AKKA

Webinar

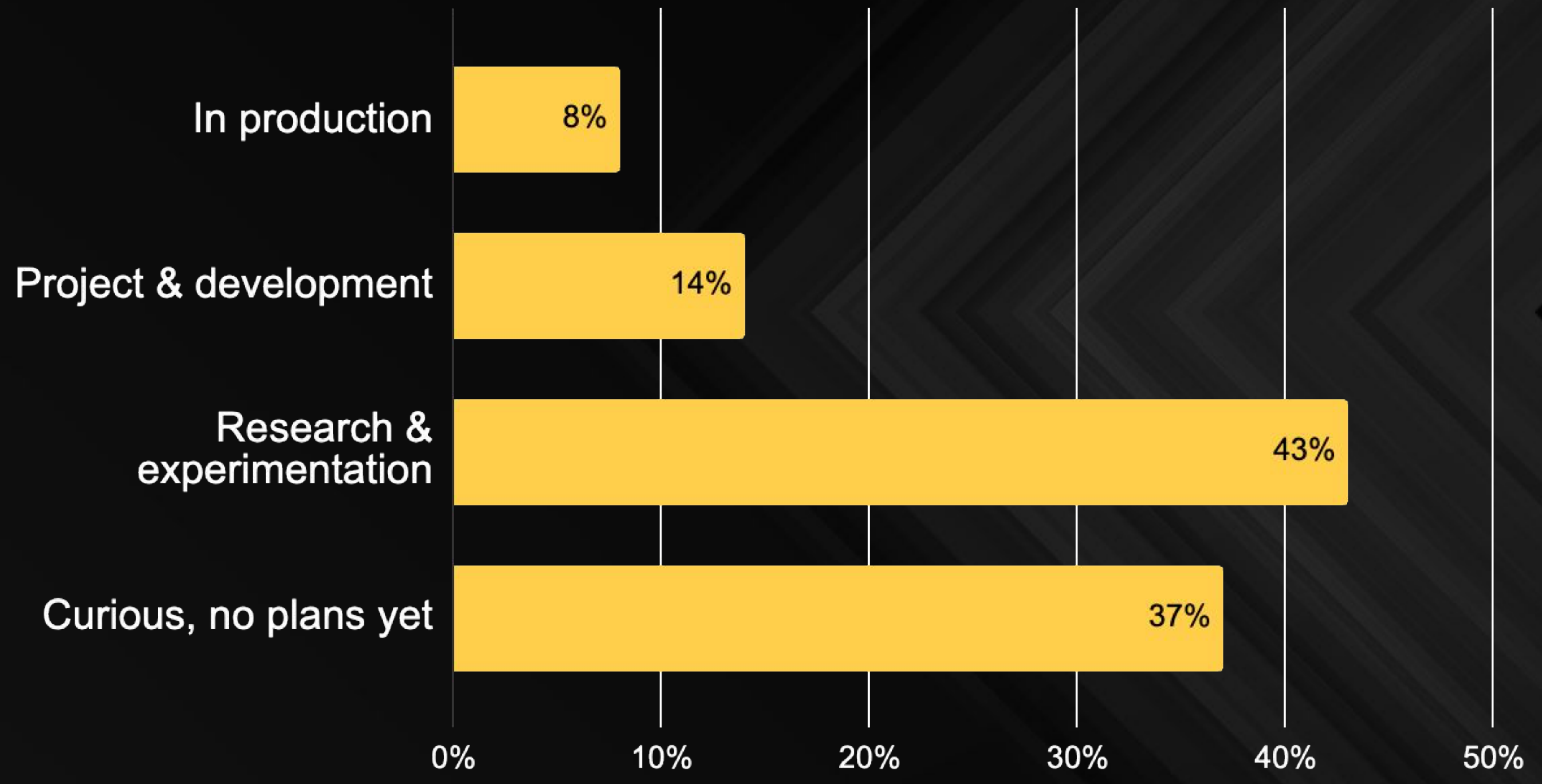
A Blueprint for Agentic AI Services

Best practices for designing and operating agentic-scale services

Accelerate delivery. Stay safe and be efficient.

Poll question #1

Where are you in your agentic journey?



n = 157 webinar poll respondents

Today’s discussion

01 Welcome

Darin Bartik, CMO, Akka

02 The agentic opportunity and the move to a-tier architecture

Richard Li, AI Expert and Entrepreneur

03 A blueprint for agentic services

Tyler Jewell, CEO, Akka

04 Agentic stories and AI in practice

Real-time video augmentation, model-driven personalization, Google Earth AI inference

05 Live Q&A

Q&A transcript and slides to be shared asap

AI is transforming our lives

AI
Assistant

A *user* app that understands natural language commands and uses a conversational AI interface to complete tasks on-demand.

ChatGPT

einstein

perplexity

Siri

AI
Agent

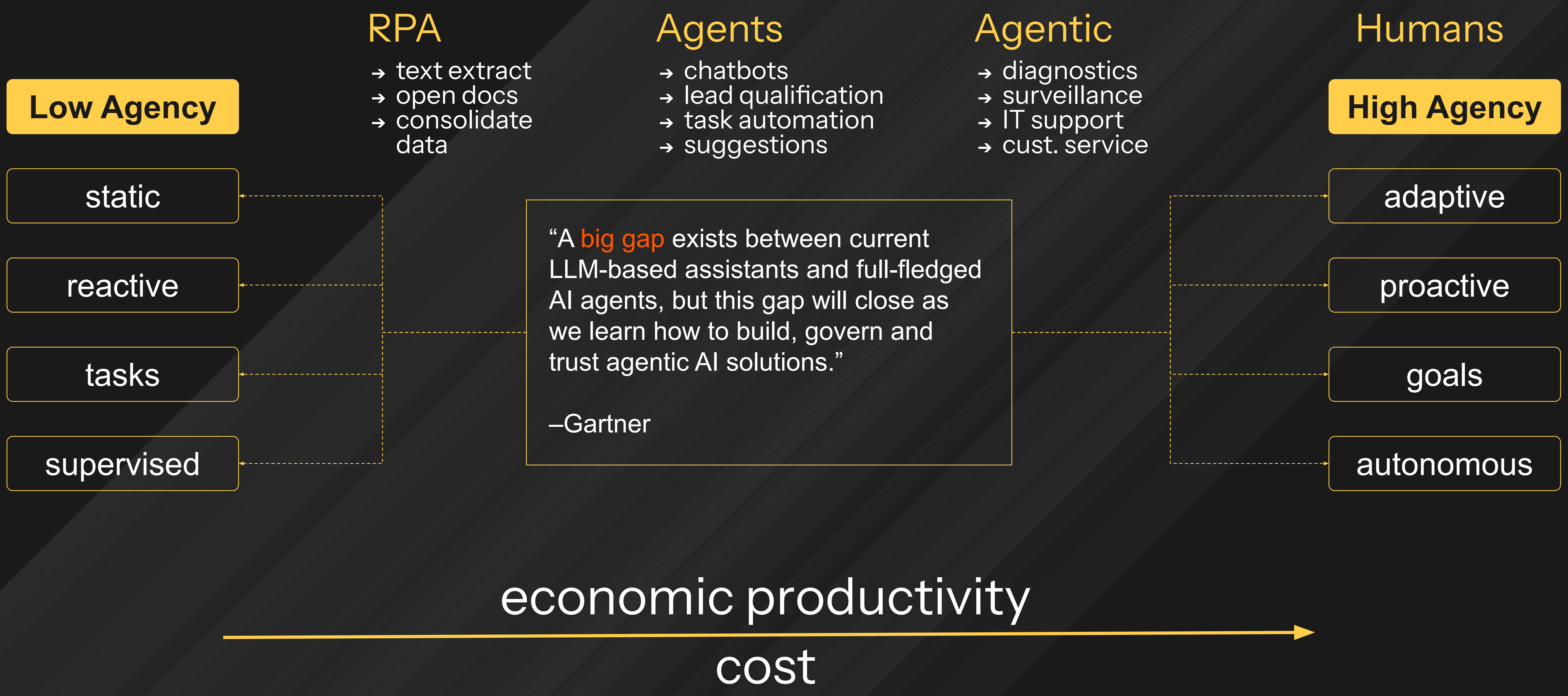
A *system* that can autonomously fulfill goals by interacting with other systems and agents.

