB VRAM for local AI, act now: grab one if you find it near MSRP, or pivot to a used RTX 3090 ($700-850) for 24GB. The 8GB models aren't worth considering for LLM work.

📚 **More on this topic:** GPU Buying Guide · GPU Prices Rising · Budget AI PC Build

Image: RTX 5060 Ti 16GB graphics card with "limited stock" warning overlay

The rumors started circulating in late December 2025: NVIDIA was killing off the RTX 5060 Ti 16GB. For anyone running local AI, this felt like a gut punch. The 16GB variant was supposed to be the affordable on-ramp to serious LLM work—finally, a new card with enough VRAM to run models beyond toy demos.

So what's actually happening? After digging through conflicting reports, official denials, and supply chain leaks, here's the situation: NVIDIA isn't officially discontinuing the card. But they're making it increasingly hard to buy one at a reasonable price. For local AI enthusiasts who've been waiting for affordable VRAM, the window is closing.

What's Actually Happening with the RTX 5060 Ti 16GB

The Rumors

In mid-January 2026, Hardware Unboxed reported that ASUS had designated the RTX 5060 Ti 16GB as "end of life" with production halted. The RTX 5070 Ti reportedly got the same treatment. According to their sources, supply had been "significantly reduced to the point of being effectively discontinued."

This followed earlier reports from Board Channels suggesting NVIDIA was preparing to pause production entirely due to GDDR7 price spikes.

The Official Denials

Both ASUS and NVIDIA pushed back hard. ASUS released a statement clarifying: “The GeForce RTX 5070 Ti and GeForce RTX 5060 Ti 16GB have not been discontinued or designated as end-of-life. ASUS has no plans to stop selling these models.”

NVIDIA added: “Demand for GeForce RTX GPUs is strong, and memory supply is constrained. We continue to ship all GeForce SKUs and are working closely with our suppliers to maximize memory availability.”

The Reality

Here’s the deal: both statements can be technically true while the practical outcome still hurts consumers. The cards aren’t “discontinued” in the sense that NVIDIA announced an EOL date. But production is being deprioritized, stock is drying up, and prices are climbing.

The RTX 5060 Ti 16GB launched at $429 MSRP in April 2025. Nine months later, street prices have hit **~$500 on eBay**—a **17% markup** over MSRP. That’s not normal market variance; it’s a card that’s becoming scarce. Some retailers are selling out entirely. This isn’t a cancellation—it’s a slow suffocation.

Image: Price chart showing RTX 5060 Ti 16GB climbing from $429 MSRP to $500 street price (17% increase)

Why NVIDIA Is Deprioritizing 16GB Cards

The culprit is GDDR7 memory, and the math is brutal.

The Memory Shortage

GDDR7 prices have spiked dramatically due to a “DRAM supercycle” driven by AI datacenter demand. Hyperscalers like Google, Microsoft, and Amazon are gobbling up memory supply for training infrastructure. AI workloads are expected to consume 20% of total DRAM production in 2026.

The result: GDDR6 and GDDR7 prices have increased by several hundred percent in recent months. Some analysts predict GPU retail prices could eventually double. NVIDIA reportedly plans to cut RTX 50 series production by 30-40% in early 2026.

The Revenue-Per-Gigabyte Problem

Here's why the 16GB cards specifically are getting squeezed. Gigabyte's CEO explained NVIDIA's allocation strategy in stark terms—it comes down to gross revenue per gigabyte of GDDR7:

GPU Model	Revenue per GB of GDDR7
RTX 5060 Ti 8GB	$47.38
RTX 5060 8GB	$37.38
RTX 5060 Ti 16GB	$26.81

The RTX 5060 Ti 16GB generates the lowest revenue per gigabyte of any RTX 50 series card. When memory is scarce, NVIDIA makes more money putting those chips into 8GB cards. The 16GB variant is getting deprioritized not because it's unpopular—but because it's too good a deal for consumers.

What This Means for Local AI Users

This couldn't come at a worse time. Local AI has hit a genuine inflection point. Models like Llama 3, Mistral, and Qwen are genuinely useful. The software stack (llama.cpp, Ollama, text-generation-webui) is mature. People are actually running local AI for real work, not just demos.

Why 16GB Matters

For LLM inference, VRAM is the bottleneck. Here's what different VRAM tiers actually get you:

VRAM	What You Can Run
8GB	7B models (Q4 quantized), barely. Forget anything larger.
16GB	7B-13B unquantized, 34B quantized, comfortable headroom
24GB	70B quantized, fine-tuning smaller models, image gen

8GB is the bare minimum for local LLMs, and "minimum" means constant frustration. You'll be squeezing models into memory, dealing with context length limitations, and watching inference crawl when you push the limits.

16GB is where local AI actually becomes practical for daily use. You can run Llama 3 8B at full precision, or fit a 34B model with Q4 quantization. It's the sweet spot for enthusiasts who want real capability without spending $2,000+.

For a deeper breakdown, see our VRAM requirements guide.

→ Use our Planning Tool to check exact VRAM for your setup.

