Used Tesla P40 for Local AI: The $200 Budget Beast

February 23, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** The NVIDIA Tesla P40 ($150-$200 on eBay) gives you 24GB of VRAM — the cheapest way to fit 14B+ models entirely on GPU. It runs Qwen2.5 14B Q4 at ~16 tok/s and Llama 7B Q4 at ~41 tok/s via llama.cpp. The catch: Pascal architecture (2016), no FP16 acceleration, no tensor cores, 347 GB/s bandwidth (1/3 of an RTX 3090), passive cooling (needs aftermarket fan), no display output, and an 8-pin EPS power connector. At ~$7/GB of VRAM it's unbeatable on price. But an RTX 3060 12GB at the same price is faster for any model that fits in 12GB. The P40 only wins when you need 24GB on a budget.

📚 **Related:** Best Used GPUs for Local AI · Best GPU Under $300 · Used RTX 3090 Buying Guide · VRAM Requirements

The NVIDIA Tesla P40 was an inference accelerator released in 2016. Nine years later, it's the cheapest 24GB GPU you can buy — $150-$200 on eBay, sometimes less.

That 24GB of VRAM lets you run 14B models entirely on GPU that wouldn't fit on a 12GB RTX 3060. It's slow by modern standards — roughly 3x slower than an RTX 3090 — but it works, and the price-per-gigabyte of VRAM is unmatched.

This is a guide for people who want 24GB of VRAM for under $250, understand the tradeoffs, and want to set it up correctly.

Specifications

Spec	Tesla P40	RTX 3060 12GB	RTX 3090
Architecture	Pascal (2016)	Ampere (2020)	Ampere (2020)
CUDA Cores	3,840	3,584	10,496
Tensor Cores	None	112 (Gen 3)	328 (Gen 3)
VRAM	**24 GB GDDR5X**	12 GB GDDR6	24 GB GDDR6X
Memory Bandwidth	347 GB/s	360 GB/s	936 GB/s

Spec	Tesla P40	RTX 3060 12GB	RTX 3090
FP16	Crippled (1/64th)	Full speed	Full speed
TDP	250W	170W	350W
Cooling	Passive (no fan)	Active	Active
Display Output	**None**	HDMI + DP	HDMI + DP
Used Price	**$150-$200**	$170-$200	$800-$1,000

The P40's FP16 runs at 1/64th the rate of FP32 — NVIDIA deliberately disabled it to differentiate from the P100. No tensor cores either. LLM inference through llama.cpp uses integer quantization (Q4, Q8), so the lack of FP16/tensor cores matters less than you'd think. The real bottleneck is memory bandwidth: 347 GB/s is roughly 1/3 of the RTX 3090.

Benchmarks

LocalScore.ai Results

Model	Params	Prompt (t/s)	Generation (t/s)	LocalScore
LLaMA 2 7B Q4_0	6.7B	833	**40.9**	282
Qwen2.5 14B Q4_K_M	14.8B	339	**15.7**	115
Qwen3-Coder-30B-A3B Q4_K_M	30.5B (MoE)	288	**29.3**	128

Direct GPU Comparison

Model	Tesla P40	RTX 3060 12GB	RTX 3090
LLaMA 7B Q4	~41 t/s	~50 t/s	~112 t/s
Qwen2.5 14B Q4_K_M	**~16 t/s**	Can't fit (12GB)	~56 t/s
30B MoE Q4	**~29 t/s**	Can't fit (12GB)	Higher

The P40 is slower than the RTX 3060 on everything that fits in 12GB. The P40 only wins when you need models that require more than 12GB of VRAM — which is precisely why you buy it.

At 16 tok/s on a 14B model, that's about 4 words per second. Comfortable for chat. Not fast, but usable.

Setup

Hardware Requirements

Power cable: The P40 uses an **8-pin EPS connector** (CPU-style), not standard PCIe power. You need a 2x PCIe 8-pin to 1x 8-pin EPS adapter ($8-15 on Amazon/eBay). Do not force a PCIe cable into the socket.

PSU: 750W+ recommended. The P40 draws up to 250W.

BIOS: Enable **"Above 4G Decoding"** in your motherboard BIOS. Without this, the P40 won't be detected.

Display: The P40 has no video outputs. You need either integrated graphics (Intel iGPU) or a cheap secondary GPU (GT 710, ~$30) for display.

Cooling: See the next section — this is mandatory.

Cooling Solutions

The P40 is passively cooled, designed for server chassis with 60+ CFM front-to-back airflow. In a desktop case, it **will** thermal throttle without aftermarket cooling. Users report temps hitting 85C and performance dropping from the expected ~16 t/s to 3-4 t/s.

Option 1: 3D-printed blower shroud (~$25-35 on Amazon, includes fan and mounting hardware). Mounts a 97x33mm blower fan directly to the heatsink. Users report temps dropping from 85C to 51C under load.

Option 2: 120mm fan duct. 3D-print or tape a duct channeling a case fan into the P40's heatsink fins. One user went from thermal-throttled 3.6 t/s to a consistent 10 t/s with a sealed 120mm fan tunnel.

Option 3: Strong case airflow. Only works in rack-mount or server-style cases with high-RPM front-to-back fans.

