Best Mini PCs for Local AI Under $300 in 2026

February 21, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** The Lenovo ThinkCentre M90q Gen 3 ($200-$250 refurbished) is the best mini PC for local AI under $300. It has a 12th-gen Intel i5-12500T and DDR5 memory, which matters more than CPU generation for LLM inference. With 32GB of DDR5 ($40-$50 upgrade), it runs 7B models at 5-8 tok/s and 3B models at 15-20 tok/s via Ollama. That's genuinely usable for chat, summarization, and light coding help. If you can stretch to $350, the MinisForum UM790 Pro with a Ryzen 9 7940HS and 32GB DDR5-5600 is the real sweet spot — 10-15 tok/s on 7B models with a Radeon 780M iGPU that can accelerate prompt processing. For always-on home AI, either machine draws 30-60W under load and costs $25-$40/year in electricity.

📚 **More on this topic:** Budget AI PC Under $500 · CPU-Only LLMs: What Actually Works · Best Models Under 3B · Run Your First Local LLM · Planning Tool

Mini PCs have gotten surprisingly capable for local AI. Not fast — nobody's confusing a $200 ThinkCentre with an RTX 3090 build — but usable. A 7B model at 8 tok/s is enough for a home AI assistant that answers questions, summarizes documents, and helps with code. All on a machine that sits silently on your desk and draws less power than a light bulb.

The catch: most buying guides recommend $400+ AMD Ryzen boxes. Those are great, but if your budget is actually $300 or less, the options are different. Here's what works, what doesn't, and what's worth stretching for.

What to Look For

Not all mini PC specs matter equally for local AI. Here's what actually drives performance, in order of importance:

RAM amount: 32GB minimum. Models load entirely into system memory for CPU inference. A 7B Q4 model needs ~5GB. A 14B Q4 needs ~9GB. The OS and Ollama need another 4-6GB. With 16GB, you're running 7B models with no headroom. With 32GB, you can run 14B models and keep browser windows open.

RAM type: DDR5 > DDR4, and it's not close. LLM token generation is memory-bandwidth-bound. DDR5-4800 delivers roughly 50% more bandwidth than DDR4-3200. In real benchmarks, that

translates to 40-60% faster inference on the same CPU. A 12th-gen Intel with DDR5 outperforms a 13th-gen Intel with DDR4 for LLM work.

CPU: Modern AMD Ryzen (7000/8000 series) or Intel (12th gen+). You need AVX2 support at minimum — every CPU from the last decade has this. AVX-512 (found in some Intel 12th/13th gen chips and Zen 4 AMD) gives a measurable boost with llama.cpp.

iGPU: AMD's Radeon 780M (in Ryzen 7840HS/8845HS) can accelerate prompt processing via Vulkan. Intel's integrated graphics are less useful for inference. Don't buy based on iGPU alone, but it's a nice bonus on AMD.

Storage: NVMe SSD, at least 512GB. Models are 4-15GB each. SATA SSDs work but load models slower.

The Picks

Best Under $200: Refurbished Business Mini PCs

These are former corporate machines that flood the refurbished market. They're built to last, well-cooled for their size, and absurdly cheap.

Model	CPU	RAM Type	Typical Refurb Price	Best For
Lenovo ThinkCentre M90q Gen 3	i5-12500T (12th gen)	**DDR5-4800**	**$170-$250**	Best value — DDR5 matters
HP ProDesk 400 G6 Mini	i5-10500T (10th gen)	DDR4-3200	$90-$170	Cheapest entry point
Dell OptiPlex 7080 Micro	i5-10500T (10th gen)	DDR4-2933	$130-$180	Good if ThinkCentre unavailable

The winner: Lenovo ThinkCentre M90q Gen 3. It's the only sub-$250 refurb with 12th-gen Intel and DDR5. That combination puts it in a different performance class from the 10th-gen DDR4 alternatives. Buy one with 16GB and a 256GB SSD for ~$200, then add a $40-$50 32GB DDR5 SO-DIMM kit and a larger NVMe if needed. Total: ~$250 for a machine that meaningfully outperforms the HP and Dell options.

The HP ProDesk 400 G6 at $100-$150 is the absolute floor for "can run a local LLM." It works. But the 10th-gen CPU and DDR4 memory roughly halve your tok/s compared to the M90q Gen 3.

If the difference between $150 and $250 matters, the HP gets you started. If you can swing the extra $100, the ThinkCentre is worth every penny.

All three support up to 64GB RAM across two SO-DIMM slots. All use standard NVMe SSDs.

