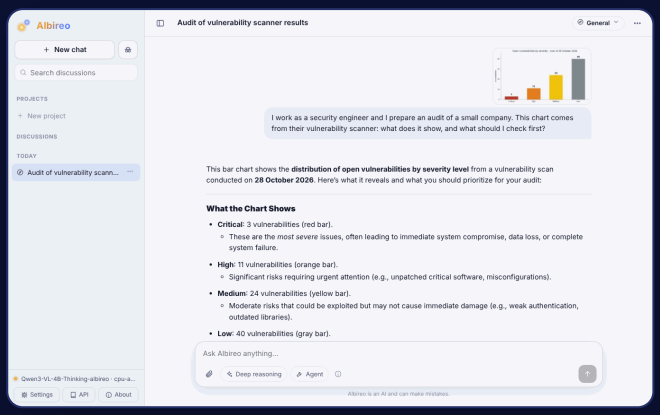
Albireo

Il vostro assistente IA privato, sul vostro computer

Albireo esegue i modelli linguistici direttamente sul vostro computer Windows 11. Le vostre conversazioni, i vostri documenti e i vostri modelli restano sulla macchina: nessun account da creare, nessun dato inviato all'esterno.

Versione **1.0.23** Piattaforma **Windows 11, x64**

Licenza **Boost Software License 1.0**



56,8
token al secondo sulla GPU di un portatile, contro 23,6 sul suo processore

22
modelli nel catalogo firmato, di 7 editori

5
motori: processori, NPU e GPU Intel, GPU NVIDIA

0
account, abbonamenti o dati inviati all'esterno

In breve

Chat
Conversazioni conservate nel browser, una modalità in incognito, una ricerca tra le conversazioni e il ragionamento dei modelli che ragionano mostrato mentre pensano.

Immagini
I modelli che leggono le immagini ricevono immagini PNG e JPEG, allegate, incollate o trascinate in una domanda.

Modalità agente
Il modello chiama strumenti che girano nella pagina: calcolatrice, JavaScript in sandbox, lettura e ricerca nei documenti, fino a 100 passi per risposta.

Lettura ad alta voce
Ogni risposta può essere letta ad alta voce dalla sintesi vocale fornita con Albireo, senza rete.

Modalità offline
Su un computer senza rete il servizio non apre alcuna connessione: il catalogo, i modelli e gli aggiornamenti arrivano da un pacchetto firmato portato su un disco.

Documenti
Word, Excel, PowerPoint, OpenDocument, PDF, RTF, HTML, testo, CSV, Markdown e codice sorgente, letti nel browser e allegati a un messaggio o a un progetto.

Progetti
Conversazioni raggruppate con istruzioni e documenti che formano il loro contesto comune, interrogato per significato e per parole.

Accesso web
Una ricerca in Wikipedia e la lettura di pagine e file PDF pubblici, facoltative e disattivate a ogni nuova conversazione.

Un'interfaccia per altri programmi
Chat, completamenti, embedding, voce e modelli seguono l'interfaccia OpenAI, con streaming e immagini, documentati in una pagina Swagger.

Inoltro a un motore esterno
Una conversazione può essere inoltrata a un motore esterno: qualsiasi interfaccia compatibile con OpenAI, il cloud di Ollama o Claude. I suoi testi sono pseudonimizzati da Pseudalis prima che lascino il computer, e un motore del computer o della rete locale resta utilizzabile offline.

Motori

Motore	Funziona su
`cpu`	Ogni processore x86-64
`cpu-avx2`	Processori con AVX2, FMA e F16C: Intel Haswell, AMD Zen e successivi
`cpu-avxvnni`	Processori con AVX-VNNI: Intel Core da Alder Lake, Intel Xeon da Sapphire Rapids, AMD Zen 5
`openvino`	NPU e GPU integrata dei processori Intel, tramite OpenVINO
`cuda`	Processori grafici NVIDIA, tramite CUDA

Velocità su hardware comune: le tecniche di Albireo

Albireo esegue i modelli con un motore proprio, progettato per i processori e le schede grafiche già in uso. Ogni tecnica qui sotto è misurata nel README del prodotto.

Pesi quantizzati, nuclei interi

I pesi restano nella mappatura in memoria del file GGUF, nella loro forma quantizzata. I nuclei interi calcolano 32 prodotti dove un vettore in virgola mobile ne calcola 8.

Lettura di un prompt: 216 token al secondo per Qwen3-VL 4B e più di 1.300 per Qwen3 0.6B (Intel Core Ultra 9 275HX, 24 thread); su due core, 230 contro 126 in virgola mobile (Qwen3 0.6B, AVX-VNNI)

Un motore per ogni tipo di processore

Motori portabile, AVX2 e AVX-VNNI per i processori, la NPU e la GPU integrata di Intel tramite OpenVINO, i processori grafici NVIDIA tramite CUDA. Il servizio li misura sulla macchina ed esegue il più veloce.