Agentforce

AI at ServiceNow

AI Agency

Capacity to make meaning from your environment



RPA
- → text extract
- → open docs
- → consolidate data

Agents
- → chatbots
- → lead qualification
- → task automation
- → suggestions

Agentic
- → diagnostics
- → surveillance
- → IT support
- → cust. service

Humans

Low Agency

static

reactive

tasks

supervised

High Agency

adaptive

proactive

goals

autonomous

> “A big gap exists between current LLM-based assistants and full-fledged AI agents, but this gap will close as we learn how to build, govern and trust agentic AI solutions.”
>
> –Gartner

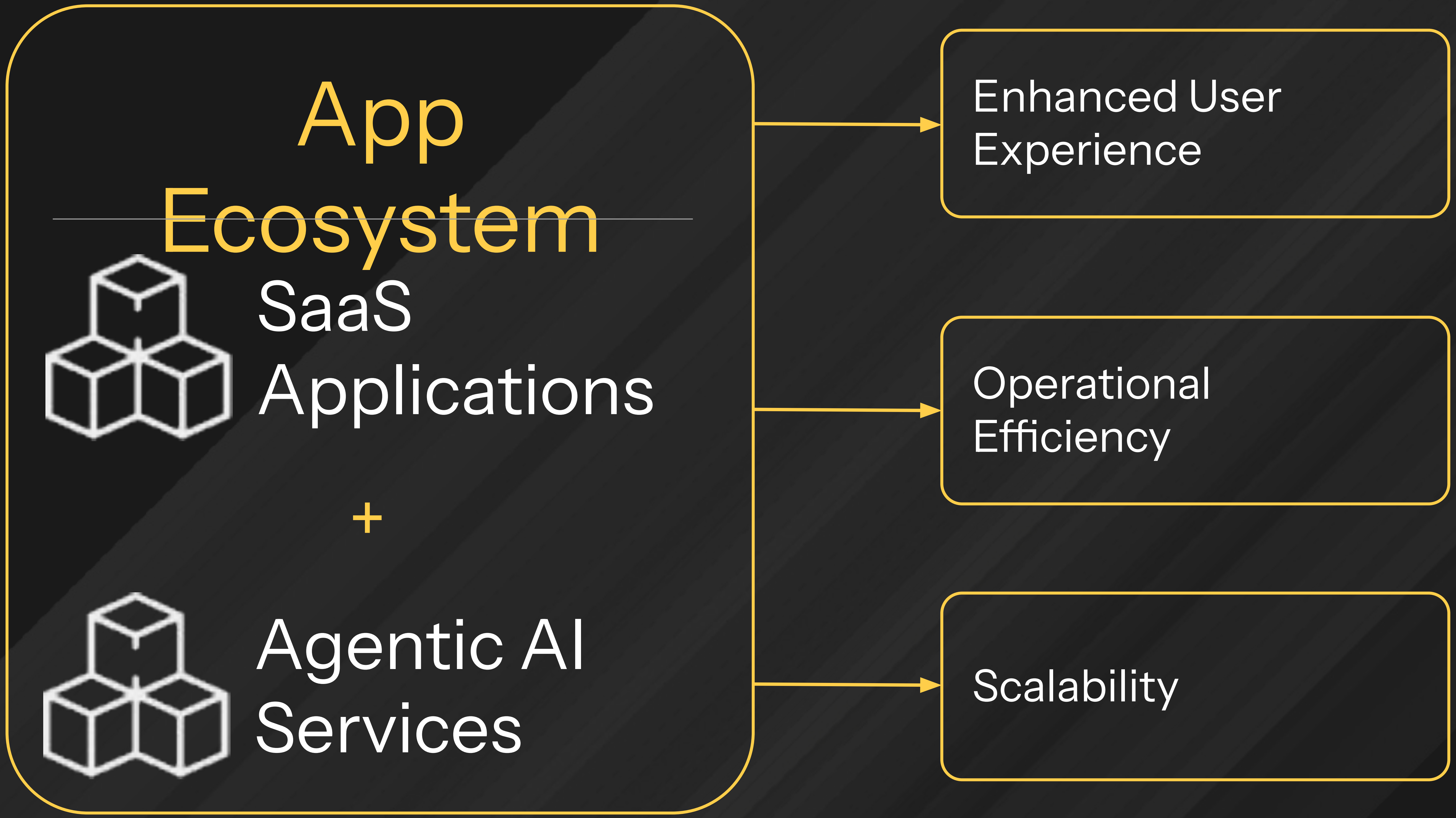
economic productivity

cost

A paradigm shift to AI-fueled app ecosystems

AI agents and apps become part of a symbiotic existence

By 2028, 33% of enterprise software applications will include agentic AI, up from less than 1% in 2024.
Gartner, *TSP 2025 Trends: Agentic AI — The Evolution of Experience, 24 February 2025*



App Ecosystem

SaaS Applications

+

Agentic AI Services

Enhanced User Experience

AI agents personalize interactions to increase satisfaction

Operational Efficiency

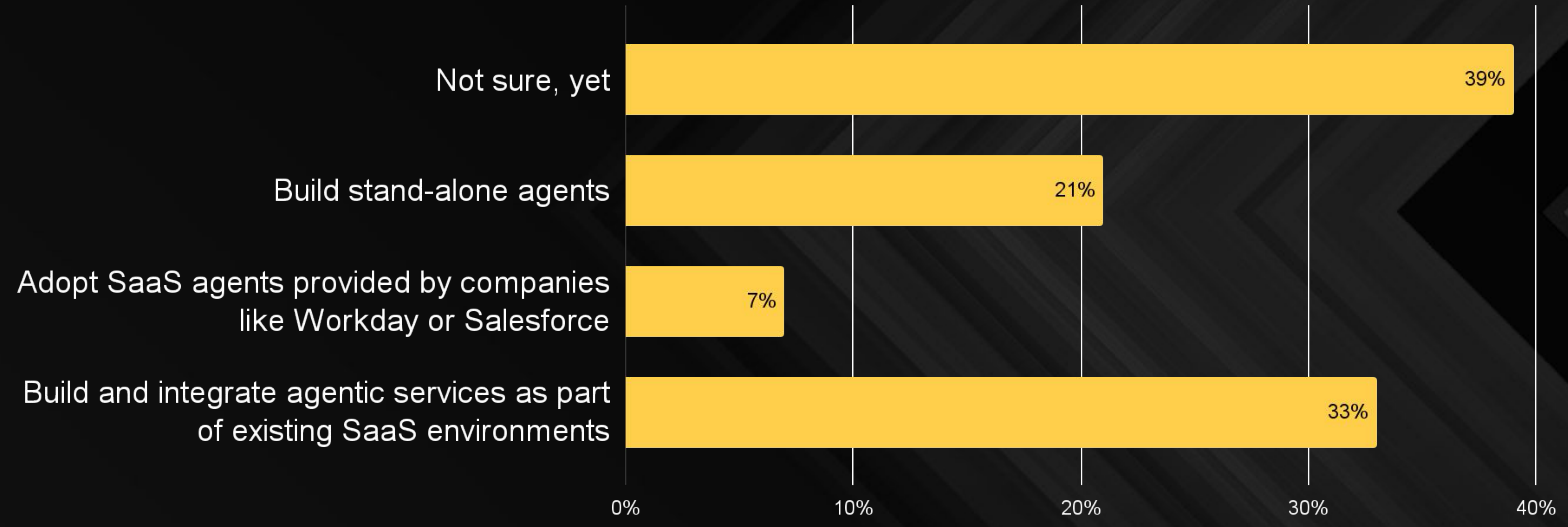
AI agents automate routine tasks to allow humans to focus on strategic initiatives

Scalability

AI-driven SaaS adapt to business needs without proportional increases in cost

Poll question #2

Which approach(es) is your organization considering about agentic apps?



n = 147 webinar poll respondents

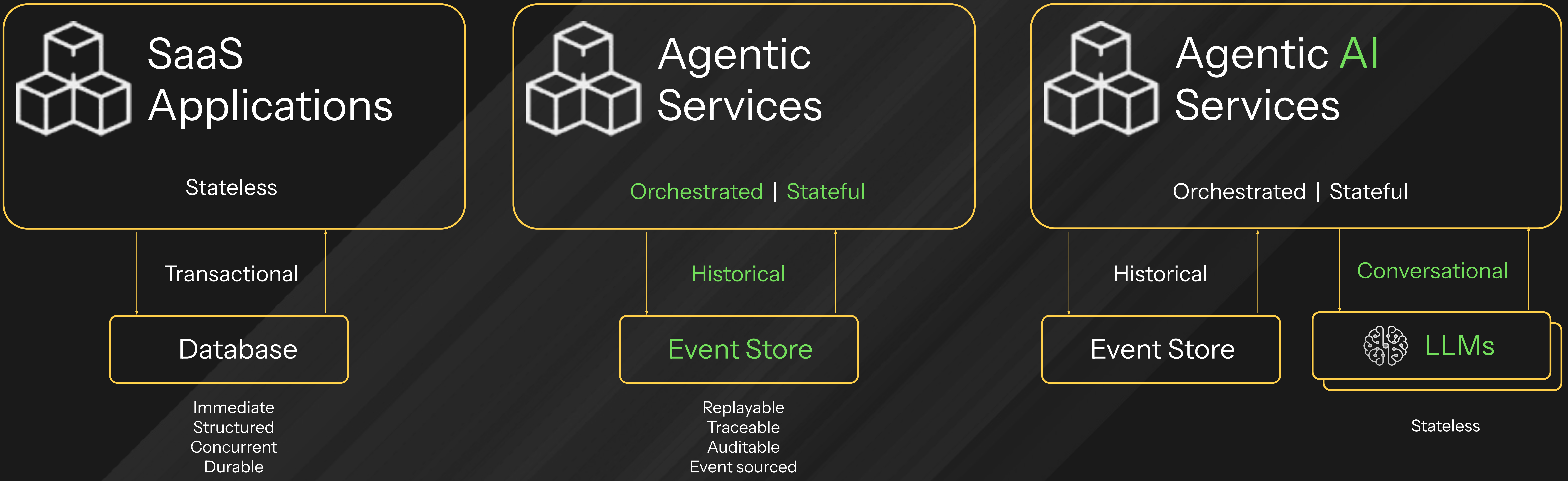
Agentic is the 5th wave of compute

Every human and device with dozens of sleepless assistants

	Mainframe	Web	Cloud	Mobile	Agentic
Users	thousands	millions	10 millions	billions	trillions
TPS	100	500	2,500	10,000	1,000,000
		5x	5x	4x	100x