The 8GB Trap

If NVIDIA successfully pushes the market toward 8GB cards, local AI enthusiasts get squeezed. The RTX 5060 Ti 8GB at $379 looks cheaper, but it's a false economy. You'll hit the VRAM wall within months as models continue growing, and you can't upgrade VRAM—only the whole card.

The irony: NVIDIA's own benchmarks show the 16GB card is "vastly superior in most areas" and "a better value" despite the $50 premium. Tom's Hardware's testing concluded that "8GB GPUs are no longer as viable as they once were."

Your Options Right Now

Let's get specific. Here's what makes sense depending on your situation:

Image: Comparison table of GPU options for local AI in 2026

GPU	VRAM	Street Price (Jan 2026)	Best For	The Catch
RTX 5060 Ti 16GB	16GB	~$500 (17% over MSRP)	New card buyers who want warranty + efficiency	Stock disappearing, prices rising
RTX 5060 Ti 8GB	8GB	$349-379	Gaming only. Skip for AI.	Insufficient VRAM for serious LLM work
Used RTX 3090	24GB	$700-850	Maximum VRAM per dollar	Used market, 350W power draw, older architecture
AMD RX 9060 XT 16GB	16GB	$349-399	Budget + Linux users	21% slower inference than RTX 5060 Ti, ROCm quirks
RTX 4090	24GB	$2,000+	Money is no object	Costs 3x a used 3090 for ~40% more speed

Option 1: Hunt for an RTX 5060 Ti 16GB Near MSRP

If you can find one at $429-450, buy it. The card delivers solid performance, has a reasonable 150W TDP, and full support for CUDA and the latest AI stack. Check Newegg, Walmart, Amazon, and Micro Center regularly.

The reality check: most cards are now selling around $500. That 17% premium might still be worth it for a new card with warranty—but don't overpay beyond that. At $550+, the value proposition breaks down.

Option 2: Used RTX 3090 (The Value King)

The RTX 3090 remains the best VRAM-per-dollar option in 2026. For $700-850 on eBay or r/hardwareswap, you get 24GB of VRAM—enough to run Llama 70B quantized or pair with another GPU for serious multi-card setups.

The card is three generations old, but 24GB is 24GB. VRAM doesn't age. The Ampere architecture has mature, stable support in every AI framework.

The tradeoffs: it's a used card (no warranty, lottery on condition), it pulls 350W under load (budget for a beefy PSU), and it won't match a 4090 or 5090 in raw speed. But for local inference where you're memory-bound anyway, the 3090 often performs within striking distance of newer cards.

We wrote a full used RTX 3090 buying guide covering where to buy, what to watch out for, and how to test your card.

Option 3: AMD RX 9060 XT 16GB

AMD's $349 MSRP makes this tempting, and you get 16GB of VRAM with RDNA 4 architecture. ROCm 6.4.1 officially supports the card, and it runs llama.cpp fine on Linux.

But here's the catch: inference speed is significantly slower than NVIDIA equivalents. The RTX 5060 Ti 8GB (not even the 16GB) delivers 21% faster token generation than the RX 9060 XT 16GB. Time to first token is 0.22 seconds on NVIDIA versus 0.57 seconds on AMD.

If you're on Linux, comfortable with ROCm, and prioritize running larger models over speed, the 9060 XT works. For Windows users or anyone who values snappy responses, stick with NVIDIA.

For the full picture, read our AMD vs NVIDIA comparison for local AI.

Option 4: Wait for RTX 50 Super?

Rumors suggest RTX 50 Super cards might arrive in Q3 2026. Maybe they'll have better VRAM configurations at reasonable prices.

Don't count on it. The GDDR7 shortage is projected to last until Q4 2027 or even 2028. Waiting means paying more later, not less. The memory crunch isn't a temporary supply chain hiccup—it's a structural shift driven by AI datacenter demand that isn't slowing down.

The Bottom Line

The RTX 5060 Ti 16GB isn't dead, but it's on life support. NVIDIA's official denials don't change the reality: stock is drying up, prices have jumped 17% over MSRP, and the company has every financial incentive to push 8GB cards instead.

If you need affordable VRAM for local AI, here's what to do:

- **Have ~$500 and want new?** Hunt for an RTX 5060 Ti 16GB. Buy immediately if you find one near MSRP ($429), accept ~$500 if you need it now.
- **Have $700-850 and comfortable with used?** Get an RTX 3090. It's 24GB for barely more than a marked-up 16GB card.
- **Budget under $400 and on Linux?** The AMD RX 9060 XT 16GB works, but expect slower inference.
- **Only considering 8GB cards?** Don't. You'll regret it within a year.

The local AI community has been asking for affordable VRAM for years. We briefly got it with the RTX 5060 Ti 16GB at $429. That window is closing. Act accordingly.

Related Guides

- GPU Buying Guide for Local AI
- NVIDIA GPU Prices Are Rising: What Local AI Enthusiasts Need to Know
- Build a Local AI PC for Under $500

Sources: Tom's Hardware, VideoCardz, TweakTown, WCCFTech, XDA Developers

Source: https://insiderllm.com/guides/rtx-5060-ti-16gb-local-ai-options/

Free guides for running AI locally