Practical Lessons from Deploying Real-World AI Agents

Clay Bavor
clay@sierra.ai

SIERRA

Meet Sierra

What we do

Sierra helps businesses build better, more human customer experiences with AI.

Our values

Trust

Customer Obsession

Craftsmanship

Intensity

Family

Who we are

Engineers, researchers, and executives from leading technology institutions

Bret Taylor

Co-Founder

Chair, OpenAI

Clay Bavor

Co-Founder

Previously VP, Google Labs

salesforce Google OpenAI

slack Meta PRINCETON UNIVERSITY

AG1

AOL

Casper

MADISONREED

minted.

NEXT

Pendulum

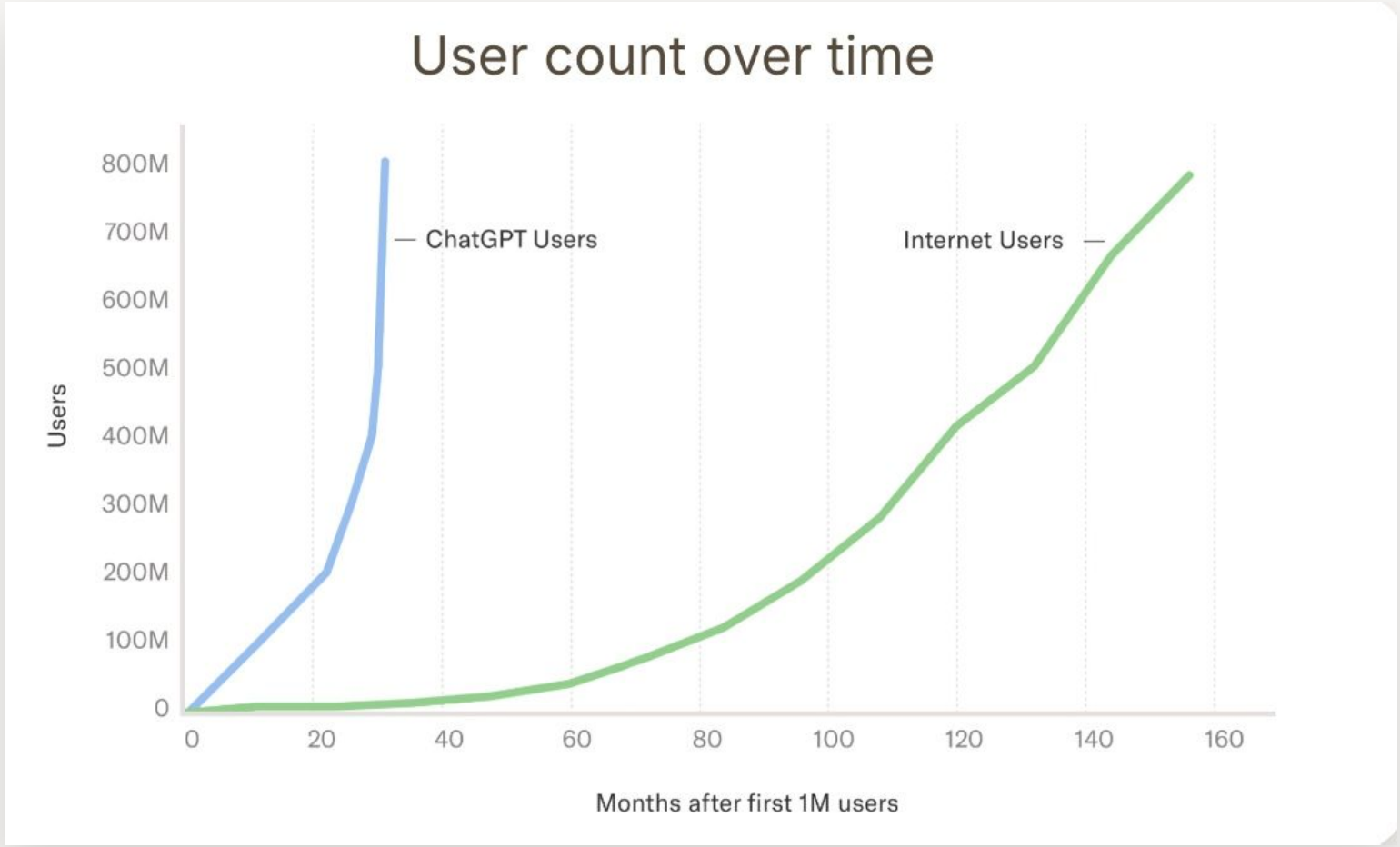
SONOS

sweetgreen

Thirdlove®

WeightWatchers

AI: the Internet movie on 5x the speed



New technology creates new experiences

Your order
has been placed.
ETA Thursday

