SILENTSWARM

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

For cities, municipalities and municipal utilities

AI for the city, on the city's own machines

SilentSwarm joins the GPUs a city already runs - in schools, libraries and offices - into one private cluster that serves and fine-tunes open language models. The data stays on the city's network, and the heat can warm the buildings it runs in.

Why a city should run it this way

The data stays in the city. The coordinator runs on a server of the city's, and the machines in the pool are the city's own. Prompts, documents and trained models never leave its network.

Use what is already paid for. A school's computer room or an office's workstations join the pool after hours. Availability windows keep a machine out of the pool while people need it.

Heat where the city needs it. Every machine's GPU energy is metered, and a building's heat demand steers the work towards the machines whose heat is wanted - a school in winter, a swimming pool all year.

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.

Availability windows per machine, and one-command onboarding behind any firewall.

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

Energy metering per machine, and the heat-demand signal that steers where work runs.

Not yet

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

Windows machines. A machine needs Linux and an NVIDIA GPU today; Windows is untested.

Search over the city's documents. There is no document connector yet: a model answers from what it was trained on and what a request contains.

Certifications such as BSI C5 or ISO 27001, and service-level guarantees. It is in pre-evaluation.

An offline install. A machine installs PyTorch from PyPI when it joins, and a coordinator without internet needs its models put on it by hand.

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