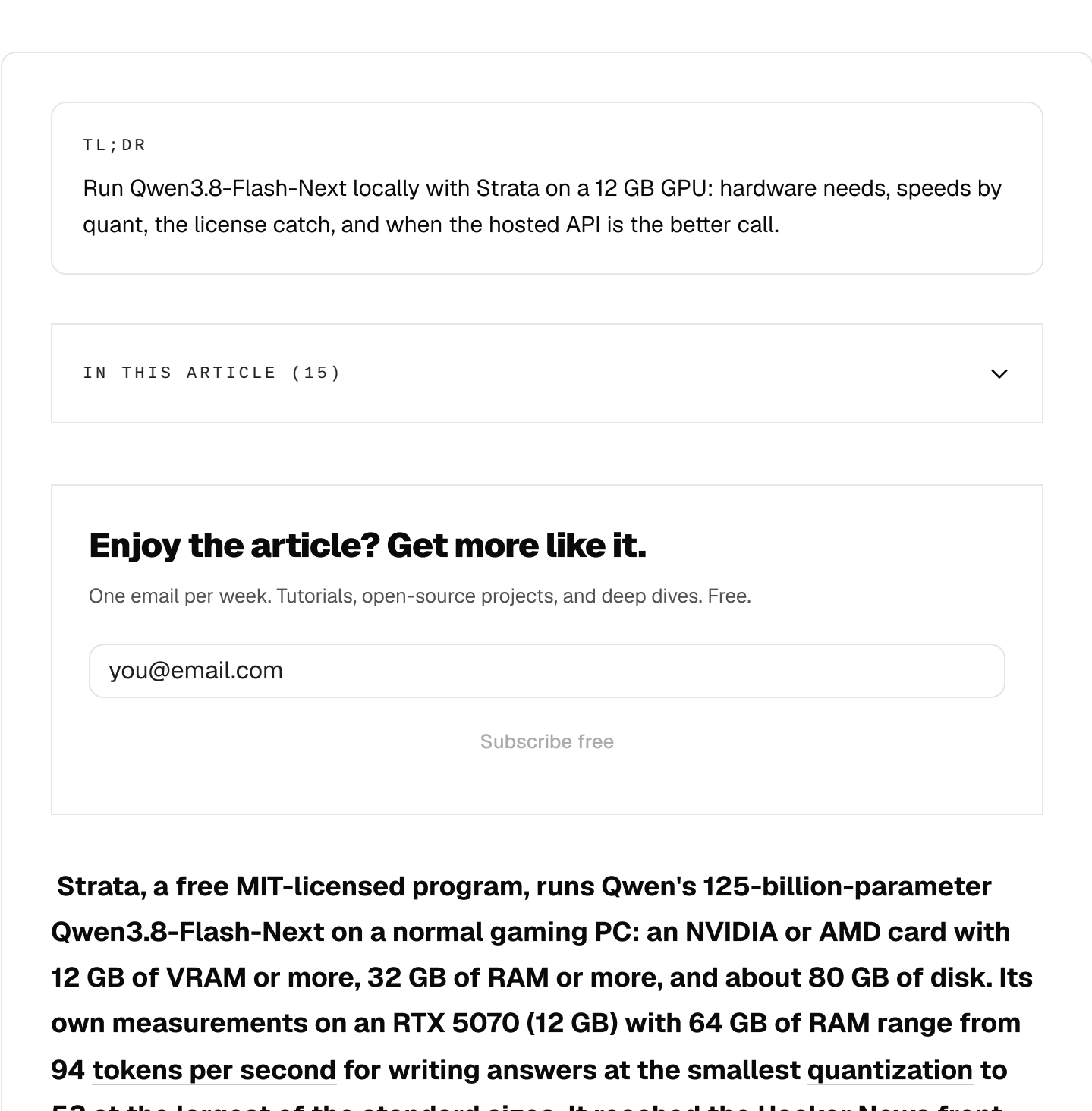


Run Qwen3.8-Flash-Next on a Gaming PC With Strata

Developers Digest • October 4, 2026 • Updated Oct 8, 2026 • 8 min read

Qwen Open Weights Local Models AI Coding Tools Open Source



TL;DR
Run Qwen3.8-Flash-Next locally with Strata on a 12 GB GPU: hardware needs, speeds by quant, the license catch, and when the hosted API is the better call.