Internet → Websites

1995

Mobile → Apps

2015

AI → Agents

2025

From a multi channel world to

single agent world

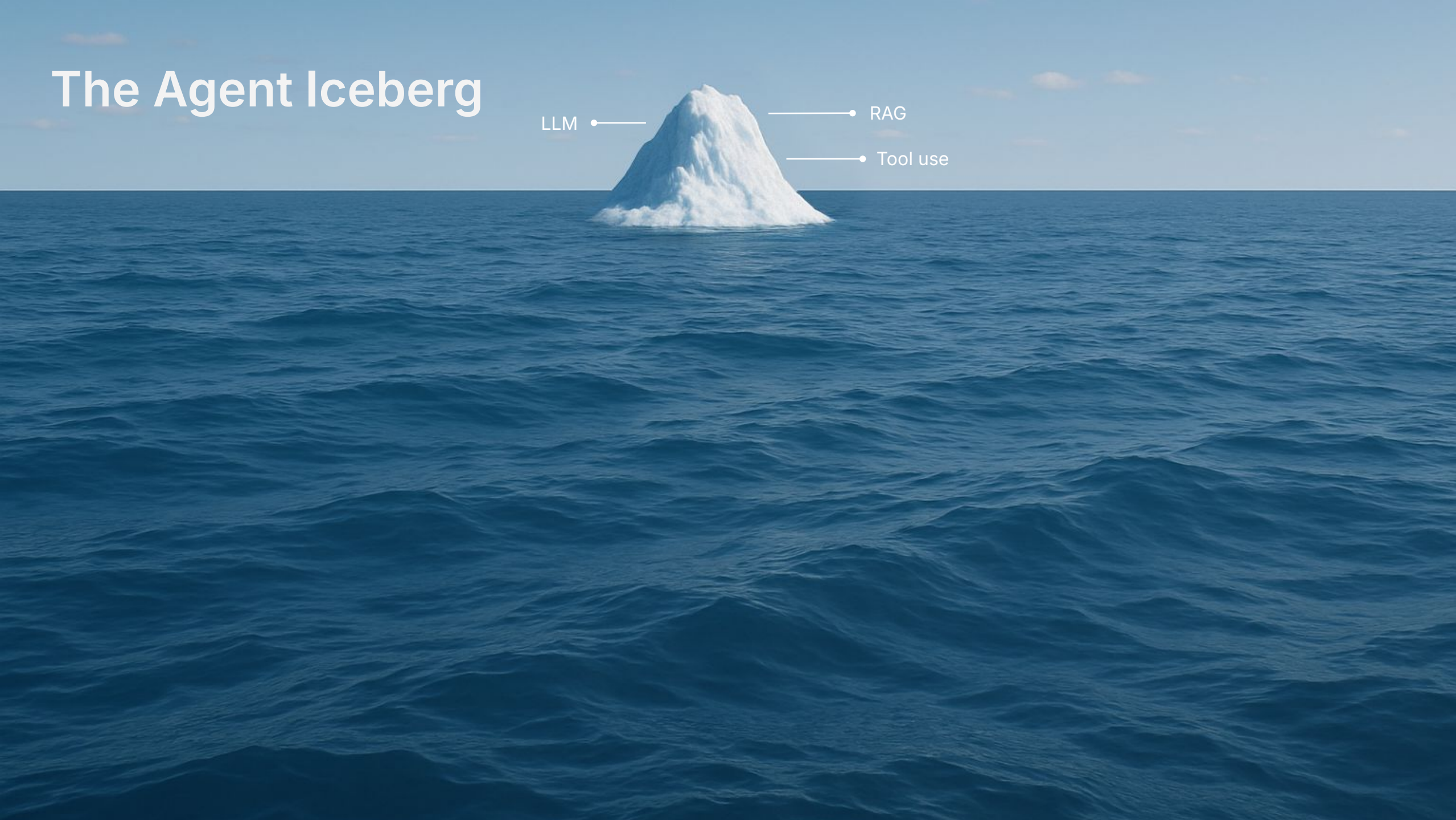
The conversation

is the interface

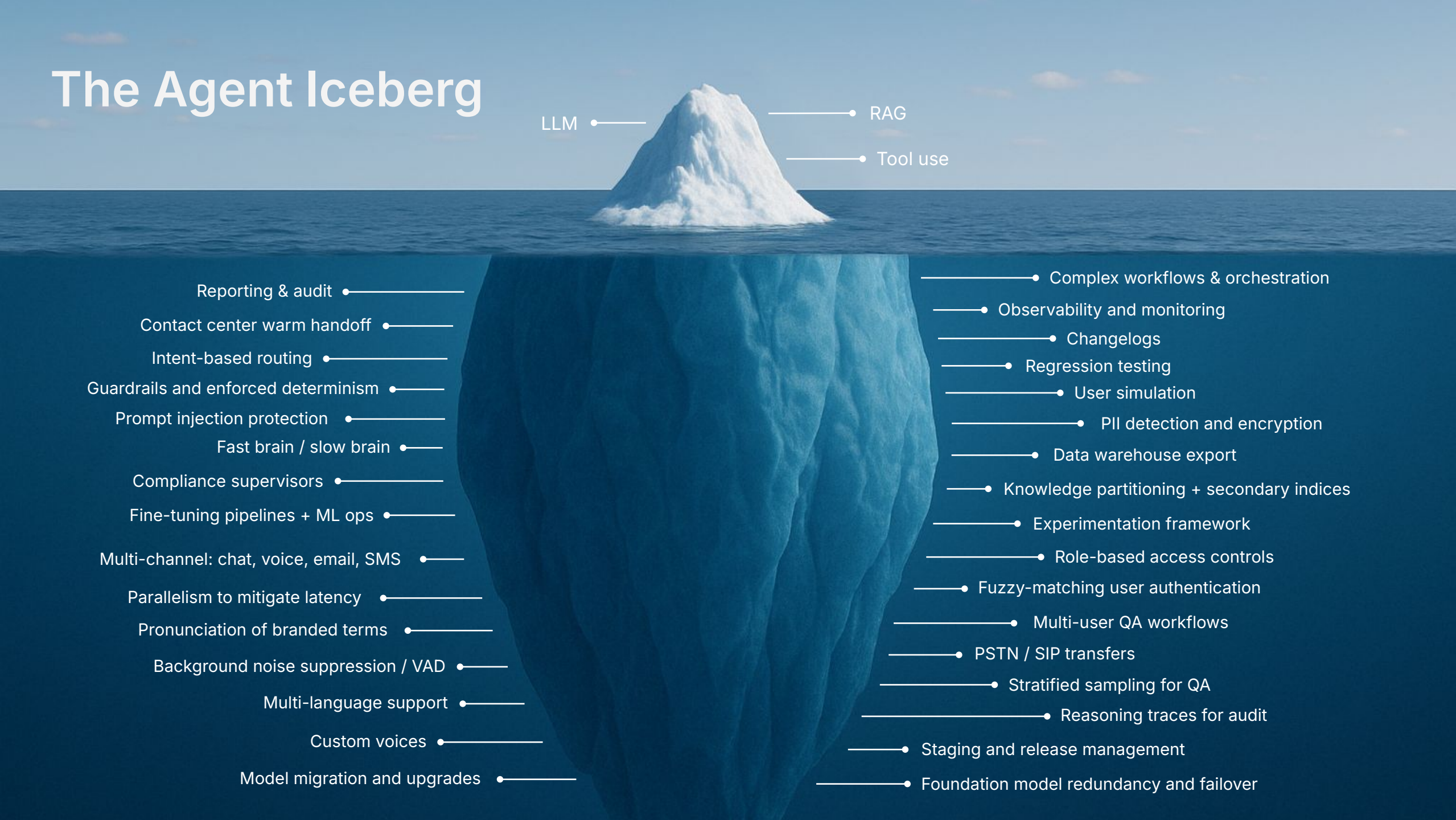
Pay for

a job well done

The Agent Iceberg



The Agent Iceberg

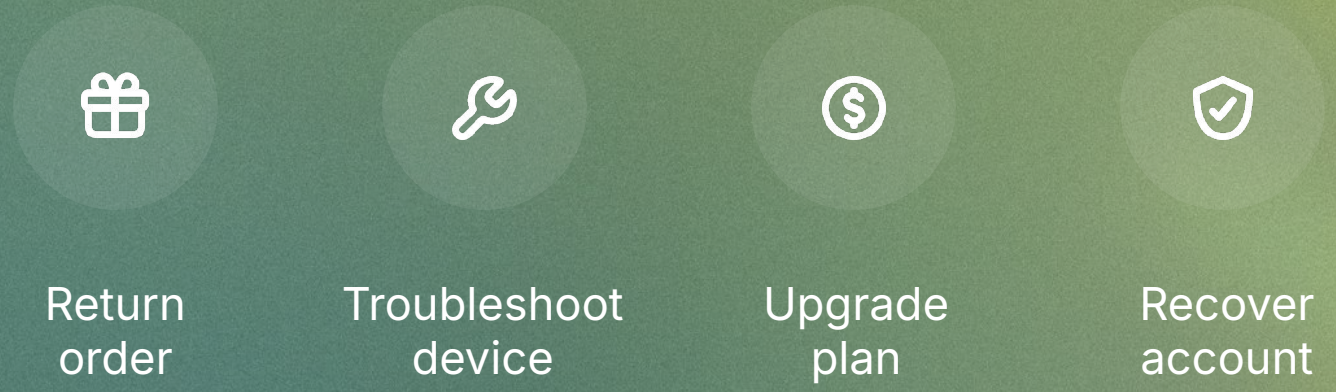


We’re in the 1997 era
of building agents

Moving from agents as technology to agents as product

Simple, not simplistic

Today’s agents are transactional

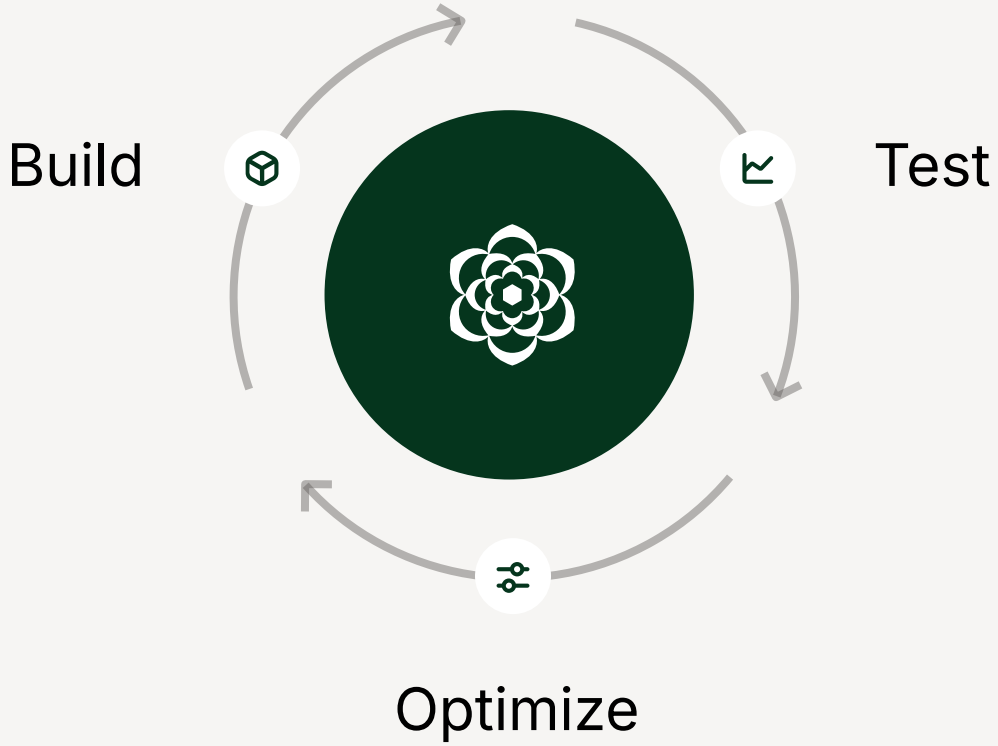


The best agents don’t resolve cases,

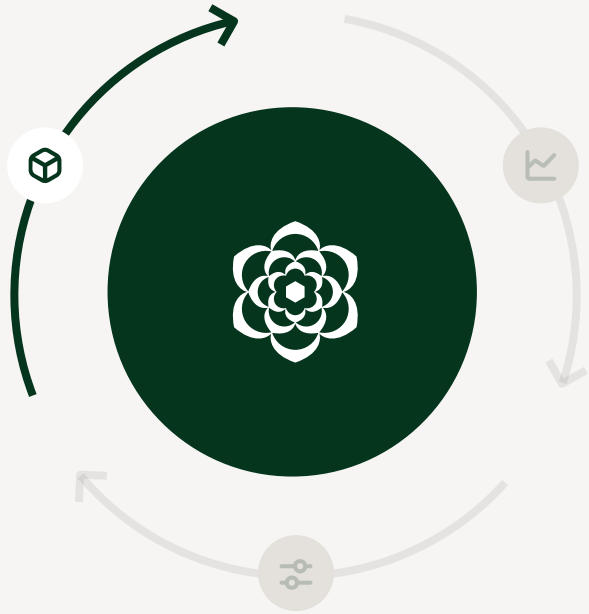
they build relationships

- Engage over time on multiple channels
- Remember your interactions
- Integrate all your enterprise data
- Proactive, not reactive