What to expect: 7B Q4 at 5-8 tok/s (M90q Gen 3) or 3-5 tok/s (10th-gen DDR4 machines). 3B models at 15-20 tok/s. 14B models at 3-5 tok/s — usable for batch tasks, not great for interactive chat.

Best at $300: Beelink EQR6

Spec	Detail
CPU	AMD Ryzen 5 6600H (6C/12T, Zen 3+)
iGPU	Radeon 660M (6 CUs, RDNA 2)
RAM	16GB LPDDR5 (soldered)
Storage	500GB NVMe SSD
Price	~$290 (Amazon)

The Beelink EQR6 is the cheapest AMD Ryzen mini PC that's actually available under $300. The Ryzen 5 6600H is a competent chip with LPDDR5 bandwidth, and $290 for a complete system is hard to argue with.

The catch: 16GB of LPDDR5, soldered. You can't upgrade the RAM. For 7B models, 16GB is tight but workable. For 14B models, forget it — you'll swap to disk and performance will crater. The Radeon 660M has only 6 CUs (half the 780M), so iGPU acceleration is minimal.

What to expect: 7B Q4 at 8-12 tok/s. Good for a dedicated chat assistant or RAG system with small models. Don't plan on running anything bigger.

Buy this if you want AMD + LPDDR5 speed at the lowest possible price and 7B models cover your needs. **Skip this if** you want to run 14B models — get the M90q Gen 3 with 32GB instead.

Worth the Stretch: MinisForum UM790 Pro ($350)

Spec	Detail
CPU	AMD Ryzen 9 7940HS (8C/16T, Zen 4)
iGPU	Radeon 780M (12 CUs, RDNA 3)
RAM	32GB DDR5-5600 (dual-channel, upgradeable)

Spec	Detail
Storage	1TB NVMe PCIe 4.0
Price	~$350 (MinisForum store on sale)

This is $50 over budget. It's also the machine I'd actually recommend for most readers.

The Ryzen 9 7940HS is the same silicon as the 7840HS (same die, higher binned). DDR5-5600 dual-channel delivers ~89 GB/s bandwidth — the highest in this roundup. The Radeon 780M with 12 RDNA 3 compute units can accelerate prompt processing via Vulkan, cutting time-to-first-token significantly. And the RAM is socketed SO-DIMM, upgradeable to 64GB if you ever need it.

What to expect: 7B Q4 at 10-15 tok/s. 14B Q4 at 5-8 tok/s. 3B models at 25-35 tok/s. Prompt processing with iGPU offloading: 50-76 tok/s (vs ~30 CPU-only). The 7B performance is genuinely comfortable for interactive chat — not instant, but you're reading as fast as it generates.

At $350, this is the same price as a refurbished ThinkCentre plus RAM upgrades plus a coffee. The performance difference is 2-3x. If your budget has any flex at all, this is the buy.

Alternatives in this range: Beelink SER7 ($380-$450, same Ryzen 7840HS chip), GMKtec NucBox K8 Plus ($399, Ryzen 8845HS).

Performance Reality Check

Time to be honest about what these machines can and can't do. CPU inference on a mini PC is not a GPU experience.

Model	Refurb Intel DDR4	M90q Gen 3 (DDR5)	AMD Ryzen DDR5	RTX 3060 12GB (for reference)
3B Q4	~8-12 tok/s	~15-20 tok/s	~25-35 tok/s	~80+ tok/s
7B Q4	~3-5 tok/s	~5-8 tok/s	~10-15 tok/s	~40+ tok/s
14B Q4	~2-3 tok/s	~3-5 tok/s	~5-8 tok/s	~22+ tok/s
32B Q4	Don't bother	Don't bother	~2-3 tok/s (64GB)	Won't fit

A few things to take away from this:

7B models are the ceiling for sub-$300 hardware. They're genuinely usable at 5-15 tok/s depending on your machine. That's slower than ChatGPT but faster than you type. For a home assistant that runs 24/7, answers questions, and summarizes documents, it works.

14B models technically run but test your patience. At 5-8 tok/s on the best AMD box, it's usable for batch processing — summarize this document, generate this code — but not great for back-and-forth chat. On the Intel refurbs, it's painful.

32B+ models: no. Don't try to run them on 32GB. Even if they fit (barely), the speed will be under 3 tok/s. That's a 20-second wait for a single paragraph. Life is too short.

The right models for this hardware: Qwen3 8B, Llama 3.2 8B, Phi-4 Mini (3.8B), and Gemma 2B/3B. These are purpose-built to be useful at small sizes. Qwen3 8B on a Ryzen mini PC is a surprisingly capable coding assistant and general-purpose chatbot.

