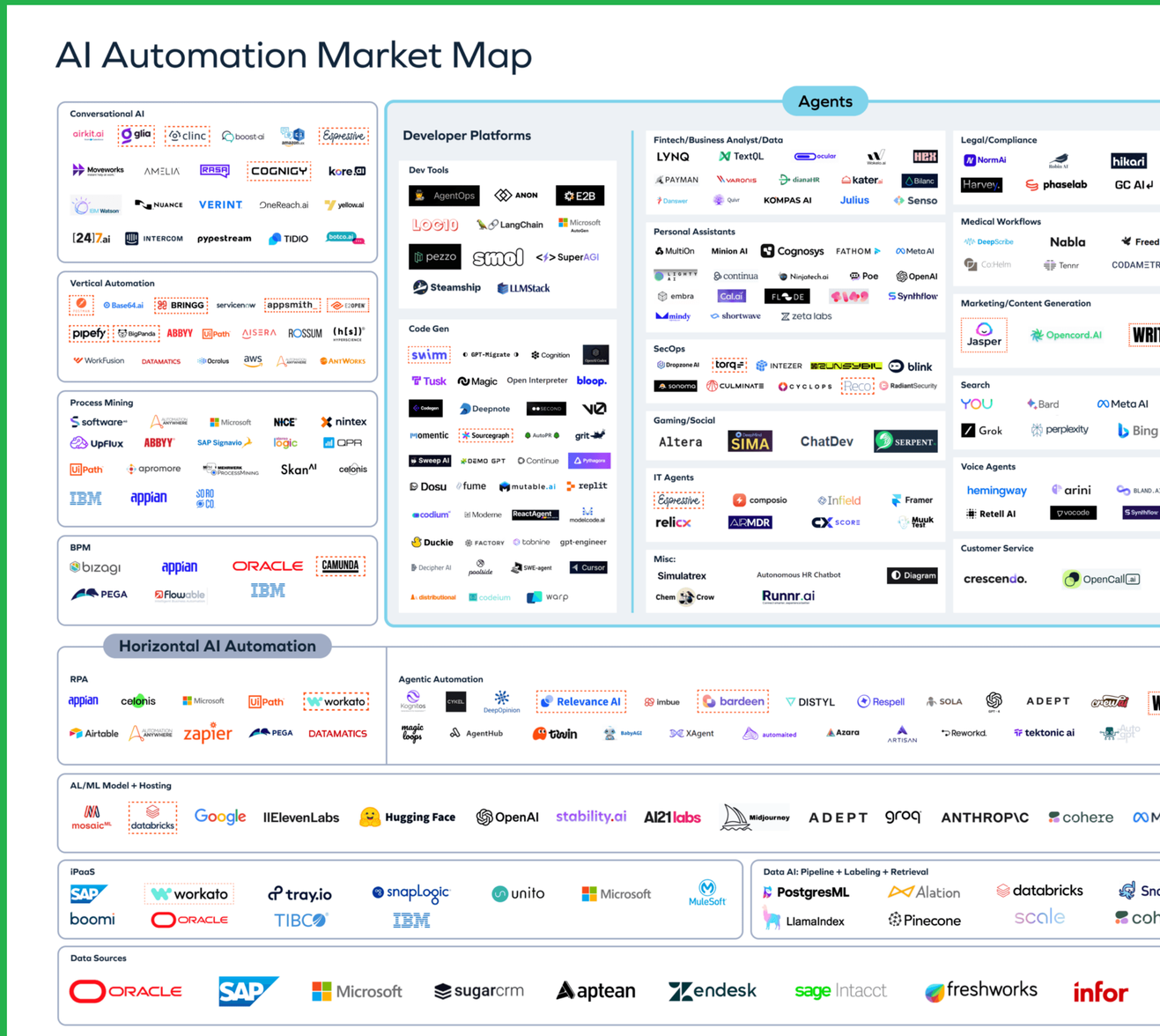
AI Agents That Don't Suck

How to build ones that actually drive business value

Robert Koch

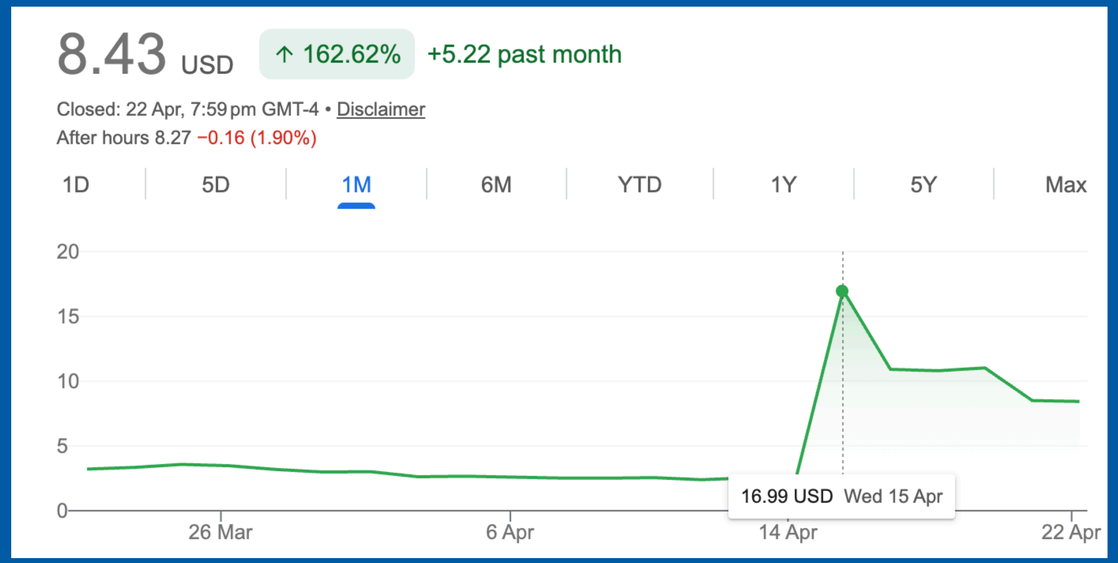
The technology is solved. People aren't.

- Foundational models can perform at human levels
- Frameworks already exist to bundle AI into a convenient package
