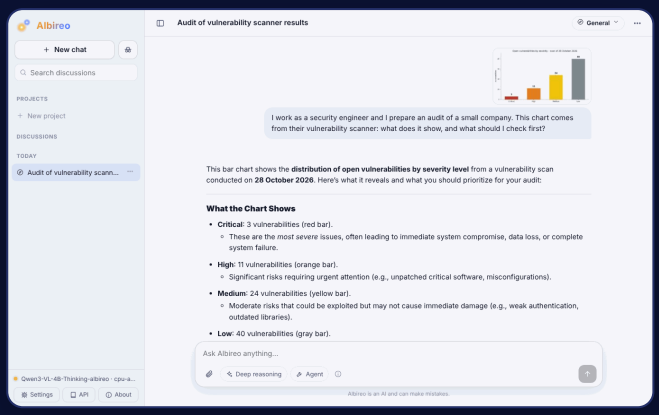
Albireo

Ваш личный ИИ-ассистент, работающий на вашем компьютере

Albireo запускает языковые модели прямо на вашем компьютере с Windows 11. Ваши беседы, документы и модели остаются на этой машине: учётная запись не нужна, и никакие данные не отправляются наружу.

Версия **1.0.23** Платформа **Windows 11, x64**

Лицензия **Boost Software License 1.0**



56,8
токена в секунду на GPU ноутбука против 23,6 на его процессоре

22
моделей в подписанном каталоге, от 7 издателей

5
движков: процессоры, NPU и GPU Intel, GPU NVIDIA

0
учётных записей, подписок и данных, отправленных наружу

Коротко

Чат
Беседы хранятся в браузере; есть режим инкогнито, поиск по беседам и показ рассуждений у моделей, которые рассуждают, пока они думают.

Изображения
Модели, читающие изображения, принимают картинки PNG и JPEG: прикреплённые, вставленные или перетащенные в вопрос.

Режим агента
Модель вызывает инструменты, работающие на странице: калькулятор, JavaScript в песочнице, чтение и поиск по документам, до 100 шагов на ответ.

Чтение вслух
Каждый ответ может быть прочитан вслух синтезатором речи, входящим в Albireo, без сети.

Автономный режим
На компьютере без сети служба не открывает ни одного соединения: каталог, модели и обновления берутся из подписанного пакета, принесённого на диске.

Документы
Word, Excel, PowerPoint, OpenDocument, PDF, RTF, HTML, текст, CSV, Markdown и исходный код читаются в браузере и прикрепляются к сообщению или проекту.

Проекты
Беседы, объединённые с инструкциями и документами, которые образуют их общий контекст, с поиском по смыслу и по словам.

Доступ к вебу
Необязательный поиск по Википедии и чтение публичных страниц и PDF-файлов; в каждой новой беседе выключен.

Интерфейс для других программ
Чат, дополнение текста, эмбеддинги, речь и список моделей следуют интерфейсу OpenAI, с потоковой выдачей и изображениями, и описаны на странице Swagger.

Ретрансляция на внешний движок
Беседу можно передать внешнему движку: любому интерфейсу, совместимому с OpenAI, облаку Ollama или Claude. Её тексты псевдонимизирует Pseudalis до того, как они покинут компьютер, а движок этого компьютера или локальной сети остаётся доступным и без подключения.

Движки

Движок	Работает на
`cpu`	Любой процессор x86-64
`cpu-avx2`	Процессоры с AVX2, FMA и F16C: Intel Haswell, AMD Zen и новее
`cpu-avxvnni`	Процессоры с AVX-VNNI: Intel Core начиная с Alder Lake, Intel Xeon начиная с Sapphire Rapids, AMD Zen 5
`openvino`	NPU и встроенный GPU процессоров Intel, через OpenVINO
`cuda`	Графические процессоры NVIDIA, через CUDA

Скорость на обычном оборудовании: технологии Albireo

Albireo запускает модели на собственном движке, созданном для процессоров и видеокарт, которые уже находятся в работе. Каждый приём ниже измерен в README продукта.

Квантованные ядра

Веса остаются в отображении файла GGUF в память, в квантованном виде. Целочисленные ядра считают 32 произведения там, где вектор чисел с плавающей точкой считает 8.

Чтение подсказки: 216 токенов в секунду для Qwen3-VL 4B и более 1 300 для Qwen3 0.6B (Intel Core Ultra 9 275HX, 24 потока); на двух ядрах 230 против 126 на числах с плавающей точкой (Qwen3 0.6B, AVX-VNNI)

Движок для каждого типа процессора

Переносимый движок, движки AVX2 и AVX-VNNI для процессоров, NPU и встроенный GPU Intel через OpenVINO, графические процессоры NVIDIA через CUDA. Служба замеряет их на машине и запускает самый быстрый.

