

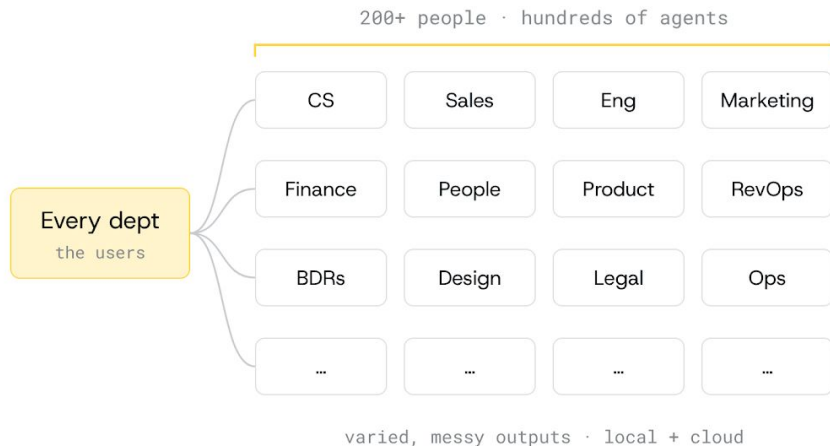
Is Your Agent Any Good?

Lessons from a wide and a deep deployment

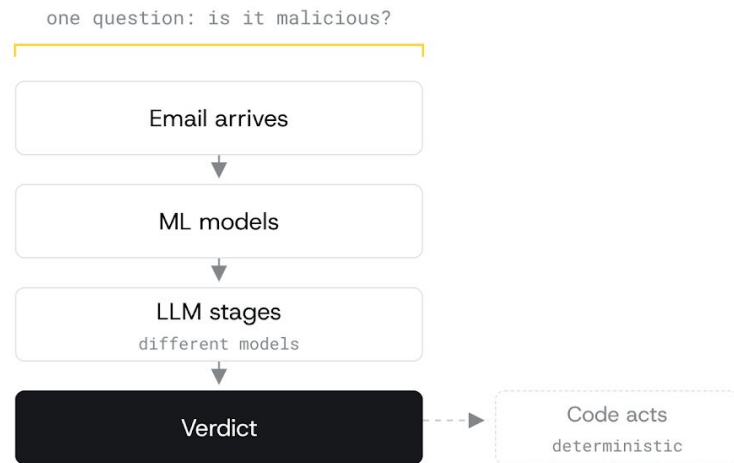
Ethan Lo, AI Enablement | Jason Jiang, Email Security
Adaptive Security

Same company, two very different deployments

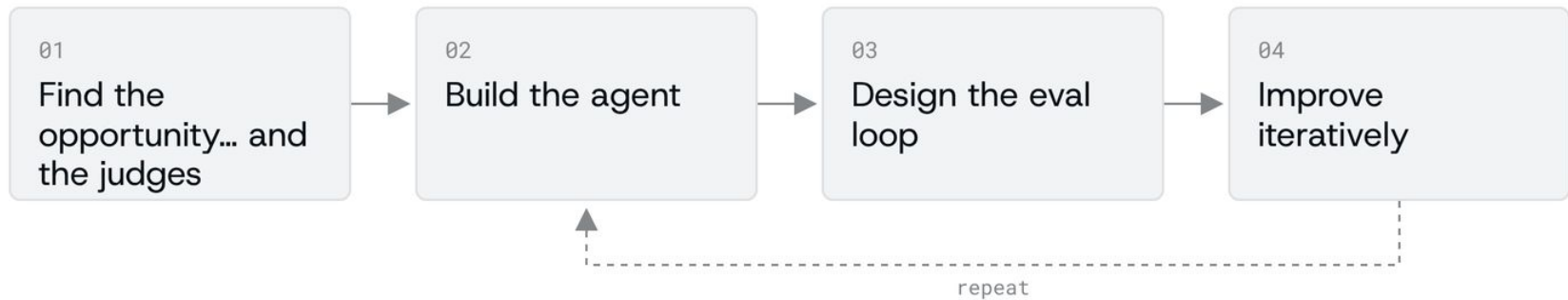
Wide: AI Enablement (Ethan)



Deep: Email Security (Jason)



Every agent goes through the same four steps:

