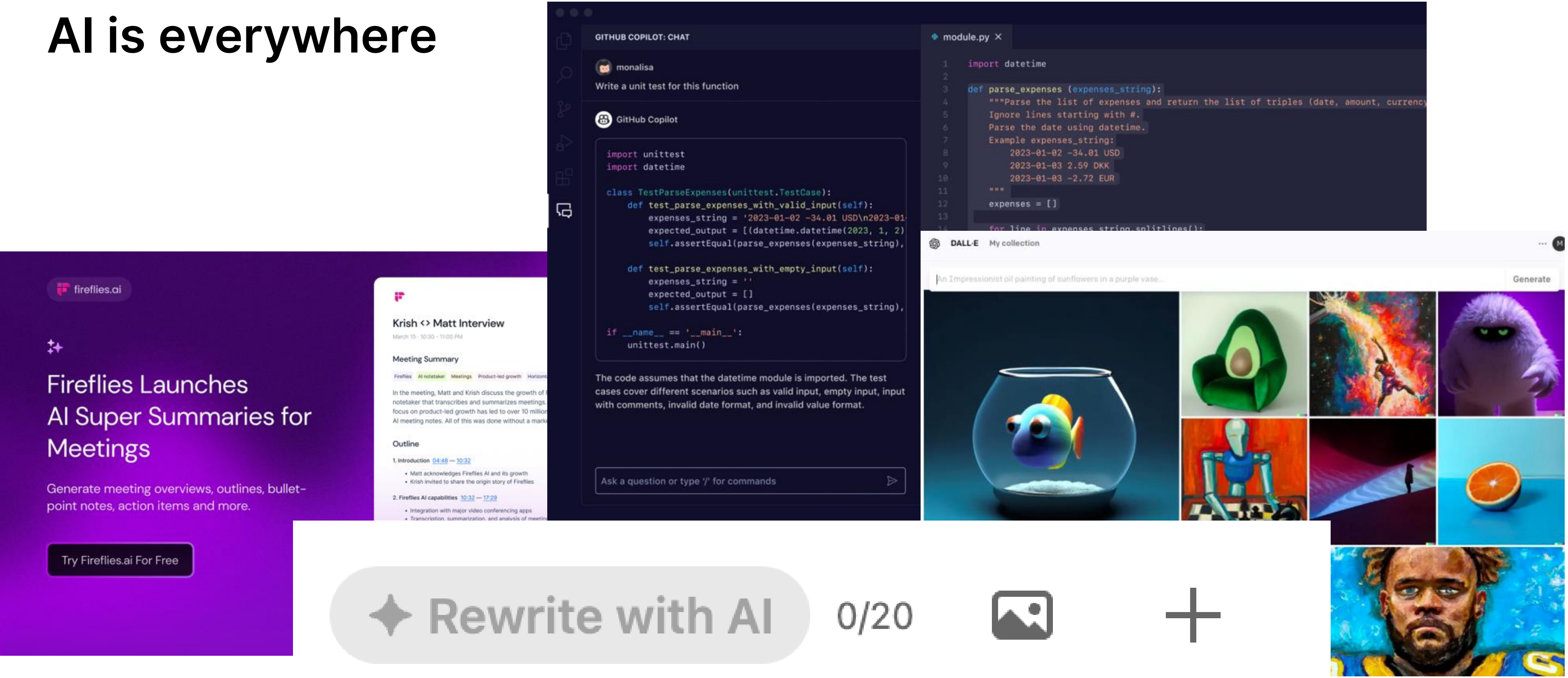
Cloudflare AI
Solution Overview

Agenda

1. Current state
2. The Cloudflare solution
3. How it works
4. Next steps

CURRENT STATE

AI is everywhere



GITHUB COPILOT: CHAT

monalisa

Write a unit test for this function

GitHub Copilot

```
import unittest
import datetime

class TestParseExpenses(unittest.TestCase):
    def test_parse_expenses_with_valid_input(self):
        expenses_string = '2023-01-02 -34.01 USD\n2023-01
        expected_output = [(datetime.datetime(2023, 1, 2)
        self.assertEqual(parse_expenses(expenses_string),

    def test_parse_expenses_with_empty_input(self):
        expenses_string = ''
        expected_output = []
        self.assertEqual(parse_expenses(expenses_string),

if __name__ == '__main__':
    unittest.main()
```

The code assumes that the datetime module is imported. The test cases cover different scenarios such as valid input, empty input, input with comments, invalid date format, and invalid value format.

Ask a question or type '/' for commands

module.py

```
import datetime

def parse_expenses (expenses_string):
    """Parse the list of expenses and return the list of triples (date, amount, currency
    Ignore lines starting with #.
    Parse the date using datetime.
    Example expenses_string:
        2023-01-02 -34.01 USD
        2023-01-03 2.59 DKK
        2023-01-03 -2.72 EUR
    """
    expenses = []

    for line in expenses_string.splitlines():
```

DALL·E My collection

An Impressionist oil painting of sunflowers in a purple vase...