IN THIS ARTICLE (15)

Enjoy the article? Get more like it.

One email per week. Tutorials, open-source projects, and deep dives. Free.

you@email.com

Subscribe free

Strata, a free MIT-licensed program, runs Qwen’s 125-billion-parameter Qwen3.8-Flash-Next on a normal gaming PC: an NVIDIA or AMD card with 12 GB of VRAM or more, 32 GB of RAM or more, and about 80 GB of disk. Its own measurements on an RTX 5070 (12 GB) with 64 GB of RAM range from 94 tokens per second for writing answers at the smallest quantization to 53 at the largest of the standard sizes. It reached the Hacker News front page on October 4. Below: what you need, which size to pick, how to point Claude Code or Codex at it, and where it breaks.

Last updated: October 8, 2026 - requirements and speed table re-checked against the Strata README, and a pointer added to the smaller local models that fit 8 to 16 GB machines. We have not run Strata ourselves. Every speed below is the Strata author’s measurement or a Hacker News commenter’s report, and the benchmark scores are Qwen’s own.

What changed on October 5

The story kept moving after publication. The Hacker News thread reached 842 points and 370 comments, the Strata repository showed 12.7k GitHub stars when checked on October 5, and today’s source-desk refresh surfaced the story again as the highest-velocity local-model lane with an existing canonical post here rather than a reason to publish a duplicate. Google Trends also says the exact product query is still tiny next to the broader lane: in the US over the past three months, `Qwen 3.8 Flash Next` summed to 151 relative points, `Qwen 3.8` to 2,082, and `local LLM` to 2,556. That is the useful framing: not “everyone is searching Strata”, but “local LLM demand is durable, and this is the current proof object.”

What the model is

Qwen3.8-Flash-Next is an open-weight mixture-of-experts model from the Qwen team, published on Hugging Face in late August. The model card lists:

- 125B parameters with 6B activated per token, plus a 51B n-gram embedding and a 4B multi-token-prediction head.
- 48 layers, 512 experts per layer with 10 routed and 1 shared active, and a vision encoder.
- A 262,144-token native context, extensible to 1 million tokens with YaRN.
- A new hybrid attention design the card calls a preview of the architecture behind Qwen4.

Qwen’s own benchmark table, which is vendor-reported, has it at 62.5 on SWE-bench Pro and 58.7 on DeepSWE 11, ahead of Qwen3.8-27B (61.7 and 42.2) and DeepSeek-V4-Flash-0731 (56.0 and 54.4). DeepSeek wins one coding row, NL2Repo-Bench, at 54.2 against 48.1. The model also lists 73.5 on Toolathlon and 91.7 on GPQA Diamond. Treat all of it as marketing until someone reproduces it. For the smaller dense sibling that fits one consumer GPU, see our Qwen3.8-27B local coding guide.

What Strata does

Strata is an installer and engine built on parts of llama.cpp. Its README says it makes a model this large fit a consumer PC by spreading the work: the most-used experts stay on the graphics card, all experts live in system RAM, and the processor works on the rest at the same time. It also uses a small draft model to guess ahead and verifies the guesses in one pass, which the README says gives the same answers 1.6 to 1.8 times sooner.

Mainline llama.cpp now has its own version of the expert cache: since October 7, `--moe-cache-mib` keeps recently used experts on the GPU while the rest stay in RAM. Our llama.cpp MoE offload guide compares it with `--n-cpu-moe` and `--cpu-moe`, using benchmarks on this same model.

Requirements from the README:

Part	Needs
Graphics card	NVIDIA RTX 20, 30, 40 or 50 series, or a supported AMD Radeon, with 12 GB of VRAM or more
RAM	32 GB or more; your RAM decides which size fits
Disk	About 80 GB free, SSD recommended
System	Windows 10 or 11, or Linux

How fast, and which size to pick

The README’s own measurements, writing a 4K-token answer and reading a 32K-token prompt:

Quant (RTX 5070, 12 GB, 64 GB RAM)	Writes answers	Reads your prompt
Q2_0	94 tokens/s	2,650 tokens/s
IQ2_XS	79 tokens/s	2,090 tokens/s
IQ3_XXS	62 tokens/s	1,750 tokens/s
IQ3_S	53 tokens/s	1,620 tokens/s
Coder	55 tokens/s	2,180 tokens/s

On an AMD RX 9070 XT (16 GB, 47 GB RAM) the same table reads 60, 52 and 44 tokens per second for Q2_0, IQ2_XS and the Coder build. The author estimates an RTX 3090 would write about 100 to 140 tokens per second. In the Hacker News thread, one commenter reported 124 tokens per second on an RTX 4090 with 128 GB of DDR5, which fits that range but is a single unverified report.

