From Evals to Experiments

How to Ship Successful AI Initiatives by Failing Cheaply

DATADOG

How do you know if your AI initiatives are working?

How *quickly* do you know if your AI initiatives are working?

Without a plan, initiatives could take months or years to measure

Bloomberg

Subs

he AI Race: Startups to Watch | AI Is Rewiring Finance | AI Unlocks Ancient Secrets | How AI Chatbots Work | Gemini Backlash

Technology

Call-Center Firm Sinks on Klarna Claims AI Is Doing Agents' Jobs

- Teleperformance shares slump on fears over AI impact
- Klarna says AI helps solve customer errands faster than humans

By Paul Jarvis and Henry Ren
February 28, 2024 at 10:58 AM UTC
Updated on February 28, 2024 at 2:24 PM UTC

Bloomberg

Subscribe

Industries | Finance

Klarna Rethinks AI Cost-Cutting Plan With Call for Real People

Sebastian Siemiatkowski *Source: Klarna Holding AB*

By Charles Daly
May 8, 2025 at 8:00 AM UTC

Copy Li

Time to course correct: **~14 months**

If we can’t fail fast, the cost of failure quickly piles up

- **Disappointed users** and subsequent hit to our reputation
- Ceding **competitive ground** by moving too slowly
- **Wasted budget** on an initiative that didn’t work
- The hidden **opportunity cost** of not working on something better

Over the last three years,

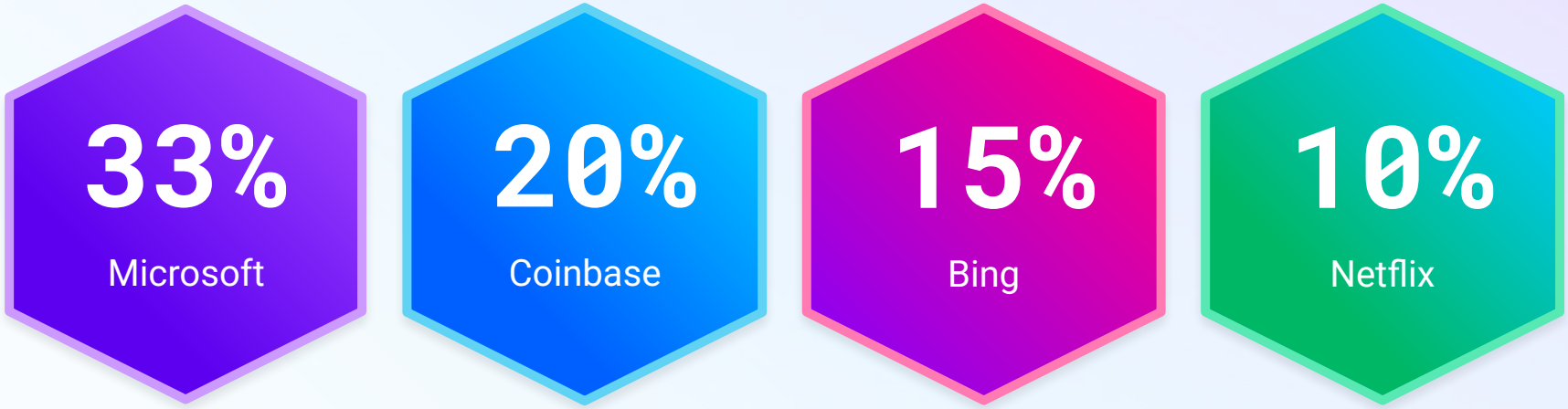
25%

of AI initiatives have delivered expected ROI

Source: IBM CEO Study, 32nd ed. (2025), n = 2,000

DATADOG

A/B tests across disciplines have similar success rates, but only take weeks to measure



Sources: Kohavi, Deng, Vermeer. “A/B Testing Intuition Busters” (2022)
Adams and Ou-Yang. “Scaling Experimentation for Machine Learning at Coinbase” (2023)

How do we bring that power of A/B testing to AI initiatives?

From Evals to Experiments

Use Evals to prioritize and tune potential solutions

Use Experiments to prove business impact and make decisions

Use the right tool for the each part of the process

Our ultimate goal is to solve a business problem

Template: We want to build for use case **X**, in order to drive **Y** business outcome

E.g., “We want to build AI-powered **customer support agents** to drive **decreased support costs**.”

This becomes our *hypothesis*, and we run an experiment to close the feedback loop.

