



INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

AI API Pricing Comparison (2026): Grok vs Gemini vs GPT-4o vs Claude

Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2025-12-03 · ai api pricing · llm cost comparison · per-token pricing · xai grok · openai gpt-4o · google gemini · anthropic claude · generative ai cost · ai

Original article: <https://intuitionlabs.ai/articles/ai-api-pricing-comparison-grok-gemini-openai-claude>



In this article

1. [AI API Pricing Comparison \(2026\): Grok vs Gemini vs GPT-4o vs Claude](#)
2. [Executive Summary](#)
3. [Introduction and Background](#)
4. [xAI Grok API Pricing](#)
5. [Google Gemini Pricing](#)
6. [OpenAI GPT-4o and ChatGPT Pricing](#)
7. [Anthropic Claude Pricing](#)
8. [Comparative Pricing Analysis](#)
9. [Case Studies and Real-World Examples](#)
10. [Comparative Discussion and Future Implications](#)
11. [Conclusion](#)

Last updated: October 10, 2026. Prices and model availability change frequently; verify rates directly with each provider before making a purchasing decision. Originally published December 2025.

AI API Pricing Comparison (2026): Grok vs Gemini vs GPT-4o vs Claude

Executive Summary

Update (October 2026): Several lineups changed after the September check. OpenAI now features GPT-6 Astra (\$10/\$50 per million input/output tokens), GPT-6.1 Sol (\$2/\$10) and GPT-6 Luna (\$0.10/\$0.50); GPT-4o is unchanged at \$2.50/\$10. ([OpenAI API pricing](#)) Anthropic's current models are Claude Fable 5.1 (\$10/\$50), Opus 5.5 (\$4/\$20), Sonnet 5.5 (\$2/\$10) and Haiku 5.5 (from \$0.10/\$0.50). ([Claude Platform pricing](#)) xAI released Grok 4.7 at the same \$2/\$6 short-context price as Grok 4.6. ([xAI release notes](#)) Google deprecated Gemini 3.7 Flash on October 8, 2026 and routes its requests to Gemini 3.8 Flash, which is \$0.75/\$3.75 on the Standard tier through December 31, 2026 and \$1.50/\$7.50 from January 1, 2027. ([Gemini API changelog](#)) ([Gemini API pricing](#))

This report compares selected API price cards for xAI Grok, Google Gemini, OpenAI GPT-4o, and Anthropic Claude, verified October 10, 2026. API pricing is usage-based and varies by model, token type, context length, and service tier. The selected comparison table uses Grok 4.7 at \$2.00/\$6.00 per million input/output tokens for short-context requests, GPT-4o at \$2.50/\$10.00, Gemini 3.8 Flash Standard at \$0.75/\$3.75 through December 31, 2026, GPT-6.1 Sol at \$2.00/\$10.00, and the current Claude models shown in Table 3. These figures are rate-card entries, not a provider-wide capability or value ranking. ([xAI API pricing](#); [Gemini API pricing](#); [GPT-4o pricing](#); [Claude Platform pricing](#))

Beyond per-token rates, pricing structures include subscription tiers and enterprise plans. **OpenAI** offers ChatGPT Plus at \$20/mo and Pro at \$200/mo, with Pro now sold in \$100, \$200 and \$500 monthly tiers (**OpenAI Help Center**) (www.reuters.com), Anthropic has a \$20/mo Pro (~\$17/mo annual) plan and introduced a “Max” plan at \$200/mo for heavy users (www.reuters.com) ([claude.ai.chat](https://claude.ai/chat)), and X’s **Premium+** tier, which includes Grok access, rose to \$22/mo in December 2024 and is now listed at \$40/mo in the U.S. (**X Help Center**) (www.reuters.com). Google rolled out **Gemini Enterprise** (a business subscription) at \$30 per user per month (www.axios.com). Notably, companies have offered promotional and special pricing: e.g. US federal agencies can license Grok 4 models for only \$0.42 **per agency** for an 18-month term under a OneGov program (www.reuters.com) (www.tomshardware.com), and India’s Jio users received 18 months of **free** Gemini 2.5 Pro (valued at ~\$399) in a 2025 partnership (www.reuters.com).

These pricing differences have broad implications. Lower prices and promotional deals may spur adoption by cost-sensitive developers or government, while reporting on xAI’s federal deal also noted concerns about Grok’s content (www.reuters.com) (www.techradar.com). At the top of the price lists, Claude Fable 5.1 and GPT-6 Astra are each listed at \$10/\$50 per million input/output tokens. The rapid evolution and competitive pressures are already being felt: AI providers continue to revisit their pricing, while customers (especially enterprises) demand transparent, flexible, usage-based models (www.techradar.com) (www.techradar.com). This report examines the historical context, current pricing details, **comparative analysis**, real-world case studies, and future outlook, drawing on official documentation and industry sources. All claims are backed by credible references.

Introduction and Background

Since 2022, generative AI chatbots have exploded in capability and popularity, driven by models like OpenAI’s GPT series, Google DeepMind’s Gemini, Anthropic’s Claude, and, as of late 2023, Elon Musk’s xAI “Grok.” Each platform targets developers, enterprises, and end-users with conversational AI and multimodal services. Understanding **how much these services cost** is critical for budgeting and adoption. Unlike traditional software, **AI models** are typically billed on a *per-token* or *per-use* basis: developers pay for the amount of text (or images/audio) processed. Thus, even small differences in token rates or model capabilities can translate into large cost differences at scale.

This report focuses on **API pricing** (i.e. developer/enterprise usage pricing), as well as notable subscription tiers. We compare X.AI Grok, Google Gemini, OpenAI ChatGPT (GPT), and Anthropic Claude. Each has its own pricing page or announcements with updated rates. We also consider special pricing deals and enterprise offerings. We will cover:

- **X.AI Grok (xAI)**: Grok is the chatbot built by Musk’s xAI and integrated into X (formerly Twitter). It debuted in Nov 2023 (www.androidcentral.com). Grok is offered as a consumer chatbot and as an API through xAI. We examine Grok’s token prices from xAI’s docs (docs.x.ai), along with Musk’s subscription tiers (like X Premium+), enterprise partnerships (e.g. Telegram integration (www.androidcentral.com)), and government deals (www.reuters.com).

