

Google launches Gemini 1.0 in Ultra, Pro and Nano sizes

Source copy kept by wheresthe.ai for the story "Google launches Gemini 1.0 in Ultra, Pro and Nano sizes" (<https://wheresthe.ai/d/google-gemini-1-0-launch-2023-12/>).

Source: <https://blog.google/technology/ai/google-gemini-ai/>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 17fe28651b1c4f1489615bcaa125c6bfa32a2d6aa52a4f859f6c9a2afcf7302b

Introducing Gemini: our largest and most capable AI model

Dec 06, 2023

|

[x.com](#)

[Facebook](#)

[LinkedIn](#)

[Mail](#)

[Copy link](#)

Making AI more helpful for everyone

Sundar Pichai

CEO of Google and Alphabet

Demis Hassabis

CEO and Co-Founder, Google DeepMind

[Share](#)

[x.com](#)

[Facebook](#)

[LinkedIn](#)

[Mail](#)

[Copy link](#)

In this article

A note from Google and Alphabet CEO Sundar Pichai: Every technology shift is an opportunity to advance scientific discovery, accelerate human progress, and improve lives. I believe the transition we are seeing right now with AI will be the most profound in our lifetimes, far bigger than the shift to mobile or to the web before it. AI has the potential to create opportunities — from the everyday to the extraordinary — for people everywhere. It will bring new waves of innovation and economic progress and drive knowledge, learning, creativity and productivity on a scale we haven't seen before. That's what excites me: the chance to make AI helpful for everyone, everywhere in the world. Nearly eight years into our journey as an AI-first company, the pace of progress is only accelerating: Millions of people are now using generative AI across our products to do things they couldn't even a year ago, from finding answers to more complex questions to using new tools to collaborate and create. At the same time, developers are using our models and infrastructure to build new generative AI applications, and startups and enterprises around the world are growing with our AI tools. This is incredible momentum, and yet, we're only beginning to scratch the surface of what's possible. We're approaching this work boldly and responsibly. That means being ambitious in our research and pursuing the capabilities that will bring enormous benefits to people and society, while building in safeguards and working collaboratively with governments and experts to address risks as AI becomes more capable. And we continue to invest in the very best tools, foundation models and infrastructure and bring them to our products and to others, guided by our AI Principles . Now, we're

taking the next step on our journey with Gemini, our most capable and general model yet, with state-of-the-art performance across many leading benchmarks. Our first version, Gemini 1.0, is optimized for different sizes: Ultra, Pro and Nano. These are the first models of the Gemini era and the first realization of the vision we had when we formed Google DeepMind earlier this year. This new era of models represents one of the biggest science and engineering efforts we've undertaken as a company. I'm genuinely excited for what's ahead, and for the opportunities Gemini will unlock for people everywhere. — Sundar

Introducing Gemini By Demis Hassabis, CEO and Co-Founder of Google DeepMind, on behalf of the Gemini team AI has been the focus of my life's work, as for many of my research colleagues. Ever since programming AI for computer games as a teenager, and throughout my years as a neuroscience researcher trying to understand the workings of the brain, I've always believed that if we could build smarter machines, we could harness them to benefit humanity in incredible ways. This promise of a world responsibly empowered by AI continues to drive our work at Google DeepMind. For a long time, we've wanted to build a new generation of AI models, inspired by the way people understand and interact with the world. AI that feels less like a smart piece of software and more like something useful and intuitive — an expert helper or assistant. Today, we're a step closer to this vision as we introduce Gemini, the most capable and general model we've ever built. Gemini is the result of large-scale collaborative efforts by teams across Google, including our colleagues at Google Research. It was built from the ground up to be multimodal, which means it can generalize and seamlessly understand, operate across and combine different types of information including text, code, audio, image and video.

Gemini is also our most flexible model yet — able to efficiently run on everything from data centers to mobile devices. Its state-of-the-art capabilities will significantly enhance the way developers and enterprise customers build and scale with AI. We've optimized Gemini 1.0, our first version, for three different sizes: Gemini Ultra — our largest and most capable model for highly complex tasks. Gemini Pro — our best model for scaling across a wide range of tasks. Gemini Nano — our most efficient model for on-device tasks.

State-of-the-art performance We've been rigorously testing our Gemini models and evaluating their performance on a wide variety of tasks. From natural image, audio and video understanding to mathematical reasoning, Gemini Ultra's performance exceeds current state-of-the-art results on 30 of the 32 widely-used academic benchmarks used in large language model (LLM) research and development. With a score of 90.0%, Gemini Ultra is the first model to outperform human experts on MMLU (massive multitask language understanding), which uses a combination of 57 subjects such as math, physics, history, law, medicine and ethics for testing both world knowledge and problem-solving abilities. Our new benchmark approach to MMLU enables Gemini to use its reasoning capabilities to think more carefully before answering difficult questions, leading to significant improvements over just using its first impression.

