IntuitionLabs

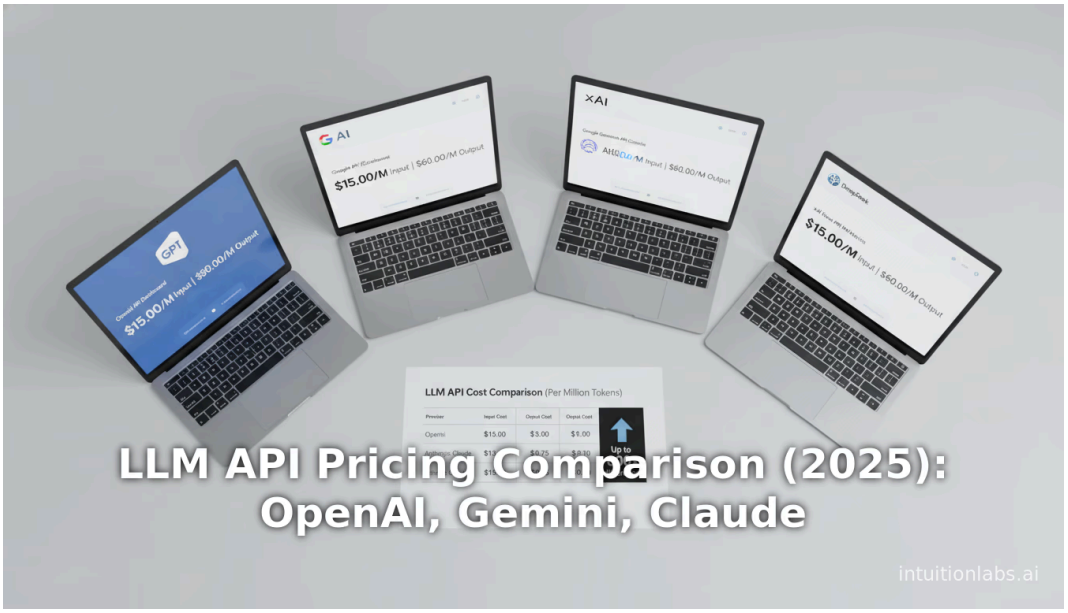
INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

LLM API Pricing 2026: OpenAI, Gemini, Claude & Grok

> 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 · 2025-10-31 · llm api pricing · openai pricing · google gemini · anthropic claude · ai cost analysis · token pricing · gpt-5 · grok api · deepseek · ai

Original article: https://intuitionlabs.ai/articles/llm-api-pricing-comparison-2025



In this article

1. LLM API Pricing 2026: OpenAI, Gemini, Claude & Grok
2. Executive Summary
3. Introduction and Background
4. OpenAI API Models and Pricing
5. Google Gemini API and Pricing
6. Anthropic Claude API Pricing
7. xAI Grok Pricing
8. DeepSeek Historical Pricing Snapshot: V3.2-Exp (September 2025)
9. Comparative Analysis
10. Case Studies and Example Scenarios
11. Pricing Trends and Future Implications
12. Conclusion

> **Last updated: October 10, 2026.** Pricing and model availability should be verified against each provider's official documentation immediately before purchase. Originally published October 2025.

LLM API Pricing 2026: OpenAI, Gemini, Claude & Grok

Executive Summary

Update (October 2026): Every provider in this comparison changed its catalog or prices after the August 2026 revision. OpenAI now lists the GPT-6 family as its latest models: gpt-6-astra at $10.00 per 1M input tokens and $50.00 per 1M output tokens, gpt-6.1-sol at $2.00/$10.00, and gpt-6-luna at $0.10/$0.50. gpt-5.6-sol dropped to a promotional $4.00/$20.00, available at least through November 21, 2026. Anthropic's current models are Claude Fable 5.1 ($10.00/$50.00), Claude Opus 5.5 ($4.00/$20.00), Claude Sonnet 5.5 ($2.00/$10.00), and Claude Haiku 5.5 ($0.10/$0.50 for prompts up to 100,000 tokens). Google added Gemini 3.8 Flash at $0.75/$3.75 through December 31, 2026, rising to $1.50/$7.50 on January 1, 2027, and Gemini 3.5 Flash-Lite at $0.30/$2.50. xAI describes grok-4.7, at $2.00/$6.00, as its most capable model. DeepSeek retired V4-Flash: its deepseek-flash model (DeepSeek-V4.1-Flash) costs $0.30/$1.20 at peak cache-miss rates and half that off-peak. For high-volume work, the lowest published rates in this comparison are GPT-5 nano at $0.05/$0.40, Gemini 2.5 Flash-Lite at $0.10/$0.40, gpt-6-luna and Claude Haiku 5.5 at $0.10/$0.50, and off-peak deepseek-flash at $0.15/$0.60 on a cache miss. All figures are standard short-context rates per 1M input/output tokens, checked on October 10, 2026. (OpenAI API pricing) (Anthropic pricing) (Gemini Developer API pricing) (xAI models) (DeepSeek API pricing)

