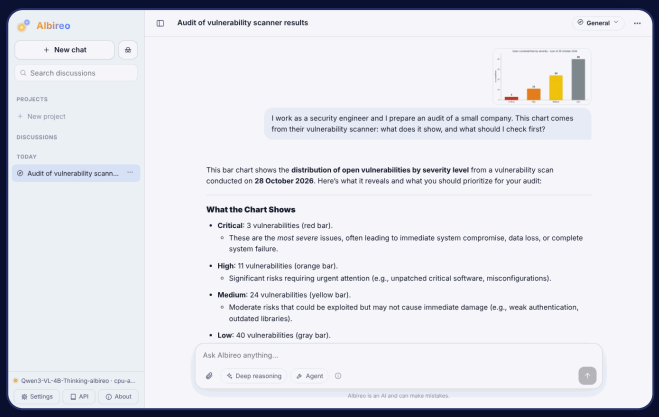
Albireo

Su asistente de IA privado, en su propio equipo

Albireo ejecuta los modelos de lenguaje directamente en su equipo con Windows 11. Sus conversaciones, sus documentos y sus modelos permanecen en la máquina: ninguna cuenta que crear, ningún dato enviado al exterior.

Versión **1.0.23** Plataforma **Windows 11, x64**

Licencia **Boost Software License 1.0**



56,8
tokens por segundo en la GPU de un portátil, frente a 23,6 en su procesador

22
modelos en el catálogo firmado, de 7 editores

5
motores: procesadores, NPU y GPU de Intel, GPU de NVIDIA

0
cuentas, suscripciones o datos enviados al exterior

En resumen

Conversación
Conversaciones guardadas en el navegador, un modo incógnito, una búsqueda entre las conversaciones y el razonamiento de los modelos que razonan mostrado mientras piensan.

Imágenes
Los modelos que leen imágenes reciben imágenes PNG y JPEG, adjuntas, pegadas o arrastradas a una pregunta.

Modo agente
El modelo llama a herramientas que se ejecutan en la página: calculadora, JavaScript aislado, lectura y búsqueda en los documentos, hasta 100 pasos por respuesta.

Lectura en voz alta
Cada respuesta puede leerse en voz alta con la síntesis de voz que acompaña a Albireo, sin red.

Modo sin conexión
En un equipo sin red el servicio no abre ninguna conexión: el catálogo, los modelos y las actualizaciones llegan en un paquete firmado traído en un disco.

Documentos
Word, Excel, PowerPoint, OpenDocument, PDF, RTF, HTML, texto, CSV, Markdown y código fuente, leídos en el navegador y adjuntados a un mensaje o a un proyecto.

Proyectos
Conversaciones agrupadas con instrucciones y documentos que forman su contexto común, consultado por significado y por palabras.

Acceso web
Una búsqueda en Wikipedia y la lectura de páginas y archivos PDF públicos, opcionales y desactivadas en cada conversación nueva.

Una interfaz para otros programas
Las conversaciones, las finalizaciones, los embeddings, la voz y los modelos siguen la interfaz de OpenAI, con transmisión e imágenes, documentados en una página Swagger.

Reenvío a un motor externo
Una conversación puede reenviarse a un motor externo: cualquier interfaz compatible con OpenAI, la nube de Ollama o Claude. Pseudalis seudonimiza sus textos antes de que salgan del equipo, y un motor del equipo o de la red local sigue disponible sin conexión.

Motores

Motor	Funciona en
`cpu`	Cualquier procesador x86-64
`cpu-avx2`	Procesadores con AVX2, FMA y F16C: Intel Haswell, AMD Zen y posteriores
`cpu-avxvnni`	Procesadores con AVX-VNNI: Intel Core desde Alder Lake, Intel Xeon desde Sapphire Rapids, AMD Zen 5
`openvino`	NPU y GPU integrada de los procesadores Intel, mediante OpenVINO
`cuda`	Procesadores gráficos de NVIDIA, mediante CUDA

Velocidad en hardware corriente: las técnicas de Albireo

Albireo ejecuta los modelos con un motor propio, diseñado para los procesadores y las tarjetas gráficas ya en servicio. Cada técnica siguiente está medida en el README del producto.

Pesos cuantizados, núcleos enteros

Los pesos permanecen en la proyección en memoria del archivo GGUF, en su forma cuantizada. Los núcleos enteros calculan 32 productos donde un vector en coma flotante calcula 8.

Lectura de un prompt: 216 tokens por segundo con Qwen3-VL 4B y más de 1300 con Qwen3 0.6B (Intel Core Ultra 9 275HX, 24 hilos); con dos núcleos, 230 frente a 126 en coma flotante (Qwen3 0.6B, AVX-VNNI)

Un motor para cada tipo de procesador

Motores portable, AVX2 y AVX-VNNI para los procesadores, la NPU y la GPU integrada de Intel mediante OpenVINO, los procesadores gráficos de NVIDIA mediante CUDA. El servicio los mide en la máquina y ejecuta el más rápido.