Agent Development Life Cycle



Build with



Versus build or buy

Every channel is digital

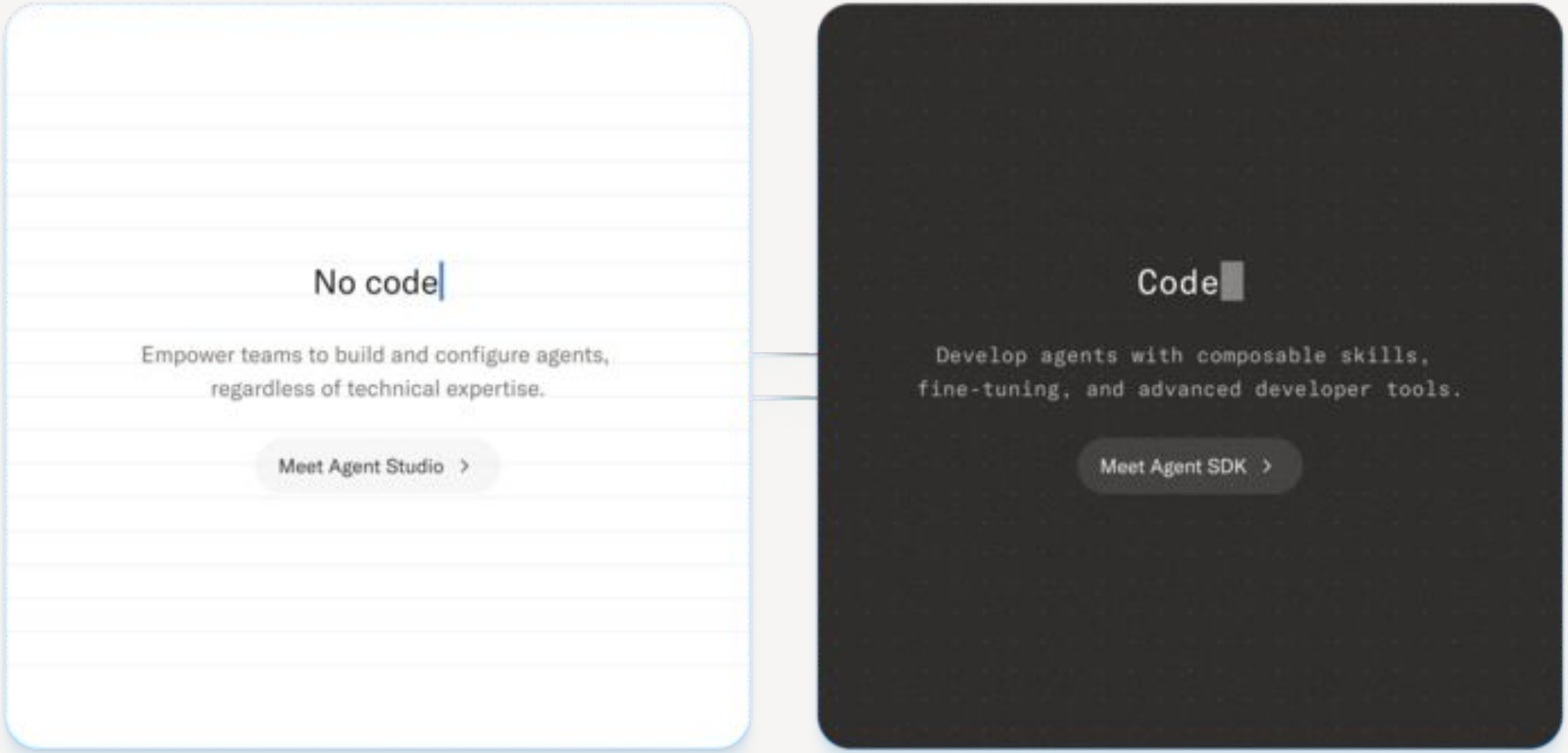
Build once deploy everywhere

Phone

Email

Chat

Technology → product



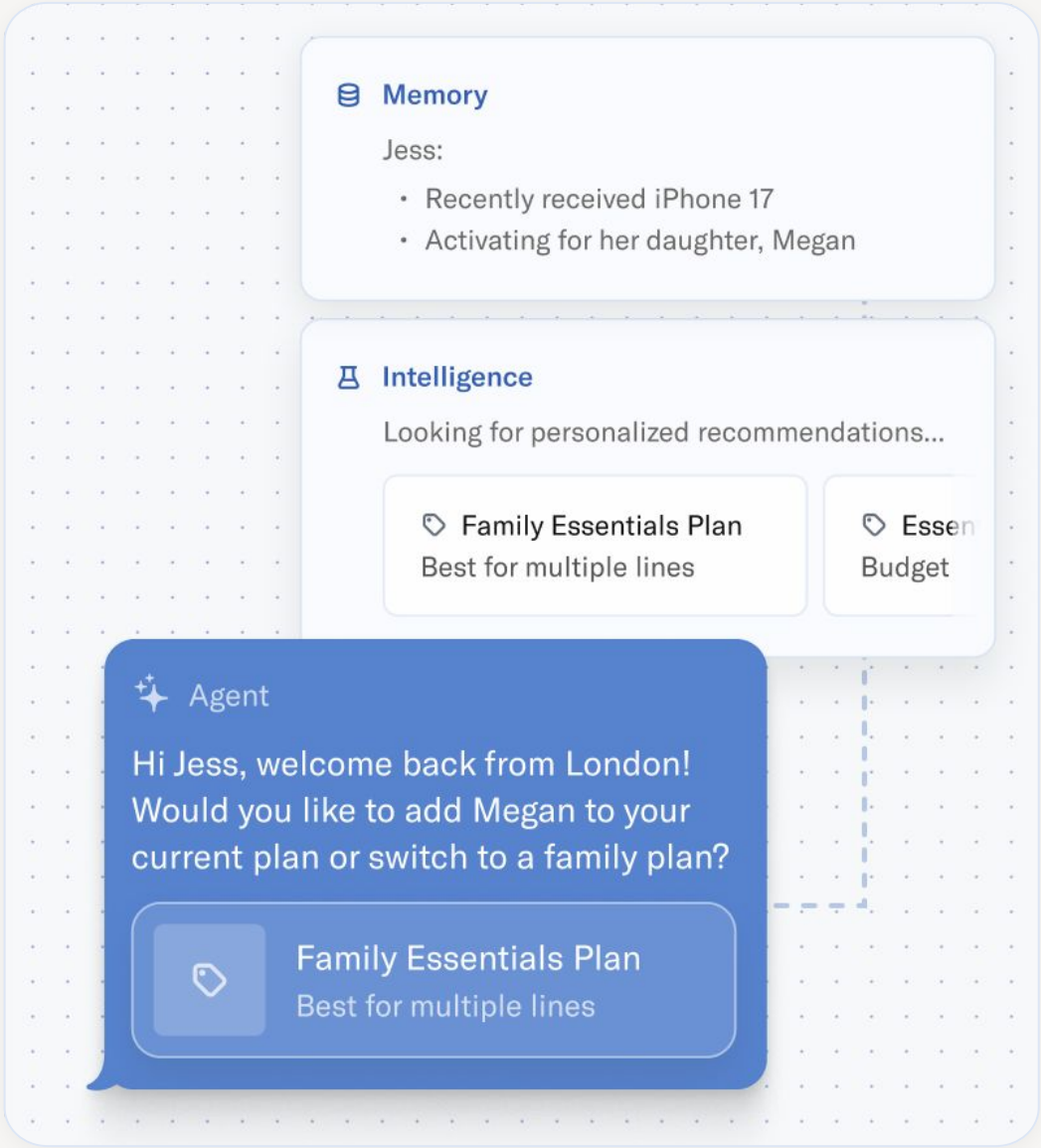
Agent Data Platform

Agent Memory

Customer Data Platform

Intelligent Decisioning

Proactive Engagement



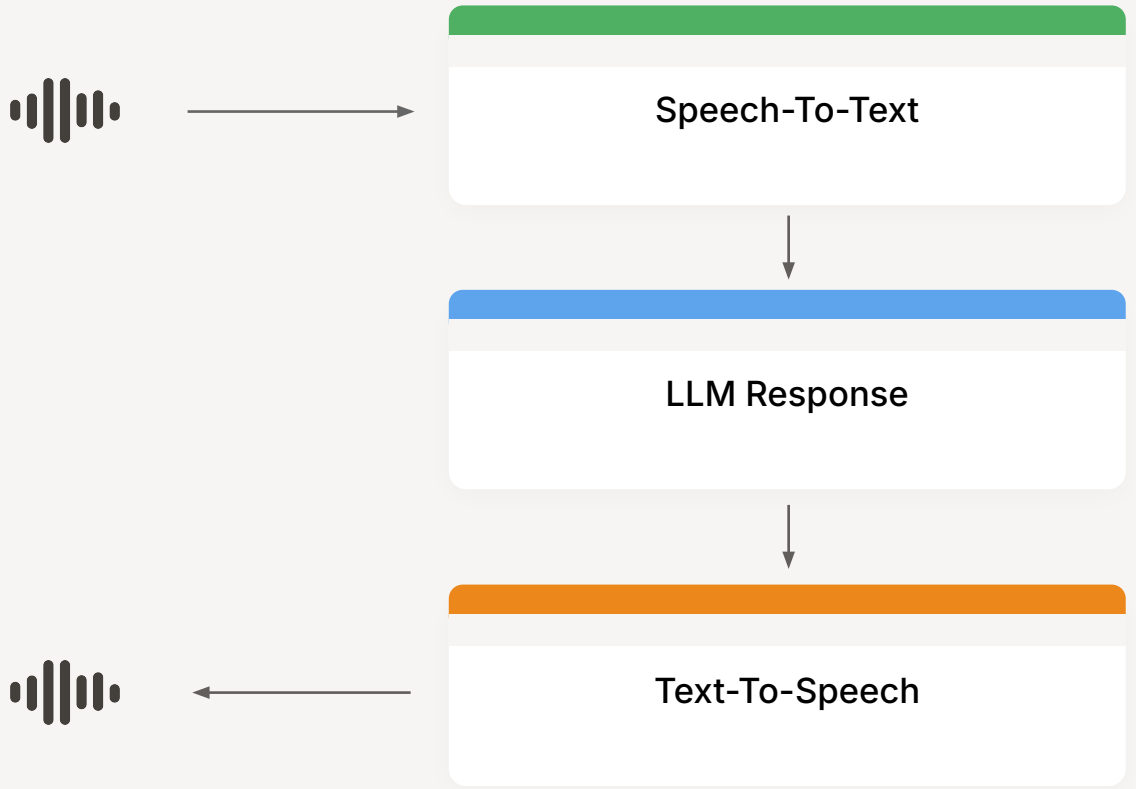
Voice architecture overview

Transcribe
Convert incoming audio to text.

Respond
Reason about the text, use business configuration, respond with text.

Synthesize
Construct audio signal from the text.

Overall challenge: production readiness
Easy to create demos, harder to make everything work consistently, with high quality and at scale.



Voice transcription

Quality measurement
WER as a metric is terrible for transcription quality in voice agents.

Audio input issues
Multiple speakers. Background TV / radios / far-field speech. Off-the-shelf solutions are generally poor.

missed	bad	acceptable	good	composite_score	wer
0.008621	0.245690	0.206897	0.538793	2.275862	0.120690
0.058608	0.267399	0.227106	0.446886	2.062271	0.340659