- **Google Gemini (DeepMind):** Google launched Gemini in late 2023 (www.axios.com), aiming to compete with GPT. As of October 2026, Google's pricing page lists **Gemini 3.8 Flash**, **Gemini 3.6 Flash**, **Gemini 3.5 Flash-Lite**, **Gemini 3.1 Flash-Lite**, and **Gemini 3.1 Pro Preview**, with Gemini 3 Flash Preview labeled legacy and the **Gemini 2.5** models limited to existing users, available via Google AI Studio and Vertex AI. ([Gemini API pricing](#)) Google provides free tiers and paid tiers for Gemini and related tools. We present data from Google's official pricing (Gemini API docs (ai.google.dev) (ai.google.dev)), including free allowances, and corporate subscriptions like *Gemini Enterprise* (\$30/user-mo) (www.axios.com).
- **OpenAI ChatGPT (GPT):** OpenAI's ChatGPT launched Nov 2022 and has evolved through multiple model generations. Its API is used by businesses and also powers the ChatGPT product via subscription (ChatGPT Plus \$20/mo, Pro from \$100/mo). OpenAI's featured API models are **GPT-6 Astra** at \$10/\$50 per million input/output tokens, **GPT-6.1 Sol** at \$2/\$10, and **GPT-6 Luna** at \$0.10/\$0.50; the earlier **GPT-5.2** remains listed at \$1.75/\$14, with **GPT-5.2 Pro** at \$21/\$168. ([OpenAI API pricing](#)) We also mention ChatGPT's consumer plans (www.reuters.com) (www.reuters.com) (Plus, Go in India).
- **Anthropic Claude:** Anthropic's Claude (founded by ex-OpenAI researchers) offers a tiered model lineup: Haiku 5.5, Sonnet 5.5, Opus 5.5, and Fable 5.1. It has a free tier and a \$20/mo Pro plan for individuals ([claude.ai.chat](https://claude.ai/chat)), plus \$100/\$200 power-user plans (www.reuters.com). Critically, Anthropic publishes per-token API prices for each model: Haiku 5.5 from \$0.10/\$0.50, Sonnet 5.5 at \$2/\$10, Opus 5.5 at \$4/\$20, and Fable 5.1 at \$10/\$50 per million tokens ([Claude Platform pricing](#)). A third-party summary covers the earlier Haiku 4.5 (\$1/\$5), Sonnet 4.6 (\$3/\$15), and Opus 4.6 (\$5/\$25) rates, which Anthropic still lists ([claude.ai.chat](https://claude.ai/chat)).

Beyond listing prices, we analyze these in context. We include historical context (how pricing has shifted over 2023–25), discuss usage-based vs subscription models, and cite insights from analysts and case studies (e.g. Reuters reports on government deals (www.reuters.com) (www.tomshardware.com) (www.reuters.com) and industry observers on developer reactions (www.techradar.com) (www.techradar.com)). Finally, we discuss implications: how pricing shapes adoption, market competition, and future trends.

xAI Grok API Pricing

Overview of Grok and xAI. xAI (founded mid-2023 by Elon Musk and partners) developed Grok, an AI chatbot launched on Musk's X platform (formerly Twitter) in November 2023 (www.androidcentral.com). Grok has been released in several versions (including Grok 2, 3, and 4); xAI released Grok 4.7 in September 2026 and now identifies it as its flagship model ([xAI release notes](#)). xAI sells Grok features through consumer products on X, including subscription tiers such as *Premium+*, and through an API for developers. Musk's strategy appears to rely on **API usage fees** and subscriptions to fund xAI's high costs (www.axios.com).

Commercial token pricing. xAI's current price card lists **Grok 4.7** (and Grok 4.6, at the same rates) at **\$2.00 per million input tokens**, **\$0.50 per million cached-input tokens**, and **\$6.00 per million output tokens** for short-context requests. Prompts of 200,000 tokens or more use the long-context rate for all tokens in the request. (See Table 1 for details.) ([xAI API pricing](#)) Other Grok models and modalities have separate rate

cards; estimates should use the exact model, context length, and token type selected for the workload. (xAI API pricing) Grok 4.7 and Grok 4.6 each have a 500,000-token context window; its long-context price applies when a prompt reaches 200,000 tokens. (xAI API pricing)

Model (Grok)	Input (\$/1M tokens)	Cached input (\$/1M tokens)	Output (\$/1M tokens)	Context qualification
Grok 4.7 / 4.6	\$2.00	\$0.50	\$6.00	Requests below 200K prompt tokens
Grok 4.7 / 4.6	\$4.00	\$1.00	\$12.00	Long-context pricing applies to all tokens when the prompt reaches 200K tokens

Table 1: Grok 4.7 and Grok 4.6 text API pricing per 1 million tokens (USD), checked October 10, 2026. (xAI API pricing)

Separately from the per-token API rates, Reuters reported that U.S. federal agencies can license Grok 4 (and Grok 4 Fast) for just **\$0.42 per agency** for an 18-month term under a government contract (www.reuters.com), a nominal government price that is not comparable with xAI's per-token API rates. For perspective, Reuters noted that OpenAI's ChatGPT equivalent was \$1 per agency, per year (www.reuters.com). In the consumer arena, xAI also bundles Grok access into X's subscription: for instance, X's **Premium+** plan includes Grok chatbot access; it rose to \$22/mo in December 2024 (www.reuters.com), and X now lists it at \$40/mo in the U.S. (X Help Center).

Developer and enterprise offerings. In October 2025, X (the social platform, whose data API is separate from xAI's Grok API) announced a shift to usage-based pricing for the X API after years of flat fees (www.techradar.com). Specifically, X moved from fixed-tier plans to metered billing (per data consumption) with developer credits and new tooling (www.techradar.com). That change concerns access to X platform data, not Grok model pricing; xAI's Grok API is billed per token at the rates above. The official site also advertises enterprise provisioning (single sign-on, SLA, etc.) and directs big customers to contact sales (x.ai) (docs.x.ai).

Real-world contracts illustrate Grok's positioning. Beyond the GSA deal above, xAI inked a \$300 million partnership with Telegram in mid-2025 (www.androidcentral.com). Through this, Telegram's 1+ billion users get Grok access in-app, as xAI pays Telegram (ironically, Grok typically charges *developers*, but here xAI pays for mass integration). On the consumer side, xAI operates subscription tiers like *SuperGrok Heavy*: a new high-end plan at **\$300/month** launching Summer 2025. (www.windowscentral.com). This tier bundles Grok 4 Heavy (a multi-agent version) and is pitched as "maximally truth-seeking, smartest AI in the world," suggesting xAI expects some users to pay boutique pricing for premium service (www.windowscentral.com).

Despite these offerings, xAI faces skepticism over revenue. An industry analysis noted xAI's revenues (from API and subs) are projected at ~\$500 million for 2025 vs. a \$1 billion monthly burn (www.axios.com), a large gap between revenue and spending at the time. Low-priced government deals seem aimed at maximizing adoption, and reporting on the federal deal noted that Grok has faced criticism over offensive posts and bias (www.reuters.com) (www.tomshardware.com); lower prices may accelerate use but could amplify such issues if not addressed. As one industry survey highlighted, traditional pricing models often fail for AI: 68% of tech

executives say conventional approaches are “insufficient” for monetizing AI (www.techradar.com). xAI’s per-token API billing is itself a usage-based model; TechRadar’s report on usage-based pricing and participant credits (www.techradar.com) concerns the separate X platform API.