To get there, we need to answer smaller questions about “how”

Evals help guide *how* we approach the build.

- What model(s) will do the best job here?
- Can we solve this affordably?
- How should our prompts/datasets look?
- How should we set parameters like temperature?

Evals = pre-production “experiments”

Try out multiple combinations of models/parameters

Use **fixed inputs** in order to compare *(e.g. expected, atypical, adversarial)*

Conduct **error analysis** to define failure modes and **metrics/evaluators**

Look at your **evaluators** plus metrics like **cost, token use, duration**

Eval strategy starts with product strategy

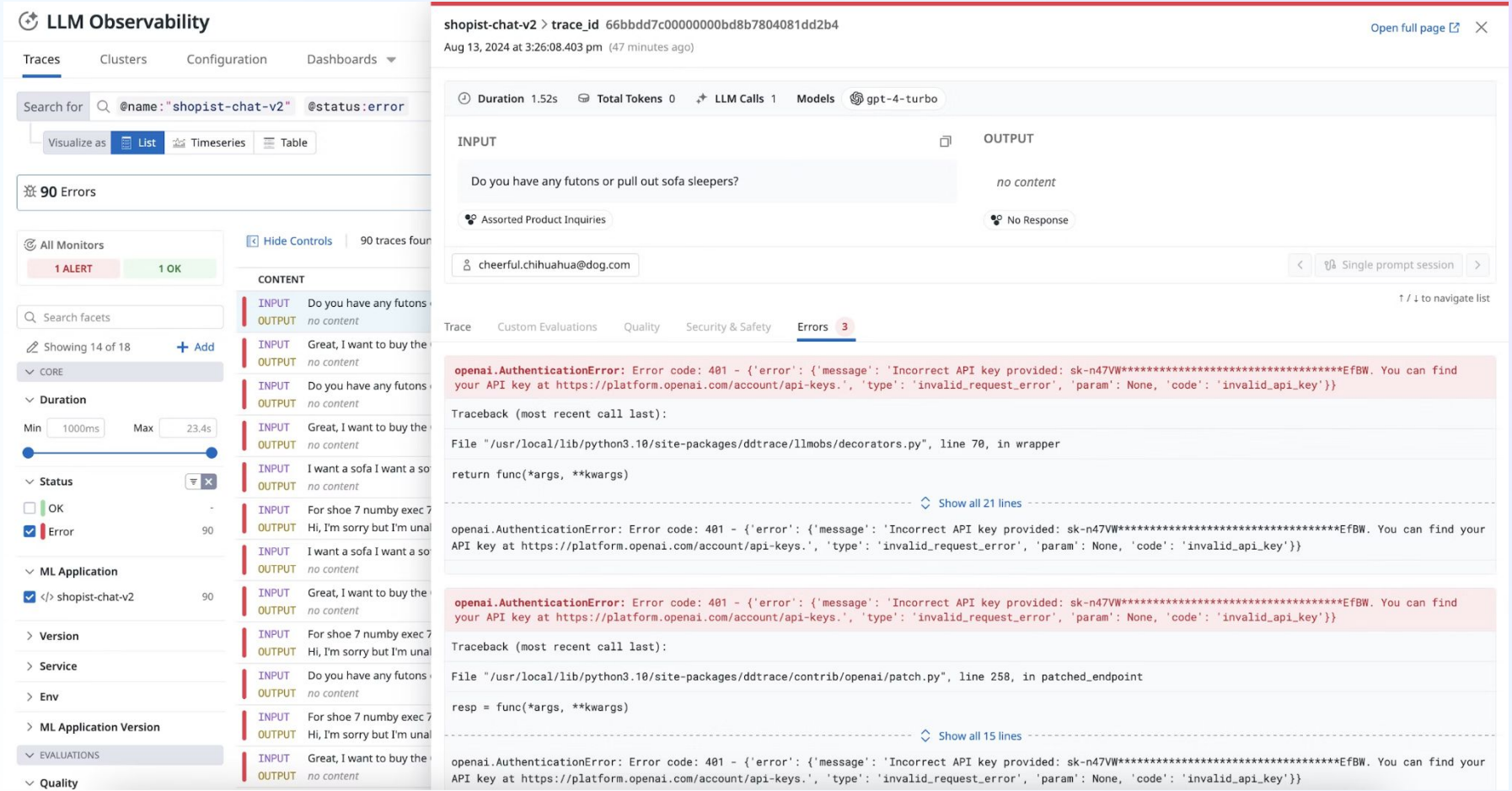
Who are you building for?

