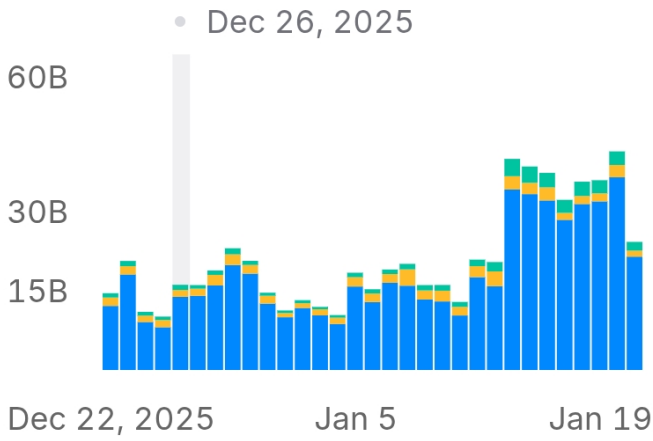
Author	openai
Context Length	131K
Reasoning	✓
Providers	21

gpt-oss-120b is an open-weight, 117B-parameter Mixture-of-Experts (MoE) language model from

Activity



Prompt	13.8B
Reasoning	1.31B
Completion	985M

Provider	DeepInfra
Latency (p50)	0.56s
Throughput (p50)	62.0 tok/s
Visualize Performance	

Pricing

Input	$0.039 / M tokens
Output	$0.19 / M tokens
Images	--

Features

Input Modalities	text
Output Modalities	text
Quantization	fp4
Max Tokens (input + output)	131K
Max Output Tokens	--
Stream cancellation	✓
Supports Tools	✓
No Prompt Training	✓

AI Model Comparison

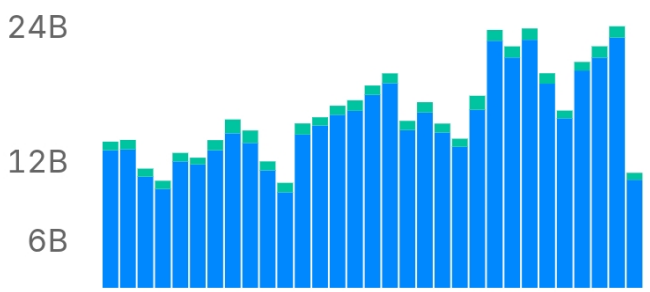
Compare GPT-4o-mini from OpenAI with other AI models on key metrics, including price, context length, and other model features.

GPT-4o-mini

Author	openai
Context Length	128K
Reasoning	✗
Providers	2

GPT-4o mini is OpenAI's newest model after GPT-4 Omni, supporting

Activity



Prompt	10.6B
Completion	527M
Reasoning	0

Provider OpenAI

Latency (p50)	0.53s
Throughput (p50)	27.0 tok/s

Visualize Performance

Pricing

Input	$0.15 / M tokens
Output	$0.60 / M tokens
Images	--

Features

Input Modalities	text, image, file
Output Modalities	text
Quantization	unknown
Max Tokens (input + output)	128K

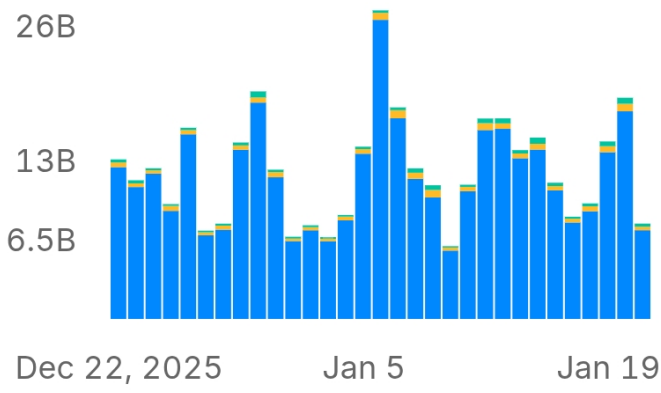
Compare GPT-5.2 from OpenAI with other AI models on key metrics, including price, context length, and other model features.

GPT-5.2

Author	openai
Context Length	400K
Reasoning	✓
Providers	2

GPT-5.2 is the latest frontier-grade model in the GPT-5 series, offering stronger agentic and long context

Activity



Provider	OpenAI
Latency (p50)	2.85s
Throughput (p50)	37.0 tok/s
Visualize Performance	

Pricing

Input	$1.75 / M tokens
Output	$14 / M tokens
Images	--

Features

Input Modalities	file, image, text
Output Modalities	text
Quantization	unknown
Max Tokens (input + output)	400K
Max Output Tokens	128K
Stream cancellation	✓
Supports Tools	✓
No Prompt Training	✓

