SILENTSWARM

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

For university labs and research groups

A 14B model on the GPUs your lab already owns

Many groups have a few workstations with 8 to 24 GB cards and no budget for one with 80 GB. SilentSwarm splits a model across the machines you have, so together they serve and fine-tune models none of them could hold alone - on hardware you run yourselves.

Why it fits a research group

Models bigger than any one card. Our benchmark model, Qwen3-14B, runs across four older GPUs of 4 and 6 GB. None of them could hold it alone.

Fine-tune, not only chat. LoRA fine-tuning runs on the same split as serving, with checkpoints along the way and an exported model at the end. A thesis project can train on its own data without a cloud account.

Data that stays put. The coordinator is a small server you run, and it needs no GPU. Prompts and training data stay on machines you run, the coordinator included.

Numbers you can cite and rerun. Every benchmark we publish names the hardware, the commit and the script, so you can repeat it on your own pool.

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

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

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

Published benchmarks with hardware, commit and scripts, on the benchmark page.

An evaluation for research groups. Terms are agreed during the evaluation.

Not yet

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

Training a large model from scratch. The split is built for fine-tuning.

Fast long prompts on old cards. On the cards above, a long prompt waits about a minute before the answer starts.

Every model family. The split is tested on the Llama, Qwen and GPT-2 families; others may need an adapter.

Support with a response time. We help pilots ourselves, as fast as we can.

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