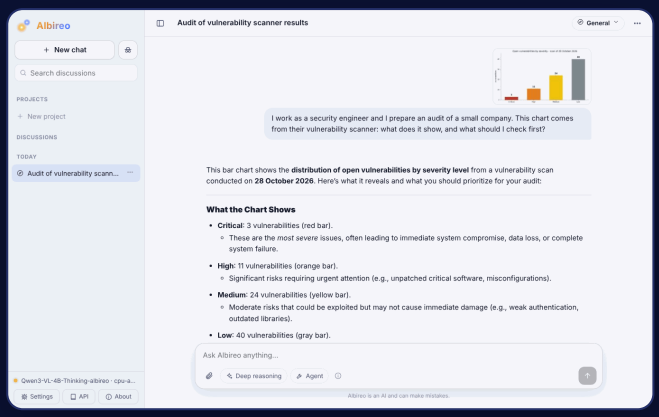
Albireo

आपका निजी *AI* सहायक, जो आपके अपने कंप्यूटर पर चलता है

Albireo सीधे आपके Windows 11 कंप्यूटर पर भाषा मॉडल चलाता है। आपकी बातचीत, दस्तावेज़ और मॉडल उसी मशीन पर रहते हैं: कोई खाता नहीं बनाना पड़ता, कोई डेटा बाहर नहीं भेजा जाता।

संस्करण 1.0.23 प्लेटफ़ॉर्म Windows 11, x64

लाइसेंस Boost Software License 1.0



56.8
टोकन प्रति सेकंड लैपटॉप GPU पर, उसके प्रोसेसर पर 23.6

22
मॉडल हस्ताक्षरित कैटलॉग में, 7 प्रकाशकों से

5
इंजन: प्रोसेसर, Intel NPU और GPU, NVIDIA GPU

0
खाता, सदस्यता या बाहर भेजा गया डेटा

एक नज़र में

चैट
बातचीत ब्राउज़र में रहती है; गुप्त मोड और बातचीत में खोज उपलब्ध है, और तर्क करने वाले मॉडल सोचते समय अपना तर्क दिखाते हैं।

चित्र
चित्र पढ़ने वाले मॉडल PNG और JPEG चित्र लेते हैं, जिन्हें प्रश्न में जोड़ा, चिपकाया या खींचकर छोड़ा जा सकता है।

एजेंट मोड
मॉडल पेज के भीतर चलने वाले टूल बुलाता है: कैलकुलेटर, सैंडबॉक्स में JavaScript, दस्तावेज़ों को पढ़ना और खोजना, प्रति उत्तर 100 स्टेप तक।

बोलकर पढ़ना
हर उत्तर को Albireo के साथ आने वाला स्पीच सिंथेसिस बिना नेटवर्क के बोलकर सुना सकता है।

ऑफ़लाइन मोड
बिना नेटवर्क वाले कंप्यूटर पर सेवा कोई कनेक्शन नहीं खोलती: कैटलॉग, मॉडल और अपडेट डिस्क पर लाए गए हस्ताक्षरित बंडल से आते हैं।

दस्तावेज़
Word, Excel, PowerPoint, OpenDocument, PDF, RTF, HTML, टेक्स्ट, CSV, Markdown और सोर्स कोड, ब्राउज़र में पढ़े जाते हैं और किसी संदेश या प्रोजेक्ट से जोड़े जाते हैं।

प्रोजेक्ट
बातचीत को उन निर्देशों और दस्तावेज़ों के साथ समूहित किया जाता है जो उनका साझा संदर्भ बनाते हैं, और जिनमें अर्थ तथा शब्दों से खोजा जा सकता है।

वेब तक पहुँच
Wikipedia में वैकल्पिक खोज और सार्वजनिक पेजों तथा PDF फ़ाइलों को पढ़ना; हर नई बातचीत में यह बंद रहता है।

दूसरे प्रोग्रामों के लिए इंटरफ़ेस
चैट कम्प्लीशन, कम्प्लीशन, एम्बेडिंग, स्पीच और मॉडल, OpenAI इंटरफ़ेस का पालन करते हैं, स्ट्रीमिंग और चित्रों के साथ, और Swagger पेज में प्रलेखित हैं।

बाहरी इंजन को रिले
किसी बातचीत को बाहरी इंजन को रिले किया जा सकता है: OpenAI-संगत कोई भी इंटरफ़ेस, Ollama का क्लाउड या Claude। कंप्यूटर से निकलने से पहले Pseudalis उसके पाठों को स्यूडोनिमाइज़ कर देता है, और इसी कंप्यूटर या स्थानीय नेटवर्क का इंजन ऑफ़लाइन भी उपयोग में रहता है।