Monitor temps: `nvidia-smi -l 1` during inference. If GPU temp exceeds 80C, you're throttling.

Driver Installation (Ubuntu)

```
sudo apt update
sudo apt install nvidia-driver-535
sudo reboot
nvidia-smi
```

The P40 (compute capability 6.1) supports CUDA 12.x with driver 535+.

Ollama

```
curl -fsSL https://ollama.com/install.sh | sh
ollama pull qwen2.5:14b-instruct-q4_K_M
ollama run qwen2.5:14b-instruct-q4_K_M
```

Ollama auto-detects the P40 via nvidia-smi. No special configuration needed.

llama.cpp with CUDA

```
git clone https://github.com/ggml-org/llama.cpp
cd llama.cpp
cmake -B build -DGGML_CUDA=ON -DCMAKE_CUDA_ARCHITECTURES=61
cmake --build build --config Release -j$(nproc)
```

The key flag is `-DCMAKE_CUDA_ARCHITECTURES=61` for Pascal. If pairing with a newer GPU, specify both: `"61;86"` for P40 + Ampere.

Performance tip: Set `GGML_CUDA_FORCE_MMQ=1` before running. This forces matrix-multiply quantized kernels optimized for non-tensor-core GPUs:

```
export GGML_CUDA_FORCE_MMQ=1
./build/bin/llama-cli -m model.gguf -ngl 99 -c 4096
```

What You Can Run

Model	Quant	VRAM	Speed	Verdict
Llama 3.2 3B	Q4	~3 GB	~48 t/s	Smooth, but small
Qwen 2.5 7B	Q8_0	~9 GB	~25 t/s	Good quality, comfortable speed
Qwen 2.5 14B	Q4_K_M	~11 GB	~16 t/s	The sweet spot — this is why you buy a P40
Qwen 2.5-Coder 14B	Q4_K_M	~11 GB	~17 t/s	Code assistance at 24GB budget
Qwen3-Coder 30B-A3B	Q4_K_M	~19 GB	~29 t/s	MoE — fast because only 3B active
32B model	Q4	~20 GB	~10 t/s	Tight fit, minimal context

The P40's sweet spot is 14B Q4_K_M models — they fit comfortably in 24GB with room for context, run at a usable speed, and deliver quality you can't get on a 12GB card.

→ Check what fits your hardware with our Planning Tool.

Practical Limitations

1. **No display output.** Need a separate GPU or iGPU for your monitor.
2. **No fast FP16.** HuggingFace Transformers pipelines that default to FP16 will be catastrophically slow. Use quantized inference (llama.cpp, Ollama) only.
3. **No tensor cores.** No hardware acceleration for mixed-precision matrix multiply.
4. **347 GB/s bandwidth.** The real performance limiter. 1/3 of the RTX 3090. You'll never exceed ~40-50 t/s on a 7B model.
5. **250W TDP.** As much power as a modern RTX 4070 Ti for a fraction of the performance.
6. **Passive cooling.** Budget $25-35 for a blower kit. Non-negotiable in a desktop case.
7. **8-pin EPS connector.** Need a specific adapter cable.
8. **Age.** Released 2016. Driver support for compute capability 6.1 is still included in CUDA 12.x, but could end in the next few years.

Decision Matrix

Buy the P40 if:

- You need 24GB VRAM and can't afford an RTX 3090
- You want to run 14B-30B quantized models fully on GPU
- You're building a headless inference server
- Total budget is under $250 (card + cooler + adapter)

Buy the RTX 3060 12GB instead if:

- You only run 7B models or smaller
- You want display output and active cooling out of the box
- You want to game or run Stable Diffusion
- Same ~$170-200 price range

Save up for the RTX 3090 if:

- Speed matters (56 t/s vs 16 t/s on 14B models)
- You want 24GB with 3x the performance
- You can afford $800-$1,000
- You want a general-purpose GPU (gaming, training, inference)

Consider two P40s if:

- You need 48GB total on an extreme budget (~$400)
- Building a multi-GPU inference server
- You don't mind 10 t/s on 32B models
- llama.cpp supports `--tensor-split 0.5,0.5` for dual-GPU

The Total Build

Component	Price
Tesla P40 (eBay)	$150-$200
Blower fan cooling kit	$25-$35

Component	Price
8-pin EPS power adapter	$8-$15
GT 710 for display (if no iGPU)	$25-$30
Total	**$208-$280**

For under $280, you get 24GB of VRAM running 14B models at conversational speed. That's the cheapest path to running models that actually know things, rather than the toy-sized 3B models that fit on budget consumer cards.

Bottom Line

The Tesla P40 is the cheapest 24GB GPU on the market at ~$7 per GB of VRAM. It lets you run models that 12GB cards can't touch. But it's slow, loud (once you add cooling), power-hungry, has no display output, and requires aftermarket modifications.

It's a budget tool for a specific job: getting 24GB of VRAM into a headless inference server for under $250. For that job, nothing else comes close on price. For everything else — gaming, training, image generation, or if speed matters — save up for the RTX 3090.

The P40 is the Honda Civic of local AI GPUs. It gets you there. It just won't be exciting.

Source: https://insiderllm.com/guides/used-tesla-p40-local-ai/