Gemini surpasses state-of-the-art performance on a range of benchmarks including text and coding.

Gemini Ultra also achieves a state-of-the-art score of 59.4% on the new MMMU benchmark, which consists of multimodal tasks spanning different domains requiring deliberate reasoning. With the image benchmarks we tested, Gemini Ultra outperformed previous state-of-the-art models, without assistance from optical character recognition (OCR) systems that extract text from images for further processing. These benchmarks highlight Gemini's native multimodality and indicate early signs of Gemini's more complex reasoning abilities. See more details in our Gemini technical report.

Gemini surpasses state-of-the-art performance on a range of multimodal benchmarks.

Next-generation capabilities Until now, the standard approach to creating multimodal models involved training separate components for different modalities and then stitching them together to roughly mimic some of this functionality. These models can sometimes be good at performing certain tasks, like describing images, but struggle with more conceptual and complex reasoning. We designed Gemini to be natively multimodal, pre-trained from the start on different modalities. Then we fine-tuned it with additional multimodal data to further refine its effectiveness. This helps Gemini seamlessly understand and reason about all kinds of inputs from the ground up, far better than existing multimodal models — and its capabilities are state of the art in nearly every domain. Learn more about Gemini's capabilities and see how it works. Sophisticated reasoning Gemini 1.0's sophisticated multimodal reasoning capabilities can help make sense of complex written and visual information. This makes it uniquely skilled at uncovering knowledge that can be difficult to discern amid vast amounts of data. Its remarkable ability to extract insights from hundreds of thousands of documents through reading, filtering and understanding information will help deliver new breakthroughs at digital speeds in many fields from science to finance.

Understanding text, images, audio and more Gemini 1.0 was trained to recognize and understand text, images, audio and more at the same time, so it better understands nuanced information and can answer questions relating to complicated topics. This makes it especially good at explaining reasoning in complex subjects like math and physics.

Advanced coding Our first version of Gemini can understand, explain and generate high-quality code in the world's most popular programming languages, like Python, Java, C++, and Go. Its ability to work across languages and reason about complex information makes it one of the leading foundation models for coding in the world. Gemini Ultra excels in several coding benchmarks, including HumanEval , an important industry-standard for evaluating performance on coding tasks, and Natural2Code, our internal held-out dataset, which uses author-generated sources instead of web-based information. Gemini can also be used as the engine for more advanced coding systems. Two years ago we presented AlphaCode , the first AI code generation system to reach a competitive level of performance in programming competitions. Using a specialized version of Gemini, we created a more advanced code generation system, AlphaCode 2 , which excels at solving competitive programming problems that go beyond coding to involve complex math and theoretical computer science.

When evaluated on the same platform as the original AlphaCode, AlphaCode 2 shows massive improvements, solving nearly twice as many problems, and we estimate that it performs better than 85% of competition participants — up from nearly 50% for AlphaCode. When programmers collaborate with AlphaCode 2 by defining certain properties for the code samples to follow, it performs even better. We're excited for programmers to increasingly use highly capable AI models as collaborative tools that can help them reason about the problems, propose code designs and assist with implementation — so they can release apps and design better services, faster. See more details in our AlphaCode 2 technical report .

More reliable, scalable and efficient We trained Gemini 1.0 at scale on our AI-optimized infrastructure using Google's in-house designed Tensor Processing Units (TPUs) v4 and v5e. And we designed it to be our most reliable and scalable model to train, and our most efficient to serve. On TPUs, Gemini runs significantly faster than earlier, smaller and less-capable models. These custom-designed AI accelerators have been at the heart of Google's AI-powered products that serve billions of users like Search, YouTube, Gmail, Google Maps, Google Play and Android. They've also enabled companies around the world to train large-scale AI models cost-efficiently. Today, we're announcing the most powerful, efficient and scalable TPU system to date, Cloud TPU v5p , designed for training cutting-edge AI models. This next generation TPU will accelerate Gemini's development and help developers and enterprise customers train large-scale generative AI models faster, allowing new products and capabilities to reach customers sooner.

A row of Cloud TPU v5p AI accelerator supercomputers in a Google data center.