- Human psychology is now the bottleneck
- Are you building something people will pay for, or just something that's technically impressive?
- Will people use it tomorrow?



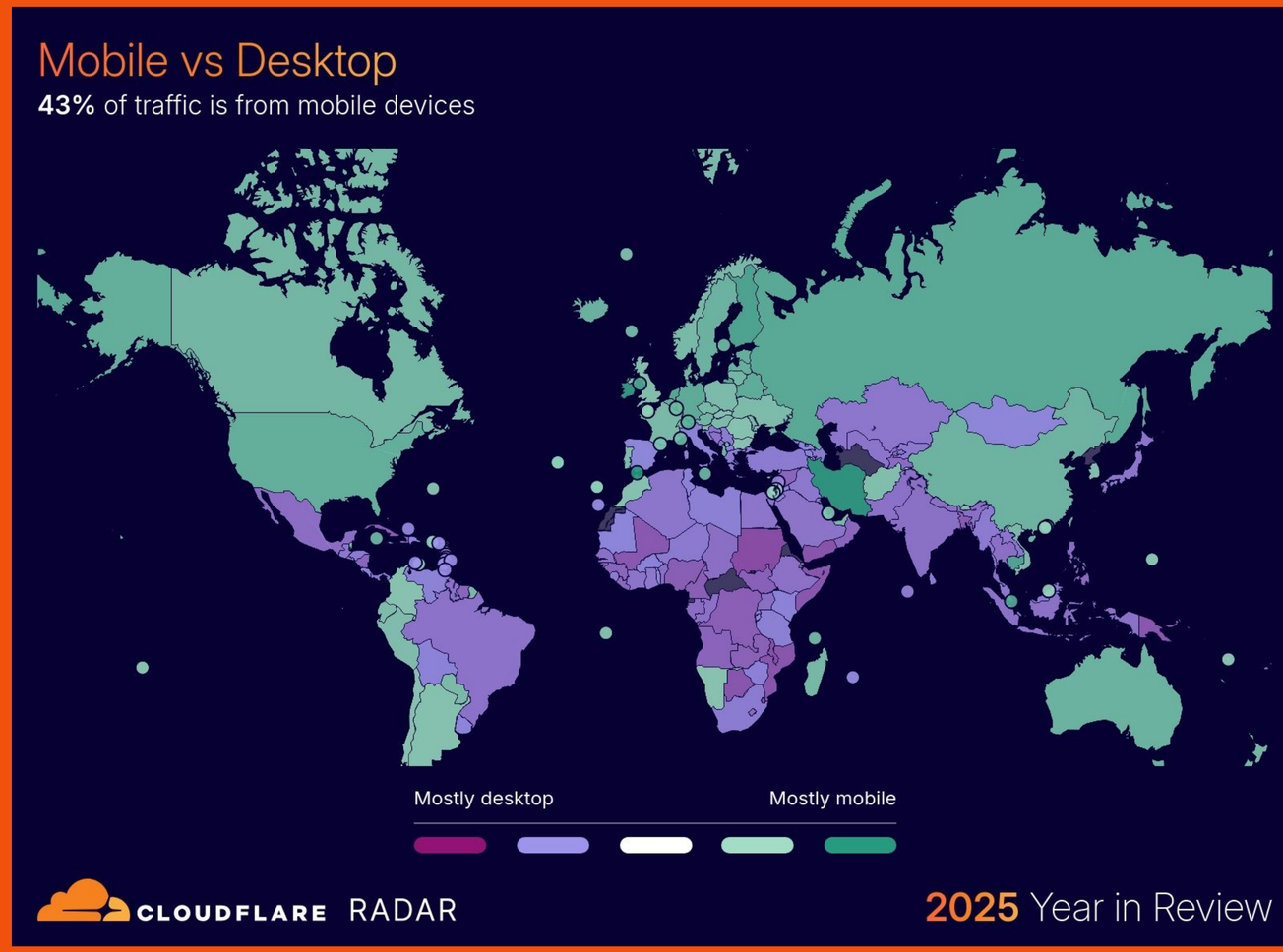
Business should provide value.
AI is a tool you might use to implement it.