What are their **jobs-to-be-done**?

How are they likely to prompt your product?

What about this may **evolve over time?**

Look at traces to identify and categorize errors



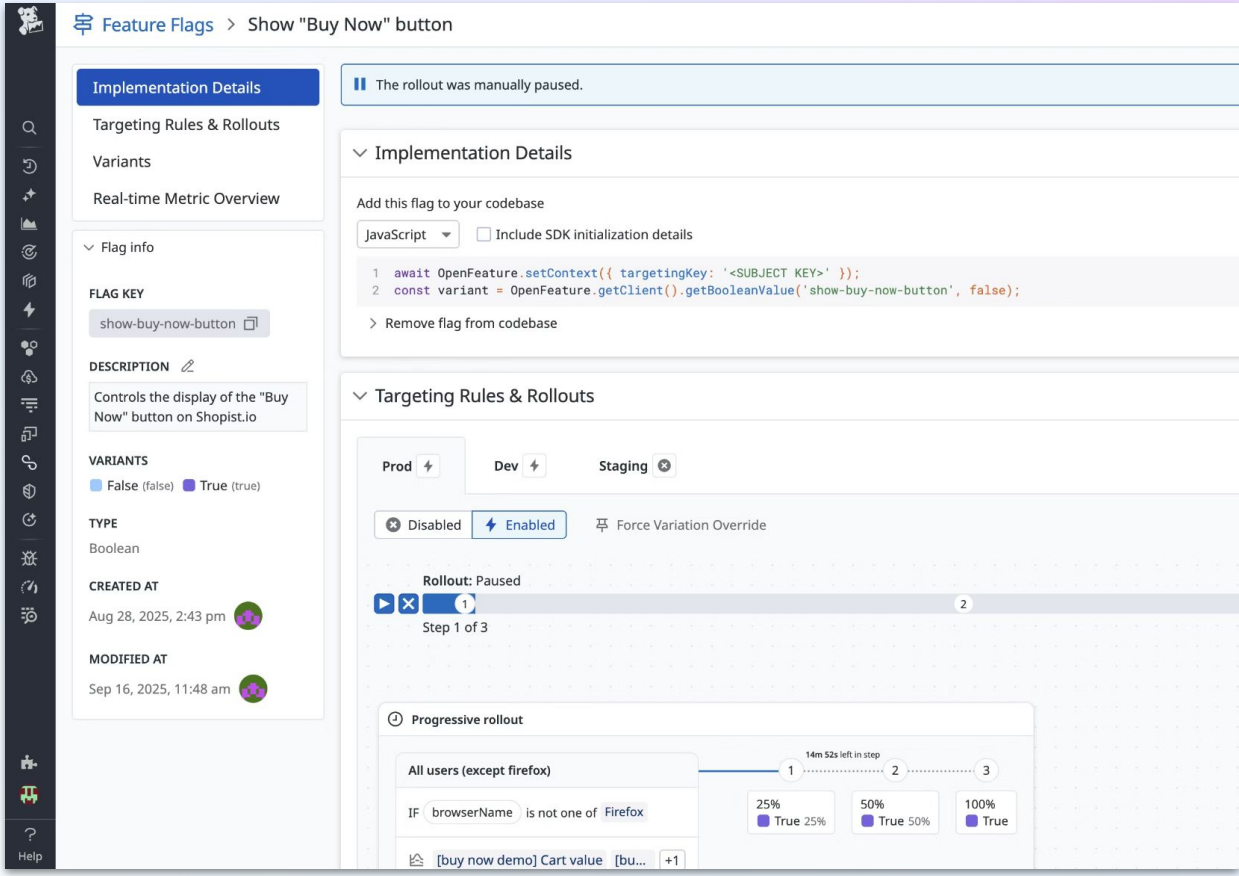
Example: Datadog Feature Flags

AI Feature #1:

Feature flag implementation

AI Feature #2:

Stale flag cleanup



DATADOG

Use errors to craft evaluators

*(*not real examples)*

AI Feature Flag Creation

Problem: Model is incorrectly implementing flag
Idea: Did the model read our documentation?