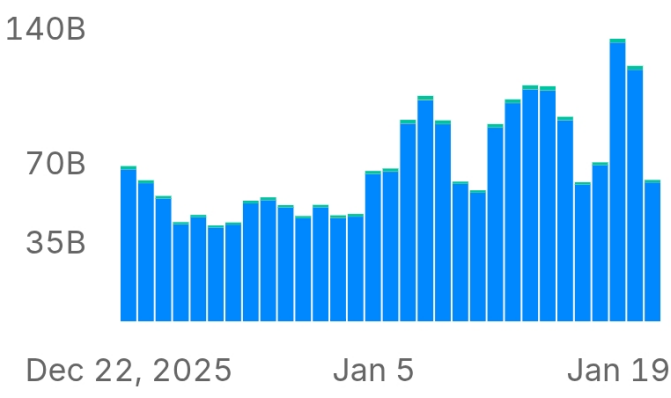
Compare Claude Sonnet 4.5 from Anthropic with other AI models on key metrics, including price, context length, and other model features.

Claude Sonnet 4.5

Author	anthropic
Context Length	1M
Reasoning	✓
Providers	3

Claude Sonnet 4.5 is Anthropic's most advanced Sonnet model to date, optimized for real-world agents

Activity



Prompt	61.4B
Completion	1.05B
Reasoning	50.5M

Provider: Anthropic

Latency (p50)	1.72s
Throughput (p50)	34.0 tok/s

Visualize Performance

Pricing

Input	$3 / M tokens
Output	$15 / M tokens
Images	--

Features

Input Modalities	text, image, file
Output Modalities	text
Quantization	unknown
Max Tokens (input + output)	1M

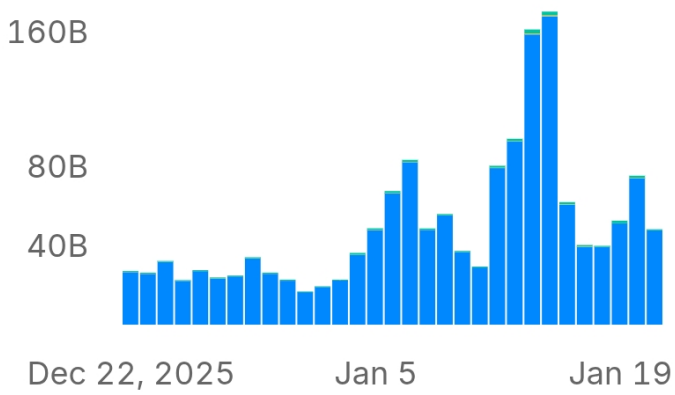
Compare Claude Opus 4.5 from Anthropic with other AI models on key metrics, including price, context length, and other model features.

Claude Opus 4.5

Author	anthropic
Context Length	200K
Reasoning	✓
Providers	3

Claude Opus 4.5 is Anthropic's frontier reasoning model optimized for complex software engineering

Activity



Prompt	47.6B
Completion	529M
Reasoning	27.7M

Provider: Anthropic

Latency (p50)	1.95s
Throughput (p50)	26.0 tok/s

Visualize Performance

Pricing

Input	$5 / M tokens
Output	$25 / M tokens
Images	--

Features

Input Modalities	file, image, text
Output Modalities	text
Quantization	unknown
Max Tokens (input + output)	200K

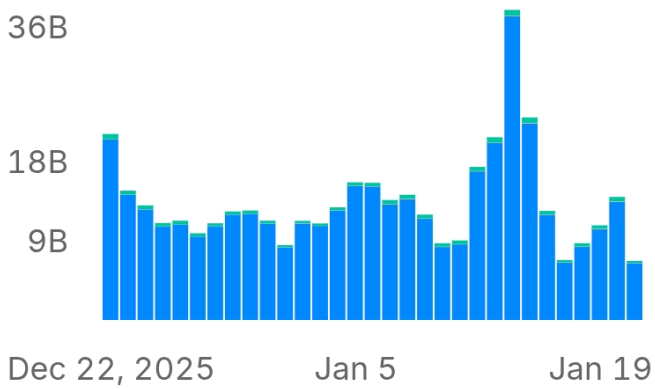
Compare Claude Haiku 4.5 from Anthropic with other AI models on key metrics, including price, context length, and other model features.