Optimization Tips

A few settings that make a real difference on constrained hardware:

Use Q4_K_M quantization. Best balance of speed and quality. Q5 and Q6 are marginally better quality but measurably slower. Q3 and below get noticeably dumber. Quantization guide.

Set thread count to physical cores. `OLLAMA_NUM_THREADS=6` for a 6-core i5, `OLLAMA_NUM_THREADS=8` for a Ryzen 8-core. Hyperthreads don't help and can hurt due to cache thrashing.

Close everything else. These machines don't have memory to spare. A browser with 10 tabs eats 2-4GB. When running a 7B model on 16GB, that matters.

Use Ollama. It's the easiest way to get started and handles quantized GGUF models well on CPU. For the AMD Ryzen machines, LM Studio can sometimes leverage the iGPU more effectively.

Run smaller specialist models. Instead of one 14B generalist, run a 3B model for quick chat and an 8B coding model for code. The 3B model at 25+ tok/s will feel responsive, and you switch models when the task demands it.

The Distributed Play

Mini PCs aren't just standalone inference boxes. They're ideal nodes in a distributed AI setup.

A $150 refurbished ThinkCentre as a coordinator node. A $350 Ryzen mini PC for inference. A Raspberry Pi 5 for sensor input or intent classification on a tiny model. Total: under $600 for a multi-node system where each machine plays to its strengths.

The Raspberry Pi 5 (8GB, ~$80) runs 1-3B models at 8-12 tok/s — too slow for primary inference, but useful for intent routing, embedding generation, or as an always-on coordinator that farms heavy work to a beefier node. At 4W idle, it costs essentially nothing to run 24/7.

If you're interested in this architecture, start with our distributed thinking network guide. The key insight: you don't need every node to be powerful. You need the right model on the right hardware for the right task.

Power and Noise: The Mini PC Advantage

This is where mini PCs genuinely excel over any desktop build.

Machine	Idle	AI Workload	Annual Cost (24/7)
Refurb Intel mini PC	8-14W	45-70W	$14-$61
AMD Ryzen mini PC	6-10W	40-80W	$10-$70
Raspberry Pi 5	3-4W	8-12W	$5-$11
Desktop + RTX 3060	60-80W	200-250W	$105-$219

At US average electricity rates ($0.18/kWh). A Ryzen mini PC running local AI 8 hours a day costs about $25-$40/year in electricity. The same workload on a desktop with a discrete GPU costs 3-5x more.

And noise? Zero. These machines are near-silent under load — small fans at low RPM, no GPU cooler whine. You can run one in a bedroom as an always-on AI assistant without hearing it.

What to Skip

Used thin clients ($30-$80): HP t630, Dell Wyse 5070, etc. They're tempting at $50, but they have 4-8GB of soldered RAM and embedded CPUs that can barely run a 3B model. Not worth the frustration unless you're building a distributed network and just need a coordinator.

Gaming mini PCs ($600+): Machines like the Beelink SER9 ($999) with Ryzen AI 9 and Radeon 890M are impressive but far above budget. Wait for prices to drop — today's $400 Ryzen 7840HS box was $600 a year ago.

Anything with less than 16GB non-upgradeable RAM: You'll regret it within a month. Models keep getting more capable at small sizes, but they still need memory to run.

The Verdict

Budget	Buy This	What You Get
$100-$150	HP ProDesk 400 G6 (refurb)	3B-7B models at 3-5 tok/s. The floor.
$200-$250	Lenovo M90q Gen 3 + 32GB DDR5	7B models at 5-8 tok/s. Best under $300.
~$290	Beelink EQR6	7B at 8-12 tok/s. AMD speed, 16GB limit.
~$350	MinisForum UM790 Pro	7B at 10-15 tok/s. The actual sweet spot.

For most readers, the Lenovo M90q Gen 3 at $200-$250 is the right buy under $300 — refurbished, DDR5, upgradeable to 64GB, and fast enough for 7B models that are genuinely useful. If you can stretch to $350, the MinisForum UM790 Pro with its Ryzen 9 7940HS doubles the performance and adds an iGPU that helps with prompt processing.

These aren't GPU-accelerated builds. They won't replace a dedicated AI rig. But for an always-on local AI assistant that sits silently on your desk, answers questions privately, costs $25/year to run, and never sends your data to the cloud — a $200-$350 mini PC is hard to beat.

Source: https://insiderllm.com/guides/best-mini-pcs-local-ai-2026/