

Cloud API-based generative AI compared: three-year total cost and four key factors

For internal discussions on “Would a cloud API be enough?” (July 2026 edition)

1. Compare four key factors

Comparison factor	Cloud API-based generative AI	Sovereign GaiXer (on-premises AI)
Three-year total cost	Grows in proportion to usage. The more established the use, the larger the bill.	Fixed at JPY 3,806,400 (excluding tax). Cost per use falls as usage grows.
Where data resides	Sent to an external cloud. Work involving confidential or personal data requires usage restrictions.	AI processing runs on the appliance. Internal data is not sent to the internet.
Deployment speed	Contracting, account design, usage guidelines and departmental permission design.	Delivered 1–2 business days after dispatch in Japan. Ready to use once powered on.
Usage charges and operation	Usage- or credit-based billing. Managing quotas and prepaid balances adds operational work.	No additional usage charges. No usage-budget cap to manage.

2. Three-year cost estimates by scenario (company with 50 employees)

Usage scenario	Cloud unit price (API list price basis)	Assumed monthly usage	Cloud three-year total	Sovereign GaiXer (on-premises AI)
Deep Research (research report generation) ChatGPT business-plan credits: 50 per task	Approx. JPY 323/task	50 people × 10 = 500 tasks	Approx. JPY 5.81M	JPY 3,806,400 (fixed)
Agent mode (automated task execution) 30 credits per message	Approx. JPY 194/message	100 tasks (10 messages each)	Approx. JPY 6.97M	JPY 3,806,400 (fixed)
Chat with internal data (RAG) API billing: approx. 20,000 characters read per question → 500-character answer	Approx. JPY 28/question	50 people × 20/day = 20,000 questions	Approx. JPY 20.04M	JPY 3,806,400 (fixed)
Contract and PDF review API billing: 30 pages (approx. 30,000 characters) → 1,000 characters	Approx. JPY 44/item	5,000 items	Approx. JPY 7.84M	JPY 3,806,400 (fixed)

The real risk of usage and credit billing is not the amount but its unpredictability

PDF reading and internal data integration pass more text (input tokens) to the AI with every conversation, so the bill grows as usage takes hold. Credits also require managing quotas and prepaid balances, which turns budget control itself into operational work.

With Sovereign GaiXer, the three-year total is JPY 3,806,400 (excluding tax) in every scenario. Three years of budget are fixed the moment you deploy.

3. “Aren't top-tier cloud models more capable?”

Yes. Cutting-edge cloud models do outperform in some situations. Even so, there are three reasons we recommend on-premises.

Reason	Explanation
① Some data cannot leave your organization, whatever the performance	For work involving confidential or personal data, cloud performance is performance you cannot use. Being usable is the precondition for performance.
② Most work does not need top-tier performance	Everyday tasks such as summarizing, drafting and searching internal documents are well handled by a dedicated environment.
③ The more a task repeats daily, the more charges pile up	In some areas, growing usage and credit charges outweigh the benefit of higher performance. Fixed pricing pays off most for routine, high-volume work.

The practical conclusion: use both

Use the cloud for advanced analysis of public information and Sovereign GaiXer for confidential, high-volume, routine work. Matching each to its role is the most rational setup for both cost and security.

4. Assumptions and sources

- OpenAI does not publish a direct yen price per credit for ChatGPT Enterprise / Edu. Enterprise / Edu purchase a shared credit pool at the contract level, and additional or overage usage is priced under the contract terms with OpenAI. For a formal estimate, check the credit rate in your contract or Order Form.
- Main credit consumption on ChatGPT business plans (Business / Enterprise / Edu): Deep Research = 50 credits per task; agent mode = 30 credits per message (source: OpenAI official rate card).
- The credit unit price was derived by cross-checking OpenAI's official ChatGPT Rate Card with the API pricing table, converting at API list prices with 1 credit ≈ USD 0.04 (source: OpenAI official API pricing page).
- Yen conversion assumes USD 1 = JPY 161.37 (1 credit ≈ JPY 6.45). Conditions for the API usage scenarios: gpt-5.5 standard (short context); 1 Japanese character ≈ 1.5 tokens.
- The Sovereign GaiXer amount is the appliance price (including hardware, software license and three-year warranty, excluding tax) for the Japan market. Running costs such as electricity are not included. Overseas availability, pricing, shipping, warranty and support require confirmation.
- On the cloud side, included seat fees, caching discounts and automatic routing to smaller models may lower the actual bill. Prices, rates and exchange rates change; always check the latest official information.
- These are approximate figures for budget planning and do not indicate the relative quality of any specific service. Unit prices are as of July 4, 2026.

Estimate with your own numbers: Enter your user count and usage frequency in the cost simulator (<https://sovereign.gaixer.com/en/simulator/>) to see a three-year comparison and payback period.