CLOUDFLARE

Serve your AI models with Cloudflare

Fast, affordable & global AI inference

AI inference-as-a-service for AI labs & model developers

Run AI models closest to your users on Cloudflare's global infrastructure, without worrying about scaling, maintaining, or paying for unused infrastructure

Workers AI is a global AI serverless inference service which means you can run AI models, on Cloudflare Network from your own code and is part of a single platform that enables companies to build and scale AI applications.

- Run your AI closer to users delivering low-latency, high-performance applications
- Integrates with Cloudflare's Vectorize (vector database) and R2 (data lake) for fast RAG
- Centralized monitoring, control & security for your AI applications with Cloudflare's AI Gateway

Our partners: Meta Deepgram stability.ai Google DeepMind Leonardo.Ai

"By hosting our voice models on Cloudflare's Workers AI, we're enabling developers to create real-time, expressive voice agents with ultra-low latency. Cloudflare's global network brings AI compute closer to users everywhere, so customers can now deliver lightning-fast conversational AI experiences without worrying about complex infrastructure."

Adam Sypniewski
CTO, Deepgram

AI-adjacent developer primitives built-in

Let your users use primitives like **WebSockets**, **WebRTC** and **QUIC** with your models without additional development or infrastructure to maintain.

Zero egress fees - pay only for inference time

Run AI inference at edge without worrying about surprise costs. Billed by the millisecond, without any egress cost - you pay only for what you use.

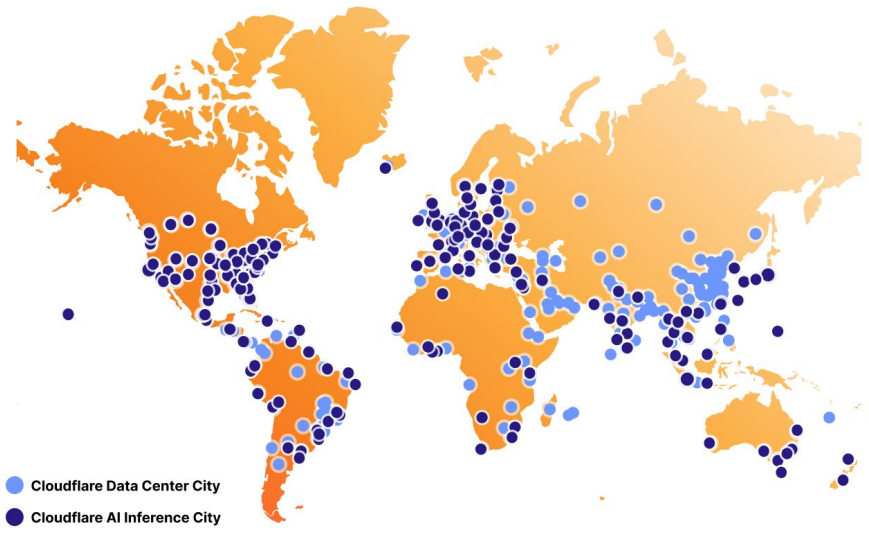
Low Latency AI

Globally distributed infrastructure to run AI models closer to users, with the latest GPU hardware, ensuring high-performance applications.

Build with Workers AI:

- Text Generation
- Summarization
- Image Generation
- Text Embeddings
- Text Classification
- Translation
- Image-To-Text
- Image Classification
- Automatic Speech Recognition
- Object Detection
- Low Rank Adaptation (LoRA)
- Function Calling

Globally distributed infrastructure for better performance



335
cities in 125+ countries, including mainland China

200+
cities with AI inferences worldwide

~50 ms
from 95% of the world's Internet-connected population

