Introducing Batch Mode, with higher rate limits and a 50% token discount.
**Learn more** (https://ai.google.dev/gemini-api/docs/batch-mode)

# Gemini models

(#gemini-2.5-pro)

## 2.5 Pro

Our most powerful thinking model with maximum response accuracy and state-of-the-art performance

- Input audio, images, video, and text, get text responses
- Tackle difficult problems, analyze large databases, and more
- Best for complex coding, reasoning, and multimodal understanding

(#gemini-2.5-flash)

## 2.5 Flash

Our best model in terms of price-performance, offering well-rounded capabilities.

- Input audio, images, video, and text, and get text responses
- Model thinks as needed; or, you can configure a thinking budget

- Best for low latency, high volume tasks that require thinking

(#gemini-2.5-flash-lite)

### 2.5 Flash-Lite

A Gemini 2.5 Flash model optimized for cost efficiency and low latency.

- Input audio, images, video, and text, and get text responses
- Most cost-efficient model supporting high throughput
- Best for real time, low latency use cases

**Note:** Gemini 2.5 Pro and 2.5 Flash come with ***thinking on by default***. If you're migrating from a non-thinking model such as 2.0 Pro or Flash, we recommend you to review the Thinking guide (/gemini-api/docs/thinking) first.

## Model variants

The Gemini API offers different models that are optimized for specific use cases. Here's a brief overview of Gemini variants that are available:

| Model variant | Input(s) | Output | Optimized for |
|---|---|---|---|
| Gemini 2.5 Pro (#gemini-2.5-pro) `gemini-2.5-pro` | Audio, images, videos, text, and PDF | Text | Enhanced thinking and reasoning, multimodal understanding, advanced coding, and more |
| Gemini 2.5 Flash (#gemini-2.5-flash) `gemini-2.5-flash` | Audio, images, videos, and text | Text | Adaptive thinking, cost efficiency |

| | | | |
|---|---|---|---|
| Gemini 2.5 Flash-Lite Preview (#gemini-2.5-flash-lite) **gemini-2.5-flash-lite-preview-06-17** | Text, image, video, audio | Text | Most cost-efficient model supporting high throughput |
| Gemini 2.5 Flash Native Audio (#gemini-2.5-flash-native-audio) **gemini-2.5-flash-preview-native-audio-dialog** & **gemini-2.5-flash-exp-native-audio-thinking-dialog** | Audio, videos, and text | Text and audio, interleaved | High quality, natural conversational audio outputs, with or without thinking |
| Gemini 2.5 Flash Preview TTS (#gemini-2.5-flash-preview-tts) **gemini-2.5-flash-preview-tts** | Text | Audio | Low latency, controllable, single- and multi-speaker text-to-speech audio generation |
| Gemini 2.5 Pro Preview TTS (#gemini-2.5-pro-preview-tts) **gemini-2.5-pro-preview-tts** | Text | Audio | Low latency, controllable, single- and multi-speaker text-to-speech audio generation |
| Gemini 2.0 Flash (#gemini-2.0-flash) **gemini-2.0-flash** | Audio, images, videos, and text | Text | Next generation features, speed, and realtime streaming. |
| Gemini 2.0 Flash Preview Image Generation (#gemini-2.0-flash-preview-image-generation) **gemini-2.0-flash-preview-image-generation** | Audio, images, videos, and text | Text, images | Conversational image generation and editing |
| Gemini 2.0 Flash-Lite (#gemini-2.0-flash-lite) **gemini-2.0-flash-lite** | Audio, images, videos, and text | Text | Cost efficiency and low latency |
| Gemini 1.5 Flash (#gemini-1.5-flash) **gemini-1.5-flash** | Audio, images, videos, and text | Text | Fast and versatile performance across a diverse variety of tasks Deprecated |

