CLOUDFLARE

SOLUTION BRIEF

Cloudflare for AI

Scale rapidly and manage costs while delivering innovative generative AI products

The future of apps is intelligent

Build locally and scale globally

Artificial intelligence (AI) businesses are booming and more developer teams are building and rapidly scaling generative AI applications. Meanwhile, coders use their own AI/ML capabilities to increase productivity, shorten learning curves, and improve the overall quality and velocity of coding projects. But they face significant and costly challenges like scaling GPU capacity, storing massive data user-generated datasets, and addressing new security gaps and data exposure.

By building on the Cloudflare network, AI companies can develop a cost-effective cloud architecture that scales at the pace they need, with industry-leading edge compute capacity powered by Cloudflare's network. Teams can reduce or eliminate egress costs while ensuring security and reliability.

Launch your next rocket ship with Cloudflare

Cloudflare for Startups is our initiative to help startups accelerate and protect their Internet properties. Our goal is to share the power of Cloudflare's network and solutions with the startup community. Cloudflare for Startups is by invitation only. Interested organizations can apply here.

Serverless AI on GPUs

Run fast, low-latency inference tasks on our global network of GPUs with Workers AI. Choose from popular models including Llama-2 and ResNet50.. Pre-built LLMs include image and text classification, similarity, and translation.

Store & search embeddings

Speed up and scale your AI Workflows with Vectorize, our vector database. Store new or existing embeddings to enable search on top of your own data for repeated use with machine learning models.

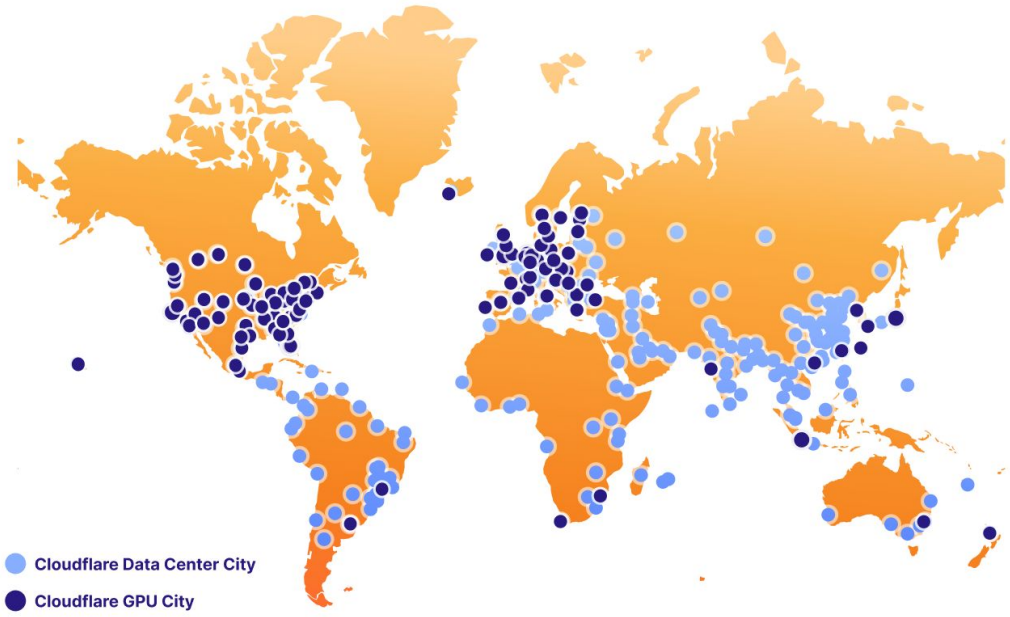
Control and protect

Protect back-end infrastructure and avoid surprise bills. The AI Gateway adds a layer of control and protection with rate-limits, caching, and visibility into how many people are using LLM applications.

Global infrastructure to power and secure your AI projects

Cloudflare is a connectivity cloud designed to make everything you connect to the Internet secure, private, fast, and reliable. Our developer platform runs on top of that network to deliver edge compute & GPU capacity wherever you need it. Your teams can improve productivity and skip time-wasting steps with a straight path from code to serverless deployment, with compute and storage available on demand.

Network and GPU locations



End of 2023