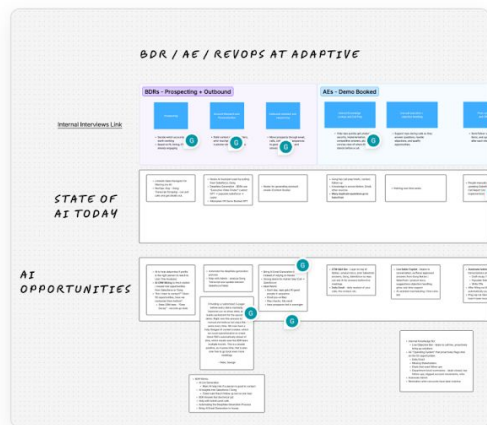


AI Enablement

Adopting AI in 2026

Find the opportunity... and the judges

what we went looking for



weeks of interviews · a flowchart per team · where AI fits

what we found

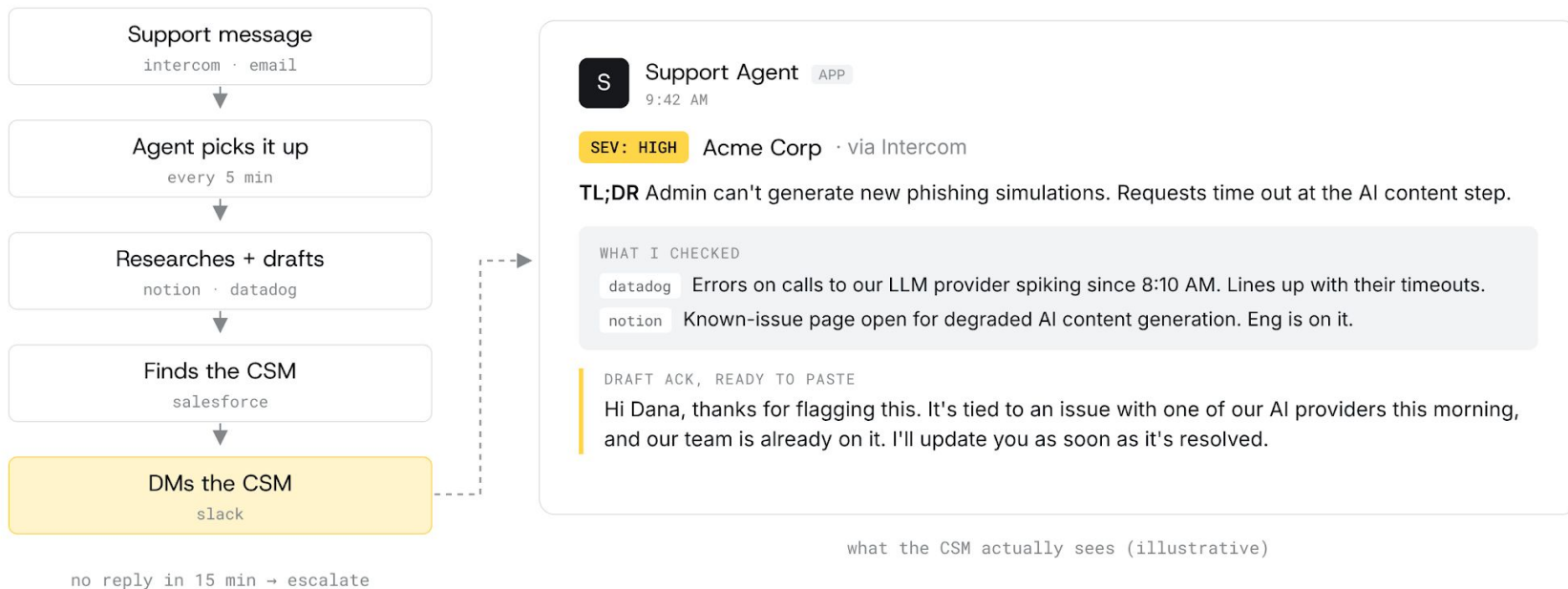
Department heads

set the bar · "you're still responsible"

