RTX 3090 vs 4070 Ti Super for Local LLMs

February 4, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Buy the used RTX 3090 ($700-800) if you want to run 32B+ models or need VRAM headroom for long contexts and future models. Buy the RTX 4070 Ti Super ($800 new) if you want warranty, lower power bills, and will stick to 7B-14B models. The 3090's 24GB beats the 4070 Ti Super's 16GB for AI workloads — VRAM capacity matters more than raw speed when your model doesn't fit.

📚 **More on this topic:** Used RTX 3090 Buying Guide · What Can You Run on 24GB VRAM · What Can You Run on 16GB VRAM · GPU Buying Guide

The RTX 3090 and RTX 4070 Ti Super sit at similar price points but make very different tradeoffs. One is a five-year-old flagship with massive VRAM. The other is a current-gen card with a warranty and better efficiency. For gaming, the 4070 Ti Super wins. For local AI, the answer depends entirely on what models you want to run.

This guide compares both cards head-to-head for LLM inference, image generation, and practical AI workloads.

Specs at a Glance

Spec	RTX 3090	RTX 4070 Ti Super
VRAM	24GB GDDR6X	16GB GDDR6X
Memory Bandwidth	936 GB/s	672 GB/s
CUDA Cores	10,496	8,448
Tensor Cores	328 (3rd gen)	264 (4th gen)
TDP	350W	285W
Architecture	Ampere (2020)	Ada Lovelace (2024)
New Price	~$1,400+ (if available)	$799
Used Price	$700-850	$650-750

Spec	RTX 3090	RTX 4070 Ti Super
Warranty	None (used)	3 years (new)

The numbers tell the story: the 3090 has 50% more VRAM and 39% more memory bandwidth. The 4070 Ti Super has newer architecture, lower power draw, and a warranty.

LLM Performance: Real-World Benchmarks

Token Generation Speed

For models that fit on both cards, the 4070 Ti Super is slightly faster due to its newer architecture:

Model	RTX 3090	RTX 4070 Ti Super
Llama 3.1 8B Q4	~87-111 tok/s	~95-120 tok/s
Qwen 2.5 14B Q4	~45-55 tok/s	~50-60 tok/s
Mistral 7B Q4	~90-115 tok/s	~100-125 tok/s

The 4070 Ti Super wins by 10-15% on raw token generation for equivalent models.

But VRAM Changes Everything

Here's where the comparison gets interesting:

Model	RTX 3090 (24GB)	RTX 4070 Ti Super (16GB)
Qwen 2.5 32B Q4 (~20GB)	~35-42 tok/s	Won't fit
DeepSeek R1 Distill 32B Q4	~35-40 tok/s	Won't fit
Llama 3.3 70B Q3 (~30GB)	~12-18 tok/s (offload)	Won't fit
Mixtral 8x7B Q4 (~26GB)	~25-30 tok/s	Won't fit

The RTX 3090 runs entire model classes that the 4070 Ti Super simply cannot load. This isn't a speed difference — it's the difference between running and not running.

Context Length Matters Too

Longer conversations eat VRAM for the KV cache. With a 14B model:

Context Length	KV Cache Size	3090 Headroom	4070 Ti Super Headroom
4K tokens	~0.8 GB	15GB+ free	7GB free
8K tokens	~1.6 GB	14GB+ free	6GB free
16K tokens	~3.2 GB	12GB+ free	4GB free (tight)
32K tokens	~6.4 GB	9GB+ free	Won't fit

The 3090 handles 32K+ context windows comfortably. The 4070 Ti Super struggles past 16K on larger models.

Image Generation Performance

For Stable Diffusion and Flux, both cards are capable but the 3090's VRAM advantage shows again:

Task	RTX 3090	RTX 4070 Ti Super
SDXL 1024x1024	~8-10 sec	~6-8 sec
SD 1.5 512x512	~3-4 sec	~2-3 sec
Flux (FP8)	~25-35 sec	~30-40 sec (tight)
Flux (FP16)	~15-20 sec	Won't fit

The 4070 Ti Super is faster for SDXL due to Ada architecture improvements. But Flux at full FP16 precision needs 22GB+ — only the 3090 can run it without quantization.