Transactional apps → Conversational agents

A fundamental shift from request-response to contextual iterations



SaaS Applications

Stateless

Transactional

Database

Immediate
Structured
Concurrent
Durable

Agentic Services

Orchestrated | Stateful

Historical

Event Store

Replayable
Traceable
Auditable
Event sourced

Agentic AI Services

Orchestrated | Stateful

Historical

Event Store

Conversational

LLMs

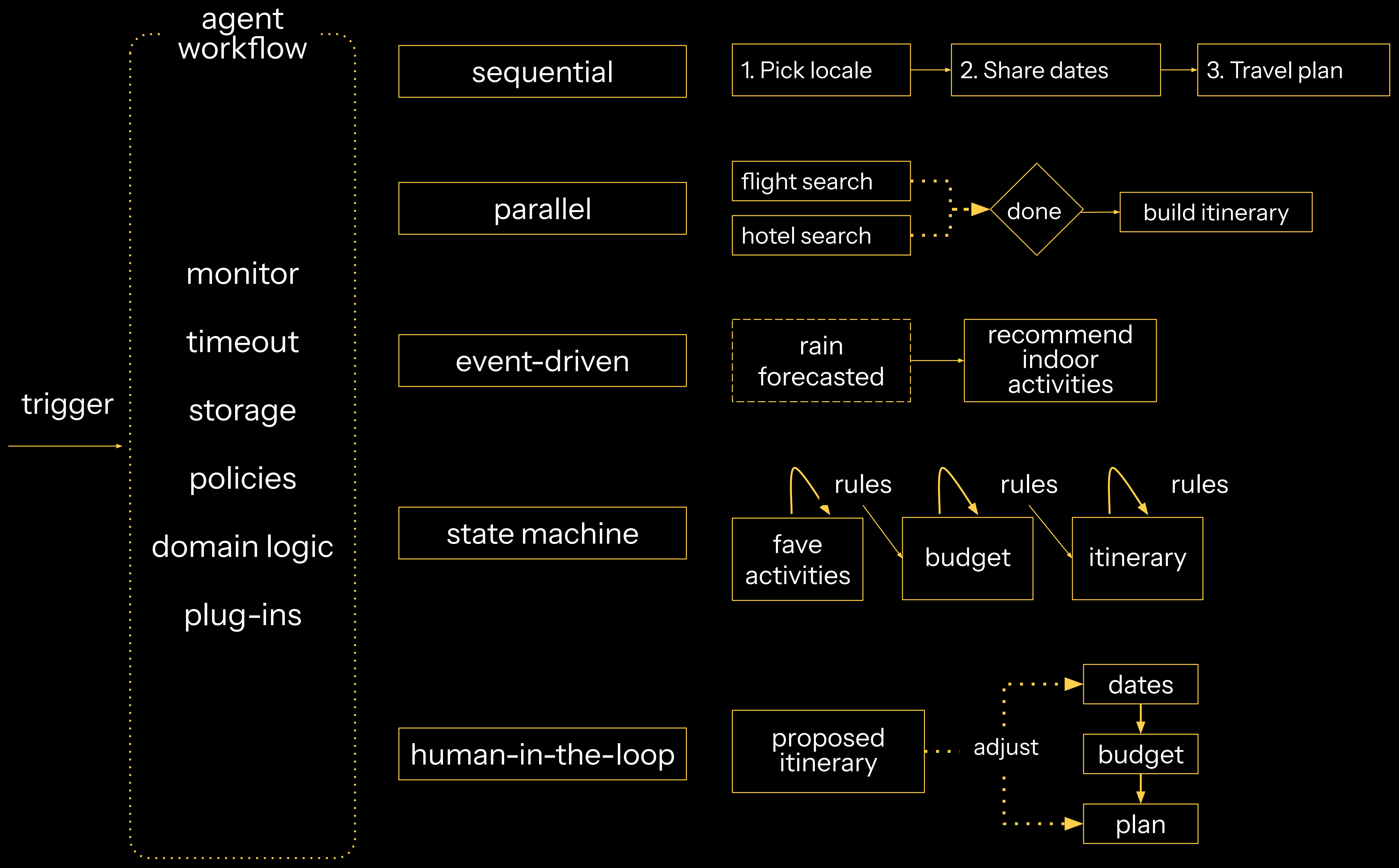
Stateless

Agents are orchestrated services

Workflows: traceable, auditable, debuggable, with point-in-time recovery

Agents are workflows

reliable execution of AI tasks with visibility into request / response data, built-in retries, and error compensation