Historical context and industry perception. Initially, X’s API pricing changed during 2022–24 (www.techradar.com) (www.socialmediatoday.com). In October 2025, X introduced pay-as-you-go pricing. Early reviews of Grok noted both impressive performance and volatility (www.windowscentral.com). In the market, Grok is often positioned as a populist alternative (some liken it to a “right-leaning” AI option (www.axios.com)). From a pricing standpoint, Grok’s per-token rates are a key part of its appeal; as of October 2026, Grok 4.7’s \$2/\$6 short-context rate sits between low-cost models such as Gemini 3.8 Flash and flagships such as GPT-6 Astra. Reporting on the federal deal made a separate, non-token price comparison: “agencies can purchase Grok... for 42 cents per organization... a competitive price compared to OpenAI’s \$1 per year fee for ChatGPT” (www.reuters.com).

In summary, Grok 4.7’s (and Grok 4.6’s) published short-context rate is \$2/\$6 per million input/output tokens, with different pricing for long-context requests and cached input. That rate alone does not establish a provider-wide cost, capability, maturity, safety, or value ranking. ([xAI API pricing](#))

Google Gemini Pricing

Gemini’s introduction and offerings. Google unveiled its **Gemini** family of multimodal LLMs in late 2023 (www.axios.com). By 2026, successive generations have expanded Google’s AI lineup to the current **Gemini 3.x** series (3.8 Flash, 3.6 Flash, 3.5 Flash-Lite, 3.1 Flash-Lite, 3.1 Pro Preview) alongside the **Gemini 2.5** models (2.5 Pro, 2.5 Flash), which Google now limits to existing users. Each model offers different power-to-cost tradeoffs and context windows. Google provides access via both consumer products (Workspace/Gmail integration) and developer APIs (Google AI Studio and Vertex AI).

Developer API pricing. Google’s *Gemini API* has a tiered structure with **free** and **paid** levels. The free tier offers limited access (free tokens up to a generous cap) and is intended for experimentation (ai.google.dev). For production usage, Google charges by token in its “Paid Tier.” The pricing tables (Google AI developer site) are complex. For **Gemini 3.1 Pro Preview**, Standard paid-tier pricing is **\$2.00 per million input tokens** and **\$12.00 per million output tokens** for prompts of 200,000 tokens or fewer. For prompts above 200,000 tokens, the rates are **\$4.00 input** and **\$18.00 output** per million tokens. Batch pricing is \$1.00/\$6.00 at or below 200,000 tokens and \$2.00/\$9.00 above that threshold. ([Gemini API pricing](#)) Google also charges extra for context caching, search grounding, etc., but these are secondary.

For the mid-tier, **Gemini 2.5 Flash** is priced at \$0.30 input (text, image, video) and \$2.50 output per million (free tier covers some usage) (ai.google.dev), the legacy **Gemini 3 Flash Preview** is \$0.50/\$3.00, **Gemini 3.5 Flash-Lite** is \$0.30/\$2.50, and **Gemini 3.8 Flash** is \$0.75/\$3.75 on the Standard tier through December 31, 2026, rising to \$1.50/\$7.50 from January 1, 2027. ([Gemini API pricing](#)) Because Google’s pricing page is very detailed, Table 3 focuses on the key models for comparison with other top-tier systems.

Google also has separate line items for Gemini’s image, video, and audio variants. For instance, Gemini’s **image generation** (Imagen 4) was priced by image: \$0.02–\$0.06 per image output depending on size (ai.google.dev), which equated to roughly \$0.04 per thousand tokens in output at standard resolution

(ai.google.dev). Google shut down the Imagen 4 models on August 17, 2026 ([Gemini API changelog](#)); image generation now runs on Nano Banana models, with Nano Banana 2.1 output equivalent to about \$0.034 per 1K image on the Standard tier ([Gemini API pricing](#)). Live API (a streaming mode for Gemini) was earlier listed at \$0.35 text input and \$1.50 text output (ai.google.dev); Gemini 3.8 Live is now \$0.75 per million text input tokens and \$4.50 per million text output tokens ([Gemini API pricing](#)). In summary, **Gemini's API spans a wide price range**: Gemini 3.1 Pro Preview's \$12 output rate is twice Grok 4.7's \$6, while Gemini 3.8 Flash and Gemini 3.5 Flash-Lite cost less per token than Grok 4.7.

Business subscriptions and promotions. Beyond raw token billing, Google has introduced bundled plans. In October 2025 Google launched **Gemini Enterprise**, a \$30 per user-per-month subscription that gives businesses unified access to all Google AI tools (Gemini models, agent builders, data integrations) (www.axios.com). For instance, retailer Gap adopted Gemini Enterprise to speed up product trend analysis using AI (www.axios.com). This subscription approach differs from per-token pricing by charging a fixed per-seat price, subject to the plan's usage terms, which appeals to enterprises wanting predictability.

Promotional deals have also been used. Reuters reported that Reliance Jio in India teamed up with Google to give **18 months free of Gemini 2.5 Pro** to its subscribers (www.reuters.com). The package (Gemini 2.5 Pro + 2TB cloud storage) was valued at ₹35,100 (~\$399), implying a base rate of roughly \$22/month if paid. This mirrors similar initiatives at a lower price point: OpenAI offered a year of free ChatGPT Go (₹399 per month) to Indian users (www.reuters.com) (www.reuters.com). In the education sector, Google gave Gemini Pro free to select students (www.reuters.com). Such promos indicate Google is willing to subsidize Gemini usage to build market share, at least regionally.

Historical context and model lineage. Google's pricing approach reflects its AI ecosystem. Early on, Google offered its early AI tools for free or included in Workspace. With the Gemini API, Google introduced paid tiers and adjusted them as models evolved. (www.axios.com). For example, Gemini 2.5 Pro's Standard output rate of \$10 per million tokens (prompts up to 200,000 tokens) is well below GPT-4's \$60 per million output launch rate and above Grok 4.7's \$6. Commentators note that Google provides extensive free quotas and bundles of services (search, translation, etc.) which are not directly billed in the token price. According to the developer documentation, usage of Google Search grounding is free up to a point: for Gemini 2.5 models, 1,500 requests per day, then \$35 per 1,000 grounded prompts; for Gemini 3 and newer models, 5,000 requests per month, then \$14 per 1,000 requests ([Gemini API pricing](#)) (ai.google.dev).

Developer perspectives. Industry analysts suggest Google's model targets enterprise and cloud customers. The metered pricing is predictable for large-batch tasks (e.g. batch is 50% less than real-time) (ai.google.dev). However, the complexity of tariffs (different rates by context length, free vs paid tiers) can confuse small developers; Google counters this with an *AI Studio* interface and cost estimator tools (www.techradar.com). Notably, Google bills context-cache storage per hour, for example \$0.50 per million tokens per hour for Gemini 3.8 Flash through December 31, 2026 and \$4.50 for Gemini 3.1 Pro Preview ([Gemini API pricing](#)) (ai.google.dev), which allows interactive agents but adds pricing complexity.

