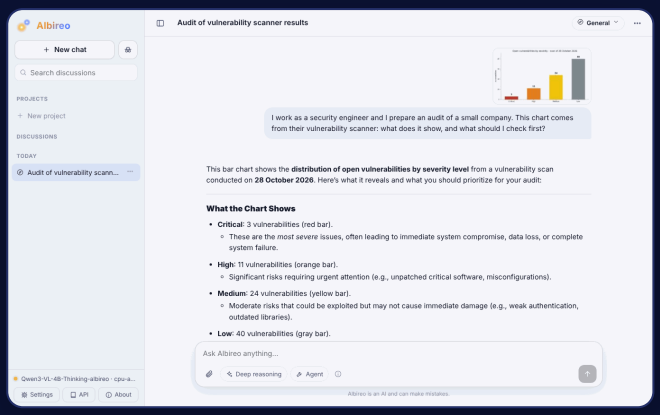
Albireo

Ihr privater KI-Assistent, auf Ihrem eigenen Computer

Albireo führt Sprachmodelle direkt auf Ihrem Windows-11-Computer aus. Ihre Unterhaltungen, Dokumente und Modelle bleiben auf dem Gerät: kein Konto, keine Daten, die nach außen gesendet werden.

Version **1.0.23** Plattform **Windows 11, x64**

Lizenz **Boost Software License 1.0**



56,8
Token pro Sekunde auf einer Laptop-GPU, gegenüber 23,6 auf ihrem Prozessor

22
Modelle im signierten Katalog, von 7 Herausgebern

5
Engines: Prozessoren, NPU und GPU von Intel, GPU von NVIDIA

0
Konten, Abonnements oder nach außen gesendete Daten

Auf einen Blick

Chat
Unterhaltungen, die im Browser bleiben, ein Inkognito-Modus, eine Suche über die Unterhaltungen und die Überlegungen der Modelle, die nachdenken, während sie denken.

Bilder
Die Modelle, die Bilder lesen, nehmen PNG- und JPEG-Bilder an, angehängt, eingefügt oder in eine Frage gezogen.

Agentenmodus
Das Modell ruft Werkzeuge auf, die in der Seite laufen: Rechner, JavaScript in einer Sandbox, Lesen und Durchsuchen der Dokumente, bis zu 100 Schritte pro Antwort.

Vorlesen
Jede Antwort kann von der mit Albireo gelieferten Sprachsynthese vorgelesen werden, ohne Netzwerk.

Offline-Modus
Auf einem Computer ohne Netzwerk öffnet der Dienst keine Verbindung: Katalog, Modelle und Updates kommen aus einem signierten Paket auf einem Datenträger.

Dokumente
Word, Excel, PowerPoint, OpenDocument, PDF, RTF, HTML, Text, CSV, Markdown und Quellcode, im Browser gelesen und einer Nachricht oder einem Projekt angehängt.

Projekte
Unterhaltungen, gruppiert mit Anweisungen und Dokumenten als gemeinsamem Kontext, durchsucht nach Bedeutung und nach Wörtern.

Webzugriff
Eine optionale Suche in Wikipedia und das Lesen öffentlicher Seiten und PDF-Dateien, in jeder neuen Unterhaltung ausgeschaltet.

Eine Schnittstelle für andere Programme
Chat-Vervollständigungen, Vervollständigungen, Embeddings, Sprache und Modelle folgen der OpenAI-Schnittstelle, mit Streaming und Bildern, dokumentiert in einer Swagger-Seite.

Relais zu einer externen Engine
Eine Unterhaltung kann an eine externe Engine weitergeleitet werden: jede OpenAI-kompatible Schnittstelle, die Cloud von Ollama oder Claude. Pseudalis pseudonymisiert ihre Texte, bevor sie den Computer verlassen, und eine Engine des Computers oder des lokalen Netzes bleibt offline nutzbar.

Engines

Engine	Läuft auf
`cpu`	Jedem x86-64-Prozessor
`cpu-avx2`	Prozessoren mit AVX2, FMA und F16C: Intel Haswell, AMD Zen und neuer
`cpu-avxvnni`	Prozessoren mit AVX-VNNI: Intel Core ab Alder Lake, Intel Xeon ab Sapphire Rapids, AMD Zen 5
`openvino`	NPU und integrierter GPU von Intel-Prozessoren, über OpenVINO
`cuda`	NVIDIA-Grafikprozessoren, über CUDA

Tempo auf gängiger Hardware: die Techniken von Albireo

Albireo führt Modelle mit einer eigenen Engine aus, entwickelt für die Prozessoren und Grafikkarten, die bereits im Einsatz sind. Jede der folgenden Techniken ist im README des Produkts gemessen.

Quantisierte Gewichte, Ganzzahlkerne

Die Gewichte bleiben in der Speicherabbildung der GGUF-Datei, in ihrer quantisierten Form. Ganzzahlkerne berechnen 32 Produkte, wo ein Gleitkommavektor 8 berechnet.

Lesen eines Prompts: 216 Token pro Sekunde für Qwen3-VL 4B und mehr als 1.300 für Qwen3 0.6B (Intel Core Ultra 9 275HX, 24 Threads); auf zwei Kernen 230 gegenüber 126 mit Gleitkommazahlen (Qwen3 0.6B, AVX-VNNI)

Eine Engine für jede Art von Prozessor

Portable, AVX2- und AVX-VNNI-Engines für Prozessoren, die NPU und die integrierte GPU von Intel über OpenVINO, NVIDIA-Grafikprozessoren über CUDA. Der Dienst misst sie auf dem Gerät und führt die schnellste aus.