| | | | |
|---|---|---|---|
| Gemini 1.5 Flash-8B (#gemini-1.5-flash-8b) **gemini-1.5-flash-8b** | Audio, images, videos, and text | Text | High volume and lower intelligence tasks **Deprecated** |
| Gemini 1.5 Pro (#gemini-1.5-pro) **gemini-1.5-pro** | Audio, images, videos, and text | Text | Complex reasoning tasks requiring more intelligence **Deprecated** |
| Gemini Embedding (#gemini-embedding) **gemini-embedding-001** | Text | Text embeddings | Measuring the relatedness of text strings |
| Imagen 4 (#imagen-4) **imagen-4.0-generate-preview-06-06** **imagen-4.0-ultra-generate-preview-06-06** | Text | Images | Our most up-to-date image generation model |
| Imagen 3 (#imagen-3) **imagen-3.0-generate-002** | Text | Images | High quality image generation model |
| Veo 2 (#veo-2) **veo-2.0-generate-001** | Text, images | Video | High quality video generation |
| Gemini 2.5 Flash Live (#live-api) **gemini-live-2.5-flash-preview** | Audio, video, and text | Text, audio | Low-latency bidirectional voice and video interactions |
| Gemini 2.0 Flash Live (#live-api-2.0) **gemini-2.0-flash-live-001** | Audio, video, and text | Text, audio | Low-latency bidirectional voice and video interactions |

You can view the rate limits for each model on the rate limits page (/gemini-api/docs/rate-limits).

## Gemini 2.5 Pro

Gemini 2.5 Pro is our state-of-the-art thinking model, capable of reasoning over complex problems in code, math, and STEM, as well as analyzing large datasets, codebases, and

documents using long context.

✦ **Try in Google AI Studio** (https://aistudio.google.com?model=gemini-2.5-pro)

### Model details

| Property | Description | |
|---|---|---|
| Model code | `gemini-2.5-pro` | |
| Supported data types | **Inputs**<br>Audio, images, video, text, and PDF | **Output**<br>Text |
| Token limits[*] (#token-size) | **Input token limit**<br>1,048,576 | **Output token limit**<br>65,536 |
| Capabilities | **Structured outputs**<br>Supported | **Caching**<br>Supported |
| | **Tuning**<br>Not supported | **Function calling**<br>Supported |
| | **Code execution**<br>Supported | **Search grounding**<br>Supported |
| | **Image generation**<br>Not supported | **Audio generation**<br>Not supported |
| | **Live API**<br>Not supported | **Thinking**<br>Supported |

| | |
|---|---|
| | **Batch API**<br>Supported |
| Versions | Read the model version patterns (/gemini-api/docs/models/gemini#model-versions) for more details.<br>• `Stable: gemini-2.5-pro`<br>• `Preview: gemini-2.5-pro-preview-06-05`<br>• `Preview: gemini-2.5-pro-preview-05-06`<br>• `Preview: gemini-2.5-pro-preview-03-25` |
| Latest update | June 2025 |
| Knowledge cutoff | January 2025 |

## Gemini 2.5 Flash

Our best model in terms of price-performance, offering well-rounded capabilities. 2.5 Flash is best for large scale processing, low-latency, high volume tasks that require thinking, and agentic use cases.

✦ **Try in Google AI Studio** (https://aistudio.google.com?model=gemini-2.5-flash)

### Model details

| Property | Description | |
|---|---|---|
| Model code | `models/gemini-2.5-flash` | |
| Supported | **Inputs**<br>Text, images, video, audio | **Output**<br>Text |

| | | |
|---|---|---|
| data types | | |
| Token limits[*] (#token-size) | **Input token limit**<br>1,048,576 | **Output token limit**<br>65,536 |
| Capabilit | **Audio generation**<br>Not supported | **Caching**<br>Supported |
| | **Code execution**<br>Supported | **Function calling**<br>Supported |
| | **Image generation**<br>Not supported | **Search grounding**<br>Supported |
| | **Structured outputs**<br>Supported | **Thinking**<br>Supported |
| | **Tuning**<br>Not supported | **Batch API**<br>Supported |
| Versions | Read the model version patterns (/gemini-api/docs/models/gemini#model-versions) for more details.<br>• Stable: `gemini-2.5-flash`<br>• Preview: `gemini-2.5-flash-preview-05-20` | |
| Latest update | June 2025 | |
| Knowledge cutoff | January 2025 | |

## Gemini 2.5 Flash-Lite Preview

A Gemini 2.5 Flash model optimized for cost efficiency and low latency.