Problem: Model is writing really inefficient code or hallucinating

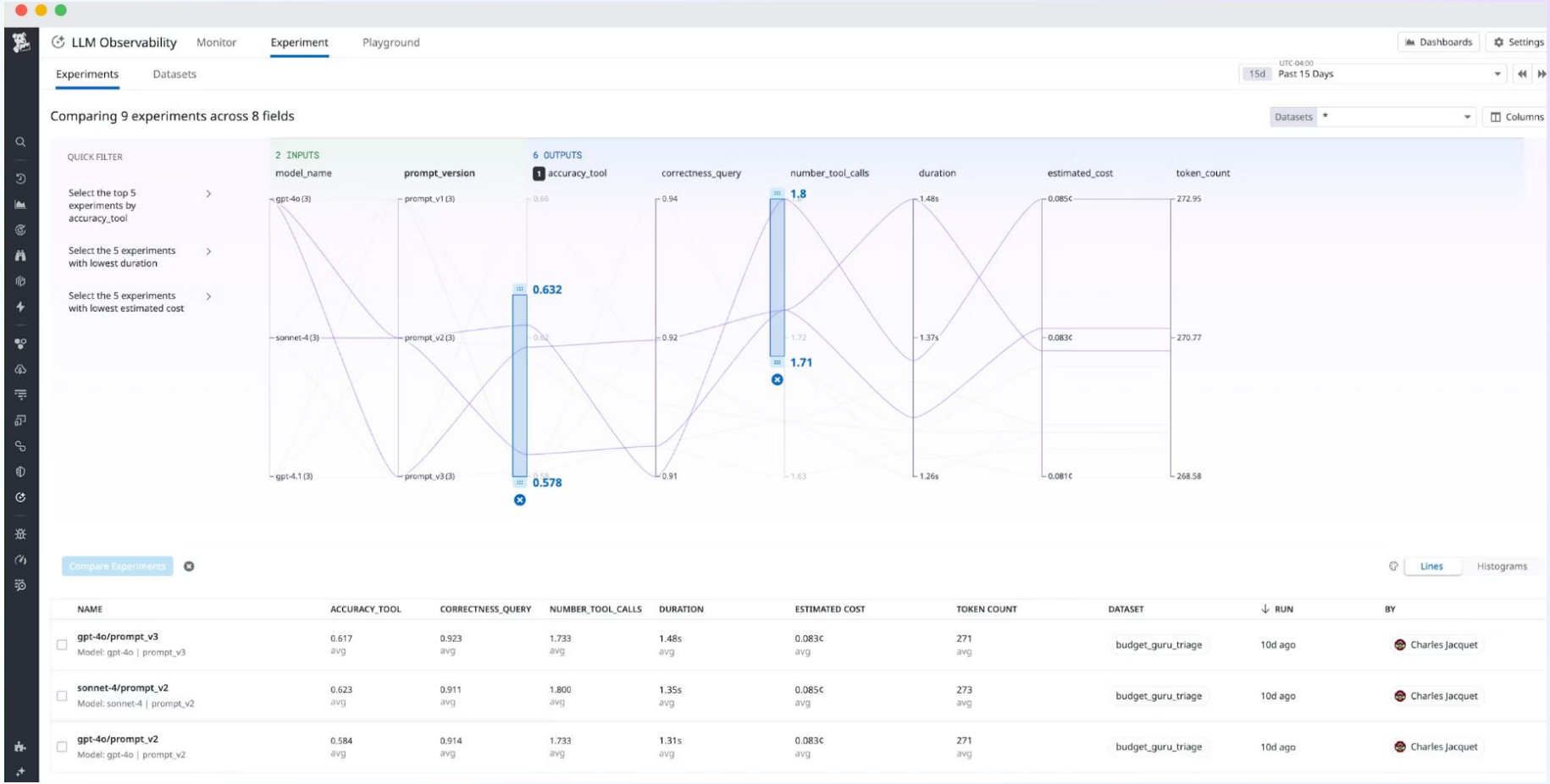
Idea: What was the size of the code change generated?

AI Stale Flag Cleanup

Problem: Model does a partial job - correctly removes some code, but leaves some behind

Idea: “LLM as judge” - ask a second LLM to check the first LLM’s work

Use evals to test prompt or model changes



Why aren’t pre-prod evals alone sufficient?