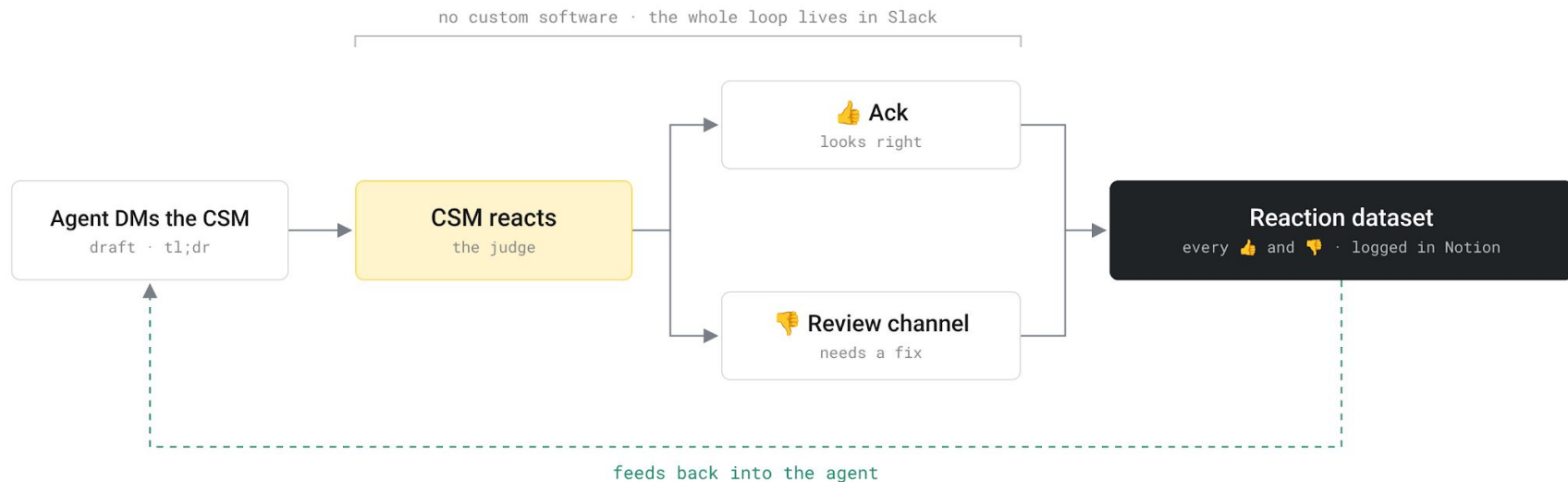
The people doing the work

can tell a good answer from a bad one

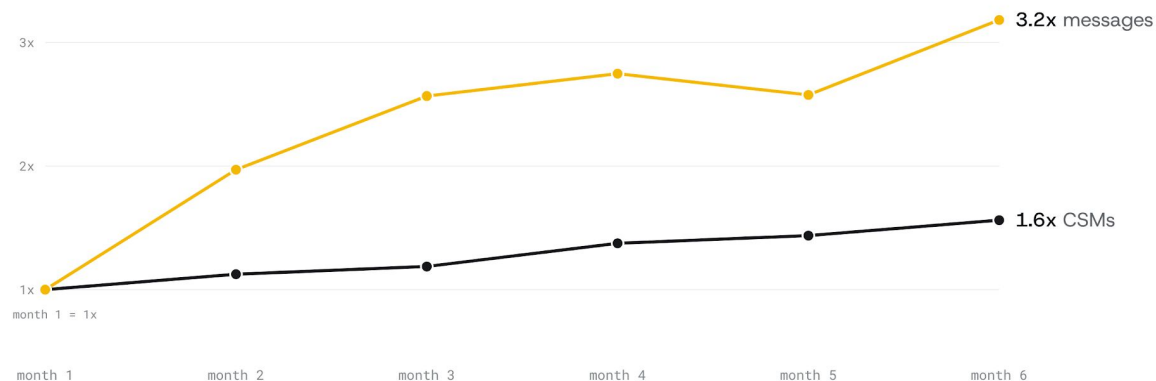
The Automated Support Bot



CS designed their own eval: an emoji



Rollout



customer messages triaged by the agent vs. CSMs on the team, per month, indexed to month 1

ESTIMATED TIME SAVED

~30%

for every CSM

CUSTOMER NPS

90+

the entire time

But then... the judges went quiet

during rollout

CSMs reacted to every draft

👍 acks it · 🗨 goes to review
fixes shipped from the 🗨s



after rollout

The reactions stopped

the bot was trusted
nobody owned the loop



follow-up poll

7 of 10 CSMs still use it

they just stopped telling us
when it was wrong

"is the automated-support bot still being used or do you guys kind of ignore it now?"

me, to our CS Owner

"hmmm let me take poll"

our CS Owner

The ground shifts under your feet

01

The judges went quiet

Once CSMs trusted the bot, they stopped reacting

02

Scale slowed it down

The support bot slowed down: from running every 5 minutes to every 20

03

The tools changed

Google added a send tool to the Gmail connector, and Claude started emailing clients!