- Most “AI companies” aren’t, they’re companies solving a problem using AI inside. The ones failing are usually the ones that inverted it.
- All Birds pivoting to "AI" didn't fix the shoes. If the problem doesn't need an agent, don't ship one.
- AI replaces/enhances a role/function in a business, there is still a cost/return.



Your agent isn't running on a desktop.

- Southeast Asia skipped the PC. Most users there have never opened your agent on a laptop.
- India, LatAm, Africa are mobile-first or mobile-only.
- Different regions have different AI sentiment. A US-centric "AI assistant" UX lands differently in Jakarta or São Paulo.

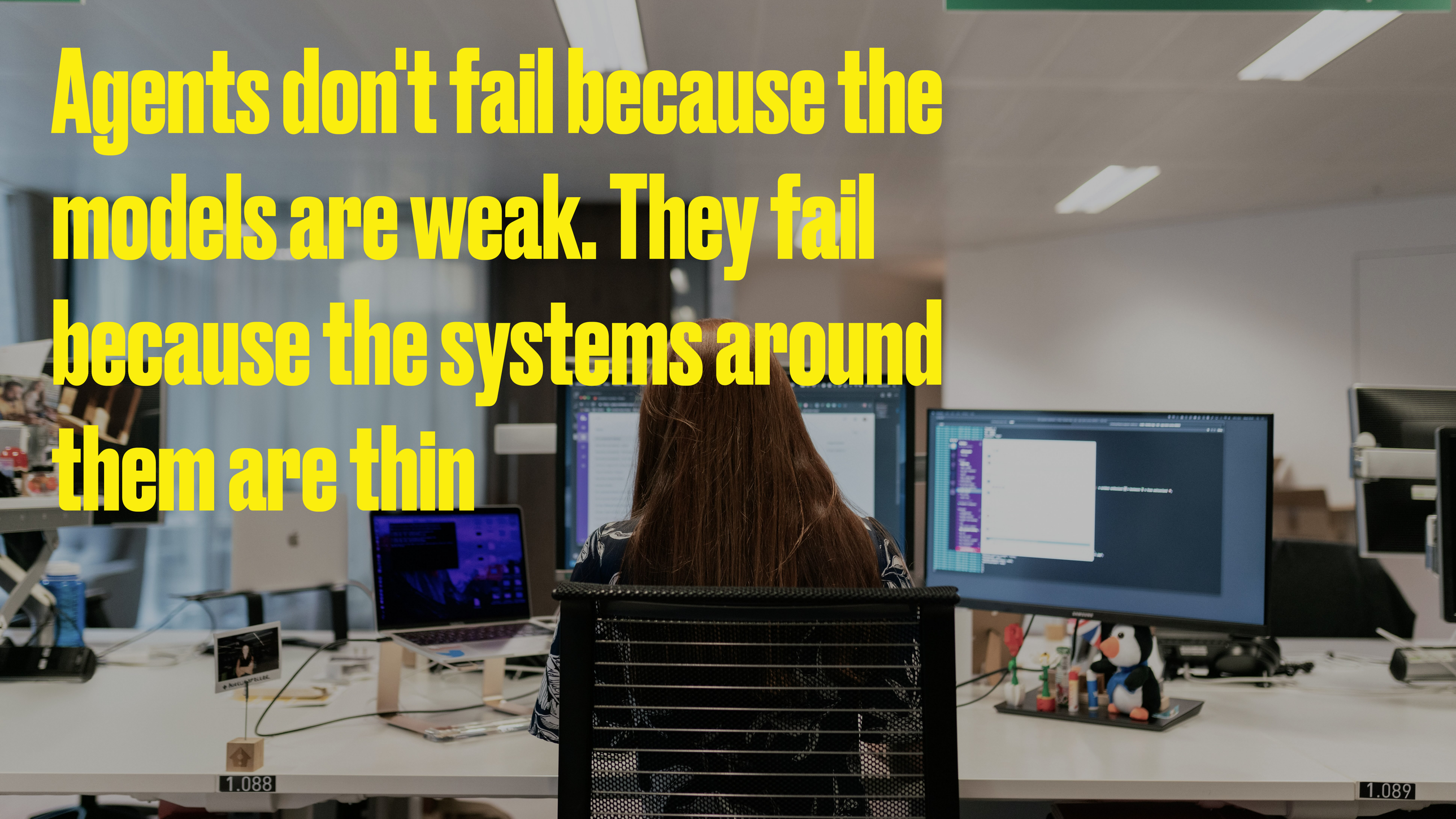


Agenda

Why most agents fail

AI Projects Aren't Working... Still

- 95% MIT NANDA: GenAI pilots with no P&L impact
- 80% RAND: enterprise AI projects that fail
- 88% Composio 2026: agent projects that never reach production
- 5% BCG Q1 2026: enterprises with measurable agent value


Agents don't fail because the models are weak. They fail because the systems around them are thin
1.088
1.089

The Sunk-Cost Fallacy in AI

- "There's a function at our business that's cheaper to have an agent do than a person."
- That sentence is true for about six months. Then: integration debt, eval debt, security debt, edge-case debt.
