OSA CON 2026 · GRAFANA LABS

Who Watches the Watcher

Benchmarking AI Agents That Debug Your Dashboards

ai-sdk
Agent Framework

o11y-bench
Benchmarking Suite

Grafana MCP
Context Server

Agent O11y
Live Observability

Speakers: Wei-Chin Call & Jun Ding | Grafana Labs

OSA Con 2026 · Nov 2

True consolidation with a full-stack agentic platform



MARKET MOMENTS

AI adoption is outpacing the ability to track, trust, and govern it.

Complexity is compounding faster than any team can track it

the way monolith → microservices once did to monitoring.

CodeRabbit; DORA State of DevOps Report

Non-engineers are already shipping production code

faster than teams can pass down the knowledge of how those systems work.

Ramp, Benjamin Levick, 2026

Autonomy is outpacing verification.

Agents are making decisions faster than anyone is checking them.

Gartner / Reuters, June 2025

Why you need both: benchmarks before you ship, and live monitoring after.

MARKET MOMENTS

AI adoption is outpacing the ability to track, trust, and govern it.

Moving Fast, Breaking More

CodeRabbit and DORA both show AI-assisted changes introduce more production issues, making change failure rate increasingly difficult to keep in check as development velocity rises.

(CodeRabbit; DORA State of DevOps Report)

Code First, Explanations Later

At Ramp, 12% of pull requests now come from non-engineers, faster than teams can pass down how those systems actually work.

(Ramp, Benjamin Levick, 2026)

Can Your AI Agents Survive 2027?

Gartner expects over 40% of agentic AI projects to be canceled by 2027, partly due to inadequate risk controls.

(Gartner / Reuters)

Why you need both: benchmarks before you ship, and live monitoring after.

THE RELIABILITY METRIC THAT MATTERS

Pass@N vs Pass^N

Pass@3

Best-of-three.

"Got it right at least once in three tries."

✓ ✗ ✗

Measures raw capability — and luck.

Pass^3

Average across three runs.

Effectively: "got it right all three times."

✓ ✓ ✓

Measures consistency — the thing that matters in production.

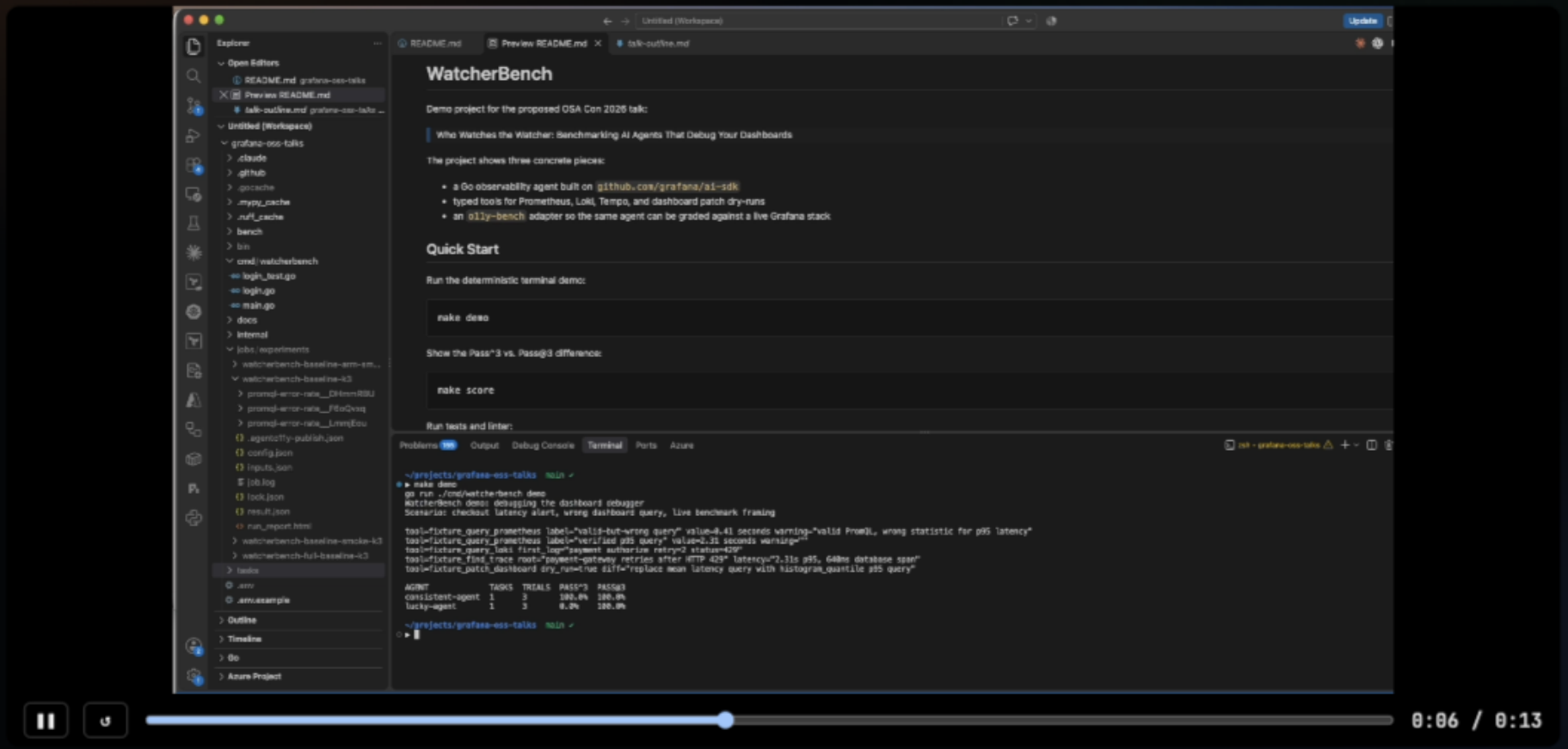
One open-source toolkit. One place to watch it run.