For LoRA training, the 3090's 24GB is a major advantage. Training at 1024x1024 resolution with reasonable batch sizes requires 20GB+.

Power and Heat

This is where the 4070 Ti Super shines:

Metric	RTX 3090	RTX 4070 Ti Super
TDP	350W	285W
Typical AI Load	300-340W	240-270W
Idle	25-35W	15-20W
Required PSU	850W	650W
Heat Output	Significant	Moderate
Noise	Loud under load	Moderate

Annual electricity cost (assuming 4 hours AI use daily, $0.12/kWh):

- RTX 3090: ~$50/year
- RTX 4070 Ti Super: ~$40/year

The difference isn't huge, but the 3090 runs hot and loud. If your setup is in a living space, this matters.

Power Supply Requirements

The RTX 3090 needs a beefy PSU:

- Minimum: 750W
- Recommended: 850W 80+ Gold
- Must use two separate 8-pin cables (not daisy-chained)

The RTX 4070 Ti Super is more forgiving:

- Minimum: 600W
- Recommended: 650W 80+ Gold
- Single 16-pin or adapter works fine

If your current PSU is under 750W, factor in a $80-120 upgrade for the 3090.

Price Analysis

Current Market Prices (February 2026)

Card	New Price	Used Price
RTX 3090	$1,400+ (rare)	$700-850
RTX 4070 Ti Super	$799	$650-750

Price Per GB of VRAM

Card	VRAM	Price	$/GB
RTX 3090 (used)	24GB	$750	$31/GB
RTX 4070 Ti Super (new)	16GB	$799	$50/GB

The 3090 delivers 50% more VRAM for roughly the same money. For AI workloads where VRAM is the limiting factor, that's exceptional value.

Total Cost of Ownership

RTX 3090 (used):

- Card: $750
- PSU upgrade (if needed): $100
- Higher electricity: +$10/year
- No warranty
- **Total: $850 + risk**

RTX 4070 Ti Super (new):

- Card: $799
- PSU: Likely no upgrade needed
- 3-year warranty included
- **Total: $799 + peace of mind**

Which Models Can You Run?

RTX 3090 (24GB) — Model Compatibility

Model Tier	Examples	Performance
7B-8B	Llama 3.1 8B, Mistral 7B	Excellent (80-110 tok/s)
13B-14B	Qwen 2.5 14B, DeepSeek R1 8B	Great (45-60 tok/s)
32B	Qwen 2.5 32B, DeepSeek R1 32B	Good (35-42 tok/s)
70B Q3-Q4	Llama 3.1 70B	Usable (12-18 tok/s with offload)
Mixtral 8x7B	MoE models	Good (25-30 tok/s)

RTX 4070 Ti Super (16GB) — Model Compatibility

Model Tier	Examples	Performance
7B-8B	Llama 3.1 8B, Mistral 7B	Excellent (95-125 tok/s)
13B-14B Q4-Q5	Qwen 2.5 14B	Great (50-60 tok/s)
13B-14B Q6-Q8	Higher quality	Tight fit
32B	—	Won't fit
70B	—	Won't fit

The 4070 Ti Super maxes out at the 14B tier. The 3090 extends into 32B and can squeeze 70B with compromises.

→ Not sure what fits? Try our Planning Tool.

The Warranty Question

This is the 3090's biggest weakness. Used cards have no warranty. If it dies in month 2, you're out $750.