✦ **Try in Google AI Studio** (https://aistudio.google.com?model=gemini-2.5-flash-lite-preview-06-17)

### Model details

| Property | Description |
|---|---|
| Model code | `models/gemini-2.5-flash-lite-preview-06-17` |
| Supported data types | **Inputs**<br>Text, images, video, and audio<br>**Output**<br>Text |
| Token limits[*] (#token-size) | **Input token limit**<br>1,000,000<br>**Output token limit**<br>64,000 |
| Capabilities | **Structured outputs**<br>Supported<br>**Caching**<br>Supported<br>**Tuning**<br>Not supported<br>**Function calling**<br>Supported<br>**Code execution**<br>Supported<br>**URL Context**<br>Supported<br>**Search grounding**<br>Supported<br>**Image generation**<br>Not supported<br>**Audio generation**<br>**Live API** |

| | |
|---|---|
| | Not supported Not supported<br><br>**Thinking**<br>Supported |
| Versions | Read the model version patterns (/gemini-api/docs/models/gemini#model-versions) for more details.<br>• Preview: `gemini-2.5-flash-lite-preview-06-17` |
| Latest update | June 2025 |
| Knowledge cutoff | January 2025 |

## Gemini 2.5 Flash Native Audio

Our native audio dialog models, with and without thinking, available through the Live API (/gemini-api/docs/live). These models provide interactive and unstructured conversational experiences, with style and control prompting.

✦ **Try native audio in Google AI Studio** (https://aistudio.google.com/app/live)

### Model details

| Property | Description |
|---|---|
| Model code | `models/gemini-2.5-flash-preview-native-audio-dialog` & `models/gemini-2.5-flash-exp-native-audio-thinking-dialog` |

| | | |
|---|---|---|
| Supported data types | **Inputs**<br>Audio, video, text | **Output**<br>Audio and text |
| Token limits[*] (#token-size) | **Input token limit**<br>128,000 | **Output token limit**<br>8,000 |
| Capabilities | **Audio generation**<br>Supported<br>**Code execution**<br>Not supported<br>**Image generation**<br>Not supported<br>**Structured outputs**<br>Not supported<br>**Tuning**<br>Not supported | **Caching**<br>Not supported<br>**Function calling**<br>Supported<br>**Search grounding**<br>Supported<br>**Thinking**<br>Supported |
| Versions | Read the model version patterns (/gemini-api/docs/models/gemini#model-versions) for more details.<br>• Preview: `gemini-2.5-flash-preview-05-20`<br>• Experimental: `gemini-2.5-flash-exp-native-audio-thinking-dialog` | |
| Latest update | May 2025 | |
| Knowledge | January 2025 | |

cutoff

## Gemini 2.5 Flash Preview Text-to-Speech

Gemini 2.5 Flash Preview TTS is our price-performant text-to-speech model, delivering high control and transparency for structured workflows like podcast generation, audiobooks, customer support, and more. Gemini 2.5 Flash rate limits are more restricted since it is an experimental / preview model.

✦ **Try in Google AI Studio** (https://aistudio.google.com/generate-speech)

### Model details

| Property | Description | |
|---|---|---|
| Model code | `models/gemini-2.5-flash-preview-tts` | |
| Supported data types | **Inputs**<br>Text | **Output**<br>Audio |
| Token limits[*] (#token-size) | **Input token limit**<br>8,000 | **Output token limit**<br>16,000 |
| Capabilities | **Structured outputs**<br>Not supported | **Caching**<br>Not supported |
| | **Tuning**<br>Not supported | **Function calling**<br>Not supported |
| | **Code execution** | **Search** |

| | |
|---|---|
| Not supported | Not supported |
| **Audio generation** | **Live API** |
| Supported | Not supported |
| **Thinking** | |
| Not supported | |

| | |
|---|---|
| Versions | Read the model version patterns (/gemini-api/docs/models/gemini#model-versions) for more details.<br>• `gemini-2.5-flash-preview-tts` |
| Latest update | May 2025 |

## Gemini 2.5 Pro Preview Text-to-Speech

Gemini 2.5 Pro Preview TTS is our most powerful text-to-speech model, delivering high control and transparency for structured workflows like podcast generation, audiobooks, customer support, and more. Gemini 2.5 Pro rate limits are more restricted since it is an experimental / preview model.