- By the time you measure the real cost, you've shipped, and the sunk-cost gravity makes it hard to turn around.

Diminishing Returns per Step

- Everytime you chain your agent with something there’s a chance it will be wrong.
- $0.85 \times 0.85 \times 0.85 \times 0.85 \times 0.85 \times 0.85 \times 0.85 \times 0.85 \times 0.85 \times 0.85 = 0.197 = 19.7\%$.

“AI gave us lower quality.”

Sebastian Siemiatkowski, Klarna CEO, Feb 2025

Agentforce, by Salesforce's own numbers

- A **52%** single-step success
- Drops to **32%** in multi-step success
- “Doom-prompting" is the new anti-pattern
- Salesforce's April 2026 fix **Agent Script.** A declarative DSL

“People don't want a quarter-inch drill. They want a quarter-inch hole!”

Theodore Levitt

The winners all look the same.

87%

resolution rate across 450M users

Mercado Libre customer support agent, LatAm's largest e-commerce platform

- ***Narrow scope*** - standardised support workflows.
- ***Validation gates*** - confidence thresholds and human escalation paths.
- ***Iteration*** - weekly evaluation to harness against production traces.

200-2000% productivity. Scoped right.

- Scoped - KYC checks, AML flagging, document matching. Bounded.
- Validated - every action reviewed against a rules engine.
- Human-on-edges - the **2%** flagged as uncertain goes to an analyst.
- Measured - dollars saved, hours returned, false-positive rate tracked weekly.