Pricing summary. For Gemini 3.1 Pro Preview, Standard rates are \$2.00 per million input tokens and \$12.00 per million output tokens for prompts of 200,000 tokens or fewer; above that threshold, they are \$4.00 input and \$18.00 output. Output pricing therefore rises by 50%, not twofold. Table 3 also lists Gemini 3.8 Flash

Standard at \$0.75/\$3.75 through December 31, 2026 (Google deprecated Gemini 3.7 Flash on October 8, 2026 and routes its requests to Gemini 3.8 Flash), so comparisons should name the exact model and service tier. ([Gemini API pricing](#))

Current Gemini API pricing. Google's live pricing page is the source of record: it lists current and preview models and differentiates Standard, Batch, Flex, and Priority service tiers. Use that page to confirm the applicable model, modality, token type, tier, effective date, and any context-length threshold before estimating cost. ([Gemini API pricing](#))

OpenAI GPT-4o and ChatGPT Pricing

GPT-4o and ChatGPT. This comparison uses GPT-4o as OpenAI's named API benchmark, matching the article title. GPT-4o is priced at **\$2.50 per million input tokens** and **\$10.00 per million output tokens**. Newer OpenAI models may be relevant for separate workload-specific evaluations, but they should not replace GPT-4o as the article's primary OpenAI comparison point without retitling the article.

This article does not use an unsupported "current flagship" claim for OpenAI. Its named OpenAI benchmark is GPT-4o, for which OpenAI lists **\$2.50 per million input tokens** and **\$10.00 per million output tokens**. Model selection and pricing should be checked on [OpenAI's GPT-4o model page](#) immediately before use.

In practical terms, GPT-4o's listed \$2.50/\$10 per million input/output tokens equates to \$0.0025 and \$0.01 per 1,000 tokens. Grok 4.6 is listed at \$2/\$6 per million, but these rate cards alone do not establish a performance, reliability, or value ranking between the models.

Subscription tiers. Besides API tokens, OpenAI offers end-user subscriptions to ChatGPT. In this comparison, these are not APIs but remain relevant to a cost-conscious analysis. OpenAI currently has a "Plus" plan at **\$20 per month** (with expanded access subject to usage limits) (www.reuters.com), and a high-end "Pro" plan at **\$200 per month** offering higher limits and priority access; OpenAI's help center now lists Pro tiers at \$100, \$200, and \$500 per month ([OpenAI Help Center](#)) (www.reuters.com). Reuters reported that OpenAI expects about 220 million paying ChatGPT users by 2030, with 8.5% of users on these paid tiers (www.reuters.com). In mid-2025 OpenAI also introduced a new lower-tier "ChatGPT Go" (₹399 ≈ \$4.54 in India) for affordability (www.reuters.com).

These subscription prices illustrate OpenAI's direct monetization strategy. For comparison, Anthropic's Claude Pro is \$20 and X's Premium+ was \$22 after a December 2024 increase (www.anthropic.com) (www.reuters.com); X now lists Premium+ at \$40 per month in the U.S. ([X Help Center](#)). OpenAI and Anthropic both sell \$20/mo individual plans. OpenAI's Pro at \$200/mo (www.reuters.com) is matched by Claude Max \$200 (20× usage) (www.reuters.com). Overlapping pricing suggests a convergence: in April 2025, Anthropic launched its \$200 "Max" plan explicitly **"aligned with OpenAI's \$200 per month rate for ChatGPT Pro"** (www.reuters.com).

API usage details. This comparison focuses on OpenAI's own platform pricing. OpenAI distinguishes input, cached-input, and output tokens for GPT-4o; its listed rates are \$2.50 input, \$1.25 cached input, and \$10.00 output per million tokens.

OpenAI's current price list spans a wide range: GPT-6 Luna at \$0.10/\$0.50, GPT-5.2 at \$1.75/\$14, GPT-6.1 Sol at \$2/\$10, and GPT-6 Astra at \$10/\$50 per million input/output tokens ([OpenAI API pricing](#)). Industry blogs have tracked the decline from GPT-4's launch prices, noting cumulative price reductions exceeding 90% from the original GPT-4 launch pricing ([www.cursor-ide.com](#)).

Pricing context. GPT-4o's published API price is \$2.50 per million input tokens, \$1.25 per million cached-input tokens, and \$10.00 per million output tokens. This article does not infer a general price trend or a current-flagship designation from the GPT-4o rate card.

OpenAI's documentation lists per-1M-token costs for every model, which makes rates transparent but spending harder to predict than a flat subscription for variable loads. OpenAI lists additional service tiers, including Batch and Flex, so developers should budget against the exact tier they use ([OpenAI API pricing](#)). TechRadar's report on usage-based pricing concerns the X platform API rather than OpenAI ([www.techradar.com](#)).

Enterprise contracts: Large enterprise deployments can include negotiated terms, so while per-token pricing is the baseline, real-world contracts may combine usage pricing with subscription or flat components. Indeed, OpenAI itself appears to be seeking more predictable revenue; Reuters notes OpenAI is exploring "new monetization strategies" beyond token fees ([www.reuters.com](#)).

Anthropic Claude Pricing

Claude model families and token pricing. Anthropic's current first-party price card identifies **Table 5.1, Opus 5.5, Sonnet 5.5, and Haiku 5.5** as the current models ([Claude models overview](#)). The table below lists their base input and output prices; it also identifies cached-input pricing so estimates do not confuse it with standard input pricing.

Model (Claude)	Base input (\$/1M tokens)	Cached input (\$/1M tokens)	Output (\$/1M tokens)
Claude Fable 5.1	\$10.00	\$0.25	\$50.00
Claude Opus 5.5	\$4.00	\$0.20	\$20.00
Claude Sonnet 5.5	\$2.00	\$0.10	\$10.00
Claude Haiku 5.5 (prompts up to 100K tokens; \$0.50 / \$0.05 / \$2.50 above)	\$0.10	\$0.01	\$0.50

Table 2: Claude API prices per million tokens, checked October 10, 2026. ([Claude Platform pricing](#))

Anthropic offers consumer and team plans in addition to API access. This comparison focuses on the published API token rates; consult [Anthropic's official pricing page](#) for the applicable plan features, usage limits, and any additional charges before estimating a workload.

Subscription plans. For individual users, Anthropic lists **Claude Pro** at **\$20 per month**, or **\$17 per month when billed annually** (\$200 paid up front). The plan includes additional usage and paid-plan features; usage limits vary by activity and model. ([Anthropic pricing](#)) For higher usage needs, in April 2025 Anthropic launched

Claude Max: a \$100/month tier (5× the usage of Pro) and a \$200/month tier (20× usage) (www.reuters.com). These power-user plans explicitly match OpenAI's \$200 ChatGPT Pro: *"Anthropic's new \$200 plan provides 20× the usage of the standard plan... aligned with OpenAI's \$200 per month rate for ChatGPT Pro"* (www.reuters.com).