Generate

fireflies.ai

Fireflies Launches AI Super Summaries for Meetings

Generate meeting overviews, outlines, bullet-point notes, action items and more.

Try Fireflies.ai For Free

Krish <> Matt Interview

March 15 · 10:30 - 11:00 PM

Meeting Summary

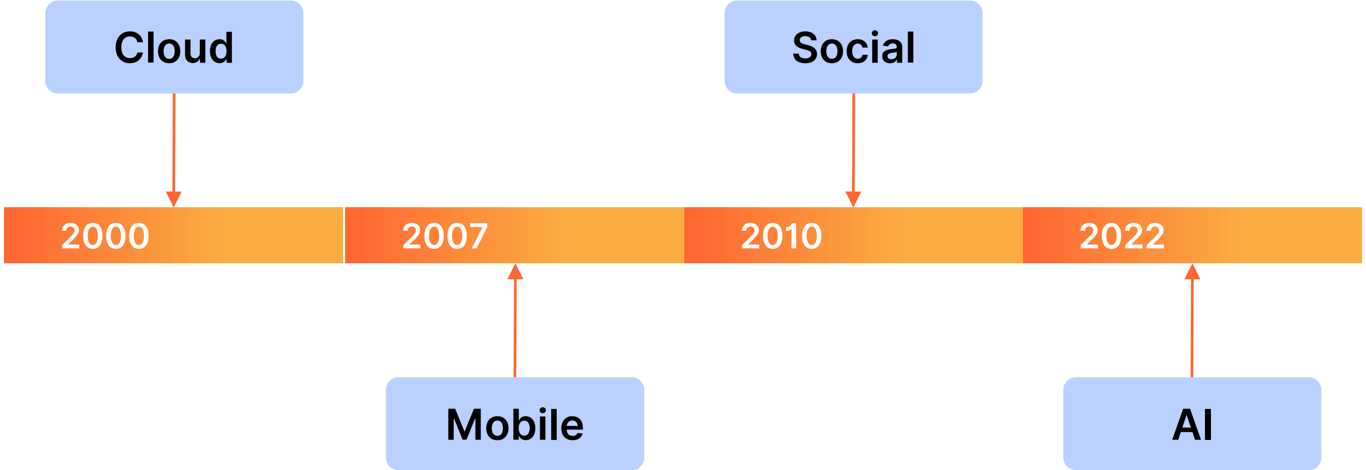
Fireflies | AI notetaker | Meetings | Product-led growth | Horizont

In the meeting, Matt and Krish discuss the growth of F
notetaker that transcribes and summarizes meetings.
focus on product-led growth has led to over 10 million
AI meeting notes. All of this was done without a mark

Outline

1. Introduction 04:48 — 10:32
   - Matt acknowledges Fireflies AI and its growth
   - Krish invited to share the origin story of Fireflies
2. Fireflies AI capabilities 10:32 — 17:29
   - Integration with major video conferencing apps
   - Transcription, summarization, and analysis of meetin

AI is the next paradigm shift



Cloud

Social

2000

2007

2010

2022

Mobile

AI

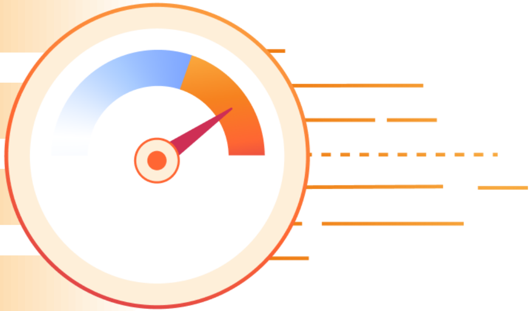
> “The ability to build a generative AI business without needing to invest significant dollars developing sophisticated models... significantly reduces (and even eliminates) the upfront costs previously associated with building an AI business.”

Source: Base10

CURRENT STATE

AI accelerates the way we work

- Write and edit content, copy, or code
- Create logos and graphics
- Generate music and videos
- Summarize documents and insights
- Generate code and documentation
- Provide customer support



All you need is 30 seconds to:

Shorten learning curve

Increase productivity

Improve accuracy

Increase efficiency

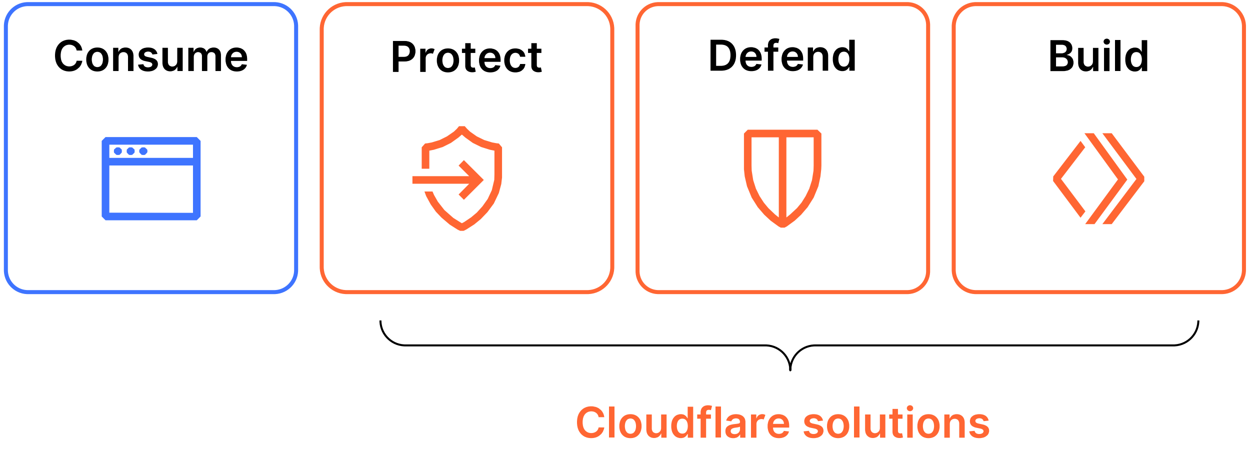
CURRENT STATE

AI initiatives face challenges

Initiative	CHALLENGES
GPUs for training and inference	High demand has led to scarcity of resources
Multi-cloud architecture	Cost of moving data across clouds
Control operational costs	Lack of visibility into usage
Reduce cyber risk	Data leaks, privacy, shadow IT

CURRENT STATE

AI impact on your business



Consume

Protect

Defend

Build

Cloudflare solutions

Where are you using AI today:

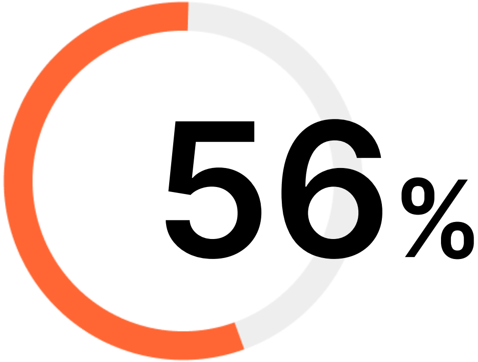
Engineering

Infrastructure

Security services

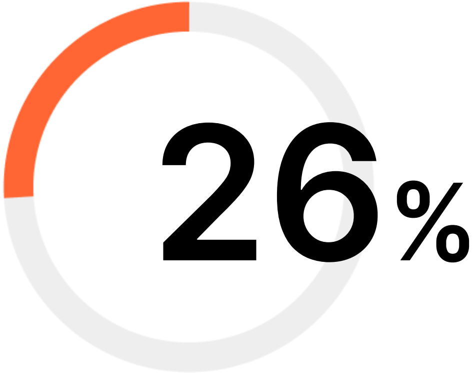
Sales & marketing

Consume



56%

Of employees have used AI-powered tools for work



26%

Of companies have an AI-policy

Source: Conference Board Survey

Section

Protect

Cloudflare can help reduce cyber risks as you experiment with AI tools

Cyber Risks

Cybersecurity & Fraud · Regulatory Compliance · Visibility · Third Party · Consumer Protection · Liability · Unreliable Outputs · Model & Output Bias · Escalating Cost

Manage Governance & Risk

Define & communicate acceptable use policies

Impact assessments per privacy & AI regulations

Monitor workforce & automated service usage via app & API traffic

Increase Security and Privacy Controls

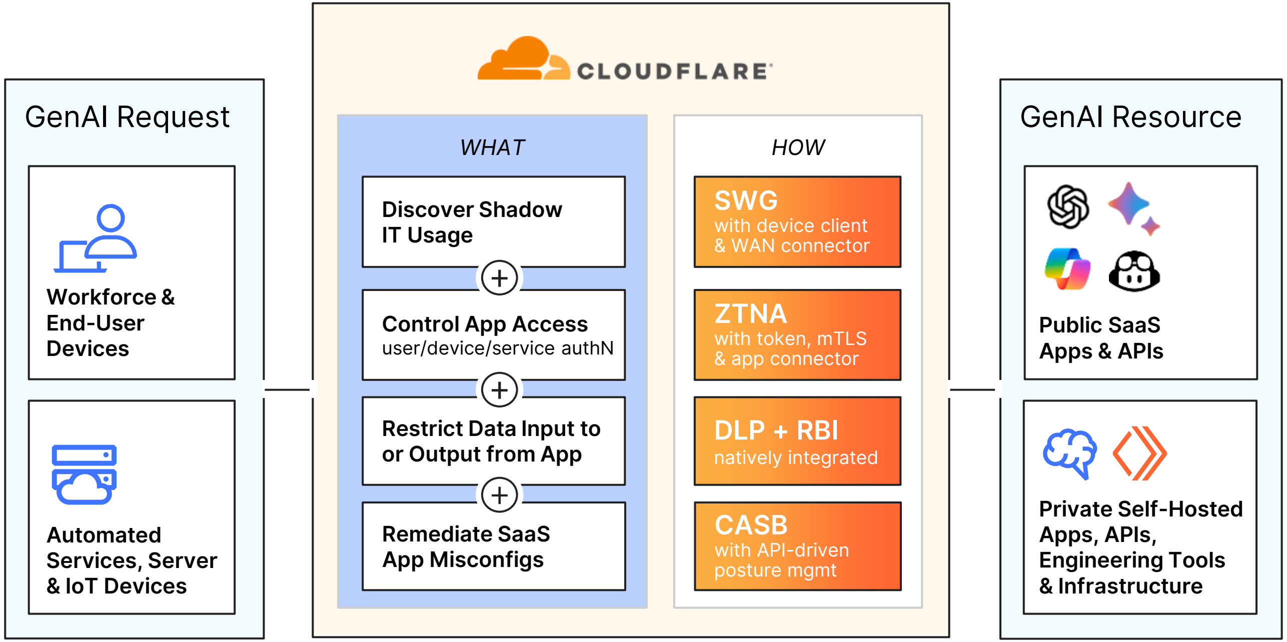
Filter data inputs & outputs for IP, confidentiality & copyright

