What is Google Gemma 4? Everything you need to know about Google's new open-source AI.

Google has launched Gemma 4, a powerful open-source AI model that supports agentic workflows, multimodal applications, and is free to use commercially.

Google has just announced **Gemma 4** , the latest generation in its line of open-source AI models. This is considered the most powerful open-source model Google has ever released, and it allows for free commercial use.

Gemma 4 is built on the same research foundation as Gemini 3. However, unlike Gemini, which is a proprietary model, Gemma 4 is released as open source under the Apache 2.0 license, allowing developers and businesses to use it freely in commercial products.

The release of Gemma 4 marks a new step in Google's AI strategy, as the company begins to push open models to compete directly with rivals such as Meta's Llama line.

What is Gemma 4?

Gemma 4 is a family of open-source AI models designed to work seamlessly across a variety of environments, from local devices to large-scale AI systems. Google's goal is to bring the capabilities of advanced models like Gemini to the development community through an open platform.

One of the highlights of Gemma 4 is its strong support for **agentic workflows** . These new models provide built-in support for function calling, structured JSON output, and system instructions. This allows developers to build AI agents that automatically perform complex tasks, handle multi-step logic, and interact with external APIs directly within their local environment.

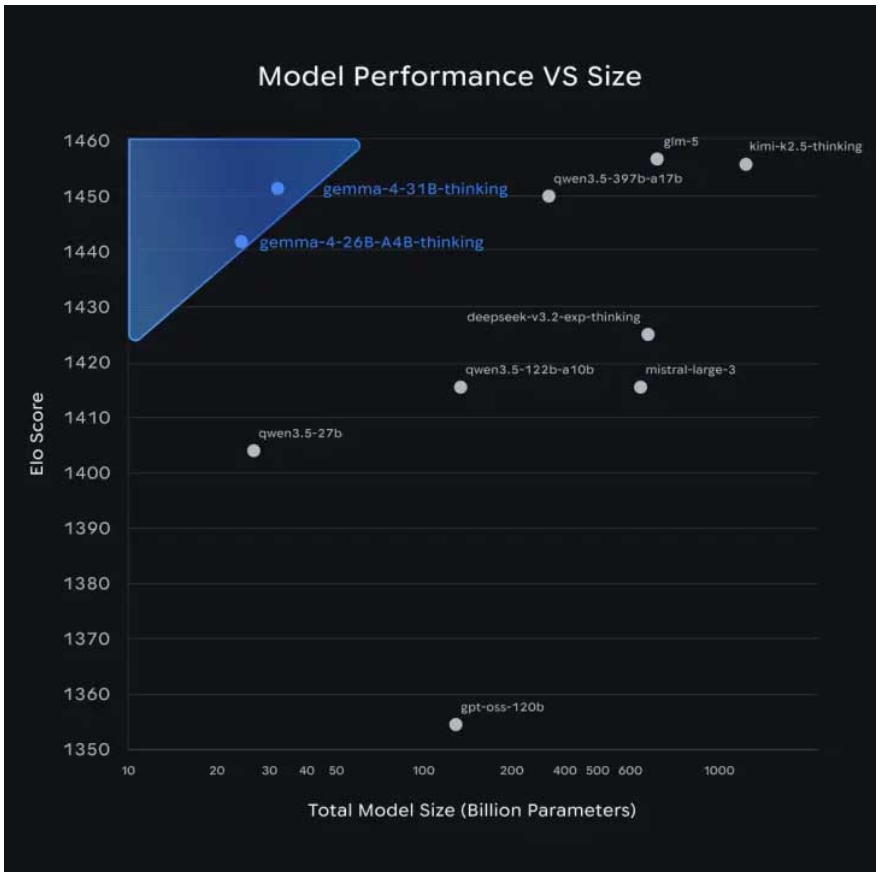
This suggests that Google is positioning Gemma 4 not just as a simple AI model, but also as a platform for building automated AI systems.

How powerful is Gemma 4?

According to Google, Gemma 4 models achieve very high performance in open-source AI rankings. The 31B Dense version currently ranks third on the Arena AI leaderboard, while the 26B version ranks sixth. Notably, these models can outperform many competitors up to 20 times larger.

Another notable point is that the 26B and 31B models can run on a single NVIDIA H100 80GB GPU. This significantly reduces hardware requirements, making deployment easier.

Additionally, Google also introduced version 26B Mixture of Experts (MoE), optimized for low latency. This model only activates about 3.8 billion parameters during inference, helping to speed up content creation. As a result, Gemma 4 can be used to build local programming assistants or real-time AI applications right on mainstream hardware.



Gemma 4 also boasts significant upgrades in multimedia capabilities. The entire model line can handle high-resolution images and video. Edge versions like the E2B and E4B even support audio input, enabling voice recognition with very low latency.

In addition, Gemma 4 also has a large context window. Edge versions support 128K token contexts, while larger models like 26B and 31B can handle up to 256K tokens. This helps the model process long documents or complex tasks more efficiently.

Compatibility and Ecosystem

One significant change in Gemma 4 lies in the Apache 2.0 license. Previous versions of Gemma had their own terms of use, meaning they weren't truly open source.

With Gemma 4, Google allows unrestricted commercial use, model customization, redistribution, and integration into products. This makes Gemma 4 a direct competitor to open-source models like Llama.

This move also shows that Google is taking its open-source AI strategy more seriously.

Gemma 4 is now compatible with many popular platforms such as Hugging Face, Ollama, and vLLM. At the same time, Google has optimized the model for various hardware platforms including NVIDIA, AMD, Qualcomm, and MediaTek.

For mobile developers, Gemma 4 is already available for testing through the AICore Developer Preview. Google also stated that these models will be compatible with Gemini Nano 4 in the future, opening up the possibility of deploying powerful AI on mobile devices.

With its high performance, AI agent support, multimodal capabilities, and true open-source licensing, Gemma 4 is becoming one of the most noteworthy AI models currently available.

Google isn't just releasing an upgrade; it's expanding its AI ecosystem in a more open direction. This gives developers access to more powerful tools and opens up many new AI applications in the future.