✦ **Try in Google AI Studio** (https://aistudio.google.com/generate-speech)

### Model details

| Property | Description |
|---|---|
| Model code | `models/gemini-2.5-pro-preview-tts` |
| Supported | **Inputs**<br>Text<br>**Output**<br>Audio |

| | |
|---|---|
| data types | |
| Token limits[*] (#token-size) | **Input token limit**<br>8,000<br>**Output token limit**<br>16,000 |
| Capabilit | **Structured outputs**<br>Not supported<br>**Caching**<br>Not supported<br>**Tuning**<br>Not supported<br>**Function calling**<br>Not supported<br>**Code execution**<br>Not supported<br>**Search**<br>Not supported<br>**Audio generation**<br>Supported<br>**Live API**<br>Not supported<br>**Thinking**<br>Not supported |
| Versions | Read the model version patterns (/gemini-api/docs/models/gemini#model-versions) for more details.<br>• `gemini-2.5-pro-preview-tts` |
| Latest update | May 2025 |

## Gemini 2.0 Flash

Gemini 2.0 Flash delivers next-gen features and improved capabilities, including superior speed, native tool use, and a 1M token context window.

✦ **Try in Google AI Studio** (https://aistudio.google.com?model=gemini-2.0-flash-001)

## Model details

| Property | Description |
|---|---|
| Model code | `models/gemini-2.0-flash` |
| Supported data types | **Inputs**<br>Audio, images, video, and text<br>**Output**<br>Text |
| Token limits[*] (#token-size) | **Input token limit**<br>1,048,576<br>**Output token limit**<br>8,192 |
| Capabilities | **Structured outputs**<br>Supported<br>**Caching**<br>Supported<br>**Tuning**<br>Not supported<br>**Function calling**<br>Supported<br>**Code execution**<br>Supported<br>**Search**<br>Supported<br>**Image generation**<br>Not supported<br>**Audio generation**<br>Not supported<br>**Live API**<br>Supported<br>**Thinking**<br>Experimental<br>**Batch API** |

| | |
|---|---|
| | Supported |
| Versions | Read the model version patterns (/gemini-api/docs/models/gemini#model-versions) for more details.<br>• Latest: `gemini-2.0-flash`<br>• Stable: `gemini-2.0-flash-001`<br>• Experimental: `gemini-2.0-flash-exp` |
| Latest update | February 2025 |
| Knowledge cutoff | August 2024 |

## Gemini 2.0 Flash Preview Image Generation

Gemini 2.0 Flash Preview Image Generation delivers improved image generation features, including generating and editing images conversationally.

✦ **Try in Google AI Studio** (https://aistudio.google.com?model=gemini-2.0-flash-preview-image-generation)

### Model details

| Property | Description | |
|---|---|---|
| Model code | `models/gemini-2.0-flash-preview-image-generation` | |
| Supported data types | **Inputs**<br>Audio, images, video, and text | **Output**<br>Text and images |

| | | |
|---|---|---|
| Token limits[*] (#token-size) | **Input token limit**<br>32,000 | **Output token limit**<br>8,192 |
| Capabilities | **Structured outputs**<br>Supported | **Caching**<br>Supported |
| | **Tuning**<br>Not supported | **Function calling**<br>Not supported |
| | **Code execution**<br>Not Supported | **Search**<br>Not Supported |
| | **Image generation**<br>Supported | **Audio generation**<br>Not supported |
| | **Live API**<br>Not Supported | **Thinking**<br>Not Supported |
| Versions | Read the model version patterns (/gemini-api/docs/models/gemini#model-versions) for more details.<br>• Preview: `gemini-2.0-flash-preview-image-generation`<br>gemini-2.0-flash-preview-image-generation is not currently supported in a number of countries in Europe, Middle East & Africa | |
| Latest update | May 2025 | |
| Knowledge cutoff | August 2024 | |

## Gemini 2.0 Flash-Lite

A Gemini 2.0 Flash model optimized for cost efficiency and low latency.

✦ **Try in Google AI Studio** (https://aistudio.google.com?model=gemini-2.0-flash-lite)

### Model details

| Property | Description |
|---|---|
| Model code | `models/gemini-2.0-flash-lite` |
| Supported data types | **Inputs**<br>Audio, images, video, and text<br>**Output**<br>Text |
| Token limits[*] (#token-size) | **Input token limit**<br>1,048,576<br>**Output token limit**<br>8,192 |
| Capabilities | **Structured outputs**<br>Supported<br>**Caching**<br>Supported<br>**Tuning**<br>Not supported<br>**Function calling**<br>Supported<br>**Code execution**<br>Not supported<br>**Search**<br>Not supported<br>**Image generation**<br>Not supported<br>**Audio generation**<br>Not supported<br>**Live API**<br>**Batch API** |

| | | |
|---|---|---|
| | Not supported | Supported |
| Versions | Read the model version patterns (/gemini-api/docs/models/gemini#model-versions) for more details.<br>• Latest: `gemini-2.0-flash-lite`<br>• Stable: `gemini-2.0-flash-lite-001` | |
| Latest update | February 2025 | |
| Knowledge cutoff | August 2024 | |

## Gemini 1.5 Flash

Gemini 1.5 Flash is a fast and versatile multimodal model for scaling across diverse tasks.