Claude Haiku 4.5

Author	anthropic
Context Length	200K
Reasoning	✓
Providers	3

Claude Haiku 4.5 is Anthropic's fastest and most efficient model, delivering near-frontier intelligence

Activity



Prompt	6.46B
Completion	228M
Reasoning	25.4M

Provider: Anthropic

Latency (p50)	0.62s
Throughput (p50)	75.0 tok/s

Visualize Performance

Pricing

Input	$1 / M tokens
Output	$5 / M tokens
Images	--

Features

Input Modalities	image, text
Output Modalities	text
Quantization	unknown
Max Tokens (input + output)	200K

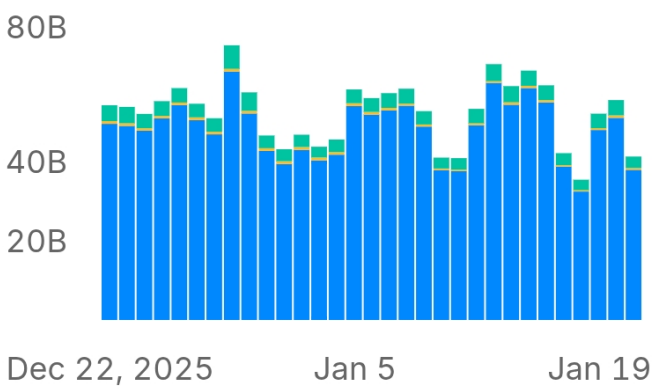
Compare Gemini 2.5 Flash from Google with other AI models on key metrics, including price, context length, and other model features.

Gemini 2.5 Flash

Author	google
Context Length	1.05M
Reasoning	✓
Providers	2

Gemini 2.5 Flash is Google's state-of-the-art workhorse model, specifically designed for advanced

Activity



Prompt	37.9B
Completion	2.89B
Reasoning	465M

Provider: Google Vertex (Global)

Latency (p50)	0.58s
Throughput (p50)	66.0 tok/s

Visualize Performance

Pricing

Input	$0.30 / M tokens
Output	$2.50 / M tokens
Images	$1.238 / K

Features

Input Modalities	file, image, text, audio, video
Output Modalities	text
Quantization	unknown
Max Tokens (input + output)	1.05M

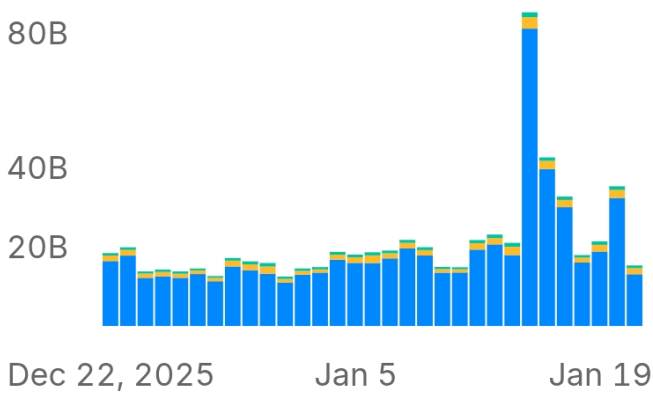
Compare Gemini 3 Pro Preview from Google with other AI models on key metrics, including price, context length, and other model features.

Gemini 3 Pro Preview

Author	google
Context Length	1.05M
Reasoning	✓
Providers	2

Gemini 3 Pro is Google's flagship frontier model for high-precision multimodal reasoning, combining

Activity



Prompt	13B
Reasoning	1.54B
Completion	678M

Provider: Google Vertex

Latency (p50)	4.31s
Throughput (p50)	76.0 tok/s

Visualize Performance

Pricing

Input	$2 / M tokens
Output	$12 / M tokens
Images	$8.256 / K

Features

Input Modalities	text, image, file, audio, video
Output Modalities	text
Quantization	unknown
Max Tokens (input + output)	1.05M

Output	$12 / M tokens
Images	$8.256 / K

Features

Input Modalities	text, image, file, audio, video
Output Modalities	text
Quantization	unknown
Max Tokens (input + output)	1.05M
Max Output Tokens	66K
Stream cancellation	✗
Supports Tools	✓
No Prompt Training	✓
Caching	✓

OpenRouter

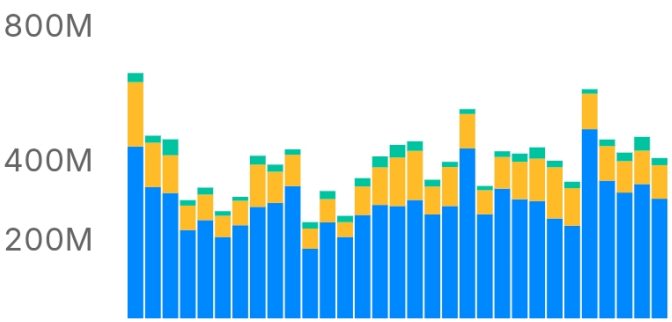
Compare Qwen3 235B A22B Thinking 2507 from Qwen with other AI models on key metrics, including price, context length, and other model features.