0.042553	0.340426	0.312057	0.304965	1.879433	0.134752

Voice Synthesis

Natural entity synthesis
Accurately synthesize addresses, phone numbers, websites by locale. Lost signal from input audio for names. How do you synth “Andrea”?

Prosody
Maintain a natural prosody across sentences.

Phrase quality
Large impact on perceived tone. Brevity is important.

Emotive range
Content should match tone, and tone should be appropriate to conversation.

SIERRA

What's in a voice?

Age

Adult, Middle-Aged, Old...

Gender

Female, Male, Nonbinary

Base Accent

American, British, French-Canadian...

Tone

Smooth, Clear, Soft, Smoky...

Pitch

Soprano, Alto, Tenor, Baritone...

Intonation

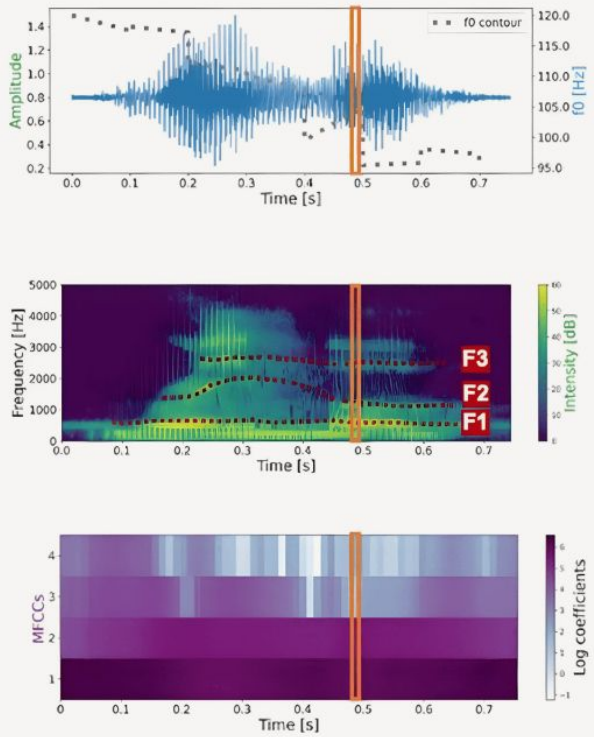
Conversational, Professional, Urban...

Speed

Fast, Quick, Slow, Relaxed...

Emotion

Upbeat, Calm, Assertive...



Regional Accent

Rhythm

Range

Dips

Lifts

Volume

Enunciation

Nasality

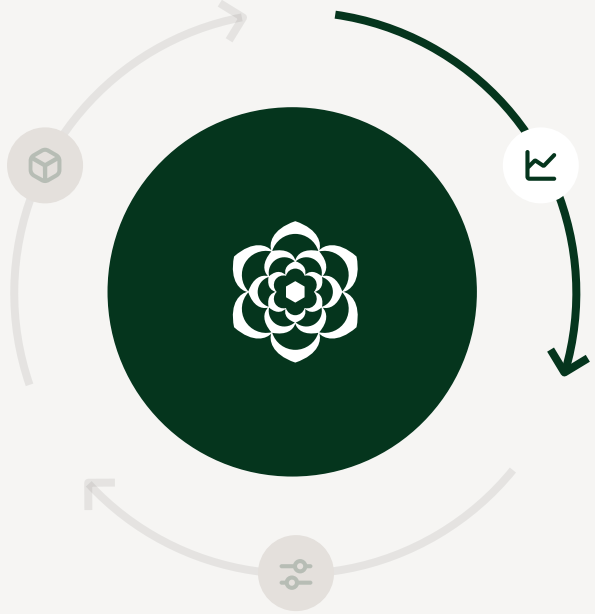
Breathiness

Constriction

Gravel

Fry

Test



New testing for new software

Back in the dark ages of May 2023,
our first user + agent simulator...

