SILENTSWARM

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

Für Teams, die GPU-Rechenleistung brauchen

GPU-Kapazität, ohne Rechenzentrum

SilentSwarm betreibt offene Sprachmodelle auf gewöhnlichen Rechnern - Ihren eigenen, denen Ihrer Partner oder Geräten, die Gebäude heizen. Sie bekommen eine OpenAI-kompatible API und wissen, welche Rechner die Arbeit erledigt haben.

Warum auf diese Weise

Größere Modelle auf kleineren Karten. Ein Modell wird in Abschnitte über mehrere Rechner verteilt. So kann ein Pool aus 24-GB-Karten ein Modell bereitstellen und feinabstimmen, das keine von ihnen allein halten könnte.

Sie entscheiden, wo es läuft. Betreiben Sie den Koordinator selbst und halten Sie Jobs auf den Rechnern Ihrer Wahl - im Büro, bei einem Partner, in der EU. Nichts geht in eine fremde Cloud, solange Sie dort keinen Rechner hinstellen.

Wärme, die genutzt wird. Ein Pool kann Compute-Heizungen enthalten - Rechner, deren Abwärme ein Gebäude heizt. Die Energie jedes Rechners wird gemessen, die Zahl hat also eine Quelle.

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.

Bereitstellung über eine OpenAI-kompatible API, auf einem Rechner oder mehreren.

Feinabstimmung über mehrere Rechner, mit Zwischenständen und exportierten Modellen.

Energiemessung pro Rechner, aus den Leistungswerten der GPUs selbst.

Veröffentlichte Benchmarks aus echten Läufen über mehrere Rechner: Qwen3-14B auf zwei PCs zu Hause, 11,6 Tokens pro Sekunde für acht Nutzer.

Noch nicht

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

Nutzungsabrechnung und Rechnungen.

Zugesicherte Servicelevel. Es ist in der Evaluierungsphase.

Ein verwaltetes Kapazitätsangebot. Sagen Sie uns, was Sie brauchen - wir sagen Ihnen ehrlich, ob ein Pilotprojekt passt.

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