Best GPU Under $300 for Local AI (2026 Picks)

February 4, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** The used RTX 3060 12GB ($170-220) is the best GPU under $300 for local AI. That 12GB of VRAM beats everything else at this price — the RX 7600 and RTX 4060 only have 8GB, which severely limits what models you can run. If you want new with warranty, the Intel Arc B580 ($250) offers competitive performance with 12GB VRAM but requires more setup on the software side.

📚 **More on this topic:** GPU Buying Guide · What Can You Run on 12GB VRAM · What Can You Run on 8GB VRAM · Budget AI PC Build

$300 is the sweet spot for budget local AI. You can get a GPU that runs 7B-14B language models at usable speeds and handles Stable Diffusion without painful waits. The catch: at this price, VRAM matters more than anything else, and most cards skimp on it.

This guide compares every worthwhile option under $300 — new and used — with real performance numbers and honest recommendations.

Why VRAM Matters More Than Speed at This Budget

Here's the uncomfortable truth about budget GPUs: a faster card with less VRAM loses to a slower card with more VRAM for AI workloads.

A 7B language model at Q4 quantization needs about 4.5GB of VRAM. A 13B model needs ~8GB. If your model doesn't fit, it either won't run or it falls back to CPU, dropping from 40 tok/s to 5 tok/s.

GPU	VRAM	What Fits
8GB cards	8GB	7B-8B models only
12GB cards	12GB	7B-14B models comfortably

That extra 4GB is the difference between running useful 14B models and being stuck at 7B forever. At under $300, prioritize VRAM over everything else.

→ Not sure what fits? Try our Planning Tool.

The Contenders

RTX 3060 12GB — Best Overall (Used: $170-220)

The RTX 3060 12GB is the budget AI champion. It's three generations old but still the best value for local AI work because of that 12GB of VRAM — more than the RTX 4060, 4070, and most cards under $500.

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:

- Llama 3.1 8B Q4: ~38-42 tok/s
- Qwen 2.5 14B Q4: ~18-22 tok/s
- Mistral 7B Q4: ~40-45 tok/s

Image Generation:

- SD 1.5 512x512: ~4-5 sec/image
- SDXL 1024x1024: ~12-15 sec/image
- Flux (NF4): ~60-80 sec/image

What you can run:

- 7B-8B models at Q4-Q8 with room for context
- 13B-14B models at Q4-Q5 (tight but works)
- SDXL comfortably, Flux with quantization

What you can't run:

- 32B+ models (need 24GB)

- 70B models at any quantization
- Flux at full FP16 precision

Why it wins: Nothing else under $300 offers 12GB of VRAM. The 3060 12GB lets you run models that simply won't load on 8GB cards.

Where to buy:

- Amazon (new): $280-330
- eBay (used): $170-220
- r/hardwareswap: $150-200

RX 7600 — Best New Card (New: $250-270)

AMD's RX 7600 is the best new GPU you can buy under $300. It's faster than the RTX 3060 in raw compute, but the 8GB VRAM is a significant limitation for AI work.

Spec	RX 7600
VRAM	8GB GDDR6
Memory Bandwidth	288 GB/s
Stream Processors	2048
TDP	165W
New Price	$250-270

LLM Performance (with ROCm):

- Llama 3.1 8B Q4: ~35-40 tok/s
- Qwen 2.5 7B Q4: ~38-42 tok/s
- 14B models: Won't fit at good quality

The ROCm Reality:

AMD GPUs use ROCm instead of CUDA. The good news: ROCm support has improved dramatically — Ollama, llama.cpp, and vLLM all work. The bad news: you may need to set `HSA_OVERRIDE_GFX_VERSION` for the 7600, and some tools still have rough edges.

```
# May be needed for RX 7600
HSA_OVERRIDE_GFX_VERSION=11.0.0 ollama serve
```

For details on AMD setup, see our AMD vs NVIDIA guide.

What you can run:

- 7B-8B models at Q4-Q6
- SDXL (tight fit)
- SD 1.5 comfortably

What you can't run:

- 13B+ models at good quality
- Flux (needs 12GB+ for practical use)
- Anything requiring long context on larger models

Why consider it: If you want a new card with warranty and primarily run 7B models, the RX 7600 is a solid choice. Just know you're trading VRAM for newness.

Intel Arc B580 — Dark Horse (New: $250)

Intel's Arc B580 is the surprise value pick. It has 12GB of VRAM — matching the RTX 3060 — and delivers competitive performance at $250 new.

Spec	Arc B580
VRAM	12GB GDDR6
Memory Bandwidth	456 GB/s
Xe Cores	20
TDP	190W
New Price	$250

LLM Performance (with IPEX-LLM/OpenVINO):

- Llama 3.1 8B Q4: ~55-65 tok/s
- Qwen 2.5 7B Q4: ~60-70 tok/s

The Catch:

Intel GPUs don't use CUDA or ROCm. You need Intel's IPEX-LLM or OpenVINO stack. This works well but requires more setup than NVIDIA:

- Ollama: Works via IPEX-LLM backend
- LM Studio: Not supported
- llama.cpp: Requires SYCL build

If you're comfortable with Linux and some tinkering, the B580 delivers excellent performance per dollar. If you want plug-and-play, stick with NVIDIA.

What you can run:

- 7B-14B models (same as RTX 3060)
- SDXL, potentially Flux
- Everything the 3060 can run

Why consider it: Same VRAM as the 3060, faster memory bandwidth, new with warranty, $250. The software ecosystem is the only downside.

