RTX 4090 vs Used RTX 3090 for Local AI: Which to Buy in 2026

February 21, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** For LLM inference, the used RTX 3090 ($700-$1,000) delivers 70-80% of the RTX 4090's token generation speed at roughly a third of the price ($2,200-$2,800 new). Both have 24GB VRAM — they run the same models at the same quality. The 4090 is faster, not more capable. For 90% of local AI hobbyists, the 3090 is the better buy. The exception: heavy image generation or regular fine-tuning, where the 4090's compute advantage actually matters. The 3090 also supports NVLink — two of them give you 48GB VRAM for ~$1,800 total, running 70B models that neither card can touch alone.

📚 **More on this topic:** Used RTX 3090 Buying Guide · What Can You Run on 24GB VRAM? · GPU Buying Guide · VRAM Requirements

This is the GPU decision that comes up more than any other in local AI communities. New RTX 4090 or used RTX 3090? Both have 24GB VRAM. Both run the same models. One costs two to three times more than the other.

People agonize over this. They shouldn't. The answer is clear for most builders — but the right answer depends on what you're actually doing with the card. Let's get into the numbers.

The Specs That Matter

Spec	Used RTX 3090	New RTX 4090
VRAM	**24GB GDDR6X**	**24GB GDDR6X**
Memory Bandwidth	936 GB/s	1,008 GB/s
CUDA Cores	10,496	16,384
Tensor TFLOPs	285	660
L2 Cache	6 MB	72 MB
TDP	350W	450W
Architecture	Ampere (GA102)	Ada Lovelace (AD102)

Spec	Used RTX 3090	New RTX 4090
NVLink Support	Yes	No
Price (Feb 2026)	**$700-$1,000**	**$2,200-$2,800**

The number that matters most for local AI is at the top: 24GB VRAM. Both cards have it. This means they load the same models at the same quantization levels. A Qwen3 32B at Q4 fits on both. Neither can run a 70B model without offloading.

The 4090 has 56% more CUDA cores, 132% more Tensor TFLOPs, and a massive 12x L2 cache advantage. But memory bandwidth — the bottleneck for LLM token generation — differs by only 8%. That 8% explains why the benchmark gap is much smaller than the spec sheet suggests.

LLM Inference: The Gap Is Smaller Than You Think

LLM token generation is memory bandwidth-bound. The model weights live in VRAM, and each token requires reading through them. More bandwidth means more tokens per second. More CUDA cores barely help.

Model	RTX 3090	RTX 4090	4090 Advantage
8B Q4_K	~87-101 tok/s	~104-128 tok/s	+20-30%
14B Q4_K	~52 tok/s	~69 tok/s	+33%
30B MoE Q4_K	~114 tok/s	~140 tok/s	+23%

Sources: Hardware Corner GPU benchmarks, llama.cpp testing.

That's the real gap. Not 2x. Not 50%. It's 20-33% faster depending on model size. The 4090's cache advantage helps more with prompt processing (the initial ingestion phase), but for the thing you actually experience — watching tokens appear — the difference is modest.

For interactive chat, both cards are well above human reading speed on every model that fits in 24GB. The 3090 at 87 tok/s on an 8B model feels instant. The 4090 at 104 tok/s also feels instant. You won't perceive the difference in a chat window.

Where the speed gap starts to matter: coding agents that chain multiple requests, batch processing documents through a RAG pipeline, or serving multiple users from a single GPU. If throughput is your bottleneck, the 4090 helps. If you're one person chatting with one model, it doesn't.

Image Generation: Where the 4090 Actually Earns Its Price

This is the one workload where the 4090 genuinely pulls away.

Workload	RTX 3090	RTX 4090	4090 Advantage
SD 1.5	16.7 it/s	21.0 it/s	+26%
SDXL	Baseline	~40-70% faster	Significant
Flux	~19 s/image	~13 s/image	+46%

Image generation is compute-bound, not bandwidth-bound. Every diffusion step hammers the Tensor Cores. The 4090 has 132% more Tensor TFLOPs, and it shows. If you're generating 50 Flux images in a batch, the difference between 16 minutes and 11 minutes is real. Over hundreds of images, it adds up to hours.

If image generation is your primary use case, the 4090 is worth considering. For LLM-first users who occasionally generate an image, it's not.