इंजन

इंजन	किस पर चलता है
`cpu`	हर x86-64 प्रोसेसर
`cpu-avx2`	AVX2, FMA और F16C वाले प्रोसेसर: Intel Haswell, AMD Zen और उनके बाद के
`cpu-avxvnni`	AVX-VNNI वाले प्रोसेसर: Alder Lake से Intel Core, Sapphire Rapids से Intel Xeon, AMD Zen 5
`openvino`	Intel प्रोसेसरों का NPU और इंटीग्रेटेड GPU, OpenVINO के ज़रिए
`cuda`	NVIDIA ग्राफ़िक्स प्रोसेसर, CUDA के ज़रिए

आम हार्डवेयर पर रफ़्तार: Albireo की तकनीकें

Albireo मॉडल को अपने इंजन से चलाता है, जो उन प्रोसेसरों और ग्राफ़िक्स कार्डों के लिए डिज़ाइन किया गया है जो पहले से उपयोग में हैं। नीचे दी गई हर तकनीक का माप उत्पाद के README में दिया गया है।

क्वांटाइज़्ड कर्नेल

वेट अपने क्वांटाइज़्ड रूप में GGUF फ़ाइल की मेमोरी मैपिंग में रहते हैं। जहाँ फ़्लोट का एक वेक्टर 8 गुणनफल निकालता है, वहाँ इंटीजर कर्नेल 32 निकालते हैं।

प्रॉम्प्ट पढ़ना: *Qwen3-VL 4B* के साथ *216* टोकन प्रति सेकंड और *Qwen3 0.6B* के साथ *1,300* से अधिक *(Intel Core Ultra 9 275HX, 24 थ्रेड)*; दो कोर पर फ़्लोट के *126* के मुक़ाबले *230 (Qwen3 0.6B, AVX-VNNI)*

हर तरह के प्रोसेसर के लिए एक इंजन

प्रोसेसर के लिए पोर्टेबल, AVX2 और AVX-VNNI इंजन, OpenVINO के ज़रिए Intel का NPU और इंटीग्रेटेड GPU, CUDA के ज़रिए NVIDIA ग्राफ़िक्स प्रोसेसर। सेवा इन्हें मशीन पर मापती है और सबसे तेज़ को चलाती है।

उत्तर लिखना: प्रोसेसर इंजन पर *Qwen3 0.6B* के साथ *80.6* से *87.7* टोकन प्रति सेकंड *(Intel Core Ultra 9 275HX, 24 थ्रेड)*

V²RAM विस्तारित मेमोरी

लाइब्रेरी की विस्तारित डिवाइस मेमोरी वेट की प्रतियों को कार्ड की मेमोरी से आगे, होस्ट मेमोरी या किसी फ़ाइल में जाने देती है, और उपयोग से पहले उन्हें वापस ले आती है।

एक साथ कई बातचीत

हर क्लाइंट के अनुरोध हर फ़ॉरवर्ड पास में एक साथ गणना किए जाते हैं, और पूरे हो चुके अनुरोध की कुंजियाँ और मान अगली बारी के लिए कैश में रहते हैं।

चार उत्तर एक साथ लिखने पर प्रति स्टेप *32* से *40%* कम समय

ड्राफ़्ट टोकन

जब संदर्भ आगे के टोकन का संकेत देता है, तो उत्तर एक स्टेप में कई टोकन लिखता है; यह स्पेक्युलेटिव डिकोडिंग का एक रूप है। मॉडल केवल वही टोकन रखता है जो वह स्वयं चुनता, इसलिए उत्तरों में सैंपलिंग का वितरण बना रहता है।

कोड दोबारा लिखते समय *19.7* के मुक़ाबले *25.5* टोकन प्रति सेकंड *(Qwen3 0.6B, दो कोर); 24* थ्रेड पर *23.6* के मुक़ाबले *26.3 (Qwen3-VL 4B)*

NVIDIA कार्ड पर गणना की गई लेयर

CUDA इंजन वेट को उनके GGUF फ़ॉर्मेट में कार्ड पर रखता है और डेंस मॉडल की लेयर वहीं गणना करता है, जनरेट किए गए टोकन के अटेंशन सहित।

23.6 के मुक़ाबले *56.8* टोकन प्रति सेकंड *(Qwen3-VL 4B, RTX 5070 Ti Laptop GPU* बनाम *24* थ्रेड*)*