trigger

agent workflow

monitor

timeout

storage

policies

domain logic

plug-ins

sequential

1. Pick locale

2. Share dates

3. Travel plan

parallel

flight search

hotel search

done

build itinerary

event-driven

rain forecasted

recommend indoor activities

state machine

rules

fave activities

rules

budget

rules

itinerary

human-in-the-loop

proposed itinerary

adjust

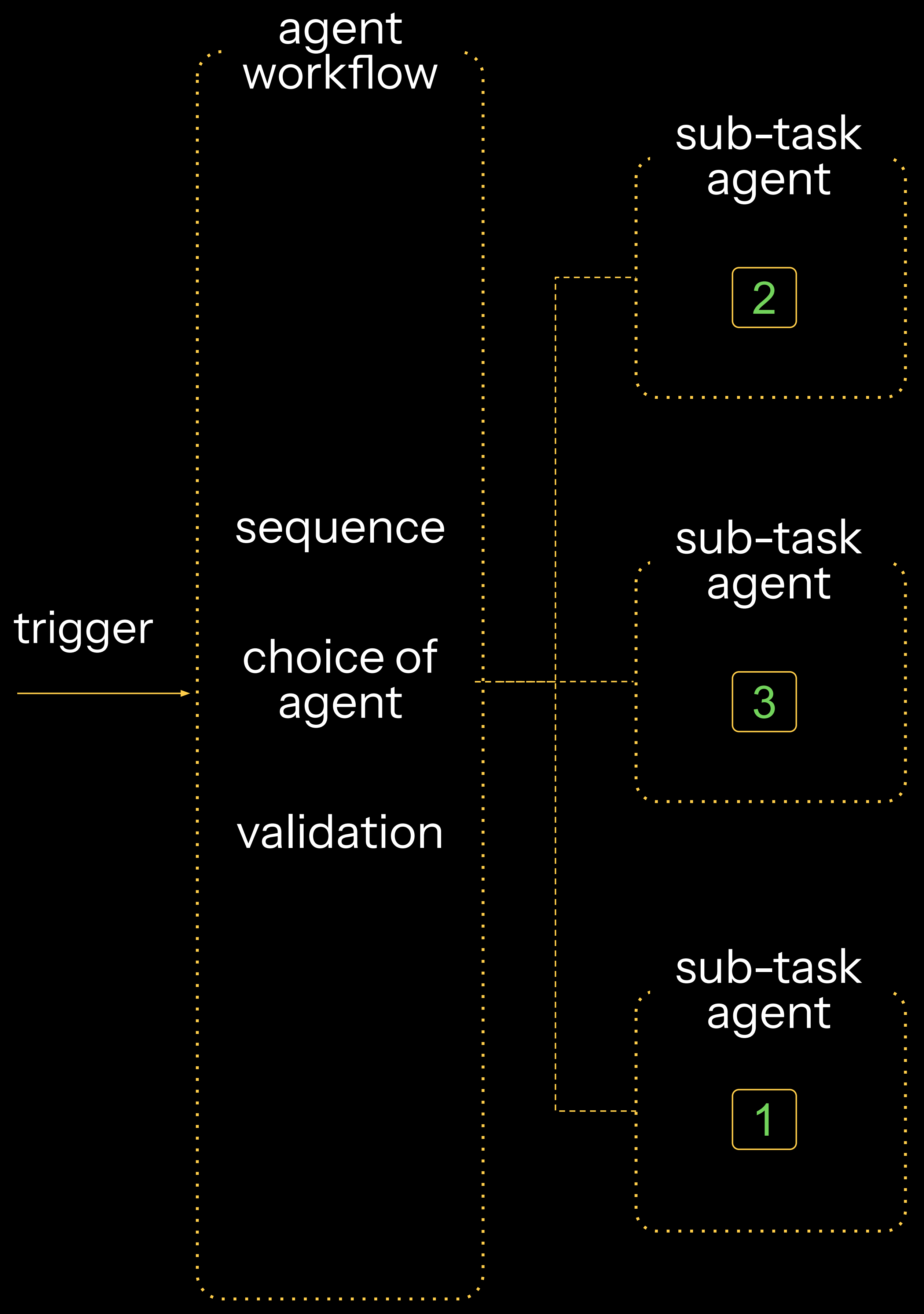
dates

budget

plan

Task chaining

AI agents break complex workflows into smaller, composable steps



trigger

agent workflow

sequence

choice of agent

validation

sub-task agent

2

sub-task agent

3

sub-task agent

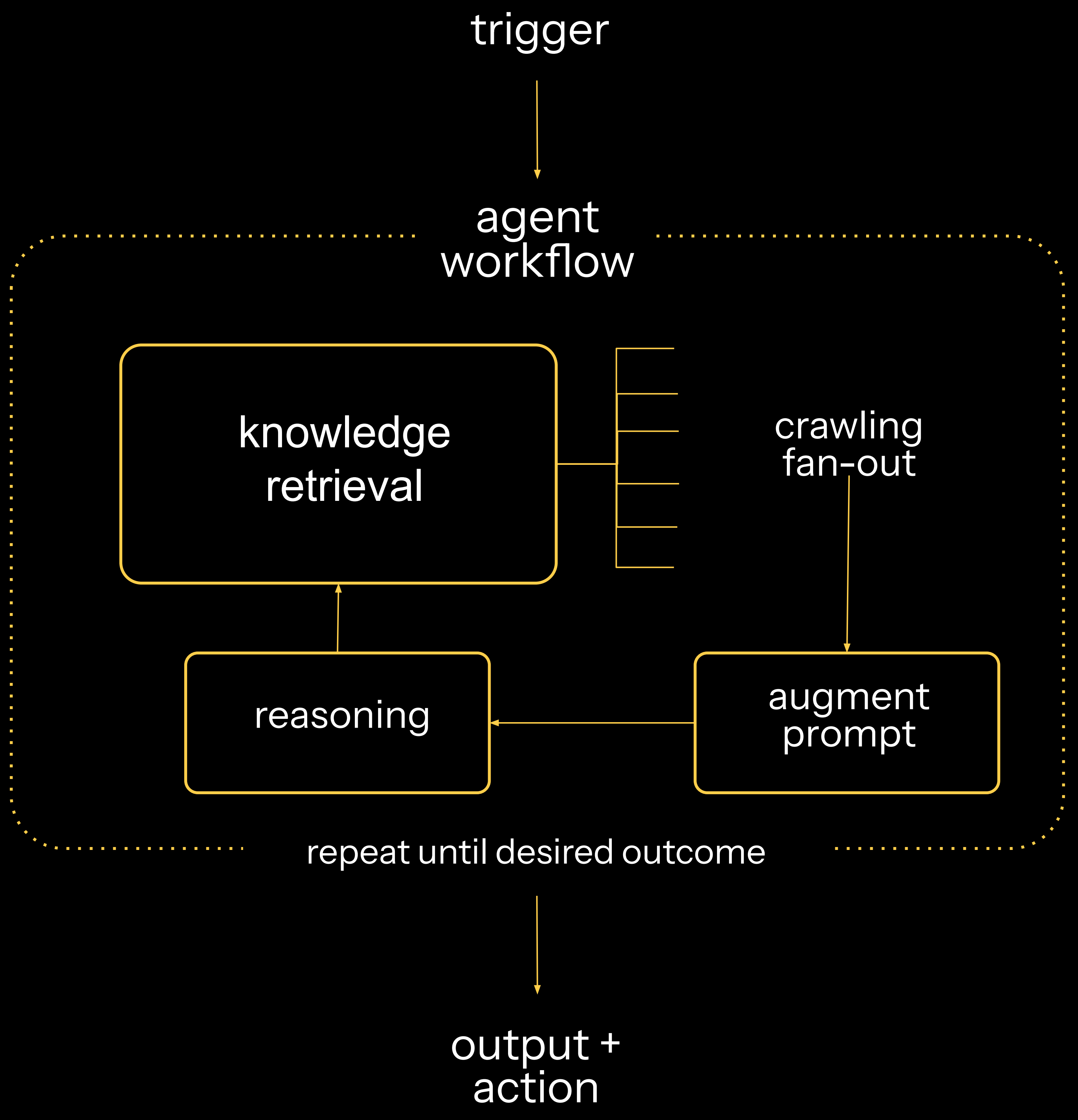
1

Agent types orchestrate levels of agency

De-coupled, event-driven patterns and control loops

Retrieve - augment

agents that combine external knowledge with reasoning and action



trigger

agent workflow

knowledge retrieval

crawling fan-out

augment prompt

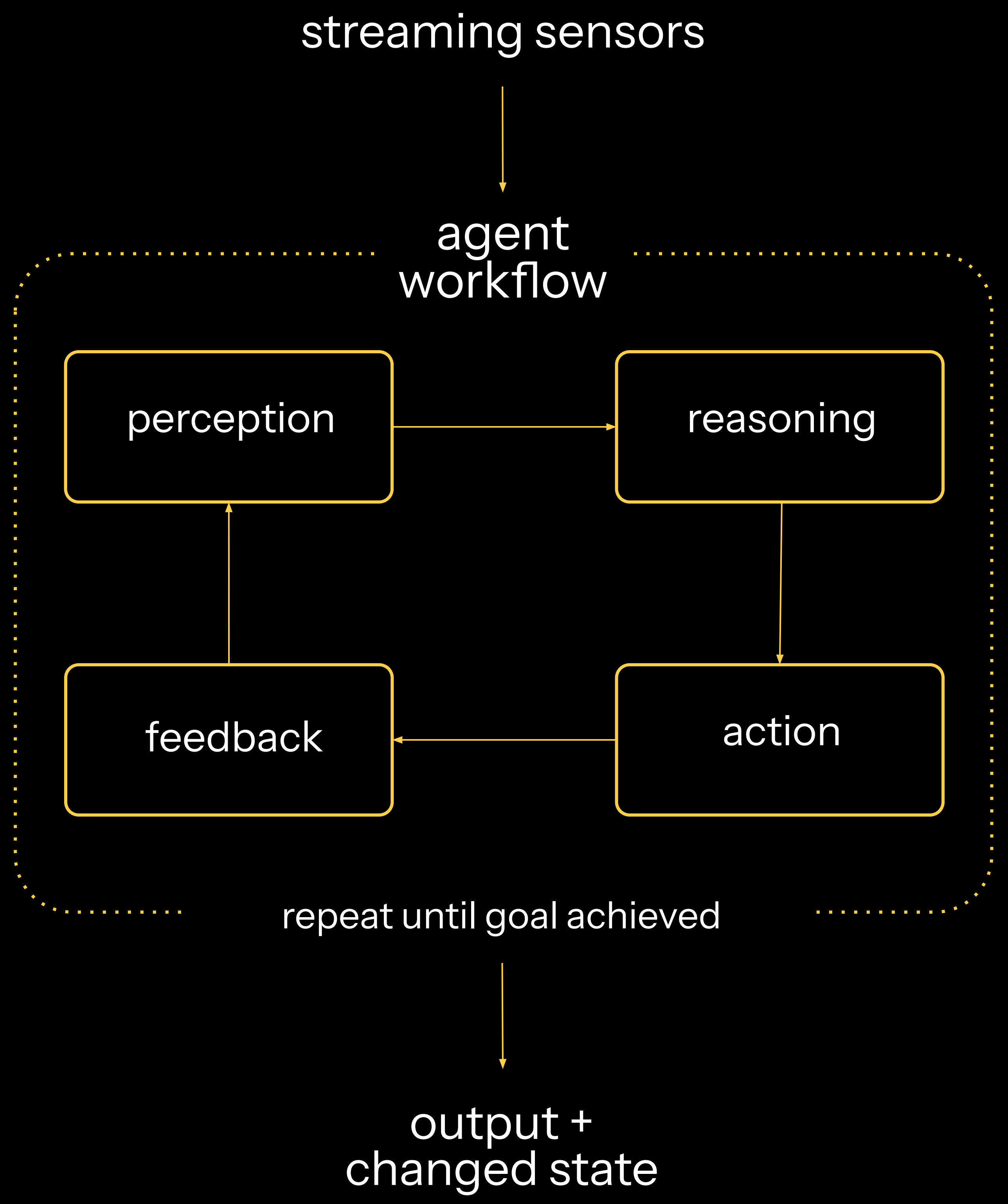