The README’s picks by RAM:

- 32 GB:** the Coder build, a coding variant with half of the experts removed. Its authors say it reaches 91% of the full model’s SWE-bench Verified score, and the README warns it is weaker outside code, including Chinese and other CJK text.
- 48 GB:** IQ2_XS, or Q2_0 if you want the fastest.
- 64 GB:** IQ2_XS is recommended; every size fits.
- 96 GB or more:** IQ3_S, or Unsloth’s roughly 4-bit UD-IQ4_XS.

Closer-to-full-quality builds exist but pay for it: the README says Unsloth’s UD-Q4_K_XL reads most of its weights from the SSD while answering and writes only 7 to 8.5 tokens per second on a 64 GB PC.

you@email.com

Subscribe

Point your coding agent at it

Strata serves an OpenAI-compatible endpoint at `http://127.0.0.1:8080/v1`, with any API key and any model name accepted. According to the README it also serves the Anthropic Messages API, so Claude Code works with `ANTHROPIC_BASE_URL=http://127.0.0.1:8080`, and Codex CLI works through the Responses API at `/v1/responses`. A reasoning setting of `off`, `low`, `medium` or `high` is exposed in the chat menu and as your app’s reasoning effort. If you already use a local model in a coding agent, this slots in the same way as in our best local coding LLMs roundup.

What people are actually saying

The strongest positive reports are practical, not benchmark-shaped: one Hacker News commenter called the smaller Qwen3.8-27B “good enough” for small and mid-sized coding tasks, while another said the model felt stronger inside a different harness. That lines up with our broader warning that agent harness choice can matter as much as the model.

The main objection is quantization. Several commenters pushed back on treating Q2 or expert-pruned builds as equivalent to the full model, which is the right skepticism: the impressive Qwen model-card scores are not automatically the scores you get from the fastest local build.

The infrastructure counter-case is still hosted inference. If you need full precision, parallel requests or predictable long-context behavior, Qwen Cloud’s price is low enough that buying RAM just to save tokens may not pencil out. For the larger economics, see our self-hosting break-even math.

Where it breaks

Quality at 2-bit. The loudest Hacker News objection is quantization. Commenters call Q2 a poor trade and note that the Coder build, with half its experts removed, will not hold up on long-horizon coding that depends on reasoning. One asks for benchmarks on the quantized builds, which the README does not publish. The “Qwen3.8-Flash-Next beats Opus 4.6 on SWE-bench Pro” numbers above are for the full-precision model, not what you run at Q2.

The first start can freeze the PC. The README says the machine may be slow or stop responding for one to three minutes while 35 to 55 GB loads into RAM, and a 70 GB download comes first.

Slow first prompt. It reads the first message of a chat at roughly one minute per 30,000 tokens, with follow-ups starting in seconds. One commenter asked for prompt-processing speed at 16K context, which is the number that matters for coding agents.

One request at a time by default. Extra requests queue unless you set `"parallel": 2`, which the README says makes each answer slower on a 12 GB card.

A paste-this-into-your-AI install. The README offers a prompt that tells your coding assistant to follow a setup doc in the repository, and a commenter immediately compared it to piping a script into bash. Read `docs/AI_SETUP.md` and the installer yourself before you let an agent run them.

Local or hosted?

Qwen’s own hosted model, Qwen3.8-Flash on Qwen Cloud, is described as the official production version of Flash-Next with a 1M-token context and built-in tools. The listed price, checked on October 4, is \$0.15 per million input tokens and \$0.47 per million output tokens, with cached input at \$0.016. That is within a cent of GLM-5.3-Flash’s first-party price in our GLM-5.3 access guide.