✦ **Try in Google AI Studio** (https://aistudio.google.com?model=gemini-1.5-flash)

### Model details

| Property | Description | |
|---|---|---|
| Model code | `models/gemini-1.5-flash` | |
| Supported data types | **Inputs**<br>Audio, images, video, and text | **Output**<br>Text |
| Token | **Input token limit**<br>1,048,576 | **Output token limit**<br>8,192 |

| | |
|---|---|
| limits[*] (#token-size) | |
| Audio/visual specs | **Maximum number of images per prompt**<br>3,600<br>**Maximum video length**<br>1 hour<br>**Maximum audio length**<br>Approximately 9.5 hours |
| Capabilities | **System instructions**<br>Supported<br>**JSON mode**<br>Supported<br>**JSON schema**<br>Supported<br>**Adjustable safety settings**<br>Supported<br>**Caching**<br>Supported<br>**Tuning**<br>Supported<br>**Function calling**<br>Supported<br>**Code execution**<br>Supported<br>**Live API**<br>Not supported |
| Versions | Read the model version patterns (/gemini-api/docs/models/gemini#model-versions) for more details.<br>• Latest: `gemini-1.5-flash-latest`<br>• Latest stable: `gemini-1.5-flash`<br>• Stable:<br>  • `gemini-1.5-flash-001`<br>  • `gemini-1.5-flash-002` |
| Deprecation date | September 2025 |

| | |
|---|---|
| Latest update | September 2024 |

## Gemini 1.5 Flash-8B

Gemini 1.5 Flash-8B is a small model designed for lower intelligence tasks.

✦ **Try in Google AI Studio** (https://aistudio.google.com?model=gemini-1.5-flash)

### Model details

| Property | Description |
|---|---|
| Model code | `models/gemini-1.5-flash-8b` |
| Supported data types | **Inputs**<br>Audio, images, video, and text<br>**Output**<br>Text |
| Token limits[*] (#token-size) | **Input token limit**<br>1,048,576<br>**Output token limit**<br>8,192 |
| Audio/visual specs | **Maximum number of images per prompt**<br>3,600<br>**Maximum video length**<br>1 hour<br>**Maximum audio length**<br>Approximately 9.5 hours |

| | |
|---|---|
| Capabilities | **System instructions**<br>Supported<br>**JSON mode**<br>Supported<br>**JSON schema**<br>Supported<br>**Adjustable safety settings**<br>Supported<br>**Caching**<br>Supported<br>**Tuning**<br>Supported<br>**Function calling**<br>Supported<br>**Code execution**<br>Supported<br>**Live API**<br>Not supported |
| Versions | Read the model version patterns (/gemini-api/docs/models/gemini#model-versions) for more details.<br>• Latest: `gemini-1.5-flash-8b-latest`<br>• Latest stable: `gemini-1.5-flash-8b`<br>• Stable:<br>  • `gemini-1.5-flash-8b-001` |
| Deprecation date | September 2025 |
| Latest update | October 2024 |

## Gemini 1.5 Pro

Try Gemini 2.5 Pro Preview (/gemini-api/docs/models/experimental-models#available-models), our most advanced Gemini model to date.

Gemini 1.5 Pro is a mid-size multimodal model that is optimized for a wide-range of reasoning tasks. 1.5 Pro can process large amounts of data at once, including 2 hours of video, 19 hours of audio, codebases with 60,000 lines of code, or 2,000 pages of text.