reasoning

repeat until desired outcome

output + action

Environment controllers

control environments in real-time for robotics, edge, and automation



streaming sensors

agent workflow

perception

reasoning

action

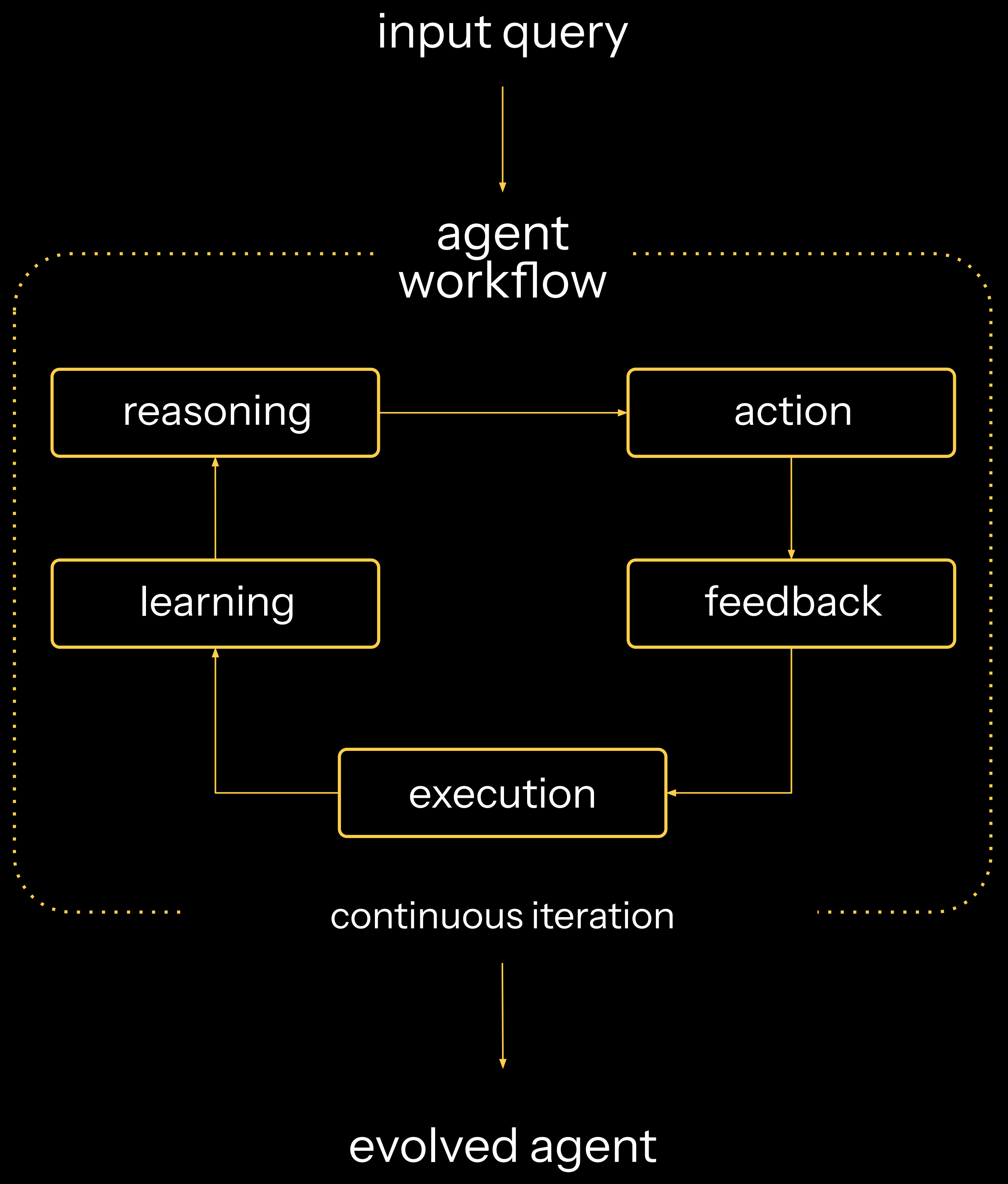
feedback

repeat until goal achieved

output + changed state

Self learning

agents that improve themselves over time through self-reflection and environment adaptation



input query

agent workflow

reasoning

action

feedback

execution

learning

continuous iteration

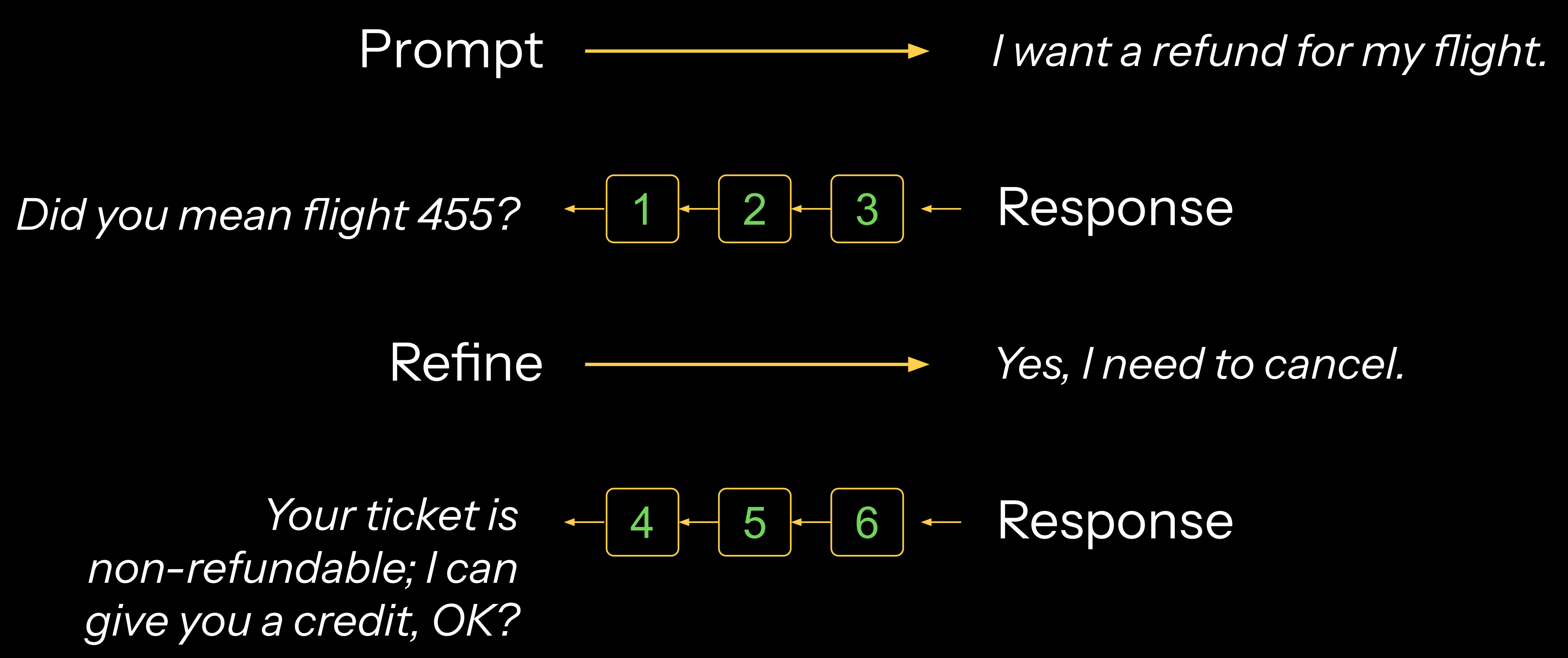
evolved agent

Conversations are stateful

Context and conversation database now a part of the agentic stack

Conversations are ongoing

each iteration
adds context



Prompt → *I want a refund for my flight.*

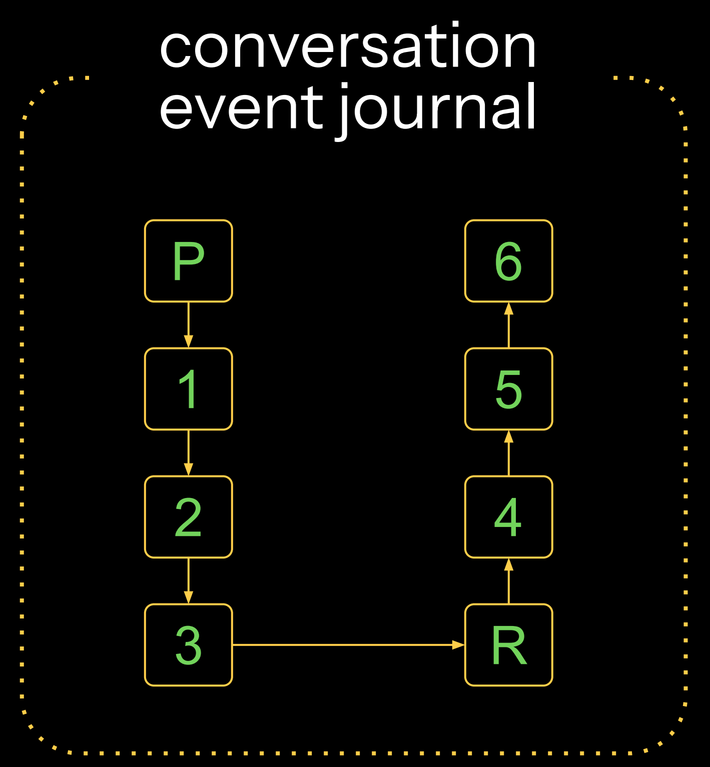
Did you mean flight 455? ← 1 ← 2 ← 3 ← Response

Refine → *Yes, I need to cancel.*

Your ticket is non-refundable; I can give you a credit, OK? ← 4 ← 5 ← 6 ← Response

