THE EXECUTION REPLAY ENGINE FOR AI AGENTS

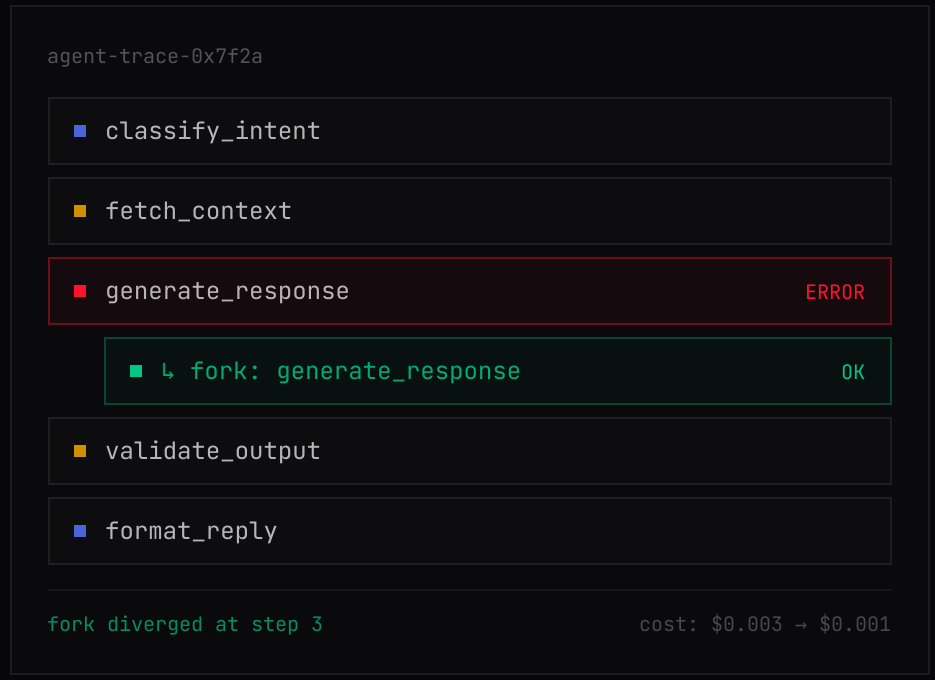
Fork any decision. Replay the future.

One decorator records every LLM call, tool invocation, and error your agent makes. When it fails, fork from any step and watch a new path diverge — with runtime guardrails that halt agents before they cascade.

"Git branching + time-travel debugger + CI gate — for AI agents."

■ PRE-SEED · RAISING $500K

SDKS SHIPPED · PYTHON & TYPESCRIPT



AI agents fail silently — three steps before the error surfaces.

Production agents make non-deterministic decisions across dozens of LLM calls and tool invocations. When one goes wrong, the bug hides behind a wall of JSON spans and the only way to "debug" it is to re-run the whole pipeline and hope.

TODAY'S REALITY

You can't reproduce a failure

Agents are stochastic. Re-running won't hit the same code path. Engineers spend hours guessing at root causes from logs that don't capture state.

avg debug time: 2–6 hrs/incident

COST

Hallucinations cascade in real time

A single bad LLM output poisons every downstream step. By the time it's caught, your agent has spent $10s, made calls it shouldn't have, and shipped a wrong answer.

3–5× cost amplification per failure

EXISTING TOOLS

Dashboards, not debuggers

LangSmith, Helicone, Langfuse show you *what* happened. None let you intervene at the failing step and replay the rest of the agent from that point forward.

$0 → debugging investment from incumbents

"Our agent failed last Tuesday in prod. We still don't know which prompt caused it. We just shipped a guard for that specific symptom."

— every AI engineer, today.

Record once. Fork forever.

Retrace is the first execution replay engine for AI agents. We capture the full causal graph of a trace, let you mutate any decision in it, and cascade-replay the entire agent from that point forward — like Git, but for non-deterministic systems.

01 · RECORD

One decorator. Every call.

Auto-instruments OpenAI, Anthropic, Gemini. Works with any framework — LangChain, CrewAI, Vercel AI SDK. 2 minutes to first trace.

02 · FORK

Branch from any span.

Pick the failing step, edit the input, and the SDK *re-executes the entire function* from that point. Not a single LLM call — the whole agent.

03 · GUARD

Halt before damage cascades.

Runtime policies (cost, loops, context, latency) send a HALT command over WebSocket the moment a threshold trips. Adaptive thresholds via LinUCB.

04 · GATE

CI/CD quality gates.

