

THE CONNECTIVE LAYER • CLOUDFLARE DEVELOPER PLATFORM

Five products. One brain, *at the edge.*

Superhuman is now Grammarly, Mail, Docs, Go, GPTZero and, since 14 September, Fathom. **Go's agents need context from all of them, in real time, per user, under tight governance.** Cloudflare's Developer Platform is the connective runtime for exactly that, and no team has to re-platform to start.

40M+

daily active users
SUPERHUMAN.COM

~100B

LLM requests a week, for GEC alone
SUPERHUMAN.ENG.BLOG

100+ / 800

agents and tools Go routes between
SUPERHUMAN.ENG.BLOG • 15 OCT 2026

14 Sep 2026

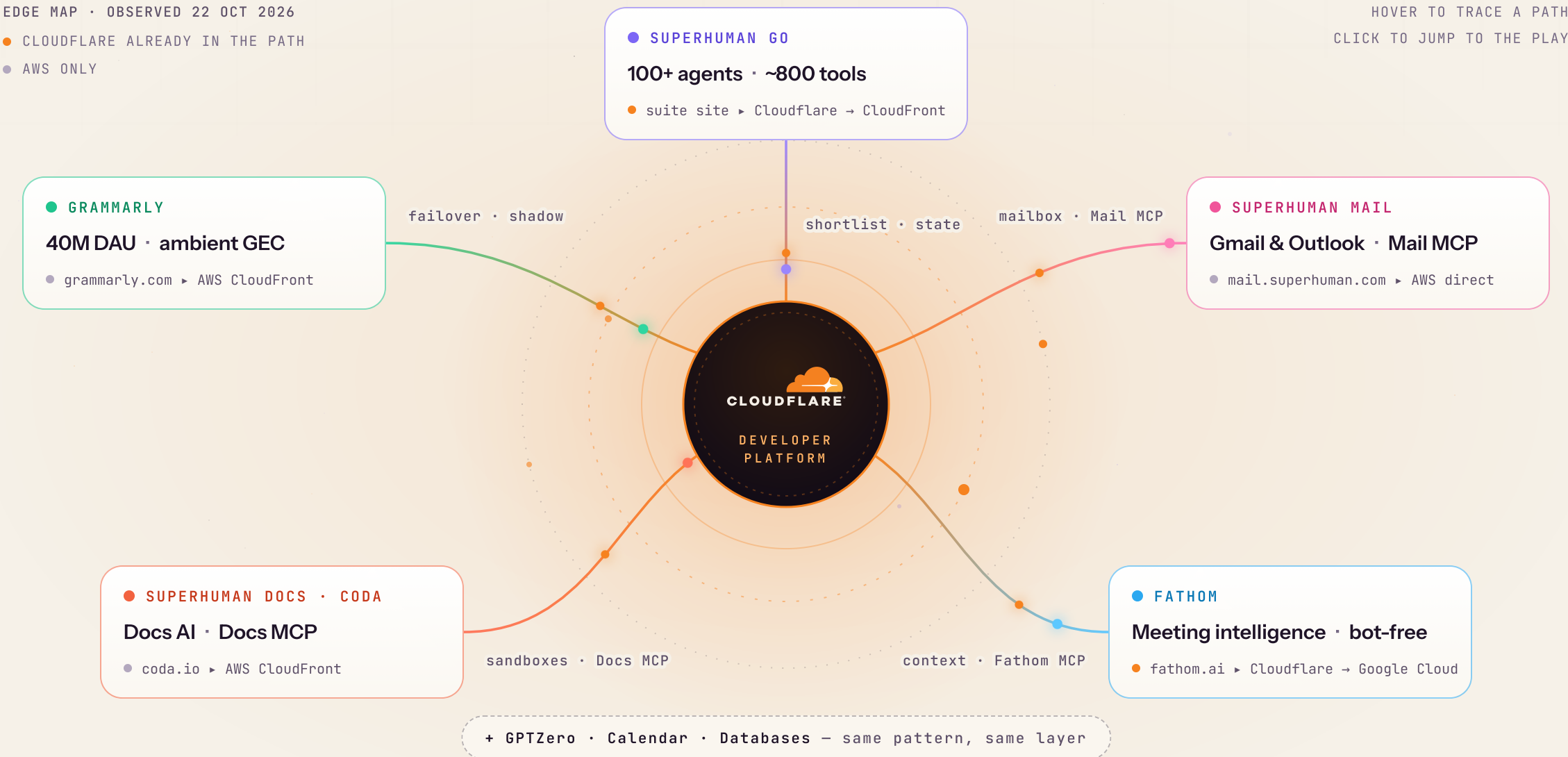
Fathom joins the suite. Meetings become agent context.

01 Unified in brand. *Not yet in stack.*

EDGE MAP • OBSERVED 22 OCT 2026

- CLOUDFLARE ALREADY IN THE PATH
- AWS ONLY

HOVER TO TRACE A PATH
CLICK TO JUMP TO THE PLAY



AGENTS SDK • DURABLE OBJECTS One stateful agent per user, mailbox, doc or meeting.	AI GATEWAY One control plane for every model call.	WORKERS AI • VECTORIZE Embeddings and inference beside the request.	WORKFLOWS • QUEUES Retryable pipelines for the flaky parts.	R2 One data plane across clouds. No egress fees.	MCP PORTALS • WORKERS VPC A governed front door, private paths back.
--	--	---	---	--	--

02 What changes in *each product line*

Each play names the primitive, the signal that justifies it, and the number we would measure.