Block access as last resort since it often triggers “user bypass” & Shadow IT activity

Reduce Attack Surface

Identity-based Zero Trust access controls to apps, engineering tools & infrastructure

Protect employees and resources



GenAI Request

Workforce & End-User Devices

Automated Services, Server & IoT Devices

CLOUDFLARE

WHAT

Discover Shadow IT Usage

+

Control App Access
user/device/service authN

+

Restrict Data Input to or Output from App

+

Remediate SaaS App Misconfigs

HOW

SWG
with device client & WAN connector

ZTNA
with token, mTLS & app connector

DLP + RBI
natively integrated

CASB
with API-driven posture mgmt

GenAI Resource

Public SaaS Apps & APIs

Private Self-Hosted Apps, APIs, Engineering Tools & Infrastructure

Protect

Gain usage visibility

Control data flows

Facilitate compliance

Improve performance

Protect data

Comply with regulations

Detect and control movement of regulated data. Push logs to SIEM for audit trails.

- ✓ GDPR
- ✓ CCPA
- ✓ GLBA
- ✓ HIPAA
- ✓ DPDP
- ✓ CPRA
- ✓ PCI DSS
- ✓ ISO

Data exposure visibility

Regain visibility and controls for sensitive data across SaaS apps, shadow IT, and emerging AI tools.

- → OpenAI
- → Bard
- → GitHub Copilot

Secure developer code

Detect and block source code in up/downloads. Find and fix misconfigurations in SaaS apps and code repositories.

- → GitHub
- → GitLab
- → Bitbucket

Section

Defend

CLOUDFLARE

Threat Actors also making “Investments” in AI

THE WALL STREET JOURNAL.

AI Experts Warn of Potential Cyberwar Facing Banking Sector

U.S. financial institutions’ machine-learning models are a potential avenue for attacks, experts said

---

Harvard
Business
Review

The New Risks ChatGPT Poses to Cybersecurity

The New York Times

A.I. Poses ‘Risk of Extinction,’ Industry Leaders Warn

Leaders from OpenAI, Google DeepMind, Anthropic and other A.I. labs warn that future systems could be as deadly as pandemics and nuclear weapons.

CNBC

Workers are secretly using ChatGPT, AI and it will pose big risks for tech leaders

---

The Verge

Microsoft and OpenAI say hackers are using ChatGPT to improve

---

THE WALL STREET JOURNAL.

AI Junk Is Starting to Pollute the Internet

Online publishers are inundated with useless article pitches as websites using AI-generated content multiply

GenAI Image created in Midjourney

AI is a new attack surface

- AI-powered phishing
- Social engineering
- Deep fakes
- Voice spoofing
- Botnet management

Firewall for AI

Addresses LLM security concerns:

- Prompt injection
- Insecure plugin design
- Sensitive information disclosure
- Excessive agency