संक्षिप्त बातचीत मेमोरी

कुंजियाँ और मान 8 बिट में रखे जाते हैं, ऐसे पेजों में जो एक ही पाठ से शुरू होने वाली बातचीत के बीच साझा होते हैं।

मानों के प्रति समूह *64* के मुक़ाबले *34* बाइट, और *210* स्थितियों पर वही अगला टोकन

Intel प्रोसेसर और GPU पर OpenVINO ग्राफ़

OpenVINO के साथ, पूरा मॉडल प्रोसेसर या Intel के इंटीग्रेटेड GPU पर एक ही ग्राफ़ के रूप में गणना किया जाता है, और NPU मैट्रिक्स गुणनफल सँभालता है। ग्राफ़ प्रॉम्प्ट का पहला पास तेज़ी से पढ़ता है; बाद के पास और उत्तर लिखने में प्रोसेसर इंजन तेज़ रहता है, और ट्रे आइकन से शुरू किया गया इंजनों का माप मशीन के लिए सबसे अच्छा विकल्प बनाए रखता है।

प्रॉम्प्ट की पहली पढ़ाई: ग्राफ़ के साथ *260* टोकन प्रति सेकंड, प्रोसेसर इंजन के साथ *216 (Qwen3-VL 4B, Intel Core Ultra 9 275HX, 24* थ्रेड*)*

विशेषताएँ

मॉडल फ़ॉर्मेट	क्वांटाइज़्ड GGUF: Q4_0 से Q8_0, Q2_K से Q6_K और IQ4_NL
आर्किटेक्चर	llama, mistral, mistral3, granite, granitemoe, phi3, qwen2, qwen3, qwen3moe, qwen3vl, qwen3vlmoe, mixtral
इमेज एनकोडर	Qwen3-VL, LLaVA-NeXT (EuroVLM) और Pixtral (Ministral 3, Devstral Small 2, Mistral Small 3.2)
Albireo के साथ आने वाले मॉडल	Qwen3-VL-4B-Instruct अपने इमेज एनकोडर के साथ (2.95 GB) और Qwen3-Embedding-0.6B (0.64 GB)
कैटलॉग	7 प्रकाशकों के 22 मॉडल, हस्ताक्षरित और अपने हैश से जाँचे गए
मेमोरी	डिफ़ॉल्ट रूप से इंस्टॉल होने वाले मॉडल के लिए 8 GB
स्थानीय इंटरफ़ेस	https://localhost:8443, OpenAI इंटरफ़ेस से संगत, Swagger के साथ प्रलेखित
बातचीत के फ़ॉर्मेट	ChatML, Mistral, Llama 3, Granite, Phi-4-mini और Foundation-Sec, नेटिव रूप से लिखे गए
अपडेट	हस्ताक्षरित फ़ीड, इंस्टॉलर आकार और SHA-256 से जाँचे जाते हैं

ऐसी सुरक्षा जिसका ऑडिट किया जा सकता है

केवल स्थानीय

सेवा 127.0.0.1 पर सुनती है और उन अनुरोधों को अस्वीकार करती है जिनका होस्ट या ऑरिजिन स्थानीय नहीं है।

कंप्यूटर की अपनी सर्टिफ़िकेट अथॉरिटी

HTTPS सर्टिफ़िकेट उस अथॉरिटी से आता है जो कंप्यूटर पर ECDSA P-256 कुंजी से बनाई गई है और नाम-प्रतिबंधों से localhost तक सीमित है।

Windows द्वारा सुरक्षित कुंजियाँ

सर्वर की निजी कुंजी और रिले की कुंजी DPAPI से एन्क्रिप्ट की जाती हैं और कभी सादे रूप में नहीं लिखी जातीं।

हस्ताक्षरित सप्लाई चेन

अपडेट फ़ीड और मॉडल कैटलॉग ECDSA P-256 से हस्ताक्षरित हैं, और हर डाउनलोड की गई फ़ाइल उसके SHA-256 से जाँची जाती है।

खुला इंटरफ़ेस

रूट OpenAI इंटरफ़ेस का पालन करते हैं और OpenAPI से वर्णित हैं, इसलिए मौजूदा क्लाइंट लाइब्रेरी बिना बदलाव के काम करती हैं।

अलग रखा गया स्पीच सिंथेसिस

सिंथेसाइज़र कम इंटीग्रिटी वाली एक अलग प्रोसेस में चलता है, बिना विशेषाधिकारों के और सीमित मेमोरी के साथ।

Boost Software License 1.0 के अंतर्गत नि:शुल्क और ओपन सोर्स। © 2026 Cydral Technology, फ़्रांस।