Agents in the real world

The Challenge: Realistic Testing

Beyond reasoning or tool use — agents must manage:

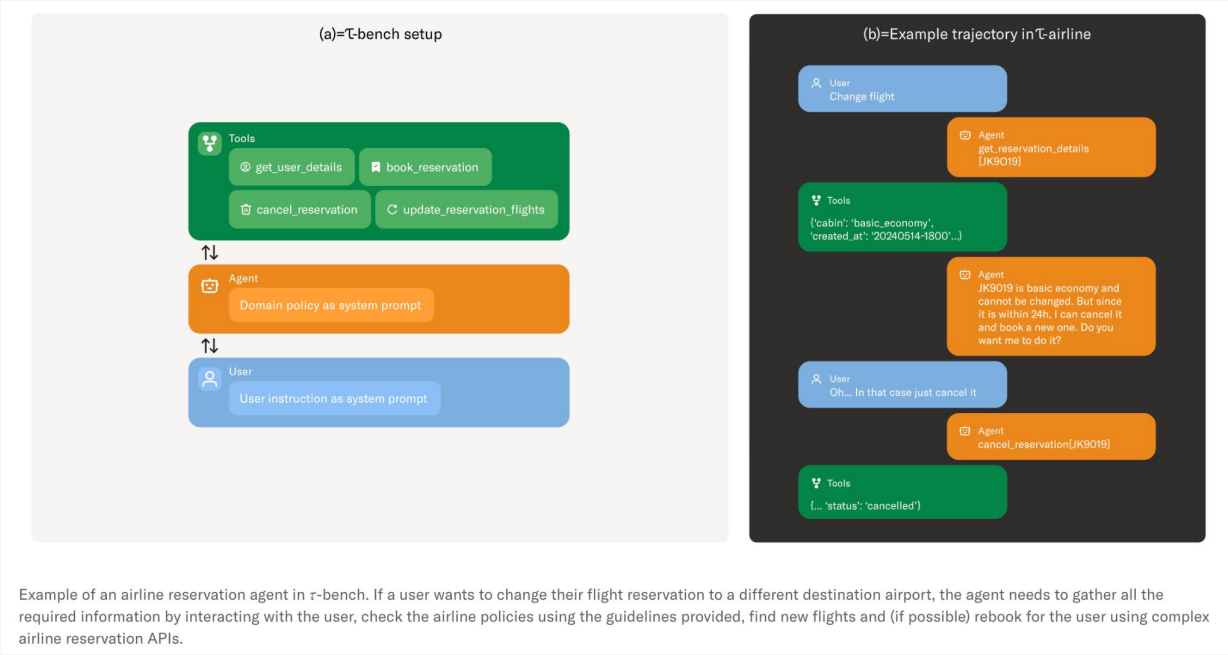
- Tool + User interaction
- Policy adherence and guardrails
- Dynamic, multi-turn conversations

Enter τ-Bench

Evaluate agents in *realistic*, *holistic* scenarios

Core Requirements for Real-World Reliability

1. Seamless interaction
2. Policy adherence
3. Consistency at scale



Example of an airline reservation agent in τ-bench. If a user wants to change their flight reservation to a different destination airport, the agent needs to gather all the required information by interacting with the user, check the airline policies using the guidelines provided, find new flights and (if possible) rebook for the user using complex airline reservation APIs.

Inside τ-Bench — Building Reliable Evaluation

1. Realistic Domains

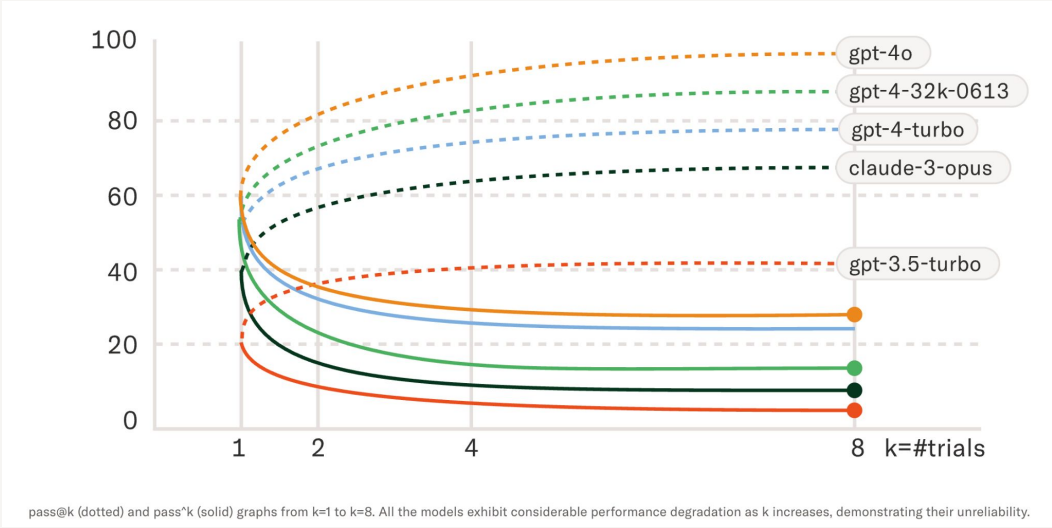
- **Complex Databases & APIs:** JSON + Python API tools for reading/writing data.
- **Domain Policies:** Documents defining rules and restrictions agents must obey.

2. Realistic & Reliable User Simulator

- **LLM-based user simulation** for authentic dialogue.
- **Dual-control environment:** Both user and agent can act via tools.
- **Reliability:** Structured task definitions anchored in a shared environment.

3. Objective Evaluation

- **Well-defined tasks**
- **Rule-based success check:** Compare final database state to ground truth.
- **Pass^k metric:**
  - Measures chance all *k* trials succeed.
  - Tests reliability under conversational variability.



pass@k (dotted) and pass^k (solid) graphs from k=1 to k=8. All the models exhibit considerable performance degradation as k increases, demonstrating their unreliability.

pass@k (dotted) vs pass^k (full)
When an agent handles millions of conversations with customers, we care about reliability!
(from 2024 paper)

Goal:

Ensure agents can **reliably, safely, and effectively** interact with a customer in real-world, policy-constrained environments.

An Industry Standard for the Whole Community

Impact and Adoption

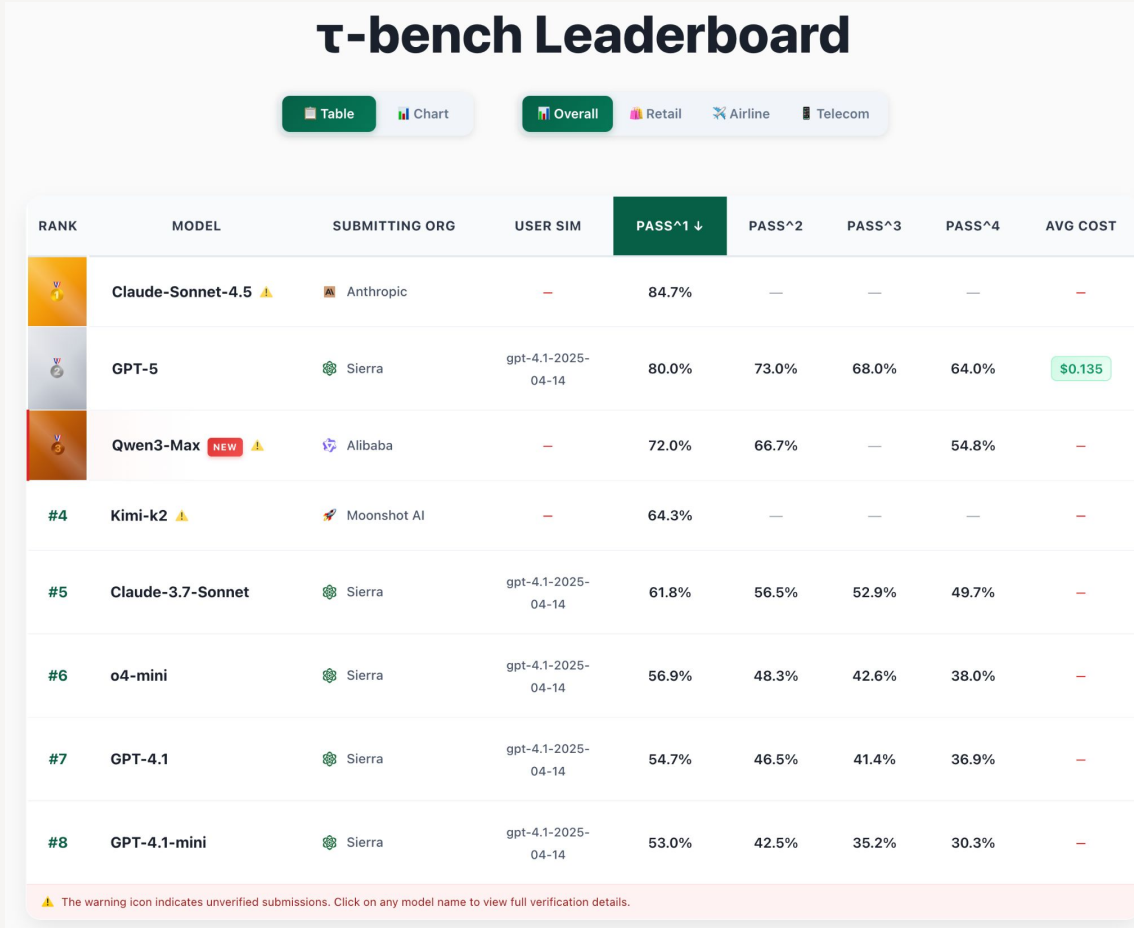
- τ-bench (and its upgrade τ^2-bench) has rapidly become the **standard for evaluating agent success, cited by major AI labs like Anthropic and OpenAI in model releases.**
- LLMs have significantly improved! Open Source LLMs are catching up!

