

Jalankan AI Anda sendiri

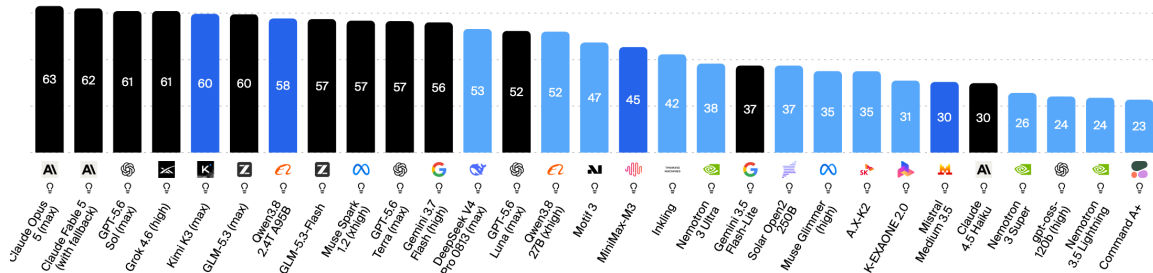
di perangkat keras milik Anda

Menjalankan sendiri berarti tidak ada yang keluar dari gedung. Tidak ada kebijakan retensi yang harus dipercaya, tidak ada ketentuan yang bisa berubah, tidak ada akun yang ditangguhkan di tengah sesi. Sampai baru-baru ini hal itu juga berarti model yang jauh lebih lemah. Sekarang tidak lagi.



● Proprietary ● Open Weights (Commercial Use Restricted) ● Open Weights

Artificial Analysis



GBR. 01 Seluruh model pada Artificial Analysis Intelligence Index, 26 Agustus 2026, berurutan peringkat. Tujuh belas dari 29 menerbitkan bobotnya. Model terbuka terbaik, Kimi K3, berada di peringkat kelima dengan 60, tiga poin di belakang Claude Opus 5. Setahun lalu selisih itu tiga belas poin. Indeks menghitung GLM-5.3 dan GLM-5.3-Flash sebagai proprietary; Z.ai menerbitkan bobot keduanya dengan lisensi MIT, jadi delegasi juga dapat menjalankannya. Sumber: artificialanalysis.ai/#intelligence.

01 MENGAPA MENJALANKAN SENDIRI

- **Data tetap di tempat.** Posisi nasional, instruksi dari ibu kota, dan apa pun yang disampaikan kepada Anda secara rahasia tidak pernah sampai ke pihak ketiga.
- **Bobot yang Anda pegang tidak bisa dinaikkan harganya atau ditarik di tengah sesi.**
- **Anda melihat seluruh jalurnya.** Laboratorium terhosting menyimpan jejak penalaran mereka. Di server sendiri Anda mencatat setiap langkah dan bisa menunjukkan kepada kolega apa yang terjadi.
- **Keterampilannya mengendap di institusi Anda** alih-alih disewa kembali setiap tahun.

02 BERAPA BIAYANYA

Satu NVIDIA DGX Spark, 128GB. Model 27B hingga 35B **\$4.699** untuk enam puluh pengguna bersamaan.

Dua unit terhubung langsung, tanpa switch. 256GB, model kelas DeepSeek pada 55 hingga 60 token per detik, konteks satu juta token. **\$9.400**

Tiga atau empat unit, 384 hingga 512GB. Model terbuka terbesar, tanpa dipangkas. **\$14 rb–19 rb**

Dibandingkan dengan berlangganan untuk setiap delegasi, perangkat keras itu balik modal dalam satu atau dua siklus, dan tetap bekerja setelah pendanaan berakhir.

03 APA YANG DIJALANKAN

MODEL	LISENSI	INDEKS
Kimi K3	Terbuka, dibatasi	60
GLM-5.3	Terbuka, MIT	60
Qwen3.8 2.4T	Terbuka, dibatasi	58
GLM-5.3-Flash	Terbuka, MIT	57
DeepSeek V4 Pro	Terbuka	53
Qwen3.8 27B	Terbuka, dibatasi	52
gpt-oss-120b	Terbuka	24

Baca lisensinya sebelum Anda berkomitmen. Beberapa model terbuka terkuat membatasi penggunaan komersial, yang bagi delegasi biasanya tidak menjadi masalah dan bagi vendor yang membangun di atasnya menjadi masalah.

Sajikan dengan vLLM atau SGLang. Keduanya menghadirkan API OpenAI, sehingga alat apa pun yang berbicara dengan ChatGPT akan berbicara dengan server Anda setelah alamat dan kuncinya diganti. Di laptop, `ollama run gpt-oss:120b`

04 SEWA ATAU BERDAULAT

- **Skala melawan kemandirian lokal.** Sistem bersama itu murah dan kuat. Ia juga menetapkan kategorinya, membangun ketergantungan, dan bisa membuat jalan keluar tertutup.
- **Kuasa melawan risiko.** Siapa yang memegang kuasa sepatutnya memikul sebagian risiko. Dengan menyewa, penyedia yang menulis ketentuannya dan Anda yang menanggung hasilnya.
- **Keterbukaan belum berarti kepercayaan.** Sistem tertutup bisa diaudit lewat kontrak. Yang Anda perlukan adalah hak permanen untuk memeriksa, dan bobot yang Anda pegang adalah cara termurah untuk mendapatkannya.
- **Kapasitas berbunga.** Model terbuka baru muncul setiap beberapa bulan. Tim yang menjalankan sistemnya sendiri mencoba tiap model dalam sepekan dan mempertahankan orang yang menguasainya. Tim yang terikat kontrak menunggu.
- **Hak untuk tidak dimodelkan.** Ada pengetahuan yang tidak layak masuk model mana pun. Kedaulatan mencakup penolakan untuk menjadi data.

Ketegangan dari lokakarya AI4PG tentang Kedaulatan Keputusan atas Alam, ETH Zurich, Juli 2026.

05 SEBELUM MEMERCAYAINYA

Indeks ini rata-rata dari sembilan evaluasi, dan model terbuka tidak tertinggal secara merata di semuanya. Kesenjangan paling lebar pada tugas penalaran tersulit dan pada halusinasi, tempat model proprietary masih unggul.

Bagi seorang negosiator, hal kedua itulah yang menentukan segalanya. Arahkan model ke dokumen Anda sendiri, minta sitasi, dan buka sitasi tersebut.

06 JIKA BELUM BISA

Gunakan layanan terhosting dengan perjanjian retensi data nol secara tertulis, dan perlakukan sebagai langkah sementara. Model terhosting untuk bahan publik, membaca, dan latihan. Perangkat keras sendiri untuk posisi apa pun yang belum publik.

GainForest dan Youth Negotiators Academy membantu delegasi melakukan ini tanpa biaya: menentukan ukuran perangkat keras, memilih model, membuatnya berjalan, dan melatih orang-orang yang akan menggunakannya.

AI YANG SELARAS DENGAN KEDAULATAN UNTUK DIPLOMASI MULTILATERAL • AI4COP.ORG

team@gainforest.net
ai4cop.org