Agents are software. Maintain them like software.

01

Use software techniques

Add monitoring/alerting to your eval. Split the agent into pieces you can test and watch on their own. Lock down connectors and review changes like any other dependency.

02

Evals are an operation, not a project

Start with humans, but be flexible when they stop annotating. Someone has to own the eval for as long as the agent runs.

03

Agents don't break like normal software

They slowly get worse, and people quietly stop trusting them. That's almost worse than an outage.

Automated Support: Now What?

the new eval

Bot drafts a reply

sent to the CSM in Slack



CSM sends their own reply

in Intercom



An LLM compares the two

classifies each pair

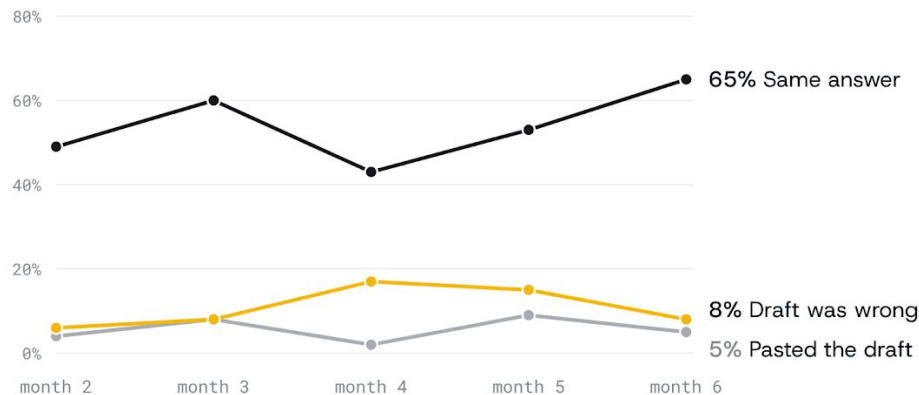


Same answer

Pasted it

Draft was wrong

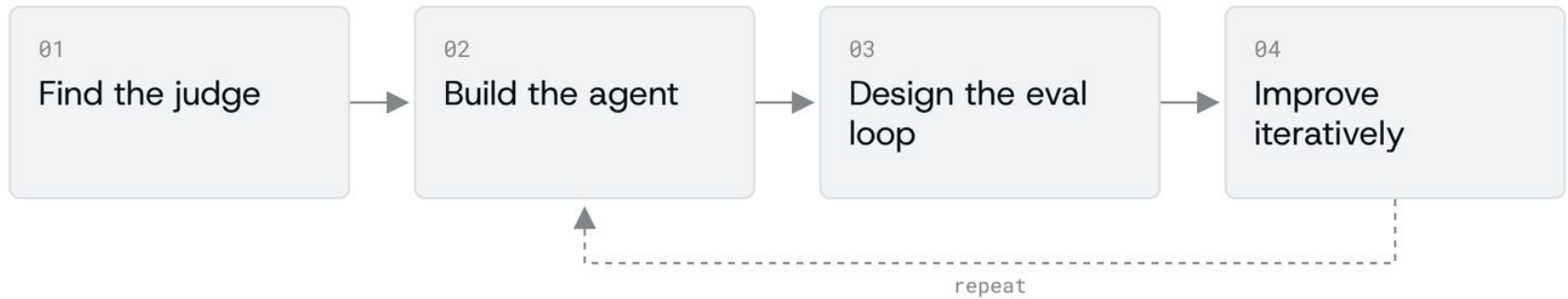
bot draft vs. what the CSM sent



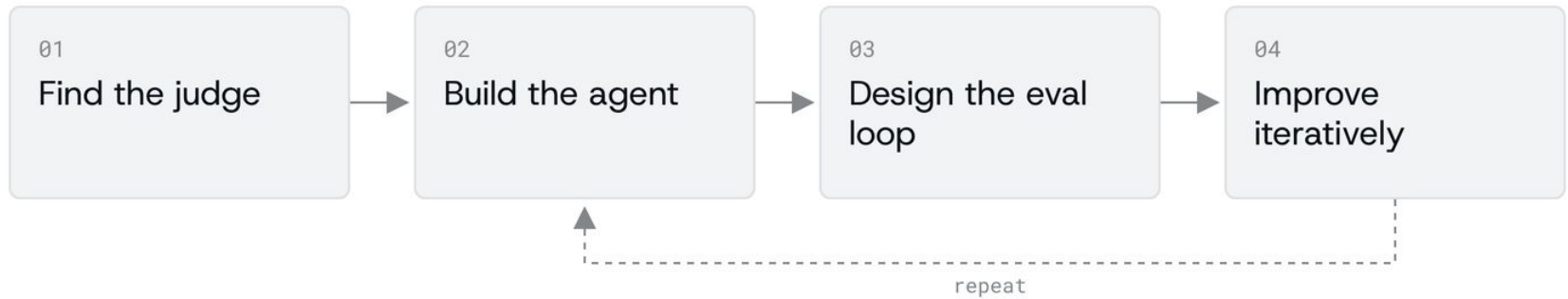
~40 comparable replies sampled per month

Email Security

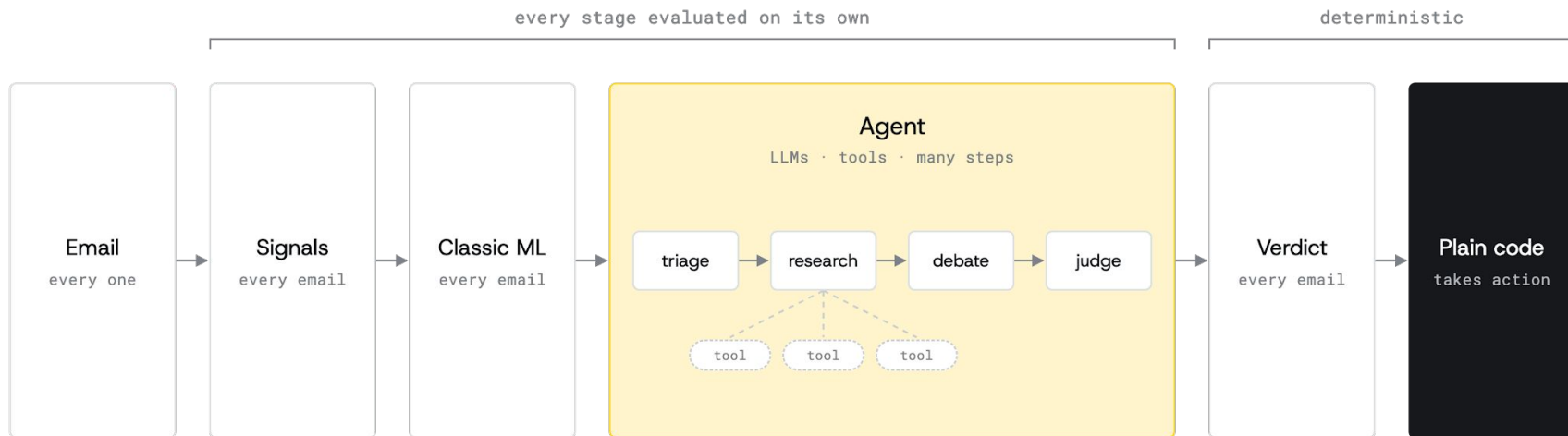
Same four steps.