By 2026, the landscape of large language model (LLM) APIs has grown intensely competitive and complex, with multiple providers offering a spectrum of model capabilities and pricing options. This report provides a detailed comparison of the API pricing for all major LLM models from OpenAI, Google’s Gemini, Anthropic’s

Claude, Elon Musk's xAI Grok, and China's DeepSeek, synthesizing the latest public information. The current provider tables in this report use the catalog and standard rates published on October 10, 2026: OpenAI's GPT-6 and GPT-5.6 families; Google's Gemini 3.8 Flash, Gemini 3.1 Pro Preview, Gemini 3.5 Flash-Lite, and Gemini 3.1 Flash-Lite; Anthropic's Claude Fable 5.1, Opus 5.5, Sonnet 5.5, and Haiku 5.5; xAI's grok-4.7; and DeepSeek's deepseek-flash (V4.1-Flash) and V4-Pro. Current rates are not directly rankable without matching the input/output mix, cache status, context band, processing tier, and, where applicable, DeepSeek's peak or off-peak period. Historical models and prices are identified as such below. (OpenAI API pricing) (Gemini Developer API pricing) (Anthropic pricing) (xAI pricing) (DeepSeek API pricing)

These differences mean that for the same task, costs can vary by orders of magnitude depending on model choice. Throughout this report, we explore the historical evolution of these pricing schemes, compare them quantitatively in tables, and discuss implications for developers and businesses choosing among providers. We also include case studies illustrating real-world cost impacts, cite independent analyses of cost-cutting trends, and outline how aggressive pricing (especially by open-source-focused players) may reshape the future economics of AI. All figures and claims here are backed by official documentation and recent technology news sources (www.reuters.com) (techcrunch.com) (docs.anthropic.com) (api-docs.deepseek.com).

Introduction and Background

Large language models have revolutionized numerous industries by enabling advanced natural language understanding and generation. Today, LLM APIs are used for chatbots, coding assistants, document summarization, translation, and more. Unlike earlier AI models, modern LLMs expose sophisticated *token-based* billing: every API call's input and output text (measured in tokens, roughly word pieces) incurs cost, aligning with a "pay-as-you-go" cloud model (www.binadox.com). This token-based pricing offers fine-grained control of costs but demands careful model selection and prompt engineering.

The major LLM providers now include **OpenAI**, **Google** (via its Gemini models), **Anthropic**, **xAI (Grok)**, and **DeepSeek**. Each has multiple model variants optimized for different trade-offs (accuracy, speed, context length). OpenAI's latest GPT-6 family includes gpt-6-astra, gpt-6.1-sol, and gpt-6-luna, and its price list still carries earlier GPT-5.x models with "Pro", "mini", and "nano" editions for different price/performance points (openai.com). Anthropic's Claude line is tiered (Haiku, Sonnet, Opus), and Google's Gemini offers "Pro" (large, multi-modal) vs "Flash" (lighter, cheaper) versions (cloud.google.com) (docs.anthropic.com). xAI's API launched with Grok 3 at $3 per million input tokens and $15 per million output tokens in April 2025; its current catalog describes grok-4.7 as its most capable model (techcrunch.com) (xAI models). DeepSeek's current API price page lists deepseek-flash (DeepSeek-V4.1-Flash) and deepseek-v4-pro; the legacy deepseek-v4-flash name is now served by V4.1-Flash. Both support thinking and non-thinking modes, have a 1M-token context length, and use cache-status and peak/off-peak pricing. (DeepSeek API pricing)

Pricing is a critical differentiator. All use per-token pricing, but rates vary widely by model. Generally, more capable models cost more per token. Historical context shows rapid evolution: for example, OpenAI halved its GPT-3.5 Turbo token price in 2023, then introduced GPT-4 (at ~10× the cost of 3.5), and in 2024 launched GPT-4o mini at just $0.15/$0.60 per million input/output, a 60% discount vs GPT-3.5 Turbo (www.reuters.com). Similarly, Chinese startups like DeepSeek entered the market with dramatically lower pricing (DeepSeek R1 debuted at $0.55/$2.19 per million, undercutting competitors by ~90% (www.techradar.com)). These shifts

reflect aggressive competition and have set new baselines. The historical discussion below should not be treated as a live provider catalog. The current model names, availability, context limits, and prices should be checked against the official provider pages linked in this article before a purchasing decision.

