Gemini Developer API Pricing

The Gemini API "free tier" is offered through the API service with lower rate limits for testing purposes. Google AI Studio usage is completely free in all available countries. The Gemini API "paid tier" comes with higher rate limits, additional features, and different data handling.

Upgrade to the Paid Tier

Gemini 2.5 Flash Preview

Try it in Google AI Studio

Our first hybrid reasoning model which supports a 1M token context window and has thinking budgets.

Preview models may change before becoming stable and have more restrictive rate limits.

	Free Tier	Paid Tier, per 1M tokens in USD
Input price	Free of charge	$0.15 (text / image / video) $1.00 (audio)
Output price	Free of charge	Non-thinking: $0.60 Thinking: $3.50
Context caching price	Not available	$0.0375 (text / image / video) $0.25 (audio) $1.00 / 1,000,000 tokens per hour
Grounding with Google Search	Free of charge, up to 500 RPD	1,500 RPD (free), then $35 / 1,000 requests
Text-to-speech (`gemini-2.5-flash-preview-tts`)	Free of charge	$0.50 (input) $10.00 (Output)
Used to improve our products	Yes	No

Gemini 2.5 Pro Preview

Try it in Google AI Studio

Our state-of-the-art multipurpose model, which excels at coding and complex reasoning tasks.

Preview models may change before becoming stable and have more restrictive rate limits.

	Free Tier	Paid Tier, per 1M tokens in USD
Input price	Not available	$1.25, prompts <= 200k tokens $2.50, prompts > 200k tokens
Output price (including thinking tokens)	Not available	$10.00, prompts <= 200k tokens $15.00, prompts > 200k
Context caching price	Not available	$0.31, prompts <= 200k tokens $0.625, prompts > 200k $4.50 / 1,000,000 tokens per hour
Grounding with Google Search	Not available	1,500 RPD (free), then $35 / 1,000 requests
Text-to-speech (`gemini-2.5-pro-preview-tts`)	Free of charge	$1.00 (input) $20.00 (Output)
Used to improve our products	Yes	No

Gemini 2.5 Flash Native Audio

Try it in Google AI Studio

Our native audio models optimized for higher quality audio outputs with better pacing, voice naturalness, verbosity, and mood.

Preview models may change before becoming stable and have more restrictive rate limits.

	Free Tier	Paid Tier, per 1M tokens in USD
Input price	Not available	$0.50 (text) $3.00 (audio / video)
Output price (including thinking tokens)	Not available	$2.00 (text) $12.00 (audio)
Used to improve our products	Yes	No

Gemini 2.5 Flash Preview TTS

Try it in Google AI Studio

Our 2.5 Flash text-to-speech audio model optimized for price-performant, low-latency, controllable speech generation.

Preview models may change before becoming stable and have more restrictive rate limits.

	Free Tier	Paid Tier, per 1M tokens in USD
Input price	Free of charge	$0.50 (text)
Output price	Free of charge	$10.00 (audio)
Used to improve our products	Yes	No

Gemini 2.5 Pro Preview TTS

Try it in Google AI Studio

Our 2.5 Pro text-to-speech audio model optimized for powerful, low-latency speech generation for more natural outputs and easier to steer prompts.

Preview models may change before becoming stable and have more restrictive rate limits.

	Free Tier	Paid Tier, per 1M tokens in USD
Input price	Not available	$1.00 (text)
Output price	Not available	$20.00 (audio)
Used to improve our products	Yes	No

Gemini 2.0 Flash

Try it in Google AI Studio

Our most balanced multimodal model with great performance across all tasks, with a 1 million token context window, and built for the era of Agents.

	Free Tier	Paid Tier, per 1M tokens in USD
Input price	Free of charge	$0.10 (text / image / video) $0.70 (audio)
Output price	Free of charge	$0.40
Context caching price	Free of charge	$0.025 / 1,000,000 tokens (text/image/video) $1.175 / 1,000,000 tokens (audio)
Context caching (storage)	Free of charge, up to 1,000,000 tokens of storage per hour	$1.00 / 1,000,000 tokens per hour
Image generation pricing	Free of charge	$0.039 per image*
Tuning price	Not available	Not available
Grounding with Google Search	Free of charge, up to 500 RPD	1,500 RPD (free), then $35 / 1,000 requests
Live API	Free of charge	Input: $0.35 (text), $0.10 (audio / image [video]) Output: $1.50 (text), $8.50 (audio)
Used to improve our products	Yes	No

[*] Image output is priced at $30 per 1,000,000 tokens. Output images up to 1024x1024px consume 1290 tokens and are equivalent to $0.039 per image.

Gemini 2.0 Flash-Lite

Try it in Google AI Studio

Our smallest and most cost effective model, built for at scale usage.