✦ **Try in Google AI Studio** (https://aistudio.google.com?model=gemini-1.5-pro)

### Model details

| Property | Description | |
|---|---|---|
| Model code | `models/gemini-1.5-pro` | |
| Supported data types | **Inputs**<br>Audio, images, video, and text | **Output**<br>Text |
| Token limits[*] (#token-size) | **Input token limit**<br>2,097,152 | **Output token limit**<br>8,192 |
| Audio/visual specs | **Maximum number of images per prompt**<br>7,200<br>**Maximum audio length**<br>Approximately 19 hours | **Maximum video length**<br>2 hours |
| Capabilities | **System instructions**<br>Supported<br>**JSON schema**<br>Supported<br>**Caching** | **JSON mode**<br>Supported<br>**Adjustable safety settings**<br>Supported<br>**Tuning** |

| | |
|---|---|
| Supported | Not supported |
| **Function calling** | **Code execution** |
| Supported | Supported |
| **Live API** | |
| Not supported | |

| | |
|---|---|
| Versions | Read the model version patterns (/gemini-api/docs/models/gemini#model-versions) for more details.<br>• Latest: `gemini-1.5-pro-latest`<br>• Latest stable: `gemini-1.5-pro`<br>• Stable:<br>  • `gemini-1.5-pro-001`<br>  • `gemini-1.5-pro-002` |
| Deprecation date | September 2025 |
| Latest update | September 2024 |

## Imagen 4

Imagen 4 is our latest image model, capable of generating highly detailed images with rich lighting, significantly better text rendering, and higher resolution output than previous models.

**Model details**

| Property | Description |
|---|---|
| Model code | **Gemini API**<br>`imagen-4.0-generate-preview-06-06` |

| | | |
|---|---|---|
| | `imagen-4.0-ultra-generate-preview-06-06` | |
| Supported data types | **Input**<br>Text | **Output**<br>Images |
| Token limits[*] (#token-size) | **Input token limit**<br>480 tokens (text) | **Output images**<br>1 (Ultra)<br>1 to 4 (Standard) |
| Latest update | June 2025 | |

## Imagen 3

Imagen 3 is our highest quality text-to-image model, capable of generating images with even better detail, richer lighting and fewer distracting artifacts than our previous models.

**Model details**

| Property | Description | |
|---|---|---|
| Model code | **Gemini API**<br>`imagen-3.0-generate-002` | |
| Supported data types | **Input**<br>Text | **Output**<br>Images |
| Token limits[*] (#token-size) | **Input token limit**<br>N/A | **Output images**<br>Up to 4 |
| Latest update | February 2025 | |

## Veo 2

Veo 2 is our high quality text- and image-to-video model, capable of generating detailed videos, capturing the artistic nuance in your prompts.

**Model details**

| Property | Description | |
|---|---|---|
| Model code | **Gemini API**<br>`veo-2.0-generate-001` | |
| Supported data types | **Input**<br>Text, image | **Output**<br>Video |
| Limits | **Text input**<br>N/A<br>**Output video**<br>Up to 2 | **Image input**<br>Any image resolution and aspect ratio up to 20MB file size |
| Latest update | April 2025 | |

## Gemini 2.5 Flash Live

The Gemini 2.5 Flash Live model works with the Live API to enable low-latency bidirectional voice and video interactions with Gemini. The model can process text, audio, and video input, and it can provide text and audio output.

✦ **Try in Google AI Studio** (https://aistudio.google.com?model=gemini-live-2.5-flash-preview)

### Model details

| Property | Description |
| --- | --- |
| Model code | `models/gemini-live-2.5-flash-preview` |
| Supported data types | **Inputs**<br>Audio, video, and text<br>**Output**<br>Text, and audio |
| Token limits[*] (#token-size) | **Input token limit**<br>1,048,576<br>**Output token limit**<br>8,192 |
| Capabilities | **Structured outputs**<br>Supported<br>**Tuning**<br>Not supported<br>**Function calling**<br>Supported<br>**Code execution**<br>Supported<br>**Search**<br>Supported<br>**Image generation**<br>Not supported<br>**Audio generation**<br>Supported<br>**Thinking**<br>Not supported |
| Versions | Read the model version patterns (/gemini-api/docs/models/gemini#model-versions) for more details.<br>• Preview: `gemini-live-2.5-flash-preview` |
| Latest update | June 2025 |

| | |
|---|---|
| | January 2025 |
| Knowledge cutoff | |

## Gemini 2.0 Flash Live

The Gemini 2.0 Flash Live model works with the Live API to enable low-latency bidirectional voice and video interactions with Gemini. The model can process text, audio, and video input, and it can provide text and audio output.

