From Clicks to Context:

Building an Open-Source Evaluation Pipeline for AI Agents

Inês Bolaños

Senior Product Analyst · PagerDuty

Let's connect!

Agentic AI Foundation

AGNTCon + MCPCon Japan

The speed of change

20 years of navigation

3 years of AI

Image generated with Nano Banana

The goal: destination over map

The hype problem: the industry has moved too fast. Everyone has a trendy new map but most of them don't know where they're going. The focus shouldn't be on having AI, but on whether it's creating meaningful value.

Image generated with Nano Banana

AI Agents vs Industrial Revolution

The cost of hype

AI Agents are powerful, but expensive

Not for every task, they're inefficient for small distances

Just like taking a Shinkansen to the house next door

Agent washing

Rebranding old tech as AI to ride the hype wave

Creates false expectations and devalues real AI innovation

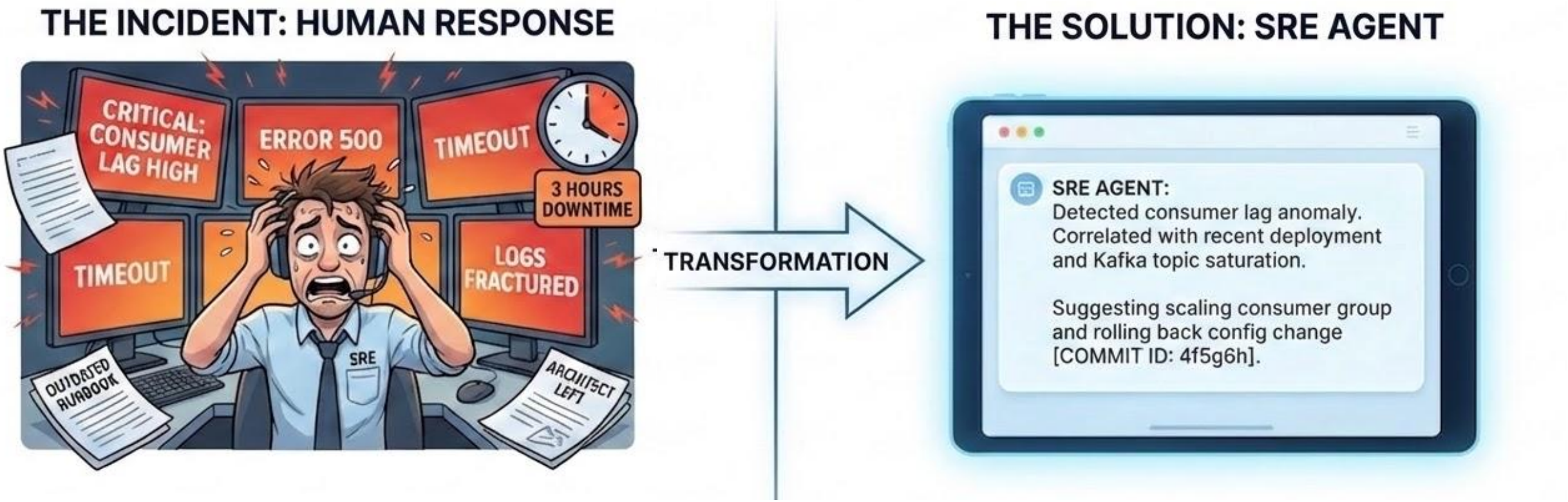
Similar to the amazing databases which were massive Excel files

Image generated with Nano Banana, icons from Freepik

Incident Management

Image generated with Nano Banana

Case study: the SRE agent



- Consumer lag issue
- Logs fractured, multiple runbooks
- 3h of downtime
- Multiple engineers

- Connected the dots immediately
- Suggested the fix quickly
- Reduced time to resolve and pressure
- Provided context aware insights

Image generated with Nano Banana

H.I.R.E.

A framework to hire, train or fire your AI Agents

The H.I.R.E. Framework

H

Humans & Habits

Change readiness & data trust.

Is the team ready to trust the machine?

Is data ready to prevent garbage out?