Limited sample size may not catch all real cases

Evaluating text output is often **subjective or qualitative**

Pre-prod metrics often **fail to correlate to business metrics**

Experiments bring AI to the real world

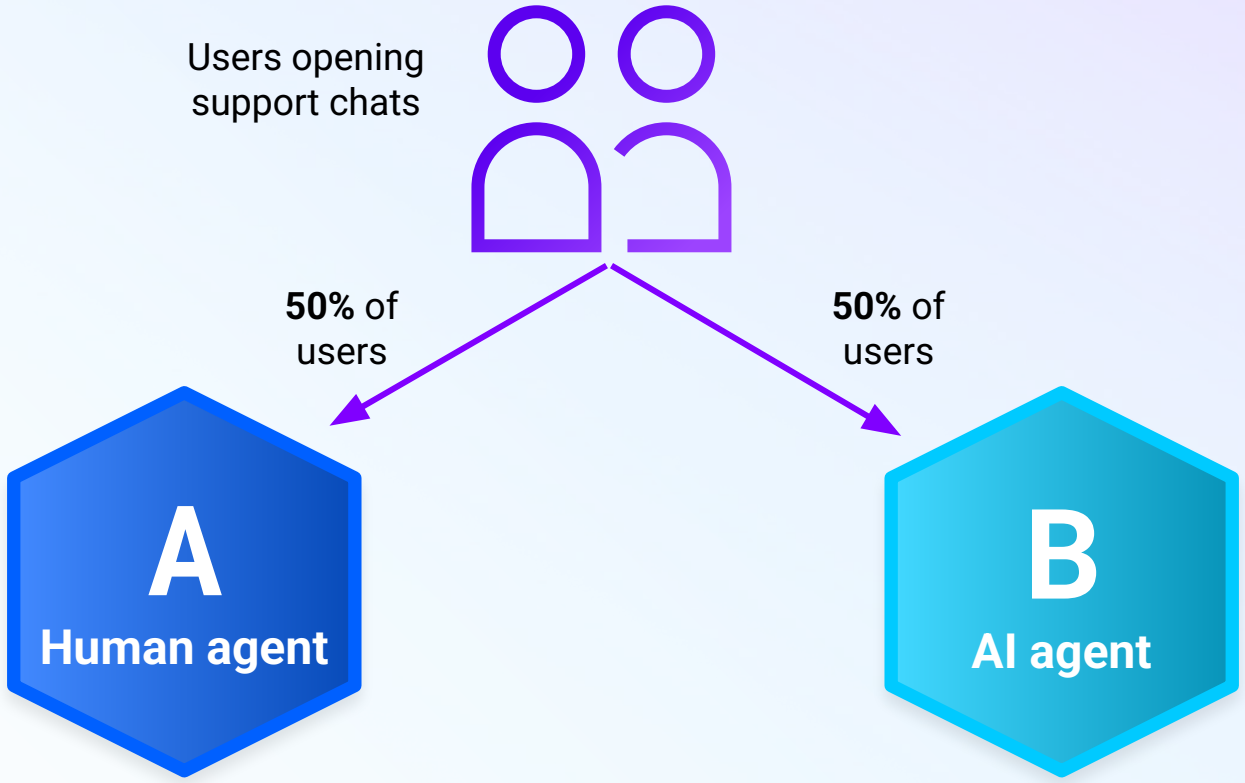
Test our solutions in **real-world environments, with real users**

Randomized, controlled trials **prove what works** *(causality)*

Tie the AI product to **business metrics like revenue or cost**

Experimenting at scale gives insights in **weeks, not quarters**

Example: Testing an AI support agent



The building blocks of an experiment plan

Three Types of Experiment Metrics

Primary Metric(s)

Business outcome

- Examples
  - Revenue
  - Completed Transactions
  - Profit Margins
  - Return/Refund rate
  - Value-driving usage (e.g. # of nights booked for a hotel)

Guardrail Metrics

Avoid negative second-order effects

- Examples
  - Repeat purchase revenue
  - Customer LTV
  - Average Order Volume
  - Support Costs
  - Error rates