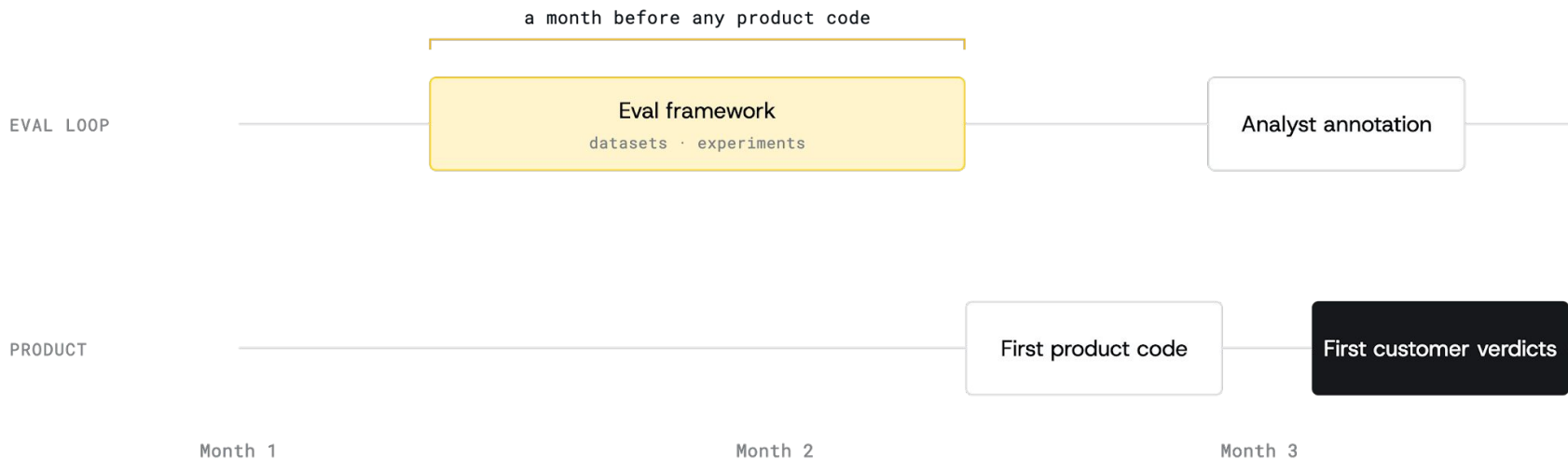
Same four steps. Different order.