Evolving the Benchmark

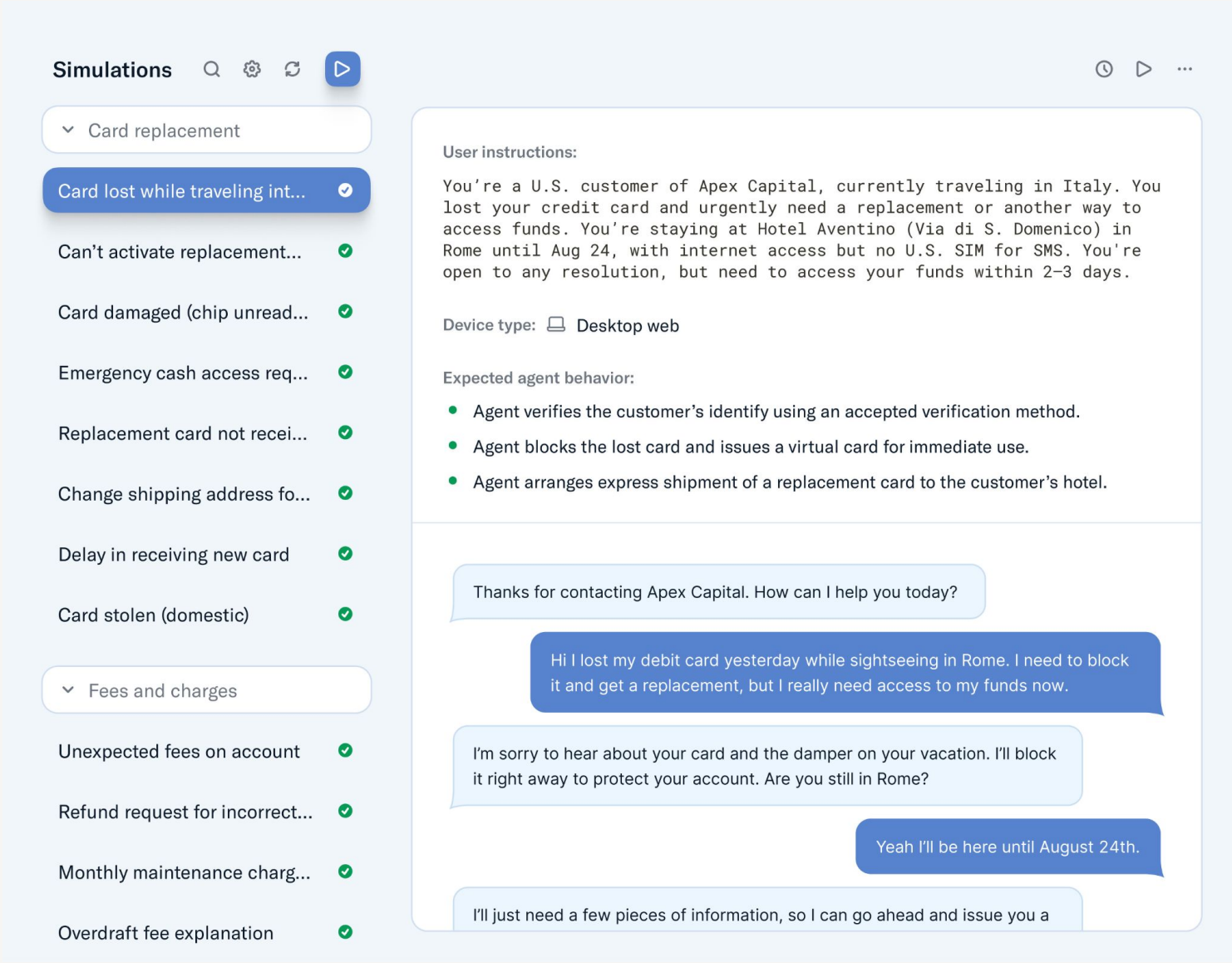
- **Transparency & Community**: The τ-bench leaderboard provides a focus point to track progress. And the τ^2-bench repo facilitate community contribution.
- **Want to contribute? Join the Tau-Bench track for the MOOC's Hackathon!**

From Research to Production

- τ-bench defines the **essential ingredients** of conversational agent evaluation.
- At Sierra, we build on this foundation, recognizing that **testing in production requires much more** — continuous, integrated systems that evolve with our agents.