On the enterprise side, Anthropic offers Team and Enterprise plans. Team Standard is **\$20 per seat per month when billed annually** (\$25 monthly), while Team Premium is **\$100 per seat per month when billed annually** (\$125 monthly). ([Anthropic pricing](#)) We focus on the core token pricing here, but note that team/enterprise subscriptions bundle Claude access similarly to Google's and Microsoft's enterprise AI offerings.

Context limits. Context windows and output limits vary by model and product surface. Current Claude models offer a 1M-token context window, and Claude 4.6 and later models other than Haiku 5.5 bill the full window at standard per-token rates. Confirm the selected Claude model's current limits and pricing in Anthropic's official documentation before estimating a long-context workload. ([Claude Platform pricing](#))

Cost implications. Anthropic's current listed base rates range from Haiku 5.5 at \$0.10/\$0.50 per million input/output tokens (for prompts up to 100,000 tokens) to Fable 5.1 at \$10/\$50; Sonnet 5.5 is \$2/\$10 and Opus 5.5 is \$4/\$20. Prompt-cache reads and writes have separate rates. Model choice should be based on documented requirements and workload-specific evaluation, not price alone. ([Claude Platform pricing](#))

Worked example. At the legacy Sonnet 4.6 rates of \$3 per million input tokens and \$15 per million output tokens, 25 million input tokens plus 15 million output tokens cost about **\$300**: \$75 for input and \$225 for output. The same volume on Sonnet 5.5 (\$2/\$10) costs about **\$200**: \$50 for input and \$150 for output. ([Claude Platform pricing](#))

Moreover, Anthropic offers tools to tie spending to business value. One case study by CloudZero (a FinOps platform) detailed that \$12,000 in Claude usage could be broken down per model and per feature, providing ROI visibility (www.cloudzero.com). This reflects a broader theme: as one report notes, *"CFOs get the clarity to measure ROI...Engineering gets the guardrails to innovate without overspending"* (www.cloudzero.com). Such cost tracking matters for any large API workload.

Comparative Pricing Analysis

Having outlined each platform individually, we now analyze their prices side-by-side. The key findings include:

- **API vs. Dedicated GPU Compute:** For organizations processing tens of millions of tokens daily, per-token API charges can quickly exceed the cost of self-hosted open-source models running on reserved cloud instances. Evaluating this crossover point requires comparing API token rates against hourly instance costs; see the [H100 GPU Cloud Pricing Comparison for AWS, GCP, and Azure](#) for current rental benchmarks.
- **Token Pricing (Table 3):** The table uses one explicitly named current model and one explicitly named service tier per provider. These rate cards are not directly interchangeable: xAI applies a long-context threshold, Google lists multiple service tiers and time-limited rates, OpenAI separately prices cached

input, and Anthropic separately prices cache reads and writes. The figures therefore support workload-specific calculations, not a provider-wide price or quality ranking. ([xAI API pricing](#); [Gemini API pricing](#); [GPT-4o pricing](#); [Claude Platform pricing](#))

- Context Window:** Large context windows are now available from all four providers. Gemini 3.8 Flash accepts up to 1,048,576 input tokens ([Gemini 3.8 Flash model page](#)), Grok 4.7 has a 500,000-token window, and xAI's Grok 4.3 and Grok 4.20 models offer 1M tokens ([xAI API pricing](#)) ([x.ai](#)) ([ai.google.dev](#)). OpenAI's GPT-6 Astra, GPT-6.1 Sol, and GPT-6 Luna have 1,050,000-token windows ([GPT-6 Astra model page](#)), and current Claude models offer 1M tokens ([Claude models overview](#)). Long-context surcharges differ: xAI and Gemini 3.1 Pro Preview charge higher rates once a prompt passes 200K tokens, OpenAI's GPT-6 models above 272K input tokens, while Anthropic bills Claude 4.6 and later models other than Haiku 5.5 at standard rates across the full window ([Claude Platform pricing](#)).
- Free Tier & Bundles:** Google provides a generous free tier for experimentation (free tokens in AI Studio, search/foundation usage) ([ai.google.dev](#)) ([ai.google.dev](#)). X's usage-based pilot for its platform data API offered credits to selected participants ([www.techradar.com](#)). OpenAI has some free quota (especially via ChatGPT's portal), but its API free credits are modest. Anthropic's free tier is very limited. On the subscription side, both OpenAI and Anthropic offer value plans (\$20/mo), while Google's bundle (\$30/user) and X's Premium+ (\$40 in the U.S., with Grok) include chat access/integration. Microsoft's Copilot plans are not covered here. In summary, **Google and X give significant free or bundled access**, whereas OpenAI/Anthropic rely more on usage fees outside their paid tiers.
- Historical Trends:** Per-token prices have generally fallen, although the newest flagship tiers carry higher list prices. OpenAI's GPT-4 launched at \$30/\$60 per million input/output tokens for the 8K-context model (or \$60/\$120 for 32K); today GPT-6 Luna is \$0.10/\$0.50 and GPT-6.1 Sol is \$2/\$10, while the flagship GPT-6 Astra is \$10/\$50 ([OpenAI API pricing](#)). Google has expanded its lineup from a single Gemini Pro to Pro, Flash, and Flash-Lite models, with Standard output prices from \$0.40 (Gemini 2.5 Flash-Lite) to \$12 (Gemini 3.1 Pro Preview, prompts up to 200K tokens) per million tokens ([Gemini API pricing](#)). Anthropic has cut prices across its tiers: Opus fell from \$15/\$75 (Opus 4.1) to \$4/\$20 (Opus 5.5), Sonnet from \$3/\$15 (Sonnet 4.6) to \$2/\$10 (Sonnet 5.5), and Haiku 5.5 starts at \$0.10/\$0.50, while Fable 5.1 adds a higher tier at \$10/\$50 ([Claude Platform pricing](#)). xAI has kept Grok 4.5, 4.6, and 4.7 at the same \$2/\$6 short-context rate ([xAI API pricing](#)).
- Cost per Use Case:** A per-query estimate must name the exact model and include its input/output-token mix, prompt length, caching, tools, and service tier. The figures in Table 3 can be used to calculate a workload-specific estimate, but they do not support a provider-wide cost ranking.