Storytelling Metrics

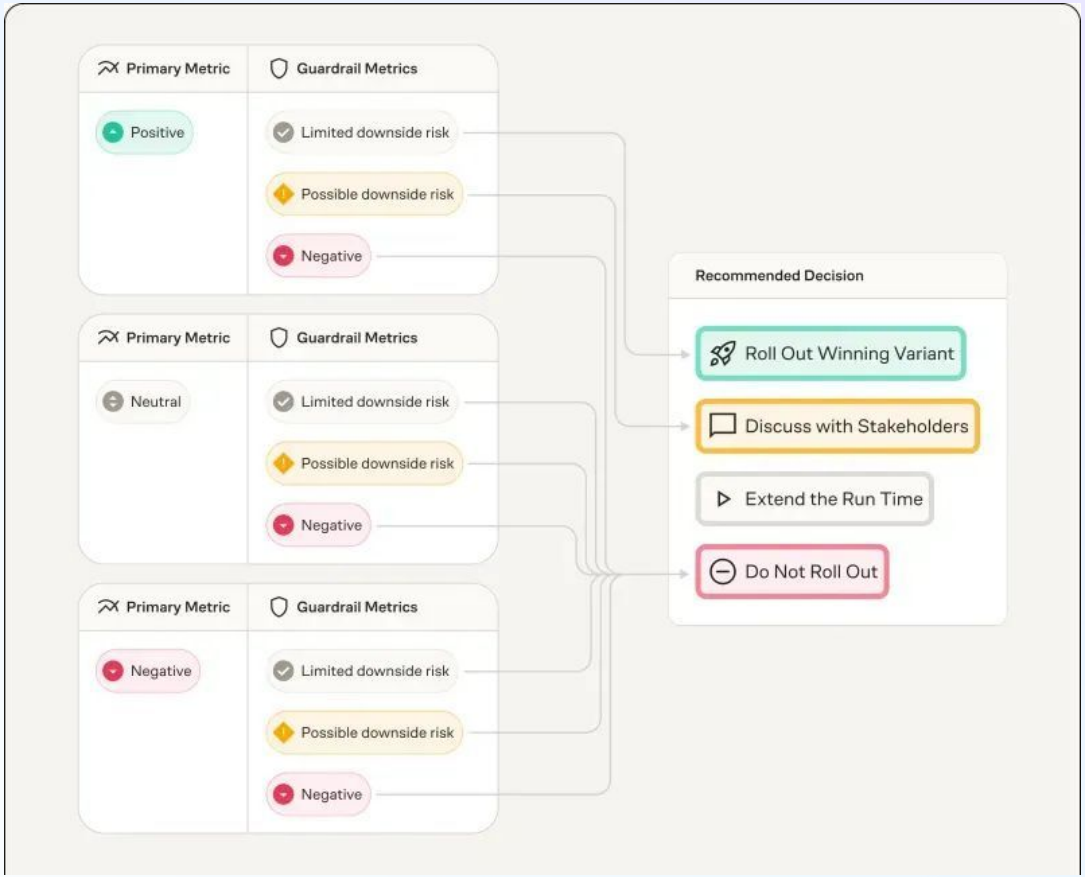
Fuel further hypotheses and future exploration

- Examples
  - Time to ticket resolution
  - Customer satisfaction
  - Latency / performance
  - Time on site