Escritura: de 80,6 a 87,7 tokens por segundo con Qwen3 0.6B y el motor del procesador (Intel Core Ultra 9 275HX, 24 hilos)

Memoria ampliada V²RAM

La memoria ampliada de la biblioteca deja que las copias de los pesos superen la memoria de la tarjeta, hacia la memoria del anfitrión o un archivo, y las trae de vuelta antes de su uso.

Varias conversaciones a la vez

Las solicitudes de todos los clientes se calculan juntas en cada pasada, y las claves y los valores de una solicitud terminada quedan en caché para el turno siguiente.

Entre un 32 y un 40 % menos de tiempo por paso con cuatro respuestas escritas a la vez

Decodificación anticipada

Una respuesta escribe varios tokens por paso cuando su contexto los anuncia. El modelo conserva los tokens que habría extraído por sí mismo, de modo que las respuestas mantienen la distribución del muestreo.

25,5 frente a 19,7 tokens por segundo al reescribir código (Qwen3 0.6B, dos núcleos); 26,3 frente a 23,6 con 24 hilos (Qwen3-VL 4B)

Capas calculadas en la tarjeta NVIDIA

El motor CUDA mantiene los pesos en la tarjeta en sus formatos GGUF y calcula allí las capas de un modelo denso, atención incluida para los tokens generados.

56,8 frente a 23,6 tokens por segundo (Qwen3-VL 4B, RTX 5070 Ti Laptop GPU frente a 24 hilos)

Una memoria de conversación compacta

Las claves y los valores se guardan en 8 bits, en páginas compartidas entre las conversaciones que empiezan con el mismo texto.

34 bytes frente a 64 por grupo de valores, con el mismo token siguiente en 210 posiciones

Un grafo de OpenVINO en el procesador y la GPU de Intel

Con OpenVINO, un modelo entero se calcula como un único grafo en el procesador o en la GPU integrada de Intel, y la NPU se encarga de productos matriciales. El grafo lee más rápido la primera pasada de un prompt; el motor del procesador sigue siendo más rápido en las pasadas siguientes y en la escritura, y la medición de los motores, lanzada desde el icono, conserva la mejor opción para la máquina.

Primera lectura de un prompt: 260 tokens por segundo con el grafo frente a 216 con el motor del procesador (Qwen3-VL 4B, Intel Core Ultra 9 275HX, 24 hilos)

Características

Formato de los modelos	GGUF cuantizado: de Q4_0 a Q8_0, de Q2_K a Q6_K e IQ4_NL
Arquitecturas	llama, mistral, mistral3, granite, granitemoe, phi3, qwen2, qwen3, qwen3moe, qwen3vl, qwen3vlmoe, mixtral
Codificadores de imágenes	Qwen3-VL, LLaVA-NeXT (EuroVLM) y Pixtral (Ministral 3, Devstral Small 2, Mistral Small 3.2)
Modelos que acompañan a Albireo	Qwen3-VL-4B-Instruct con su codificador de imágenes (2,95 GB) y Qwen3-Embedding-0.6B (0,64 GB)
Catálogo	22 modelos de 7 editores, firmado y comprobado con sus hashes
Memoria	8 GB para el modelo instalado por defecto
Interfaz local	https://localhost:8443, compatible con la interfaz de OpenAI, documentada con Swagger
Formatos de conversación	ChatML, Mistral, Llama 3, Granite, Phi-4-mini y Foundation-Sec, escritos de forma nativa
Actualizaciones	Canal firmado, instaladores comprobados por tamaño y SHA-256

Una seguridad auditable

Solo local

El servicio escucha en 127.0.0.1 y rechaza las solicitudes cuyo anfitrión u origen no es local.

Una autoridad de certificación del equipo

El certificado HTTPS procede de una autoridad creada en el equipo con una clave ECDSA P-256, limitada a localhost por restricciones de nombre.

Claves protegidas por Windows

La clave privada del servidor y la clave del reenvío se cifran con DPAPI y nunca se escriben en claro.

Una cadena de suministro firmada

El canal de actualizaciones y el catálogo de modelos se firman con ECDSA P-256, y cada archivo descargado se comprueba con su SHA-256.

Una interfaz abierta

Las rutas siguen la interfaz de OpenAI y se describen con OpenAPI, de modo que las bibliotecas cliente existentes funcionan sin cambios.

Síntesis de voz aislada

El sintetizador se ejecuta en un proceso separado de baja integridad, sin privilegios y acotado en memoria.

Más información
cydral-technology.eu/es/products/albireo/

Descargas
cydral-technology.eu/es/downloads/

Contacto
cydral-technology.eu/es/contact/

Libre y gratuito bajo la Boost Software License 1.0. © 2026 Cydral Technology, Francia.