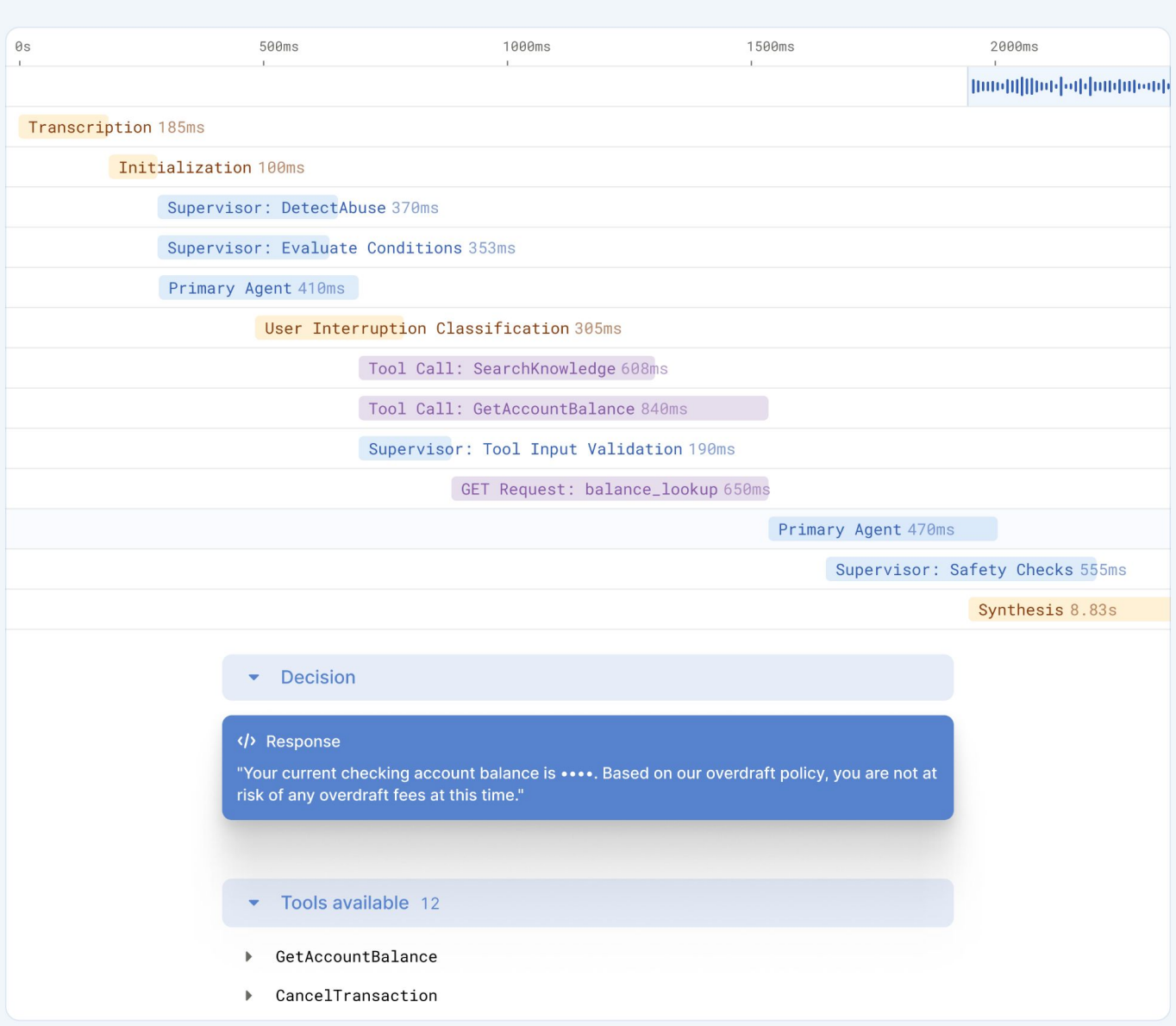


Sims

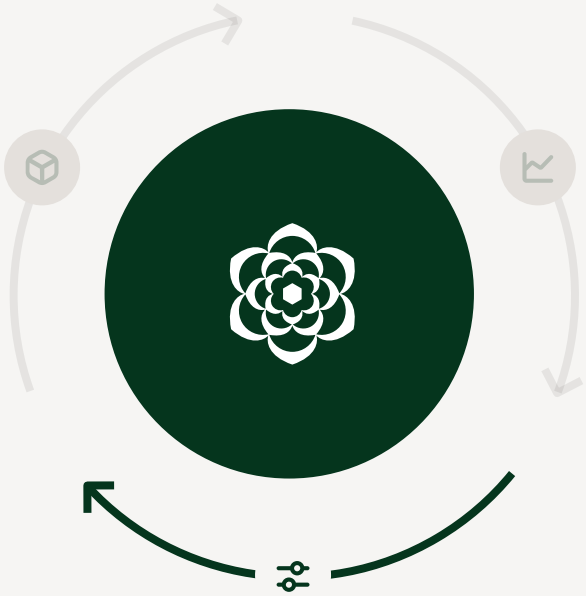

Simulations
Card replacement
Card lost while traveling int...
Can't activate replacement...
Card damaged (chip unread...
Emergency cash access req...
Replacement card not recei...
Change shipping address fo...
Delay in receiving new card
Card stolen (domestic)
Fees and charges
Unexpected fees on account
Refund request for incorrect...
Monthly maintenance charg...
Overdraft fee explanation
User instructions:
You're a U.S. customer of Apex Capital, currently traveling in Italy. You lost your credit card and urgently need a replacement or another way to access funds. You're staying at Hotel Aventino (Via di S. Domenico) in Rome until Aug 24, with internet access but no U.S. SIM for SMS. You're open to any resolution, but need to access your funds within 2–3 days.
Device type: Desktop web
Expected agent behavior:
Agent verifies the customer's identify using an accepted verification method.
Agent blocks the lost card and issues a virtual card for immediate use.
Agent arranges express shipment of a replacement card to the customer's hotel.
Thanks for contacting Apex Capital. How can I help you today?
Hi I lost my debit card yesterday while sightseeing in Rome. I need to block it and get a replacement, but I really need access to my funds now.
I'm sorry to hear about your card and the damper on your vacation. I'll block it right away to protect your account. Are you still in Rome?
Yeah I'll be here until August 24th.
I'll just need a few pieces of information, so I can go ahead and issue you a

Voice Sims compilation

Traces


0s
500ms
1000ms
1500ms
2000ms
Transcription 185ms
Initialization 100ms
Supervisor: DetectAbuse 370ms
Supervisor: Evaluate Conditions 353ms
Primary Agent 410ms
User Interruption Classification 305ms
Tool Call: SearchKnowledge 608ms
Tool Call: GetAccountBalance 840ms
Supervisor: Tool Input Validation 190ms
GET Request: balance_lookup 650ms
Primary Agent 470ms
Supervisor: Safety Checks 555ms
Synthesis 8.83s
Decision
Response
"Your current checking account balance is ••••. Based on our overdraft policy, you are not at risk of any overdraft fees at this time."
Tools available 12
GetAccountBalance
CancelTransaction

