

Google releases Gemma 4 open models under Apache 2.0

Source copy kept by wheresthe.ai for the story "Google releases Gemma 4 open models under Apache 2.0" (<https://wheresthe.ai/d/google-gemma-4-open-models-2026-04/>).

Source: <https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/>

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

SHA-256 of the saved text: 5759c0156960913f9648cf91dc4c75b33856dde03984ed42d8e142e41462db7e

Gemma 4: Byte for byte, the most capable open models

Apr 02, 2026

|

x.com

Facebook

LinkedIn

Mail

Copy link

Clement Farabet

VP of Research, Google DeepMind

Olivier Lacombe

Director Product Management, Google Deepmind

Share

x.com

Facebook

LinkedIn

Mail

Copy link

. Inlining them here makes them available in the DOM for the page. -->

Your browser does not support the audio element.

Listen to article

[[duration]] minutes

This content is generated by Google AI. Generative AI is experimental

Voice

Speed

Voice

Speed 0.75X 1X 1.5X 2X

Today, we are introducing Gemma 4 — our most intelligent open models to date. Purpose-built for advanced reasoning and agentic workflows, Gemma 4 delivers an unprecedented level of intelligence-per-parameter. This breakthrough builds on incredible community momentum: since the launch of our first generation, developers have downloaded Gemma over 400 million times, building a vibrant Gemmaverse of more than 100,000 variants. We listened closely to what innovators need next to push the boundaries of AI, and Gemma 4 is our answer:

breakthrough capabilities made widely accessible under an Apache 2.0 license .

Open model performance vs size on Arena.ai 's chat arena as of 4/1.

Built from the same world-class research and technology as Gemini 3, Gemma 4 is the most capable model family you can run on your hardware. They complement our Gemini models, giving developers the industry's most powerful combination of both open and proprietary tools. Industry-leading capabilities and mobile-first AI We are releasing Gemma 4 in four versatile sizes: Effective 2B (E2B) , Effective 4B (E4B) , 26B Mixture of Experts (MoE) and 31B Dense . The entire family moves beyond simple chat to handle complex logic and agentic workflows. Our larger models deliver state-of-the-art performance for their sizes, with the 31B model currently ranking as the #3 open model in the world on the industry-standard Arena AI text leaderboard , and the 26B model securing the #6 spot. There, Gemma 4 outcompetes models 20x its size. For developers, this new level of intelligence-per-parameter means achieving frontier-level capabilities with significantly less hardware overhead.

At the edge, our E2B and E4B models redefine on-device utility, prioritizing multimodal capabilities, low-latency processing and seamless ecosystem integration over raw parameter count. Powerful, accessible, open To power the next generation of pioneering research and products, we've sized the Gemma 4 models specifically to run and fine-tune efficiently on hardware — from billions of Android devices worldwide, to laptop GPUs, all the way up to developer workstations and accelerators. By using these highly optimized models, you can fine-tune Gemma 4 to achieve state-of-the-art performance on your specific tasks. We've already seen incredible success with this approach; for instance, INSAIT created a pioneering Bulgarian-first language model (BgGPT), and we worked with Yale University on Cell2Sentence-Scale to discover new pathways for cancer therapy, among many others. Here is what makes Gemma 4 our most capable open model family yet: Advanced reasoning: Capable of multi-step planning and deep logic, Gemma 4 demonstrates significant improvements in math and instruction-following benchmarks that require it. Agentic workflows: Native support for function-calling, structured JSON output, and native system instructions enables you to build autonomous agents that can interact with different tools and APIs and execute workflows reliably. Code generation: Gemma 4 supports high-quality offline code, turning your workstation into a local-first AI code assistant. Vision and audio: All models natively process video and images, supporting variable resolutions, and excelling at visual tasks like OCR and chart understanding. Additionally, the E2B and E4B models feature native audio input for speech recognition and understanding. Longer context: Process long-form content seamlessly. The edge models feature a 128K context window, while the larger models offer up to 256K, allowing you to pass repositories or long documents in a single prompt. 140+ languages: Natively trained on over 140 languages, Gemma 4 helps developers build inclusive, high-performance applications for a global audience. Versatile models for diverse hardware We are releasing the Gemma 4 model weights in sizes tailored for specific hardware and use cases, ensuring you get frontier-class reasoning wherever you need it: 26B and 31B models: Frontier intelligence, offline on your personal computers Optimized to provide researchers and developers with state-of-the-art reasoning on accessible hardware, our unquantized bfloat16 weights fit efficiently on a single 80GB NVIDIA H100 GPU. For local setups, quantized versions run natively on consumer GPUs to power your IDEs, coding assistants and agentic workflows. Our 26B Mixture of Experts (MoE) focus on latency, activating only 3.8 billion of its total parameters during inference to deliver exceptionally fast tokens-per-second, while our 31B Dense is maximizing raw quality and provides a powerful foundation for fine-tuning.