RTX 4060 — Skip It (New: $290-310)

The RTX 4060 looks appealing — it's new, efficient, and has great CUDA support. But for AI work, it's a bad deal.

Spec	RTX 4060
VRAM	8GB GDDR6
Memory Bandwidth	272 GB/s
CUDA Cores	3072
TDP	115W
New Price	$290-310

The Problem:

8GB of VRAM in 2026 is not enough for serious local AI work. You're stuck at 7B models with limited context. The RTX 3060 12GB costs less used and runs larger models.

Skip unless: You specifically need a low-power card for a small form factor build and will only ever run 7B models.

RTX 3060 Ti — Also Skip (Used: $180-240)

The 3060 Ti is faster than the 3060 in gaming, but worse for AI because it only has 8GB of VRAM.

Spec	RTX 3060 Ti
VRAM	8GB GDDR6X
Memory Bandwidth	448 GB/s
CUDA Cores	4864
TDP	200W
Used Price	$180-240

Higher bandwidth and more CUDA cores don't help when your model doesn't fit. The 3060 12GB is the better AI card despite being "lower tier" in gaming.

RTX 3070 — Only If Cheap (Used: $220-280)

The RTX 3070 has 8GB of VRAM like the 3060 Ti. It's faster but has the same VRAM limitation.

Spec	RTX 3070
VRAM	8GB GDDR6
Memory Bandwidth	448 GB/s
CUDA Cores	5888
TDP	220W
Used Price	$220-280

Consider only if: You find one under $200 and will only run 7B models. Otherwise, the 3060 12GB is still the better choice.

Head-to-Head Comparison

GPU	VRAM	LLM Speed (8B)	Max Model (Q4)	Price	Best For
RTX 3060 12GB	12GB	~40 tok/s	14B	$170-220 used	**Best overall**
Arc B580	12GB	~60 tok/s	14B	$250 new	Tinkerers wanting new
RX 7600	8GB	~38 tok/s	8B	$260 new	AMD fans, 7B only
RTX 4060	8GB	~45 tok/s	8B	$300 new	Low power builds
RTX 3060 Ti	8GB	~50 tok/s	8B	$200 used	Skip for AI
RTX 3070	8GB	~55 tok/s	8B	$250 used	Skip for AI

Buying Used: What to Look For

Used GPUs are the best value for AI, but you need to buy smart.

Where to Buy

eBay — Safest option. 30-day Money Back Guarantee covers any problems.

- Filter for 99%+ seller feedback
- Look for actual photos of the card, not stock images
- Check that it's in the correct listing category (Consumer Electronics)

r/hardwareswap — Best prices, more risk.

- Only buy from users with confirmed trades
- Always use PayPal Goods & Services
- Get timestamps showing the card working

Facebook Marketplace — Good for local pickup.

- Inspect before paying
- Test if possible
- Meet in public places

Red Flags to Avoid

- Price significantly below market (scam)
- Stock photos instead of actual card photos
- New account with no history
- Seller pushing for PayPal Friends & Family or Venmo
- Multiple identical cards (mining operation)

After Purchase

1. Verify the card in GPU-Z (check VRAM, model, specs)
2. Run FurMark for 30+ minutes (thermal stress test)
3. Run an actual AI workload (Ollama with a 7B model)
4. Monitor temperatures — GPU should stay under 83°C, memory under 100°C

For complete buying guidance, see our Used GPU Buying Guide.

What You Can Actually Run Under $300

With a 12GB card (RTX 3060 or Arc B580):

Model	Size	Performance	Use Case
Llama 3.1 8B	4.5GB	38-42 tok/s	General assistant
Qwen 2.5 7B	4.5GB	40-45 tok/s	Coding, multilingual
Mistral 7B	4.5GB	40-45 tok/s	Fast chat
Qwen 2.5 14B Q4	8.5GB	18-22 tok/s	Better reasoning
DeepSeek R1 Distill 8B	4.5GB	35-40 tok/s	Math, reasoning
SDXL	~6.5GB	12-15 sec	Image generation

With an 8GB card (RX 7600, RTX 4060):

Model	Size	Performance	Use Case
Llama 3.1 8B Q4	4.5GB	35-45 tok/s	General assistant
Qwen 2.5 7B Q4	4.5GB	38-42 tok/s	Coding

Model	Size	Performance	Use Case
14B models	—	Won't fit	—
SDXL	~6.5GB	Tight fit	Limited headroom

The 12GB cards open up the entire 14B tier, which is a meaningful step up in quality for coding, reasoning, and complex tasks.

The Bottom Line

Buy the RTX 3060 12GB used ($170-220). Nothing else under $300 matches its combination of VRAM, CUDA compatibility, and proven reliability.

If you want new with warranty: The Intel Arc B580 ($250) offers 12GB VRAM and strong performance, but requires more software setup. Worth it if you're comfortable with Linux.

If you're AMD-only: The RX 7600 ($260) works, but 8GB limits you to 7B models. Only buy if you're committed to AMD.

Skip: RTX 4060, RTX 3060 Ti, RTX 3070 — all have only 8GB VRAM, which isn't enough for serious local AI work in 2026.

The hierarchy is simple: 12GB beats 8GB every time. Buy the 3060 12GB, install Ollama, and start running real models.

Related Guides

- GPU Buying Guide for Local AI
- What Can You Run on 12GB VRAM?
- What Can You Run on 8GB VRAM?
- Build a Local AI PC for Under $500
- Used GPU Buying Guide

Get notified when we publish new guides.

Subscribe — free, no spam

Source: https://insiderllm.com/guides/best-gpu-under-300-local-ai/

Free guides for running AI locally