OpenAI API Models and Pricing

OpenAI offers several model families and service tiers. As of October 10, 2026, its pricing page lists the GPT-6 family (gpt-6-astra, gpt-6.1-sol, and gpt-6-luna) as its latest models; prices differ by model, service tier, cache status, and whether a request uses long context (more than 272K input tokens). (OpenAI API pricing)

- **gpt-6-astra**: The highest-priced model in OpenAI's latest GPT-6 family.
- *Input:* $10.00 per 1M tokens.
- *Cached input:* $1.00 per 1M tokens.
- *Output:* $50.00 per 1M tokens.
- **gpt-6.1-sol**: The mid-priced GPT-6 model.
- *Input:* $2.00 per 1M tokens.
- *Cached input:* $0.10 per 1M tokens.
- *Output:* $10.00 per 1M tokens.
- **gpt-6-luna**: The lowest-priced GPT-6 model.
- *Input:* $0.10 per 1M tokens.
- *Cached input:* $0.01 per 1M tokens.
- *Output:* $0.50 per 1M tokens.
- **GPT-5.2**: An earlier GPT-5 flagship that remains on OpenAI's price list, aimed at reasoning and agentic tasks.
- *Input:* $1.75 per 1M tokens.
- *Cached input:* $0.175 per 1M tokens.
- *Output:* $14.00 per 1M tokens.
- **GPT-5.2 Pro**: The premium tier of GPT-5.2 for maximum capability.
- *Input:* $21.00 per 1M tokens.
- *Output:* $168.00 per 1M tokens.
- **GPT-5 mini**: A smaller, cheaper model for well-defined tasks.
- *Input:* $0.25 per 1M.
- *Cached:* $0.025 per 1M.
- *Output:* $2.00 per 1M.
- **GPT-5 nano**: The smallest and cheapest variant.
- *Input:* $0.05 per 1M.
- *Cached:* $0.005 per 1M.
- *Output:* $0.40 per 1M (openai.com) (openai.com).

For example, sending 100K input tokens and receiving 100K output tokens on gpt-5.6-terra at the standard short-context rate costs $0.20 for input plus $1.20 for output, or $1.40 before any cache, service-tier, or regional-processing adjustments. Historical model prices should be treated as historical snapshots rather than a current catalog.

For comparison and thoroughness, below is a summary table of OpenAI's key current models and rates:

Model (OpenAI)	Description	Input ($/M tokens)	Output ($/M tokens)	Notes
gpt-6-astra	Highest-priced GPT-6 model	$10.00	$50.00	Standard short-context rate
gpt-6.1-sol	Mid-priced GPT-6 model	$2.00	$10.00	Standard short-context rate
gpt-6-luna	Lowest-priced GPT-6 model	$0.10	$0.50	Standard short-context rate
gpt-5.6-sol	Previous flagship family	$4.00	$20.00	Promotional standard short-context rate, at least through November 21, 2026
gpt-5.6-terra	Lower-cost GPT-5.6 model	$2.00	$12.00	Standard short-context rate
gpt-5.6-luna	Lowest-cost GPT-5.6 model	$0.20	$1.20	Standard short-context rate

Source: OpenAI API pricing. Note: fine-tuning, Batch, Flex, Fast mode, long-context, caching, and regional-processing prices can differ from the standard short-context rates shown above.

OpenAI's pricing has trended sharply downward over the past two years. The GPT-4o mini launched in mid-2024 at $0.15/$0.60 per MTok, a 60% reduction from GPT-3.5 Turbo (www.reuters.com). The GPT-6 family continues that trend at the low end: gpt-6-luna is listed at $0.10/$0.50, below the gpt-5.6-luna rate of $0.20/$1.20, while gpt-6-astra at $10.00/$50.00 is the highest-priced model in the family. Organizations should compare applicable context bands, cached-input rates, and service tiers rather than treating historical GPT-5 prices as current.

Google Gemini API and Pricing

Google provides its LLMs through **Vertex AI** (Google Cloud) under the "Gemini" brand. Google's Gemini API catalog changes frequently; as of October 2026 it describes Gemini 3.8 Flash as its most intelligent Flash model, alongside Gemini 3.1 Pro Preview and the 3.x Flash-Lite models. Google uses different standard, batch, Flex, and Priority prices for some models; Gemini 3.1 Pro Preview also changes rates when a prompt exceeds 200,000 tokens. (Gemini API pricing)

- **Gemini 2.5 Pro**:
- *Input tokens:* $1.25 per 1M (for ≤200K input), $2.50 per 1M (for >200K).
- *Text output (response & reasoning):* $10 per 1M (≤200K input), $15 per 1M (>200K) (cloud.google.com).
- **Gemini 2.5 Flash** (multimodal smaller model):
- *Text/Image/Video input:* $0.30 per 1M.
- *Audio input:* $1.00 per 1M.

- *Text output:* $2.50 per 1M (cloud.google.com).

