How Much Does It Cost to Run LLMs Locally?

February 4, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** A capable local AI setup costs $200-800 for hardware (one-time) plus $5-15/month in electricity. If you use AI moderately (1-2 hours daily), local pays for itself in 3-6 months vs ChatGPT Plus. Heavy users break even faster. The real math: a $750 RTX 3090 setup replaces $240/year of ChatGPT Plus subscriptions forever.

📚 **More on this topic:** Budget AI PC Under $500 · GPU Buying Guide · Local LLMs vs ChatGPT · Best Used GPUs 2026 · Planning Tool

Running LLMs locally has real costs — hardware, electricity, and your time. But so do cloud APIs and subscriptions. The question is: when does local AI actually save money?

This guide breaks down every cost involved and shows you exactly when running your own models beats paying for cloud services.

The Cost Categories

One-Time Costs (Hardware)

- GPU
- CPU (if upgrading)
- RAM (if upgrading)
- Power supply (if upgrading)
- Storage (if upgrading)

Ongoing Costs (Electricity)

- GPU power draw during inference
- System idle power
- Cooling (included in GPU draw)

Hidden Costs

- Your time setting up
- Troubleshooting
- Model quality differences

Hardware Costs: What You Actually Need

Budget Setup ($200-350)

Component	Cost	What It Gets You
RTX 3060 12GB (used)	$180	7B-14B models
16GB DDR4 RAM	$30	Enough for most use
Total	**$210**	

If your existing PC has a 500W+ PSU and a modern CPU (anything from last 8 years), add a used RTX 3060 12GB and you're running local AI.

What you can run:

- Llama 3.1 8B at ~40 tok/s
- Qwen 2.5 14B at ~20 tok/s
- Stable Diffusion XL
- Whisper speech-to-text

Mid-Range Setup ($500-800)

Component	Cost	What It Gets You
RTX 3090 (used)	$750	24GB VRAM, 32B models
850W PSU (if needed)	$100	Powers the 3090
32GB DDR4 RAM	$60	Headroom for offloading
Total	**$750-910**	

The RTX 3090 setup handles everything except the largest 70B models at good quality.

What you can run:

- Qwen 2.5 32B at ~40 tok/s
- DeepSeek R1 Distill 32B at ~35 tok/s
- Llama 3.3 70B at ~15 tok/s (with offloading)
- Flux image generation at full quality

High-End Setup ($1,600-2,000)

Component	Cost	What It Gets You
RTX 4090	$1,600	Fastest 24GB option
1000W PSU	$150	Powers the 4090
64GB DDR5 RAM	$150	Maximum offloading
Total	**$1,900**	

The RTX 4090 runs everything the 3090 can, but 30-50% faster.

Complete New Build ($500)

If you need an entire computer, see our Budget AI PC Under $500 guide:

Component	Cost
Used office PC (i5/i7)	$150
RTX 3060 12GB (used)	$180
32GB RAM upgrade	$60
650W PSU	$60
500GB SSD	$50
Total	**$500**

Electricity Costs: The Ongoing Expense

Power Draw by GPU

GPU	Inference Draw	Idle Draw
RTX 3060 12GB	140-170W	15-25W
RTX 3080	280-320W	20-30W
RTX 3090	300-350W	25-35W
RTX 4060 Ti	140-165W	10-15W
RTX 4090	350-450W	20-30W

System overhead: Add ~100W for CPU, RAM, drives, etc.

Monthly Cost Calculation

Formula:

```
Monthly cost = (GPU watts + system watts) × hours/day × 30 days × electricity rate ÷ 1000
```

Example: RTX 3090, 3 hours daily use, $0.15/kWh

```
(350W + 100W) × 3 hours × 30 days × $0.15 ÷ 1000 = $6.08/month
```

Monthly Cost Table

GPU	1 hr/day	3 hrs/day	8 hrs/day
RTX 3060 12GB	$1.22	$3.65	$9.72
RTX 3080	$1.89	$5.67	$15.12
RTX 3090	$2.03	$6.08	$16.20
RTX 4060 Ti	$1.17	$3.51	$9.36
RTX 4090	$2.48	$7.43	$19.80

Based on $0.15/kWh US average. Adjust for your local rate.

