

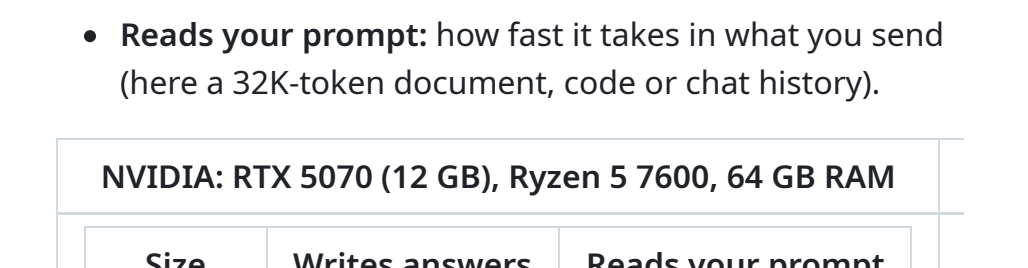
Niko1221	Merge f-1441-perf (Respons...	296b59d · 13 hours ago	
.github	chore(github): issue fo...		last week
bench	arena: pin by need (ST...		yesterday
cmake	gfx1103 (Radeon 780...		2 days ago
data	Formats cost nothing ...		2 days ago
docs	serve: the config's rea...		13 hours ago
include/strata	pinned: load_experts...		yesterday
metal	Metal: gdn_step fits its...		15 hours ago
ref	Strata: Qwen3.8-Flash...		3 weeks ago
serve	Merge f-1441-perf (Re...		13 hours ago
src	pool_tasks_test: skipp...		13 hours ago
sycl	sycl: prompt GDN rec...		yesterday
tests	Merge f-qwen36 into f...		yesterday
third_party/ggml	Strata: Qwen3.8-Flash...		3 weeks ago
tools	Merge f-1441-perf (Re...		13 hours ago
.dockerignore	Docker build (#96): th...		2 weeks ago
.gitattributes	Quant yardstick: tools...		2 days ago
.gitignore	chore(git): add standa...		last week
AGENTS.md	docs: a short Contribu...		last week
BENCH.bat	Benchmark mode: BE...		4 days ago
CMakeLists.txt	Version 0.1.44.1: proje...		13 hours ago
Dockerfile	0.1.44: version, engin...		yesterday
LICENSE	License: MIT (the ggm...		2 weeks ago
README.de.md	README translations: ...		17 hours ago
README.es.md	README translations: ...		17 hours ago
README.fr.md	README translations: ...		17 hours ago
README.ja.md	README translations: ...		17 hours ago
README.md	README: how to instal...		17 hours ago
README.pt-BR.md	README translations: ...		17 hours ago
README.zh-CN.md	README translations: ...		17 hours ago
SECURITY.md	web: a Model settings ...		2 weeks ago
SETUP.bat	Engine 0.1.18; setup e...		2 weeks ago
START-HERE.bat	setup: AMD on Windo...		2 weeks ago
UPDATE.bat	Fix PowerShell 5.1 co...		5 days ago
bench.sh	Benchmark mode: BE...		4 days ago
chat.py	Strata: Qwen3.8-Flash...		3 weeks ago
docker-entrypoint.sh	docker: warn, not stop...		5 days ago
flake.lock	Nix environment		yesterday
flake.nix	rm git dep		yesterday
requirements.txt	setup: reproducible in...		2 weeks ago
setup.py	Merge f-1441-perf (Re...		13 hours ago
setup.sh	setup on macOS: met...		2 days ago
update.sh	setup: checksum verifi...		last week

Strata

English · 简体中文 · 日本語 · Deutsch · Français · Español · Português

Run a 125-billion-parameter AI model on your own gaming PC

NVIDIA or AMD graphics card (12 GB or more) · Windows or Linux · free and open source



A voxel pagoda garden, 1 shot prompt running on an RTX 5070 with Strata (IQ3_S, 128K context) · [full video \(49 s\)](#)

Strata runs [Qwen3.8-Flash-Next](#) on a normal PC. This is a large, smart AI model that usually needs a server. It chats, writes code, reads pictures and works with your apps and coding agents. Nothing leaves your PC.

