RTX 3060 vs 3060 Ti vs 3070 for Local AI

February 8, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** The RTX 3060 12GB ($180-230 used) beats the 3060 Ti and 3070 for most local AI work — despite being the slowest card. Why? VRAM. The 3060's 12GB runs 14B parameter models at Q4. The 3060 Ti and 3070 have only 8GB, capping them at 7-8B models. For LLM inference, a model that fits on your card at 20 tok/s beats a model that doesn't fit at 0 tok/s. The 3070 only wins if you're focused on image generation speed or only run 7-8B models. The 3060 Ti is the worst choice for AI — less VRAM than the 3060, less speed than the 3070, more expensive than both for what you get.

📚 **Related:** Best GPU Under $300 · VRAM Requirements · GPU Buying Guide · Budget AI PC Build

This comparison makes no sense on paper. The RTX 3060 is the weakest card. Fewer CUDA cores, lower memory bandwidth, cheapest price. In any normal GPU ranking, it sits at the bottom of these three.

For local AI, it's the best of the three. And it's not close.

The reason is a single number: 12GB of VRAM. NVIDIA gave the 3060 more VRAM than the 3060 Ti or 3070 — a quirk of the product stack that makes the "worst" card the most capable for LLM inference.

The Specs

Spec	RTX 3060 12GB	RTX 3060 Ti 8GB	RTX 3070 8GB
VRAM	**12GB GDDR6**	8GB GDDR6	8GB GDDR6
Memory Bus	192-bit	256-bit	256-bit
Memory Bandwidth	360 GB/s	448 GB/s	448 GB/s
CUDA Cores	3,584	4,864	5,888
TDP	170W	200W	220W
Used Price (Feb 2026)	**$180-230**	$220-280	$280-350

Spec	RTX 3060 12GB	RTX 3060 Ti 8GB	RTX 3070 8GB
New Price	$280-330	Discontinued	Discontinued

The 3060 Ti and 3070 are better GPUs by every conventional metric except the one that matters most for AI: VRAM capacity.

For LLM Inference: The 3060 Wins

LLM inference is bottlenecked by VRAM, not compute. A model that doesn't fit in your GPU's memory either doesn't run at all or falls back to CPU offloading — dropping from 30+ tok/s to 2-5 tok/s.

What Each Card Can Run

Model	Size at Q4	RTX 3060 12GB	RTX 3060 Ti 8GB	RTX 3070 8GB
Llama 3.1 8B	~5 GB	Runs	Runs	Runs
Qwen 2.5 7B	~5 GB	Runs	Runs	Runs
Phi-4 14B	~9 GB	**Runs**	Won't fit	Won't fit
Qwen 2.5 14B	~9 GB	**Runs**	Won't fit	Won't fit
Gemma 3 12B	~8 GB	**Runs**	Tight fit	Tight fit
DeepSeek R1 Distill 14B	~9 GB	**Runs**	Won't fit	Won't fit

The 3060 runs 14B models. The other two don't. That's the entire argument.

14B models are meaningfully better than 7-8B models for reasoning, coding, and following complex instructions. Qwen 2.5 14B and Phi-4 14B punch well above what you'd expect from their size. Being locked out of this tier is a real limitation.

Inference Speeds

For models that fit on all three cards:

Model	RTX 3060	RTX 3060 Ti	RTX 3070
Llama 3.1 8B Q4	~35 tok/s	~45 tok/s	~55 tok/s

Model	RTX 3060	RTX 3060 Ti	RTX 3070
Qwen 2.5 7B Q4	~38 tok/s	~48 tok/s	~58 tok/s
Mistral 7B Q4	~40 tok/s	~50 tok/s	~60 tok/s

The 3070 is roughly 55-60% faster than the 3060 on identical models. That's a real difference — 55 tok/s feels snappier than 35 tok/s for interactive chat.

But here's the number that matters:

Model	RTX 3060	RTX 3060 Ti	RTX 3070
Qwen 2.5 14B Q4	~20 tok/s	Can't run	Can't run
Phi-4 14B Q4	~18 tok/s	Can't run	Can't run

20 tok/s on a 14B model versus 0 tok/s. The 3060 runs these models. The others don't. A slower card running a smarter model beats a faster card running a dumber model every time.

→ Not sure what fits? Try our Planning Tool.

For Image Generation: The 3070 Wins

Image generation flips the equation. SDXL and Flux need 8GB minimum — all three cards clear that bar. After that, speed is what matters: more CUDA cores and bandwidth mean faster renders.

Task	RTX 3060	RTX 3060 Ti	RTX 3070
SDXL 512x512 (20 steps)	~12 sec	~9 sec	~7 sec
SDXL 1024x1024 (20 steps)	~35 sec	~25 sec	~20 sec
Flux Dev (20 steps)	~45 sec	~35 sec	~28 sec

The 3070 generates images nearly twice as fast as the 3060. If you're iterating on prompts — generating dozens of variations to find the right one — that speed difference compounds. Twenty seconds per image versus thirty-five adds up over a session.

The 3060's extra VRAM helps for one specific image gen task: **LoRA training**. Fine-tuning SDXL LoRAs benefits from the extra memory headroom. If you're training, not just generating, the 3060's 12GB gives you more room for larger batch sizes.