ai-sdk · o11y-bench · Grafana MCP Server · Agent O11y

CAPABILITY	AI-SDK	O11Y-BENCH	GRAFANA MCP	AGENT O11Y
Build a tool-calling agent in Go	✓			
Grades against a live Grafana stack		✓		
Query metrics, logs, traces, dashboards			✓	
Compare runs: Pass@3 vs Pass^3		✓		✓
Watch production cost, evals, and guards				✓
100% open source	✓	✓	✓	

DEMO 1

Build + verify



ai-sdk, o11y-bench, Grafana MCP Server

Reuse o11y-bench's rubrics, send only scores: completely free

Experiments view compares runs side by side in Grafana Cloud, with **Pass@3** / **Pass^3** built into Advanced Metrics

No new evals: this demo reuses o11y-bench's own grading rubrics

COMBO	SENDS CONVERSATIONS?	SCORED BY	COST
1	Yes	Cloud evaluators	Billable
2	Yes	Local, free logic	Billable (export)
3	**No, scores only**	Local, o11y-bench rubrics	**Free***

* Telemetry charges apply

ECOSYSTEM MILESTONE · GRAFANA LABS

o11y-bench Joins Fireworks SII & Hits v1.1

Standardized real-world evaluation for observability models and autonomous SRE agents.

INDUSTRY INDEX

Fireworks Specialized Intelligence

- **Featured SRE Benchmark:** Selected in Fireworks.ai's index for autonomous SRE.
- **Domain-Specific Focus:** Tests complex telemetry reasoning and multi-hop root-cause analysis.
- **Ecosystem Validation:** Grades real-world model reliability across live observability stacks.

V1.1 RELEASE

Expanded Scenarios & Dataset

- **Broader SRE Coverage:** Includes distributed traces, metric spikes, and log triage.
- **Live Sandbox Evals:** Tests agents directly against active Grafana environments.
- **Pass@k & Pass^k:** Evaluates single-run capability and multi-attempt consistency.

