SILENTSWARM

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

Für Hersteller von Wärmepumpen und Heizungen

Wärme aus Rechenleistung, gesteuert per Software

Eine GPU wandelt fast jedes Watt, das sie aufnimmt, in Wärme um. SilentSwarm gibt diesen GPUs bezahlte Arbeit, wo immer Ihre Geräte stehen, und zeigt Ihnen, wie viel Wärme jedes einzelne geliefert hat.

Warum Rechenleistung in den Heizkreis gehört

Derselbe Strom, zwei Produkte. Jede Kilowattstunde treibt zuerst eine Berechnung an und heizt danach das Gebäude. Die GPU ist eine Elektroheizung, die nebenbei bezahlte Arbeit leistet.

Eine Quelle, die der Wärmepumpe liegt. Ein flüssigkeitsgekühlter GPU-Kreis ist weit wärmer als die Außenluft im Winter. Je wärmer die Quelle, desto weniger muss der Verdichter anheben.

Viele kleine Standorte, ein Cluster. In keinen Keller passt ein Rechenzentrum. SilentSwarm verteilt ein Modell auf Rechner in verschiedenen Gebäuden, sodass viele bescheidene Geräte zusammen einen nützlichen Cluster ergeben.

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

Energiemessung pro Gerät, pro Flotte und insgesamt, aus den Leistungswerten der GPUs selbst.

Die Wärmebedarfs-API und eine Platzierung, die Geräte bevorzugt, deren Gebäude Wärme braucht.

Einrichtung mit einem Befehl für Geräte hinter jedem Router, mit einer Zustandsnote pro Gerät.

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

Noch nicht

Dem Bedarf folgen, indem die GPU-Leistung begrenzt wird. Dafür braucht es Rechte auf Treiberebene, die der Worker heute bewusst nicht hat.

Abrechnung und Auszahlung für Rechenleistung, die Ihre Flotte verkauft.

Geeichte Wärmemessung. Die Werte stammen von den Sensoren der GPUs, nicht von einem zertifizierten Wärmemengenzähler.

Hardware. Wir schreiben die Software; Gerät, Hydraulik und Wärmepumpe kommen von Ihnen.

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