Host your own AI

on hardware you control

Self-hosting means nothing leaves the building. No retention policy to trust, no terms that can change, no account suspended mid-session. Until recently it also meant a much weaker model. That is no longer true.

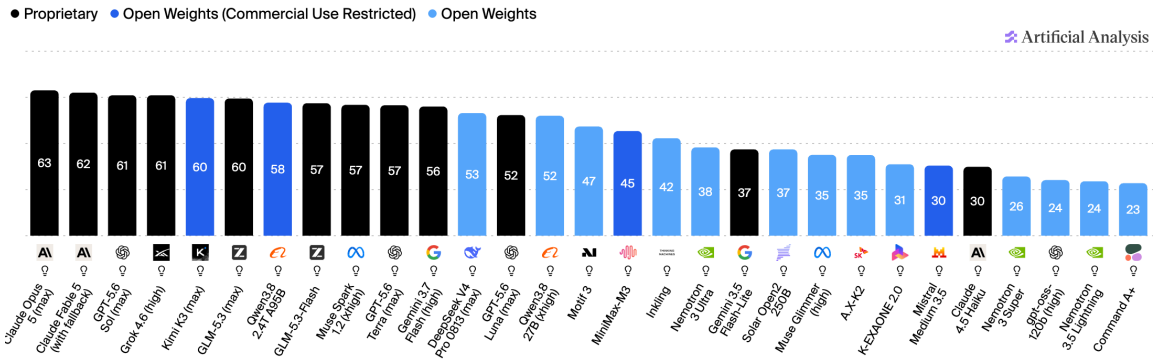

FIG. 01 **Every model on the Artificial Analysis Intelligence Index, 26 August 2026, in rank order.** Seventeen of the 29 publish their weights. The best open one, Kimi K3, is fifth at 60, three points behind Claude Opus 5. A year ago that gap was thirteen points. Artificial Analysis counts GLM-5.3 and GLM-5.3-Flash as proprietary; Z.ai publishes both sets of weights under MIT, so a delegation can run them too. Source: artificialanalysis.ai/#intelligence.

01 WHY RUN IT YOURSELF

- **The data stays put.** National positions, instructions from a capital, and anything told to you in confidence never reach a third party.
- **Weights you hold cannot be repriced or withdrawn** in the middle of a session.
- **You can see the whole path.** Hosted labs keep their reasoning traces. On your own server you log every step and can show a colleague what happened.
- **The skill lands in your institution** instead of being rented back every year.

02 WHAT IT COSTS

One NVIDIA DGX Spark, 128GB. A 27B to 35B model for sixty concurrent users. **$4,699**

Two linked directly, no switch. 256GB, a DeepSeek-class model at 55 to 60 tokens a second, one million token context. **$9,400**

Three or four units, 384 to 512GB. The largest open models, unpruned. **$14k to $19k**

Set against a subscription for every delegate, the hardware pays for itself inside a cycle or two, and it keeps working after the funding ends.

03 WHAT TO RUN

MODEL	LICENCE	INDEX
Kimi K3	Open, restricted	**60**
GLM-5.3	Open, MIT	**60**
Qwen3.8 2.4T	Open, restricted	**58**
GLM-5.3-Flash	Open, MIT	**57**
DeepSeek V4 Pro	Open	**53**
Qwen3.8 27B	Open, restricted	**52**
gpt-oss-120b	Open	**24**

Read the licence before you commit. Several of the strongest open models restrict commercial use, which for a delegation is usually fine and for a vendor building on top of one is not.

Serve it with vLLM or SGLang. Both speak the OpenAI API, so any tool that talks to ChatGPT talks to your server after a change of address and key. On a laptop,

```
ollama run gpt-oss:120b
```

04 RENTED OR SOVEREIGN

- **Scale against local agency.** A shared system is cheap and strong. It also sets the categories, builds dependence, and can make exit impossible.
- **Power against risk.** Whoever holds the power should carry part of the risk. Rented, the provider writes the terms and you live with the result.
- **Openness is not yet trust.** Closed systems can be audited under contract. What you need is a standing right to check, and weights you hold are the cheapest way to get one.
- **Capacity compounds.** New open models land every few months. A team running its own stack tries each one in a week and keeps the people who know how. A team on a contract waits.
- **The right to stay unmodelled.** Some knowledge belongs in no model. Sovereignty includes the refusal to become data.

Tensions from the AI4PG workshop on Decision Sovereignty in Nature, ETH Zurich, July 2026.

05 BEFORE YOU TRUST IT

The index averages nine evaluations, and open models do not trail evenly across them. The gap is widest on the hardest reasoning tasks and on hallucination, where proprietary models are still ahead.

For a negotiator that second one decides everything. Point the model at your own documents, ask for citations, and open them.

06 IF YOU CANNOT YET

Use a hosted service with a zero data retention agreement in writing, and treat it as a stopgap. Hosted models for public material, reading, and rehearsal. Your own hardware for any position that is not yet public.

GainForest and the Youth Negotiators Academy help delegations do this at no cost: sizing the hardware, choosing a model, getting it serving, and training the people who will use it.

team@gainforest.net
ai4cop.org