Section

Build

Elements of AI applications and infrastructure

Train

- Store training data
- Create and store model
- Fine tune models

Secure

- Secure access to AI services
- Protect private data

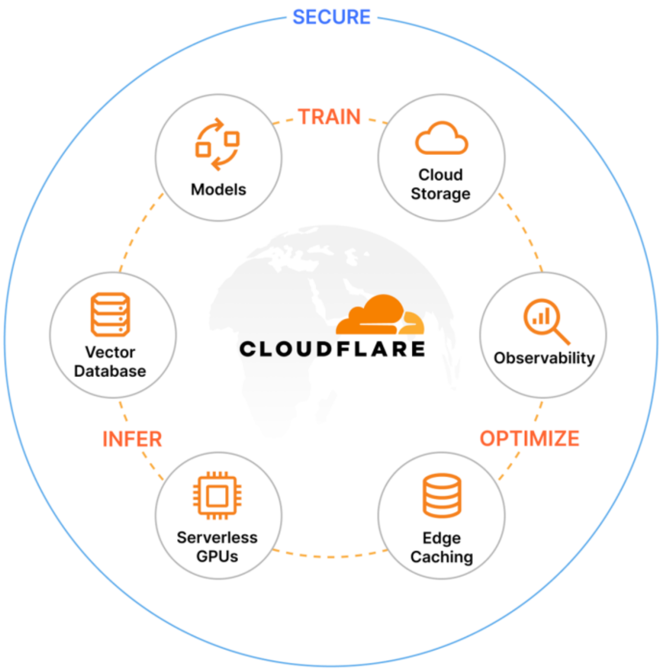
Deploy

- Generate embeddings
- Deploy model
- Run inference at scale

Optimize

- Observe and analyze usage
- Control traffic & optimize cost

A holistic, end-to-end AI solution



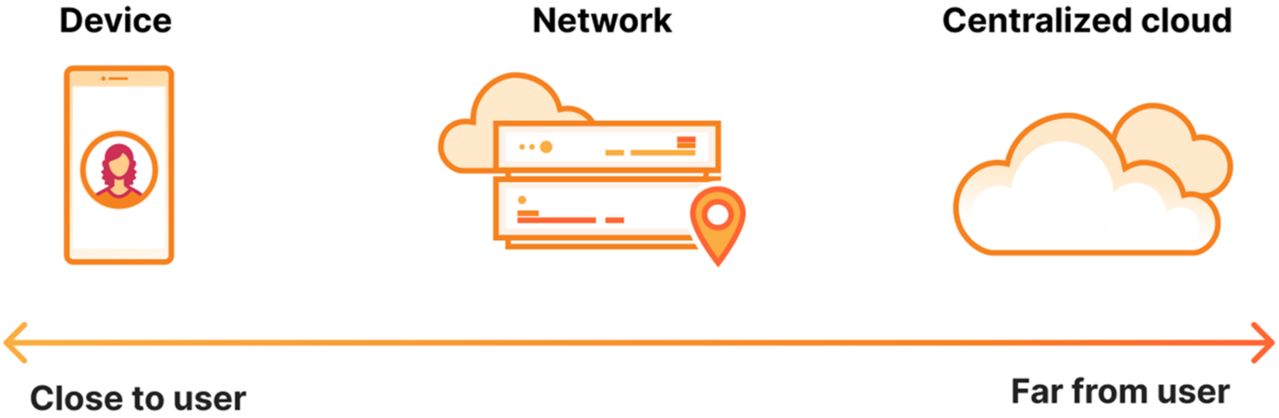
1) TRAIN
Zero Egress Cloud Storage with **R2**

2) INFER
Generate, store, and search embeddings with **Workers AI** and **Vectorize**