One CLI call (`retrace eval gate`) returns pass/fail with a threshold. Block prod deploys when agent quality regresses against a baseline.

```python
# Python — works with any agent
import retrace

@retrace.record(name="my-agent", resumable=True)
def run_agent(prompt):
    plan   = llm.plan(prompt)
    facts  = tools.search(plan.query)
    answer = llm.synth(plan, facts)
    return answer

run_agent("What is quantum computing?")
# every span, cost, error, output captured.
# fork & cascade replay from the dashboard.
```

PYTHON SDK | TYPESCRIPT SDK | RUST CLI | GITHUB ACTION | MCP SERVER

Beyond observability.

Every layer below is shipping today. The wedge is fork & cascade replay — but the platform around it is what makes us defensible.

FLAGSHIP

Fork & Cascade Replay

Select any span. Modify the input. The SDK re-executes the entire function from that decision, with context flowing into every downstream LLM call. Side-by-side diff with cost and latency deltas.

Information theory · Causal graph · Sub-100ms span retrieval

RUNTIME

Guardrails

5 policy types. Cost / loop / context / latency / hallucination — all enforced via real-time HALT.

ANALYSIS

Hallucination Detection

KL divergence + entropy on every output. Tiered cheap-to-deep.

GRAPH

Critical Path

Betweenness centrality on the trace DAG finds the bottleneck span.

CI/CD

Eval Gates

Regression detection in pull requests. Pass/fail in pipeline.

DATA

RLHF / DPO Export

Fork outcomes become training pairs. Your debugging compounds.

SEARCH

Semantic Search

pgvector HNSW + halfvec. Describe a bug in English, find every matching

DRIFT

Drift Detection

Recent vs baseline statistical comparison. Alerts before users notice.

SESSIONS

Multi-Agent Graphs

Vector-clock causal ordering across agent boundaries.

SHARE

Tapes

Publish any trace as a shareable URL. Built-in replay player.

The market is forming right now.

AI observability is being recreated from scratch because legacy APM tools (Datadog, NewRelic) can't reason about non-deterministic systems. Whoever owns the agent debugger wins this category.

TAM	$54B	Global AI infrastructure + observability spend by 2030 (Grand View, IDC).
SAM	$11B	LLMOps & AI-agent tooling — the slice we directly compete in.
SOM	$420M	~70K AI eng teams worldwide × $500/mo average reachable in 3 yrs.

TAILWINDS

- **Agents went production in 2025.**
  Claude 4, GPT-5, tool-use everywhere — every Series-B company now ships an agent.
- **Non-determinism breaks every existing tool.**
  $1.4B/yr spent on APM that can't trace LLM decisions. The category is being reset.
- **Regulation is coming.**
  EU AI Act + US Executive Orders mandate audit trails for AI decisions. Compliance is becoming required infra.

Bottom-up SaaS. Land with devs. Expand with platform.

Self-serve subscription, usage-aware caps, Enterprise for governance. Free tier is generous on purpose — single-decorator install, instant value, viral inside engineering teams.

FREE

$0 /mo

Land the developer.

+ 5,000 traces/mo
+ 10 tapes · 7d retention
+ Fork & replay
+ Guardrails (2 policies)

PRO · PRIMARY WEDGE

$29 /mo

For solo devs shipping AI.

+ Unlimited traces & tapes
+ 90d retention
+ Semantic search
+ Eval gates · cost analytics

TEAMS

$99 /mo

Per-seat, expansion lever.

+ 10 seats · fork sweeps
+ Audit logs
+ 50 guardrail policies
+ Collab + sharing

ENTERPRISE

Custom

Governance & compliance.

+ AI governance · LTL verification
+ Adaptive guardrails (LinUCB)
+ SOC 2 · SLA · SSO
+ Dedicated support

TARGET CAC

< $120

TARGET LTV

~$2,400

LTV:CAC

20×

GROSS MARGIN

78–84%

Product is live. Now we scale.

We are launch-stage — the entire platform is shipped end-to-end. The pre-seed round funds GTM and the first developer-relations hire to convert that surface into paying customers.

SHIPPED

28

Database tables, pgvector HNSW indexes, BullMQ queues.

SDKS LIVE

v0.2.2

Python & TypeScript published to PyPI & npm.

CLI BINARIES

5 platforms

Rust CLI for macOS, Linux x86/ARM, Windows.

BUILD STATUS

100%

CI/CD live, monitored, prod on DigitalOcean + Cloudflare.

ROADMAP

Q4 '25 · DONE

v0.1 — Record & Replay

SDKs, API, dashboard, fork engine.