Real-World Usage Patterns

Light use (research, occasional questions): 1 hour/day

- RTX 3060: ~$1/month
- RTX 3090: ~$2/month

Moderate use (daily coding assistant): 3 hours/day

- RTX 3060: ~$4/month
- RTX 3090: ~$6/month

Heavy use (all-day development, content creation): 8 hours/day

- RTX 3060: ~$10/month
- RTX 3090: ~$16/month

Cloud API Costs: What You're Comparing Against

Subscription Services

Service	Monthly Cost	What You Get
ChatGPT Plus	$20/month	GPT-4, DALL-E, limited usage
Claude Pro	$20/month	Claude 3.5/4, limited usage
Gemini Advanced	$20/month	Gemini Pro/Ultra
GitHub Copilot	$10/month	Code completion

Pay-Per-Token APIs

Model	Input Cost	Output Cost
GPT-4o	$2.50/1M tokens	$10.00/1M tokens
GPT-4o mini	$0.15/1M tokens	$0.60/1M tokens
Claude 3.5 Sonnet	$3.00/1M tokens	$15.00/1M tokens
Claude 3.5 Haiku	$0.25/1M tokens	$1.25/1M tokens

What Does Usage Actually Cost?

Average conversation: ~2,000 tokens total (input + output)

API Model	Cost per Chat	100 Chats	1000 Chats
GPT-4o	$0.025	$2.50	$25.00
Claude 3.5 Sonnet	$0.036	$3.60	$36.00
GPT-4o mini	$0.0015	$0.15	$1.50

Coding session: ~50,000 tokens (lots of code context)

API Model	Cost per Session	20 Sessions/Month
GPT-4o	$0.63	$12.50
Claude 3.5 Sonnet	$0.90	$18.00
GPT-4o mini	$0.04	$0.75

Break-Even Analysis: When Local Pays Off

vs ChatGPT Plus ($20/month)

Hardware Cost	Monthly Electric	Break-Even
$200 (3060)	$4	13 months
$500 (3060 + build)	$4	31 months
$750 (3090)	$6	54 months

But consider:

- ChatGPT Plus has usage limits
- Local has no limits — run 24/7 if you want
- Local works offline
- No data leaves your machine

vs OpenAI API (Heavy User)

A developer using 2M tokens/day on GPT-4o:

- API cost: ~$25/day = $750/month

Break-even for a $750 RTX 3090 setup: **1 month**

Even using cheaper GPT-4o mini at 2M tokens/day:

- API cost: ~$1.50/day = $45/month

Break-even: **~17 months**

vs Claude Pro ($20/month)

Same math as ChatGPT Plus. The question is whether local models meet your quality needs.

Local models that compete:

- Qwen 2.5 32B — comparable to GPT-4o mini for many tasks
- DeepSeek R1 Distill 32B — strong reasoning
- Llama 3.3 70B — approaches GPT-4 quality

The Real Cost Comparison

Total Cost of Ownership (3 Years)

Option	Year 1	Year 2	Year 3	Total
ChatGPT Plus	$240	$240	$240	$720
RTX 3060 setup	$260	$48	$48	$356
RTX 3090 setup	$822	$72	$72	$966

Assumes moderate use (3 hrs/day)

Break-Even Chart

```
ChatGPT Plus cumulative cost:
Month 1:  $20   Month 6:  $120   Month 12: $240
Month 18: $360  Month 24: $480   Month 36: $720
```

```
RTX 3060 setup ($200 + $4/month):
Month 1:  $204  Month 6:  $224   Month 12: $248
Month 18: $272  Month 24: $296   Month 36: $344

Break-even: ~13 months
Savings after 3 years: $376
```

