Best GPU Under $500 for Local AI (2026 Picks)

February 4, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** The RTX 4060 Ti 16GB ($450 new) is the best GPU under $500 for local AI if you want warranty and efficiency. The used RTX 3080 10GB ($350-400) offers better raw performance but only 10GB VRAM. For pure VRAM per dollar, a used RTX 3060 12GB ($170-220) still beats everything. Your choice depends on whether you value warranty, VRAM, or speed.

📚 **More on this topic:** Best GPU Under $300 · GPU Buying Guide · What Can You Run on 16GB VRAM · What Can You Run on 12GB VRAM

$500 is the sweet spot where local AI gets serious. You can run 14B-32B models at usable speeds, handle Stable Diffusion XL without compromise, and even squeeze some 70B models with offloading. The question is which card gives you the best balance of VRAM, speed, and reliability.

This guide compares every worthwhile option under $500 — new and used — with real benchmarks and clear recommendations.

The $500 GPU Landscape

At this budget, you have two paths:

1. **New cards with warranty** — RTX 4060 Ti 16GB, RX 7700 XT
2. **Used cards with more power** — RTX 3080, RTX 3060 12GB

The tradeoff: used cards offer more raw performance, but you lose warranty protection and take on reliability risk.

GPU	VRAM	New Price	Used Price	Best For
RTX 4060 Ti 16GB	16GB	$449-499	$380-420	Balanced new option
RTX 3080 10GB	10GB	N/A	$350-400	Raw speed on budget
RTX 3060 12GB	12GB	$280-330	$170-220	Best VRAM/dollar

GPU	VRAM	New Price	Used Price	Best For
RX 7700 XT	12GB	$400-450	$320-360	AMD alternative

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

RTX 4060 Ti 16GB — Best New Card Under $500

The RTX 4060 Ti 16GB is NVIDIA's answer to the "we need more VRAM" complaints. It's the same Ada Lovelace architecture as the 8GB version, but with double the memory.

Spec	RTX 4060 Ti 16GB
VRAM	16GB GDDR6
Memory Bandwidth	288 GB/s
CUDA Cores	4352
TDP	165W
New Price	$449-499

LLM Performance

Model	Speed	VRAM Usage
Llama 3.1 8B Q4	~55-65 tok/s	~5GB
Qwen 2.5 14B Q4	~28-35 tok/s	~9GB
Qwen 2.5 14B Q6	~25-30 tok/s	~11GB
DeepSeek R1 Distill 14B Q4	~28-32 tok/s	~9GB
Qwen 2.5 32B Q3	~12-15 tok/s	~15GB (tight)

The 16GB opens up the entire 14B tier at high quality (Q6-Q8) and can squeeze 32B models at aggressive quantization.

Image Generation

Task	Performance
SDXL 1024x1024	~5-6 sec
SD 1.5 512x512	~2-3 sec
Flux (NF4/FP8)	~30-40 sec

Pros

- 16GB VRAM — enough for 14B at Q8 and 32B at Q3
- Low power draw (165W) — works with most PSUs
- Modern architecture with excellent CUDA support
- 3-year warranty from NVIDIA/AIB
- Quiet and cool under load

Cons

- Low memory bandwidth (288 GB/s) hurts tok/s
- Expensive per GB of VRAM vs used options
- 32B models are a tight fit

Buy if: You want a new card with warranty, run 14B models primarily, and value efficiency over raw speed.

Used RTX 3080 10GB — Best Performance Under $500

The RTX 3080 10GB was the 2020 flagship. Used prices have dropped to $350-400, making it an incredible value for raw compute power.

Spec	RTX 3080 10GB
VRAM	10GB GDDR6X
Memory Bandwidth	760 GB/s
CUDA Cores	8704
TDP	320W

Spec	RTX 3080 10GB
Used Price	$350-400

LLM Performance

Model	Speed	VRAM Usage
Llama 3.1 8B Q4	~70-85 tok/s	~5GB
Qwen 2.5 14B Q4	~35-42 tok/s	~9GB (tight)
Mistral 7B Q4	~75-90 tok/s	~4.5GB
32B models	Won't fit	—

The 3080's higher bandwidth delivers 25-30% faster token generation than the 4060 Ti on models that fit. But 10GB limits you to 14B at Q4 with minimal context headroom.

Image Generation

Task	Performance
SDXL 1024x1024	~4-5 sec
SD 1.5 512x512	~2 sec
Flux (NF4)	~25-30 sec

Pros

- Excellent tok/s thanks to high bandwidth
- Great for 7B-8B models at maximum speed
- Handles SDXL and image gen well
- Much cheaper than equivalent new cards

Cons

- 10GB VRAM limits model options
- 320W TDP requires beefy PSU (750W+)
- No warranty (used)
- Runs hot and loud

Buy if: You primarily run 7B-8B models and want maximum speed, or you'll use it for image generation where 10GB is adequate.

RTX 3060 12GB — Best VRAM Per Dollar

The RTX 3060 12GB is still the budget champion. At $170-220 used, nothing matches its VRAM-per-dollar ratio.

Spec	RTX 3060 12GB
VRAM	12GB GDDR6
Memory Bandwidth	360 GB/s
CUDA Cores	3584
TDP	170W
Used Price	$170-220
New Price	$280-330

LLM Performance

Model	Speed	VRAM Usage
Llama 3.1 8B Q4	~38-42 tok/s	~5GB
Qwen 2.5 14B Q4	~18-22 tok/s	~9GB
Mistral 7B Q4	~40-45 tok/s	~4.5GB
DeepSeek R1 Distill 14B Q4	~18-22 tok/s	~9GB

Slower than the 3080 and 4060 Ti, but 12GB means you can run 14B models with comfortable context headroom.

