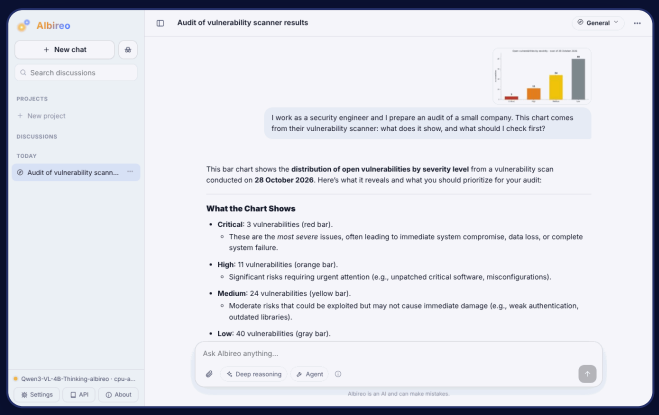
Albireo

Votre assistant IA privé, sur votre propre ordinateur

Albireo exécute les modèles de langage directement sur votre ordinateur Windows 11. Vos conversations, vos documents et vos modèles restent sur la machine : aucun compte à créer, aucune donnée envoyée à l'extérieur.

Version **1.0.23** Plateforme **Windows 11, x64**

Licence **Boost Software License 1.0**



56,8
jetons par seconde sur le GPU d'un portable, contre 23,6 sur son processeur

22
modèles dans le catalogue signé, de 7 éditeurs

5
moteurs : processeurs, NPU et GPU Intel, GPU NVIDIA

0
compte, abonnement ou donnée envoyée à l'extérieur

En bref

Discussion
Des discussions conservées dans le navigateur, un mode incognito, une recherche dans les discussions, et le raisonnement des modèles qui raisonnent affiché pendant leur réflexion.

Images
Les modèles qui lisent les images reçoivent des images PNG et JPEG, jointes, collées ou déposées dans une question.

Mode agent
Le modèle appelle des outils qui s'exécutent dans la page : calculatrice, JavaScript en bac à sable, lecture et recherche dans les documents, jusqu'à 100 étapes par réponse.

Lecture à voix haute
Chaque réponse peut être lue à voix haute par la synthèse vocale fournie avec Albireo, sans réseau.

Mode hors ligne
Sur un ordinateur sans réseau, le service n'ouvre aucune connexion : le catalogue, les modèles et les mises à jour viennent d'un paquet signé apporté sur un disque.

Documents
Word, Excel, PowerPoint, OpenDocument, PDF, RTF, HTML, texte, CSV, Markdown et code source, lus dans le navigateur et joints à un message ou à un projet.

Projets
Des discussions regroupées avec des instructions et des documents qui forment leur contexte commun, interrogé par le sens et par les mots.

Accès web
Une recherche dans Wikipédia et la lecture de pages et de fichiers PDF publics, en option, désactivées à chaque nouvelle discussion.

Une interface pour d'autres programmes
Les discussions, les complétions, les embeddings, la parole et les modèles suivent l'interface OpenAI, flux et images compris, documentés dans une page Swagger.

Relais vers un moteur externe
Une discussion peut être relayée vers un moteur externe : toute interface compatible OpenAI, le cloud d'Ollama ou Claude. Ses textes sont pseudonymisés par Pseudalis avant de quitter l'ordinateur, et un moteur de l'ordinateur ou du réseau local reste utilisable hors ligne.

Moteurs

Moteur	Fonctionne sur
`cpu`	Tout processeur x86-64
`cpu-avx2`	Processeurs avec AVX2, FMA et F16C : Intel Haswell, AMD Zen et suivants
`cpu-avxvnni`	Processeurs avec AVX-VNNI : Intel Core à partir d'Alder Lake, Intel Xeon à partir de Sapphire Rapids, AMD Zen 5
`openvino`	NPU et GPU intégré des processeurs Intel, par OpenVINO
`cuda`	Processeurs graphiques NVIDIA, par CUDA

Vitesse sur matériel courant : les techniques d'Albireo

Albireo exécute les modèles avec son propre moteur, conçu pour les processeurs et les cartes graphiques déjà en service. Chaque technique ci-dessous est mesurée dans le README du produit.

Poids quantifiés, noyaux entiers

Les poids restent dans la projection en mémoire du fichier GGUF, sous leur forme quantifiée. Les noyaux entiers calculent 32 produits là où un vecteur de flottants en calcule 8.

Lecture d'un prompt : 216 jetons par seconde pour Qwen3-VL 4B et plus de 1 300 pour Qwen3 0.6B (Intel Core Ultra 9 275HX, 24 fils) ; sur deux cœurs, 230 contre 126 en flottants (Qwen3 0.6B, AVX-VNNI)

Un moteur par type de processeur

Des moteurs portable, AVX2 et AVX-VNNI pour les processeurs, le NPU et le GPU intégré d'Intel par OpenVINO, les processeurs graphiques NVIDIA par CUDA. Le service les mesure sur la machine et exécute le plus rapide.