Scrittura: da 80,6 a 87,7 token al secondo per Qwen3 0.6B con il motore del processore (Intel Core Ultra 9 275HX, 24 thread)

Memoria estesa V²RAM

La memoria estesa della libreria lascia che le copie dei pesi superino la memoria della scheda, verso la memoria dell'host o un file, e le riporta prima del loro uso.

Più conversazioni insieme

Le richieste di tutti i client sono calcolate insieme a ogni passata, e le chiavi e i valori di una richiesta conclusa restano in cache per il turno successivo.

Dal 32 al 40% di tempo in meno per passo con quattro risposte scritte insieme

Decodifica anticipata

Una risposta scrive più token per passo quando il suo contesto li annuncia. Il modello tiene i token che avrebbe estratto da sé, così le risposte conservano la distribuzione del campionamento.

25,5 contro 19,7 token al secondo nella riscrittura di codice (Qwen3 0.6B, due core); 26,3 contro 23,6 su 24 thread (Qwen3-VL 4B)

Strati calcolati sulla scheda NVIDIA

Il motore CUDA tiene i pesi sulla scheda nei loro formati GGUF e vi calcola gli strati di un modello denso, attenzione compresa per i token generati.

56,8 contro 23,6 token al secondo (Qwen3-VL 4B, RTX 5070 Ti Laptop GPU contro 24 thread)

Una memoria di conversazione compatta

Chiavi e valori sono tenuti a 8 bit, in pagine condivise tra le conversazioni che iniziano con lo stesso testo.

34 byte contro 64 per gruppo di valori, con lo stesso token successivo su 210 posizioni

Un grafo OpenVINO sul processore e sulla GPU Intel

Con OpenVINO, un intero modello è calcolato come un unico grafo sul processore o sulla GPU integrata di Intel, e la NPU prende in carico prodotti matriciali. Il grafo legge più velocemente il primo passaggio di un prompt; il motore del processore resta più veloce nei passaggi successivi e in scrittura, e la misura dei motori, avviata dall'icona, mantiene la scelta migliore per la macchina.

Prima lettura di un prompt: 260 token al secondo con il grafo contro 216 con il motore del processore (Qwen3-VL 4B, Intel Core Ultra 9 275HX, 24 thread)

Caratteristiche

Formato dei modelli	GGUF quantizzato: da Q4_0 a Q8_0, da Q2_K a Q6_K e IQ4_NL
Architetture	llama, mistral, mistral3, granite, granitemoe, phi3, qwen2, qwen3, qwen3moe, qwen3vl, qwen3vlmoe, mixtral
Codificatori di immagini	Qwen3-VL, LLaVA-NeXT (EuroVLM) e Pixtral (Ministral 3, Devstral Small 2, Mistral Small 3.2)
Modelli forniti con Albireo	Qwen3-VL-4B-Instruct con il suo codificatore di immagini (2,95 GB) e Qwen3-Embedding-0.6B (0,64 GB)
Catalogo	22 modelli di 7 editori, firmato e verificato con gli hash
Memoria	8 GB per il modello installato per impostazione predefinita
Interfaccia locale	https://localhost:8443, compatibile con l'interfaccia OpenAI, documentata con Swagger
Formati di conversazione	ChatML, Mistral, Llama 3, Granite, Phi-4-mini e Foundation-Sec, scritti nativamente
Aggiornamenti	Feed firmato, programmi di installazione verificati per dimensione e SHA-256

Una sicurezza verificabile

Solo locale

Il servizio ascolta su 127.0.0.1 e rifiuta le richieste il cui host o la cui origine non è locale.

Un'autorità di certificazione del computer

Il certificato HTTPS proviene da un'autorità creata sul computer con una chiave ECDSA P-256, limitata a localhost da vincoli di nome.

Chiavi protette da Windows

La chiave privata del server e la chiave dell'inoltro sono cifrate con DPAPI e non sono mai scritte in chiaro.

Una catena di fornitura firmata

Il feed degli aggiornamenti e il catalogo dei modelli sono firmati con ECDSA P-256, e ogni file scaricato è verificato con il suo SHA-256.

Un'interfaccia aperta

Le rotte seguono l'interfaccia OpenAI e sono descritte con OpenAPI, così le librerie client esistenti funzionano senza modifiche.

Sintesi vocale isolata

Il sintetizzatore gira in un processo separato a bassa integrità, senza privilegi e limitato in memoria.

Libero e gratuito sotto la Boost Software License 1.0. © 2026 Cydral Technology, Francia.