

Workers AI

Fast, affordable & global open-source AI inference

Workers AI Overview

Build and deploy your AI applications

Workers AI is a global AI serverless inference service and is part of a single platform that enables companies to build, deploy, store and monitor scalable AI applications.

- Running AI closer to users delivers low-latency, high-performance applications
- Prebuilt use case templates & 50+ AI models for your business needs
- Integrates with Cloudflare's Vectorize, vector database for lightning fast RAG
- Centralized monitoring, control & security for your AI applications with Cloudflare's AI Gateway

Available with prebuilt AI application templates or choose from best of breed AI models to build your custom AI applications.



Workers AI: Inference as a Service

Global AI inference as a service, accelerated by our partners:



Meta



Run Production AI Applications



Build on one platform

Build AI applications in one platform that includes data storage & serverless inference that runs closest to users.



Minimize Operational Costs


Managed AI inference enables faster AI deployment with less complexity. Free engineers from infrastructure management for more valuable work.




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

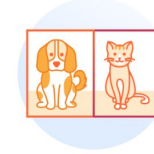
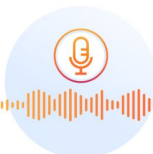



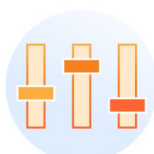
Image Classification




Automatic Speech Recognition



Object Detection



Low Rank Adaptation (LoRA)



Function Calling

Globally distributed infrastructure for better performance