Hidden Costs: What the Math Doesn't Show

Setup Time

First-time setup: 2-8 hours

- Installing drivers, CUDA, Ollama
- Downloading models
- Troubleshooting if something breaks

Value of your time: If you bill $50/hour, 4 hours of setup = $200

Quality Differences

Local models in 2026 are good, but not always GPT-4/Claude-level:

Task	Local 32B	GPT-4o
Simple questions	Equivalent	Equivalent
Coding (common)	90% as good	Reference
Coding (obscure)	70% as good	Reference
Complex reasoning	80% as good	Reference
Creative writing	Equivalent	Equivalent

For most tasks, local models are good enough. For cutting-edge capabilities, you may still need cloud APIs occasionally.

Reliability

- Cloud APIs: 99.9% uptime, always updated
- Local: Your hardware, your problem

Budget $0-50/year for potential repairs or troubleshooting.

When Local AI Makes Financial Sense

Strong Yes

- **Heavy daily use** — 4+ hours/day of LLM interaction
- **API-heavy development** — would spend $50+/month on tokens
- **Privacy requirements** — data can't leave your machine
- **Offline needs** — travel, unreliable internet
- **Already have hardware** — gaming PC with decent GPU

Maybe

- **Moderate use** — 1-2 hours/day (break-even takes longer)
- **Need cutting-edge models** — may still need cloud API access
- **Tight budget** — upfront cost is significant

Probably Not

- **Light occasional use** — free tiers of ChatGPT/Claude are enough
- **Always need latest models** — local lags behind cloud
- **Zero technical tolerance** — setup requires some effort

Cost-Optimized Builds

The $200 AI Upgrade

Already have a PC with 500W+ PSU:

Item	Cost
RTX 3060 12GB (used)	$180
16GB RAM (if needed)	$30
Total	**$180-210**

Break-even vs ChatGPT Plus: ~12 months **Runs:** 7B-14B models, Stable Diffusion

The $500 Complete Build

Need an entire computer:

Item	Cost
Used office PC (Dell Optiplex, HP Z-series)	$150
RTX 3060 12GB (used)	$180
650W PSU	$60
32GB RAM	$60
500GB NVMe SSD	$50
Total	**$500**

Break-even vs ChatGPT Plus: ~31 months **Runs:** 7B-14B models, Stable Diffusion

The $800 Power Build

Want to run 32B models:

Item	Cost
RTX 3090 (used)	$750
850W PSU	$100
Total	**$850**

Break-even vs ChatGPT Plus: ~54 months **Runs:** Everything up to 70B with offloading

---

The Bottom Line

Local AI costs:

- One-time: $200-900 for hardware
- Ongoing: $4-16/month for electricity

Cloud AI costs:

- Ongoing: $20/month (subscriptions) or variable (API)

Break-even:

- Budget setup: 12-18 months vs ChatGPT Plus
- Mid-range setup: 36-54 months vs ChatGPT Plus
- Heavy API user: 1-6 months

The real value of local:

- No usage limits
- Complete privacy
- Works offline
- One-time cost, unlimited use

For moderate to heavy users, local AI pays for itself within 1-3 years and keeps saving money forever after. The upfront investment is real, but so are the long-term savings.

Start with a used RTX 3060 12GB for $180. Run it for a year. If you use it daily, you've already saved money. If local AI becomes essential to your workflow, upgrade to a RTX 3090 and never pay for cloud again.

Related Guides

- Budget AI PC Under $500
- GPU Buying Guide for Local AI
- Local LLMs vs ChatGPT
- Best Used GPUs for Local AI 2026
- What Can You Run on 12GB VRAM?
- Local AI Planning Tool — VRAM Calculator

Get notified when we publish new guides.

Subscribe — free, no spam

Source: https://insiderllm.com/guides/cost-to-run-llms-locally/