3) OPTIMIZE
Gain visibility and improve performance via **AI Gateway**

4) SECURE
Protect and Defend with **Zero Trust** and **API Security**

Most AI workloads are going to run near the device, close to the user



Device

Network

Centralized cloud

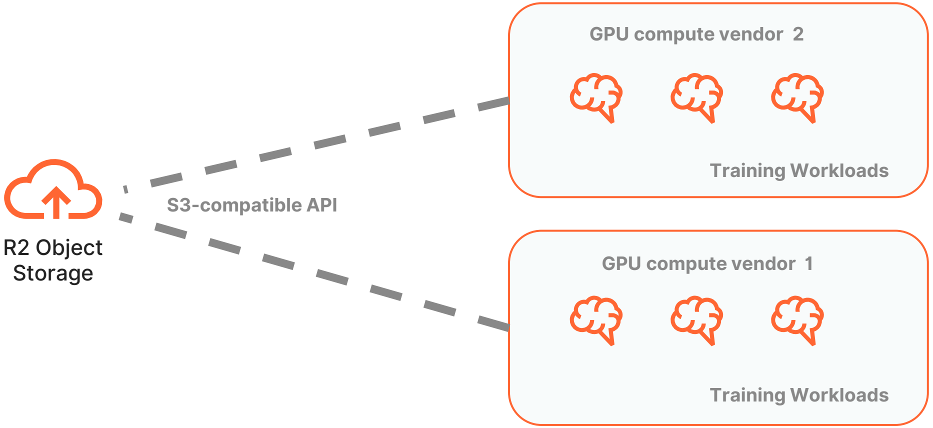
Close to user

Far from user

CLOUDFLARE

Unlock the best prices across cloud providers

Store training data in **R2** to move between clouds with zero egress fees



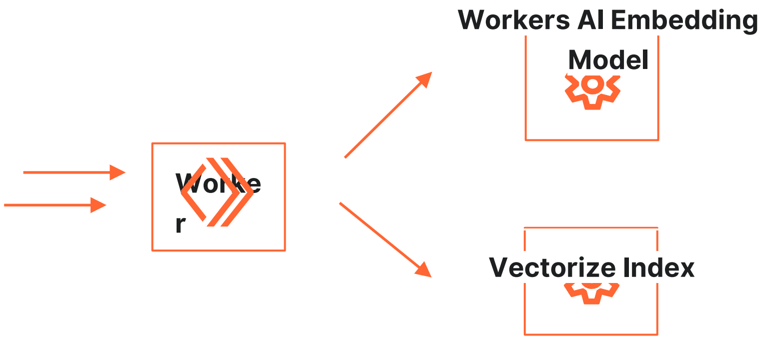
Train

Move training data across clouds without egress fees

Eliminate the need for complex data transfers or duplication of data sets

Partnerships so you can do more with your data

Generate, store, and search vectors



Workers AI Embedding Model

Worker

Vectorize Index

Sample models

Llama-3

whisper

M2M100

DistilBERT-SST-2

ResNet

Infer

Use built-in models from **Workers AI** to generate embeddings

Store embeddings in **Vectorize**

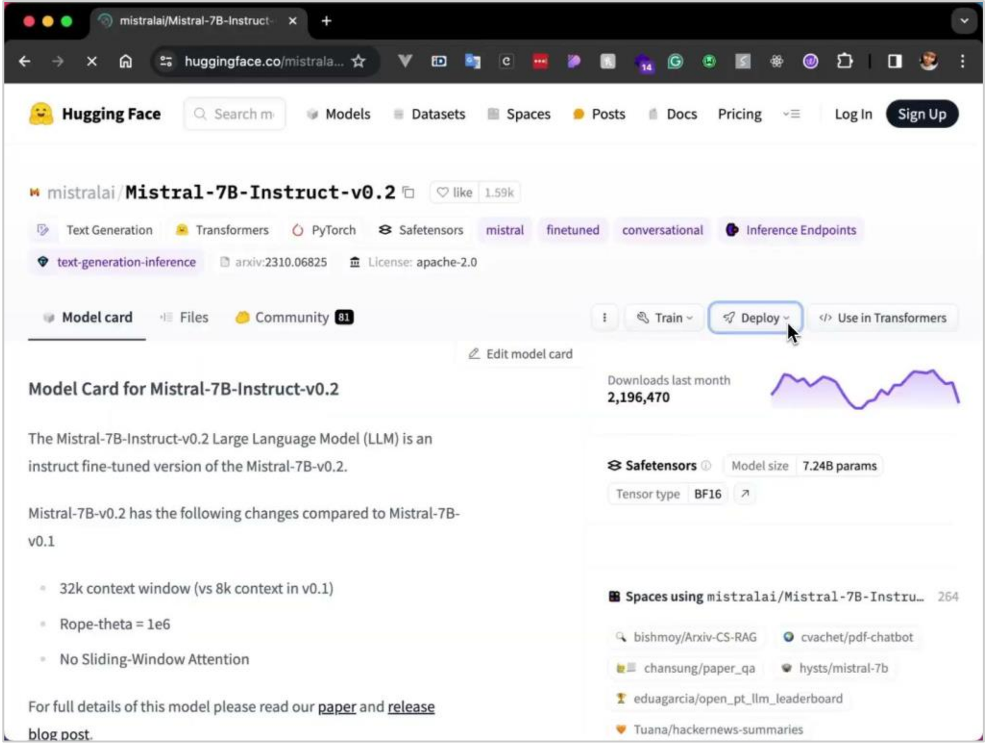
Transform each query into an embedding with **Workers AI**

Query **Vectorize** for similar vectors

Retrieve source content from **R2** or **KV**

Build with a library of off the shelf models and deploy directly from HuggingFace

- Generate text
- Classify text
- Create images from text prompts
- Translate text
- Classify images



Model Card for Mistral-7B-Instruct-v0.2

The Mistral-7B-Instruct-v0.2 Large Language Model (LLM) is an instruct fine-tuned version of the Mistral-7B-v0.2.

Mistral-7B-v0.2 has the following changes compared to Mistral-7B-v0.1

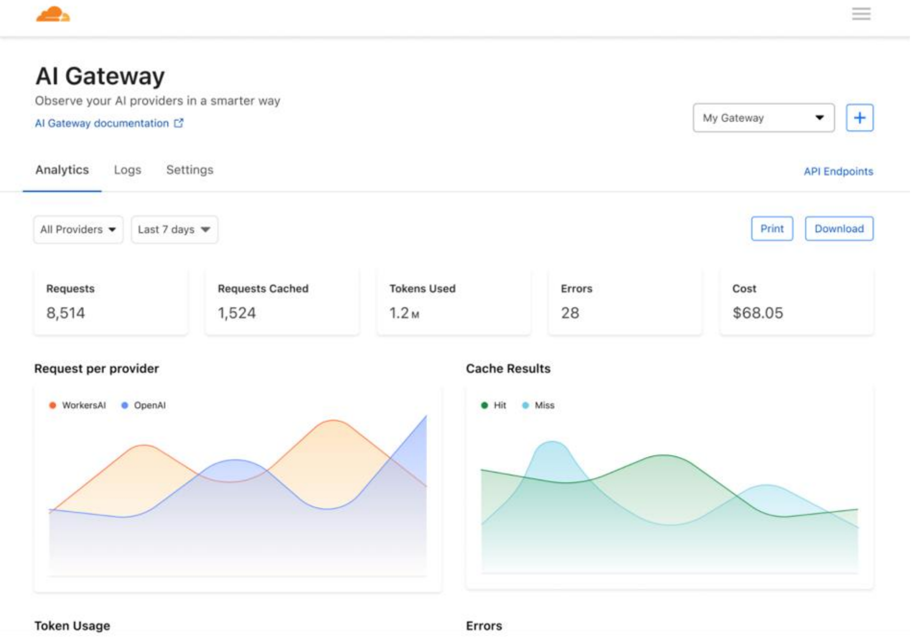
- 32k context window (vs 8k context in v0.1)
- Rope-theta = 1e6
- No Sliding-Window Attention

For full details of this model please read our paper and release blog post.

See our documentation for a full list of all 40+ models

Observe and control AI applications

AI Gateway is a full proxy with caching, rate limiting, request retries, analytics, and tracking of AI usage to mitigate data loss and AI integrity.



AI Gateway

Observe your AI providers in a smarter way

AI Gateway documentation

Requests	Requests Cached	Tokens Used	Errors	Cost
8,514	1,524	1.2 M	28	$68.05

Request per provider

Cache Results

Token Usage

Errors

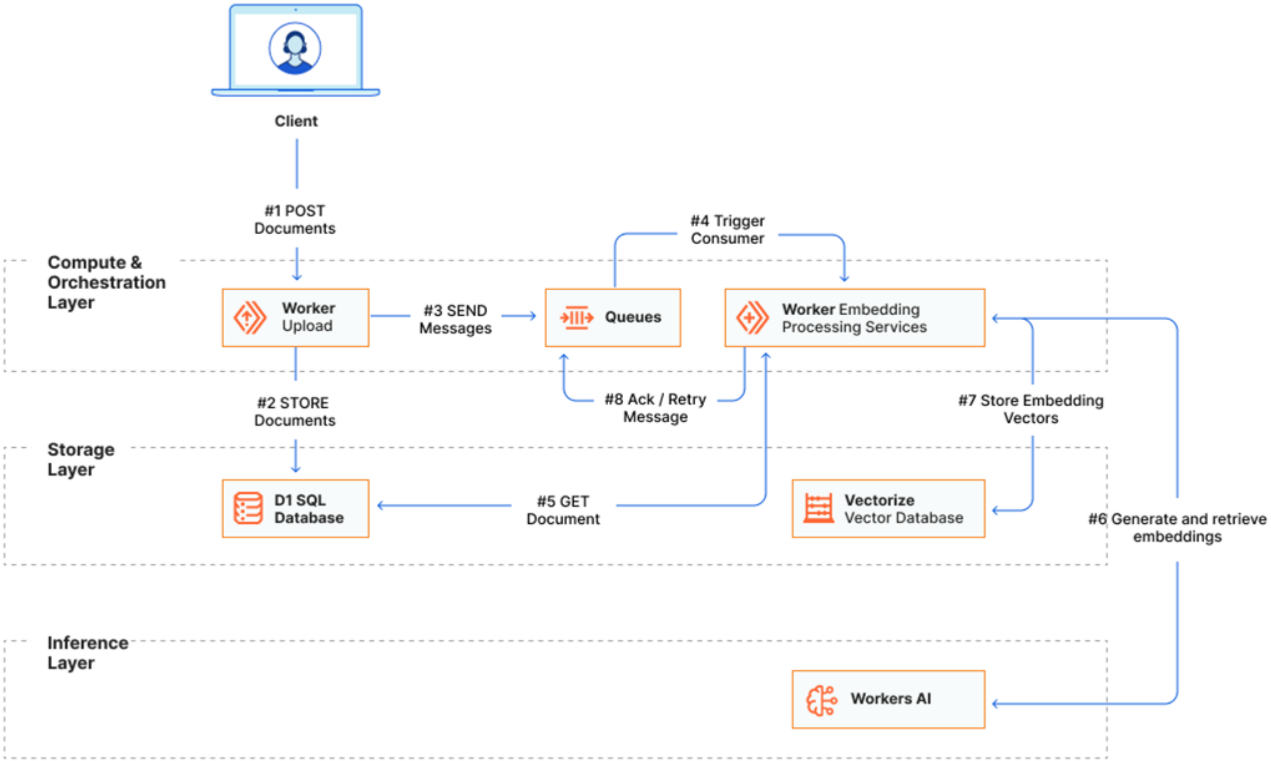
Optimize

Observe traffic patterns, usage, logs, and cost

Provide visibility into traffic & malicious actors

Alleviate the burden of cost & speed with caching and rate-limiting

RAG application architecture



Common RAG use cases

Conversational applications including chatbot, text, and email

Cloudflare products

D1
Queues
Workers
WorkersAI
Vectorize

AI products expand what you can build on Cloudflare

Developer Products

Compute

Workers

Pages

Data Storage

R2

D1

Workers KV

Durable Objects

Developer Services

Images

Stream

Workers Analytics Engine

Queues

AI Products

Workers AI

Vectorize

AI Gateway

Platform

Cloudflare Network Infrastructure

- **Global Edge:** 310+ cities, 95% of Internet users within 50ms, 13,000 interconnects, 248 Tbps capacity, China Network
- **Building Blocks:** SSL/TLS, mTLS, Authoritative/Recursive DNS, DNSSEC, DNS over HTTP, L4-7 over Wireguard
- **Compliance/Privacy:** ISO, SOC, PCI, GDPR-compliant Logs & Analytics, Data Localization Suite

Why Cloudflare AI?

NETWORK
SCALE

EASE
OF USE

NO TRADE
OFFS

Meta

Microsoft

databricks

NVIDIA

Hugging Face

OUR PLATFORM

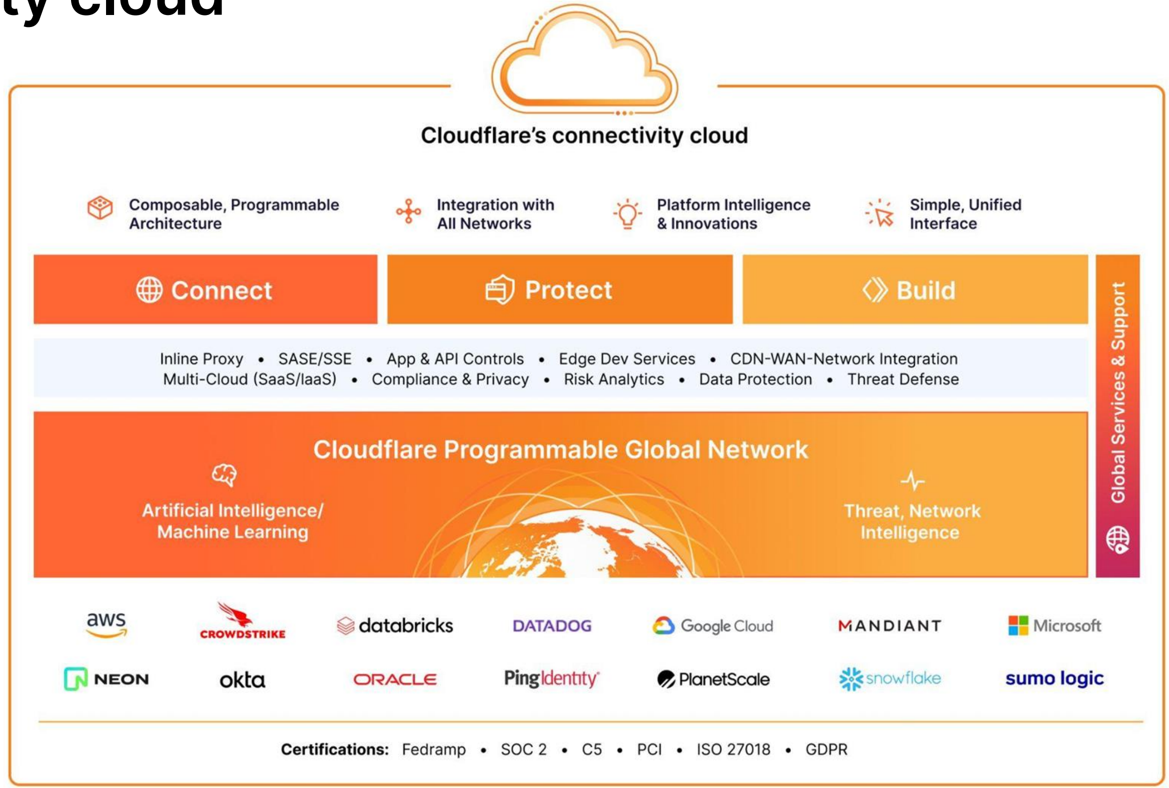
CLOUDFLARE

Cloudflare’s connectivity cloud

With Cloudflare, organizations can:

- **Connect** users, networks, apps and clouds globally
- **Protect** data, apps, infrastructure, and users everywhere
- **Build** innovative digital services and experiences anywhere

With security, speed, programmability and resilience



Cloudflare’s connectivity cloud

Composable, Programmable Architecture

Integration with All Networks

Platform Intelligence & Innovations

Simple, Unified Interface

Connect

Protect

Build

Global Services & Support

Inline Proxy • SASE/SSE • App & API Controls • Edge Dev Services • CDN-WAN-Network Integration

Multi-Cloud (SaaS/IaaS) • Compliance & Privacy • Risk Analytics • Data Protection • Threat Defense

Cloudflare Programmable Global Network

Artificial Intelligence/ Machine Learning

Threat, Network Intelligence

aws • CROWDSTRIKE • databricks • DATADOG • Google Cloud • MANDIANT • Microsoft

NEON • okta • ORACLE • PingIdentity • PlanetScale • snowflake • sumo logic

Certifications: Fedramp • SOC 2 • C5 • PCI • ISO 27018 • GDPR

CLOUDFLARE

> **This makes it incredibly simple for any developer to start building with AI, even with zero machine-learning expertise. Now you can deploy an AI-powered app in minutes.**

- Ben’s Bites