Optimize

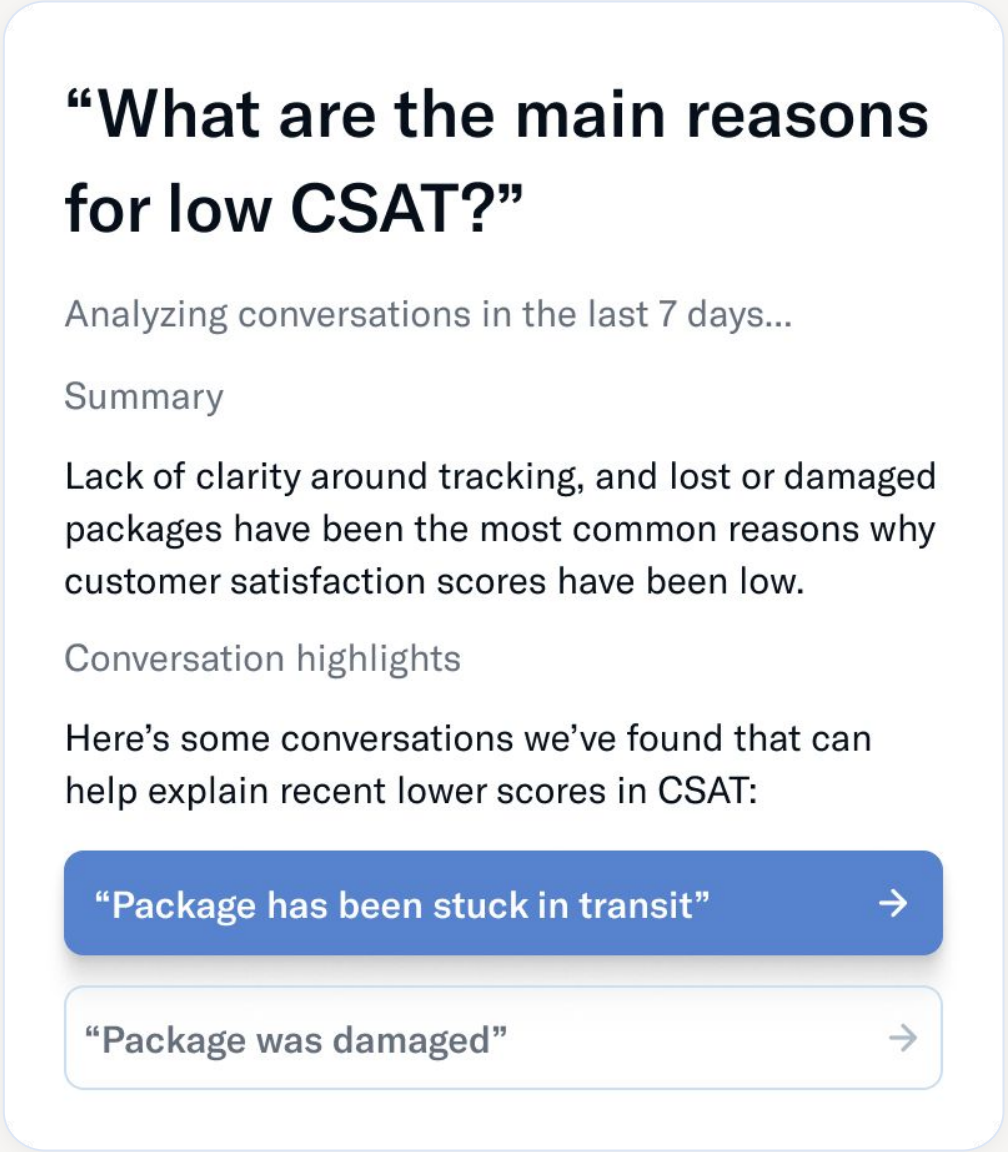


Use AI to improve AI

Insights

Explorer

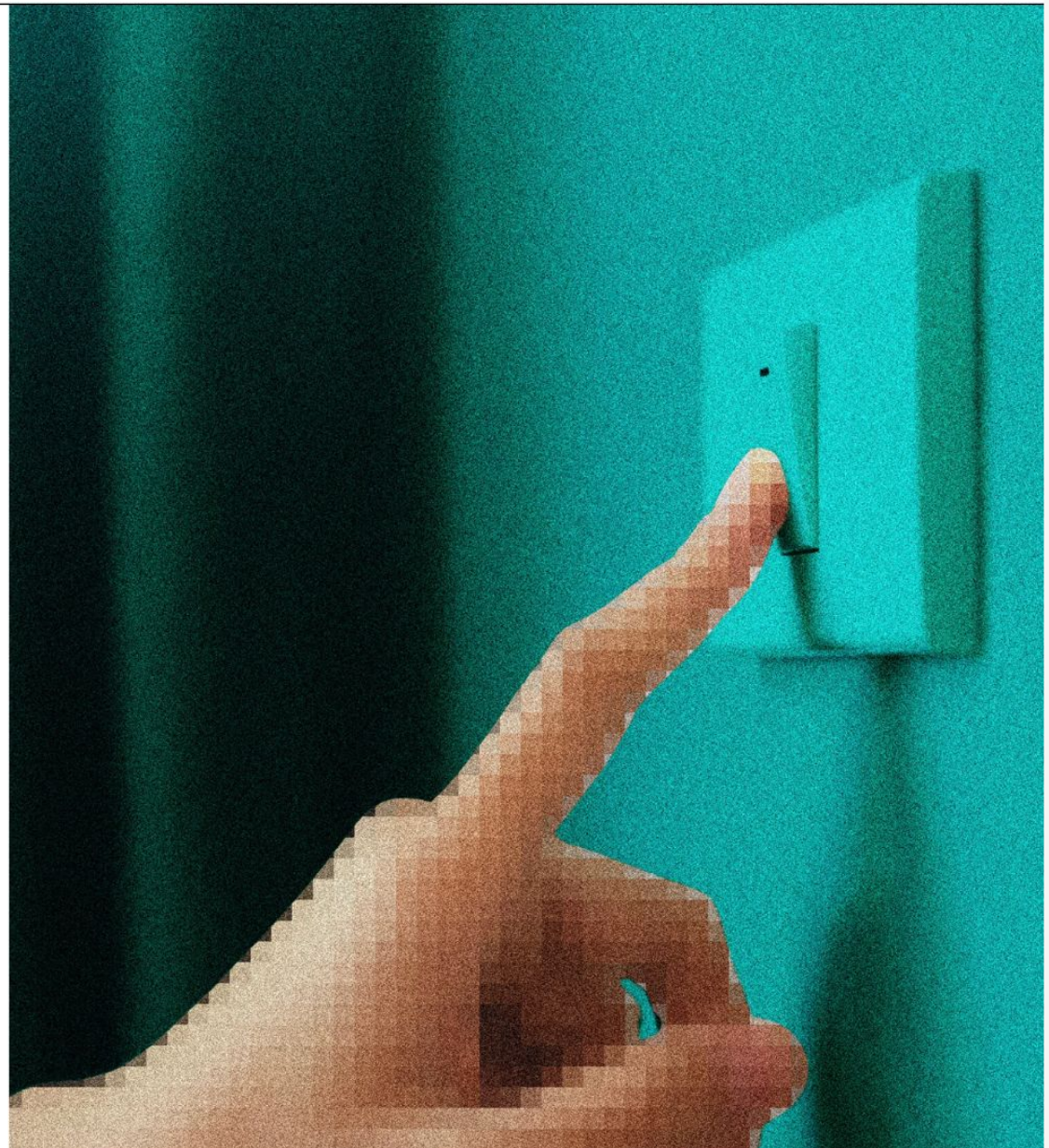
Expert Answers

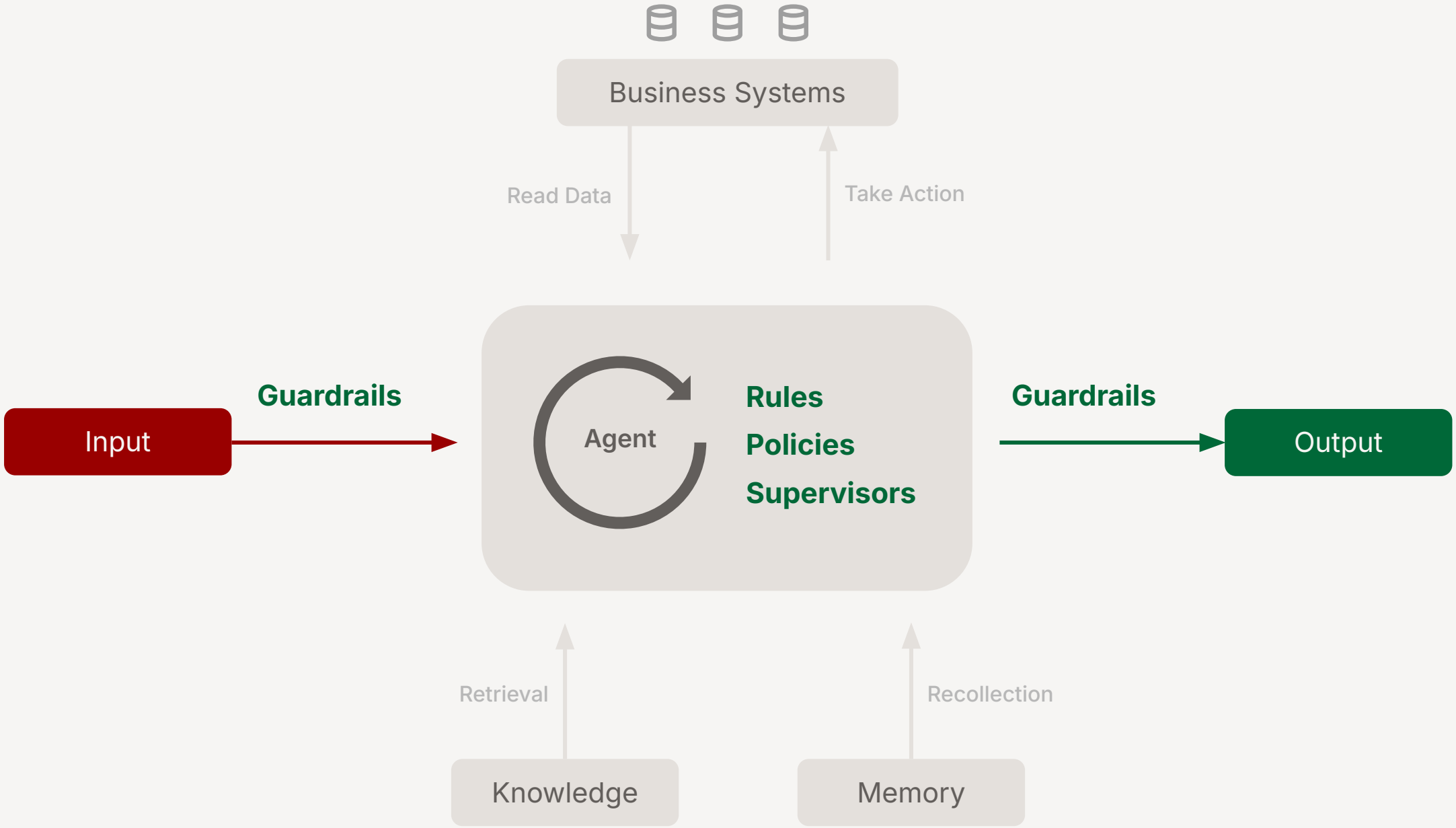


MATT BURGESS SECURITY AUG 6, 2025 9:00 AM

Hackers Hijacked Google's Gemini AI With a Poisoned Calendar Invite to Take Over a Smart Home

For likely the first time ever, security researchers have shown how AI can be hacked to create real-world havoc, allowing them to turn off lights, open smart shutters, and more.




Business Systems
Read Data
Take Action
Input
Guardrails
Agent
Rules
Policies
Supervisors
Guardrails
Output
Retrieval
Recollection
Knowledge
Memory


agents
AI vulnerabilities
classic software vulnerabilities

Thank you.

(And also, we are hiring. Email me at clay@sierra.ai)

SIERRA