These models were evaluated against a large collection of different datasets and metrics to cover different aspects of text generation. See additional benchmarks in our model card .

E2B and E4B models: A new level of intelligence for mobile and IoT devices Engineered from the ground up for maximum compute and memory efficiency, these models activate an effective 2 billion and 4 billion parameter footprint during inference to preserve RAM and battery life. In close collaboration with our Google Pixel team and mobile hardware leaders like Qualcomm Technologies and MediaTek, these multimodal models run completely offline with near-zero latency across edge devices like phones, Raspberry Pi, and NVIDIA Jetson Orin Nano. Android developers can now prototype agentic flows in the AICore Developer Preview today for forward-compatibility with Gemini Nano 4. An open-source license You gave us feedback, and we listened. Building the future of AI requires a collaborative approach, and we believe in empowering the developer ecosystem without restrictive barriers. That's why Gemma 4 is released under a commercially permissive Apache 2.0 license. This open-source license provides a foundation for complete developer flexibility and digital sovereignty; granting you complete control over your data, infrastructure, and models. It allows you to build freely and deploy securely across any environment, whether on-premises or in the cloud.

Built on a foundation of trust and safety These models undergo the same rigorous infrastructure security protocols as our proprietary models. By choosing Gemma 4, enterprises and sovereign organizations gain a trusted, transparent foundation that delivers state-of-the-art capabilities while meeting the highest standards for security and reliability. An ecosystem of choices Start experimenting in seconds: Get instant access to Gemma 4 and begin building right away. Explore Gemma 4 in Google AI Studio (31B and 26B MoE) or in Google AI Edge Gallery (E4B and E2B). For Android development , use it to power Agent Mode in Android Studio , and start building apps for production on Android with the ML Kit GenAI Prompt API . Use your favorite tools: With day-one support for Hugging Face (Transformers, TRL, Transformers.js, Candle), LiteRT-LM , vLLM, llama.cpp, MLX , Ollama , NVIDIA NIM and NeMo , LM Studio , Unsloth , SGLang, Cactus , Baseten , Docker , MaxText, Tunix, Keras, you have the flexibility to choose the best tools for your project. Download the models: Get the model weights from Hugging Face , Kaggle or Ollama . Customize Gemma 4 to your specific needs: Train and adapt the model using your preferred platform, like Google Colab, Vertex AI or even your gaming GPU. Scale to production on Google Cloud: While local on-device inference is ideal for offline use, Google Cloud removes all compute ceilings. Deploy your way through Vertex AI, Cloud Run , GKE , Sovereign Cloud, TPU-accelerated serving and the highest compliance guarantees for regulated workloads. Learn more about getting started on Google Cloud here . Accelerate your AI development across multiple hardware platforms: Gemma 4 is optimized for industry-leading hardware out of the box. Experience maximum performance on NVIDIA AI infrastructure from NVIDIA Jetson Orin Nano to Blackwell GPUs, integrate with AMD GPUs via the open-source ROCm™ stack, or deploy on Trillium and Ironwood TPUs for massive scale and efficiency. Compete for impact: Join the Gemma 4 Good Challenge on Kaggle to build products that create meaningful, positive change in the world.

Posted in:

Related stories

Safety & Security

Developing Gemini 4 Argon responsibly and safely

By

Anca Dragan

& Raluca Ada Popa

Developer tools

Gemini 4 Argon is now available for developers.

Gemini models

Gemini 4 Argon: Built to solve your hardest problems

By

Koray Kavukcuoglu

Google DeepMind

EmbeddingGemma 2: an open, lightweight multimodal embedding model

By

Sahil Dua

& Henrique Schechter Vera

AI

The latest AI news we announced in September 2026

By

Blog Team

Gemini models

Gemini 4 Argon: our next era of frontier intelligence

By

Koray Kavukcuoglu