


DEVELOPERS

Gemma: Introducing new state-of-the-art open models

Gemma is built for responsible AI development from the same research and technology used to create Gemini models.

Feb 21, 2024 · 3 min read

 Share



Jeanine Banks

VP & GM, Developer X and
DevRel



Tris Warkentin

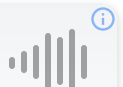
Director, Google DeepMind



Gemma



Listen to article 7 minutes



At Google, we believe in [making AI helpful for everyone](#). We have a long history of contributing innovations to the open community, such as with [Transformers](#), [TensorFlow](#), [BERT](#), [T5](#), [JAX](#), [AlphaFold](#), and [AlphaCode](#). Today, we're excited to introduce a new generation of open models from Google to assist developers and researchers in building AI responsibly.

Gemma open models

[Gemma](#) is a family of lightweight, state-of-the-art [open models](#) built from the same research and technology used to create the [Gemini](#) models. Developed by Google DeepMind and other teams across Google, Gemma is inspired by Gemini, and the name reflects the Latin *gemma*, meaning "precious stone." Accompanying our model weights, we're also releasing tools to support developer innovation, foster collaboration, and guide responsible use of Gemma models.

Gemma is available worldwide, starting today. Here are the key details to know:

- We're releasing model weights in two sizes: [Gemma 2B](#) and [Gemma 7B](#). Each size is released with pre-trained and instruction-tuned variants.
- A new [Responsible Generative AI Toolkit](#) provides guidance and essential tools for creating safer AI applications with Gemma.
- We're providing toolchains for inference and supervised fine-tuning (SFT) across all major frameworks: JAX, PyTorch, and TensorFlow through native [Keras 3.0](#).
- Ready-to-use [Colab](#) and [Kaggle notebooks](#), alongside integration with popular tools such as [Hugging Face](#), [MaxText](#), [NVIDIA NeMo](#) and [TensorRT-LLM](#), make it easy to get started with Gemma.
- [Pre-trained and instruction-tuned Gemma models](#) can run on your laptop, workstation, or Google Cloud with easy deployment on [Vertex AI](#) and [Google Kubernetes Engine](#) (GKE).
- Optimization across multiple AI hardware platforms ensures industry-leading performance, including [NVIDIA GPUs](#) and [Google Cloud TPUs](#).
- [Terms of use](#) permit responsible commercial usage and distribution for all organizations, regardless of size.

State-of-the-art performance at size

Gemma models share technical and infrastructure components with [Gemini](#), our largest and most capable AI model widely available today. This enables Gemma 2B and 7B to achieve best-in-class performance for

their sizes compared to other open models. And Gemma models are capable of running directly on a developer laptop or desktop computer. Notably, Gemma surpasses significantly larger models on key benchmarks while adhering to our rigorous standards for safe and responsible outputs. See the [technical report](#) for details on performance, dataset composition, and modeling methodologies.



Responsible by design

Gemma is designed with our [AI Principles](#) at the forefront. As part of making Gemma pre-trained models safe and reliable, we used automated techniques to filter out certain personal information and other sensitive data from training sets. Additionally, we used extensive fine-tuning and reinforcement learning from human feedback (RLHF) to align our instruction-tuned models with responsible behaviors. To understand and reduce the risk profile for Gemma models, we conducted robust evaluations including manual red-teaming, automated adversarial testing, and assessments of model capabilities for dangerous activities. These evaluations are outlined in our [Model Card](#). ¹

We're also releasing a new [Responsible Generative AI Toolkit](#) together with Gemma to help developers and researchers prioritize building safe and responsible AI applications. The toolkit includes:

- **Safety classification:** We provide a [novel methodology](#) for building robust safety classifiers with minimal examples.
- **Debugging:** A model [debugging tool](#) helps you investigate Gemma's behavior and address potential issues.
- **Guidance:** You can access best practices for model builders based on Google's experience in developing and deploying large language models.

Optimized across frameworks, tools and hardware

You can fine-tune Gemma models on your own data to adapt to specific application needs, such as summarization or retrieval-augmented generation (RAG). Gemma supports a wide variety of tools and systems:

- **Multi-framework tools:** Bring your favorite framework, with reference implementations for inference and fine-tuning across multi-framework Keras 3.0, native PyTorch, JAX, and Hugging Face Transformers.
- **Cross-device compatibility:** Gemma models run across popular device types, including laptop, desktop, IoT, mobile and cloud, enabling broadly accessible AI capabilities.
- **Cutting-edge hardware platforms:** We've [partnered with NVIDIA to optimize Gemma for NVIDIA GPUs](#), from data center to the cloud to local RTX AI PCs, ensuring industry-leading performance and integration with cutting-edge technology.
- **Optimized for Google Cloud:** Vertex AI provides a broad MLOps toolset with a range of tuning options and one-click deployment using built-in inference optimizations. Advanced customization is available with fully-managed Vertex AI tools or with self-managed GKE, including deployment to cost-efficient infrastructure across GPU, TPU, and CPU from either platform.


Free credits for research and development

Gemma is built for the open community of developers and researchers powering AI innovation. You can start working with Gemma today using free access in Kaggle, a free tier for Colab notebooks, and \$300 in credits for first-time Google Cloud users. Researchers can also apply for [Google Cloud credits](#) of up to a collective \$500,000 to accelerate their projects.

Getting started

You can explore more about Gemma and access quickstart guides on ai.google.dev/gemma.

As we continue to expand the Gemma model family, we look forward to introducing new variants for diverse applications. Stay tuned for events and opportunities in the coming weeks to connect, learn and build with Gemma.

We're excited to see what you create! 

POSTED IN:

Developers

AI

[More Information](#)



-
- 1 Google adheres to rigorous data filtering practices to ensure fair evaluation. Our models exclude benchmark data from training sets, ensuring the integrity of benchmark comparisons.

[Collapse](#) 