The Gemini Developer API lists **Gemini 3 Flash Preview** at paid-tier **Standard** rates of $0.50 per 1M text/image/video input tokens ($1.00 for audio) and $3.00 per 1M output tokens, including thinking tokens. **Gemini 3.1 Flash-Lite** is listed at $0.25 per 1M text/image/video input tokens ($0.50 for audio) and $1.50 per 1M output tokens under Standard pricing. Preview models can change before general availability. Gemini 3.8 Flash is listed at $0.75 per 1M input tokens and $3.75 per 1M output tokens through December 31, 2026, and at $1.50/$7.50 from January 1, 2027; Gemini 3.6 Flash carries the same rates. Gemini 3.5 Flash-Lite costs $0.30 per 1M input tokens (text, image, video, or audio) and $2.50 per 1M output tokens. Google now labels Gemini 3 Flash Preview a legacy model. (Gemini Developer API pricing)

In plain terms, calling Gemini 2.5 Pro with a moderate prompt (e.g. 100K tokens) costs about $0.125 (input) + $1.00 (output) = $1.125 per 100K tokens, doubling if the prompt exceeds 200K. For Gemini 2.5 Flash, the cost is much lower: only $0.03 input + $0.25 output per 100K. Google's strategy offers a "Pro" model competitive with premium pricing and a very cheap "Flash" model for volume tasks. Gemini 2.5 Pro below thresholds at $1.25/$10 per 1M is lower than the $3/$15 that xAI charged for Grok 3 at its April 2025 API launch (www.linkedin.com).

Google also allows **"grounding"** with Google Search/Web to enrich responses, billed up to $35 per 1K grounded queries (cloud.google.com). However, the base token costs above generally suffice for textual tasks.

Google Gemini Pricing Summary (per 1M tokens):

Model	Input (≤200K / >200K)	Output	Notes
Gemini 3.8 Flash	$0.75	$3.75	Paid-tier Standard rate through December 31, 2026; $1.50 input and $7.50 output from January 1, 2027
Gemini 3.5 Flash-Lite	$0.30 text/image/video/audio	$2.50	Published paid-tier Standard rate
Gemini 3 Flash Preview	$0.50 text/image/video; $1.00 audio	$3.00	Published paid-tier Standard rate; preview model Google now labels legacy
Gemini 3.1 Pro Preview	$2.00 / $4.00	$12.00 / $18.00	Published paid-tier rates; both change above 200K prompt tokens
Gemini 3.1 Flash-Lite	$0.25 text/image/video; $0.50 audio	$1.50	Published paid-tier Standard rate
Gemini 2.5 Pro	$1.25 / $2.50	$10.00 / $15.00	Published paid-tier rates; both change above 200K prompt tokens
Gemini 2.5 Flash	$0.30 text/image/video; $1.00 audio	$2.50	Published paid-tier standard rate
Gemini 2.5 Flash-Lite	$0.10 text/image/video; $0.30 audio	$0.40	Published paid-tier standard rate

Source: Gemini Developer API pricing (paid-tier rates checked October 10, 2026) and Google Cloud Vertex AI pricing docs (cloud.google.com). Notes: Above 200K input tokens, both input and output rates increase for Pro models.

In addition to token-pricing, Google's Gemini (like Anthropic) often benefits from integration discounts if using Google Cloud infrastructure. Overall, direct per-token comparisons require matching the input/output mix, prompt-length band, cache status, processing tier, and platform. The current published rates for Gemini 3.1 Pro Preview and OpenAI's GPT-6 family should be checked against their official pricing pages before making a purchasing decision. (Gemini Developer API pricing) (OpenAI API pricing)

Anthropic Claude API Pricing

Anthropic's Claude family (*Haiku*, *Sonnet*, *Opus*) targets safety and reliability. Anthropic uniquely offers **prompt caching discounts** in its pricing model (where repeat queries get cheaper). As of October 2026, Anthropic's current models are Claude Fable 5.1, Claude Opus 5.5, Claude Sonnet 5.5, and Claude Haiku 5.5, and its price list still carries the earlier Claude 5 models. The base, non-cache rates below are first-party Claude API rates; prompt caching, Batch API use, data residency, and Fast mode can change effective pricing. (Anthropic model pricing)

- **Claude Fable 5.1**:
- Base input: $10 per 1M tokens.
- Output: $50 per 1M tokens.
- **Claude Opus 5.5**:
- Base input: $4 per 1M tokens.
- Output: $20 per 1M tokens.
- **Claude Sonnet 5.5**:
- Base input: $2 per 1M tokens.
- Output: $10 per 1M tokens.
- **Claude Haiku 5.5**:
- Base input: $0.10 per 1M tokens for prompts up to 100,000 tokens ($0.50 above).
- Output: $0.50 per 1M tokens for prompts up to 100,000 tokens ($2.50 above).
- **Earlier models still listed**: Claude Fable 5 at $10/$50, Claude Opus 5 at $5/$25, Claude Sonnet 5 at $2/$10 (Anthropic made its launch pricing the standard price), and Claude Haiku 4.5 at $1/$5 per 1M input/output tokens.

