AI

Our next-generation model: Gemini 1.5

The model delivers dramatically enhanced performance, with a breakthrough in long-context understanding across modalities.

Feb 15, 2024 · 8 min read

Share

Sundar Pichai
CEO of Google and Alphabet

Demis Hassabis
CEO of Google DeepMind

In this story

Listen to article 11 minutes

A note from Google and Alphabet CEO Sundar Pichai:

Last week, we rolled out our most capable model, Gemini 1.0 Ultra, and took a significant step forward in making Google products more helpful, starting with Gemini Advanced. Today, developers and Cloud customers can begin building with 1.0 Ultra too — with our Gemini API in AI Studio and in Vertex AI.

Our teams continue pushing the frontiers of our latest models with safety at the core. They are making rapid progress. In fact, we're ready to introduce the next generation: Gemini 1.5. It shows dramatic improvements across a number of dimensions and 1.5 Pro achieves comparable quality to 1.0 Ultra, while using less compute.

This new generation also delivers a breakthrough in long-context understanding. We've been able to significantly increase the amount of information our models can process — running up to 1 million tokens consistently, achieving the longest context window of any large-scale foundation model yet.

Longer context windows show us the promise of what is possible. They will enable entirely new capabilities and help developers build much more useful models and applications. We're excited to offer a limited preview of this experimental feature to developers and enterprise customers. Demis shares more on capabilities, safety and availability below.

— Sundar

Introducing Gemini 1.5

By Demis Hassabis, CEO of Google DeepMind, on behalf of the Gemini team

This is an exciting time for AI. New advances in the field have the potential to make AI more helpful for billions of people over the coming years. Since introducing Gemini 1.0, we've been testing, refining and enhancing its capabilities.

Today, we're announcing our next-generation model: Gemini 1.5.

Gemini 1.5 delivers dramatically enhanced performance. It represents a step change in our approach, building upon research and engineering innovations across nearly every part of our foundation model development and infrastructure. This includes making Gemini 1.5 more efficient to train and serve, with a new Mixture-of-Experts (MoE) architecture.

The first Gemini 1.5 model we're releasing for early testing is Gemini 1.5 Pro. It's a mid-size multimodal model, optimized for scaling across a wide-range of tasks, and performs at a similar level to 1.0 Ultra, our largest model to date. It also introduces a breakthrough experimental feature in long-context understanding.

Gemini 1.5 Pro comes with a standard 128,000 token context window. But starting today, a limited group of developers and enterprise customers can try it with a context window of up to 1 million tokens via AI Studio and Vertex AI in private preview.

As we roll out the full 1 million token context window, we're actively working on optimizations to improve latency, reduce computational requirements and enhance the user experience. We're excited for people to try this breakthrough capability, and we share more details on future availability below.

These continued advances in our next-generation models will open up new possibilities for people, developers and enterprises to create, discover and build using AI.

Context lengths of leading foundation models

Highly efficient architecture

Gemini 1.5 is built upon our leading research on Transformer and MoE architecture. While a traditional Transformer functions as one large neural network, MoE models are divided into smaller "expert" neural networks.

Depending on the type of input given, MoE models learn to selectively activate only the most relevant expert pathways in its neural network. This specialization massively enhances the model's efficiency.

Google has been an early adopter and pioneer of the MoE technique for deep learning through research such as Sparsely-Gated MoE, GShard-Transformer, Switch-Transformer, M4 and more.

Our latest innovations in model architecture allow Gemini 1.5 to learn complex tasks more quickly and maintain quality, while being more efficient to train and serve. These efficiencies are helping our teams iterate, train and deliver more advanced versions of Gemini faster than ever before, and we're working on further optimizations.

Greater context, more helpful capabilities

An AI model's "context window" is made up of tokens, which are the building blocks used for processing information. Tokens can be entire parts or subsections of words, images, videos, audio or code. The bigger a model's context window, the more information it can take in and process in a given prompt — making its output more consistent, relevant and useful.