Framework designed by Inês Bolaños

H

Human & Habits

Implementing AI Agents is more about cultural change than about technical deployment.

1 Change readiness

You're asking an engineer under pressure to trust a machine's fix

2 Garbage in, garbage out

Don't build on messy, siloed data. A confident garbage fix at 2 AM breaks trust instantly.

3 Let it be proactive

We want the agent to be a teammate, so we make it a proactive responder instead of a dumb search bar

Go / No-Go: Is our team ready to trust a new hire to improve a process that already works?

The H.I.R.E. Framework

H

Humans & Habits

Change readiness & data trust.

Is the team ready to trust the machine?

Is data ready to prevent garbage out?

I

Impact & Intent

The job description & ROI.

Would its impact justify a full-time hire?

Framework designed by Inês Bolaños

I

Impact & Intent

If this agent was a human, would its impact justify a full-time salary? What is the expected ROI?

Writing a job description prevents agent washing and clarifies intent:

RESPONSIBILITIES

- First responder to incidents
- Gathers diagnostic data automatically
- Understands troubleshooting steps

BOUNDARIES

- Acts as a co-pilot, not the captain
- Waits for explicit human approval for critical actions
- Operates within defined access and permission

PERFORMANCE

- Measurable reduction in resolution time (eg: 20% faster)
- Increased team efficiency and reduced toil
- Clear ROI based on saved time and resources

The H.I.R.E. Framework

H

Humans & Habits

Change readiness & data trust.

Is the team ready to trust the machine?

Is data ready to prevent garbage out?

I

Impact & Intent

The job description & ROI.

Would its impact justify a full-time hire?

R

Rules · Responses · Reasoning

The Shinkansen vs. walking decision.

Predictable rules → automation.
Response generation → chatbot
Reasoning → agent.

Framework designed by Inês Bolaños

R · Rules vs. Responses vs. Reasoning

The locomotive vs. walking decision: don't take a bullet train to cross the street.

RULES

→ Automation

Predictable, repeatable steps. Cheaper, faster, safer.

e.g. CPU hits 95% = the same three steps, every time.

RESPONSES

→ Chatbot

Generates words, not work. Locked in conversation.

e.g. summarize a doc or answer an FAQ.

REASONING

→ AI Agent

Handles ambiguity, adapts to context, and acts.

e.g. a complex outage: piece together fractured logs under pressure.

The H.I.R.E. Framework

H

Humans & Habits

Change readiness & data trust.

Is the team ready to trust the machine?

Is data ready to prevent garbage out?

I

Impact & Intent

The job description & ROI.

Would its impact justify a full-time hire?

R

Rules · Responses · Reasoning

The Shinkansen vs. walking decision.

Predictable rules → automation.
Response generation → chatbot
Reasoning → agent.