Écriture : 80,6 à 87,7 jetons par seconde pour Qwen3 0.6B avec le moteur du processeur (Intel Core Ultra 9 275HX, 24 fils)

Mémoire étendue V²RAM

La mémoire étendue de la bibliothèque laisse les copies des poids dépasser la mémoire de la carte, vers la mémoire de l'hôte ou un fichier, et les ramène avant leur emploi.

Plusieurs conversations à la fois

Les requêtes de tous les clients sont calculées ensemble à chaque passe, et les clés et les valeurs d'une requête terminée restent en cache pour le tour suivant.

32 à 40 % de temps en moins par pas avec quatre réponses écrites ensemble

Décodage préemptif

Une réponse écrit plusieurs jetons par pas lorsque son contexte les annonce. Le modèle garde les jetons qu'il aurait tirés lui-même, si bien que les réponses conservent la distribution de l'échantillonnage.

25,5 contre 19,7 jetons par seconde à la réécriture de code (Qwen3 0.6B, deux cœurs) ; 26,3 contre 23,6 sur 24 fils (Qwen3-VL 4B)

Des couches calculées sur la carte NVIDIA

Le moteur CUDA garde les poids sur la carte dans leurs formats GGUF et y calcule les couches d'un modèle dense, attention comprise pour les jetons générés.

56,8 contre 23,6 jetons par seconde (Qwen3-VL 4B, RTX 5070 Ti Laptop GPU contre 24 fils)

Une mémoire de conversation compacte

Les clés et les valeurs sont gardées en 8 bits, dans des pages partagées entre les conversations qui commencent par le même texte.

34 octets contre 64 par groupe de valeurs, avec le même jeton suivant sur 210 positions

Un graphe OpenVINO sur le processeur et le GPU Intel

Avec OpenVINO, un modèle entier est calculé comme un seul graphe sur le processeur ou sur le GPU intégré d'Intel, et le NPU prend en charge des produits matriciels. Le graphe lit plus vite la première passe d'un prompt ; le moteur du processeur reste plus rapide sur les passes suivantes et à l'écriture, et la mesure des moteurs, lancée depuis l'icône, retient le meilleur choix pour la machine.

Première lecture d'un prompt : 260 jetons par seconde avec le graphe contre 216 avec le moteur du processeur (Qwen3-VL 4B, Intel Core Ultra 9 275HX, 24 fils)

Caractéristiques

Format des modèles	GGUF quantifié : Q4_0 à Q8_0, Q2_K à Q6_K et IQ4_NL
Architectures	llama, mistral, mistral3, granite, granitemoe, phi3, qwen2, qwen3, qwen3moe, qwen3vl, qwen3vlmoe, mixtral
Encodeurs d'images	Qwen3-VL, LLaVA-NeXT (EuroVLM) et Pixtral (Ministral 3, Devstral Small 2, Mistral Small 3.2)
Modèles fournis avec Albireo	Qwen3-VL-4B-Instruct avec son encodeur d'images (2,95 Go) et Qwen3-Embedding-0.6B (0,64 Go)
Catalogue	22 modèles de 7 éditeurs, signé et vérifié par empreintes
Mémoire	8 Go pour le modèle installé par défaut
Interface locale	https://localhost:8443, compatible avec l'interface OpenAI, documentée avec Swagger
Formats de conversation	ChatML, Mistral, Llama 3, Granite, Phi-4-mini et Foundation-Sec, écrits nativement
Mises à jour	Flux signé, installateurs vérifiés par leur taille et leur SHA-256

Une sécurité auditable

Local seulement

Le service écoute sur 127.0.0.1 et refuse les requêtes dont l'hôte ou l'origine n'est pas local.

Une autorité de certification propre à l'ordinateur

Le certificat HTTPS vient d'une autorité créée sur l'ordinateur avec une clé ECDSA P-256, restreinte à localhost par des contraintes de nom.

Des clés protégées par Windows

La clé privée du serveur et la clé du relais sont chiffrées avec DPAPI et ne sont jamais écrites en clair.

Une chaîne d'approvisionnement signée

Le flux de mises à jour et le catalogue des modèles sont signés en ECDSA P-256, et chaque fichier téléchargé est vérifié par son SHA-256.

Une interface ouverte

Les routes suivent l'interface OpenAI et sont décrites avec OpenAPI, si bien que les bibliothèques clientes existantes fonctionnent telles quelles.

Une synthèse vocale isolée

Le synthétiseur s'exécute dans un processus séparé de faible intégrité, sans privilèges et borné en mémoire.

En savoir plus
cydral-technology.eu/fr/products/albireo/

Téléchargements
cydral-technology.eu/fr/downloads/

Contact
cydral-technology.eu/fr/contact/

Libre et gratuit sous la Boost Software License 1.0. © 2026 Cydral Technology, France.