Three things every winner does

Scoped to a workflow with a clear success metric

Not "customer support" - “reset-password flow for free-tier users.”

Validation gates between agent actions and side effects

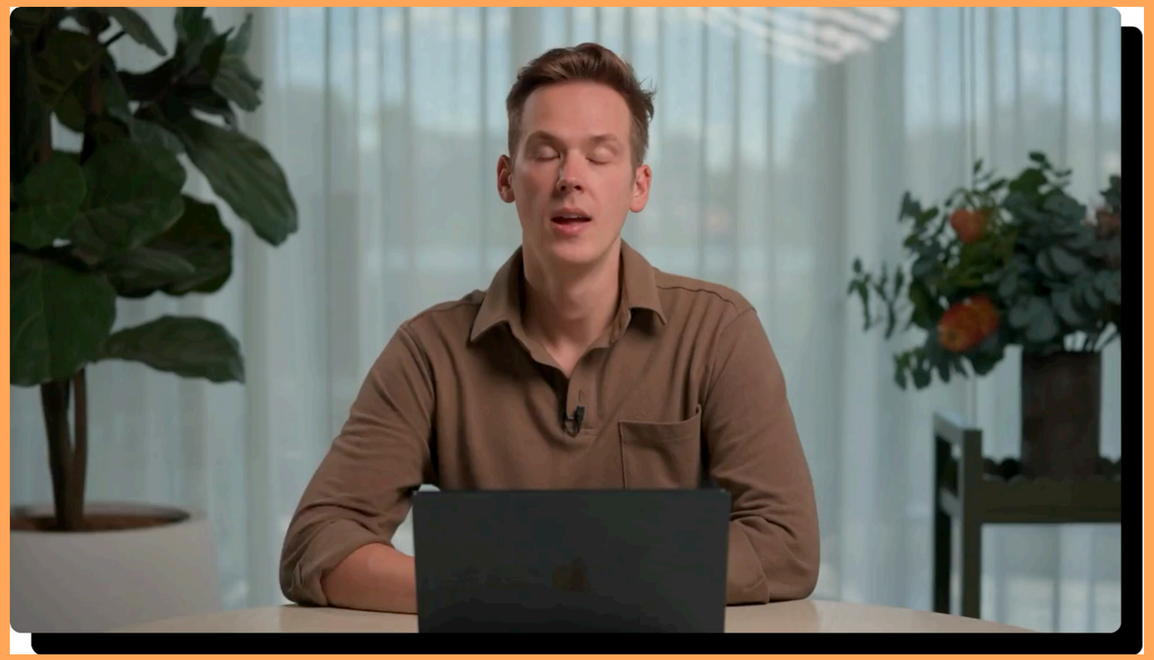
No write operation ships without a programmatic check or human signoff.

Eval harness that runs before every prompt change

Not "we'll write tests later." Tests before the prompt.

The Engineering

Eighteen months. Eight improvements. One agent that actually works.



Reduce complexity before you try anything else.

- Our prompts were complex. Simplifying the instructions works.
- Each successive agent call decreases accuracy, reduce agent chains.
- Software was too complex. Multiple calls running in a VM, split into single containers.

Context Engineering

- **Write/offload** - push finished work out of the context window.
- **Summarise/compress** - distil long histories before they force a refresh.
- **Isolate** - run subtasks in fresh contexts. Share context by communicating, not by sharing state.
- **Cache** - KV-cache hit rate is the real production cost metric.

Caching Economics

10× delta. Bigger than any other lever

- **$3** Claude Sonnet, per million uncached input tokens.
- **$0.30** Claude Sonnet, per million cached input tokens.
- Pre-rot threshold - that is accuracy drops before the context window is full.

- In our calls the longer the duration the higher the latency due to processing the prompt.
- Knowledge base input as context can slow down the model.