Запись ответа: от 80,6 до 87,7 токена в секунду для Qwen3 0.6B на процессорном движке (Intel Core Ultra 9 275HX, 24 потока)

Расширенная память V²RAM

Расширенная память устройства из библиотеки позволяет копиям весов выходить за пределы памяти карты, в память хоста или в файл, и возвращает их до того, как они понадобятся.

Несколько бесед одновременно

Запросы всех клиентов вычисляются вместе в каждом прямом проходе, а ключи и значения завершённого запроса остаются в кэше для следующей реплики.

На 32–40 % меньше времени на шаг при четырёх ответах, написанных вместе

Черновые токены

Ответ пишет несколько токенов за шаг, когда контекст их предсказывает; это форма спекулятивного декодирования. Модель оставляет только те токены, которые выбрала бы сама, так что ответы сохраняют распределение выборки.

25,5 против 19,7 токена в секунду при переписывании кода (Qwen3 0.6B, два ядра); 26,3 против 23,6 на 24 потоках (Qwen3-VL 4B)

Слои, вычисляемые на карте NVIDIA

Движок CUDA держит веса на карте в форматах GGUF и вычисляет там слои плотной модели, включая внимание для генерируемых токенов.

56,8 против 23,6 токена в секунду (Qwen3-VL 4B, RTX 5070 Ti Laptop GPU против 24 потоков)

Компактная память беседы

Ключи и значения хранятся в 8 битах, в страницах, общих для бесед, которые начинаются с одного и того же текста.

34 байта против 64 на группу значений, с тем же следующим токеном на 210 позициях

Граф OpenVINO на процессоре и GPU Intel

С OpenVINO вся модель вычисляется как единый граф на процессоре или на встроенном GPU Intel, а NPU берёт на себя матричные произведения. Граф быстрее читает первый проход подсказки; процессорный движок остаётся быстрее на следующих проходах и при записи ответа, а измерение движков, запускаемое со значка, оставляет лучший выбор для данной машины.

Первое чтение подсказки: 260 токенов в секунду с графом против 216 с процессорным движком (Qwen3-VL 4B, Intel Core Ultra 9 275HX, 24 потока)

Характеристики

Формат моделей	Квантованный GGUF: от Q4_0 до Q8_0, от Q2_K до Q6_K и IQ4_NL
Архитектуры	llama, mistral, mistral3, granite, granitemoe, phi3, qwen2, qwen3, qwen3moe, qwen3vl, qwen3vlmoe, mixtral
Кодировщики изображений	Qwen3-VL, LLaVA-NeXT (EuroVLM) и Pixtral (Ministral 3, Devstral Small 2, Mistral Small 3.2)
Модели в составе Albireo	Qwen3-VL-4B-Instruct со своим кодировщиком изображений (2,95 ГБ) и Qwen3-Embedding-0.6B (0,64 ГБ)
Каталог	Моделей: 22, издателей: 7; всё подписано и сверено с хешами
Память	8 ГБ для модели, устанавливаемой по умолчанию
Локальный интерфейс	https://localhost:8443, совместим с интерфейсом OpenAI, документирован в Swagger
Форматы бесед	ChatML, Mistral, Llama 3, Granite, Phi-4-mini и Foundation-Sec, реализованы нативно
Обновления	Подписанная лента, установщики проверяются по размеру и SHA-256

Безопасность, которую можно проверить

Только локально

Служба слушает 127.0.0.1 и отклоняет запросы, у которых хост или источник не локальные.

Удостоверяющий центр компьютера

Сертификат HTTPS выдаёт центр, созданный на компьютере с ключом ECDSA P-256 и ограниченный по именам адресом localhost.

Ключи под защитой Windows

Закрытый ключ сервера и ключ ретрансляции зашифрованы DPAPI и никогда не записываются в открытом виде.

Подписанная цепочка поставки

Лента обновлений и каталог моделей подписаны ECDSA P-256, а каждый загруженный файл сверяется с его SHA-256.

Открытый интерфейс

Маршруты следуют интерфейсу OpenAI и описаны в OpenAPI, так что существующие клиентские библиотеки работают без изменений.

Изолированный синтез речи

Синтезатор работает в отдельном процессе с низким уровнем целостности, без привилегий и с ограничением памяти.

Бесплатно и с открытым кодом под лицензией Boost Software License 1.0. © 2026 Cydral Technology, Франция.