EVALUATION IMPACT

Model Performance & Insights

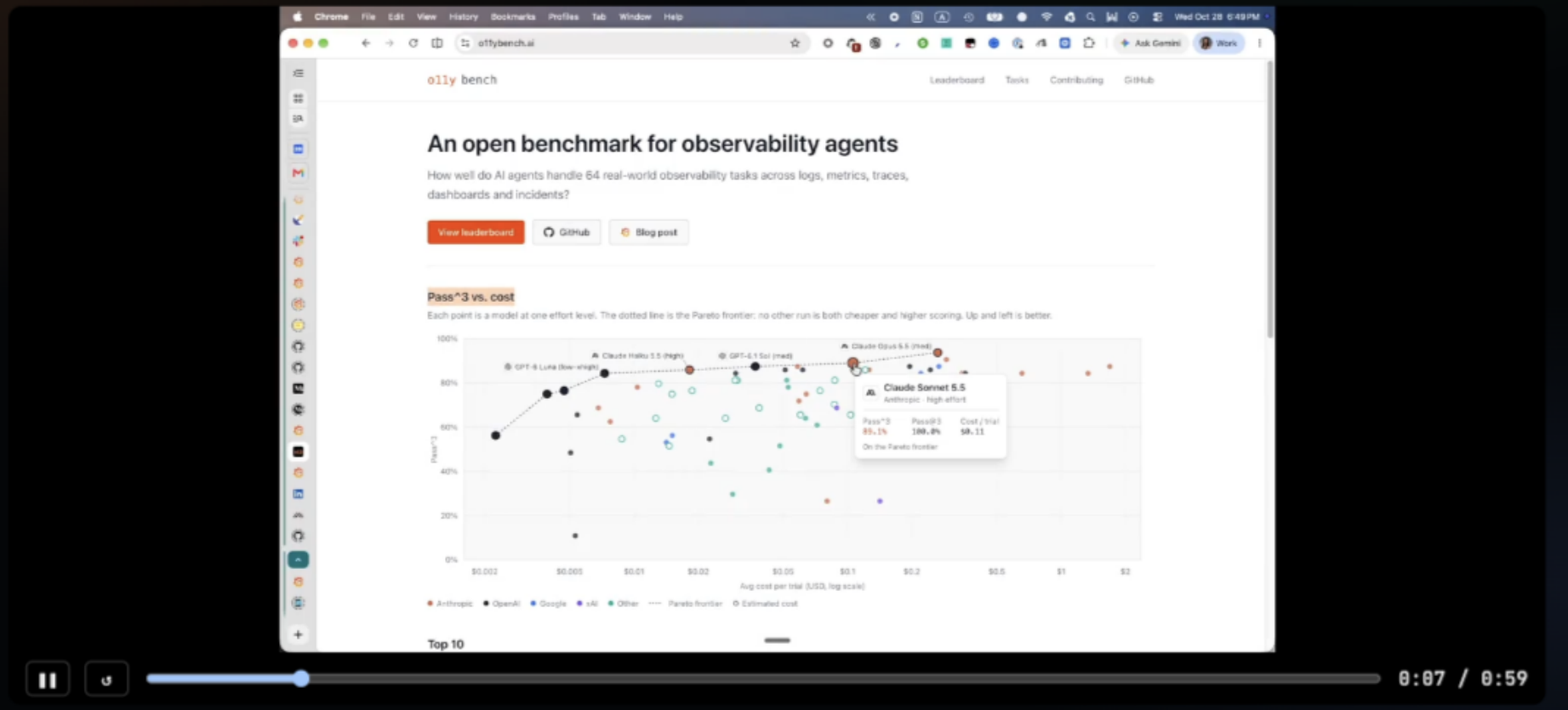
- **Reasoning vs Cost:** Shows trade-offs between frontier models and tuned open LLMs.
- **Tool-Calling Precision:** Exposes hallucinated queries vs accurate telemetry navigation.
- **Continuous Guardrails:** Allows teams to benchmark and guard agents prior to rollout.

PRIMARY REFERENCES & RESOURCES

- **Grafana Announcement:** grafana.com/blog/evaluate-ai-models-for-observability-o11y-bench...
- **Fireworks.ai SRE Index:** fireworks.ai/specialized-intelligence-index/... | Socials: linkd.in/...