For Mixed Workloads

Most people don't exclusively do LLM inference or exclusively do image generation. If you want one card for everything:

RTX 3060 12GB: Best flexibility. Runs 14B LLMs, handles SDXL/Flux, has headroom for LoRA training, and costs the least. Slower at everything, but capable of everything.

RTX 3070 8GB: Best speed on smaller models. If you know you'll stick to 7-8B LLMs and want faster image generation, the 3070 is a better experience. But you're permanently locked out of 14B models.

RTX 3060 Ti 8GB: The awkward middle. Same 8GB VRAM limitation as the 3070 but 17% slower. Costs more than the 3060, does less. It exists because NVIDIA needed a product between the 3060 and 3070 — not because anyone specifically needs this combination of specs.

The 3060 Ti Problem

The RTX 3060 Ti is the worst choice of the three for AI work. Here's why:

- **Less VRAM than the 3060:** 8GB vs 12GB. Same model limitations as the 3070.
- **Slower than the 3070:** 17% fewer CUDA cores, same memory bandwidth.
- **More expensive than the 3060:** $220-280 vs $180-230 for less capability.
- **Same VRAM as the 3070:** No advantage over the card above it.

The 3060 Ti makes sense for gaming, where its higher CUDA core count and bandwidth translate directly to more FPS. For AI, it falls between two chairs. If VRAM matters (LLMs), get the 3060. If speed matters (image gen), get the 3070. The 3060 Ti doesn't win either category.

The one exception: If someone offers you a 3060 Ti for $180 or less — close to 3060 prices — it's a fine card. Same 8GB limitation, but faster than the 3060 for models that fit. Just don't pay a premium for it over the 3060 when AI is your primary use case.

Used Prices and Value (February 2026)

Card	Used Price	Price per GB VRAM	Value Rating
RTX 3060 12GB	$180-230	**$15-19/GB**	Best value
RTX 3060 Ti 8GB	$220-280	$28-35/GB	Poor value
RTX 3070 8GB	$280-350	$35-44/GB	Fair (for speed)

The 3060 12GB has the best price-per-gigabyte of any GPU currently available for local AI. At $200, you get 12GB of VRAM — the same amount as an RTX 4060 that costs $350+ new.

Where to Buy

All three cards are widely available used since they were popular gaming GPUs:

- **eBay:** Largest selection, 30-day buyer protection. Best for safe purchases.
- **r/hardwareswap:** Typically $20-50 cheaper than eBay. Use PayPal Goods & Services only.
- **Facebook Marketplace:** Best for local pickup and inspection.

For detailed buying advice, see the used GPU buying guide.

The Real Competition: 3060 vs 3090

Before buying any of these three cards, consider the bigger picture.

Card	VRAM	Used Price	Largest Model (Q4)
RTX 3060 12GB	12GB	$180-230	14B
RTX 3090 24GB	24GB	$800-900	32B (70B with offload)

The used RTX 3090 costs 4x more but delivers:

- **2x the VRAM** (24GB vs 12GB)
- **2.6x the memory bandwidth** (936 vs 360 GB/s)
- **32B model capability** (Qwen 2.5 32B, the current sweet spot for local AI)
- **70B models** with some CPU offloading

If you can stretch your budget to $800, the 3090 is a different league. If $200-300 is your ceiling, the 3060 12GB is the best you can do — and it's genuinely good at that price.

The middle ground doesn't exist. Between the 3060 12GB at $200 and the 3090 at $800, there's nothing with significantly more VRAM that's worth buying. The RTX 4060 Ti 16GB ($400-450 new) has 16GB but you're paying 2x the price for 33% more VRAM with slightly better speed. The math doesn't work.

Recommendations

Buy the RTX 3060 12GB if:

- LLM inference is your primary use case
- You want to run 14B parameter models
- Budget is under $250
- You want the best value per dollar
- You're building a budget AI PC

Buy the RTX 3070 8GB if:

- Image generation speed is your priority
- You only need 7-8B LLMs (and you're sure about that)
- You found one under $300 (at $350 it's overpriced for AI)
- You also game and want better FPS

Skip the RTX 3060 Ti because:

- Same 8GB VRAM as the 3070 but slower
- More expensive than the 3060 with less VRAM
- Doesn't win any AI-relevant category
- Only worth it if priced at or below 3060 levels

The Bottom Line

The RTX 3060 12GB is the best budget GPU for local AI in 2026. Not because it's fast — the 3070 is 55% faster on identical models. Because it runs models the 3070 can't.

At $180-230 used, 12GB of VRAM gets you into the 14B model tier: Qwen 2.5 14B, Phi-4 14B, DeepSeek R1 Distill 14B. These are genuinely useful models for coding, writing, and reasoning. Being stuck at 7-8B models on an 8GB card means missing the biggest quality jump in the local AI model lineup.

The 3070 is the better card by every metric except the one that counts. Buy it for image generation or gaming. Buy the 3060 for AI.

📚 **Hardware guides:** Best GPU Under $300 · Used GPU Buying Guide · VRAM Requirements · Used RTX 3090 Guide

📚 **What can you run:** 8GB VRAM Guide · 12GB VRAM Guide · Budget AI PC Build

---

Source: https://insiderllm.com/guides/rtx-3060-vs-3060ti-vs-3070-local-ai/

Free guides for running AI locally