Through a series of machine learning innovations, we've increased 1.5 Pro's context window capacity far beyond the original 32,000 tokens for Gemini 1.0. We can now run up to 1 million tokens in production.

This means 1.5 Pro can process vast amounts of information in one go — including 1 hour of video, 11 hours of audio, codebases with over 30,000 lines of code or over 700,000 words. In our research, we've also successfully tested up to 10 million tokens.

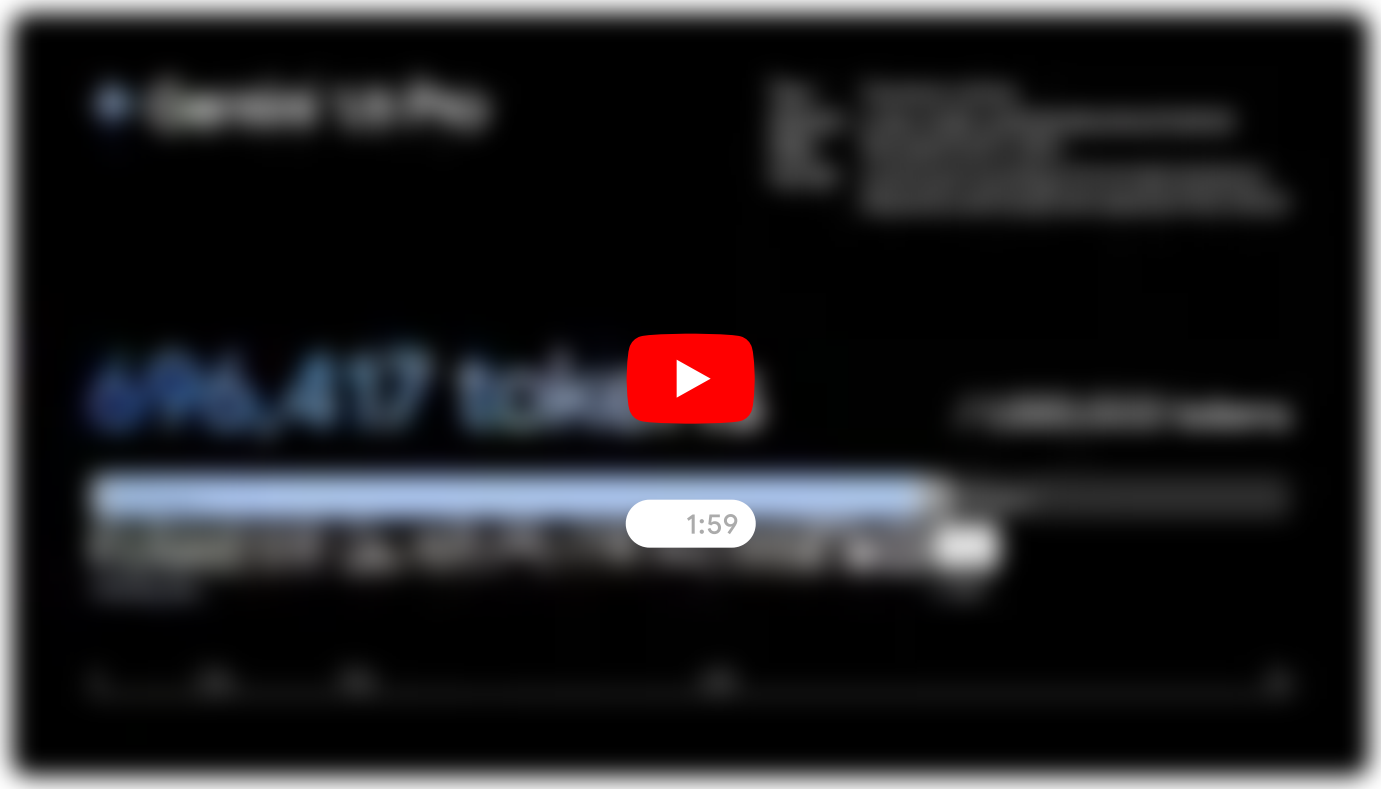
Complex reasoning about vast amounts of information

1.5 Pro can seamlessly analyze, classify and summarize large amounts of content within a given prompt. For example, when given the 402-page transcripts from Apollo 11's mission to the moon, it can reason about conversations, events and details found across the document.

