SILENTSWARM

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

Für Hochschullabore und Forschungsgruppen

Ein 14B-Modell auf den GPUs, die Ihr Labor schon besitzt

Viele Gruppen haben einige Workstations mit Karten von 8 bis 24 GB und kein Budget für eine mit 80 GB. SilentSwarm verteilt ein Modell auf die vorhandenen Rechner, sodass sie gemeinsam Modelle bereitstellen und feinabstimmen, die keiner von ihnen allein halten könnte - auf Hardware, die Sie selbst betreiben.

Warum es zu einer Forschungsgruppe passt

Modelle, größer als jede einzelne Karte. Unser Benchmark-Modell Qwen3-14B läuft auf vier älteren GPUs mit 4 und 6 GB. Keine davon könnte es allein halten.

Feinabstimmen, nicht nur chatten. LoRA-Feinabstimmung läuft auf derselben Aufteilung wie die Bereitstellung, mit Zwischenständen unterwegs und einem exportierten Modell am Ende. Eine Abschlussarbeit kann auf eigenen Daten trainieren, ohne Cloud-Konto.

Daten, die bleiben, wo sie sind. Der Koordinator ist ein kleiner Server, den Sie selbst betreiben, und er braucht keine GPU. Prompts und Trainingsdaten bleiben auf Rechnern, die Sie betreiben - den Koordinator eingeschlossen.

Zahlen zum Zitieren und Nachmessen. Jeder Benchmark, den wir veröffentlichen, nennt Hardware, Commit und Skript - Sie können ihn auf Ihrem eigenen Pool wiederholen.

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

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

LoRA-Feinabstimmung über mehrere Rechner, mit Zwischenständen und exportierten Modellen.

Veröffentlichte Benchmarks mit Hardware, Commit und Skripten, auf der Benchmark-Seite.

Eine Evaluierung für Forschungsgruppen. Die Bedingungen vereinbaren wir während der Evaluierung.

Noch nicht

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

Ein großes Modell von Grund auf trainieren. Die Aufteilung ist für die Feinabstimmung gebaut.

Schnelle lange Prompts auf alten Karten. Auf den Karten oben wartet ein langer Prompt etwa eine Minute, bis die Antwort beginnt.

Jede Modellfamilie. Getestet ist die Aufteilung mit den Familien Llama, Qwen und GPT-2; andere brauchen eventuell einen Adapter.

Support mit zugesicherter Antwortzeit. Pilotprojekten helfen wir selbst, so schnell wir können.

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