	Free Tier	Paid Tier, per 1M tokens in USD
Input price	Free of charge	$0.075
Output price	Free of charge	$0.30
Context caching price	Not available	Not available
Context caching (storage)	Not available	Not available
Tuning price	Not available	Not available
Grounding with Google Search	Not available	Not available
Used to improve our products	Yes	No

Imagen 3

Try it in Google AI Studio

Our state-of-the-art image generation model, available to developers on the paid tier of the Gemini API.

	Free Tier	Paid Tier, per image in USD
Image price	Not available	$0.03
Used to improve our products	Yes	No

Veo 2

Try the API

Our state-of-the-art video generation model, available to developers on the paid tier of the Gemini API.

	Free Tier	Paid Tier, per second in USD
Video price	Not available	$0.35
Used to improve our products	Yes	No

Gemma 3

Try Gemma 3

Our lightweight, state-of the art, open model built from the same technology that powers our Gemini models.

	Free Tier	Paid Tier, per 1M tokens in USD
Input price	Free of charge	Not available
Output price	Free of charge	Not available
Context caching price	Free of charge	Not available
Context caching (storage)	Free of charge	Not available
Tuning price	Not available	Not available
Grounding with Google Search	Not available	Not available
Used to improve our products	Yes	No

Gemma 3n

Try Gemma 3n

Our open model built for efficient performance on everyday devices like mobile phones, laptops, and tablets.

	Free Tier	Paid Tier, per 1M tokens in USD
Input price	Free of charge	Not available
Output price	Free of charge	Not available
Context caching price	Free of charge	Not available
Context caching (storage)	Free of charge	Not available
Tuning price	Not available	Not available
Grounding with Google Search	Not available	Not available
Used to improve our products	Yes	No

Gemini 1.5 Flash

Try it in Google AI Studio

Our fastest multimodal model with great performance for diverse, repetitive tasks and a 1 million token context window.

	Free Tier	Paid Tier, per 1M tokens in USD
Input price	Free of charge	$0.075, prompts <= 128k tokens $0.15, prompts > 128k tokens
Output price	Free of charge	$0.30, prompts <= 128k tokens $0.60, prompts > 128k tokens
Context caching price	Free of charge, up to 1 million tokens of storage per hour	$0.01875, prompts <= 128k tokens $0.0375, prompts > 128k tokens
Context caching (storage)	Free of charge	$1.00 per hour
Tuning price	Token prices are the same for tuned models Tuning service is free of charge.	Token prices are the same for tuned models Tuning service is free of charge.
Grounding with Google Search	Not available	$35 / 1K grounding requests
Used to improve our products	Yes	No

Gemini 1.5 Flash-8B

Try it in Google AI Studio

Our smallest model for lower intelligence use cases, with a 1 million token context window.

	Free Tier	Paid Tier, per 1M tokens in USD
Input price	Free of charge	$0.0375, prompts <= 128k tokens $0.075, prompts > 128k tokens
Output price	Free of charge	$0.15, prompts <= 128k tokens $0.30, prompts > 128k tokens
Context caching price	Free of charge, up to 1 million tokens of storage per hour	$0.01, prompts <= 128k tokens $0.02, prompts > 128k tokens
Context caching (storage)	Free of charge	$0.25 per hour
Tuning price	Token prices are the same for tuned models Tuning service is free of charge.	Token prices are the same for tuned models Tuning service is free of charge.
Grounding with Google Search	Not available	$35 / 1K grounding requests
Used to improve our products	Yes	No

Gemini 1.5 Pro

Try it in Google AI Studio

Our highest intelligence Gemini 1.5 series model, with a breakthrough 2 million token context window.

	Free Tier	Paid Tier, per 1M tokens in USD
Input price	Free of charge	$1.25, prompts <= 128k tokens $2.50, prompts > 128k tokens
Output price	Free of charge	$5.00, prompts <= 128k tokens $10.00, prompts > 128k tokens
Context caching price	Not available	$0.3125, prompts <= 128k tokens $0.625, prompts > 128k tokens
Context caching (storage)	Not available	$4.50 per hour
Tuning price	Not available	Not available
Grounding with Google Search	Not available	$35 / 1K grounding requests
Used to improve our products	Yes	No

Text Embedding 004

Our state-of-the-art text embedding model.

	Free Tier	Paid Tier, per 1M tokens in USD
Input price	Free of charge	Not available
Output price	Free of charge	Not available
Tuning price	Not available	Not available
Used to improve our products	Yes	No

[*] Google AI Studio usage is free of charge in all available regions. See Billing FAQs for details.

[**] Prices may differ from the prices listed here and the prices offered on Vertex AI. For Vertex prices, see the Vertex AI pricing page.

[***] If you are using dynamic retrieval to optimize costs, only requests that contain at least one grounding support URL from the web in their response are charged for Grounding with Google Search. Costs for Gemini always apply. Rate limits are subject to change.

Last updated 2025-06-06 UTC.