

Aloje a sua própria IA

em hardware que você controla

O autoalojamento significa que nada sai do edifício. Nenhuma política de retenção em que confiar, nenhuns termos que possam mudar, nenhuma conta suspensa a meio de uma sessão. Até há pouco significava também um modelo muito mais fraco. Já não é assim.

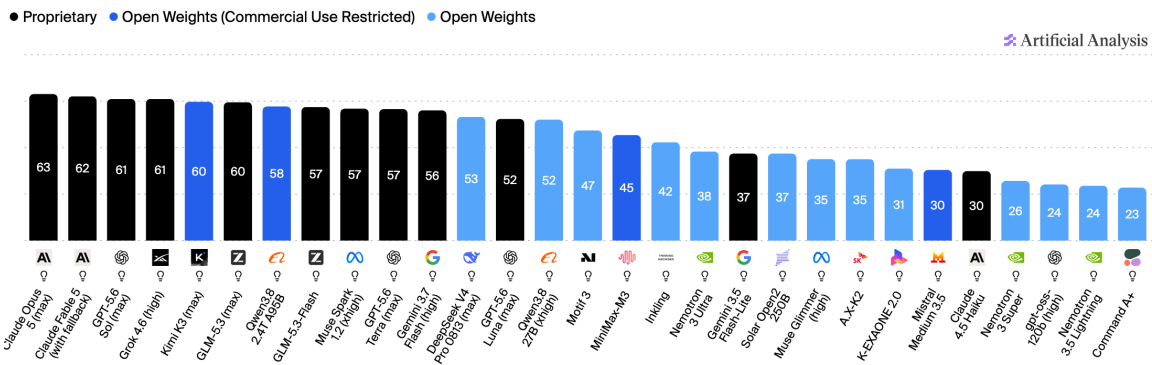


FIG. 01 Todos os modelos do Artificial Analysis Intelligence Index, 26 de agosto de 2026, por ordem de classificação. Dezassete dos 29 publicam os seus pesos. O melhor aberto, Kimi K3, está em quinto com 60, a três pontos do Claude Opus 5. Há um ano essa diferença era de treze pontos. O índice conta o GLM-5.3 e o GLM-5.3-Flash como proprietários; a Z.ai publica ambos os pesos sob licença MIT, pelo que uma delegação também os pode executar. Fonte: artificialanalysis.ai/#intelligence.

01 PORQUÊ EXECUTÁ-LA POR CONTA PRÓPRIA

- Os dados ficam onde estão. Posições nacionais, instruções de uma capital e tudo o que lhe for dito em confiança nunca chegam a terceiros.
- Pesos que você detém não podem ser reavaliados nem retirados a meio de uma sessão.
- Vê todo o percurso. Os laboratórios hospedados guardam os seus rastros de raciocínio. No seu próprio servidor regista cada passo e pode mostrar a um colega o que aconteceu.
- A competência fica na sua instituição em vez de ser alugada todos os anos.

02 QUANTO CUSTA

Um NVIDIA DGX Spark, 128GB. Um modelo de 27B a 35B para sessenta utilizadores simultâneos. **4.699 \$**

Dois ligados diretamente, sem switch. 256GB, um modelo da classe DeepSeek a 55 a 60 tokens por segundo, contexto de um milhão de tokens. **9.400 \$**

Três ou quatro unidades, 384 a 512GB. Os maiores modelos abertos, sem poda. **14 a 19 mil \$**

Em comparação com uma assinatura para cada delegado, o hardware paga-se a si mesmo dentro de um ou dois ciclos, e continua a funcionar depois de terminado o financiamento.

03 O QUE EXECUTAR

MODELO	LICENÇA	ÍNDICE
Kimi K3	Aberto, restrito	60
GLM-5.3	Aberto, MIT	60
Qwen3.8 2.4T	Aberto, restrito	58
GLM-5.3-Flash	Aberto, MIT	57
DeepSeek V4 Pro	Aberto	53
Qwen3.8 27B	Aberto, restrito	52
gpt-oss-120b	Aberto	24

Leia a licença antes de se comprometer. Vários dos modelos abertos mais robustos restringem o uso comercial, o que para uma delegação em geral não é problema, mas para um fornecedor que construa sobre um deles é.

Sirva-o com vLLM ou SGLang. Ambos apresentam a API da OpenAI, pelo que qualquer ferramenta que converse com o ChatGPT conversará com o seu servidor após uma alteração de endereço e chave. Num portátil, `ollama run gpt-oss:120b`

04 ALUGADO OU SOBERANO

- A escala contra a autonomia local. Um sistema partilhado é barato e forte. Também fixa as categorias, cria dependência e pode tornar a saída impossível.
- O poder contra o risco. Quem detém o poder deveria suportar parte do risco. Alugado, o fornecedor escreve os termos e você vive com o resultado.
- A abertura ainda não é confiança. Os sistemas fechados podem ser auditados por contrato. O que é preciso é um direito permanente de verificar, e pesos que você detém são a forma mais barata de o ter.
- A capacidade acumula-se. Novos modelos abertos surgem de poucos em poucos meses. Uma equipa com infraestrutura própria experimenta cada um numa semana e mantém quem sabe fazê-lo. Uma equipa com contrato espera.
- O direito de não ser modelado. Há conhecimento que não pertence a modelo nenhum. A soberania inclui a recusa de se tornar dados.

Tensões do seminário AI4PG sobre soberania de decisão na natureza, ETH Zurique, julho de 2026.

05 ANTES DE CONFIAR

O índice é uma média de nove avaliações, e os modelos abertos não ficam uniformemente atrás em todas elas. A diferença é maior nas tarefas de raciocínio mais difíceis e na alucinação, onde os modelos proprietários ainda estão à frente.

Para um negociador, esse segundo ponto decide tudo. Aponte o modelo para os seus próprios documentos, peça citações e abra-as.

06 SE AINDA NÃO PUDER

Use um serviço hospedado com um acordo escrito de retenção zero de dados, e trate-o como medida provisória. Modelos hospedados para material público, leitura e ensaio. O seu próprio hardware para qualquer posição que ainda não seja pública.

A GainForest e a Youth Negotiators Academy assessoram delegações sem custo: dimensionar o hardware, escolher um modelo, colocá-lo em serviço e capacitar quem o vai utilizar.

IA ALINHADA À SOBERANIA PARA A DIPLOMACIA MULTILATERAL • AI4COP.ORG

team@gainforest.net
ai4cop.org