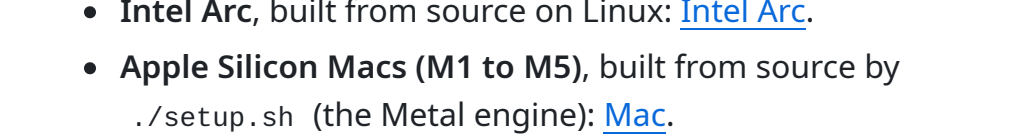
How fast is it?

We measured it on two ordinary gaming PCs. A token is about ¾ of a word.

- Writes answers:** how fast the reply appears in a short chat. 60 tokens per second is faster than you can read.
- Reads your prompt:** how fast it takes in what you send (here a 32K-token document, code or chat history).

NVIDIA: RTX 5070 (12 GB), Ryzen 5 7600, 64 GB RAM			
	Size	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

NVIDIA: Q2_0 with engine 0.1.36, the other rows with 0.1.26 (4K answers, 32K prompts). The full tables are in [DETAILS.md](#). A card with more VRAM is faster: an RTX 3090 (24 GB) should write about 100-140 tokens per second. Long chats and other cards: [speed of each model](#), [community results](#).



Strata is free. If it runs well on your PC, a coffee keeps the work on it going.

What you need

Graphics card	NVIDIA GeForce RTX 20, 30, 40 or 50 series, or AMD Radeon RX 7900 XT / XTX, RX 7800 XT / 7700 XT, RX 9060 XT, RX 9070 / 9070 XT, Radeon AI PRO R9700 or RX 6800 / 6900 series. It needs 12 GB of VRAM or more.
RAM	32 GB or more. Your RAM decides which model fits. 64 GB runs every size.
Disk	About 80 GB free. Use an SSD if you can: the first start is much faster.
System	Windows 10 / 11 or Linux, and a current graphics driver from NVIDIA or AMD.

The installer sets up everything else. Two or three cards can share the model ([multi-GPU](#)).

Experimental, written and tested by community members on their own machines:

- Older graphics cards** (Tesla P40 / V100, GTX 10, Radeon VII / M150, RX 6700 XT, RX 5500 XT): [Older GPUs](#).
- Intel Arc**, built from source on Linux: [Intel Arc](#).
- Apple Silicon Macs (M1 to M5)**, built from source by `./setup.sh` (the Metal engine): [Mac](#).
- AMD Ryzen AI** (Max Strix Halo), built from source on Linux: [Strix Halo](#).
- Older processors without AVX2**: they work, but slowly. [Older CPUs](#).

The full list: [docs/INSTALL.md](#).

Install

Let your AI set it up

Do you use an AI coding assistant (Claude Code, Cursor, Codex, GitHub Copilot, ...)? Paste this into it:

Set up Strata on this PC for me: <https://github.com/Niko1221/strata>

It checks your graphics card, RAM and disk and picks the model that fits. Then it installs and starts it and tells you how to connect your apps. AI tools can also install, start and stop Strata through its [MCP server](#).

Or do it yourself

[Download Strata](#) and unzip it (or `git clone` it). **Windows**: double-click `START-HERE.bat`. **Linux**: run `./setup.sh` in the Strata folder.

Mac (Apple Silicon, M1 to M5, macOS 15 or newer): install Apple's Command Line Tools once (`xcode-select --install` in Terminal; Xcode itself is not needed). Then open Terminal in the Strata folder and run `./setup.sh`. It builds the Mac engine on your Mac (about 7 minutes), asks the same questions, downloads the model and starts it. Next time, `./setup.sh` starts it right away. Which size fits your Mac and how fast it is: [docs/MAC.md](#).

The steps are the same for NVIDIA and AMD. The installer finds your card and sets up the right engine for it. It asks you a few questions:

- which model and which size,
- how much context (how much text the model keeps in mind),
- whether it should read pictures.