E

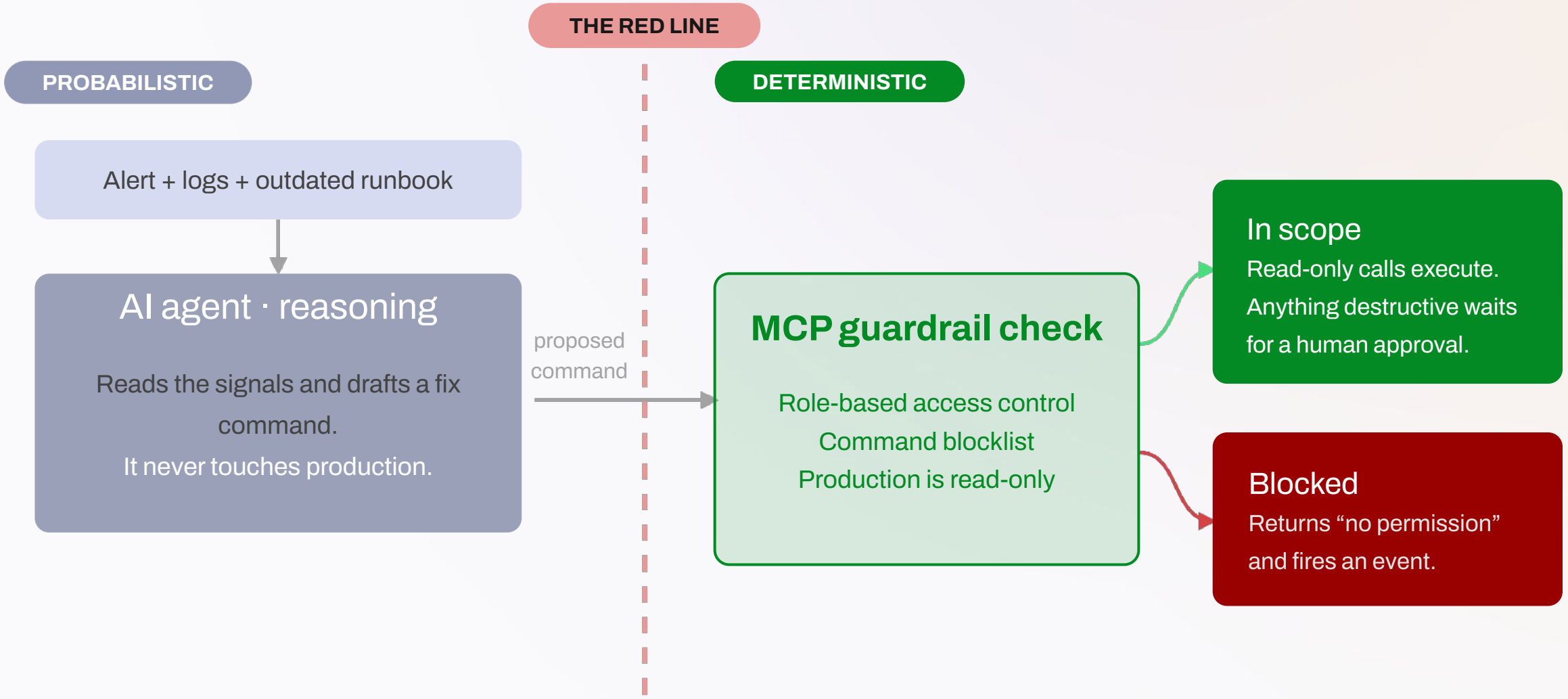
Exposure & Execution

The worst-case scenario.

