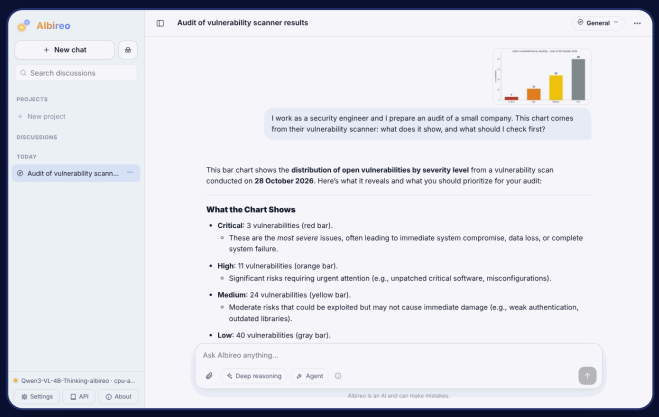
Albireo

您的私有 *AI* 助手，就运行在您自己的电脑上

Albireo 直接在您的 Windows 11 电脑上运行语言模型。您的对话、文档和模型都留在这台机器上：无需创建账户，任何数据都不会发送到外部。

版本 **1.0.23**　　平台 **Windows 11, x64**

许可证 **Boost Software License 1.0**



56.8
笔记本 GPU 上每秒词元数，其处理器上为 23.6

22
签名目录中的模型数，来自 7 家发布方

5
个引擎：处理器、Intel NPU 与 GPU、NVIDIA GPU

0
账户、订阅或外发数据

概览

对话
对话保存在浏览器中，提供无痕模式和跨对话搜索；会推理的模型在思考时会显示其推理过程。

图像
能读图的模型接受 PNG 和 JPEG 图片，可以附加、粘贴或拖入问题。

智能体模式
模型调用在页面内运行的工具：计算器、沙箱化的 JavaScript、文档的读取与搜索，每个回答最多 100 步。

朗读
每个回答都可以由 Albireo 自带的语音合成朗读，无需网络。

离线模式
在没有网络的电脑上，服务不会发起任何连接：目录、模型和更新都来自通过磁盘带入的签名离线包。

文档
Word、Excel、PowerPoint、OpenDocument、PDF、RTF、HTML、文本、CSV、Markdown 和源代码，在浏览器中读取，并可附加到消息或项目。

项目
把对话与指令、文档归为一组，构成共享上下文，可按语义和关键词检索。

网络访问
可选的维基百科搜索，以及公开网页和 PDF 文件的读取；每个新对话默认关闭。

面向其他程序的接口
聊天补全、文本补全、嵌入、语音和模型列表均遵循 OpenAI 接口，支持流式输出和图像，并以 Swagger 页面提供文档。

转发到外部引擎
对话可以转发到外部引擎：任何兼容 OpenAI 的接口、Ollama 云或 Claude。其文本在离开电脑之前先由 Pseudalis 做假名化处理；本机或局域网中的引擎在离线时仍可使用。

引擎

引擎	运行于
`cpu`	所有 x86-64 处理器
`cpu-avx2`	支持 AVX2、FMA 和 F16C 的处理器：Intel Haswell、AMD Zen 及更新型号
`cpu-avxvnni`	支持 AVX-VNNI 的处理器：Alder Lake 起的 Intel Core、Sapphire Rapids 起的 Intel Xeon、AMD Zen 5
`openvino`	Intel 处理器的 NPU 与集成 GPU，通过 OpenVINO
`cuda`	NVIDIA 图形处理器，通过 CUDA

常见硬件上的速度：Albireo 的技术

Albireo 用自己的引擎运行模型，这个引擎专为已在使用的处理器和显卡而设计。下面每项技术的实测数据都写在产品的 README 中。

量化内核

权重以量化形式留在 GGUF 文件的内存映射中。整数内核一次计算 32 个乘积，而浮点向量一次只算 8 个。

读取提示：*Qwen3-VL 4B* 每秒 *216* 个词元，*Qwen3 0.6B* 每秒超过 *1,300* 个（*Intel Core Ultra 9 275HX*，*24* 线程）；在两个核心上为每秒 *230* 个，浮点运算为 *126* 个（*Qwen3 0.6B*，*AVX-VNNI*）

