AI & Machine Learning

Gemma is now available on Google Cloud

February 21, 2024

Today, Google announced Gemma, a family of lightweight, state-of-the art open models built from the same research and technology that we used to create the Gemini models. We're pleased to share that Google Cloud customers can get started today customizing and building with Gemma models in Vertex AI and running them on Google Kubernetes Engine (GKE). The release of Gemma and our updated platform capabilities are the next phase of our commitment to making AI more open and accessible to developers on Google Cloud.

Gemma is now available on Google Cloud

Gemma models share technical and infrastructure components with our capable Gemini models. This enables Gemma models to achieve best-in-class performance for their sizes compared to other open models. We are releasing weights in two sizes: Gemma 2B and Gemma 7B. Each size is released with pre-trained and instruction-tuned variants to enable both research and development.

Gemma supports tools that Google Cloud developers love and use today, including Colab and Kaggle notebooks, as well as frameworks like JAX, PyTorch, Keras 3.0, and Hugging Face Transformers. Gemma models can run on a laptop, workstation, or

Vertex AI and run it on GKE. To maximize industry-leading performance, we have collaborated with NVIDIA to optimize Gemma for NVIDIA GPUs.

Unlocking the power of Gemma in Vertex AI

Gemma joins over 130 models in Vertex AI Model Garden, including our recently announced expanded access to Gemini: Gemini 1.0 Pro, 1.0 Ultra, and 1.5 Pro models.

By using Gemma models on Vertex AI, developers can take advantage of an end-to-end ML platform that makes tuning, managing, and monitoring models simple and intuitive. With Vertex AI, builders can reduce operational overhead and focus on creating bespoke versions of Gemma that are optimized for their use case. For example, using Gemma models on Vertex AI, developers can:

- Build generative AI apps for lightweight tasks such as text generation, summarization, and Q&A
- Enable research and development using lightweight-but-customized models for exploration and experimentation
- Support real-time generative AI use cases that require low latency, such as streaming text

power AI applications of all sizes.

Scale from prototype to production with Gemma on GKE

GKE provides tools to build custom apps, from prototyping simple projects to rolling them out at enterprise scale. Today, developers can also deploy Gemma directly on GKE to create their own gen AI apps for building prototypes or testing model capabilities:

- Deploy custom, fine-tuned models in portable containers alongside applications using familiar toolchains
- Customize model serving and infrastructure configurations without the need to provision or maintain nodes
- Integrate AI Infrastructure fast with the ability to scale to meet the most demanding training and inference scenarios

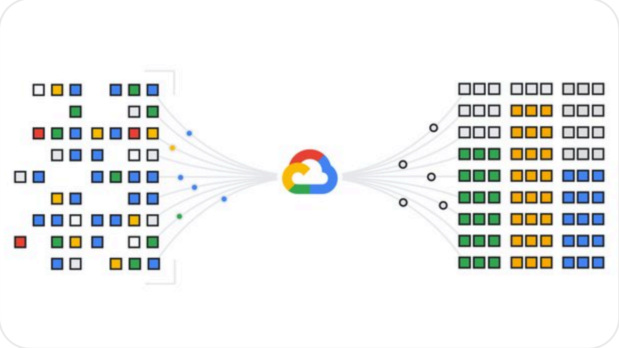
GKE delivers efficient resource management, consistent ops environments, and autoscaling. In addition, it helps enhance these environments with easy orchestration of Google Cloud AI accelerators, including GPUs and TPUs, for faster training and inference when building generative AI models.

Google Cloud today

You can start working with Gemma models today on Google Cloud in Vertex AI and GKE. For more information about Gemma, access quickstart guides on ai.google.dev/gemma.

Posted in AI & Machine Learning

Related articles



AI & Machine Learning

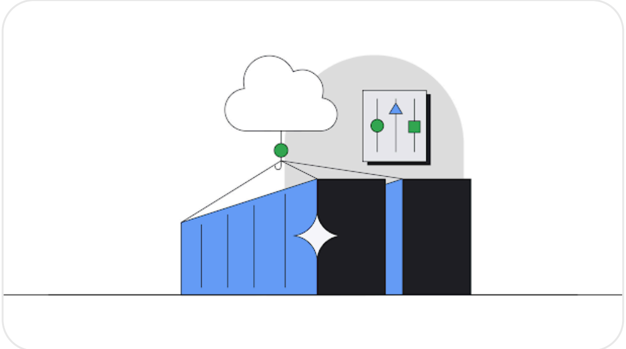
Enterprise Connect 2024 - Bringing AI to the Contact Center

By Kevin Shatzkamer • 4-minute read

Data Analytics

How Palo Alto Networks uses BigQuery ML to automate resource classification

By Gunjan Patel • 5-minute read



Claude 3 Haiku are now generally available on Vertex AI

By Warren Barkley • 4-minute read

with Ray and Kueue

By Andrew Sy Kim • 3-minute read