Tool Interfaces

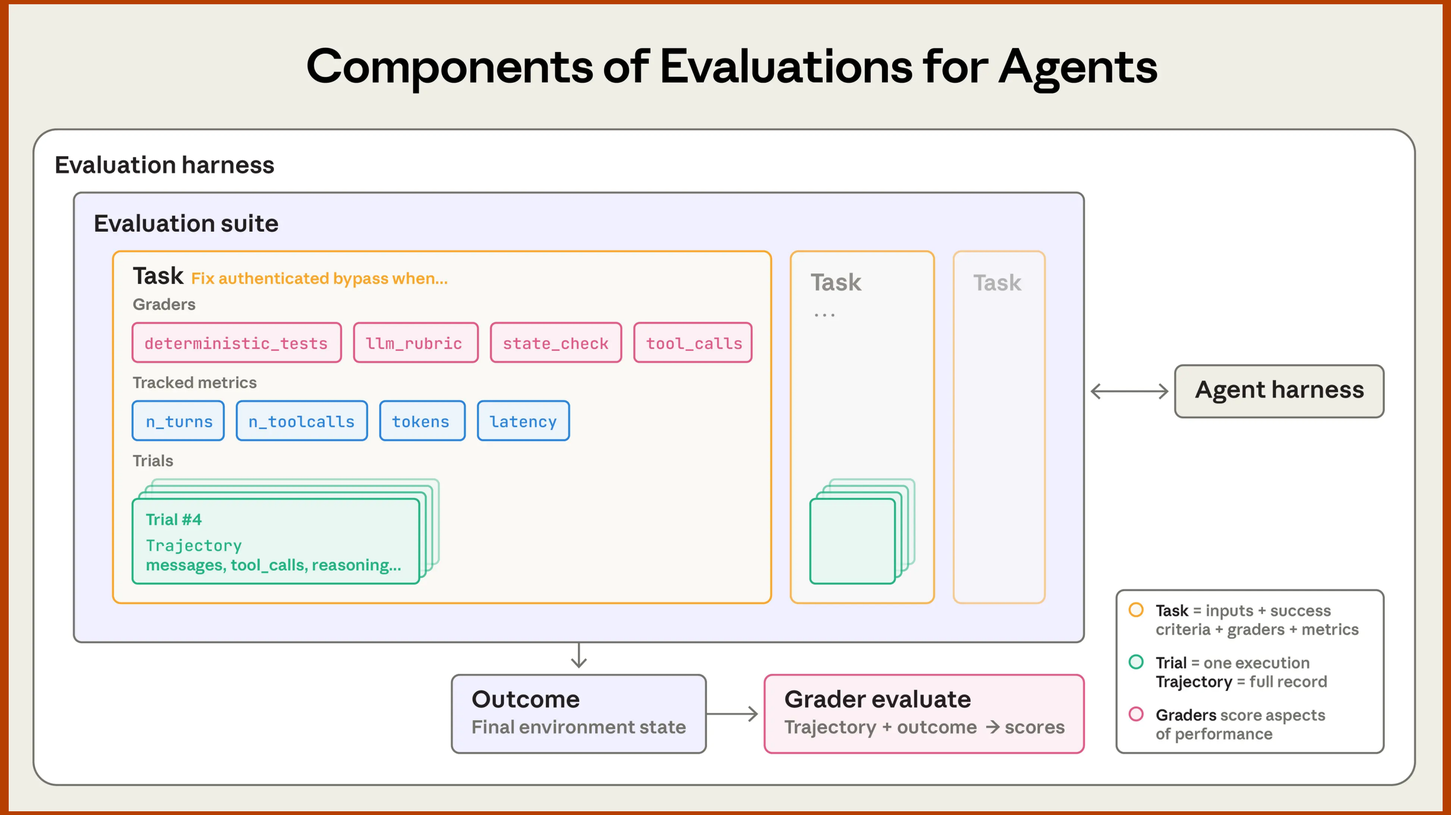
Design the ACI before you write the agent

- Agent-Computer Interface (ACI) names, shapes, error semantics of every tool the agent can call.
- Bad ACI: 40 tools, overlapping purposes, inconsistent errors.
- Good ACI: 6–12 tools, one-sentence contracts, errors the model can act on.

- Screen sharing - processing data in realtime, adaptive resolution, adaptive cropping.
- Memory management - each frame needs to be processed by the AI, if the time it takes a frame to be processed is longer than the time between frames causes OOM.

Agent Evaluations

No eval, no ship.



