Latest LLM API model strings across five providers

As of February 9, 2026, here are the exact API model identifiers, endpoints, and pricing for OpenAI, Anthropic, xAI, Google, and DeepSeek. All five providers support streaming and OpenAI-compatible API formats (native or via compatibility layers), making LiteLLM orchestration straightforward. The landscape has shifted significantly in late 2025: OpenAI launched the GPT-5 family, Anthropic reached Claude Opus 4.6, xAI shipped Grok 4.1, Google previewed Gemini 3, and DeepSeek unified on V3.2 at remarkably low prices.

OpenAI: GPT-5.2 is the current flagship

Model String	Role	Input / Output (per 1M tokens)	Context
`gpt-5.2`	Flagship reasoning (Dec 2025)	**$1.75 / $14.00**	400K
`gpt-5.2-2025-12-11`	Pinned snapshot	$1.75 / $14.00	400K
`gpt-5.2-pro`	Max-compute variant	$21.00 / $168.00	400K
`gpt-5-mini`	Cost-efficient reasoning	$0.25 / $2.00	400K
`gpt-5-nano`	Cheapest reasoning	$0.05 / $0.40	400K
`gpt-4.1`	Best non-reasoning model	$2.00 / $8.00	~1M
`gpt-4.1-mini`	Fast non-reasoning	$0.40 / $1.60	~1M

Base endpoint: `https://api.openai.com/v1/chat/completions` (Chat Completions) or `https://api.openai.com/v1/responses` (newer Responses API). pricepertoken **Streaming:** fully supported via SSE (`"stream": true`). **LiteLLM prefix:** use model strings directly (e.g., `gpt-5.2`).

Key notes: GPT-5.2 supports a `reasoning_effort` parameter with values `none` , `low` , `medium` , `high` , `xhigh` OpenAI — set to `none` to use it as a fast non-reasoning model. Prompt caching gives **90% off input costs** automatically (e.g., $0.175/M cached for gpt-

5.2). The Responses API is now recommended by OpenAI AI/ML API and required for `gpt-5.2-codex` models. OpenAI Rate limits are tiered by spend; batch API available for async workloads at lower cost.

Anthropic: Claude Opus 4.6 just launched February 5

Model String	Role	Input / Output (per 1M tokens)	Context
`claude-opus-4-6`	Latest flagship (Feb 5, 2026)	**$5.00 / $25.00**	200K (1M beta)
`claude-sonnet-4-5-20250929`	Best value mid-tier	$3.00 / $15.00	200K (1M beta)
`claude-haiku-4-5-20251001`	Fast and cheap	$1.00 / $5.00	200K
`claude-opus-4-5-20251101`	Previous flagship	$5.00 / $25.00	200K
`claude-sonnet-4-20250514`	Previous Sonnet	$3.00 / $15.00	200K

Base endpoint: `https://api.anthropic.com/v1/messages` . **Streaming:** fully supported via SSE. **LiteLLM prefix:** `anthropic/claude-opus-4-6` , etc.

Required headers:

```
x-api-key: YOUR_KEY
anthropic-version: 2023-06-01
```

Key notes: Opus 4.6 uses a **simplified naming scheme** — no date suffix, just `claude-opus-4-6` . Digital Applied digitalapplied It introduces "adaptive thinking" with `thinking: {type: "adaptive"}` and effort levels (low/medium/high/max). digitalapplied For 1M context, add header `anthropic-beta: context-1m-2025-08-07` Claude API Docs (requires Tier 4). Claude API Docs Batch API at 50% discount. Claude API Docs Breaking change on Opus 4.6: **assistant prefilling is disabled** (returns 400). Digital Applied Claude API Docs Rate limits use a 4-tier system based on deposit amount, measured across RPM, ITPM, and OTPM per organization. AI Free API Also available on **AWS Bedrock** (`anthropic.claude-opus-4-6-`

`v1:0`) and **Vertex AI** (`claude-opus-4-6@latest`).

---

xAI: Grok 4.1 Fast leads an OpenAI-compatible API

Model String	Role	Input / Output (per 1M tokens)	Context
`grok-4-1-fast-reasoning`	Latest fast reasoning	**$0.20 / $0.50**	2M
`grok-4-1-fast-non-reasoning`	Latest fast non-reasoning	$0.20 / $0.50	2M
`grok-4-0709`	Flagship deep reasoning	$3.00 / $15.00	256K
`grok-3-beta`	Previous gen standard	$3.00 / $15.00	131K
`grok-3-mini-beta`	Lightweight reasoning	$0.30 / $0.50	131K
`grok-code-fast-1`	Coding specialist	$0.20 / $1.50	256K

Base endpoint: `https://api.x.ai/v1/chat/completions` . **Streaming:** fully supported. promptfoo **LiteLLM prefix:** `xai/grok-4-1-fast-reasoning` , etc.