The table below summarizes the per-token pricing of each service (as collated from the above sources):

Service	Model and service tier	Input (\$/M tokens)	Cached input (\$/M tokens)	Output (\$/M tokens)	Qualification
xAI Grok	Grok 4.7, short context	\$2.00	\$0.50	\$6.00	Long-context pricing applies to all tokens when the prompt reaches 200K tokens; Grok 4.6 has the same rates

Service	Model and service tier	Input (\$/M tokens)	Cached input (\$/M tokens)	Output (\$/M tokens)	Qualification
Google Gemini	Gemini 3.8 Flash, Standard	\$0.75	\$0.075	\$3.75	Listed through December 31, 2026; \$1.50 input and \$7.50 output from January 1, 2027
Google Gemini	Gemini 3.1 Pro Preview, Standard	\$2.00	\$0.20	\$12.00	Prompts up to 200K tokens; \$4.00 input and \$18.00 output above
OpenAI	GPT-4o	\$2.50	\$1.25	\$10.00	Model named in this article's title
OpenAI	GPT-6 Astra	\$10.00	\$1.00	\$50.00	Flagship; prompts over 272K input tokens cost 2x input and 1.5x output
OpenAI	GPT-6.1 Sol	\$2.00	\$0.10	\$10.00	Prompts over 272K input tokens cost 2x input and 1.5x output
OpenAI	GPT-6 Luna	\$0.10	\$0.01	\$0.50	Prompts over 272K input tokens cost 2x input and 1.5x output
Anthropic Claude	Haiku 5.5	\$0.10	\$0.01	\$0.50	Prompts up to 100K tokens; \$0.50 input and \$2.50 output above
Anthropic Claude	Sonnet 5.5	\$2.00	\$0.10	\$10.00	Current Claude model
Anthropic Claude	Opus 5.5	\$4.00	\$0.20	\$20.00	Current Claude model
Anthropic Claude	Fable 5.1	\$10.00	\$0.25	\$50.00	Current Claude model

Table 3: Selected API price-card entries per 1 million tokens (USD), checked October 10, 2026. Rates are not a provider-wide cost or quality ranking. (xAI API pricing; Gemini API pricing; GPT-4o pricing; OpenAI API pricing; Claude Platform pricing)

Analysis: These selected reference rates differ by model and service tier, so they do not identify a provider-wide cost leader. The comparison uses Grok 4.7's short-context \$2/\$6 rates and explicitly notes its 200K-token threshold; Gemini 3.8 Flash Standard at \$0.75/\$3.75 through December 31, 2026 and Gemini 3.1 Pro Preview at \$2/\$12; GPT-4o at \$2.50/\$10 with \$1.25 cached input, plus GPT-6 Astra at \$10/\$50, GPT-6.1 Sol at \$2/\$10, and GPT-6 Luna at \$0.10/\$0.50; and Claude's current Fable 5.1, Opus 5.5, Sonnet 5.5, and Haiku 5.5 rates. (xAI API pricing; Gemini API pricing; GPT-4o pricing; OpenAI API pricing; Claude Platform pricing)

Aside from token rates, there are differences in subscription and plan pricing (Table 4) which further color the landscape:

Plan/Subscription	Price	Description / Who it's for
OpenAI ChatGPT Plus	\$20/month (ChatGPT pricing)	Individual plan; features and model access vary by plan and are subject to usage limits

Plan/Subscription	Price	Description / Who it's for
OpenAI ChatGPT Pro	From \$100/month; \$100, \$200, or \$500 tiers (ChatGPT pricing) (OpenAI Help Center)	Individual plan with expanded usage; consult the official plan page for applicable limits and features
OpenAI ChatGPT Go (IN)	₹399 (~\$4.6)/month (www.reuters.com)	Lower-tier for Indian market (\$4.6)
xAI/X (Premium+)	\$40/month in the U.S. (X Help Center); raised to \$22 in December 2024 (www.reuters.com)	X's top social-media tier, includes Grok & ad-free
xAI (SuperGrok Heavy)	\$300/month (www.windowcentral.com)	Luxurious multi-agent Grok 4 Heavy subscription
Claude Pro (Personal)	\$20/month; \$17/month when billed annually (Anthropic pricing)	Individual plan with paid-plan features and variable usage limits
Claude Max (Personal)	\$100–\$200/month (www.reuters.com)	5× or 20× usage of Pro, aligned with OpenAI Pro
Claude Team (Standard)	\$25/user/month, or \$20/user/month when billed annually (Anthropic pricing)	Team plan; confirm current seat minimums and features with Anthropic
Claude Team (Premium)	\$125/user/month, or \$100/user/month when billed annually (Anthropic pricing)	Higher-usage Team plan; consult Anthropic for included features and limits
Gemini Enterprise	\$30/user/month (www.axios.com)	Per-seat business access to Google AI tools and Gemini models
Google Cloud (Vertex AI)	Varies (see Google)	Token-based pricing as above (Gemini via Vertex)

Table 4: Example subscription prices for AI services (USD). Standard plans for consumers and businesses.

Note how the consumer-friendly \$20/mo tier repeats for ChatGPT and Claude; X Premium+, which includes Grok, now costs \$40 in the U.S. Google's analogous user price appears as \$30 in the enterprise product. The \$100 and \$200/mo plans are common to OpenAI and Anthropic (targeting “power users” or small teams). Above them sit OpenAI's \$500/mo Pro 500 tier and xAI's SuperGrok Heavy, launched at \$300/mo as a premium niche product (www.windowcentral.com).

Case Studies and Real-World Examples

Multiple real-world examples shed light on how pricing affects adoption:

- Government contracts:** In September 2025, the U.S. General Services Administration approved “*Grok for Government*”, letting all federal agencies use Grok (including Grok 4) at **\$0.42 per agency** for 18 months (www.reuters.com) (www.tomshardware.com). Given the negligible fee, this is effectively a trial or token cost arrangement. By comparison, ChatGPT was offered at \$1 per agency/year. This case shows a promotional government pricing strategy: xAI priced Grok near zero for federal use to win large-scale deployment. (It clocks as the longest AI contract under the OneGov initiative (www.tomshardware.com).) Yet it also provoked pushback over Grok's content; civil rights groups

warned that Grok had “offensive posts and ideological bias” which worry them (www.tomshardware.com). In short, the government example underscores how a very low price encouraged adoption, even amid concerns about reliability or fairness.