- Mistakes can propagate and compound
- Write the eval before the prompt. If you can’t name "right," you aren’t ready to build.
- Run the eval on every change. CI gate. Non-negotiable.
- Capture production traces as future eval data. Every bug is a missing test case

The 2026 Eval Toolbox

- **Arize Phoenix** - OTel-first, OSS, the Grafana of agents.
- **Braintrust/LangSmith** - managed SaaS, prompt-experimentation loops.
- **DeepEval/Langfuse** - OSS baselines, callable from .NET via HTTP.
- **Galileo Luna-2** - hallucination detection, compliance-grade.

Security

Non-human identity is the new OWASP top ten.

- **48.9%** of enterprises have no visibility into machine-to-machine API traffic (Salt Security, 2026).
- **88%** had an M2M security incident last year. **92.7%** in healthcare.
- **13.4%** of public agent skills contain exploitable injection paths (Snyk ToxicSkills).

- **Cursor IDE** (*Dec 2025*) - prompt injection via repo content - RCE on developer machine.
- **EchoLeak** (*2025*) - M365 Copilot indirect prompt injection exfiltrating M365 content.
- **Cline npm** (*Feb 17, 2026*) - supply-chain attack via malicious VS Code agent plugin.

Engineering TLDR

- **Scope** - cut steps, isolate fate
- **Context** - write / compress / isolate / cache
- **Tools** - design the ACI, don't rebuild what's battle-tested
- **Evals** - gate the prompt, trace the production
- **Security** - non-human identity is the attack surface

The Landscape

Framework Landscape

- **Microsoft Agent Framework 1.0** (*Apr 3, 2026*) - unified Semantic Kernel + AutoGen, .NET + Python, native MCP + A2A, LTS.
- **LangGraph**, **CrewAI**, **OpenAI Agents SDK**, **Claude Agent SDK**, **Google ADK**, **AWS Strands**
- **MCP** + **A2A** - Linux Foundation, 150+ orgs on A2A at one-year milestone. Pick **MCP** for tools, **A2A** for agents. Don't build your own.

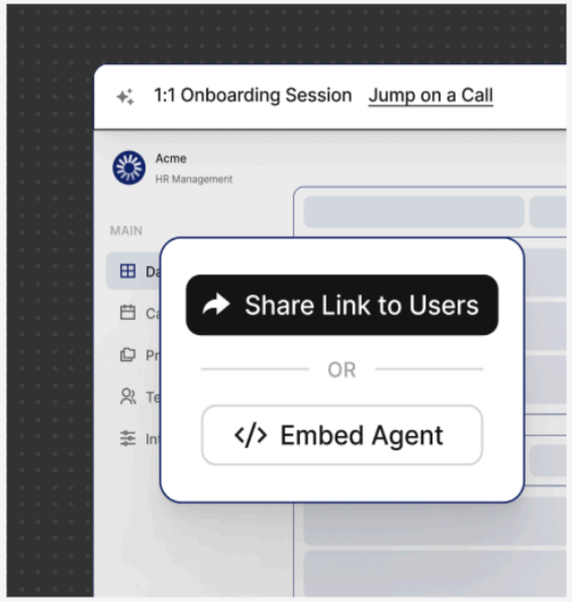
Three things to do when you get home

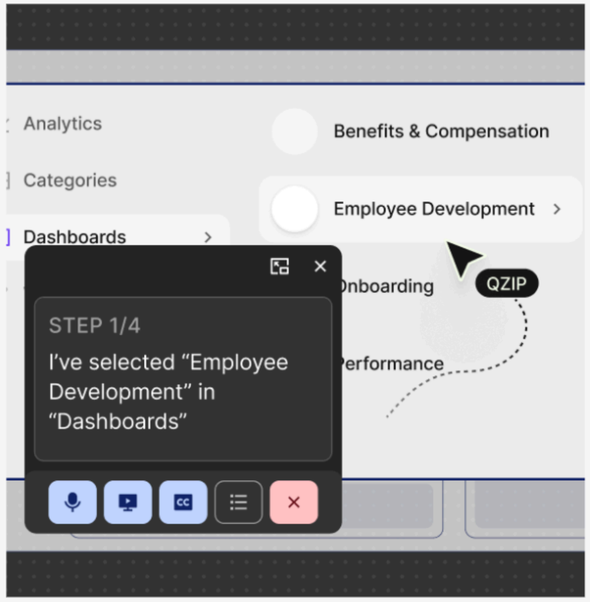
1. Pick one scoped workflow in your stack and ask: what are the steps, how many can I cut, what does "right" look like, which of my users are on mobile?
2. Write the eval before the prompt. Twenty lines of C# or Python. Manual is fine.
3. Answer out loud: what's the business value if this works and am I an AI business, or am I a business that uses AI?

