

Aloje su propia IA

en hardware que usted controla

El autoalojamiento significa que nada sale del edificio. Ninguna política de conservación en la que confiar, ninguna condición que pueda cambiar, ninguna cuenta suspendida en mitad de una sesión. Hasta hace poco también significaba un modelo mucho más débil. Ya no es así.

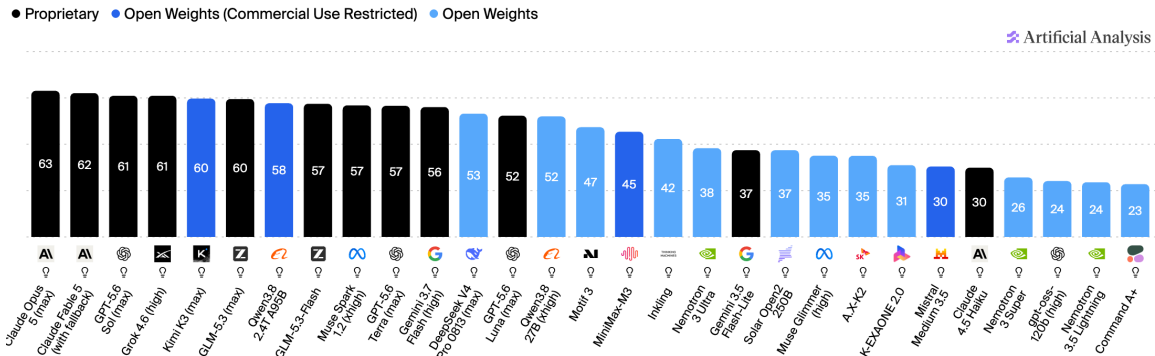


FIG. 01 Todos los modelos del Artificial Analysis Intelligence Index, 26 de agosto de 2026, en orden de clasificación. Diecisiete de los 29 publican sus pesos. El mejor abierto, Kimi K3, ocupa el quinto lugar con 60, a tres puntos de Claude Opus 5. Hace un año esa brecha era de trece puntos. El índice cuenta GLM-5.3 y GLM-5.3-Flash como propietarios; Z.ai publica ambos pesos bajo licencia MIT, así que una delegación también puede ejecutarlos. Fuente: artificialanalysis.ai/#intelligence.

01 POR QUÉ EJECUTARLA POR CUENTA PROPIA

- **Los datos se quedan donde están.** Las posiciones nacionales, las instrucciones de una capital y todo lo que le confíen nunca llegan a un tercero.
- **Unos pesos que usted tiene no pueden encarecerse ni retirarse** en mitad de una sesión.
- **Puede ver todo el recorrido.** Los laboratorios alojados se guardan sus trazas de razonamiento. En su propio servidor usted registra cada paso y puede mostrar a un colega qué ocurrió.
- **La capacidad queda en su institución** en lugar de alquilarse cada año.

02 CUÁNTO CUESTA

Una NVIDIA DGX Spark, 128GB. Un modelo de 27B a 35B para sesenta usuarios simultáneos. **4.699 \$**

Dos conectadas directamente, sin conmutador. 256GB, un modelo de la clase DeepSeek a 55 a 60 tokens por segundo, contexto de un millón de tokens. **9.400 \$**

Tres o cuatro unidades, 384 a 512GB. Los mayores modelos abiertos, sin poder. **14 a 19 mil \$**

Comparado con una suscripción para cada delegado, el hardware se amortiza en uno o dos ciclos, y sigue funcionando después de que termine la financiación.

03 QUÉ EJECUTAR

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

Lea la licencia antes de comprometerse. Varios de los modelos abiertos más sólidos restringen el uso comercial, lo que para una delegación suele estar bien y para un proveedor que construye sobre uno de ellos no.

Sírvalo con vLLM o SGLang. Ambos presentan la API de OpenAI, de modo que cualquier herramienta que hable con ChatGPT hablará con su servidor tras cambiar la dirección y la clave. En un portátil, `ollama run gpt-oss:120b`

04 ALQUILADO O SOBERANO

- **La escala frente a la agencia local.** Un sistema compartido es barato y potente. También fija las categorías, crea dependencia y puede hacer imposible la salida.
- **El poder frente al riesgo.** Quien tiene el poder debería cargar con parte del riesgo. Alquilado, el proveedor escribe las condiciones y usted vive con el resultado.
- **La apertura todavía no es confianza.** Los sistemas cerrados pueden auditarse por contrato. Lo que falta es un derecho permanente a comprobar, y unos pesos propios son la forma más barata de tenerlo.
- **La capacidad se acumula.** Cada pocos meses aparecen nuevos modelos abiertos. Un equipo con infraestructura propia prueba cada uno en una semana y conserva a quienes saben hacerlo. Un equipo con contrato espera.
- **El derecho a no ser modelado.** Hay conocimiento que no pertenece a ningún modelo. La soberanía incluye la negativa a convertirse en datos.

Tensiones del taller AI4PG sobre soberanía en las decisiones sobre la naturaleza, ETH Zürich, julio de 2026.

05 ANTES DE CONFIAR

El índice promedia nueve evaluaciones, y los modelos abiertos no van a la zaga de manera uniforme en todas ellas. La brecha es mayor en las tareas de razonamiento más difíciles y en las alucinaciones, donde los modelos propietarios siguen por delante.

Para un negociador, ese segundo aspecto lo decide todo. Dirija el modelo a sus propios documentos, pídale citas y ábralas.

06 SI AÚN NO PUEDE

Use un servicio alojado con un acuerdo escrito de conservación cero de datos, y trátelo como solución provisional. Modelos alojados para material público, lectura y ensayo. Su propio hardware para cualquier posición que aún no sea pública.

GainForest y la Youth Negotiators Academy asesoran a las delegaciones sin costo alguno: dimensionar el hardware, elegir un modelo, ponerlo en servicio y formar a quienes lo utilizarán.

IA ALINEADA CON LA SOBERANÍA PARA LA DIPLOMACIA MULTILATERAL · AI4COP.ORG

team@gainforest.net
ai4cop.org