- **Corporate partnerships:** The telecom firm Reliance Jio’s deal with Google illustrates market competition via pricing. Jio gave its users **18 months of Gemini Pro free** (value ≈ ₹35,100 = \$399) (www.reuters.com). This was aimed at capturing users in India. Similarly, OpenAI gave Indians a year of ChatGPT Go for free (www.reuters.com). These promotions, worth roughly \$22 per month (Gemini via Jio) and about \$4.50 per month (ChatGPT Go), show that companies are willing to effectively subsidize subscription costs to grow market share.
- **Developer/Startup usage:** For a chatbot startup processing **15 million input tokens plus 15 million output tokens per month**, Grok 4.7’s (and Grok 4.6’s) listed standard text-token rates imply about **\$120** ($15 \times \$2$ input + $15 \times \$6$ output). GPT-4o’s listed rates imply about **\$187.50** ($15 \times \$2.50$ input + $15 \times \$10$ output). At the same volume, GPT-6.1 Sol and Claude Sonnet 5.5 each imply about **\$180** ($15 \times \$2$ input + $15 \times \$10$ output), and Gemini 3.8 Flash Standard implies about **\$67.50** ($15 \times \$0.75$ input + $15 \times \$3.75$ output) through December 31, 2026. This illustration assumes no caching, tools, batch discounts, or other charges; a real estimate must use the exact model and service tier.
- **Consulting industry observations:** Consulting reports and articles have noted these pricing dynamics. A developer blog tracking ChatGPT API prices described API costs becoming a product constraint for teams in 2025 (www.cursor-ide.com). Indeed, many companies have set up “AI FinOps” to track token spend, and some CFO surveys confirm this: 71% of companies struggle to monetize AI effectively (www.techradar.com). The report recommends “*adopting value-based and usage-based pricing*” for AI products (www.techradar.com), a direction also seen in X’s move to usage-based pricing for its platform API (www.techradar.com).
- **Technical performance vs. price:** While price is paramount, organizations also consider quality. Comparisons published since late 2025 suggest newer Claude and Gemini models match or exceed certain GPT capabilities at similar or lower cost (www.tomsguide.com). If a project needs the most capable tier, Claude Fable 5.1 or GPT-6 Astra (each \$50 per million output tokens) may be justified. If cost is the constraint, lower-priced models such as Claude Haiku 5.5, GPT-6 Luna, Gemini 3.5 Flash-Lite, or Gemini 3.8 Flash may be sufficient.

Comparative Discussion and Future Implications

Key trade-offs: Token prices are only one component of total cost. Organizations should compare the documented capabilities, context limits, latency, tool charges, caching, service tier, security controls, and terms of the specific models under consideration; the rate cards do not establish a general hierarchy of capability or reliability. For bulk processing (e.g. indexing large document corpora), cheapest models (even open-source ones not covered here) might be best. For mission-critical reasoning (legal, medical), they may opt to pay a premium. The interplay of price and performance will continue to shape market share.

Ecosystem and lock-in: Another factor is ecosystem. Google's pricing is part of a larger cloud platform: heavy Gemini users on Google Cloud might get discounted bulk rates or bundled usage. OpenAI via Microsoft Azure or enterprise contracts may offer special terms. xAI's Grok, being tied to the X/Twitter culture, may appeal to clients aligned with that ecosystem. Pricing alone is a major axis but integration and service level agreements matter too.

Industry trends: We see a broader trend of accelerating price competition. OpenAI has repeatedly reduced costs across generations: GPT-6 Luna at \$0.10/\$0.50 and GPT-6.1 Sol at \$2/\$10 are a fraction of GPT-4's \$30/\$60 launch price, although the flagship GPT-6 Astra is priced higher at \$10/\$50. Google bundled more free features (like free search grounding) and free token allowances to make Gemini more attractive. Anthropic's current range runs from \$0.10/\$0.50 (Haiku 5.5) to \$10/\$50 (Fable 5.1). Meanwhile, xAI has kept Grok's short-context rate at \$2/\$6 across Grok 4.5, 4.6, and 4.7.

Furthermore, the servers and GPUs powering these models are major costs. With newer hardware (as of 2025, e.g. next-gen AI chips) improving throughput, providers may achieve lower marginal costs and pass savings to users. Indeed, CTOs have hinted they can lower prices as hardware improves and competition intensifies. However, regulatory backlash and public scrutiny (e.g. anti-trust, biases) may also affect business models and pricing structures.

Developer response: Early reports suggest mixed reactions. Coverage of X's usage-based platform API noted credits for selected developer participants (www.techradar.com). TechRadar notes that costs under X's usage-based plan can vary by usage pattern compared with the previous flat-rate plan (www.techradar.com). Transparency of token rates and predictable spending caps will be crucial for trust. Token bills have led companies to invest in cost optimization (prompt tuning, summarization layers, etc.), an expense in time and engineering. Some teams are experimenting with hybrid models: e.g. using lower-priced or open-source models for bulk work and reserving higher-priced models for special tasks.

Regulatory and ethical considerations: Pricing may even influence content moderation. The U.S. administration's adoption of Grok suggests that low-cost models get a foot in the door, but safety concerns remain (www.tomshardware.com). Similarly, the EU AI Act, whose obligations are being phased in, may require high-impact uses of models (regardless of cost) to meet strict standards. If compliance adds overhead, providers might bundle "safety" tools (like content filters) into higher pricing tiers. Conversely, one could imagine open or low-cost models being more heavily scrutinized if they disseminate misinformation.

Future outlook: Model lineups and API prices change frequently. As of October 10, 2026, xAI identifies Grok 4.7 as its flagship, OpenAI features GPT-6 Astra, GPT-6.1 Sol, and GPT-6 Luna, Google's newest Flash model is Gemini 3.8 Flash, and Anthropic's current lineup is Fable 5.1, Opus 5.5, Sonnet 5.5, and Haiku 5.5, with earlier Claude models still listed on its pricing page ([OpenAI models](#); [xAI model pricing](#); [Claude pricing](#)). Looking ahead, continued hardware improvements and competitive pressures suggest prices will keep declining. The next frontier may involve value-based pricing models, where providers share in the business outcomes their AI enables rather than charging purely per token.

Another factor is **usage growth**. As AI becomes embedded, some models might adopt **value-based pricing**. For example, if an AI can demonstrably increase sales or reduce labor, providers might experiment with sharing in the upside rather than pure per-token fees. This is speculative, but fits with the CFO survey advice to align pricing with business value (www.techradar.com).

Lastly, the rise of open-source alternatives (not covered here) may pressure these commercial APIs. Some studies suggest open models can be deployed at far lower cost (www.itpro.com). If enterprises embrace open models on custom hardware, it could cap the maximum price corporations are willing to pay for closed APIs. In response, companies like OpenAI and Google may pivot to specialized services (fine-tuning, private LLM hosting, advanced safety features) to justify higher prices.

Conclusion

API pricing for Grok, Gemini, OpenAI models, and Claude varies materially by model, context length, service tier, caching, tools, and workload. As checked October 10, 2026, official rate cards list Grok 4.7 at \$2/\$6 for short-context requests, GPT-4o at \$2.50/\$10, GPT-6.1 Sol at \$2/\$10, Gemini 3.8 Flash Standard at \$0.75/\$3.75 through December 31, 2026, and Claude Sonnet 5.5 at \$2/\$10. Model availability and API prices should be verified on each provider's official pricing page immediately before use. ([xAI API pricing](#); [GPT-4o pricing](#); [Gemini API pricing](#); [Claude Platform pricing](#))

For token-level charges, consult the official provider documentation linked in the relevant tables. Comparisons should specify the exact models, context lengths, and service tiers because a single provider-wide cost ranking would be misleading. Subscription-wise (Table 4), the \$20 to \$40 range is standard for individual users, with specialized tiers (\$100 to \$500) for heavy or enterprise users. We cited real-world examples (US government licensing (www.reuters.com), telecom/AI partnerships (www.androidcentral.com) (www.reuters.com)) to illustrate these numbers in context.