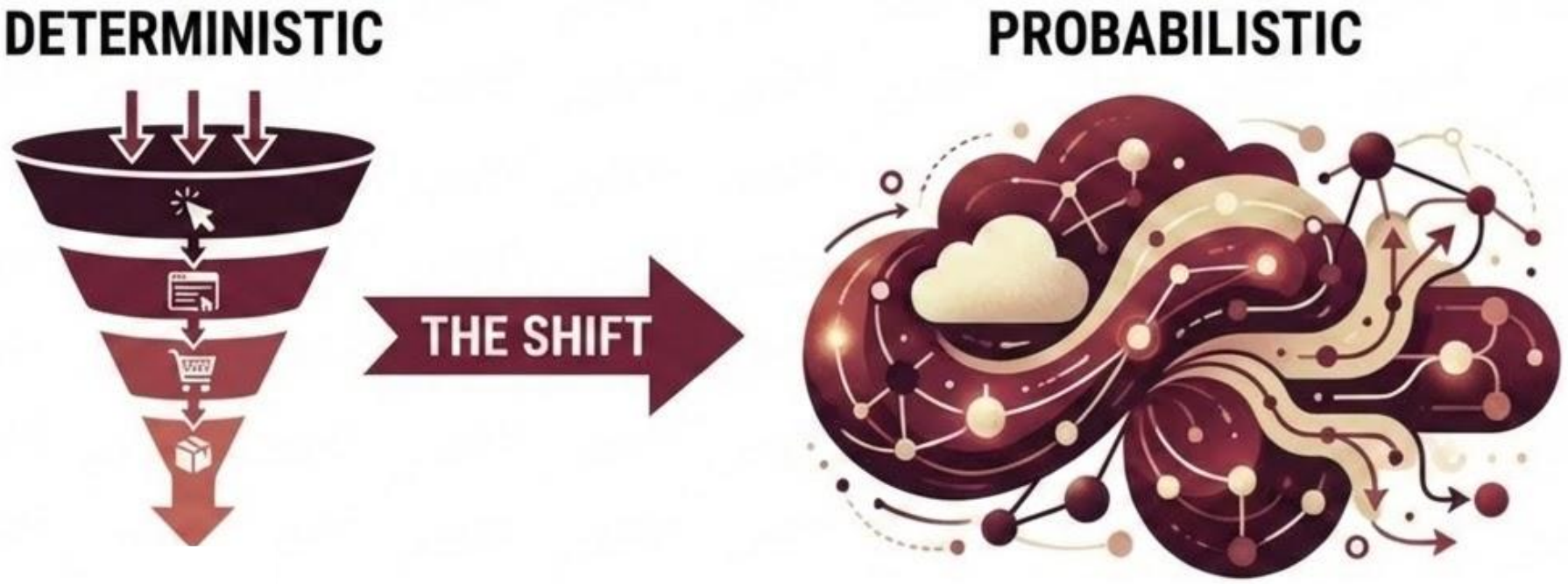
Guardrails live at the MCP boundary

Framework designed by Inês Bolaños

E · Exposure & Execution



New Product Analytics: Probabilistic Software



We are moving from tracking button clicks and feature usage to AI interactions that are similar to human relationships.

We now have full customer transcripts, but must evaluate unstructured data as rigorously as quantitative data.

Technical concepts are now tied directly to user experience.

Images generated with Nano Banana

New AI metrics

Traditional metrics are misleading in the AI context

User Correction Rate (UCR)

How much of the solution does the user have to delete or rewrite before solving the problem?

Low user correction rate → high precision

New AI metrics

Traditional metrics are misleading in the AI context

User Correction Rate (UCR)

Ping Pong Rate (PPR)

Average number of back-and-forth prompts until the user actually gets the correct answer or executes the fix.

Metric should be close to one while still allowing the optimal ratio of clarifying questions.

New AI metrics

Traditional metrics are misleading in the AI context

User Correction Rate (UCR)

Ping Pong Rate (PPR)

Fine, Do It Yourself Rate (DIY)

How often users abandon the chat or execute actions in the traditional UI to solve the problem.

Ultimate churn metric for AI Agents.

New AI metrics

Traditional metrics are misleading in the AI context

User Correction Rate (UCR)

Ping Pong Rate (PPR)

Fine, Do It Yourself Rate (DIY)

Instant Resolution Rate (IRR)

How many times does the AI Agent give an accurate answer? No editsfollow up questions or alternative actions in the UI. Proves system accuracy and human trust.

New AI metrics

Traditional metrics are misleading in the AI context

User Correction Rate (UCR)

Ping Pong Rate (PPR)

Fine, Do It Yourself Rate (DIY)

Instant Resolution Rate (IRR)

Red Line Rate (RLR)

How often does the agent execute an action, suggest a command that hits a guardrail, fails a security check or gets blocked? Events suggested by the agent that the MCP's permission check rejects.

Turning agent conversations into product insights

A three-step pipeline that labels both sides of agent conversations, turning product usage into a routing fix, a new category or a new feature

STEP 1

Deterministic telemetry

Categorize conversations with fixed rules from the agent's own source code: the controls it offers and the actions it takes. Zero model drift and zero hallucination.

Deterministic coverage: this metric indicates whether users are increasingly asking for things outside the agent's contract. The trend is the signal itself.

We label what we know.

Data pipeline designed by Inês Bolaños

Turning agent conversations into product insights

A three-step pipeline that labels both sides of agent conversations, turning product usage into a routing fix, a new category or a new feature

STEP 1

Deterministic telemetry

We label what we know.

STEP 2

LLM supervised classification

A cheap model checks each freeform message against the fixed category list, returning a reasoning and confidence score. Below the confidence threshold, a stronger model gives a 2nd opinion (only what it can't solve reaches the HITL queue).