Gemini 1.5 Pro can understand, reason about and identify curious details in the 402-page transcripts from Apollo 11's mission to the moon.

Better understanding and reasoning across modalities

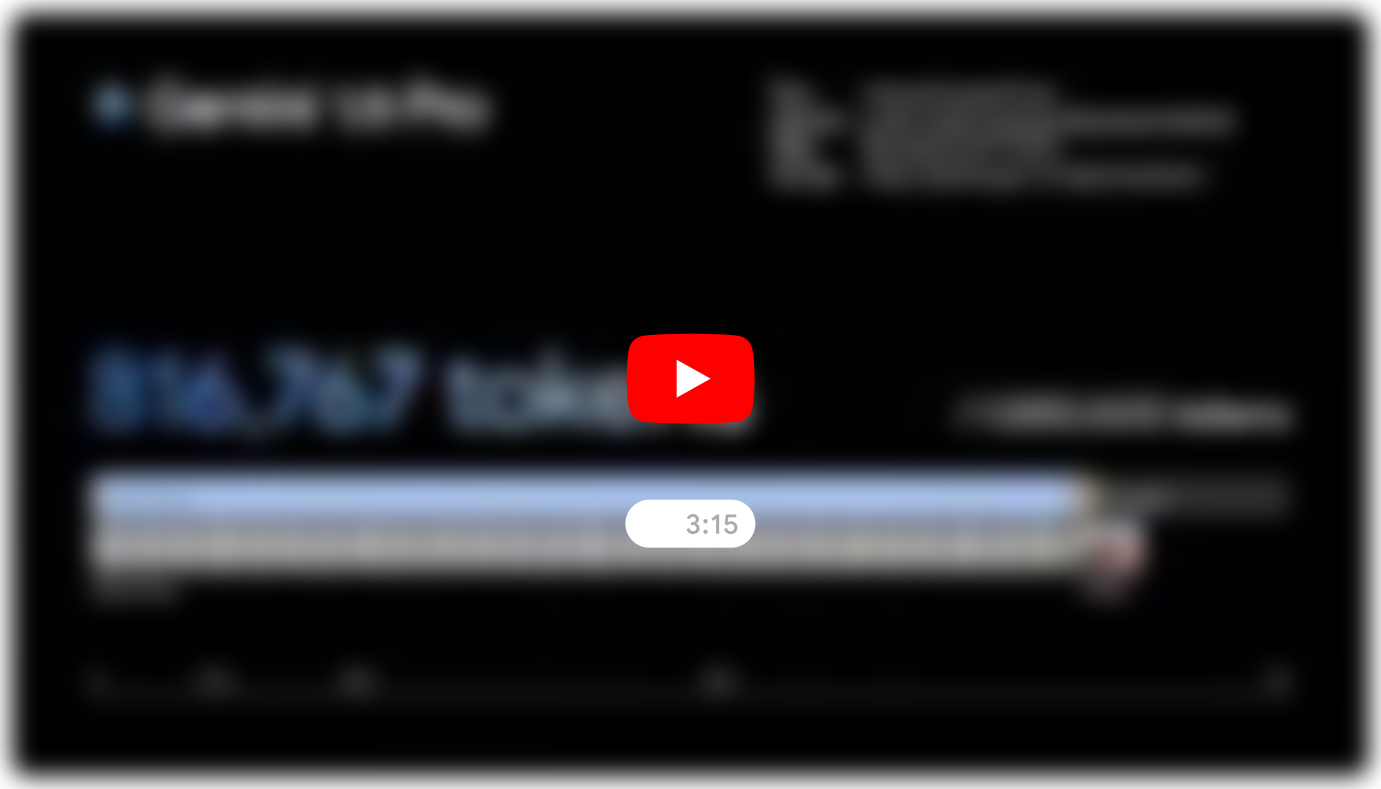
1.5 Pro can perform highly-sophisticated understanding and reasoning tasks for different modalities, including video. For instance, when given a 44-minute silent Buster Keaton movie, the model can accurately analyze various plot points and events, and even reason about small details in the movie that could easily be missed.