Choose	When
Strata, local	Your code cannot leave the machine, you already own a 12 GB or larger card and 64 GB of RAM, or you want an offline fallback
Qwen Cloud, hosted	You need the full-precision model, steady long-context agent runs, or several requests at once
A smaller local model	You have 8 to 16 GB of RAM total, which is the commenter objection that most of this hardware is out of reach

If you land in that last row, the current local-model shortlist names the pick for each memory size, from Gemma 4 E4B on 8 GB up to Qwen3.8-27B on a 24 GB card.

One license point if you build a product on it: the model ships under the Qwen Community License 1.0, not Apache or MIT. Its text says a licensee that runs a “Model as a Service” or “AI Work Assistant” business must get a separate license from Qwen first. Strata’s MIT license covers the engine, not the model weights. Read the license file before you ship anything on top of it.

FAQ

Can I run a 125B model on an RTX 4090?

According to Strata’s README, yes: it needs a 12 GB or larger NVIDIA or AMD card plus 32 GB or more of system RAM, and it runs the model by keeping the busiest experts on the card and the rest in RAM. The author estimates 100 to 140 tokens per second on a 24 GB RTX 3090, and one Hacker News commenter reports 124 on a 4090 with 128 GB of RAM.

How much RAM do I need for Qwen3.8-Flash-Next?

The README recommends 32 GB for the Coder build, 48 GB for IQ2_XS or Q2_0, 64 GB for IQ2_XS as the sweet spot, and 96 GB or more for the larger sizes.

Does it work with Claude Code?

Strata serves an Anthropic-compatible endpoint, and the README gives `ANTHROPIC_BASE_URL=http://127.0.0.1:8080` for Claude Code. We have not tested it ourselves. Our Claude Code permissions guide covers keeping an agent on a short leash.

Is Qwen3.8-Flash-Next free to use commercially?

The weights are free to download, but the Qwen Community License 1.0 requires a separate license from Qwen if you run a hosted model service or a coding or office assistant product on it. Strata itself is MIT licensed.

Continue Reading

RTX Spark for Local LLMs: Will It Run a Coding Agent? - the bandwidth ceiling for this 6B-active model on NVIDIA’s new 128 GB laptops

Best Local AI Models in 2026 - the current local-model shortlist by hardware tier

llama.cpp MoE Offload: --n-cpu-moe vs the New Expert Cache - the same expert-caching idea, now built into mainline llama.cpp

Qwen3.8-27B for Local Agentic Coding - the dense 27B sibling that fits a single consumer GPU

The Best Local Coding LLMs of 2026 - where the local options stack up

Qwen 3.8 Max Release Guide - the hosted flagship from the same team

Where to Run GLM-5.3 Free and Cheap - the closest hosted price comparison

Budget AI Coding Models Compared - cost per task for the cheap tier

Sources

Source	URL
Strata README (speeds, requirements, model picks), fetched October 5, 2026	https://github.com/Niko1221/Strata
Qwen3.8-Flash-Next model card, fetched October 4, 2026	https://huggingface.co/Qwen/Qwen3.8-Flash-Next
Qwen Community License 1.0 (model license text)	https://huggingface.co/Qwen/Qwen3.8-Flash-Next/raw/main/LICENSE
Qwen3.8-Flash on Qwen Cloud, pricing checked October 4, 2026	https://www.qwencloud.com/models/qwen3.8-flash
Hacker News discussion, read October 5, 2026	https://news.ycombinator.com/item?id=49952495
Google Trends query cluster (Qwen 3.8 Flash Next, Qwen3.8 Flash, Qwen 3.8, local LLM, RTX 4090 AI), US past 3 months, checked October 5, 2026	https://trends.google.com/trends/

Next issue: this week’s AI dev stories, ranked

Each story links to its primary source and says what it means for builders. Then new videos, new articles, and what to watch next week. Free, one email a week.

you@email.com

Subscribe

READ NEXT ON LOCAL AND OPEN-WEIGHT MODELS

Qwen3.8-27B vs Opus 4.6 Max: Local Agentic Coding Benchmarks
Qwen3.8-27B is a 27B dense Apache-2.0 model that beats...
10 min read

The Best Local Coding LLMs in 2026: Run Enterprise-Grade AI Without the Cloud
Best local coding LLMs in 2026: which models to run on what...
8 min read

Qwen 3.8 Max Ships: 2.4T MoE, 1M Context, \$2/\$6 per MTok, Open Weights Next Week
Qwen 3.8 Max is a 2.4T-parameter MoE with 95B activ...
8 min read

NEW HERE? START WITH

Best local models Every AI tool comparison

Share

Twitter/X LinkedIn Reddit Hacker News

Email Copy Cite

Suggest an edit Save

Discuss this article on Twitter/X

Developers Digest

Technical content at the intersection of AI and development. Building with AI agents, Claude Code, and modern dev tools - then showing you exactly how it works.

