SILENTSWARM

Pre-evaluation. The source code is not public. People evaluating it run it on their own machines.

For teams that need GPU compute

GPU capacity, without the data centre

SilentSwarm runs open language models across ordinary machines - your own, your partners', or units that heat buildings. You get an OpenAI-compatible API, and you know which machines did the work.

Why run it this way

Bigger models on smaller cards. A model is split into stages across machines, so a pool of 24 GB cards can serve and fine-tune a model none of them could hold alone.

You decide where it runs. Run the coordinator yourself and keep jobs on the machines you chose - in your office, at a partner, in the EU. Nothing goes to a third-party cloud unless you put a machine there.

Heat that gets used. A pool can include compute heaters, machines whose waste heat warms a building. The energy every machine draws is metered, so the figure has a source.

Measured on two home PCs

Qwen3-14B in 4-bit, split across two ordinary PCs on one home internet connection: two GTX 1050 Ti cards with 4 GB each in one, two GTX 1660 SUPER cards with 6 GB each in the other. Measured on 2026-10-19 through the public chat API.

3.3
tokens per second for one user

11.6
tokens per second for eight users at once

5.4
tokens per second on code edits, guessing ahead

9.2
tokens per second on code edits, one PC

Small samples on old cards: these numbers show the shape, not a guarantee. The benchmark page has the method, the raw results and the limits, including a minute of waiting before the first token of a 1,500-token prompt.

What exists, and what does not yet

Available today

A self-hosted coordinator with pools, organizations, scoped API tokens and an audit log.

Serving through an OpenAI-compatible API, across one machine or several.

Fine-tuning split across machines, with checkpoints and exported models.

Energy metering per machine, from the GPUs' own power readings.

Published benchmarks from real runs across machines: Qwen3-14B on two home PCs, 11.6 tokens per second for eight users.

Not yet

Selling idle capacity. Credits for busy GPU time, then payouts.

Usage billing and invoices.

Service-level guarantees. It is in pre-evaluation.

A managed capacity offer. Tell us what you need, and we will tell you honestly whether a pilot fits.

Apply for the evaluation: silentswarm.nexpatch.ai/evaluation

How it handles your data: silentswarm.nexpatch.ai/trust · Benchmarks: silentswarm.nexpatch.ai/docs/02_experiments/benchmarks/

NexPatch AI UG (haftungsbeschränkt) · Fabian Franz · Neumarkt 31, 04109 Leipzig, Germany · Amtsgericht Leipzig, HRB 45089 L · info@nexpatch.ai