Gemini 1.5 Pro can identify a scene in a 44-minute silent Buster Keaton movie when given a simple line drawing as reference material for a real-life object.

Relevant problem-solving with longer blocks of code

1.5 Pro can perform more relevant problem-solving tasks across longer blocks of code. When given a prompt with more than 100,000 lines of code, it can better reason across examples, suggest helpful modifications and give explanations about how different parts of the code works.



Gemini 1.5 Pro can reason across 100,000 lines of code giving helpful solutions, modifications and explanations.

Enhanced performance

When tested on a comprehensive panel of text, code, image, audio and video evaluations, 1.5 Pro outperforms 1.0 Pro on 87% of the benchmarks used for developing our large language models (LLMs). And when compared to 1.0 Ultra on the same benchmarks, it performs at a broadly similar level.

Gemini 1.5 Pro maintains high levels of performance even as its context window increases. In the Needle In A Haystack (NIAH) evaluation, where a small piece of text containing a particular fact or statement is purposely placed within a long block of text, 1.5 Pro found the embedded text 99% of the time, in blocks of data as long as 1 million tokens.

Gemini 1.5 Pro also shows impressive “in-context learning” skills, meaning that it can learn a new skill from information given in a long prompt, without needing additional fine-tuning. We tested this skill on the Machine Translation from One Book (MTOB) benchmark, which shows how well the model learns from information it’s never seen before. When given a grammar manual for Kalamang, a language with fewer than 200 speakers worldwide, the model learns to translate English to Kalamang at a similar level to a person learning from the same content.

As 1.5 Pro's long context window is the first of its kind among large-scale models, we're continuously developing new evaluations and benchmarks for testing its novel capabilities.

For more details, see our Gemini 1.5 Pro technical report.

Extensive ethics and safety testing

In line with our AI Principles and robust safety policies, we're ensuring our models undergo extensive ethics and safety tests. We then integrate these research learnings into our governance processes and model development and evaluations to continuously improve our AI systems.

Since introducing 1.0 Ultra in December, our teams have continued refining the model, making it safer for a wider release. We've also conducted novel research on safety risks and developed red-teaming techniques to test for a range of potential harms.

In advance of releasing 1.5 Pro, we've taken the same approach to responsible deployment as we did for our Gemini 1.0 models, conducting extensive evaluations across areas including content safety and representational harms, and will continue to expand this testing. Beyond this, we're developing further tests that account for the novel long-context capabilities of 1.5 Pro.

Build and experiment with Gemini models

We're committed to bringing each new generation of Gemini models to billions of people, developers and enterprises around the world responsibly.

Starting today, we're offering a limited preview of 1.5 Pro to developers and enterprise customers via AI Studio and Vertex AI. Read more about this on our Google for Developers blog and Google Cloud blog.

We'll introduce 1.5 Pro with a standard 128,000 token context window when the model is ready for a wider release. Coming soon, we plan to introduce pricing tiers that start at the standard 128,000 context window and scale up to 1 million tokens, as we improve the model.

Early testers can try the 1 million token context window at no cost during the testing period, though they should expect longer latency times with this experimental feature. Significant improvements in speed are also on the horizon.

Developers interested in testing 1.5 Pro can sign up now in AI Studio, while enterprise customers can reach out to their Vertex AI account team.

Learn more about Gemini's capabilities and see how it works.

Get more stories from Google in your inbox.

Email address

Your information will be used in accordance with Google's privacy policy.

Subscribe

Try Gemini 1.5 Pro

Developers interested in testing 1.5 Pro can sign up now in AI Studio.

Learn more

POSTED IN:

AI Developers Google Cloud Google DeepMind