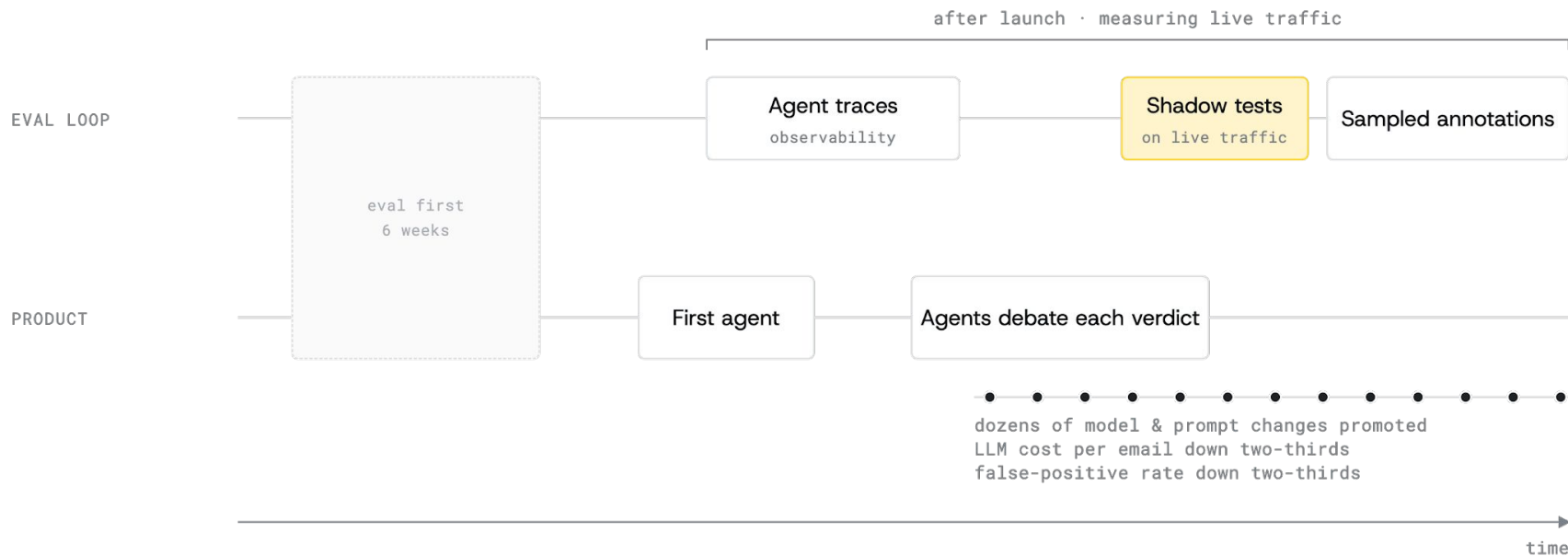
Email detection pipeline overview



We spent over a month on just the eval loop



We kept building the eval loop after launch



Shadow testing on live traffic

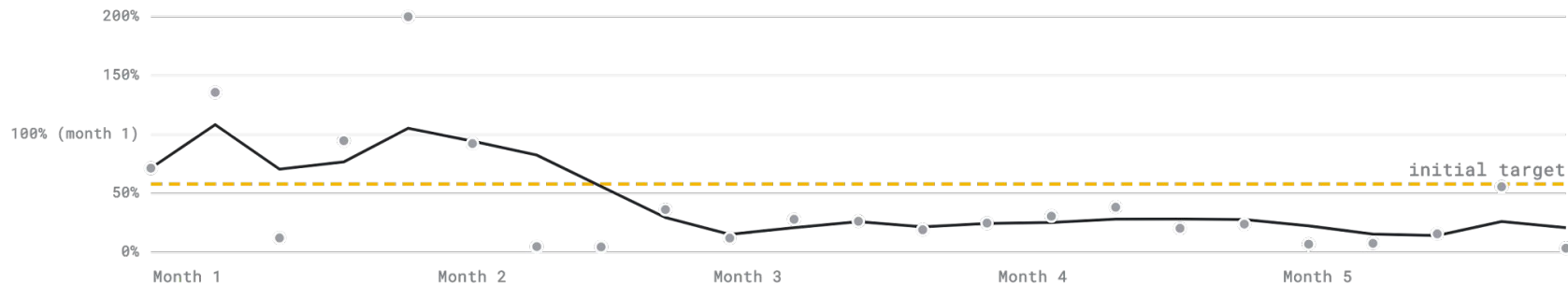
	OFFLINE GATES	SHADOW TEST · LIVE TRAFFIC
Cost per email	4–15% cheaper predicted	17.5% cheaper held
Tail latency	45% faster p99 · predicted	37% faster p99 · held
False positives	no change within noise	36% more confirmed inbox false positives
Threats caught	1% fewer within run-to-run noise	4% fewer relative to production, same emails

The cheapest model wasn't the one we shipped

	MODEL A	MODEL B · CHEAPER PER TOKEN
Cost per email	22% cheaper held in shadow test	37% cheaper smoke test
Threats caught	no change within noise	9% fewer significant recall loss
False positives	no change within noise	13% fewer smoke test
	Shipped	Rejected before it touched production

Fewer false positives, at a third of the cost

False-positive rate 4-week rolling · dots = weekly · % of month 1 average



LLM cost per email weekly · % of month 1 average



We need to constantly adapt our process

Human labels a day 4-week average



ANNOTATION

3 analysts label ~1,000 emails a day

labels up 2.6x · mail up 11x

NON-DETERMINISM

The judge disagrees with itself on 15–18% of borderline emails

so every eval scores each email multiple times

NEVER DONE

The eval loop is an operation, not a project

as our product improves, so too does the evaluation process

We need to constantly adapt our process

Human labels a day 4-week average



ANNOTATION

3 analysts label ~1,000 emails a day

labels up 2.6x · mail up 11x

NON-DETERMINISM

The judge disagrees with itself on 15–18% of borderline emails

so every eval scores each email multiple times

NEVER DONE

The eval loop is an operation, not a project

as our product improves, so too does the evaluation process

Before you deploy your next agent

	WIDE · AGENT FIRST	DEEP · JUDGE FIRST
01 Find your judge	CSMs	Engineers first, Security Analysts at scale
02 Give them a way to judge that fits their day	Slack emoji → LLM Judge	Custom-built annotation queue
03 Treat evals as an operation, not a project	Things change; watch and adjust	Constantly iterating alongside our product

Questions?

Ethan Lo | Jason Jiang | Adaptive Security