Conversational sessions

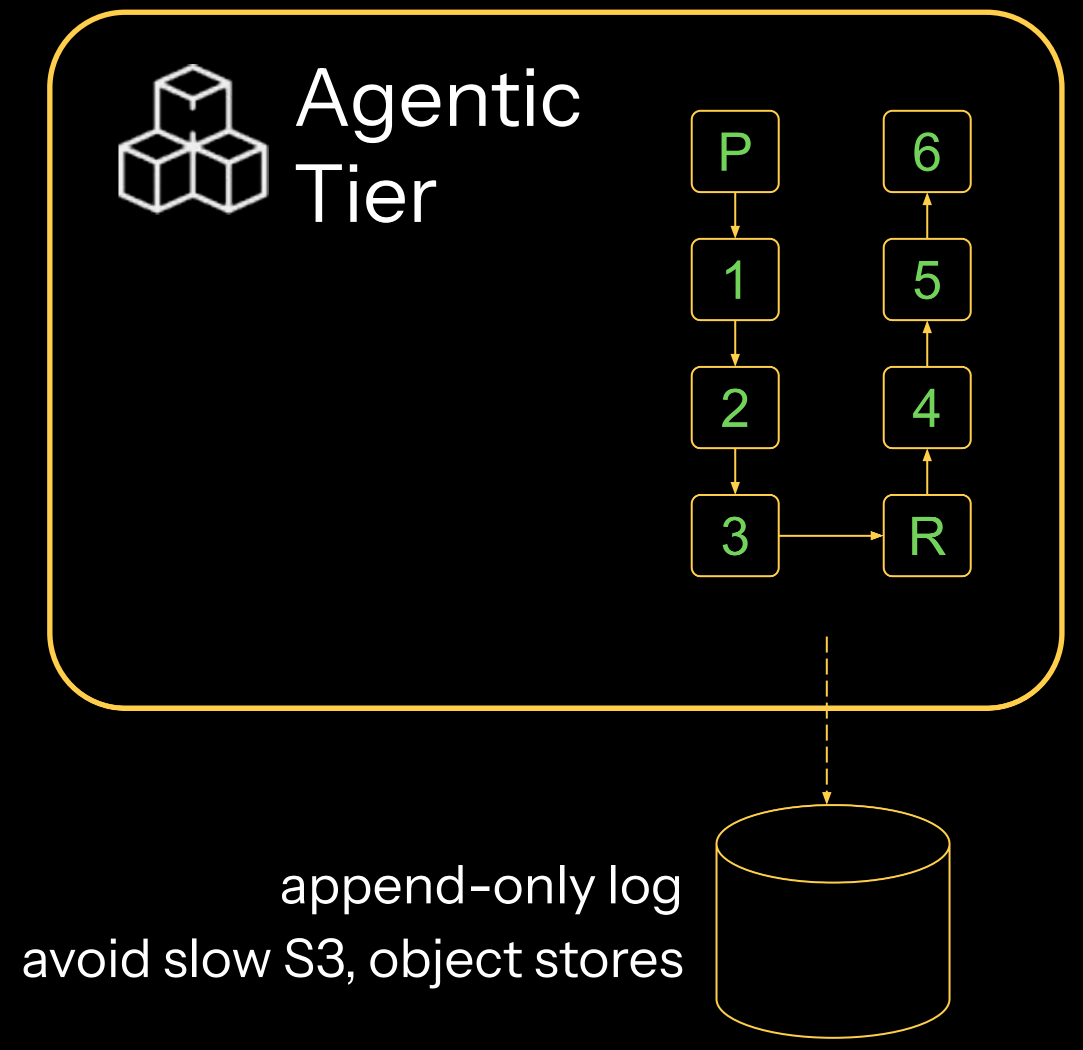
journaled sequences
for context and recovery



conversation
event journal

Conversational persistence

in-memory, durable journals
for speed + resilience



Agentic
Tier

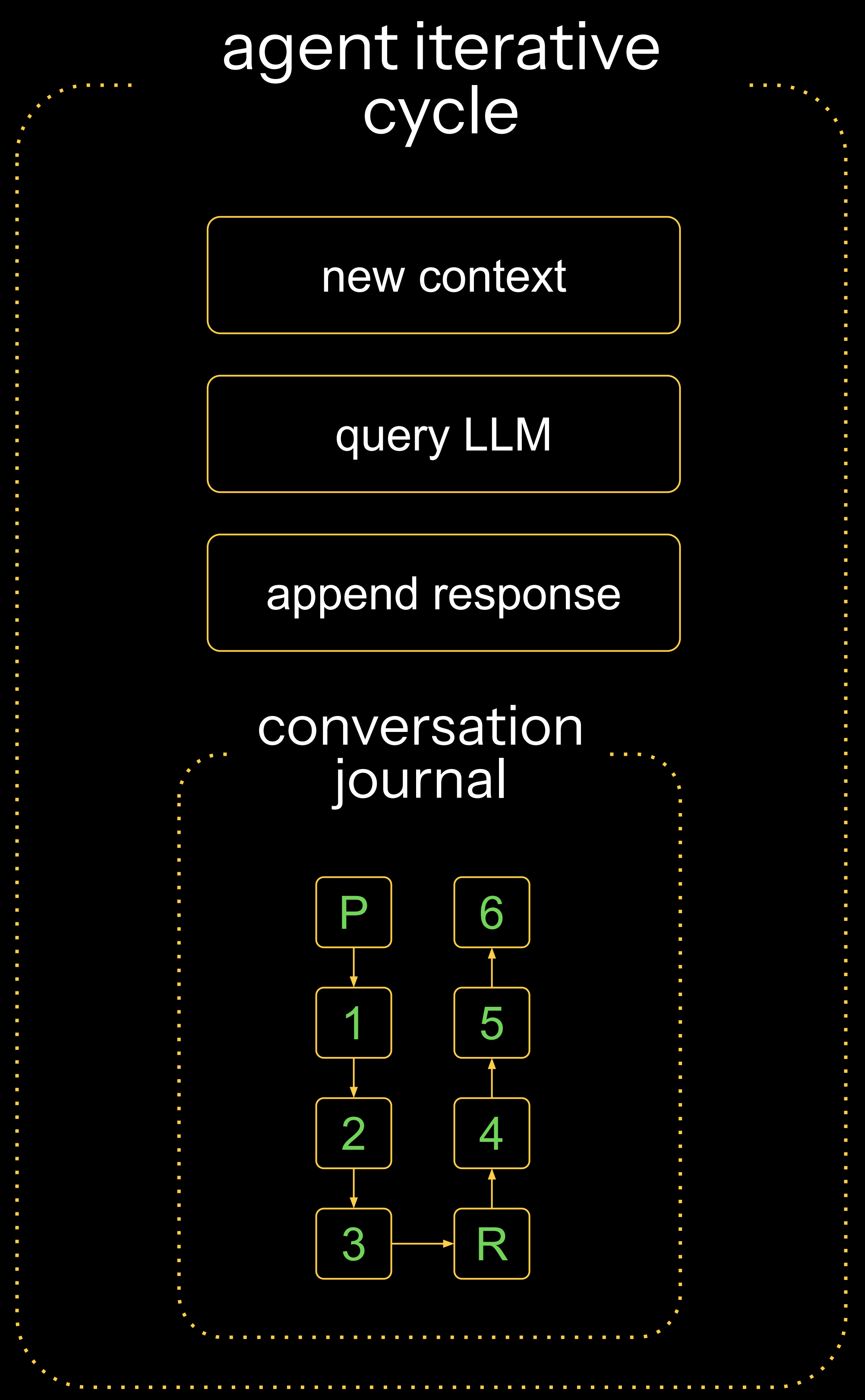
append-only log
avoid slow S3, object stores

