

Hébergez votre propre IA

sur du matériel que vous contrôlez

L'auto-hébergement signifie que rien ne quitte les locaux. Pas de politique de conservation à laquelle faire confiance, pas de conditions susceptibles de changer, pas de compte suspendu en pleine session. Jusqu'à récemment, cela signifiait aussi un modèle beaucoup moins performant. Ce n'est plus le cas.

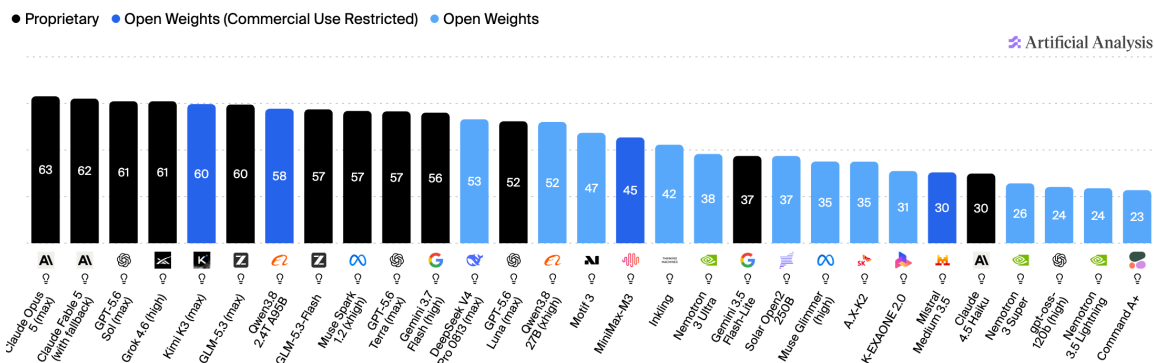


FIG. 01 Tous les modèles de l'Artificial Analysis Index, au 26 août 2026, classés par rang. Dix-sept des 29 publient leurs poids. Le meilleur modèle ouvert, Kimi K3, est cinquième à 60, à trois points de Claude Opus 5. Il y a un an, cet écart était de treize points. L'indice compte GLM-5.3 et GLM-5.3-Flash comme propriétaires; Z.ai publie les deux jeux de poids sous licence MIT, une délégation peut donc les exploiter aussi. Source: artificialanalysis.ai/#intelligence.

01 POURQUOI L'EXPLOITER SOI-MÊME

- **Les données restent sur place.** Positions nationales, instructions d'une capitale et tout ce qui vous est confié ne parviennent jamais à un tiers.
- **Des poids que vous détenez ne peuvent être ni retarifés ni retirés** au milieu d'une session.
- **Vous voyez tout le cheminement.** Les laboratoires hébergés gardent leurs traces de raisonnement. Sur votre serveur, vous journalisez chaque étape et pouvez montrer à un collègue ce qui s'est passé.
- **La compétence reste dans votre institution** au lieu d'être louée chaque année.

02 COMBIEN CELA COÛTE

Un NVIDIA DGX Spark, 128GB. Un modèle de 27B à 35B pour soixante utilisateurs simultanés. **4 699 \$**

Deux unités reliées directement, sans commutateur. 256GB, un modèle de la classe DeepSeek à 55 à 60 jetons par seconde, contexte d'un million de jetons. **9 400 \$**

Trois ou quatre unités, 384 à 512GB. Les plus grands modèles ouverts, sans élagage. **14 à 19 k\$**

Comparé à un abonnement pour chaque délégué, le matériel s'amortit en un cycle ou deux, et il continue de fonctionner une fois le financement terminé.

03 QUEL MODÈLE EXPLOITER

MODÈLE	LICENCE	INDICE
Kimi K3	Ouvert, usage restreint	60
GLM-5.3	Ouvert, MIT	60
Qwen3.8 2.4T	Ouvert, usage restreint	58
GLM-5.3-Flash	Ouvert, MIT	57
DeepSeek V4 Pro	Ouvert	53
Qwen3.8 27B	Ouvert, usage restreint	52
gpt-oss-120b	Ouvert	24

Lisez la licence avant de vous engager. Plusieurs des modèles ouverts les plus performants restreignent l'usage commercial, ce qui ne pose pas de problème pour une délégation, mais en pose un pour un fournisseur qui construit par-dessus.

Servez-le avec vLLM ou SGLang. Les deux présentent l'API d'OpenAI: tout outil qui dialogue avec ChatGPT dialoguera avec votre serveur après un changement d'adresse et de clé. Sur un ordinateur portable, `ollama run gpt-oss:120b`

04 LOUÉ OU SOUVERAIN

- **L'échelle contre l'autonomie locale.** Un système partagé est économique et performant. Il fixe aussi les catégories, installe la dépendance et peut rendre la sortie impossible.
- **Le pouvoir contre le risque.** Qui détient le pouvoir devrait porter une part du risque. En location, le fournisseur écrit les conditions et vous en assumez le résultat.
- **L'ouverture n'est pas encore la confiance.** Les systèmes fermés peuvent être audités par contrat. Ce qu'il vous faut, c'est un droit permanent de vérifier, et des poids que vous détenez sont le moyen le moins cher de l'obtenir.
- **La capacité se cumule.** De nouveaux modèles ouverts paraissent tous les quelques mois. Une équipe qui exploite sa propre infrastructure les essaie en une semaine et garde les personnes qui savent faire. Une équipe sous contrat attend.
- **Le droit de rester non modélisé.** Certains savoirs n'ont leur place dans aucun modèle. La souveraineté inclut le refus de devenir une donnée.

Tensions issues de l'atelier AI4PG sur la souveraineté décisionnelle dans la nature, ETH Zurich, juillet 2026.

05 AVANT DE S'Y FIER

L'indice est une moyenne de neuf évaluations, et les modèles ouverts n'y sont pas en retrait de manière uniforme. L'écart est le plus marqué sur les tâches de raisonnement les plus difficiles et sur les hallucinations, où les modèles propriétaires restent en tête. Pour un négociateur, ce second point décide de tout. Orientez le modèle vers vos propres documents, demandez des citations, et ouvrez-les.

06 SI CE N'EST PAS POSSIBLE

Recourez à un service hébergé assorti d'un accord écrit de conservation nulle des données, et traitez-le comme une mesure transitoire. Les modèles hébergés pour le matériel public, la lecture et la répétition. Votre propre matériel pour toute position qui n'est pas encore publique.

GainForest et la Youth Negotiators Academy conseillent gratuitement les délégations: dimensionnement du matériel, choix d'un modèle, mise en service et formation des personnes qui l'utiliseront.

L'IA AU SERVICE DE LA SOUVERAINETÉ POUR LA DIPLOMATIE MULTILATÉRALE
· AI4COP.ORG

team@gainforest.net
ai4cop.org