Mitigating the risk:

- Buy from eBay for 30-day Money Back Guarantee
- Stress test thoroughly within the return window

- Check thermal paste and pad condition
- Monitor memory junction temps (keep under 100°C)

The 4070 Ti Super's warranty advantage:

- 3 years coverage from NVIDIA/AIB partner
- RMA process if anything fails
- Peace of mind for professional use

If you're using the card for paid work or can't afford to lose $750, the warranty has real value.

Use Case Recommendations

Buy the RTX 3090 If:

- **You want to run 32B models** — Qwen 2.5 32B, DeepSeek R1 Distill 32B, Mixtral 8x7B
- **You need long context windows** — 32K+ tokens for document analysis, RAG
- **You plan to fine-tune or train LoRAs** — 24GB makes this practical
- **You want Flux at full quality** — FP16 needs 22GB+
- **Budget is the priority** — Best VRAM-per-dollar available
- **You're comfortable buying used** — Can handle testing and potential issues

Buy the RTX 4070 Ti Super If:

- **7B-14B models meet your needs** — Most users fall here
- **You want a warranty** — 3 years of coverage
- **Power efficiency matters** — 65W less draw, less heat, less noise
- **Your PSU is under 750W** — No upgrade needed
- **You prefer new hardware** — Known good condition
- **You also game** — Better rasterization performance

Skip Both and Consider:

- **RTX 4090 ($1,599)** — If you need 24GB with modern architecture and can afford it
- **RTX 3060 12GB ($180-220 used)** — If budget is very tight and 14B models are sufficient
- **Dual 3090s** — If you need 48GB for 70B models at good quality (complex setup)

Real-World Scenarios

Scenario 1: Coding Assistant

You want a local coding model for daily development work.

Winner: Either works, 4070 Ti Super edges out

DeepSeek Coder 14B and Qwen 2.5 Coder 14B fit on both cards. The 4070 Ti Super runs them slightly faster with less power and heat. Unless you want 32B coding models, save the hassle of used hardware.

Scenario 2: Running 70B Models

You want to run Llama 3.1 70B or Qwen 72B locally.

Winner: RTX 3090 (but barely)

The 3090 can run 70B at Q3-Q4 with CPU offloading at 12-18 tok/s. It's slow but functional. The 4070 Ti Super can't run 70B at all. However, if 70B is your primary goal, consider dual 3090s or saving for an RTX 4090.

Scenario 3: Image Generation Focus

You primarily run Stable Diffusion and Flux.

Winner: Depends on Flux priority

For SDXL, the 4070 Ti Super is faster. For Flux at FP16, only the 3090 works. If you're doing professional image work with Flux, the 3090's VRAM is essential.

Scenario 4: Future-Proofing

You want a card that will remain useful as models grow.

Winner: RTX 3090

Models are getting bigger, not smaller. The 3090's 24GB will remain relevant longer than 16GB. When 20B models become the new 7B, the 3090 will still run them.

The Bottom Line

RTX 3090 (used, $700-850): Buy this if VRAM matters to you. The 24GB opens up 32B models, long contexts, Flux at full precision, and future headroom. Accept the used hardware risk, power requirements, and noise.

RTX 4070 Ti Super (new, $799): Buy this if 14B models are enough. You get a warranty, lower power, and slightly faster inference on smaller models. Accept the 16GB VRAM ceiling.

The decision framework is simple: **count the parameters of the models you'll actually run**. If your answer includes 32B or above, buy the 3090. If your answer is 14B or below, buy the 4070 Ti Super.

For most hobbyists experimenting with local AI, 14B models like Qwen 2.5 14B and DeepSeek R1 Distill 14B deliver excellent results. The 4070 Ti Super handles these well. But if you're serious about local AI and want room to grow, the 3090's VRAM advantage is hard to ignore at this price.

Related Guides

- Used RTX 3090 Buying Guide
- What Can You Run on 24GB VRAM?
- What Can You Run on 16GB VRAM?
- GPU Buying Guide for Local AI
- VRAM Requirements for Local LLMs

Get notified when we publish new guides.

Subscribe — free, no spam