为每类处理器配备引擎

面向处理器的通用、AVX2 和 AVX-VNNI 引擎，通过 OpenVINO 使用的 Intel NPU 与集成 GPU，通过 CUDA 使用的 NVIDIA 图形处理器。服务会在本机上实测，并运行最快的那个。

写出回答：*Qwen3 0.6B* 使用处理器引擎时每秒 *80.6* 到 *87.7* 个词元（*Intel Core Ultra 9 275HX*，*24* 线程）

V²RAM 扩展内存

核心库的设备扩展内存让权重副本可以超出显卡显存，进入主机内存或文件，并在使用之前提前取回。

同时进行多个对话

所有客户端的请求在每次前向计算中一并处理，已完成请求的键和值会保留在缓存中，供下一轮使用。

四个回答同时生成时，每步耗时减少 *32%* 到 *40%*

草稿词元

当上下文足以预示后续内容时，回答每一步会写出多个词元，这是推测解码的一种形式。模型只保留它自己本来也会采样到的词元，因此回答保持采样分布不变。

重写代码时每秒 *25.5* 个词元，对比 *19.7* 个（*Qwen3 0.6B*，两个核心）；*24* 线程上为 *26.3* 对比 *23.6*（*Qwen3-VL 4B*）

在 NVIDIA 显卡上计算的层

CUDA 引擎把权重以 GGUF 格式保留在显卡上，并在显卡上计算稠密模型的各层，生成词元时的注意力也包括在内。

每秒 *56.8* 个词元，对比 *23.6* 个（*Qwen3-VL 4B*，*RTX 5070 Ti Laptop GPU* 对比 *24* 线程）

紧凑的对话记忆

键和值以 8 位保存，并按页在开头文本相同的对话之间共享。

每组值 *34* 字节，对比 *64* 字节，且 *210* 个位置上的下一个词元完全相同

Intel 处理器与 GPU 上的 OpenVINO 计算图

借助 OpenVINO，整个模型作为一张计算图在处理器或 Intel 集成 GPU 上计算，NPU 则承担矩阵乘积。计算图读取提示的第一遍更快；处理器引擎在后续各遍和写出回答时仍然更快。从托盘图标启动的引擎测速会为这台机器保留最佳选择。

首次读取提示：计算图每秒 *260* 个词元，处理器引擎为 *216* 个（*Qwen3-VL 4B*，*Intel Core Ultra 9 275HX*，*24* 线程）

规格

模型格式	量化 GGUF：Q4_0 至 Q8_0、Q2_K 至 Q6_K 以及 IQ4_NL
架构	llama, mistral, mistral3, granite, granitemoe, phi3, qwen2, qwen3, qwen3moe, qwen3vl, qwen3vlmoe, mixtral
图像编码器	Qwen3-VL、LLaVA-NeXT（EuroVLM）和 Pixtral（Ministral 3、Devstral Small 2、Mistral Small 3.2）
Albireo 自带的模型	Qwen3-VL-4B-Instruct 及其图像编码器（2.95 GB），以及 Qwen3-Embedding-0.6B(0.64 GB)
目录	来自 7 家发布方的 22 个模型，均经签名并按哈希校验
内存	默认安装的模型需要 8 GB
本地接口	https://localhost:8443，兼容 OpenAI 接口，以 Swagger 提供文档
对话格式	ChatML、Mistral、Llama 3、Granite、Phi-4-mini 和 Foundation-Sec，均为原生实现
更新	签名的更新源，安装程序按大小和 SHA-256 校验

可以审计的安全性

仅限本地

服务监听 127.0.0.1，并拒绝主机名或来源不是本地的请求。

本机的证书颁发机构

HTTPS 证书由在本机上以 ECDSA P-256 密钥创建的机构签发，并通过名称约束限定为 localhost。

由 Windows 保护的密钥

服务器私钥和转发密钥都用 DPAPI 加密，绝不以明文写入。

签名的供应链

更新源和模型目录以 ECDSA P-256 签名，每个下载的文件都按其 SHA-256 校验。

开放的接口

各路由遵循 OpenAI 接口并以 OpenAPI 描述，现有客户端库无需修改即可使用。

隔离的语音合成

合成器运行在独立的低完整性进程中，没有特权，且内存受限。

免费且开源，采用 Boost Software License 1.0。© 2026 Cydral Technology，法国。