Key notes: xAI's API is **fully OpenAI-compatible** — swap `base_url` and API key, everything else works. xAI The `grok-4-1-fast-reasoning` model offers a **2M token context window** at extremely competitive pricing ($0.20/$0.50), costgoat making it the cheapest frontier-class model per token in this comparison. AI Free API Automatic prompt caching gives 50-75% savings on repeat prefixes. Alias convention: `grok-4-1-fast` and `grok-4-1-fast-latest` both point to the reasoning variant. promptfoo xAI Grok-4 (non-fast) is always-reasoning with no `reasoning_effort` parameter; only `grok-3-mini` variants support `reasoning_effort` (`"low"` or `"high"`). xAI Docs xAI Regional endpoints available at `us-east-1.api.x.ai` and `eu-west-1.api.x.ai` . New users get **$25 in free credits**. AI Free API

Google Gemini: version 3 is in preview, 2.5 is stable

Model String	Role	Input / Output (per 1M tokens)	Context
`gemini-3-flash-preview`	Latest preview (Dec 2025)	**$0.50 / $3.00**	1M
`gemini-3-pro-preview`	Most powerful preview	$2.00 / $12.00	1M
`gemini-2.5-pro`	Stable production flagship	$1.25 / $10.00	1M
`gemini-2.5-flash`	Stable best value	$0.30 / $2.50	1M
`gemini-2.5-flash-lite`	Cheapest stable option	$0.10 / $0.40	1M

Base endpoint (Google AI): `https://generativelanguage.googleapis.com/v1beta/models/{MODEL}:generateContent?key={KEY}`. **OpenAI-compatible:** `https://generativelanguage.googleapis.com/v1beta/openai`. Microsoft Learn **Streaming:** use `:streamGenerateContent` endpoint or `stream=True` via OpenAI compat. **LiteLLM prefix:** `gemini/gemini-2.5-flash`, etc.

Key notes: Gemini 3 models are **preview only** — for production LiteLLM pipelines, `gemini-2.5-pro` and `gemini-2.5-flash` are the recommended stable choices. Gemini 2.0 models are **deprecated and shut down March 31, 2026** Firebase — migrate now. Pro models have **2x pricing beyond 200K tokens**; Flash models keep flat pricing regardless of context length. Costgoat Thinking tokens are billed as output. Free tier available for most models (rate-limited). google Rate limits are tier-based (Free → Tier 1/2/3) and viewable per-project in AI Studio. google Also accessible via **Vertex AI** for enterprise use with different endpoint format.

DeepSeek: two model strings, one stunningly low price

Model String	Role	Input / Output (per 1M tokens)	Context
`deepseek-chat`	General purpose (V3.2, non-thinking)	**$0.28 / $0.42**	128K

`deepseek-reasoner`	Chain-of-thought reasoning (V3.2, thinking)	**$0.28 / $0.42**	128K

Base endpoint: `https://api.deepseek.com/chat/completions` (also `https://api.deepseek.com/v1/chat/completions`). **Streaming:** fully supported. **LiteLLM prefix:** `deepseek/deepseek-chat`, etc.

Key notes: DeepSeek offers just two model strings — both backed by **DeepSeek-V3.2** in different modes. DeepSeek It is **fully OpenAI SDK-compatible** (change `base_url` and key). deepseek Automatic context caching drops input cost to **$0.028/M** on cache hits — a 90% discount, making this by far the cheapest provider. Costgoat No fixed rate limits; the system queues during high traffic. DeepSeek Max output is **8K tokens** for `deepseek-chat` and **64K** for `deepseek-reasoner`. Costgoat DeepSeek V4 is expected around mid-February 2026 but has not launched yet. GIGAZINE +2 **Data residency concern:** all data stored on servers in mainland China, subject to Chinese law. CODECS.COM Multiple governments have banned the consumer app on government devices, though API access itself remains globally available. New users get **5M free tokens**. Costgoat

---

Side-by-side pricing comparison for LiteLLM routing

This table compares the "best general-purpose" model from each provider — the one you'd likely use as your primary in a multi-LLM router:

Provider	Recommended Model	Input $/1M	Output $/1M	Context	Relative Cost Index
DeepSeek	`deepseek-chat`	$0.28	$0.42	128K	**1.0×** (baseline)
xAI	`grok-4-1-fast-reasoning`	$0.20	$0.50	2M	1.0×
Google	`gemini-2.5-flash`	$0.30	$2.50	1M	4.0×
OpenAI	`gpt-5.2`	$1.75	$14.00	400K	26×
Anthropic	`claude-opus-4-6`	$5.00	$25.00	200K	48×

For cost-sensitive routing, **DeepSeek and xAI Grok 4.1 Fast are roughly 25-50× cheaper on output tokens than OpenAI or Anthropic flagships.** Google's Flash models sit in between. A smart LiteLLM strategy would route simple tasks to DeepSeek/Grok Fast and

reserve GPT-5.2 or Opus 4.6 for complex reasoning.

LiteLLM configuration quick-reference

For a working LiteLLM setup calling all five in parallel, you need these environment variables and model mappings:

```
# litellm_config.yaml
model_list:
  - model_name: openai-flagship
    litellm_params:
      model: gpt-5.2
      api_key: os.environ/OPENAI_API_KEY

  - model_name: anthropic-flagship
    litellm_params:
      model: anthropic/claude-opus-4-6
      api_key: os.environ/ANTHROPIC_API_KEY

  - model_name: xai-flagship
    litellm_params:
      model: xai/grok-4-1-fast-reasoning
      api_key: os.environ/XAI_API_KEY
      api_base: https://api.x.ai/v1

  - model_name: gemini-flagship
    litellm_params:
      model: gemini/gemini-2.5-pro
      api_key: os.environ/GEMINI_API_KEY

  - model_name: deepseek-flagship
    litellm_params:
      model: deepseek/deepseek-chat
      api_key: os.environ/DEEPSEEK_API_KEY
      api_base: https://api.deepseek.com
```

All five support streaming, function calling, and JSON mode. The critical implementation detail is that **Anthropic requires the `anthropic-version: 2023-06-01` header** (LiteLLM handles this automatically), and **Gemini's native API uses a different request format** (LiteLLM translates). xAI and DeepSeek are natively OpenAI-compatible and need only a `base_url` swap. DEV Community +2