Schreiben: 80,6 bis 87,7 Token pro Sekunde für Qwen3 0.6B mit der Prozessor-Engine (Intel Core Ultra 9 275HX, 24 Threads)

Erweiterter Speicher V²RAM

Der erweiterte Gerätespeicher der Bibliothek lässt die Kopien der Gewichte über den Speicher der Karte hinausgehen, in den Host-Speicher oder eine Datei, und holt sie vor ihrer Verwendung zurück.

Mehrere Unterhaltungen zugleich

Die Anfragen aller Clients werden in jedem Vorwärtsdurchlauf gemeinsam berechnet, und die Schlüssel und Werte einer beendeten Anfrage bleiben für den nächsten Zug im Cache.

32 bis 40 % weniger Zeit pro Schritt bei vier gemeinsam geschriebenen Antworten

Vorausschauendes Dekodieren

Eine Antwort schreibt mehrere Token pro Schritt, wenn ihr Kontext sie ankündigt. Das Modell behält die Token, die es selbst gezogen hätte, sodass die Antworten die Verteilung des Samplings behalten.

25,5 gegenüber 19,7 Token pro Sekunde beim Umschreiben von Code (Qwen3 0.6B, zwei Kerne); 26,3 gegenüber 23,6 auf 24 Threads (Qwen3-VL 4B)

Schichten, auf der NVIDIA-Karte berechnet

Die CUDA-Engine hält die Gewichte in ihren GGUF-Formaten auf der Karte und berechnet dort die Schichten eines dichten Modells, für die erzeugten Token einschließlich der Attention.

56,8 gegenüber 23,6 Token pro Sekunde (Qwen3-VL 4B, RTX 5070 Ti Laptop GPU gegenüber 24 Threads)

Kompaktes Gesprächsgedächtnis

Schlüssel und Werte werden in 8 Bit gehalten, in Seiten, die sich Unterhaltungen mit gleichem Anfangstext teilen.

34 Bytes gegenüber 64 pro Wertegruppe, mit demselben nächsten Token auf 210 Positionen

Ein OpenVINO-Graph auf Prozessor und GPU von Intel

Mit OpenVINO wird ein ganzes Modell als ein einziger Graph auf dem Prozessor oder auf der integrierten GPU von Intel berechnet, und die NPU übernimmt Matrixprodukte. Der Graph liest den ersten Durchlauf eines Prompts schneller; die Prozessor-Engine bleibt bei den folgenden Durchläufen und beim Schreiben schneller, und die Messung der Engines, über das Symbol gestartet, behält die beste Wahl für die Maschine.

Erstes Lesen eines Prompts: 260 Token pro Sekunde mit dem Graphen gegenüber 216 mit der Prozessor-Engine (Qwen3-VL 4B, Intel Core Ultra 9 275HX, 24 Threads)

Merkmale

Modellformat	Quantisiertes GGUF: Q4_0 bis Q8_0, Q2_K bis Q6_K und IQ4_NL
Architekturen	llama, mistral, mistral3, granite, granitemoe, phi3, qwen2, qwen3, qwen3moe, qwen3vl, qwen3vlmoe, mixtral
Bild-Encoder	Qwen3-VL, LLaVA-NeXT (EuroVLM) und Pixtral (Ministral 3, Devstral Small 2, Mistral Small 3.2)
Mit Albireo gelieferte Modelle	Qwen3-VL-4B-Instruct mit seinem Bild-Encoder (2,95 GB) und Qwen3-Embedding-0.6B (0,64 GB)
Katalog	22 Modelle von 7 Herausgebern, signiert und anhand ihrer Hashes geprüft
Arbeitsspeicher	8 GB für das standardmäßig installierte Modell
Lokale Schnittstelle	https://localhost:8443, kompatibel mit der OpenAI-Schnittstelle, dokumentiert mit Swagger
Gesprächsformate	ChatML, Mistral, Llama 3, Granite, Phi-4-mini und Foundation-Sec, nativ geschrieben
Updates	Signierter Feed, Installationsprogramme nach Größe und SHA-256 geprüft

Überprüfbare Sicherheit

Nur lokal

Der Dienst lauscht auf 127.0.0.1 und lehnt Anfragen ab, deren Host oder Ursprung nicht lokal ist.

Eine Zertifizierungsstelle des Computers

Das HTTPS-Zertifikat stammt von einer auf dem Computer erzeugten Stelle mit einem ECDSA-P-256-Schlüssel, durch Namensbeschränkungen auf localhost begrenzt.

Von Windows geschützte Schlüssel

Der private Schlüssel des Servers und der Schlüssel des Relais sind mit DPAPI verschlüsselt und werden nie im Klartext geschrieben.

Eine signierte Lieferkette

Der Update-Feed und der Modellkatalog sind mit ECDSA P-256 signiert, und jede heruntergeladene Datei wird anhand ihres SHA-256 geprüft.

Eine offene Schnittstelle

Die Routen folgen der OpenAI-Schnittstelle und sind mit OpenAPI beschrieben, sodass vorhandene Client-Bibliotheken unverändert funktionieren.

Isolierte Sprachsynthese

Der Synthesizer läuft in einem getrennten Prozess niedriger Integrität, ohne Privilegien und im Speicher begrenzt.

Frei und quelloffen unter der Boost Software License 1.0. © 2026 Cydral Technology, Frankreich.