DEMO 2

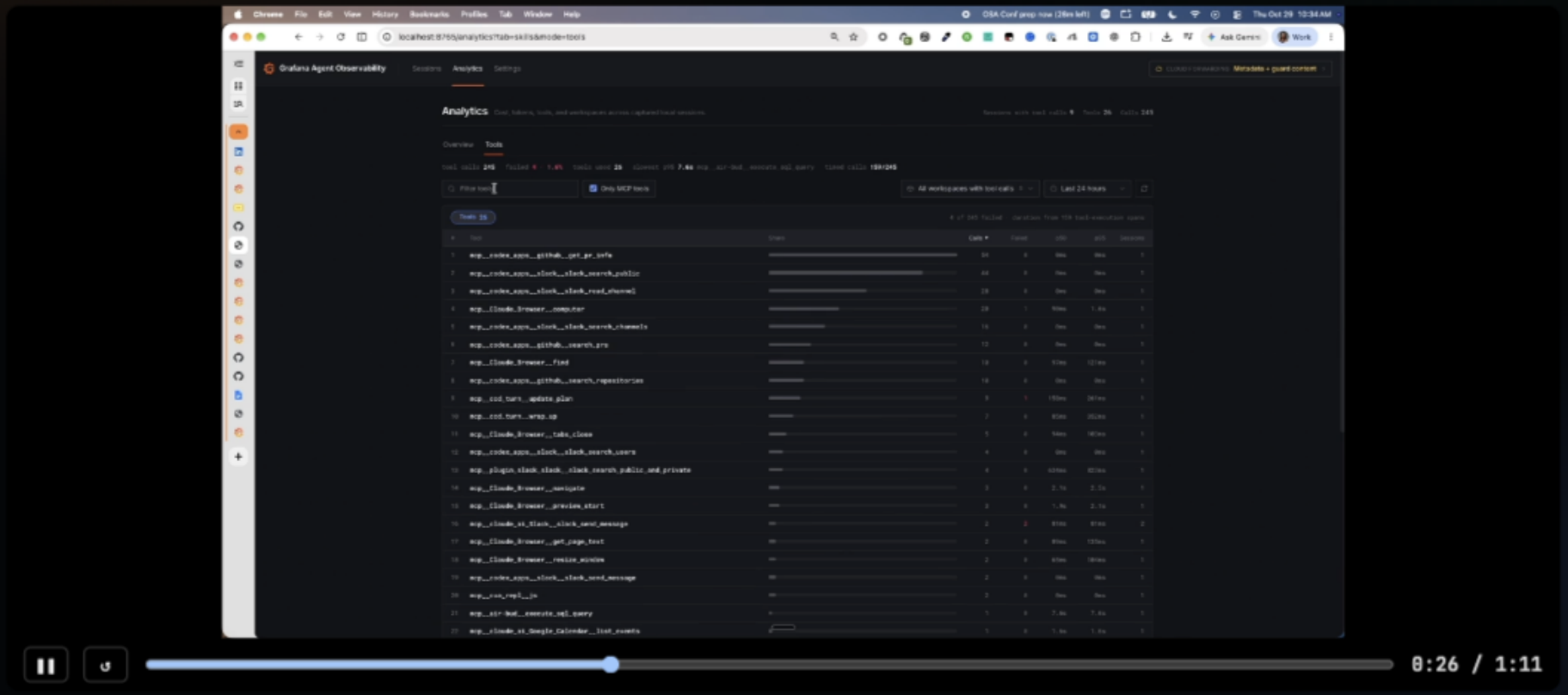
o11ybench.ai



0:07 / 0:59

DEMO 3

Local Coding Agent Observability with agento11y kiro --local



AI Gateway

in progress

One endpoint for every model provider — and the one place where keys, policy and observability get wired up.

INTERNAL CALLERS

- Grafana Assistant
- K6 Agentic Testing
- Investigations
- Any internal service

```
aisdk.GenerateText(ctx, model, …)
// same API, model ID
```

AI Gateway

gateway/providerwire

MIDDLEWARE — THE VALUE

agentobservability	Every generation recorded; preflight policy hooks
enrichment	Approved tenant, region and request metadata
prometheus	Call volume, latency, stream timing, token usage
fallback + registry	Backup models after eligible failures
gateway/catalog	Stable public names: default · fast · balanced

AUTH — INTERNAL
The gateway holds the provider API keys. Callers never see them.

AUTH — EXTERNAL
You pass provider auth; we forward it. We never hold your keys.

PROVIDERS

- Anthropic
- Amazon Bedrock
- OpenAI
- OpenAI-compatible
- Open models · vLLM

Internally

Products ship LLM features without the wiring — no key management, no per-team tracking, no provider drift.

Externally, with Agent Observability

Bring your own provider auth, proxy through the gateway, and get instrumentation, cost attribution and traces in Grafana Cloud for free.

Build it. Then prove it.

- A demo proves your agent CAN work once. A benchmark (experiment) proves whether it works EVERY time.
- Watch for silent failures: syntactically valid, semantically wrong. They don't crash, they mislead.
- Pass@3 is easy to achieve. Pass^3 is your target. Ship against the second one.
- Reliable agents require both benchmarks/experiments and live production monitoring.

github.com/grafana/ai-sdk	github.com/grafana/o11y-bench	o11ybench.ai leaderboard + task explorer
github.com/grafana/agento11y	github.com/grafana/benchmark-experiments-example demo for this talk	https://play.grafana.org/a/grafana-assistant-app/pages/who-watches-the-watcher-osa-con-2026-3262f01eef54 presentation for this talk

Questions?

Get started with Grafana Cloud FREE

Free Forever Plan

$0 / Always free • No payment ever

All Grafana Cloud Features • Usage capped to Free tier limits • Community Support

What you get in our Free Tier:

Metrics: 10k Billable Series, 14 days retention
Logs, Traces & Profiles: 50 GBs each, 14 days retention
Visualization & IRM: 3 Active users each
k6 Performance Testing: 500 Virtual User Hours, 14 days retention
Frontend Observability: 250k Sessions

Grafana Assistant: 3 active AI users (40M tokens / user / mo) & 25M tokens per month for service account usage
Grafana Investigations: Automatically triggered investigations (alerts, incidents) draw from the system-initiated token pool (25M tokens per month). Investigations you run yourself count toward per-user Assistant allowances (40M tokens / user / mo)
Agent Observability: Limited to 30k generations per month & 25M tokens per month for system-initiated usage (Evals & Guards). Telemetry usage counts toward Grafana Cloud free tier allowances

Create your free account

Scan QR code to sign up

Instant access • No credit card required