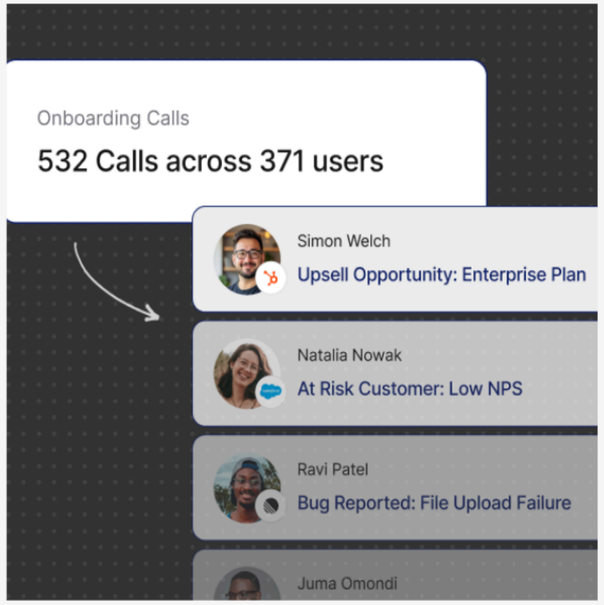
Quarterzip automates follow up and customer interaction

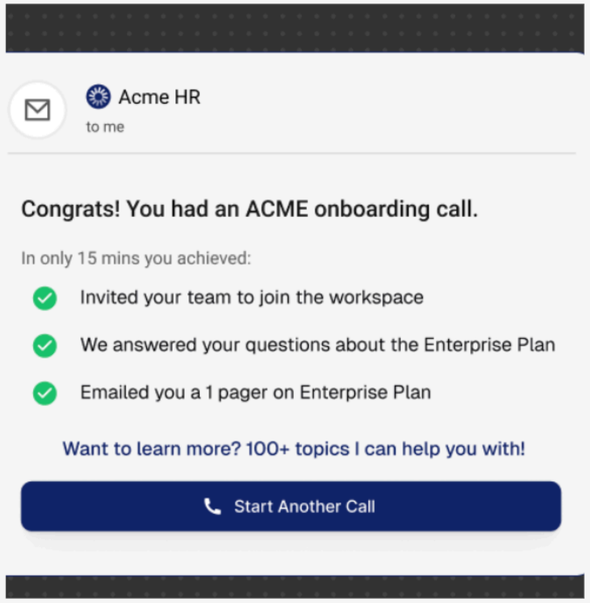
Connect with your customers anywhere.

In product or out.









Leverage our embeddable link or SDK to integrate Quarterzip into your messaging and/or product to start connecting with customers seamlessly.

We step in to handle the hard parts, guiding users through their journey. Actions can be done via the SDK or via MCP integration.

Automatically push insights into Hubspot and Salesforce and never miss an opportunity again. Provide **product teams** clarity on bugs and feature opportunities with real customer voice & screen actions.

Take customers on the journey. We create multiple touchpoints & push **account expansion** when paid features would help users meet their goals.

Successfully activating users across many industries

Our performance on average:

20 Min+
Avg time users spend with Quarterzip agents

11x ROI
Average customer ROI

87%
Average cost reduction vs human-led onboarding

1.8x
Calls per users, indicating return usage throughout and post onboarding.

33+
Supported languages our agents speak.

94%
Calls are multipurpose. 57% cover onboarding along with technical support, sales enablement and feature discovery

Apollo

Sales Intelligence

"Quarterzip changes the math on activation. We can give every user a guided, personalized experience without throwing headcount at the problem."

James Boone
Sr. Director, Customer Onboarding & Engagement

Grab — Transport and Food Delivery

eleos — Health-Tech

Verifast — Digital Verification

VARICON — Construction

bloomreach — Marketing Automation

+ more

Thankyou!

Forward Deploy Engineering Lead

@kochie
linkedin.com/in/rkkochie

NDC Sydney 2026