Known category, missed input: the agent has the capability, but the routing didn't fire it. Phrasing gap, not feature gap.
Unknown category: the input doesn't match any category. It's labeled as *other* and sent to the next step.

We catch what we're missing.

Data pipeline designed by Inês Bolaños

Turning agent conversations into product insights

A three-step pipeline that labels both sides of agent conversations, turning product usage into a routing fix, a new category or a new feature

STEP 1

Deterministic telemetry

We label what we know.

STEP 2

LLM supervised classification

We catch what we're missing.

STEP 3

LLM + unsupervised classification

A stronger model first audits the *other* category: known categories wrongly labeled go back to step 2, ambiguous ones, or below the threshold, to HITL queue. The unknown residual is clustered using BERTopic for reproducibility as inputs grow and an LLM just labels the clusters.

Product signal: track trends in new categories. A stable, high-volume category can be a candidate for a new feature, feeding the deterministic telemetry.

We discover what we don't know.

Data pipeline designed by Inês Bolaños

Turning agent conversations into product insights

A three-step pipeline that labels both sides of agent conversations, turning product usage into a routing fix, a new category or a new feature

STEP 1

Deterministic telemetry

We label what we know.

STEP 2

LLM supervised classification

We catch what we're missing.

STEP 3

LLM + unsupervised classification

We discover what we don't know.

$ **Cost-aware tiering:** rules are free, the cheap model handles most of the cases, the strong model sees only low-confidence rows, humans see only what both models can't solve.

FEEDBACK LOOP: each step feeds the one before it

Step 2 → 1: better routing turns *known category, missed input* into rules, raising **deterministic coverage**.

Step 3 → 2: recurring emergent themes are promoted into the **supervised category** list or even to a new feature, feeding the **deterministic telemetry**.

Evaluate the evaluator: every LLM layer is scored against a human-labeled gold set, and the confidence threshold is calibrated on it, not guessed

Process runs on a schedule, snapshots track metrics over time: deterministic coverage, routing quality, resolution status, emergent topics

Tone analysis: understanding not just *what* users ask but ***how*** they ask it. Reading sentiment turns frustration into fixes and satisfaction into direction.

Empathy drives adoption

Corporate Incident Report (INC-8092)

Incident ID: INC-8892
Date/Time: 2026-10-27 02:14 UTC
Severity: SEV-1 (Critical)
Status: Resolved
Description: Primary database cluster db-prod-us-east-1 unavailable due to unexpected failover and subsequent cascading replication lag. **Impacted services:** All customer-facing APIs, Web App, Mobile App.
Resolution: Initiated emergency failover to secondary region. Restored from point-in-time recovery snapshot. Root cause analysis (RCA) pending.

The lightness after the fire

The workday ends, you pour a beer,
and the pager screams right in your ear.
You dig up a runbook from years ago,
its author has left; where did they go?

The logs don't match, the minutes fly,
the losses grow as time slips by.
You ask the team, but the night is deep
you fix it at dawn, too tired to sleep.

Images generates with Nano Banana, poem generated by Gemini

The AI Shift

~~Stop measuring **clicks**~~. Start measuring **friction removed**.

Technology is accessible. **Cultural change** is the true challenge.

The **tools** have changed. Our **function** hasn't.

Our purpose is to provide **direction**

Image generated with Nano Banana

The AI agent lifecycle

Don’t just buy the trendiest new map.

Validate your destination.

GitHub: HIRE-AI-framework

Framework designed by Inês Bolaños, image generated with Nano Banana

The AI agent lifecycle

Make every AI agent pitch go through the **H.I.R.E.** test:

HIRE if it solves a real human problem, handles true ambiguity and has safe guardrails.

TRAIN if it's struggling with context but ROI is there.

Having the courage to know exactly when to **FIRE** an AI agent is what makes a great product strategy.

GitHub: HIRE-AI-framework

Framework designed by Inês Bolaños, image generated with Nano Banana

From Clicks to Context:

Building an Open-Source Evaluation Pipeline for AI Agents

Inês Bolaños

Senior Product Analyst · PagerDuty

Let's connect!

Agentic AI Foundation

AGNTCon + MCPCon Japan