Google Cloud

Bringing Gemini to Organizations Everywhere

Permalink: https://cloud.google.com/blog/products/ai-machine-learning/bringing-gemini-to-organizations-everywhere

Thomas Kurian, CEO, Google Cloud

Throughout 2023, we have introduced incredible new AI innovations to our customers and the broader developer and user community, including: AI Hypercomputer to train and serve generative AI models; Generative AI support in Vertex AI, our Enterprise AI platform; Duet AI in Google Workspace; and Duet AI for Google Cloud. We have shipped a number of new capabilities in our AI-optimized infrastructure with notable advances in GPUs, TPUs, ML software and compilers, workload management and others; many innovations in Vertex AI; and an entire new suite of capabilities with Duet AI agents in Google Workspace and Google Cloud Platform.

Already, we have seen tremendous developer and user growth. For example, between Q2 and Q3 this year, the number of active gen AI projects on Vertex AI grew by more than 7X. Leading brands like Forbes, Formula E, and Spotify are using Vertex AI to build their own agents, and Anthropic, AI21 Labs, and Cohere are training their models. The breadth and creativity of applications that customers are developing is breathtaking. Fox Sports is creating more engaging content. Priceline is building a digital travel concierge. Six Flags is building a digital concierge. And Estée Lauder is building a digital brand manager.

Today, we are introducing a number of important new capabilities across our AI stack in support of Gemini, our most capable and general model yet. 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 in the same way humans see, hear, read, listen, and talk about many different types of information simultaneously.

Google Cloud's unified AI stack

Starting today, Gemini is part of a vertically-integrated and vertically-optimized AI technology stack that consists of several important pieces - all of which have been engineered to work together:

- **Super-scalable AI infrastructure:** Google Cloud offers leading AI-optimized infrastructure for companies, the same used by Google, to train and serve models. We offer this infrastructure to you in our cloud regions as a service, to run in your data centers with Google Distributed Cloud, and on the edge. Our entire AI infrastructure stack was built with systems-level codesign to boost efficiency and productivity across AI training, tuning, and serving.
- **World-class models:** We continue to deliver a range of AI models with different skills. In late 2022, we launched our Pathways Language Model (PaLM), quickly followed by PaLM 2, and we are now delivering Gemini Pro. We have also introduced domain specific models like Med-PaLM and Sec-PaLM.

- **Vertex AI - Leading enterprise AI platform for developers:** To help developers build agents and integrate gen AI into their applications, we have rapidly enhanced Vertex AI, our AI development platform. Vertex AI helps customers discover, customize, augment, deploy, and manage agents built using the Gemini API, as well as a curated list of more than 130 open-source and third-party AI models that meet Google's strict enterprise safety and quality standards. Vertex AI leverages Google Cloud's built-in data governance and privacy controls, and also provides tooling to help developers use models responsibly and safely. Vertex AI also provides Search and Conversation, tools that use a low code approach to developing sophisticated search and conversational agents that can work across many channels.
- **Duet AI - Assistive AI Agents for Workspace and Google Cloud:** Duet AI is our AI-powered collaborator that provides users with assistance when they use Google Workspace and Google Cloud. Duet AI in Google Workspace, for example, helps users write, create images, analyze spreadsheets, draft and summarize emails, and chat messages, and summarize meetings. Duet AI in Google Cloud, for example, helps users code, deploy, scale, and monitor applications, as well as identify and accelerate resolution of cybersecurity threats.

We are excited to make announcements across each of these areas:

Bolstering our world-class infrastructure

As gen AI models have grown in size and complexity, so have their training, tuning, and inference requirements. As a result, the demand for high-performance, highly-scalable, and cost-efficient AI infrastructure for training and serving models is increasing exponentially.

This isn't just true for our customers, but Google as well. TPUs have long been the basis for training and serving AI-powered products like YouTube, Gmail, Google Maps, Google Play, and Android. In fact, Gemini was trained on, and is served, using TPUs.

Last week, we announced Cloud TPU v5p, our most **powerful, scalable, and flexible** AI accelerator to date. TPU v5p is 4X more scalable than TPU v4 in terms of total available FLOPs per pod. Earlier this year, we announced the general availability of Cloud TPU v5e. With 2.7X inference performance per dollar improvements in an industry benchmark over the previous generation TPU v4, it is our most **cost-efficient** TPU to date.

We also announced our AI Hypercomputer, a groundbreaking supercomputer architecture that employs an integrated system of performance-optimized hardware, open software, leading ML frameworks, and flexible consumption models. AI Hypercomputer has a wide range of accelerator options, including multiple classes of 5th generation TPUs and NVIDIA GPUs.

