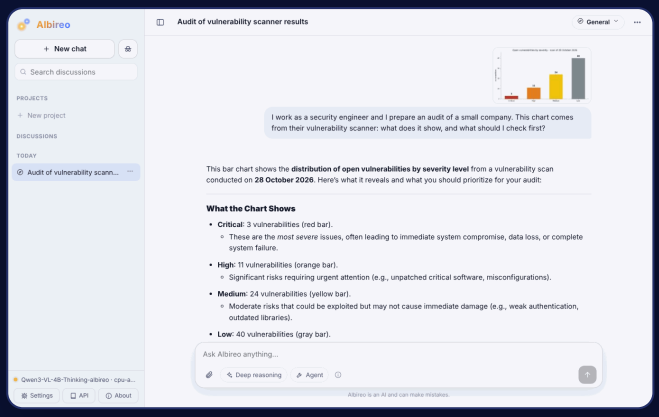
Albireo

あなた自身のコンピューターで動く、プライベートな *AI* アシスタント

Albireo は、あなたの Windows 11 コンピューターの上で直接、言語モデルを動かします。会話、文書、モデルはそのマシンにとどまります。アカウントの作成は不要で、データが外部に送信されることもありません。

バージョン **1.0.23**　　プラットフォーム **Windows 11, x64**

ライセンス **Boost Software License 1.0**



56.8
ノート PC の GPU での毎秒トークン数。プロセッサーでは 23.6

22
署名付きカタログのモデル数。提供元は 7

5
エンジン：プロセッサー、Intel の NPU と GPU、NVIDIA GPU

0
アカウント、サブスクリプション、外部に送るデータ

概要

チャット
会話はブラウザーに保存されます。シークレットモード、会話の横断検索に対応し、推論するモデルでは考えている間の推論過程を表示します。

画像
画像を読めるモデルは、添付、貼り付け、ドロップされた PNG と JPEG の画像を受け取ります。

エージェントモード
モデルがページ内で動くツールを呼び出します。電卓、サンドボックス化された JavaScript、文書の読み取りと検索。1 回の回答につき最大 100 ステップです。

読み上げ
各回答は、Albireo に付属する音声合成でネットワークなしに読み上げられます。

オフラインモード
ネットワークのないコンピューターでは、サービスは一切接続を開きません。カタログ、モデル、アップデートは、ディスクで持ち込む署名付きバンドルから取得します。

文書
Word、Excel、PowerPoint、OpenDocument、PDF、RTF、HTML、テキスト、CSV、Markdown、ソースコードをブラウザーで読み込み、メッセージやプロジェクトに添付できます。

プロジェクト
会話を、共有コンテキストとなる指示や文書とともにまとめ、意味と語句の両方で検索できます。

ウェブアクセス
任意で使える Wikipedia 検索と、公開ページや PDF ファイルの読み取り。新しい会話では毎回オフから始まります。

他のプログラム向けのインターフェース
チャット補完、補完、埋め込み、音声、モデル一覧が OpenAI インターフェースに準拠し、ストリーミングと画像に対応。Swagger ページで文書化されています。

外部エンジンへの中継
会話を外部エンジンに中継できます。OpenAI 互換のインターフェース、Ollama のクラウド、Claude に対応しています。テキストはコンピューターを出る前に Pseudalis が仮名化し、このコンピューターやローカルネットワーク上のエンジンはオフラインでも使えます。

エンジン

エンジン	動作環境
`cpu`	すべての x86-64 プロセッサー
`cpu-avx2`	AVX2、FMA、F16C 対応プロセッサー：Intel Haswell、AMD Zen 以降
`cpu-avxvnni`	AVX-VNNI 対応プロセッサー：Alder Lake 以降の Intel Core、Sapphire Rapids 以降の Intel Xeon、AMD Zen 5
`openvino`	Intel プロセッサーの NPU と内蔵 GPU（OpenVINO 経由）
`cuda`	NVIDIA グラフィックスプロセッサー（CUDA 経由）

一般的なハードウェアでの速さ：Albireo の技術

Albireo は、すでに稼働しているプロセッサーとグラフィックカードのために設計された独自のエンジンでモデルを動かします。以下の各技術の測定値は、製品の README に記載しています。

量子化カーネル

重みは量子化された形のまま、GGUF ファイルのメモリマッピング上に置かれます。浮動小数点のベクトルが 8 個の積を計算するところを、整数カーネルは 32 個計算します。

プロンプトの読み込み：Qwen3-VL 4B で毎秒 216 トークン、Qwen3 0.6B で毎秒 1,300 トークン超（Intel Core Ultra 9 275HX、24 スレッド）。2 コアでは毎秒 230 トークン、浮動小数点では 126（Qwen3 0.6B、AVX-VNNI）

