SILENTSWARM

**Evaluierungsphase.** Der Quellcode ist nicht öffentlich. Wer es evaluiert, betreibt es auf eigenen Rechnern.

Für Städte, Gemeinden und Stadtwerke

# KI für die Stadt, auf den eigenen Rechnern der Stadt

SilentSwarm verbindet die GPUs, die eine Stadt schon betreibt - in Schulen, Bibliotheken und Ämtern - zu einem privaten Cluster, das offene Sprachmodelle bereitstellt und feinabstimmt. Die Daten bleiben im Netz der Stadt, und die Wärme kann die Gebäude heizen, in denen gerechnet wird.

## Warum eine Stadt es so betreiben sollte

**Die Daten bleiben in der Stadt.** Der Koordinator läuft auf einem Server der Stadt, und die Rechner im Pool gehören der Stadt. Prompts, Dokumente und trainierte Modelle verlassen ihr Netz nie.

**Nutzen, was schon bezahlt ist.** Der Computerraum einer Schule oder die Workstations eines Amts kommen nach Feierabend in den Pool. Verfügbarkeitsfenster halten einen Rechner aus dem Pool heraus, solange Menschen ihn brauchen.

**Wärme, wo die Stadt sie braucht.** Die GPU-Energie jedes Rechners wird gemessen, und der Wärmebedarf eines Gebäudes lenkt die Arbeit zu den Rechnern, deren Wärme gebraucht wird - in die Schule im Winter, ins Schwimmbad das ganze Jahr.

## Gemessen auf zwei PCs zu Hause

Qwen3-14B in 4 Bit, verteilt auf zwei gewöhnliche PCs an einem privaten Internetanschluss: im einen zwei GTX 1050 Ti mit je 4 GB, im anderen zwei GTX 1660 SUPER mit je 6 GB. Gemessen am 19.10.2026 über die öffentliche Chat-API.

**3,3**
Tokens pro Sekunde für eine Person

**11,6**
Tokens pro Sekunde für acht Personen gleichzeitig

**5,4**
Tokens pro Sekunde bei Code-Änderungen, mit Vorausraten

**9,2**
Tokens pro Sekunde bei Code-Änderungen, ein PC

Kleine Stichproben auf alten Karten: Die Zahlen zeigen die Größenordnung, keine Garantie. Die Benchmark-Seite nennt Methode, Rohdaten und Grenzen - auch die Minute Wartezeit bis zum ersten Token bei einem Prompt mit 1.500 Tokens.

## Was es gibt - und was noch nicht

### Heute verfügbar

**Ein selbst betriebener Koordinator** mit Pools, Organisationen, API-Tokens mit Rechten und einem Prüfprotokoll.

**Verfügbarkeitsfenster** pro Rechner und Einrichtung mit einem Befehl hinter jeder Firewall.

**Offene Modelle**, bereitgestellt über eine OpenAI-kompatible API und über mehrere Rechner hinweg feinabgestimmt.

**Energiemessung** pro Rechner und das Wärmebedarfs-Signal, das lenkt, wo gerechnet wird.

### Noch nicht

**Verkauf freier Kapazität.** Zuerst Guthaben für ausgelastete GPU-Zeit, dann Auszahlungen.

**Windows-Rechner.** Ein Rechner braucht heute Linux und eine NVIDIA-GPU; Windows ist ungetestet.

**Suche in den Dokumenten der Stadt.** Einen Dokumenten-Connector gibt es noch nicht: Ein Modell antwortet aus dem, worauf es trainiert wurde, und aus dem, was eine Anfrage enthält.

**Zertifizierungen** wie BSI C5 oder ISO 27001 und zugesicherte Servicelevel. Es ist in der Evaluierungsphase.

**Eine Offline-Installation.** Ein Rechner lädt beim Beitritt PyTorch von PyPI, und einem Koordinator ohne Internet müssen die Modelle von Hand aufgespielt werden.

## Für die Evaluierung bewerben: silentswarm.nexpatch.ai/de/evaluation

Wie es mit Ihren Daten umgeht: silentswarm.nexpatch.ai/de/trust · Benchmarks: silentswarm.nexpatch.ai/docs/02_experiments/benchmarks/

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