✦ **Try in Google AI Studio** (https://aistudio.google.com?model=gemini-2.0-flash-live-001)

### Model details

| Property | Description | |
|---|---|---|
| Model code | `models/gemini-2.0-flash-live-001` | |
| Supported data types | **Inputs**<br>Audio, video, and text | **Output**<br>Text, and audio |
| Token limits[*] (#token-size) | **Input token limit**<br>1,048,576 | **Output token limit**<br>8,192 |
| Capabilit | **Structured outputs**<br>Supported | **Tuning**<br>Not supported |
| | **Function calling**<br>Supported | **Code execution**<br>Supported |

| | | |
|---|---|---|
| | **Search**<br>Supported | **Image generation**<br>Not supported |
| | **Audio generation**<br>Supported | **Thinking**<br>Not supported |
| Versions | Read the model version patterns (/gemini-api/docs/models/gemini#model-versions) for more details.<br>• Preview: `gemini-2.0-flash-live-001` | |
| Latest update | April 2025 | |
| Knowledge cutoff | August 2024 | |

## Gemini Embedding

The Gemini Embedding model achieves a SOTA performance (https://deepmind.google/research/publications/157741/) across many key dimensions including code, multi-lingual, and retrieval.

**Model details**

| Property | Description | |
|---|---|---|
| Model code | **Gemini API**<br>`gemini-embedding-001` | |
| Supported data types | **Input**<br>Text | **Output**<br>Text embeddings |

| | Input token limit | Output dimension size |
|---|---|---|
| Token limits[*] (#token-size) | 2,048 | Flexible, supports: 128 - 3072, Recommended: 768, 1536, 3072 |
| Versions | Read the model version patterns (/gemini-api/docs/models/gemini#model-versions) for more details.<br>• Stable: `gemini-embedding-001`<br>• Preview: `gemini-embedding-exp-03-07` | |
| Latest update | June 2025 | |

## Legacy Embedding Models

### Text Embedding (Legacy)

**Note:** Gemini Embedding is the newest version of the Embedding model. If you're creating a new project, use Gemini Embedding (/gemini-api/docs/models#gemini-embedding).

Text embeddings (/gemini-api/docs/embeddings) are used to measure the relatedness of strings and are widely used in many AI applications.

#### Model details

| Property | Description |
|---|---|
| Model code | **Gemini API**<br>`models/text-embedding-004` |

| | | |
|---|---|---|
| Supported data types | **Input**<br>Text | **Output**<br>Text embeddings |
| Token limits[*] (#token-size) | **Input token limit**<br>2,048 | **Output dimension size**<br>768 |
| Rate limits[**] (#rate-limits) | 1,500 requests per minute | |
| Adjustable safety settings | Not supported | |
| Deprecation date | January 2026 | |
| Latest update | April 2024 | |

## AQA

You can use the AQA model to perform Attributed Question-Answering (/gemini-api/docs/semantic_retrieval) (AQA)–related tasks over a document, corpus, or a set of passages. The AQA model returns answers to questions that are grounded in provided sources, along with estimating answerable probability.

### Model details

| Property | Description | |
|---|---|---|
| Model code | `models/aqa` | |
| Supported data types | **Input**<br>Text | **Output**<br>Text |

| | | |
|---|---|---|
| Supported language | English | |
| Token limits[*] (#token-size) | **Input token limit** | **Output token limit** |
| | 7,168 | 1,024 |
| Rate limits[**] (#rate-limits) | 1,500 requests per minute | |
| Adjustable safety settings | Supported | |
| Latest update | December 2023 | |

See the examples (/examples) to explore the capabilities of these model variations.

[*] A token is equivalent to about 4 characters for Gemini models. 100 tokens are about 60-80 English words.

## Model version name patterns

Gemini models are available in either *stable*, *preview*, or *experimental* versions. In your code, you can use one of the following model name formats to specify which model and version you want to use.

### Latest stable

Points to the most recent stable version released for the specified model generation and variation.

To specify the latest stable version, use the following pattern: `<model>-<generation>-<variation>`. For example, `gemini-2.0-flash`.

### Stable

Points to a specific stable model. Stable models usually don't change. Most production apps should use a specific stable model.

To specify a stable version, use the following pattern: `<model>-<generation>-<variation>-<version>`. For example, `gemini-2.0-flash-001`.

### Preview

Points to a preview model which may not be suitable for production use, come with more restrictive rate limits, but may have billing enabled.

To specify a preview version, use the following pattern: `<model>-<generation>-<variation>-<version>`. For example, `gemini-2.5-pro-preview-06-05`.

Preview models are not stable and availability of model endpoints is subject to change.