Anthropic's price page also lists other available and limited-availability models; the comparison below includes Fable alongside the Opus, Sonnet, and Haiku tier representatives. (Anthropic pricing)

These rates assume standard (5-minute) caching or direct usage. Anthropic explains that "5m cache writes" cost 1.25× base input, and "cache reads" cost 0.1× base on most models, 0.05× on Claude Opus 5.5 and Sonnet 5.5, and 0.025× on Claude Fable 5.1 (docs.anthropic.com).

In simplified terms, Claude Opus 5.5 is priced at $4/$20 per million tokens, Claude Sonnet 5.5 at $2/$10, and Claude Haiku 5.5 at $0.10/$0.50 for prompts up to 100,000 tokens. Direct price comparisons require the same input/output mix, cache status, context band, service tier, and platform; DeepSeek's current rates also vary by

cache status and peak/off-peak period.

For illustrative clarity, one can also view Claude batch rates (50% off) or long-context premium (detailed on Anthropic's site); however, the base token costs above capture the primary differences.

Anthropic Claude Pricing Summary (per 1M tokens):

Model	Input (base)	Output (base)	Notes
Claude Fable 5.1	$10.00	$50.00	Anthropic's recommended model for the most demanding reasoning work
Claude Opus 5.5	$4.00	$20.00	Current Opus model
Claude Sonnet 5.5	$2.00	$10.00	Current Sonnet model
Claude Haiku 5.5	$0.10 / $0.50	$0.50 / $2.50	Current small model; higher rates apply to prompts over 100,000 tokens
Claude Opus 5	$5.00	$25.00	Previous Opus model
Claude Sonnet 5	$2.00	$10.00	Previous Sonnet model
Claude Haiku 4.5	$1.00	$5.00	Previous small model

Source: Anthropic API pricing docs (docs.anthropic.com). Note **cache writes** (~x1.25-2x) and **cache hits** (~x0.1 on most models, lower on Fable 5.1, Opus 5.5, and Sonnet 5.5) impact effective cost for repeated prompts, and Anthropic offers a 50% discount on input/output under its Batch API (not shown above for brevity).

In practice, published base input/output rates for Claude Opus 5.5, Sonnet 5.5, and Haiku 5.5 (prompts up to 100,000 tokens) are $4/$20, $2/$10, and $0.10/$0.50 per million tokens, respectively. Cache, Batch API, data-residency, Fast mode, and any applicable context-band charges can change the effective bill. Teams can route workloads among tiers based on their quality, latency, and cost requirements. (Anthropic pricing)

xAI Grok Pricing

xAI (Elon Musk's AI startup) began public Grok releases in 2023. The earlier Grok 3 and Grok 4 entries are not the current xAI catalog. xAI's documentation describes `grok-4.7` as its most capable model and recommends it for code and chat. Its standard short-context rate is $2.00 per 1M input tokens, $0.50 cached input, and $6.00 output, with a 500,000-token context window. Once a prompt reaches 200,000 tokens, xAI applies its long-context rates to all tokens in the request. (xAI pricing) (xAI models)

Current xAI pricing should be compared only with like-for-like assumptions. grok-4.7's long-context rates are $4.00 per 1M input tokens, $1.00 cached input, and $12.00 output, and the US regional endpoint bills grok-4.7 and grok-4.6 at 1.1 times the global rates. grok-4.3 offers a 1M-token context window at $1.25 input and $2.50 output, and it is one of the models eligible for xAI's 20% Batch API discount. (xAI pricing)

Grok Pricing Summary (per 1M tokens):

Model (xAI Grok)	Input	Output	Context (tokens)	Notes
grok-4.7	$2.00	$6.00	500,000	Most capable current model; standard short-context rate
grok-4.6	$2.00	$6.00	500,000	Previous flagship; standard short-context rate
grok-4.3	$1.25	$2.50	1,000,000	Lower-cost 1M-context model; standard short-context rate
grok-build-0.1	$1.00	$2.00	256,000	Current build-oriented model; standard short-context rate

Source: xAI API pricing. All prices are USD per 1M tokens; long-context rates, caching, and tool charges may differ.

Direct price rankings require like-for-like assumptions about context band, cache status, service tier, and input/output mix. xAI's pricing page should be checked for current model availability and tool charges before choosing a model.

DeepSeek Historical Pricing Snapshot: V3.2-Exp (September 2025)

DeepSeek is a Chinese AI startup that gained fame for open and affordable models. Its initial DeepSeek-R1 (open-source) hit the market in January 2025, offering GPT-4 class performance at a fraction of the usual price (www.techradar.com). DeepSeek then evolved through its V3 series. In September 2025, **DeepSeek-V3.2-Exp (Experimental)** was documented with the following historical Chat and Reasoner rates and a 128K-token context window. DeepSeek's current catalog and pricing are deepseek-flash (V4.1-Flash) and V4-Pro; see the comparative table for their current peak cache-miss rates. (DeepSeek API pricing)

- **Input tokens:** $0.028 per 1M (cache hit), $0.28 per 1M (cache miss) (api-docs.deepseek.com).
- **Output tokens:** $0.42 per 1M (api-docs.deepseek.com).