Press Enter each time the recommended answer. Then it downloads the model (about 70 GB) and starts it. If the download stops, run it again: it continues where it left off. Your browser opens the Strata app at `http://127.0.0.1:8080`.

While the model starts, your PC can be slow or stop responding for 1-3 minutes (longest the first time). Strata loads 35-55 GB into your RAM and locks part of it for the graphics card. This is normal. Wait, and don't close the window. The window shows what Strata is doing.

Next time, run `START-HERE.bat` (or `./setup.sh`) again. It starts right away and downloads nothing twice. Close its window to stop the model. `UPDATE.bat` (`./update.sh`) updates Strata without starting it. Updating, Docker, several cards, where the files go and every option: [docs/INSTALL.md](#).

Is it fast on your PC? Run `BENCH.bat` (`./bench.sh`): about 5-10 minutes, it measures chat, 4K/16K prompts and an agent-state session with your own settings and writes a report to share ([how](#)).

Which model should I pick?

The installer recommends one for your RAM. The same model comes in several sizes, compressed more or less. Smaller sizes are faster. Larger sizes are a bit smarter.

Your RAM	Take	Why
32 GB	Coder	it fits 32 GB, and it is made for code (with a 24 GB card, Q2_0 and IQ2_XS run too)
48 GB	IQ2_XS (or Q2_0, the fastest)	the larger sizes do not fit
64 GB	IQ2_XS (recommended), or IQ3_XXS / IQ3_S	every size fits; IQ3_S is the best and the slowest
96 GB or more	IQ3_S, or Unsloth's UD-IQ4_XS (~4-bit)	room for the largest sizes with everything else open

- Coder**: a coding version with half of the experts removed. It reaches 91% of the full model's SWE-bench Verified score (measured by its authors) and fits 32 GB of RAM. It is weaker outside code, including Chinese and other CJK text (#438). For those, take Q2_0, IQ2_XS or IQ3_S, which keep every expert.
- Swift 1.5**: a fine-tune that thinks for a much shorter time before it answers. You get the answer sooner, at about the same quality.
- Unsloth UD-IQ4_XS**: Unsloth's ~4-bit version, between IQ3_S and UD-Q4_K_XL in quality. A 94 GB download. With less than ~80 GB of RAM, Strata reads part of it from the SSD while it answers, so it is slower there (an NVMe SSD helps).
- Unsloth UD-Q4_K_XL, UD-Q5_K_XL, UD-Q6_K_XL and Q8_0**: 4- to 8-bit, closer to the full model (93-95% the same next token as Q8_0). Fast where every expert fits in RAM (~88 GB for UD-Q4_K_XL up to ~136 GB for Q8_0); with less RAM Strata reads part of them from the SSD while it answers, which is much slower (7-8.5 tokens/s for UD-Q4_K_XL on a 64 GB PC).
- OrcaRouter's Uncensored IQ3_XXS**: you set it up by hand. It is not in the installer's menu.

Sizes, downloads and what fits where: [docs/MODELS.md](#). To add another model later, run `SETUP.bat` (Linux: `./setup.sh --setup`).

Using it



The Strata app's Monitor (left) while a coding agent writes the pagoda garden from the video (right)

- In the browser**: open `http://127.0.0.1:8080`. It has Chat, a live Monitor of the model and your GPU/CPU/RAM, and About with the settings and addresses.
- Your apps and coding agents**: add an "OpenAI-compatible" provider with the base URL `http://127.0.0.1:8080/v1`. Any API key and any model name work.
 - Apps that use Anthropic's API: `http://127.0.0.1:8080/v1/messages` (Claude Code: `ANTHROPIC_BASE_URL=http://127.0.0.1:8080`).
 - Codex CLI and other apps that use the OpenAI Responses API: `/v1/responses` ([setup](#)).
- Thinking**: choose **off**, **low**, **medium** or **high** in the chat menu or in your app's "reasoning effort". Off is the fastest. High is best for hard questions.
- Pictures**: say yes to "Images?" in setup. Then click **Picture** in the chat, or attach pictures in your app. AMD cards read pictures on Linux through the processor; on Windows they can't yet.
- From your phone or another PC**: `START-HERE.bat --setup --host 0.0.0.0 --api-key <secret>`. Always set a key.
- One request at a time**: by default Strata answers one request, and the others wait. To answer several at once, set `"parallel": 2` ([BATCHING.md](#)). On a 12 GB card this makes each answer slower.
- Long prompts**: Strata reads the first message of a chat in full, about 1 minute per 30,000 tokens. Follow-up messages start in seconds.

