NEWS RELEASE

Cloudflare Enhances AI Inference Platform with Powerful GPU Upgrade, Faster Inference, Larger Models, Observability, and Upgraded Vector Database

2024-09-26

Workers AI is the easiest place to build and scale AI applications; can now deploy larger models and handle more complex AI tasks

SAN FRANCISCO--(BUSINESS WIRE)-- **Cloudflare, Inc.** (NYSE: NET), the leading connectivity cloud company, today announced powerful new capabilities for Workers AI, the serverless AI platform, and its suite of AI application building blocks, to help developers build faster, more powerful and more performant AI applications. Applications built on Workers AI can now benefit from faster inference, bigger models, improved performance analytics, and more. Workers AI is the easiest platform to build global AI applications and run AI inference close to the user, no matter where in the world they are.

As large language models (LLMs) become smaller and more performant, network speeds will become the bottleneck to customer adoption and seamless AI interactions. Cloudflare's globally distributed network helps to minimize network latency, setting it apart from other networks that are typically made up of concentrated resources in limited data centers. Cloudflare's serverless inference platform, Workers AI, now has GPUs in more than 180 cities around the world, built for global accessibility to provide low latency times for end users all over the world. With this network of GPUs, Workers AI has one of the largest global footprints of any AI platform, and has been designed to run AI inference locally as close to the user as possible and help keep customer data closer to home.

“As AI took off last year, no one was thinking about network speeds as a reason for AI latency, because it was still a novel, experimental interaction. But as we get closer to AI becoming a part of our daily lives, the network, and milliseconds, will matter,” said Matthew Prince, co-founder and CEO, Cloudflare. “As AI workloads shift from training to inference, performance and regional availability are going to be critical to supporting the next phase of AI. Cloudflare is the most global AI platform on the market, and having GPUs in cities around the world is going to be what takes AI from a novel toy to a part of our everyday life, just like faster Internet did for smartphones.”

Cloudflare is also introducing new capabilities that make it the easiest platform to build AI applications with:

- Upgraded performance and support for larger models: Now, Cloudflare is enhancing their global network with more powerful GPUs for Workers AI to upgrade AI inference performance and run inference on significantly larger models like Llama 3.1 70B, as well as the collection of Llama 3.2 models with 1B, 3B, 11B (and 90B soon). By supporting larger models, faster response times, and larger context windows, AI applications built on Cloudflare’s Workers AI can handle more complex tasks with greater efficiency – thus creating natural, seamless end-user experiences.
- Improved monitoring and optimizing of AI usage with persistent logs: New persistent logs in AI Gateway, available in open beta, allow developers to store users’ prompts and model responses for extended periods to better analyze and understand how their application performs. With persistent logs, developers can gain more detailed insights from users’ experiences, including cost and duration of requests, to help refine their application. Over two billion requests have traveled through AI Gateway since launch last year.
- Faster and more affordable queries: Vector databases make it easier for models to remember previous inputs, allowing machine learning to be used to power search, recommendations, and text generation use-cases. Cloudflare’s vector database, Vectorize, is now generally available, and as of August 2024 now supports indexes of up to five million vectors each, up from 200,000 previously. Median query latency is now down to 31 milliseconds (ms), compared to 549 ms. These improvements allow AI applications to find relevant information quickly with less data processing, which also means more affordable AI applications.

To learn more, please check out the resources below:

- Blog: **Cloudflare’s Bigger, Better, Faster AI platform**
- Blog: **Making Workers AI faster and more efficient: Performance optimization with KV cache compression and speculative decoding**
- Join us online for demos, product announcements, and more at our first Builder Day Live Stream today, September 26 at 11am PT. Register at **https://builderday.pages.dev** .

About Cloudflare

Cloudflare, Inc. (NYSE: NET) is the leading connectivity cloud company on a mission to help build a better Internet. It

empowers organizations to make their employees, applications and networks faster and more secure everywhere, while reducing complexity and cost. Cloudflare's connectivity cloud delivers the most full-featured, unified platform of cloud-native products and developer tools, so any organization can gain the control they need to work, develop, and accelerate their business.

Powered by one of the world's largest and most interconnected networks, Cloudflare blocks billions of threats online for its customers every day. It is trusted by millions of organizations – from the largest brands to entrepreneurs and small businesses to nonprofits, humanitarian groups, and governments across the globe.

Learn more about Cloudflare's connectivity cloud at **cloudflare.com/connectivity-cloud** . Learn more about the latest Internet trends and insights at **https://radar.cloudflare.com** .

Follow us: **Blog** | **X** | **LinkedIn** | **Facebook** | **Instagram**

Forward-Looking Statements

This press release contains forward-looking statements within the meaning of Section 27A of the Securities Act of 1933, as amended, and Section 21E of the Securities Exchange Act of 1934, as amended, which statements involve substantial risks and uncertainties. In some cases, you can identify forward-looking statements because they contain words such as "may," "will," "should," "expect," "explore," "plan," "anticipate," "could," "intend," "target," "project," "contemplate," "believe," "estimate," "predict," "potential," or "continue," or the negative of these words, or other similar terms or expressions that concern Cloudflare's expectations, strategy, plans, or intentions. However, not all forward-looking statements contain these identifying words. Forward-looking statements expressed or implied in this press release include, but are not limited to, statements regarding the capabilities and effectiveness of Workers AI, AI Gateway, Vectorize, R2, and Cloudflare's other products and technology, the benefits to Cloudflare's customers from using Workers AI, AI Gateway, Vectorize, R2, and Cloudflare's other products and technology, the timing of when Workers AI, AI Gateway, Vectorize, R2, or any of its related features will be generally available to all current and potential Cloudflare customers, Cloudflare's technological development, future operations, growth, initiatives, or strategies, and comments made by Cloudflare's CEO and others. Actual results could differ materially from those stated or implied in forward-looking statements due to a number of factors, including but not limited to, risks detailed in Cloudflare's filings with the Securities and Exchange Commission (SEC), including Cloudflare's Quarterly Report on Form 10-Q filed on August 1, 2024, as well as other filings that Cloudflare may make from time to time with the SEC.

The forward-looking statements made in this press release relate only to events as of the date on which the statements are made. Cloudflare undertakes no obligation to update any forward-looking statements made in this press release to reflect events or circumstances after the date of this press release or to reflect new information or

the occurrence of unanticipated events, except as required by law. Cloudflare may not actually achieve the plans, intentions, or expectations disclosed in Cloudflare's forward-looking statements, and you should not place undue reliance on Cloudflare's forward-looking statements.

Cloudflare, Inc.
Daniella Vallurupalli
Vice President, Head of Global Communications
press@cloudflare.com

Source: Cloudflare, Inc.