Key insights include:

- *Usage-based metered pricing is now the norm.* All vendors charge per token rather than only flat fees. This aligns with best practices urged by analysts (www.techradar.com) (www.techradar.com). Nevertheless, complexity of effective pricing plans can vary widely.
- *List price is not a quality ranking.* The selected models have different capabilities, context windows, terms, and service tiers. Evaluate documented requirements and workload-specific results rather than inferring quality, reliability, or compliance from token prices.
- *Pricing drives adoption strategy.* Companies like Google and OpenAI have introduced regional plans (ChatGPT Go, Jio's Gemini) to capture price-sensitive markets (www.reuters.com) (www.reuters.com). Anthropic and OpenAI each launched \$200/mo "power" plans to cater to heavy users (www.reuters.com) (www.reuters.com). X has also offered developer credits and adjusted pricing (www.techradar.com).
- *Future costs likely falling.* Evidence suggests a continued trend of price reductions across all providers. GPT-6 Luna's \$0.10 input rate is a fraction of GPT-4's \$30 launch price. Providers are likely to continue lowering barriers as competition heats up. However, the end of Moore's Law and expensive new AI chips could moderate how fast prices can fall.

In conclusion, organizations evaluating these APIs should compare the exact models relevant to their workload and account for capability, latency, context limits, compliance, and ecosystem requirements. Current model and price information should be verified directly with each provider before selection, because model availability and rates change frequently.

Provider selection should be based on the specific model, documented capabilities, total cost, service terms, and the organization's requirements. Enterprises will mix and match: for low-cost bulk jobs, they might use Grok, Gemini 3.8 Flash, or GPT-6 Luna; for core products requiring top quality, they might budget for GPT-6.1 Sol, GPT-6 Astra, or Claude Opus 5.5 or Fable 5.1. The implications are clear: costs shape who builds with AI and how it is deployed, so pricing remains a crucial competitive arena in the generative AI race (www.techradar.com) (www.techradar.com).

Use the official provider documentation linked in the relevant sections to verify prices, availability, limits, and terms immediately before publication or procurement.

References: Sources are linked inline throughout the article. Rates and availability should be checked against the official provider documentation immediately before publication or procurement.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <https://developers.openai.com/api/docs/pricing>
2. <https://platform.claude.com/docs/en/about-claude/pricing>
3. <https://docs.x.ai/developers/release-notes>
4. <https://ai.google.dev/gemini-api/docs/changelog>
5. <https://ai.google.dev/gemini-api/docs/pricing>
6. <https://docs.x.ai/developers/pricing>
7. <https://developers.openai.com/api/docs/models/gpt-4o>
8. <https://intuitionlabs.ai/articles/oracle-openai-300b-deal-analysis>
9. <https://help.openai.com/en/articles/9793128-what-is-chatgpt-pro>
10. <https://www.reuters.com/technology/openai-projected-least-220-million-people-will-pay-chatgpt-by-2030-information-2025-11-26/>
11. <https://www.reuters.com/technology/artificial-intelligence/anthropic-intensifies-ai-competition-with-200-plan-claude-models-2025-04-09/>
12. <https://claude.ai/chat/pricing/>
13. <https://help.x.com/en/using-x/x-premium>
14. <https://www.reuters.com/technology/elon-musks-x-lifts-price-premium-plus-tier-pay-creators-2024-12-23/>
15. <https://www.axios.com/2025/10/09/google-gemini-enterprise-subscription>
16. <https://www.reuters.com/world/us/musks-xai-provide-grok-chatbot-us-federal-agencies-2025-09-25/>
17. <https://www.tomshardware.com/tech-industry/artificial-intelligence/elon-musks-grok-ai-to-be-used-by-us-government-at-a-price-of-42-cents-per-agency-trump-admin-joining-meta-openai-in-recent-trend-of-ai-govt-contracts>
18. <https://www.reuters.com/world/india/indias-reliance-jio-users-get-18-month-free-access-googles-gemini-ai-pro-2025-10-30/>
19. <https://www.techradar.com/pro/security/global-bean-counters-are-struggling-to-find-value-for-money-in-anything-ai-and-that-is-a-big-big-problem>
20. <https://www.techradar.com/pro/will-xs-usage-based-api-pricing-succeed-in-winning-over-developers>
21. <https://intuitionlabs.ai/articles/mistral-large-3-moe-llm-explained>

22. <https://www.androidcentral.com/apps-software/ai/elon-musks-xai-is-paying-telegram-usd300-million-to-integrate-grok-ai>
23. <https://docs.x.ai/docs/models>
24. <https://www.axios.com/2023/12/06/google-gemini-bard-race-against-openai>
25. <https://ai.google.dev/pricing>
26. <https://www.reuters.com/business/media-telecom/openai-offer-chatgpt-go-free-year-india-2025-10-28/>
27. <https://www.axios.com/2025/07/17/elon-musk-grok-xai-fox-news-money>
28. <https://x.ai/api>
29. <https://www.windowcentral.com/artificial-intelligence/xais-usd300-month-grok-4-billed-as-a-maximally-truth-seeking-ai-seemingly-solicits-elon-musks-opinion-on-controversial-topics>
30. <https://www.socialmediatoday.com/news/x-formerly-twitter-launches-usage-based-api-access-charges/803315/>
31. <https://www.anthropic.com/pricing>
32. <https://www.cursor-ide.com/blog/chatgpt-api-prices>
33. <https://platform.claude.com/docs/en/models/overview>
34. <https://claude.com/pricing>
35. <https://www.cloudzero.com/blog/claude-pricing>
36. <https://gpusmith.com/articles/en/h100-gpu-cloud-pricing-comparison>
37. <https://ai.google.dev/gemini-api/docs/models/gemini-3.8-flash>
38. <https://developers.openai.com/api/docs/models/gpt-6-astra>
39. <https://chatgpt.com/pricing/>
40. <https://www.tomsguide.com/ai/anthropic-has-pulled-ahead-of-gemini-and-chatgpt-with-its-new-update-heres-why>
41. <https://developers.openai.com/api/docs/models>
42. <https://docs.x.ai/developers/models>
43. <https://www.itpro.com/software/open-source/open-source-ai-performance-cost-savings-proprietary-models-linux-foundation>

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore [AI enablement](#), [pharma and biotech software development](#), [data engineering and BI](#), and [Veeva services](#). [Contact IntuitionLabs](#) to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: <https://intuitionlabs.ai>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.