Transactional **Event Stream**

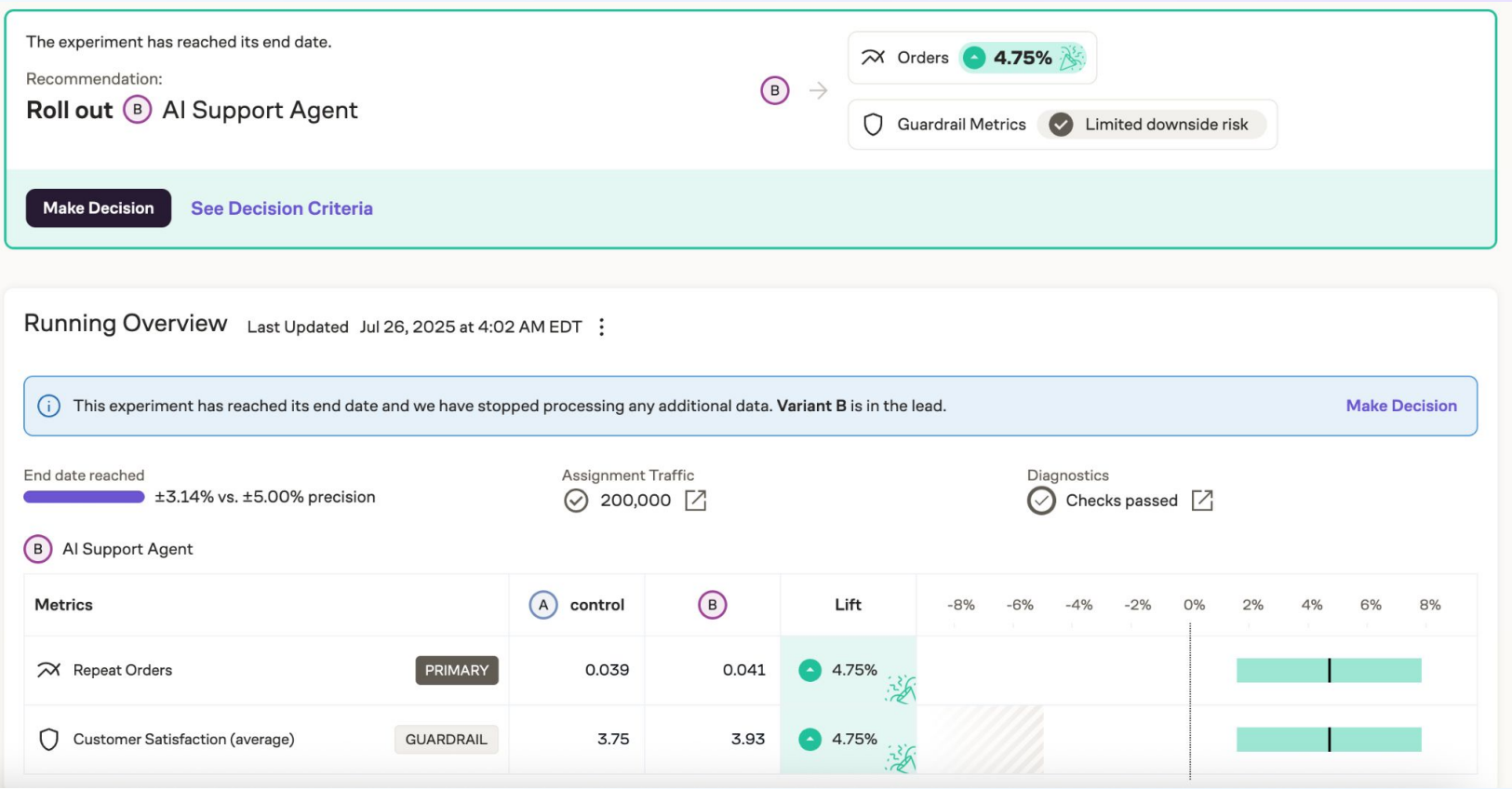
DATADOG

One more step: make decisions in advance

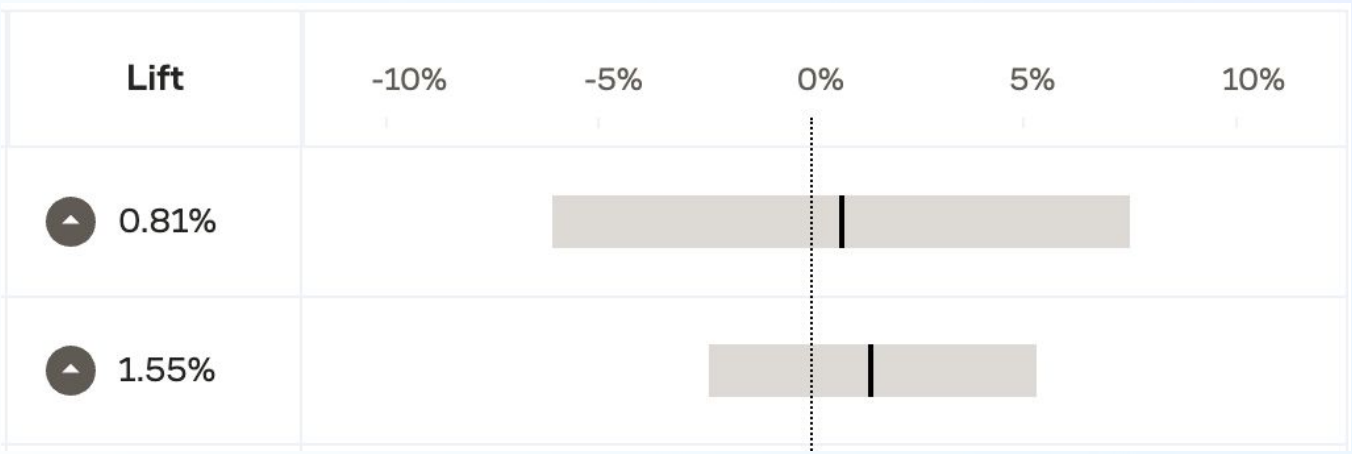


DATADOG

Example experiment results

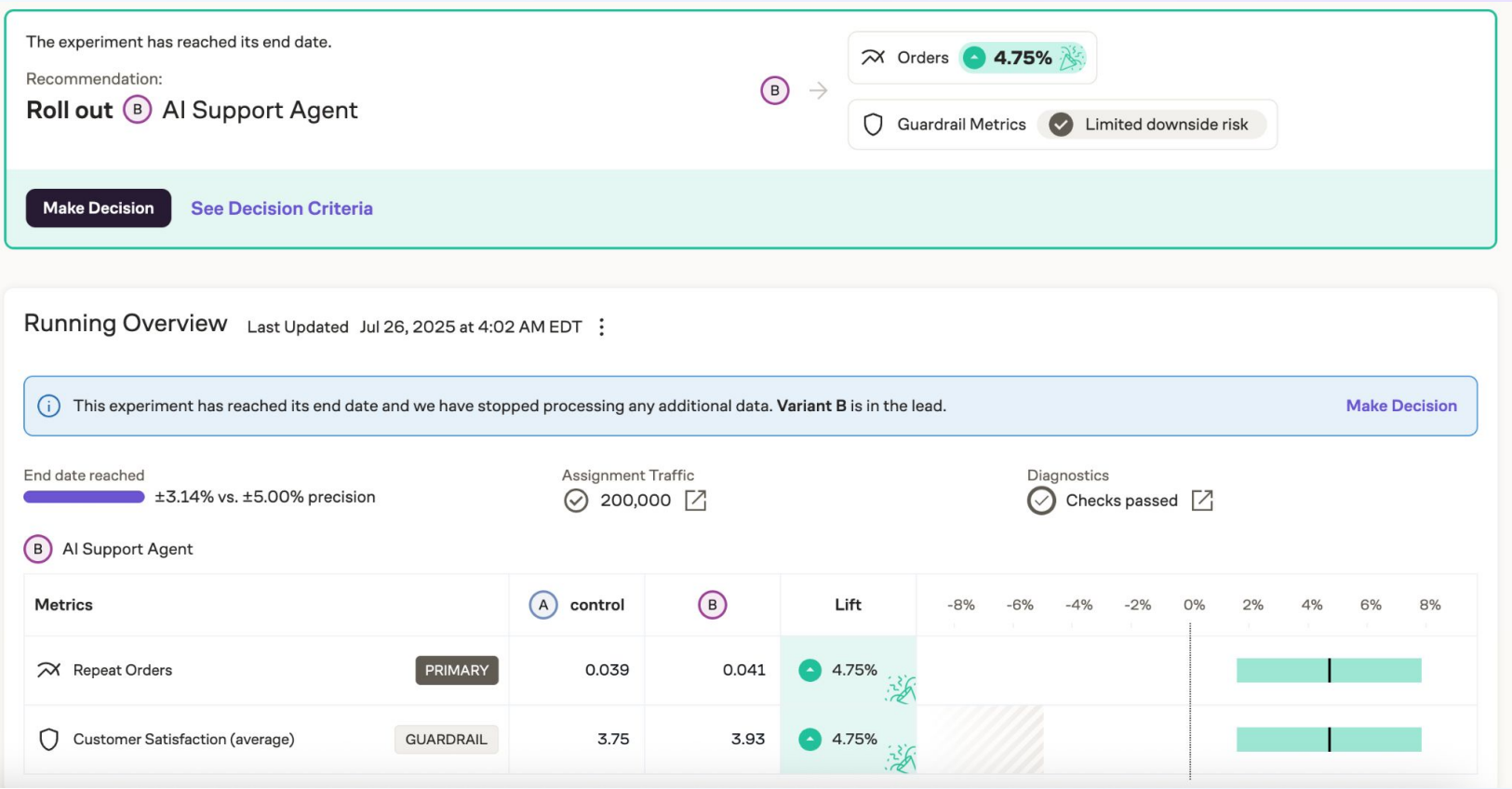


Stats hack: look at confidence intervals



DATADOG

Output: Confident business decisions



Combine Evals + Experiments

Don’t fall prey to long feedback loops

Define success metrics as business outcomes

Run A/B tests to go from idea to proof

Thank you!

Ryan Lucht
ryan.lucht@datadoghq.com
linkedin.com/in/ryanlucht

DATADOG