

Starting April 29, 2025, Gemini 1.5 Pro and Gemini 1.5 Flash models are not available in projects that have no prior usage of these models, including new projects. For details, see [Model versions and lifecycle](#) (/vertex-ai/generative-ai/docs/learn/model-versions#legacy-stable).

Gemini 2.0 Flash

Release Notes

Gemini 2.0 Flash delivers next-generation features and improved capabilities designed for the agentic era, including superior speed, built-in tool use, multimodal generation, and a 1M token context window. Gemini 2.0 Flash improves upon our previous Flash model and offers enhanced quality at similar speeds.

For even more detailed technical information on Gemini 2.0 Flash (such as performance benchmarks, information on our training datasets, efforts on sustainability, intended usage and limitations, and our approach to ethics and safety), see the [model card for Gemini 2.0 Flash](#) (https://storage.googleapis.com/model-cards/documents/gemini-2-flash.pdf).

2.0 Flash

Image generation
Preview

Live API
Preview

◆ Try in Vertex AI

(https://console.cloud.google.com/vertex-ai/generative/multimodal/create/text?model=gemini-2.0-flash-001)

🔗 View in Model Garden

(https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemini-2.0-flash-001)

➤ (Preview) Deploy example app

(https://console.cloud.google.com/vertex-ai/studio/multimodal?suggestedPrompt=How%20does%20AI%20work&deploy=true&model=gemini-2.0-flash-001)



Note: To use the "Deploy example app" feature, you need a Google Cloud project with billing and Vertex AI API enabled.




Model ID


gemini-2.0-flash

Supported inputs &

Inputs: Text, code, images, audio, video
Outputs: Text

outputs			
Token limits		Maximum input tokens: 1,048,576 Maximum output tokens: 8,192 (default)	
Capabilities			
		Supported	Not supported
<u>Grounding with Google Search</u> (/vertex-ai/generative-ai/docs/grounding/grounding-with-google-search)		<u>Function calling</u> (/vertex-ai/generative-ai/docs/multimodal/function-calling)	<u>Live API</u> (/vertex-ai/generative-ai/docs/live-api)
<u>Code execution</u> (/vertex-ai/generative-ai/docs/multimodal/code-execution)		<u>Count Tokens</u> (/vertex-ai/generative-ai/docs/multimodal/get-token-count)	 <u>Thinking</u> (/vertex-ai/generative-ai/docs/thinking)
<u>Tuning</u> (/vertex-ai/generative-ai/docs/models/tune-models)		<u>Context caching</u> (/vertex-ai/generative-ai/docs/context-cache/context-cache-overview)	
<u>System instructions</u> (/vertex-ai/generative-ai/docs/learn/prompts/system-instruction-introduction)		<u>Vertex AI RAG Engine</u> (/vertex-ai/generative-ai/docs/rag-engine/rag-overview)	
<u>Structured output</u> (/vertex-ai/generative-ai/docs/multimodal/control-generated-output)		<u>Chat completions</u> (/vertex-ai/generative-ai/docs/migrate/openai/overview)	
<u>Batch prediction</u> (/vertex-ai/generative-ai/docs/multimodal/batch-prediction-gemini)			
Usage types			
		Supported	Not supported
<u>Provisioned Throughput</u> (/vertex-ai/generative-ai/docs/provisioned-throughput)		<u>Dynamic shared quota</u> (/vertex-ai/generative-ai/docs/dsq)	<u>Fixed quota</u> (/vertex-ai/generative-ai/docs/quotas)
Input size limit500 MB			
Technical specifications	Images 	<ul style="list-style-type: none">Maximum images per prompt: 3,000Maximum image size: 7 MBMaximum tokens per minute (TPM) per project:<ul style="list-style-type: none">High/Medium/Default media resolution:<ul style="list-style-type: none">US/Asia: 40 M	