Providing our latest breakthrough models

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

Today, we also introduced an upgraded version of our image model, Imagen 2, our most advanced text-to-image technology. This latest version delivers improved photorealism, text

rendering, and logo generation capabilities so you can easily create images with text overlays and generate logos.

In addition, building on our efforts around domain-specific models with Med-PaLM, we are excited to announce MedLM, our suite of medically-tuned models. MedLM is available to allowlist customers in Vertex AI, bringing customers the power of Google's foundation models tuned with medical expertise.

Supercharging the Vertex AI platform with Gemini

Today, we are announcing Gemini Pro is now available in preview on Vertex AI. It empowers developers to build new and differentiated agents that can process information across text, code, images, and video at this time. Vertex AI helps you deploy and manage agents to production, automatically evaluate the quality and trustworthiness of agent responses, as well as monitor and manage them.

Vertex AI gives you comprehensive support for Gemini, with the ability to discover, customize, augment, manage, and deploy agents built against the Gemini API, including:

- Multiple ways to customize agents built with Gemini using your own data, including prompt engineering, adapter based fine tuning such as Low-Rank Adaptation (LoRA), reinforcement learning from human feedback (RLHF), and distillation.
- Augmentation tools that enable agents to use embeddings to retrieve, understand, and act on real-world information with configurable retrieval augmented generation (RAG) building blocks. Vertex AI also offers extensions to take actions on behalf of users in third-party applications.
- Grounding to improve quality of responses from Gemini and other AI models by comparing results against high-quality web and enterprise data sources.
- A broad set of controls that help you to be safe and responsible when using gen AI models, including Gemini.

In addition to Gemini support in Vertex AI, today we're also announcing:

- Automatic Side by Side (Auto SxS), an automated tool to compare models. Auto SxS is faster and more cost-efficient than manual model evaluation, as well as customizable across various task specifications to handle new generative AI use cases.
- The addition of Mistral, ImageBind, and DITO into Vertex AI's Model Garden, continuing our commitment to an open model ecosystem.
- We will soon be bringing Gemini Pro into Vertex AI Search and Conversation to help you create engaging, production-grade applications quickly.

Expanding Duet AI's capabilities

With Duet AI, we are committed to helping our customers boost productivity, gain competitive advantages, and ultimately improve their bottom line. Today, Duet AI for Developers and Duet AI in Security Operations are generally available, and we will be incorporating Gemini across our Duet AI portfolio over the next few weeks.

Duet AI for Developers helps users code faster with AI code completion, code generation, and chat in multiple integrated development environments (IDEs). It streamlines repetitive developer tasks and processes with shortcuts for common tasks, including unit test generation and code explanation, speeds troubleshooting and issue remediation, and it helps reduce

context-switching. Duet AI also expedites skills-based learning by giving users the ability to ask questions using natural language chat.

Today, we're also announcing that more than 25 code-assist and knowledge-base partners will contribute datasets specific to their platforms, so users of Duet AI for Developers can receive AI assistance based on partners' coding and data models, product documentation, best practices, and other useful enterprise resources.

Duet AI in Security Operations, Google Cloud's unified security operations platform, can enable defenders to more effectively protect their organizations from cyberattacks. Security teams can elevate their skills and help accelerate threat detection, investigation, and response using the power of gen AI. With Duet AI in Security Operations, we are offering AI assistance first in Chronicle, where users can search vast amounts of data in seconds with custom queries generated from natural language, reduce time-consuming manual reviews, quickly surface critical context by leveraging automatic summaries of case data and alerts, and improve response time using recommendations for next steps to support incident remediation.

Google owns the entire Duet AI technology stack, from the infrastructure and foundation models to the top-level integration and user experience. We're proud that our engineers and researchers uniquely collaborate to bring our latest AI technology breakthroughs to customers with a consistent, unified product experience. Early next year, we plan to expand Duet AI across our portfolio, including Duet AI in BigQuery, Looker, our database products, Apigee, and more.

Propelling the next generation of AI solutions

In addition to these new capabilities across our vertically-integrated AI technology stack, we have competitive pricing that makes Gemini accessible to more organizations, and are expanding our indemnification to help protect you from copyright concerns.

The release of Gemini, combined with our portfolio of super-scalable AI infrastructure, Vertex AI, and Duet AI offers a comprehensive and powerful cloud for developers and customers. With these innovations, Google Cloud is propelling the next generation of AI-powered agents across every industry, empowering organizations to build, use, and successfully adopt gen AI to fuel their digital transformations.