300+ videos 30K+ GitHub stars 850+ articles

Subscribe

YouTube GitHub Twitter/X

Related Tools

AI Coding

Open source

DeepSeek-TUI

Open-source (MIT) terminal coding agent, renamed Codewhale, with Plan, Work, and Operate modes...

[View Tool](#)

AI Coding

Alder

Open-source AI pair programming in your terminal. Works with any LLM - Claude, GPT, Gemini, local...

[View Tool](#)

AI Coding

Z

Zed

High-performance code editor built in Rust with native AI integration. Sub-millisecond input latency...

[View Tool](#)

AI Coding

Cline

Open-source autonomous coding agent inside VS Code. Creates files, runs commands, and can use a...

[View Tool](#)

Related Guides

Guide

Getting Started with DevDigest CLI

Install the dd CLI and scaffold your first AI-powered app in...

[Getting Started](#)

Guide

Interactive Mode - Claude Code

Real-time prompt loop with history, completions, and...

[Claude Code](#)

Guide

Claude Code Complete Course

A complete, citation-backed Claude Code course with...

[ai-development](#)

Related Videos

10 min read

Owen

Owen3.8-27B vs Opus 4.6 Max: Local Agentic Coding Benchmarks

Owen3.8-27B is a 27B dense Apache-2.0 model that beats Opus 4.6 Max on SWE-bench Pro (61.7 vs 53... Benchmarks, hardware...

[Video](#) · March 21, 2025

8 min read

local llm

Open Source Perplexity Clone with Lepton AI in...

In this video, I'm taking a quick look at Lepton Search, an ope...

[Video](#) · January 30, 2024

8 min read

glim

LLAMA 2 - Get Started With Meta's Newest...

#LLAMA-2 #Meta #AI in this video I show you LLAMA 2...

[Video](#) · July 20, 2023

Related Posts

10 min read

Owen

Owen3.8-27B vs Opus 4.6 Max: Local Agentic Coding Benchmarks

Owen3.8-27B is a 27B dense Apache-2.0 model that beats Opus 4.6 Max on SWE-bench Pro (61.7 vs 53... Benchmarks, hardware...

[August 15, 2026](#)

8 min read

local llm

The Best Local Coding LLMs in 2026: Run Enterprise-Grade AI Without the Cloud

Best local coding LLMs in 2026: which models to run on what hardware, how they benchmark, owned...

[June 10, 2026](#)

8 min read

News

Owen 3.8 Max Ships: 2.4T MoE, 1M Context, \$2/\$6 per MTok, Open Weights Next Week

Owen 3.8 Max is a 2.4T-parameter MoE with 95B active, a 1M context window and \$2/\$6 per million...

[August 3, 2026](#)

13 min read

glim

GLM-5.3 Free and Cheap: Every Route and Price (October 2026)

GLM-5.3 has no free hosted tier. Cheapest on October 9: Wafer via OpenRouter at \$0.049/\$3.39... Flash from \$0.075/\$0.25. Z...

[August 14, 2026](#)

11 min read

AI Models

Budget AI Coding Models, October 2026: Haiku 5.5 vs GPT-6 Luna

Claude Haiku 5.5 and GPT-6 Luna both list at \$0.10/\$0.50 per MTok. The price tie breaks at...

[July 31, 2026](#)

10 min read

Open Weights

How to Run Kolibri Locally: GGUF, MLX and Hardware Picks

Kolibri 1 runs locally from 28.6 GB (Q2_K) to 831 GB (Q8_0). Which quant fits your RAM or Mac, measure...

[October 9, 2026](#)

Build with the member tools

Chat, image and voice generation, memory, and more run on one universal credit balance across every Developers Digest app. Sign up free and get 25 credits, no card required.

Start free →

See pricing