Q1 '26 · DONE

v0.2 — Platform

Guardrails, evals, governance, MCP, billing.

Q2 '26 · NOW

Launch + 25 design partners

Public launch, ProductHunt, content, Discord.

Q3 '26

$10K MRR

200 paying teams. Enterprise pilot #1.

Q4 '26

Seed-ready

$30K MRR, > 50% MoM growth, ready to raise.

They log. We replay.

Every existing LLMOps tool is a dashboard layered on top of structured logs. None of them let you go back in time, fork a decision, and re-run the agent. That's our wedge — and the moat compounds with every recorded trace.

CAPABILITY	LANGSMITH	LANGFUSE	HELICONE	DATADOG LLM	RETRACE
Trace recording	+ Yes	+ Yes	+ Yes	+ Yes	+ Yes
Fork & cascade replay	– No	– No	– No	– No	+ Yes
Runtime guardrails (HALT)	– No	– No	– No	– No	+ Yes
Hallucination detection (KL/entropy)	– No	– No	– No	– No	+ Yes
Critical-path graph analysis	– No	– No	– No	– No	+ Yes
Framework-agnostic SDK	– LangChain-locked	+ Yes	+ Proxy only	+ Agent	+ Yes
CI eval gate	+ Yes	+ Yes	– No	– No	+ Yes
RLHF / DPO export	– No	– No	– No	– No	+ Yes

DEFENSIBILITY

REPLAY ENGINE

DISTRIBUTION

Three curves are crossing.

Agents went production in 2025; the tools to debug them did not. The next two years are when the category gets reset — and we want to be the default record button when that happens.

01

The agent inflection point

OpenAI Operator, Claude Computer Use, Devin, Cursor agents. In 12 months agents moved from demo to ARR. Every Series-B+ company is shipping one. None can debug them.

02

Legacy APM is wrong abstraction

Datadog/NewRelic are designed for deterministic services. LLM calls don't have stack traces. The category is being recreated from scratch — and incumbents are five years late to build replay.

03

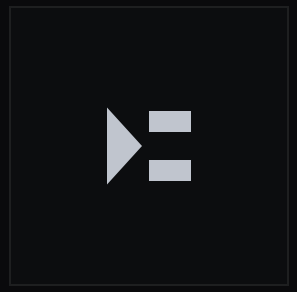
AI-native compliance is mandatory

EU AI Act (Aug 2026 enforcement) requires traceable AI decisions. SOC 2 auditors now ask for LLM logs. Audit trails are becoming required infrastructure — and we ship them by default.

The window is narrow. Whoever owns the agent debugger inherits the category. We are early enough to define the primitives — fork, cascade, gate — and late enough that the buyer (the AI engineer) already exists in volume.

Built solo. Shipped fast.

Retrace is a one-person product as of today — every line of the platform, the SDKs, the CLI, the dashboard, and the docs. The pre-seed funds the first two hires.