Agentic AI augmentation cycle

Agents slow down on each iteration as context grows

Agents start fast

small prompts, small conversations generate quicker responses



agent iterative cycle

new context

query LLM

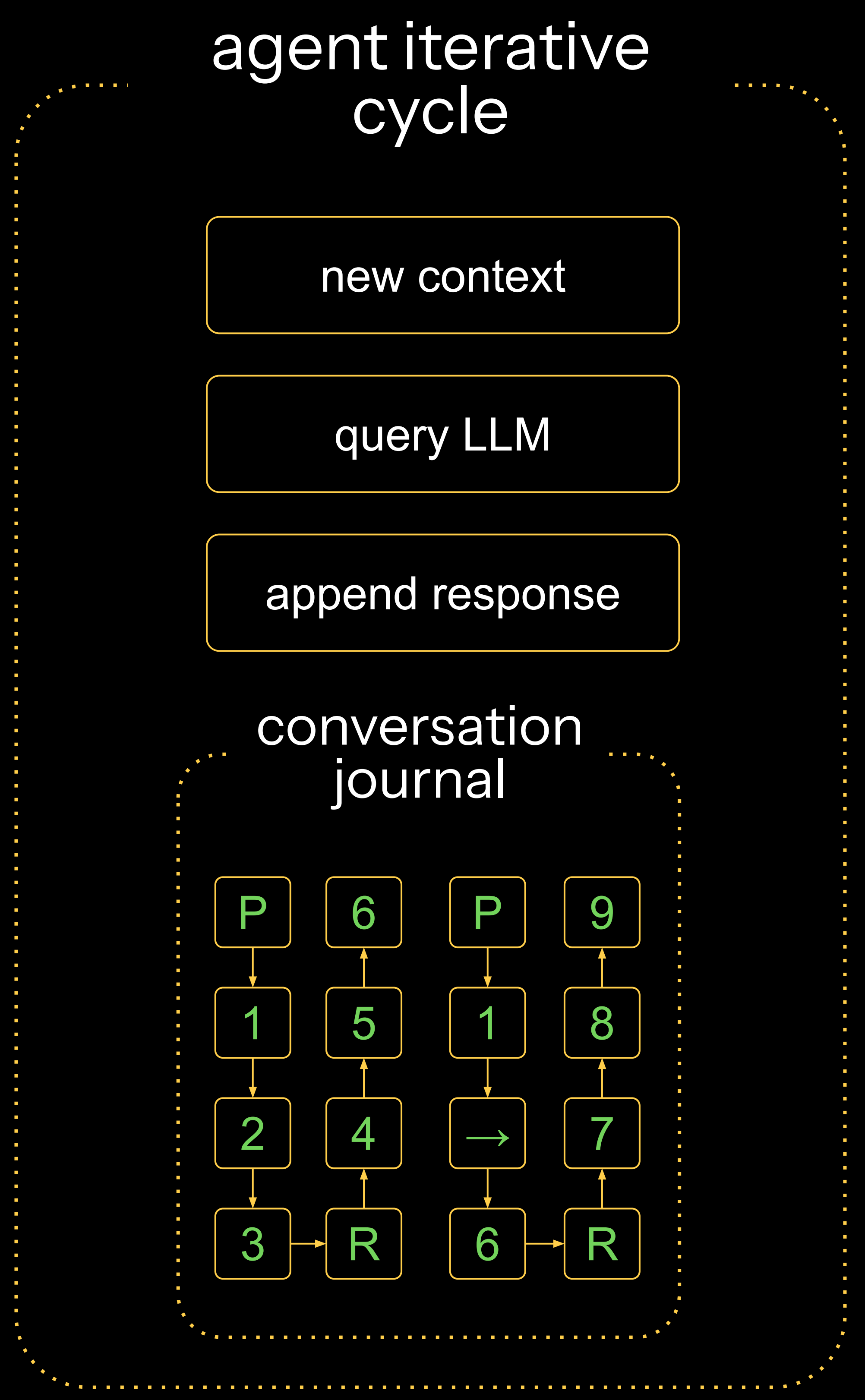
append response

conversation journal

P 1 2 3 R 4 5 6

Agent iterations grow slower

Conversations and prompts grow, eventually hitting LLM token cap



agent iterative cycle

new context

query LLM

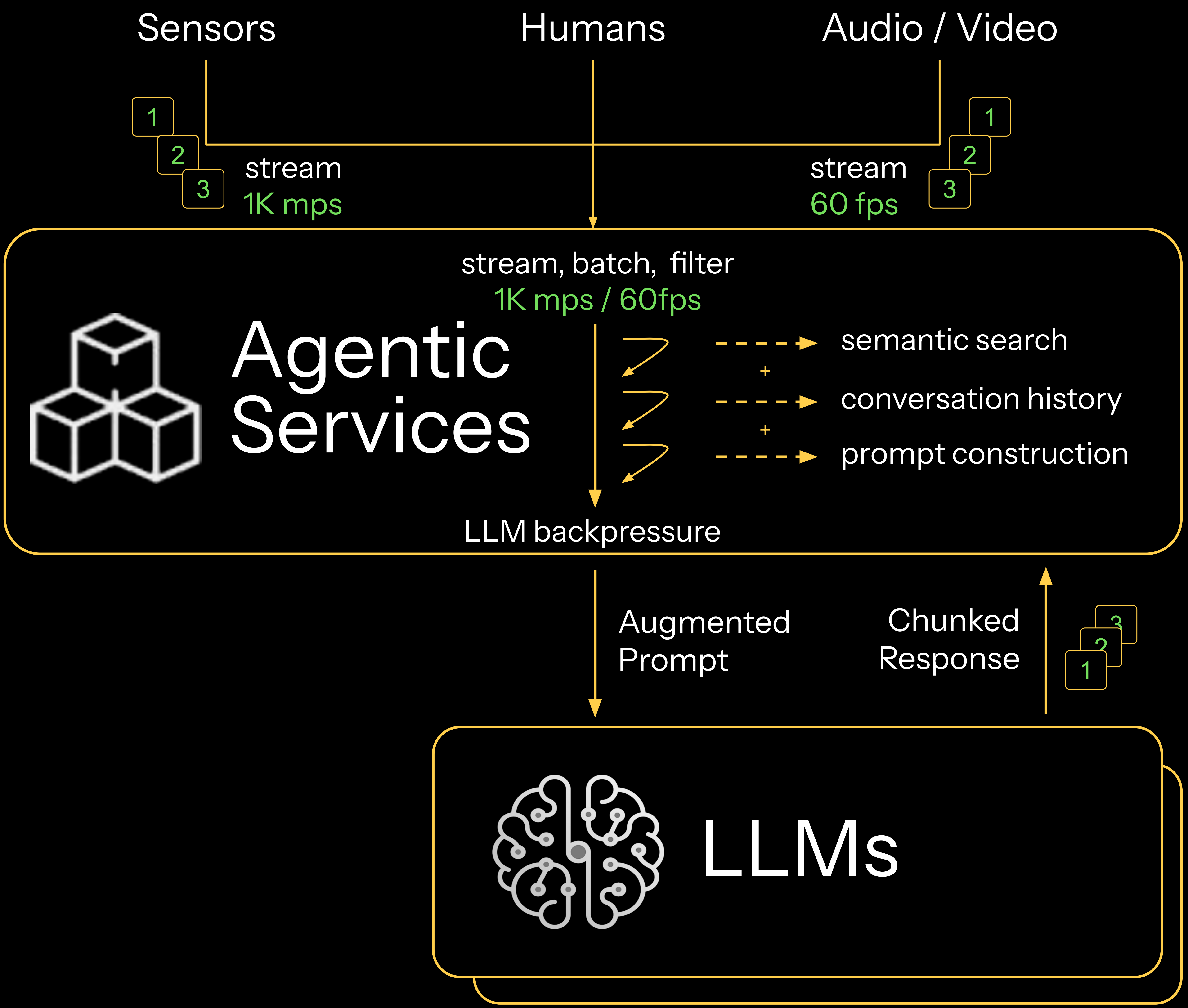
append response

conversation journal

P 1 2 3 R 4 5 6 P 1 → 6 R 7 8 9

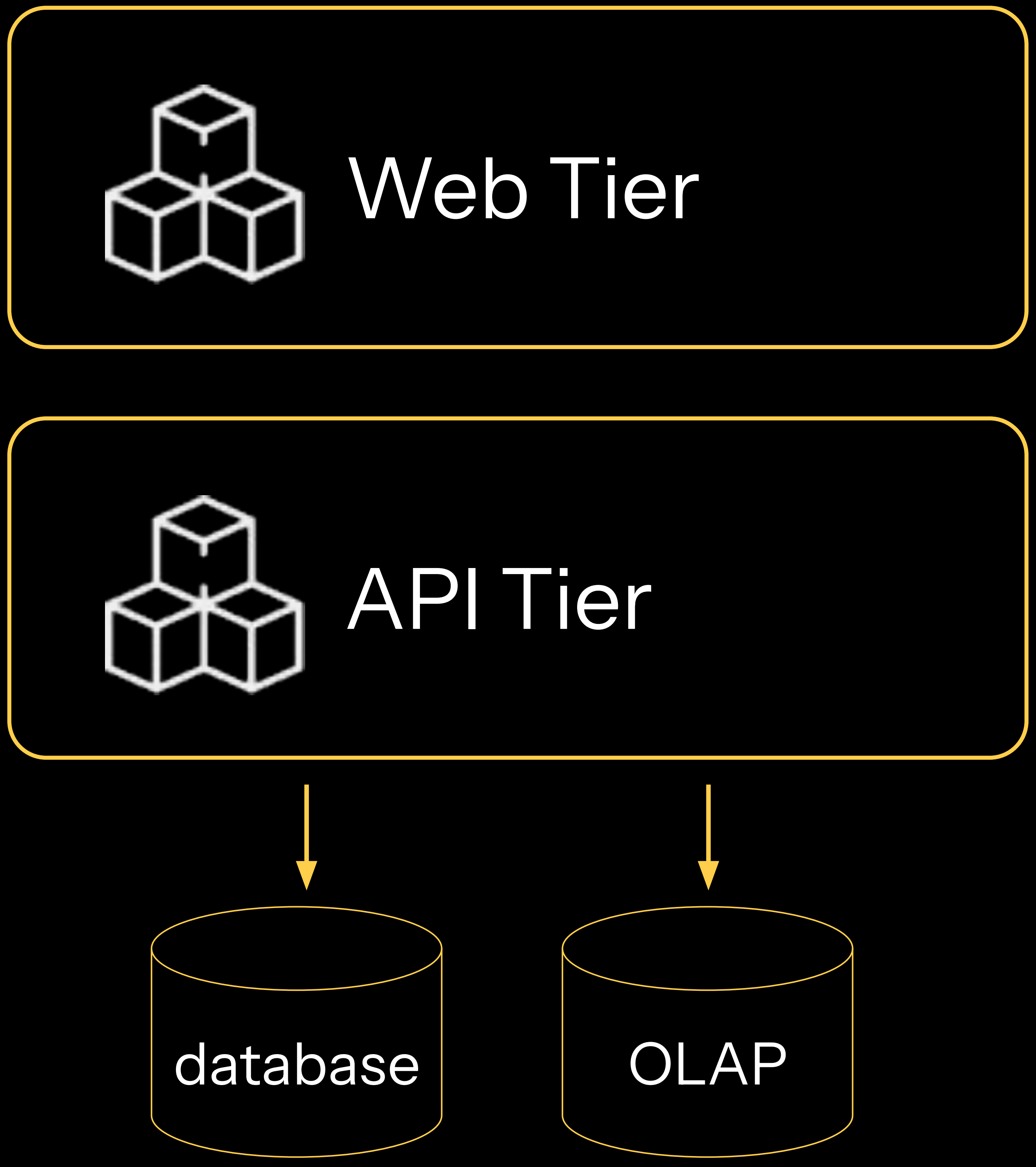
Augment at streaming speeds

Agents augment from a continuous stream of inputs without overloading themselves or their LLMs