	<ul style="list-style-type: none">• EU: 10 M• Low media resolution:<ul style="list-style-type: none">• US/Asia: 10 M• EU: 2.6 M• Supported MIME types: image/png, image/jpeg, image/webp
<div>Documents</div> <div></div>	<ul style="list-style-type: none">• Maximum number of files per prompt: 3,000• Maximum number of pages per file: 1,000• Maximum file size per file: 50 MB• Maximum tokens per minute (TPM) per project¹:<ul style="list-style-type: none">• US/Asia: 3.4 M• EU: 3.4 M• Supported MIME types: application/pdf, text/plain
<div>Video</div> <div></div>	<ul style="list-style-type: none">• Maximum video length (with audio): Approximately 45 minutes• Maximum video length (without audio): Approximately 1 hour• Maximum number of videos per prompt: 10• Maximum tokens per minute (TPM):<ul style="list-style-type: none">• High/Medium/Default media resolution:<ul style="list-style-type: none">• US/Asia: 38 M• EU: 10 M• Low media resolution:<ul style="list-style-type: none">• US/Asia: 10 M• EU: 2.5 M• Supported MIME types: video/x-flv, video/quicktime, video/mpeg, video/mpegs, video/mpg, video/mp4, video/webm, video/wmv, video/3gpp
<div>Audio</div> <div></div>	<ul style="list-style-type: none">• Maximum audio length per prompt: Approximately 8.4 hours, or up to 1 million tokens• Maximum number of audio files per prompt: 1• Speech understanding for: Audio summarization, transcription, and translation• Maximum tokens per minute (TPM):<ul style="list-style-type: none">• US/Asia: 3.5 M• EU: 3.5 M• Supported MIME types: audio/x-aac, audio/flac, audio/mp3, audio/m4a, audio/mpeg, audio/mpga, audio/mp4, audio/opus, audio/pcm, audio/wav, audio/webm
Parameter defaults	<ul style="list-style-type: none">• Temperature: 0.0-2.0 (default 1.0)

		<ul style="list-style-type: none">• topP: 0.0-1.0 (default 0.95)• topK: 64 (fixed)• candidateCount: 1-8 (default 1)		
Supported regions	Model availability	Global <ul style="list-style-type: none">• global	United States <ul style="list-style-type: none">• us-central1• us-east1• us-east4• us-east5• us-south1• us-west1• us-west4	Europe <ul style="list-style-type: none">• europe-central2• europe-north1• europe-southwest1• europe-west1• europe-west4• europe-west8• europe-west9
	(Includes dynamic shared quota & Provisioned Throughput)			
	ML processing	United States <ul style="list-style-type: none">• Multi-region	Europe <ul style="list-style-type: none">• Multi-region	
See Data residency (/vertex-ai/generative-ai/docs/learn/data-residency) for more information.				
Knowledge cutoff date	June 2024			
Versions	gemini-2.0-flash-001 <ul style="list-style-type: none">• Launch stage: Generally available• Release date: February 5, 2025• Discontinuation date: February 5, 2026			
Security controls	Online prediction	Data residency (at rest) Supported	VPC Service Controls Supported	
		Customer-managed encryption keys (CMEK) Supported	Access Transparency (AXT) Supported	
	Batch prediction	Data residency (at rest) Supported	VPC Service Controls Supported	
		Customer-managed encryption keys (CMEK) Not supported	Access Transparency (AXT) Not supported	
	Tuning	Data residency (at rest) Supported	Customer-managed encryption keys (CMEK) Supported	

	VPC Service Controls	Access Transparency (AXT)
	Supported	Not supported
	See Security controls (/vertex-ai/generative-ai/docs/security-controls) for more information.	
Pricing	See Pricing (/vertex-ai/generative-ai/pricing).	

Except as otherwise noted, the content of this page is licensed under the [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/) (https://creativecommons.org/licenses/by/4.0/), and code samples are licensed under the [Apache 2.0 License](https://www.apache.org/licenses/LICENSE-2.0) (https://www.apache.org/licenses/LICENSE-2.0). For details, see the [Google Developers Site Policies](https://developers.google.com/site-policies) (https://developers.google.com/site-policies). Java is a registered trademark of Oracle and/or its affiliates.

Last updated 2025-07-26 UTC.