DeepSeek introduced a **cache mechanism**: if a prompt (or subprompt) is already used ("cache hit"), input cost is only $0.028/M; otherwise $0.28/M. This means repeated queries become almost free, incentivizing stateful use. Notably, DeepSeek **cut all its prices by ~50% in Sep 2025** (www.reuters.com) compared to the prior V3.2-beta (the Reuters piece states "reducing API pricing by over 50%" (www.reuters.com)).

For context, DeepSeek's original R1 model had pricing around $0.55/$2.19 (input/output) (www.techradar.com). Thus V3.2-Exp's $0.28/$0.42 rates were dramatically lower than the GPT-5.2 prices used in this historical comparison; they are not comparable to the current GPT-6 catalog. TechRadar reports: *"DeepSeek-R1 debuted at $0.55 input/$2.19 output"* and comments that LLM pricing "has cratered" since (www.techradar.com). DeepSeek's new prices validate that trend.

In September 2025, these V3.2-Exp rates were unusually low, but they are not a basis for a current universal API price ranking. For example, processing 1M tokens of input and 1M output (a huge request) would cost just $0.28 + $0.42 = $0.70 total (cache-miss). Even with repeated use (cache hits), cost is only $0.028 + $0.42 = $0.448. By comparison, OpenAI's cheapest model (GPT-5 nano) would cost $0.05 + $0.40 = $0.45 (similar), and GPT-5 nano's 400K-token context window (OpenAI GPT-5 nano model) is larger than the 128K context DeepSeek listed for V3.2-Exp (api-docs.deepseek.com).

DeepSeek V3.2-Exp Pricing Summary (per 1M tokens):

Mode	Input (cache-hit)	Input (cache-miss)	Output	Context
DeepSeek Chat	$0.028	$0.28	$0.42	128K tokens
DeepSeek Reasoner	$0.028	$0.28	$0.42	128K tokens

Source: DeepSeek API documentation (Sept 2025) (api-docs.deepseek.com). Note: caching heavily affects input pricing.

The aggressive pricing is a deliberate strategy. As quoted by Reuters, DeepSeek aims "to match or exceed rivals' performance at reduced costs" to "solidify its position" (www.reuters.com). Industry commentary underscores this "price war": Chinese models (DeepSeek, Baidu's Ernie) have pushed token costs near zero, challenging Western providers (www.techradar.com). In fact, TechRadar notes *"China is commoditizing AI faster than the West can monetize it"*, since free open models threaten the business case for paid ones (www.techradar.com).

Comparative Analysis

To compare across providers, consider example per-token costs. Table below contrasts representative **high-capacity** and **low-cost** models:

Provider	Model	Input ($/M)	Output ($/M)	Context (tokens)	Remarks
OpenAI	gpt-6-astra	$10.00	$50.00	See provider documentation	Latest-family standard short-context rate
OpenAI	gpt-6.1-sol	$2.00	$10.00	See provider documentation	Latest-family standard short-context rate
OpenAI	gpt-6-luna	$0.10	$0.50	See provider documentation	Lowest-cost latest-family standard short-context rate
OpenAI	gpt-5.6-sol	$4.00	$20.00	See provider documentation	Promotional standard short-context rate, at least through November 21, 2026
OpenAI	gpt-5.6-terra	$2.00	$12.00	See provider documentation	GPT-5.6 standard short-context rate
OpenAI	gpt-5.6-luna	$0.20	$1.20	See provider documentation	GPT-5.6 standard short-context rate
Google	Gemini 3.8 Flash	$0.75	$3.75	See provider documentation	Paid-tier Standard rate through December 31, 2026; doubles on January 1, 2027
Google	Gemini 3.1 Pro Preview	$2.00-$4.00*	$12-$18*	See provider documentation	Gemini Developer API paid-tier Standard rates
Google	Gemini 3.5 Flash-Lite	$0.30	$2.50	See provider documentation	Gemini Developer API paid-tier Standard rate