プロセッサーの種類ごとのエンジン

プロセッサー向けの汎用、AVX2、AVX-VNNI の各エンジン、OpenVINO 経由の Intel の NPU と内蔵 GPU、CUDA 経由の NVIDIA グラフィックスプロセッサー。サービスがそのマシンで計測し、最も速いものを使います。

回答の書き出し：プロセッサーエンジンの Qwen3 0.6B で毎秒 80.6〜87.7 トークン（Intel Core Ultra 9 275HX、24 スレッド）

V²RAM 拡張メモリ

ライブラリのデバイス拡張メモリにより、重みのコピーはカードのメモリを超えてホストメモリやファイルに置かれ、使われる前に先回りして戻されます。

複数の会話を同時に

すべてのクライアントのリクエストを各フォワードパスでまとめて計算し、完了したリクエストのキーとバリューは次のターンのためにキャッシュに残します。

4 つの回答を同時に書くと 1 ステップあたりの時間が 32〜40% 短縮

ドラフトトークン

文脈から続きが予測できるとき、回答は 1 ステップで複数のトークンを書きます。投機的デコードの一種です。モデルは自分でも選んだはずのトークンだけを採用するため、回答はサンプリングの分布を保ちます。

コードの書き直しで毎秒 25.5 トークン、従来は 19.7（Qwen3 0.6B、2 コア）。24 スレッドでは 26.3、従来は 23.6（Qwen3-VL 4B）

NVIDIA カード上で計算する層

CUDA エンジンは重みを GGUF 形式のままカード上に保持し、密なモデルの各層を、生成トークンの注意機構も含めてカード上で計算します。

毎秒 56.8 トークン、従来は 23.6（Qwen3-VL 4B、RTX 5070 Ti Laptop GPU と 24 スレッドの比較）

コンパクトな会話メモリ

キーとバリューは 8 ビットで保持され、同じテキストで始まる会話の間でページ単位で共有されます。

値のグループあたり 64 バイトが 34 バイトに。210 位置で次トークンは同一

Intel のプロセッサーと GPU 上の OpenVINO グラフ

OpenVINO では、モデル全体を 1 つのグラフとしてプロセッサーまたは Intel の内蔵 GPU で計算し、NPU は行列積を受け持ちます。グラフはプロンプトの最初のパスをより速く読み込みます。後続のパスと回答の書き出しではプロセッサーエンジンのほうが速く、アイコンから実行するエンジンの計測が、そのマシンに最適な選択を残します。

プロンプトの最初の読み込み：グラフで毎秒 260 トークン、プロセッサーエンジンでは 216（Qwen3-VL 4B、Intel Core Ultra 9 275HX、24 スレッド）

仕様

モデル形式	量子化 GGUF:Q4_0〜Q8_0、Q2_K〜Q6_K、IQ4_NL
アーキテクチャ	llama, mistral, mistral3, granite, granitemoe, phi3, qwen2, qwen3, qwen3moe, qwen3vl, qwen3vlmoe, mixtral
画像エンコーダー	Qwen3-VL、LLaVA-NeXT（EuroVLM）、Pixtral（Ministral 3、Devstral Small 2、Mistral Small 3.2）
Albireo に付属するモデル	Qwen3-VL-4B-Instruct とその画像エンコーダー（2.95 GB）、Qwen3-Embedding-0.6B(0.64 GB)
カタログ	7 の提供元による 22 モデル。署名付きで、ハッシュと照合済み
メモリ	標準でインストールされるモデルで 8 GB
ローカルインターフェース	https://localhost:8443、OpenAI インターフェース互換、Swagger で文書化
会話フォーマット	ChatML、Mistral、Llama 3、Granite、Phi-4-mini、Foundation-Sec をネイティブに実装
アップデート	署名付きフィード。インストーラーはサイズと SHA-256 で検証

監査できるセキュリティ

ローカル限定

サービスは 127.0.0.1 で待ち受け、ホストまたはオリジンがローカルでないリクエストを拒否します。

コンピューター固有の認証局

HTTPS 証明書は、そのコンピューター上で ECDSA P-256 鍵を使って作られた認証局が発行し、名前制約によって localhost に限定されます。

Windows が保護する鍵

サーバーの秘密鍵と中継用の鍵は DPAPI で暗号化され、平文で書き込まれることはありません。

署名されたサプライチェーン

アップデートフィードとモデルカタログは ECDSA P-256 で署名され、ダウンロードされた各ファイルは SHA-256 と照合されます。

オープンなインターフェース

ルートは OpenAI インターフェースに準拠し、OpenAPI で記述されているため、既存のクライアントライブラリがそのまま使えます。

隔離された音声合成

合成器は、特権を持たずメモリを制限した、低整合性の別プロセスで動作します。

Boost Software License 1.0 による無料のオープンソース。© 2026 Cydral Technology、フランス。