● Superhuman Go

START HERE

ROUTING AT 800 TOOLS

SIGNAL Your research: routing F1 drops 16–23% from 10 to 110 agents; embedding shortlists beat flat routing by 6–11%. Smaller, locally run embedders are on your list to explore.

- Workers AI embeddings + Vectorize as the tool shortlist, beside the request.
- Agents SDK + Durable Objects give every user session its own stateful agent.
- McpAgent hosts remote MCP servers with OAuth for third-party and customer-built agents.

PROVE IT F1 vs. your flat-routing baseline • p95 shortlist latency

● Fathom

MEETINGS → SHARED CONTEXT

SIGNAL Meeting data as agent context is the deal thesis. Fathom already sits behind Cloudflare, with a Google Cloud origin; Grammarly, Docs and Mail sit on AWS.

- Queues + Workflows run transcribe, summarize and CRM/Slack sync, retried per step.
- R2 for transcripts and audio, readable from AWS and GCP with no egress fees.
- Vectorize for cross-meeting search per tenant; Deepgram models on Workers AI.

PROVE IT Meeting end → context in Go • pipeline recovery rate

● Superhuman Mail

MAILBOX STATE AT USER SCALE

SIGNAL A client over Gmail and Outlook, now MCP-enabled: rate-limited, push-driven, per-account work where one slow mailbox must not touch another.

- A Durable Object per mailbox serializes sync, holds push state and fires snooze and follow-up alarms.
- Queues absorb provider bursts and fan out to Go agents.
- Mail MCP on Workers with OAuth, fronted by an MCP portal.

PROVE IT p95 sync lag • provider 429s absorbed

● Superhuman Docs

AGENTS AS CO-AUTHORS

SIGNAL Docs MCP is live and Go lets customers add their own agents: untrusted code and tools running next to customer data.

- Workers for Platforms gives each tenant an isolated Worker for customer-built agents.
- Sandbox SDK + Containers for heavy tools; Browser Run for agents that act on the web.
- Durable Objects + WebSockets for agent presence inside a live doc.

PROVE IT Agent cold start • blast radius of one tenant

● Grammarly

AMBIENT AI AT 100B A WEEK

SIGNAL An in-house gateway hardened from ~1K to ~100K RPS across EKS + vLLM and Databricks; a silent upstream change once degraded quality mid-test.

- Start off the hot path: AI Gateway on the long tail of new models, experiments and Go's frontier calls.
- Workers as the front door for shadow traffic and A/B routing, the pattern you built by hand.
- R2 as a shared eval and training data plane across AWS and Databricks.

PROVE IT Provider-failover time • cost per 1M long-tail calls

● The whole suite

ONE FRONT DOOR, ONE MEMORY

SIGNAL Mail MCP, Docs MCP and Fathom's Public API & MCP: three endpoints across AWS and GCP.

- MCP server portals put one URL, SSO, tool-level controls and an audit log in front of all three.
- Durable Objects + namespaced Vectorize hold permission-aware memory across mail, docs and meetings.
- AI Search for managed RAG over whatever lands in R2.

PROVE IT One endpoint, one audit trail, tool-level controls

03 Prove it in *90 days*

Three bounded pilots. Each ends in a number your team already knows how to measure.

30 DAYS

Shortlist bake-off

Replay Go's routing eval set against Workers AI embeddings and Vectorize. Same harness, same models, one variable changed.

EXIT ► F1 AT OR ABOVE TODAY, P95 RETRIEVAL LOWER

→

60 DAYS

Stitch the context

Fathom to R2 to a Workflow to a Go context API. One MCP portal in front of Mail, Docs and Fathom.

EXIT ► MEETING TO AGENT-READY CONTEXT, MEASURED END TO END

→

90 DAYS

Gate the models

AI Gateway on Go and experiment traffic: fallback, caching, cost tracking and spend limits. Then decide on the GEC long tail.

EXIT ► COST AND RELIABILITY READ-OUT FOR YOUR STAFF

→

THE ASK

Bring your routing eval set. *We'll bring a Worker.*

A 45-minute working session with the Go platform and Fathom infrastructure leads.

HOW WE KNOW • EDGE MAP

Observed 22 Oct 2026 from public DNS (NS and A records) and HTTP response headers for [superhuman.com](#), [mail.superhuman.com](#), [docs.superhuman.com](#), [coda.io](#), [grammarly.com](#), [fathom.ai](#), [api.fathom.ai](#). Origins are inferred from [via](#) headers and IP ownership. This shows the path, not contract status.

HOW WE KNOW • PRODUCT FACTS

[superhuman.com/about](#) for the suite and usage figures. [blog.superhuman.com](#) for the Fathom acquisition (14 Sep 2026), Go agent orchestration (15 Oct 2026), GEC inference at scale, and Go cross-platform architecture.

LIMITS OF THIS DOCUMENT

Cloudflare capabilities checked against [developers.cloudflare.com](#) in Oct 2026. Architectures shown are proposals, not descriptions of Superhuman's internal systems. No non-public information was used.

CLOUDFLARE × SUPERHUMAN

PREPARED FOR SUPERHUMAN ENGINEERING LEADERSHIP • DRAFT FOR DISCUSSION