Provider	Model	Input ($/M)	Output ($/M)	Context (tokens)	Remarks
Google	Gemini 3.1 Flash-Lite	$0.25 text/image/video; $0.50 audio	$1.50	See provider documentation	Gemini Developer API paid-tier Standard rate
Google	Gemini 2.5 Pro	$1.25-$2.50*	$10-$15*	1M	Tiered pricing
Google	Gemini 2.5 Flash	$0.30	$2.50	1M	Cheap, versatile
Anthropic	Claude Fable 5.1	$10.00	$50.00	1M	Current top-tier model
Anthropic	Claude Opus 5.5	$4.00	$20.00	1M	Current Opus model
Anthropic	Claude Sonnet 5.5	$2.00	$10.00	1M	Current Sonnet model
Anthropic	Claude Haiku 5.5	$0.10*	$0.50*	1M	Current small model; prompts up to 100,000 tokens
xAI	grok-4.7	$2.00	$6.00	500K	Most capable current model; standard short-context rate
DeepSeek	deepseek-flash (V4.1-Flash)	$0.30*	$1.20	1M	Current peak cache-miss input rate
DeepSeek	deepseek-v4-pro	$1.32*	$3.96	1M	Current peak cache-miss input rate

* Gemini Pro rates rise above 200K prompt tokens and Claude Haiku 5.5 rates rise above 100,000 prompt tokens. DeepSeek's off-peak rates are half of peak rates, and cache-hit input is priced separately.

This comparison should not collapse cache-hit, cache-miss, peak/off-peak, long-context, or service-tier rates into one ranking. A cache-miss deepseek-flash input token at its published peak rate ($0.30 per 1M) is more expensive than OpenAI gpt-6-luna input ($0.10 per 1M), while DeepSeek's off-peak or cache-hit rates are lower. Evaluate a provider with the request mix and service conditions that apply to the workload.

Price per 100K tokens (example): A 100K-input, 100K-output standard short-context request costs $0.20 + $1.20 = **$1.40** on OpenAI gpt-5.6-terra. Direct cross-provider comparisons require the same stated input/output split, cache status, context band, and service tier; provider tables should not be converted into a single ranking without those assumptions.

The table is a rate reference, not a universal capability ranking or a complete bill estimate. Model choice should account for workload quality requirements as well as input/output volume, cache behavior, context length, service tier, tools, and any regional or platform-specific charges.

Case Studies and Example Scenarios

Case 1: Customer Support Chatbot. A monthly total-token estimate is insufficient to calculate an API bill because input and output are charged separately and cache status can change the input rate. For example, a workload with 8M short-context input tokens and 2M output tokens would cost $32 + $40 = $72 on OpenAI gpt-5.6-sol at its promotional standard rates, or $16 + $20 = $36 on gpt-6.1-sol. The same split would cost $16 + $24 = $40 on gpt-5.6-terra, before cache, batch, Flex, long-context, tools, or regional-processing adjustments.

Case 2: Enterprise Document Summarization. Assume 100 documents per month, each with 60K input tokens and a 6K-token summary: 6M input tokens and 0.6M output tokens total. At first-party standard rates, Claude Opus 5.5 would cost $24 for input plus $12 for output, or $36; Claude Sonnet 5.5 would cost $12 plus $6, or $18; and OpenAI gpt-5.6-terra would cost $12 plus $7.20, or $19.20. Actual bills can differ with caching, batch processing, long-context requests, or tool use.

Real-World Example: Bing Chat Integration. Microsoft documented that the new AI-enhanced Bing used GPT-4 and that it worked with the model for months before GPT-4's public release in March 2023. OpenAI introduced the distinct o1 reasoning-model family in September 2024, so it was not the model used in Bing's 2023 launch. This article therefore uses publicly posted API rates as its comparison baseline rather than making claims about undisclosed internal agreements or discounts. (Microsoft Responsible AI for the new Bing) (OpenAI: Learning to reason with LLMs)

Illustrative daily estimate. A 50,000-token daily workload needs an explicit input/output split. If it consists of 25,000 input and 25,000 output tokens, gpt-5.6-terra's published short-context rates produce $0.05 of input cost plus $0.30 of output cost, or $0.35 per day. At that same split, 100 bots running for 30 days would cost about $1,050 before caching, Batch, Flex, Fast mode, long-context, tool, or regional-processing charges. (OpenAI API pricing)

Pricing Trends and Future Implications

The rapid evolution in LLM pricing reflects both market competition and technological advances. Key trends and implications include:

- **Pricing pressure and model segmentation:** Provider pricing changes frequently, and catalogs segment by capability, latency, context length, caching, and processing mode. Current published rates show substantial variation even within one provider's model family, but they do not establish that any provider's models are "effectively free" or support a universal forecast for future price cuts. (OpenAI API pricing) (Anthropic pricing) (DeepSeek API pricing)
- **Differentiated Offerings:** Providers segment their catalogs by price, capability, context, latency, and modality. OpenAI's latest GPT-6 family includes astra, sol, and luna models; Google offers Pro and Flash-family models; Anthropic offers Opus, Sonnet, and Haiku; and xAI lists grok-4.7, grok-4.6, and grok-build-0.1. Organizations can route workloads to models that meet their quality, latency, and cost requirements rather than applying a single model everywhere. (OpenAI pricing) (Gemini Developer API pricing) (Anthropic pricing) (xAI pricing)