Built with responsibility and safety at the core At Google, we're committed to advancing bold and responsible AI in everything we do. Building upon Google's AI Principles and the robust safety policies across our products, we're adding new protections to account for Gemini's multimodal capabilities. At each stage of development, we're considering potential risks and working to test and mitigate them. Gemini has the most comprehensive safety evaluations of any Google AI model to date, including for bias and toxicity. We've conducted novel research into potential risk areas like cyber-offense, persuasion and autonomy, and have applied Google Research's best-in-class adversarial testing techniques to help identify critical safety issues in advance of Gemini's deployment. To identify blindspots in our internal evaluation approach, we're working with a diverse group of external experts and partners to stress-test our models across a range of issues. To diagnose content safety issues during Gemini's training phases and ensure its output follows our policies, we're using benchmarks such as Real Toxicity Prompts , a set of 100,000 prompts with varying degrees of toxicity pulled from the web, developed by experts at the Allen Institute for AI. Further details on this work are coming soon. To limit harm, we built dedicated safety classifiers to identify, label and sort out content involving violence or negative stereotypes, for example. Combined with robust filters, this layered approach is designed to make Gemini safer and more inclusive for everyone. Additionally, we're continuing to address known challenges for models such as factuality, grounding, attribution and corroboration. Responsibility and safety will always be central to the development and deployment of our models. This is a long-term commitment that requires building collaboratively, so we're partnering with the industry and broader ecosystem on defining best practices and setting safety and security benchmarks through organizations like MLCommons , the Frontier Model Forum and its AI Safety Fund , and our Secure AI Framework (SAIF) , which was designed to help mitigate security risks specific to AI systems across the public and private sectors. We'll continue partnering with researchers, governments and civil society groups around the world as we develop Gemini.

Making Gemini available to the world Gemini 1.0 is now rolling out across a range of products and platforms:

Gemini Pro in Google products We're bringing Gemini to billions of people through Google products. Starting today, Bard will use a fine-tuned version of Gemini Pro for more advanced reasoning, planning, understanding and more. This is the biggest upgrade to Bard since it launched. It will be available in English in more than 170 countries and territories, and we plan to expand to different modalities and support new languages and locations in the near future. We're also bringing Gemini to Pixel . Pixel 8 Pro is the first smartphone engineered to run Gemini Nano, which is powering new features like Summarize in the Recorder app and rolling out in Smart Reply in Gboard, starting with WhatsApp, Line and KakaoTalk

1 — with more messaging apps coming next year. In the coming months, Gemini will be available in more of our products and services like Search, Ads, Chrome and Duet AI. We're already starting to experiment with Gemini in Search, where it's making our Search Generative Experience (SGE) faster for users, with a 40% reduction in latency in English in the U.S., alongside improvements in quality. Building with Gemini Starting on December 13, developers and enterprise customers can access Gemini Pro via the Gemini API in Google AI Studio or Google Cloud Vertex AI . Google AI Studio is a free, web-based developer tool to prototype and launch apps quickly with an API key. When it's time for a fully-managed AI platform, Vertex AI allows customization of Gemini with full data control and benefits from additional Google Cloud features for enterprise security, safety, privacy and data governance and compliance. Android developers will also be able to build with Gemini Nano, our most efficient model for on-device tasks, via AICore, a new system capability available in Android 14, starting on Pixel 8 Pro devices. Sign up for an early preview of AICore . Gemini Ultra coming soon For Gemini Ultra, we're currently completing extensive trust and safety checks, including red-teaming by trusted external parties, and further refining the model using fine-tuning and reinforcement learning from human feedback (RLHF) before making it broadly available. As part of this process, we'll make Gemini Ultra available to select customers, developers, partners and safety and responsibility experts for early experimentation and feedback before rolling it out to developers and enterprise customers early next year. Early next year, we'll also launch Bard Advanced , a new, cutting-edge AI experience that gives you access to our best models and capabilities, starting with Gemini Ultra.

The Gemini era: enabling a future of innovation This is a significant milestone in the development of AI, and the start of a new era for us at Google as we continue to rapidly innovate and responsibly advance the capabilities of our models. We've made great progress on Gemini so far and we're working hard to further extend its capabilities for future versions, including advances in planning and memory, and increasing the context window for processing even more information to give better responses. We're excited by the amazing possibilities of a world responsibly empowered by AI — a future of innovation that will enhance creativity, extend knowledge, advance science and transform the way billions of people live and work around the world.

Collection

More about Gemini

Explore our collection to find out more about Gemini, the most capable and general model we've ever built.

See more

Get the latest news from Google in your inbox

Sign up for our newsletters with product updates, event information, special offers, and more.

Done. Just one step more.

Check your inbox to confirm your subscription.

You can also subscribe with a different email address .

Your information will be used in accordance with Google's privacy policy. You may opt out at any time.

Posted in:

1

Updated on Dec 13 to include additional messaging apps

Related stories

Quantum computing

Bringing the magic: Google breaks open a bottleneck in quantum computing

By

Emma Rosenfeld

& Craig Gidney

Google Research

How we're helping the world prepare for natural disasters

Our north star is simple yet ambitious: No one should be surprised by a natural disaster.

By

Yossi Matias

& Kate Brandt

Pixel

Making your Pixel devices last longer through design, repair, and reuse

By

Steven Nickel

Google DeepMind

TIME has named Google SynthID one of the Best Inventions of 2026.

Google Research

Research@ 2026: Novel capabilities and intelligence

By

Google Research

Google DeepMind

We're powering European scientific discovery with advanced AI compute and frontier models.

By

Maureen Costello