More: [where your chats are stored](#), [the API](#).

Optional: more speed on one card when the whole model fits in RAM

Two opt-in settings (by RubenTorney, #1904 and #1905). They help when you run one graphics card and your RAM holds every expert (Strata's arena mode, for example Q2_0 or IQ3_XXS on a 64 GB PC). Open your config, `strata-<model>.json` in the Strata data folder ([where](#)), and add:

- Faster decode**: `--adapt-async`, `"1"` to the `"args"` list. The expert cache then updates on a helper thread instead of pausing the answer. Measured on an RTX 5070 (Windows, Q2_0): +4-5% decode, +7-10% on code and long documents. Answers can differ slightly from the default (they stay the same from run to run).
- Faster prompt reading on Windows**: `"STRATA_PIN_BY_NEED": "1"` in the `"env"` block. Windows can lock only about half the RAM for the graphics card; this locks the experts the card actually needs first. Measured on the RTX 5070: +4.4% prompt reading.

```
"args": [ ..., "--adapt-async", "1" ],
"env": { "STRATA_PIN_BY_NEED": "1" }
```

Restart Strata (`START-HERE.bat` or `./setup.sh`) after the change. To undo, remove the two entries.

Something went wrong?

- My PC froze the first time Strata started**. This is normal while it loads the model. Wait, and don't close the window. Still frozen after 10 minutes? Restart the PC, close other programs and try again, or pick a smaller size.
- It stopped while downloading or installing**. Run `START-HERE.bat` (or `./setup.sh`) again. It continues where it stopped.
- It's very slow and the disk light keeps blinking, or it says "the engine stopped unexpectedly"**. Your PC does

About

Qwen3.8-Flash-Next on any consumer hardware: one-click install for Windows / Linux. Strata infers engine, OpenAI/Anthropic API on localhost, optional image input.

- Readme
 - MIT license
 - Security policy
 - Activity
 - 24.2k stars
 - 130 watching
 - 2.4k forks
- Report repository

Releases 51

Strata 0.1.44.1 Latest 12 hours ago

+ 50 releases

Contributors 261



+ 247 contributors

Languages

C++ 64.8% Python 21.5% CUDA 10.2% Metal 1.5% CMake 0.8% JavaScript 0.3% Other 0.9%

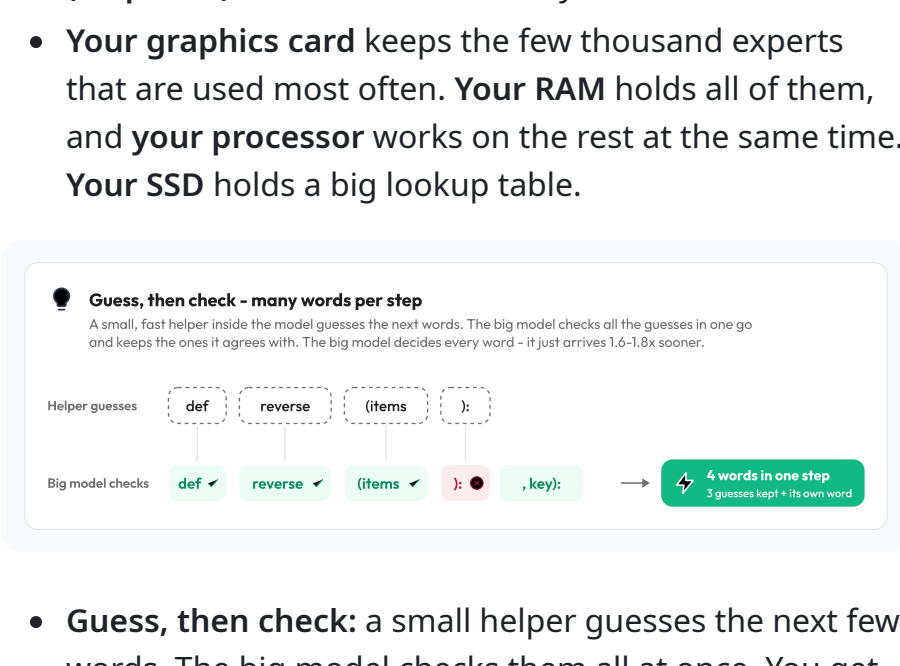
not have enough free RAM. Close other programs (browsers use a lot), or pick a smaller size (Q2_0 or IQ2_XS).