Yash Bogam

FOUNDER & ENGINEER

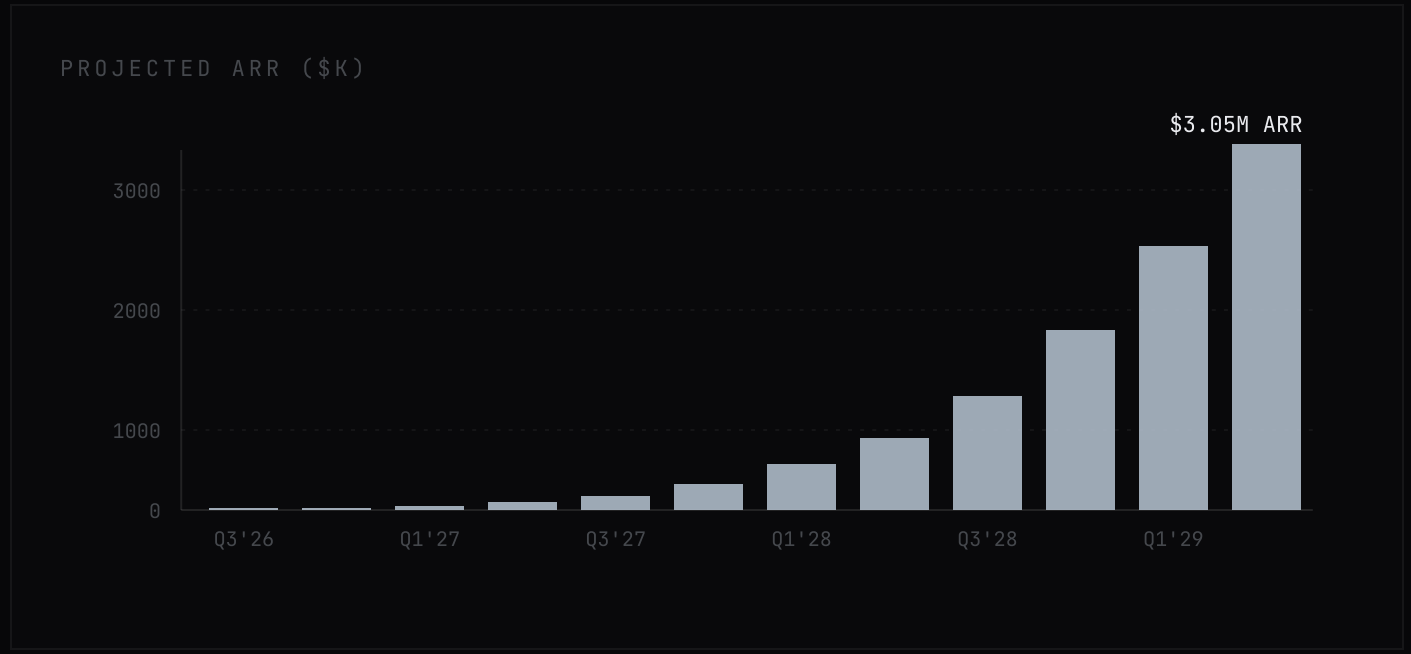
Designed and shipped the full Retrace platform end-to-end — Fastify + Drizzle API, Next.js 16 frontend, Python & TypeScript SDKs, Rust CLI, GitHub Action, MCP server, and the production deployment on DigitalOcean + Cloudflare. Background in distributed systems and ML tooling, with a focus on causal-graph analysis and pgvector-backed semantic search. Operates the entire stack — from kernel-level container limits up to OKLCH-tuned UI tokens.

- 28 TABLES · PGVECTOR HNSW
- V0.2.2 SDKS IN PROD
- RUST CLI · 5 PLATFORMS
- FORK + CASCADE REPLAY ENGINE
- RUNTIME GUARDRAILS + LTL SAFETY

FIRST HIRES THIS ROUND

Path to $3M ARR in 36 months.

Conservative bottom-up model: free-tier conversion holds at developer-tool benchmarks (~3%); per-team ACV grows as we move teams from Pro into Teams & Enterprise.



EOY 2026

$60K ARR

~200 paying teams. Seed-ready.

EOY 2027

$600K ARR

~1.5K teams + 3 Enterprise.

EOY 2028

$3.0M ARR

~8K teams + 12 Enterprise.

GROSS MARGIN

78–84%

Stable across the plan.

Full data room: cohort model, conversion benchmarks, infra cost curve, scenario analysis — available in due diligence.

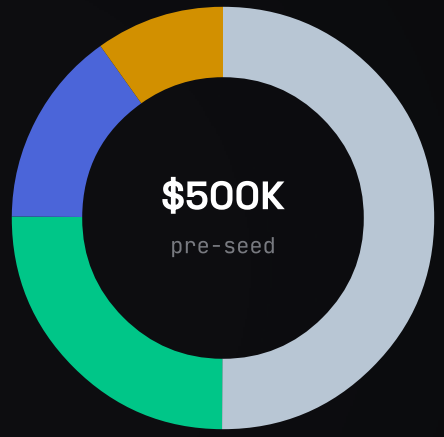
RAISING

$500K Pre-seed · SAFE

18 months of runway to reach **$30K MRR**, hire a founding engineer and a developer-relations lead, and prove the wedge with 25 design-partner customers — putting us in a position to raise a seed of $2–3M on Q4 2027 metrics.

RUNWAY 18 months	MILESTONE $30K MRR
CUSTOMERS 25 design partners	NEXT ROUND Seed $2–3M

USE OF FUNDS



- **Engineering** — 2 hires: replay engine, scheduler. — 50%
- **GTM & DevRel** — Content, demos, design partners. — 25%
- **Infrastructure & AI gateway** — Compute, Postgres + pgvector, LLM cost. — 15%
- **Ops · Legal · SOC 2** — Compliance is required infra. — 10%

LET'S BUILD THE AGENT DEBUGGER.

Yash Bogam · Founder, Retrace

https://retrace.yashbogam.me https://github.com/yash1511-bogam/retrace-sdk hello@yashbogam.me