SILENTSWARM

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

For heat pump and heating manufacturers

Heat from compute, run by software

A GPU turns nearly every watt it draws into heat. SilentSwarm puts those GPUs to paid work wherever your units are installed, and tells you how much heat each one delivered.

Why compute belongs in the heating loop

Same electricity, two products. Every kilowatt-hour runs a computation first and heats the building after. The GPU is an electric heater that does paid work on the side.

A source the heat pump likes. A liquid-cooled GPU loop runs far warmer than outdoor air in winter. The warmer the source, the less the compressor has to lift.

Many small sites, one cluster. No basement holds a data centre. SilentSwarm splits a model across machines in different buildings, so many modest units add up to one useful cluster.

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

Energy metering per unit, per fleet and in total, from the GPUs' own power readings.

The heat-demand API and the placement that favours units whose building wants heat.

One-command onboarding for units behind any router, with a health grade per unit.

Open models served through an OpenAI-compatible API and fine-tuned across machines.

Not yet

Following the demand by capping GPU power. That needs driver-level rights on the unit, which the worker deliberately does not have today.

Billing and payouts for compute sold from your fleet.

Calibrated heat metering. The figures come from the GPUs' sensors, not from a certified heat meter.

Hardware. We write the software; the unit, the hydraulics and the heat pump are yours.

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