- **Two-Part Pricing and Ecosystem Effects:** Tiered pricing (volume discounts, batch APIs) and multi-modal features complicate direct comparisons. For instance, Google charges extra for “grounded” search queries, while Anthropic’s long-context pricing premium and caching create effective nonlinear cost curves (cloud.google.com) (docs.anthropic.com). This means vendors might not strictly compete on raw token price, but on value (e.g. freshness, search, tool use). Pricing strategy becomes part of lock-in: e.g. Google bundling lower Gemini rates for Cloud users. We already see enterprises bundling LLM spend into cloud contracts (Azure, AWS Bedrock multi-LLM billing) to manage cost predictability (www.binadox.com).
- **Use of Tokens vs. Alternatives:** Some providers experiment with non-token billing. For example, OpenAI offers a “request-based” pricing for certain APIs (e.g. Vision API in dollar per image rather than per token). Large-scale use cases might evolve hybrid pricing (e.g. fixed-price content creation subscription). However, the fundamental token model will likely remain primary for text generation at least through 2026. Tools and prompt optimization (e.g. GPT’s “cost consciousness” like avoiding ChatGPT answer verbosity) will become best practices for controlling token spend.
- **Long-Term Outlook:** Provider catalogs and prices can change quickly. Published prices alone do not establish a reliable forecast for future prices, market structure, or provider subsidy practices. Teams should monitor official pricing pages and re-estimate their own workloads when models, context bands, processing modes, or tool charges change.

In sum, **cost optimization** is now a core part of the AI developer playbook. Our analysis suggests that enterprises must carefully match model selection to use-case, balancing “best model” vs “acceptable model” based on token budgets. For example, routing routine tasks to a cheaper model and reserving the most expensive model for the tasks that need it lowers spend under pay-as-you-go token billing (www.binadox.com). The pricing data compiled here is a starting point for that analysis.

Conclusion

This October 2026 update (originally published October 2025) describes an LLM API market with **material cost differentials**. Published rates alone do not establish a provider-wide price or capability ranking: actual charges depend on the selected model, input/output mix, cache status, context length, service tier, tools, region, and platform. The DeepSeek figures in this article use the provider’s current deepseek-flash (V4.1-Flash) and V4-Pro catalog, but they still depend on cache status and whether usage occurs during peak or off-peak hours.

Crucially, these rates are dynamic. New model releases, volume agreements, and one-off promotions will further shift the playing field. We recommend that AI consumers continuously re-evaluate pricing, consider multi-provider strategies, and leverage specialized offerings (batch APIs, caching, long-context) to optimize costs.

The current model and rate references above are based on the providers’ official pricing documentation and should be rechecked immediately before publication or purchase. (OpenAI API pricing) (Anthropic pricing) (xAI pricing) (DeepSeek API pricing) Continued transparency from providers and third-party tracking will be essential for navigating the ongoing AI pricing revolution.

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://ai.google.dev/gemini-api/docs/pricing
4. https://docs.x.ai/developers/models
5. https://api-docs.deepseek.com/quick_start/pricing/
6. https://intuitionlabs.ai/articles/deepseek-inference-cost-explained
7. https://docs.x.ai/developers/pricing
8. https://www.reuters.com/technology/artificial-intelligence/openai-unveils-cheaper-small-ai-model-gpt-4o-mini-2024-07-18/
9. https://techcrunch.com/2025/04/09/elon-musks-ai-company-xai-launches-an-api-for-grok-3/
10. https://docs.anthropic.com/en/docs/about-claude/pricing
11. https://api-docs.deepseek.com/quick_start/pricing
12. https://intuitionlabs.ai/articles/glm-4-6-open-source-coding-model
13. https://www.binadox.com/blog/llm-api-pricing-comparison-2025-complete-cost-analysis-guide/
14. https://intuitionlabs.ai/articles/what-is-context-engineering
15. https://openai.com/bn-BD/api/pricing/
16. https://cloud.google.com/vertex-ai/generative-ai/pricing?hl=he
17. https://www.techradar.com/pro/why-baidus-ernie-matters-more-than-deepseek
18. https://intuitionlabs.ai/articles/agentic-ai-temporal-orchestration
19. https://www.linkedin.com/pulse/xai-launches-grok-3-api-four-pricing-tiers-intensifying-%E6%9D%B0-%E9%82%93-q1qic
20. https://docs.anthropic.com/en/docs/about-claude/pricing?%3F%3F%3F__hstc=43401018.71aa366c60c32c7e3032e45be702fadd.1753488000320.1753488000321.1753488000322.1
21. https://www.reuters.com/technology/deepseek-releases-model-it-calls-intermediate-step-towards-next-generation-2025-09-29/
22. https://developers.openai.com/api/docs/models/gpt-5-nano
23. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2023/04/RAI-for-the-new-Bing-April-2023.pdf
24. https://openai.com/index/learning-to-reason-with-llms/

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.