Qwen3 235B A22B Thinking 2

Author	qwen
Context Length	262K
Reasoning	✓
Providers	7

Qwen3-235B-A22B-Thinking-2507 is a high-performance, open-weight Mixture-of-Experts (MoE)

Activity



Dec 22, 2025 Jan 5 Jan 19

Prompt	302M
Reasoning	84.8M
Completion	18.1M

Provider Chutes

Latency (p50)	1.70s
Throughput (p50)	63.0 tok/s

Visualize Performance

Pricing

Input	$0.11 / M tokens
Output	$0.60 / M tokens
Images	--

Features

Input Modalities	text
Output Modalities	text

Max Tokens (input + output)	262K
Max Output Tokens	262K
Stream cancellation	✓
Supports Tools	✓
No Prompt Training	✗
Caching	✗

+

Add model

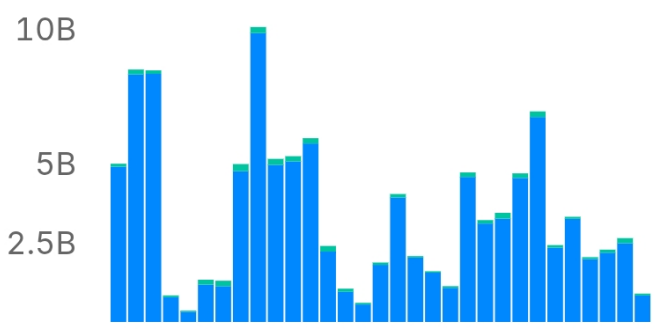
Compare Qwen3 VL 235B A22B Instruct from Qwen with other AI models on key metrics, including price, context length, and other model features.

Qwen3 VL 235B A22B Instruc

Author	qwen
Context Length	262K
Reasoning	✗
Providers	9

Qwen3-VL-235B-A22B Instruct is an open-weight multimodal model that unifies strong text generation

Activity



Prompt	837M
Completion	58.1M
Reasoning	365

Provider DeepInfra

Latency (p50)	0.74s
Throughput (p50)	22.0 tok/s

Visualize Performance

Pricing

Input	$0.20 / M tokens
Output	$1.20 / M tokens
Images	--

Features

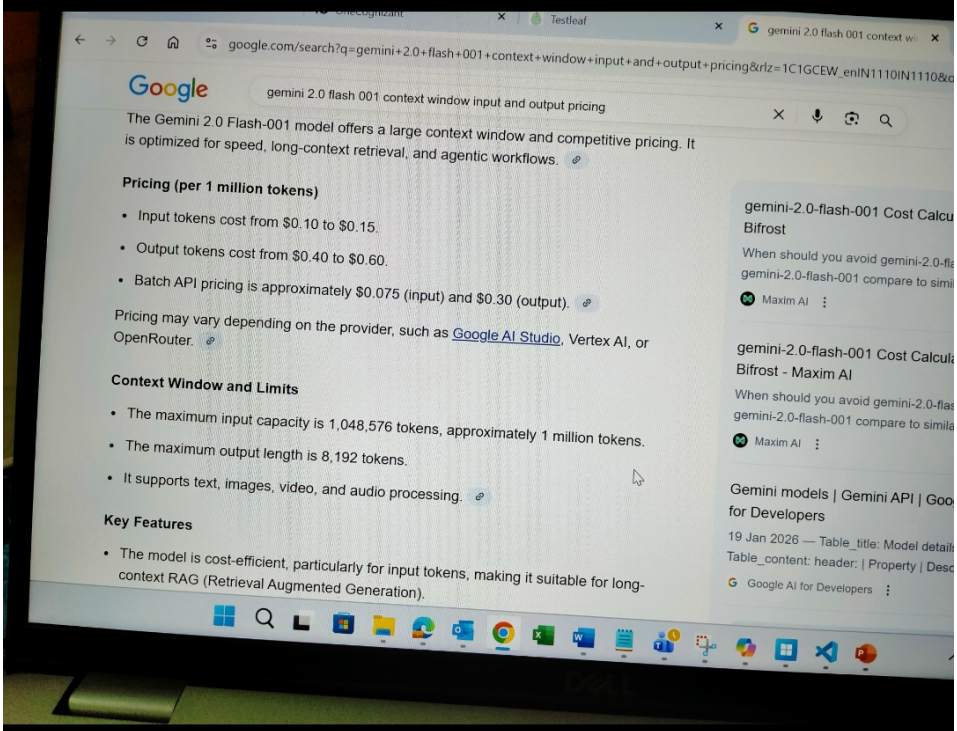
Input Modalities	text, image
Output Modalities	text
Quantization	fp8
Max Tokens (input + output)	262K

Pricing

Input	$0.20 / M tokens
Output	$1.20 / M tokens
Images	--

Features

Input Modalities	text, image
Output Modalities	text
Quantization	fp8
Max Tokens (input + output)	262K
Max Output Tokens	--
Stream cancellation	✓
Supports Tools	✗
No Prompt Training	✓
Caching	✗



Google Lens

Share Favorite Edit Delete More