Fine-Tuning: Same Story as Image Gen

Fine-tuning is compute-bound. The 4090 delivers 1.3-1.9x higher throughput depending on the framework and precision:

- BERT Base fine-tuning (FP16): 4090 at 297 tokens/s vs 3090 at 172 tokens/s (1.7x)
- LoRA/QLoRA training: both handle models up to ~20-30B, but the 4090 finishes in roughly half the time
- The 4090 has native FP8 Tensor Core support that the 3090 lacks, enabling further speedups with compatible frameworks

If you fine-tune regularly — weekly, not once-a-year — the 4090 saves meaningful time. If you fine-tune occasionally, the 3090 gets the job done; it just takes longer.

The Money Math

This is where the decision gets clear.

Used RTX 3090: ~$700-$1,000 (eBay, r/hardwareswap, local deals). Lower end for older FE cards, higher for newer AIB models. Check our used RTX 3090 buying guide for what to look for.

New RTX 4090: ~$2,200-$2,800. Original MSRP was $1,599, but that price is long gone. DRAM shortages and tariff pressures have pushed real retail prices well above MSRP.

The price gap: $1,200-$2,100. What could you do with that?

Option	Cost	What It Gets You
Second used RTX 3090 + NVLink bridge	~$800-$1,100	48GB VRAM — run 70B models
128GB DDR4/DDR5 RAM upgrade	~$150-$300	Better CPU offloading for huge models
Entire second machine (used workstation + GPU)	~$1,000-$1,500	Distributed inference, dedicated tasks
The savings	$0	In your pocket

The most compelling option: **two used 3090s with NVLink.** The RTX 3090 is the only consumer GeForce card that supports NVLink. The 4090 does not. The 5090 does not. Two NVLink'd 3090s give you 48GB of pooled VRAM for roughly the price of a single 4090.

48GB opens up 70B models at Q4 — a category of model neither single card can run. That's not a speed upgrade. That's a capability upgrade. A 70B model at Q4 on 48GB is smarter than any model a single 24GB card can run at any speed.

The catch: you need a motherboard with two x16 PCIe slots with proper spacing, a 1000W+ PSU, and willingness to deal with two 3-slot cards in one case. It's not for everyone. But if you have the setup for it, the value is hard to argue with.

→ Not sure what fits? Try our Planning Tool.

Power: Not a Factor

GPU	Inference Draw	Annual Cost (8hr/day)
RTX 3090	~310W	~$163
RTX 4090	~275W	~$144
Difference		**~$19/year**

At US average electricity rates ($0.18/kWh), the annual power cost difference is about $19 for 8 hours of daily use. Even at 24/7 operation, it's ~$55/year. The 4090 is actually more power-efficient per token due to its architecture, but the absolute savings don't move the needle on a purchase decision.

Don't let anyone tell you the 3090's power draw is a reason to buy a 4090. The price difference would take 60+ years to recoup in electricity savings.

What About the RTX 5090?

The RTX 5090 launched at $1,999 MSRP with 32GB GDDR7 and 1,792 GB/s bandwidth — nearly 2x the bandwidth of either card here. Real street prices are $2,900-$3,500+. It's a generational leap for LLM inference.

If your budget is $2,500+ and you can actually find one at a reasonable price, the 5090's 32GB VRAM and bandwidth advantage make it the better buy over the 4090. But at 3-5x the cost of a used 3090, it's a different conversation entirely — and one for a different article.

The Verdict

Buy the used RTX 3090 if:

- LLM inference is your primary use case (chat, coding, RAG)
- You're budget-conscious (most of our readers)
- You want to spend savings on a second card for 48GB VRAM
- You're comfortable buying used (here's how)
- You want the most capability per dollar — which is the whole game in local AI

Buy the RTX 4090 if:

- Image generation is your primary workload (Flux, SDXL, ComfyUI)
- You fine-tune models regularly (weekly+)
- You're running a multi-user inference server where throughput matters
- Budget isn't the constraint, simplicity is
- You want a single new card with warranty

For 90% of local AI builders, the used RTX 3090 is the better buy. Same VRAM. Same models. Same quality. 70-80% of the speed. A third of the price. And the option to go to 48GB via NVLink that the 4090 can never match.

The 4090 is a great GPU. It's just not $1,500 better for what most people do with local AI. Spend the difference on more VRAM, more RAM, or more hardware. In this game, capability beats speed.