Why It Still Wins on Value

Metric	RTX 3060 12GB	RTX 4060 Ti 16GB	RTX 3080 10GB
Price	$200	$450	$375
VRAM	12GB	16GB	10GB

Metric	RTX 3060 12GB	RTX 4060 Ti 16GB	RTX 3080 10GB
$/GB	$16.67	$28.12	$37.50

If budget is your primary constraint, the 3060 12GB delivers more usable VRAM per dollar than anything else.

Buy if: Budget is tight, you want 14B model capability, and you can tolerate slower speeds.

RX 7700 XT — AMD Alternative

The RX 7700 XT is AMD's answer to the RTX 4060 Ti. Same 12GB as the 3060, but with modern architecture and better bandwidth.

Spec	RX 7700 XT
VRAM	12GB GDDR6
Memory Bandwidth	432 GB/s
Stream Processors	3456
TDP	245W
New Price	$400-450

The ROCm Reality

AMD GPUs use ROCm instead of CUDA. Support has improved dramatically, but it's not seamless:

```
# May be needed for RX 7700 XT
HSA_OVERRIDE_GFX_VERSION=11.0.0 ollama serve
```

What works well:

- Ollama (mostly)
- llama.cpp with ROCm build
- PyTorch with ROCm

What's rough:

- LM Studio (limited support)
- Some GGUF quantizations
- Debugging when things break

LLM Performance (with ROCm)

Model	Speed
Llama 3.1 8B Q4	~45-55 tok/s
Qwen 2.5 14B Q4	~22-28 tok/s

Performance is solid when it works, but expect more setup hassle than NVIDIA.

Buy if: You're comfortable with Linux, want 12GB new with warranty, and don't mind troubleshooting ROCm.

Head-to-Head Comparison

GPU	VRAM	8B Speed	14B Q4	Context @ 14B	Price	Value
RTX 4060 Ti 16GB	16GB	60 tok/s	✓ (28 tok/s)	16K+	$450	Good
RTX 3080 10GB	10GB	80 tok/s	Tight	4K	$375	Good
RTX 3060 12GB	12GB	40 tok/s	✓ (20 tok/s)	8K	$200	Best
RX 7700 XT	12GB	50 tok/s	✓ (25 tok/s)	8K	$425	OK

Which GPU Should You Buy?

Buy the RTX 4060 Ti 16GB if:

- You want a **new card with warranty**
- You'll primarily run **14B models**
- You want to experiment with **32B at Q3**
- Your PSU is **under 650W**

- **Quiet operation** matters

Buy the used RTX 3080 if:

- **7B-8B models** are your focus
- You want **maximum tok/s** on smaller models
- You'll do **image generation** heavily
- You have a **750W+ PSU**
- You're comfortable with **used hardware risk**

Buy the RTX 3060 12GB if:

- **Budget is the priority** ($200 vs $450)
- You want **14B capability** cheaply
- Slower speeds are **acceptable**
- You'd rather save money for a **future upgrade**

Buy the RX 7700 XT if:

- You prefer **AMD** and use Linux
- You want **12GB new with warranty**
- You're comfortable with **ROCm troubleshooting**

The $500 Strategy Question

Here's the real decision: do you spend your full budget now, or save some for later?

Option A: Spend $450 on RTX 4060 Ti 16GB

- Gets you 14B+ capability today
- Warranty protection
- No upgrade path for a while

Option B: Spend $200 on RTX 3060 12GB, save $250

- Same 14B capability (slower)
- Money saved toward future RTX 3090 or 4080
- Can upgrade when prices drop

For many people, Option B makes more sense. The 3060 12GB handles everything the 4060 Ti can except 32B models (which barely fit anyway). Save the difference for when used RTX 3090s drop to $600 or RTX 4080s hit the used market.

Power Supply Requirements

GPU	Minimum PSU	Recommended PSU
RTX 4060 Ti 16GB	550W	650W 80+ Gold
RTX 3080 10GB	700W	850W 80+ Gold
RTX 3060 12GB	500W	550W 80+ Bronze
RX 7700 XT	600W	700W 80+ Gold

If your PSU is under 650W, factor in a $80-120 upgrade when considering the RTX 3080.

The Bottom Line

Best overall under $500: RTX 4060 Ti 16GB ($450) — 16GB VRAM, warranty, efficiency, and enough power for 14B models at good quality.

Best value: RTX 3060 12GB used ($170-220) — Same 14B capability for less than half the price. Save the difference for a future upgrade.

Best speed: RTX 3080 10GB used ($350-400) — Fastest tok/s in this range, but 10GB limits model options.

Skip: The RX 7700 XT unless you specifically need AMD. The software hassle isn't worth it when NVIDIA options exist at similar prices.

The decision framework: if you value **warranty and VRAM**, buy the 4060 Ti 16GB. If you value **budget and upgradability**, buy the 3060 12GB and bank the savings. If you value **speed on smaller models**, buy the used 3080.

Related Guides

- Best GPU Under $300 for Local AI
- GPU Buying Guide for Local AI
- What Can You Run on 16GB VRAM?
- What Can You Run on 12GB VRAM?
- RTX 3090 vs RTX 4070 Ti Super

Get notified when we publish new guides.

Subscribe — free, no spam