### Experimental

Points to an experimental model which may not be suitable for production use and come with more restrictive rate limits. We release experimental models to gather feedback and get our latest updates into the hands of developers quickly.

To specify an experimental version, use the following pattern: `<model>-<generation>-<variation>-<version>`. For example, `gemini-2.0-pro-exp-02-05`.

Experimental models are not stable and availability of model endpoints is subject to change.

## Experimental models

In addition to stable models, the Gemini API offers experimental models which may not be suitable for production use and come with more restrictive rate limits.

We release experimental models to gather feedback, get our latest updates into the hands of developers quickly, and highlight the pace of innovation happening at Google. What we learn from experimental launches informs how we release models more widely. An experimental model can be swapped for another without prior notice. We don't guarantee that an experimental model will become a stable model in the future.

## Previous experimental models

As new versions or stable releases become available, we remove and replace experimental models. You can find the previous experimental models we released in the following section along with the replacement version:

| Model code | Base model | Replacement version |
|---|---|---|
| `gemini-embedding-exp-03-07` | Gemini Embedding | `gemini-embedding-001` |
| `gemini-2.5-flash-preview-04-17` | Gemini 2.5 Flash | `gemini-2.5-flash-preview-05-20` |
| `gemini-2.0-flash-exp-image-generation` | Gemini 2.0 Flash | `gemini-2.0-flash-preview-image-generation` |
| `gemini-2.5-pro-preview-06-05` | Gemini 2.5 Pro | `gemini-2.5-pro` |
| `gemini-2.5-pro-preview-05-06` | Gemini 2.5 Pro | `gemini-2.5-pro` |
| `gemini-2.5-pro-preview-03-25` | Gemini 2.5 Pro | `gemini-2.5-pro` |
| `gemini-2.0-flash-thinking-exp-01-21` | Gemini 2.5 Flash | `gemini-2.5-flash-preview-04-17` |
| `gemini-2.0-pro-exp-02-05` | Gemini 2.0 Pro Experimental | `gemini-2.5-pro-preview-03-25` |
| `gemini-2.0-flash-exp` | Gemini 2.0 Flash | `gemini-2.0-flash` |
| `gemini-exp-1206` | Gemini 2.0 Pro | `gemini-2.0-pro-exp-02-05` |
| `gemini-2.0-flash-thinking-exp-1219` | Gemini 2.0 Flash Thinking | `gemini-2.0-flash-thinking-exp-01-21` |

| | | |
|---|---|---|
| `gemini-exp-1121` | Gemini | `gemini-exp-1206` |
| `gemini-exp-1114` | Gemini | `gemini-exp-1206` |
| `gemini-1.5-pro-exp-0827` | Gemini 1.5 Pro | `gemini-exp-1206` |
| `gemini-1.5-pro-exp-0801` | Gemini 1.5 Pro | `gemini-exp-1206` |
| `gemini-1.5-flash-8b-exp-0924` | Gemini 1.5 Flash-8B | `gemini-1.5-flash-8b` |
| `gemini-1.5-flash-8b-exp-0827` | Gemini 1.5 Flash-8B | `gemini-1.5-flash-8b` |

## Supported languages

Gemini models are trained to work with the following languages:

- Arabic (`ar`)
- Bengali (`bn`)
- Bulgarian (`bg`)
- Chinese simplified and traditional (`zh`)
- Croatian (`hr`)
- Czech (`cs`)
- Danish (`da`)
- Dutch (`nl`)
- English (`en`)
- Estonian (`et`)
- Finnish (`fi`)
- French (`fr`)
- German (`de`)

- Greek (`el`)
- Hebrew (`iw`)
- Hindi (`hi`)
- Hungarian (`hu`)
- Indonesian (`id`)
- Italian (`it`)
- Japanese (`ja`)
- Korean (`ko`)
- Latvian (`lv`)
- Lithuanian (`lt`)
- Norwegian (`no`)
- Polish (`pl`)
- Portuguese (`pt`)
- Romanian (`ro`)
- Russian (`ru`)
- Serbian (`sr`)
- Slovak (`sk`)
- Slovenian (`sl`)
- Spanish (`es`)
- Swahili (`sw`)
- Swedish (`sv`)
- Thai (`th`)
- Turkish (`tr`)
- Ukrainian (`uk`)
- Vietnamese (`vi`)

Last updated 2025-07-15 UTC.