- **It says port 8080 is already in use.** Strata is already running. Look for its window.

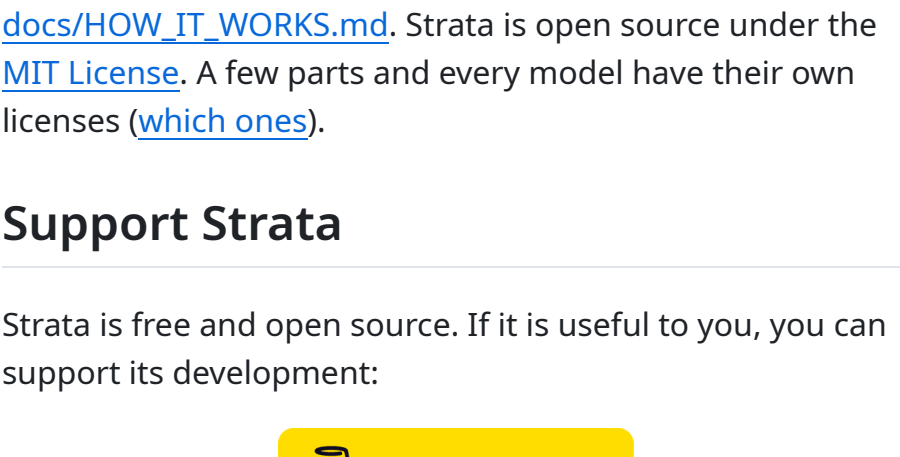
More problems and their fixes: [docs/TROUBLESHOOTING.md](#). Still stuck? Open an [issue](#) and attach `strata-<model>.log` from the Strata folder. Found a security problem? Report it privately: [SECURITY.md](#).

How does it work?

Models like this one usually run on servers with hundreds of gigabytes of graphics memory. Your graphics card has 12-24 GB. Strata makes the model fit by **sharing the work across your whole PC**. Think of a kitchen: the things you use all the time stay on the counter, and the rest waits in the pantry.



- **The model is a team of 24,576 small specialists** ("experts"). Each word needs only 10 of them.
- **Your graphics card** keeps the few thousand experts that are used most often. **Your RAM** holds all of them, and **your processor** works on the rest at the same time. **Your SSD** holds a big lookup table.



- **Guess, then check:** a small helper guesses the next few words. The big model checks them all at once. You get the same answer, 1.6-1.8x sooner.
- **Long texts are read in big pieces** (up to 8,192 tokens at a time), at over 1,000 tokens per second.

The longer explanation: [docs/HOW_IT_WORKS.md](#). Every part and its numbers: [the details](#) and the [paper](#).

Credits and license

The model is [Qwen3.8-Flash-Next](#) by the Qwen team. It was compressed by [ISTA-DASLab](#), UkisAI (Swift 1.5) and Unsloth. Strata uses parts of [llama.cpp](#) / [ggml](#). All credits: [docs/HOW_IT_WORKS.md](#). Strata is open source under the [MIT License](#). A few parts and every model have their own licenses ([which ones](#)).

Support Strata

Strata is free and open source. If it is useful to you, you can support its development:

