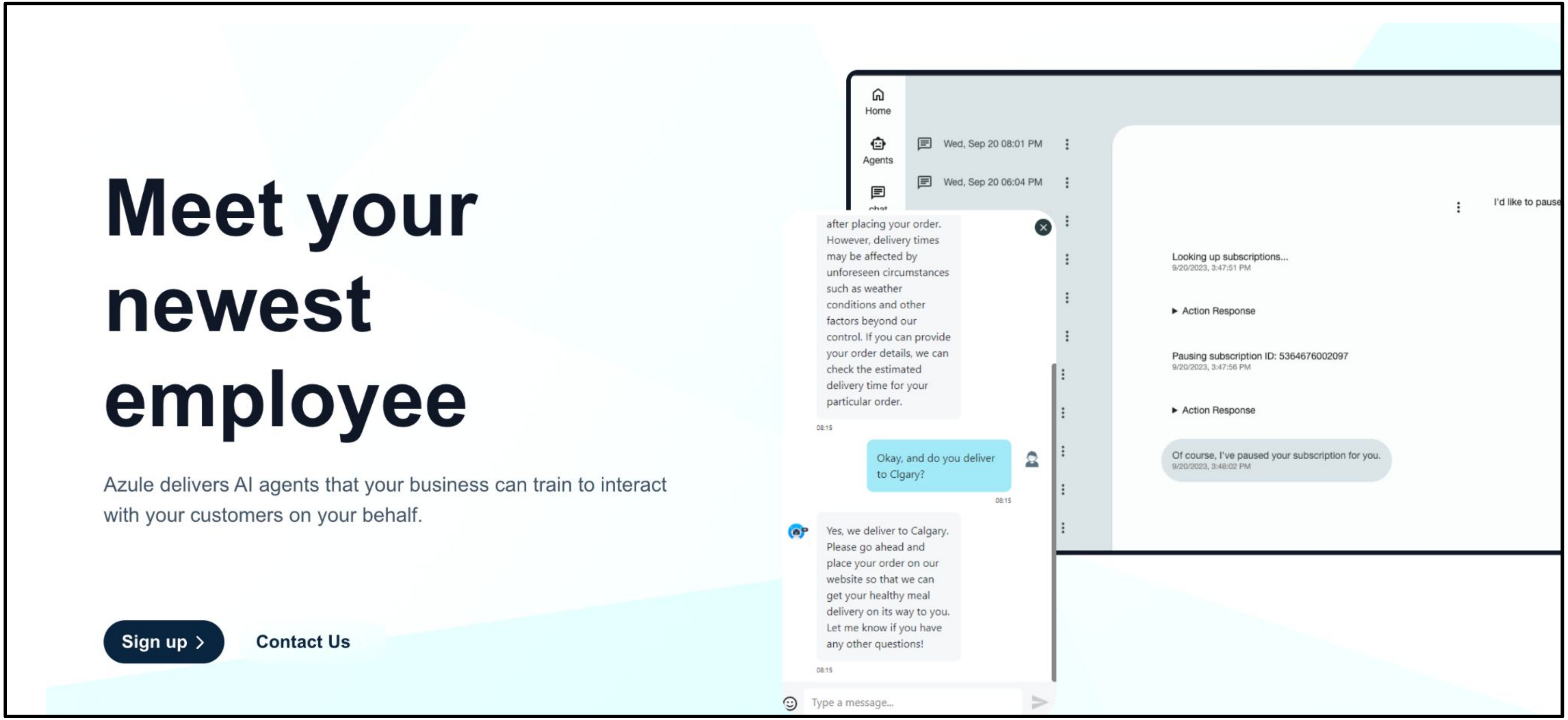
CLOUDFLARE

AI Inference at the Edge

Logan Grasby
Developer Educator, AI at Cloudflare

Nvidia GTC 2024



Meet your newest employee

Azule delivers AI agents that your business can train to interact with your customers on your behalf.

after placing your order. However, delivery times may be affected by unforeseen circumstances such as weather conditions and other factors beyond our control. If you can provide your order details, we can check the estimated delivery time for your particular order.

08:15

Okay, and do you deliver to Clgary?

08:15

Yes, we deliver to Calgary. Please go ahead and place your order on our website so that we can get your healthy meal delivery on its way to you. Let me know if you have any other questions!

08:15

I'd like to pause

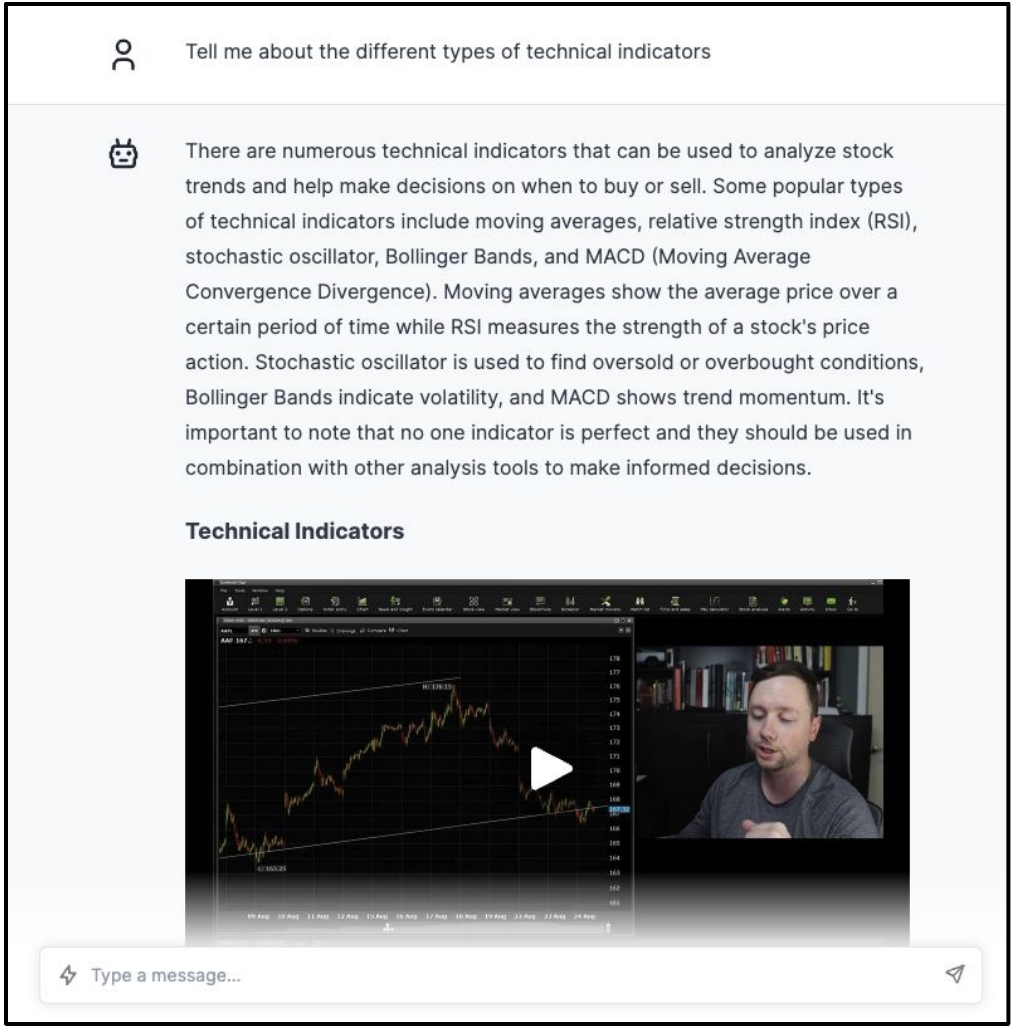
Looking up subscriptions...
9/20/2023, 3:47:51 PM

▸ Action Response

Pausing subscription ID: 5364676002097
9/20/2023, 3:47:56 PM

▸ Action Response

Of course, I've paused your subscription for you.
9/20/2023, 3:48:02 PM



Tell me about the different types of technical indicators

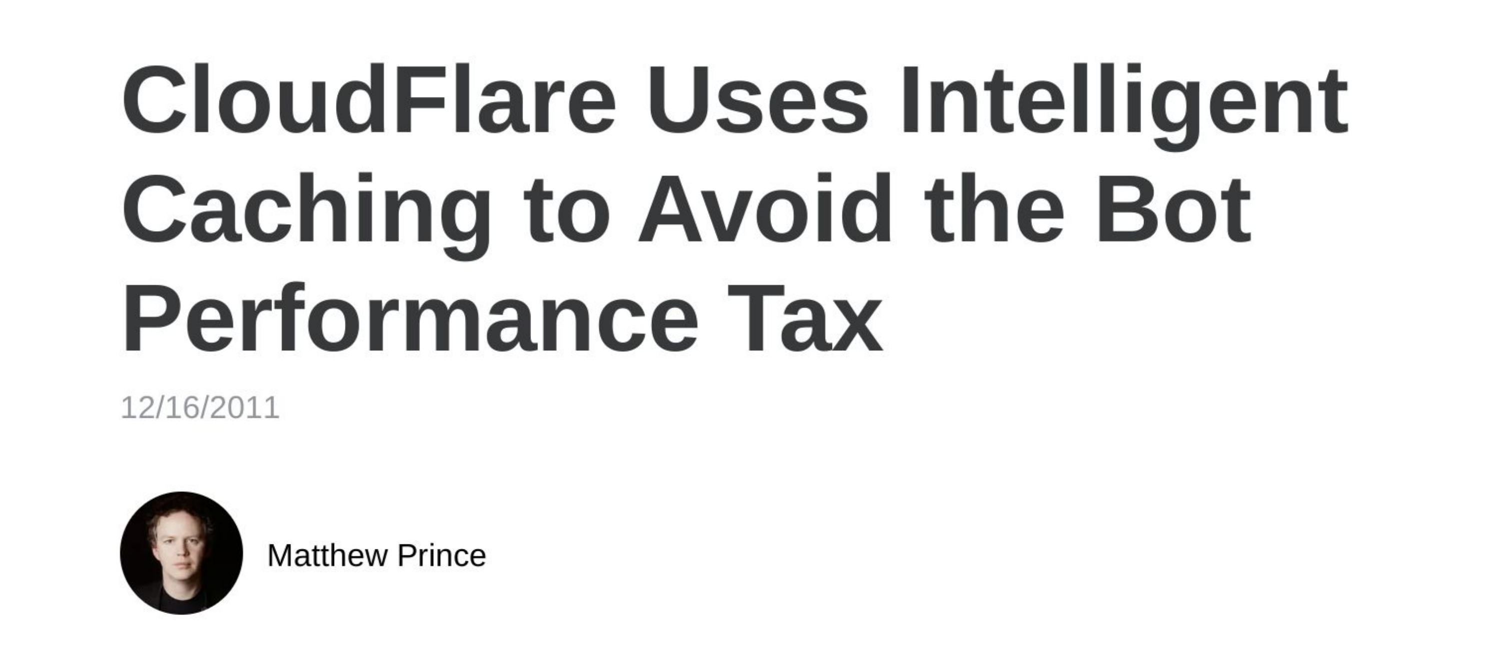
There are numerous technical indicators that can be used to analyze stock trends and help make decisions on when to buy or sell. Some popular types of technical indicators include moving averages, relative strength index (RSI), stochastic oscillator, Bollinger Bands, and MACD (Moving Average Convergence Divergence). Moving averages show the average price over a certain period of time while RSI measures the strength of a stock's price action. Stochastic oscillator is used to find oversold or overbought conditions, Bollinger Bands indicate volatility, and MACD shows trend momentum. It's important to note that no one indicator is perfect and they should be used in combination with other analysis tools to make informed decisions.

Technical Indicators

Today Cloudflare serves on average over 55 million HTTP requests per second

- 20% of internet traffic
- 3 Terabytes of DNS queries per day
- 8 Billion web pages crawled every 2 weeks

A history of AI/ML at Cloudflare



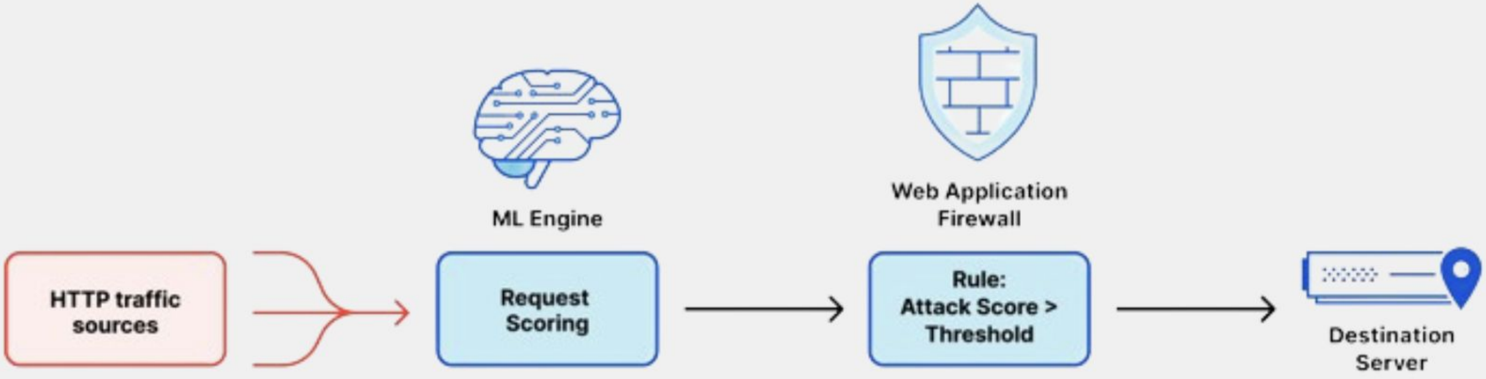
CloudFlare Uses Intelligent Caching to Avoid the Bot Performance Tax

12/16/2011

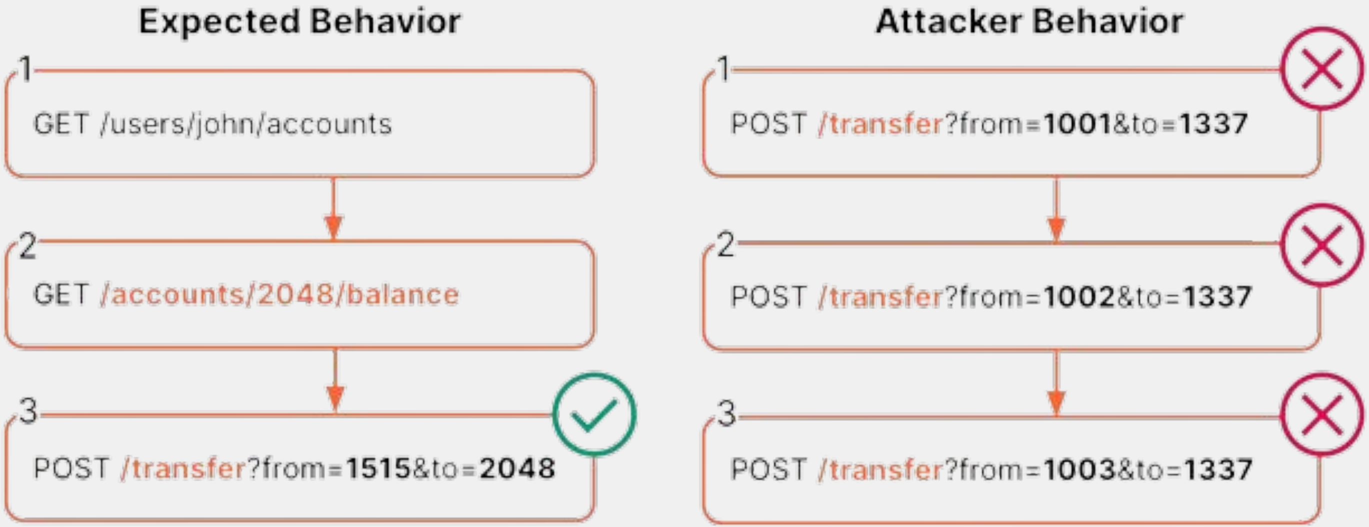
Matthew Prince

A history of AI/ML at Cloudflare

AI/ML models power products like WAF and Bot Management at the edge



AI/ML at Cloudflare today



Expected Behavior

1 GET /users/john/accounts

2 GET /accounts/2048/balance

3 POST /transfer?from=1515&to=2048 ✓

Attacker Behavior

1 POST /transfer?from=1001&to=1337 ✗

2 POST /transfer?from=1002&to=1337 ✗

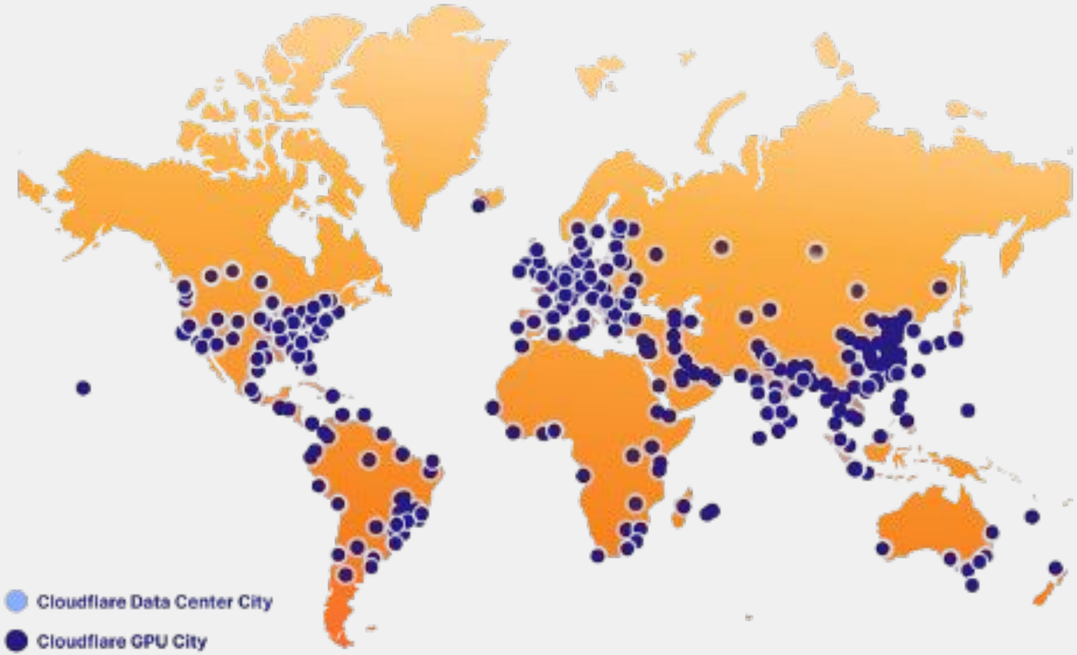
3 POST /transfer?from=1003&to=1337 ✗

Deploying GPUs around the world

- GPUs in 125+ locations globally
- Covering Cloudflare’s entire network
- Global AI inference at the edge

AI Inference at the Edge

Deploying GPUs around the world



End of 2024

Run inference on region: Earth

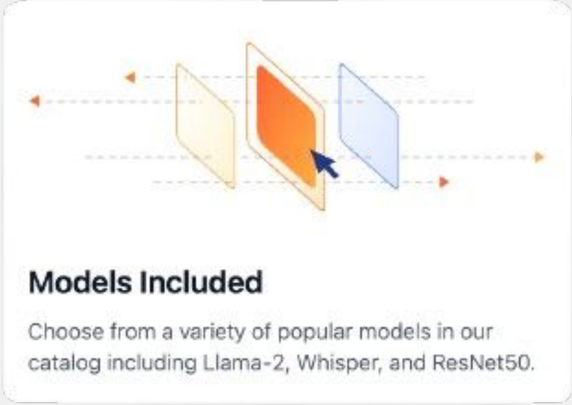
Build and deploy ambitious **AI applications** to Cloudflare's global network

AI Inference at the Edge

AI accessible to everyone

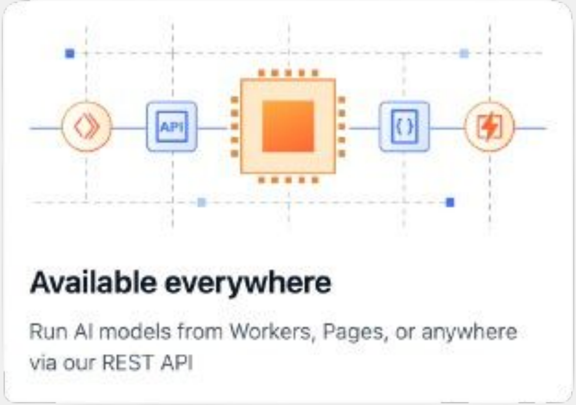
Serverless AI on GPUs

Run generative AI tasks on our global network of NVIDIA GPUs with no extra setup.



Models Included

Choose from a variety of popular models in our catalog including Llama-2, Whisper, and ResNet50.



Available everywhere

Run AI models from Workers, Pages, or anywhere via our REST API

AI accessible to everyone

Supercharge with Vectorize

Generate and store embeddings in a globally distributed vector database.

AI Gateway

Improve reliability and scalability with caching, rate limiting, and analytics.

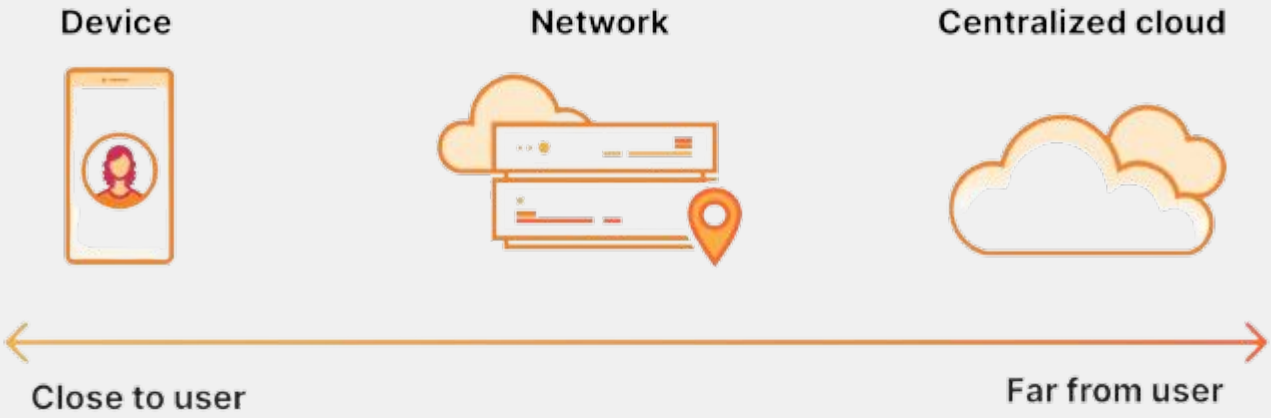
Train with R2

Build multi-cloud training architectures with free egress.

Announced this week:

Gemma 2b and 7b

Why does AI inference belong on the edge?



Lowering inference latency

Smart placement and GPU inference

- Avoid costly round trips to a centralized service

Worker
You
Back-end

CLOUDFLARE

Lowering inference latency

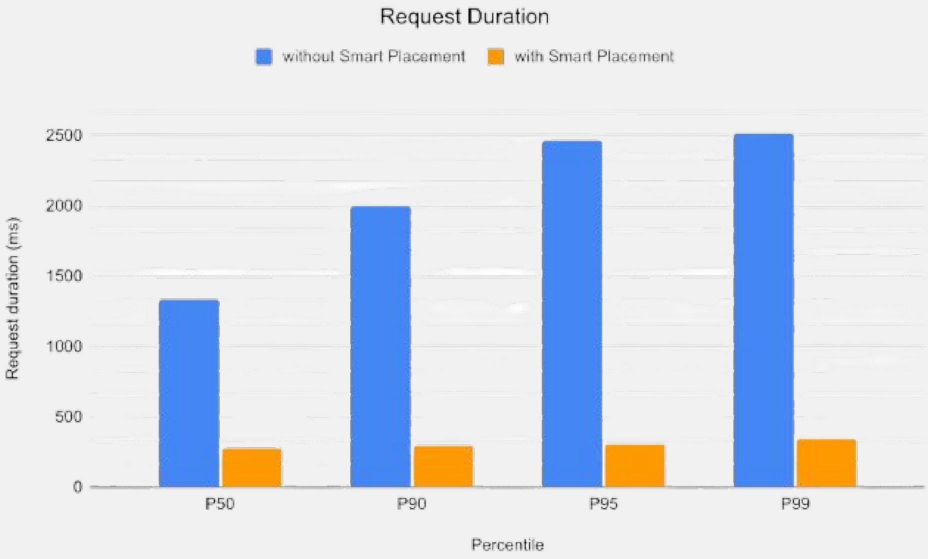
You
Worker
Back-end

Smart placement and GPU inference

- Avoid costly round trips to a centralized service

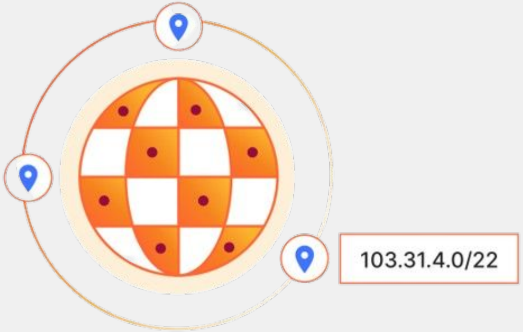
Lowering inference latency

Smart placement and GPU inference



Helping products stay compliant

- GDPR
- EU AI Act
- Fine tuning with customer data



Enhanced protection with AI at the edge

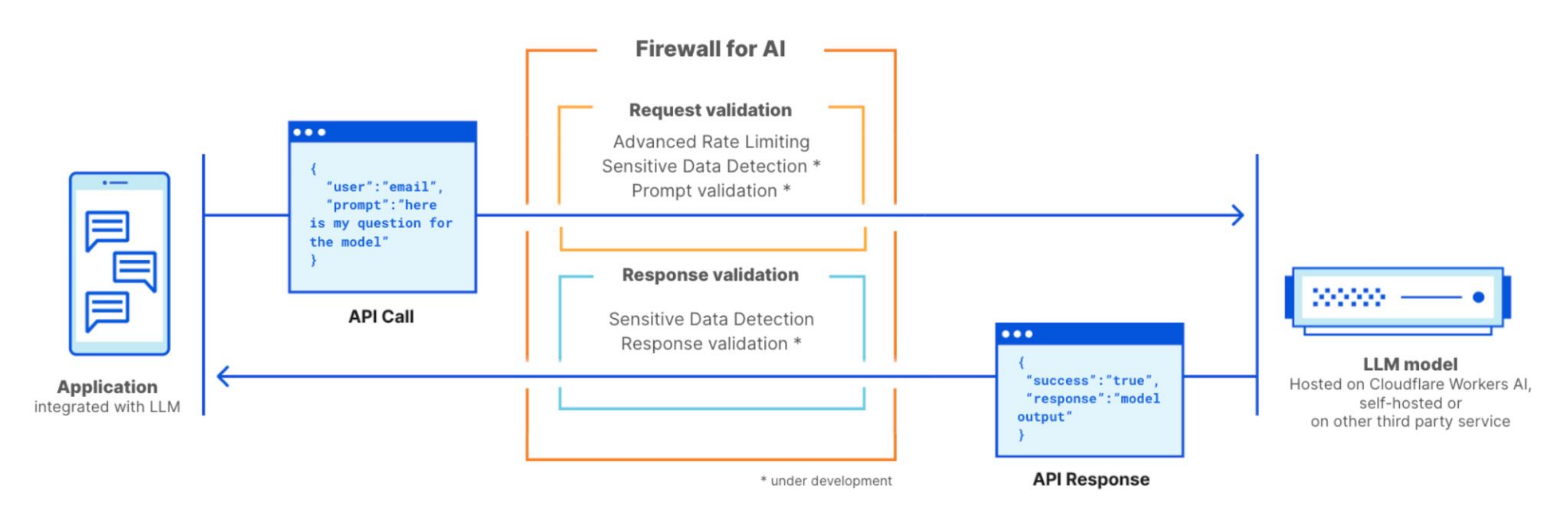
Firewall for AI

LLM Security Concerns

- Prompt Injection
- Insecure Plugin Design
- Sensitive Information Disclosure
- Excessive Agency

CLOUDFLARE

Enhanced protection with AI at the edge



Firewall for AI

Request validation

Advanced Rate Limiting
Sensitive Data Detection *
Prompt validation *

Response validation

Sensitive Data Detection
Response validation *

* under development

```
{
  "user":"email",
  "prompt":"here
is my question for
the model"
}
```

API Call

```
{
 "success":"true",
 "response":"model
output"
}
```

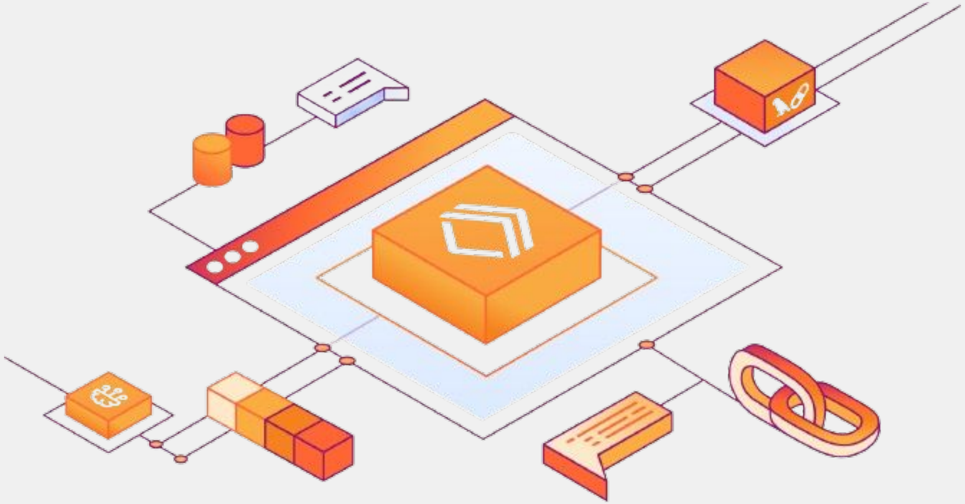
API Response

Application
integrated with LLM

LLM model
Hosted on Cloudflare Workers AI,
self-hosted or
on other third party service

Not just AI, a whole developer platform

Build full stack, instantly scalable applications



Storage Solutions

- Durable Objects
- D1 (SQLite)
- KV
- R2 (Zero egress object storage)

Bring your frontend framework

- Cloudflare Pages

Secure your applications

- API Gateway
- Bot Management

Connect with us!

Logan Grasby

Developer Educator, AI at Cloudflare

𝕏 @LoganGrasby

in linkedin.com/in/logangrasby

At NVIDIA GTC

Booth 1634

Online

meet-us.pages.dev