BIZCOMMUNITY

Google releases Gemini, says it’s the next generation AI model

By Lindsey Schutters

6 Dec 2023

There's a new entrant in the AI race and, unlike traditional models that train separate components for different modalities and then combine them, Gemini is designed to be natively multimodal. It is pre-trained from the outset on various modalities and then fine-tuned with additional multimodal data to enhance its effectiveness.

This innovative approach sets Gemini apart from its competitors, including OpenAI’s GPT-4, the previous leader in the MMLU. Gemini is expected to be the most powerful AI ever built, boasting sophisticated multimodal capabilities that allow it to master human-style conversations, language, and content, understand and interpret images, code prolifically and effectively, drive data and analytics, and be used by developers to create new AI apps and APIs.

"Every technology shift is an opportunity to advance scientific discovery, accelerate human progress, and improve lives. I believe the transition we are seeing right now with AI will be the most profound in our lifetimes, far bigger than the shift to mobile or to the web before it," said Google and Alphabet CEO Sundar Pichai in the announcement blog post.

5 ways Google Health is using AI in Africa

24 Oct 2023

"AI has the potential to create opportunities — from the everyday to the extraordinary — for people everywhere. It will bring new waves of innovation and economic progress and drive knowledge, learning, creativity and productivity on a scale we haven’t seen before."

Sophisticated reasoning

Gemini's sophisticated multimodal reasoning capabilities, enabling it to decipher complex written and visual information, suggests that the future of AI might be more general-purpose than the tools we have today. As such, Gemini represents not just a significant advancement in AI technology, but also a potential paradigm shift in the industry.

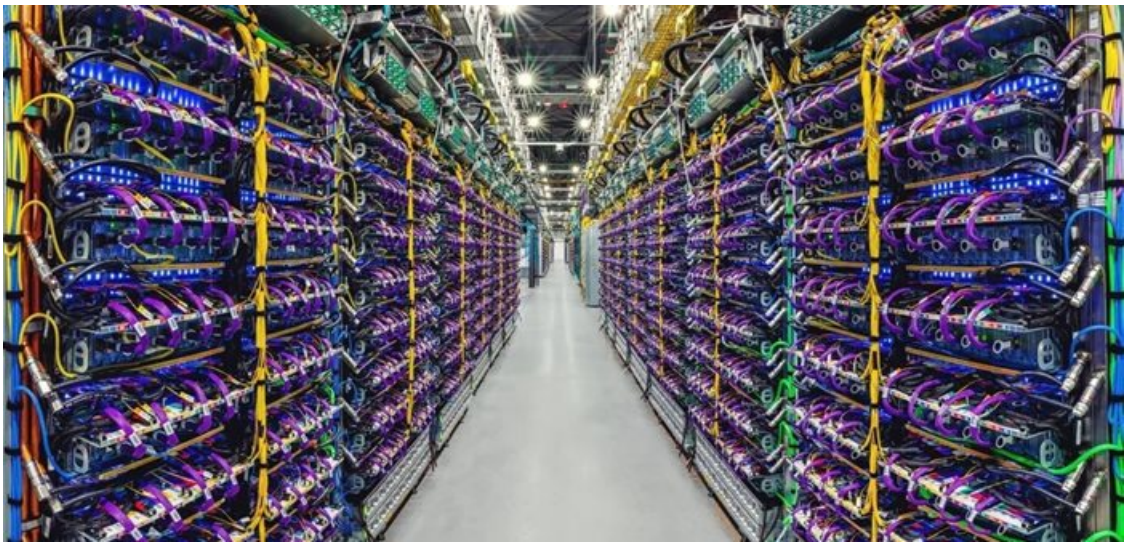
Like all current LLMs, Gemini has the ability to extract insights from hundreds of thousands of documents by reading, filtering, and understanding information, and Google promises to deliver new breakthroughs at digital speeds in various fields, from science to finance.

"We're taking the next step on our journey with Gemini, our most capable and general model yet, with state-of-the-art performance across many leading benchmarks. Our first version, Gemini 1.0, is optimised for different sizes: Ultra, Pro and Nano. These are the first models of the Gemini era and the first realisation of the vision we had when we formed Google DeepMind earlier this year," Pichai explained.

"This new era of models represents one of the biggest science and engineering efforts we've undertaken as a company. I'm genuinely excited for what's ahead, and for the opportunities Gemini will unlock for people everywhere."

The hardware behind Gemini

Gemini 1.0 was trained at scale on Google's AI-optimised infrastructure using the company's in-house designed Tensor Processing Units (TPUs) v4 and v5e. Google explains this as the most reliable and scalable model to train, and the most efficient to serve.



A row of Cloud TPU v5p AI accelerator supercomputers in a Google data centre.

On TPUs, Gemini runs significantly faster than earlier, smaller, and less-capable models. These custom-designed AI accelerators have been integral to Google's AI-powered products that serve billions of users, such as Search, YouTube, Gmail, Google Maps, Google Play, and Android. They have also enabled companies worldwide to train large-scale AI models cost-efficiently.

Google also announced its most powerful, efficient, and scalable TPU system to date, Cloud TPU v5p, designed for training cutting-edge AI models. This next-generation TPU will accelerate Gemini’s development and help developers and enterprise customers train large-scale generative AI models faster. This will allow new products and capabilities to reach customers sooner, marking a significant stride in the field of artificial intelligence.

ABOUT LINDSEY SCHUTTERS

Lindsey is the editor for ICT, Construction&Engineering and Energy&Mining at Bizcommunity

- DCDT overhauls radio frequency spectrum policy - 31 May 2024
- Vodacom goes to war against spectrum pooling - 30 May 2024
- Icasa extends deadline for digital migration regulations review - 27 May 2024
- HPE takes aim at Cisco, emphasises partner ecosystem and AI focus - 24 May 2024
- OpenAI inks News Corp deal, Google threatens to cut news funding - 23 May 2024

View my profile and articles...

For more, visit: https://www.bizcommunity.com