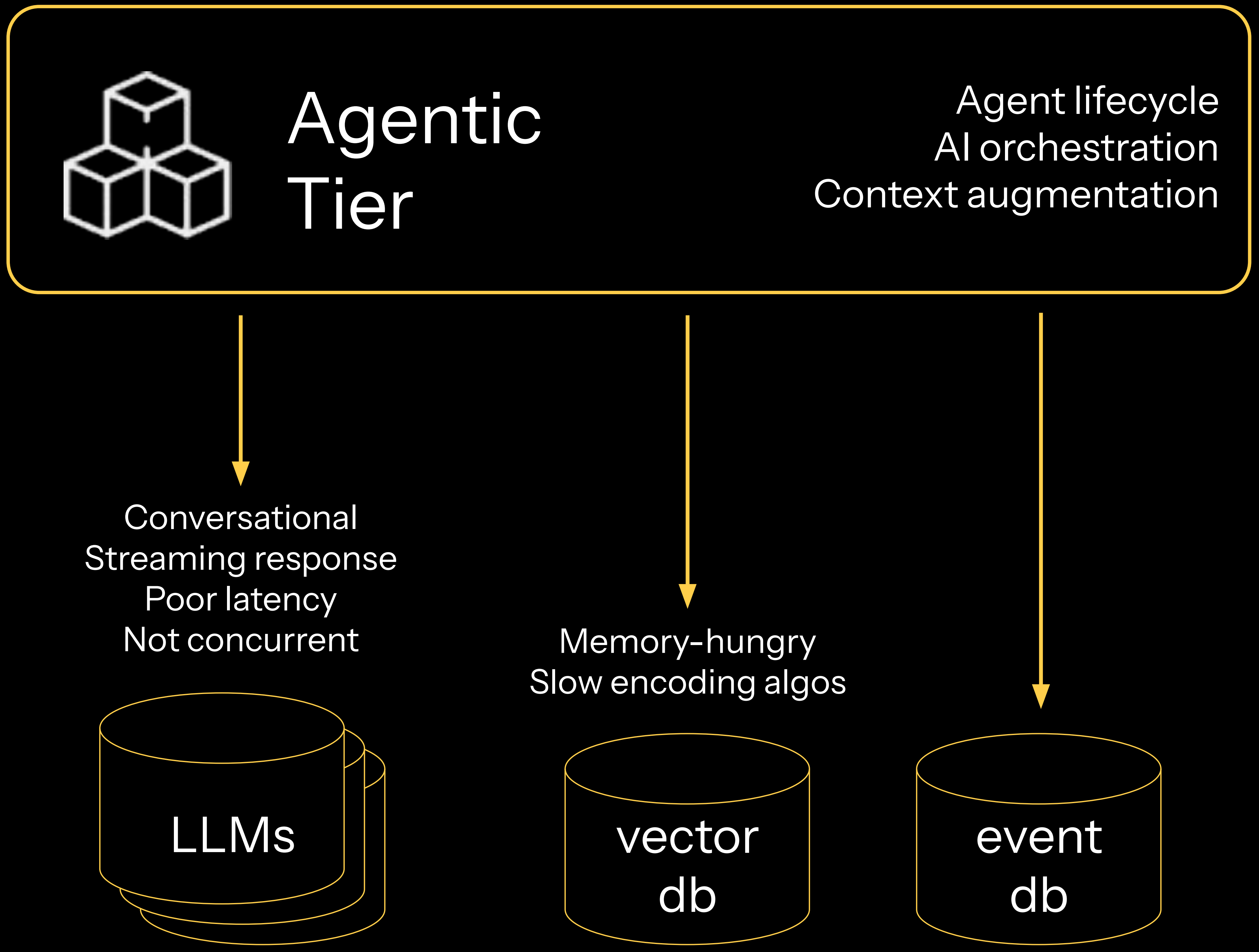


From n-tier to a-tier architecture

Humans + devices augmented with dozens of agent assistants that never sleep



TPS	thousands	hundreds
p(99) latency	10-50ms	50-300 ms



TPS	100x	5x	100x
p(99) latency	150 - 3000ms	50-200ms	5-150ms

Agentic scale requires efficiency

More txs: each slower, less predictable and more costly

	SaaS	Agentic
Users	billions	20x
TPS	10,000	100x
p(99) Latency	10-80ms	15-400x
Cost / LLM tx	cheap	10-10,000x

Mar 25: the best performing LLM @ 86% MMLU accuracy costs $98 / 1M tokens, or ~850,000x more expensive than the average database transaction. The worst performing LLM @ 36% MMLU accuracy costs $.01 / 1M tokens, or 7x more expensive.

Bumpy path from POC to production

52%

fail to reach production

8+ months

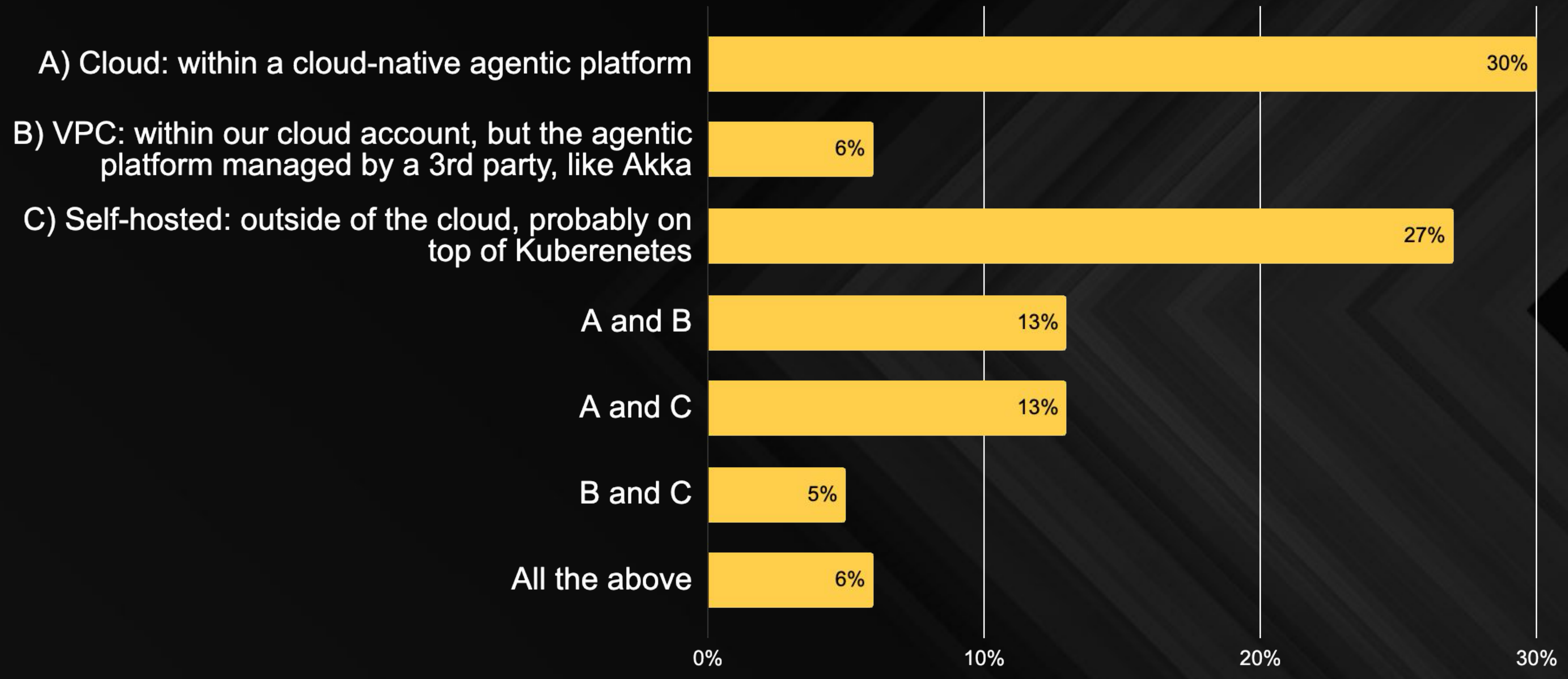
POC to production

> “Leaders reported that only 48% of AI POCs (Proof Of Concept) make it into production, and they take an average of 8.2 months to go from POC to production.”

Gartner

Poll question #3

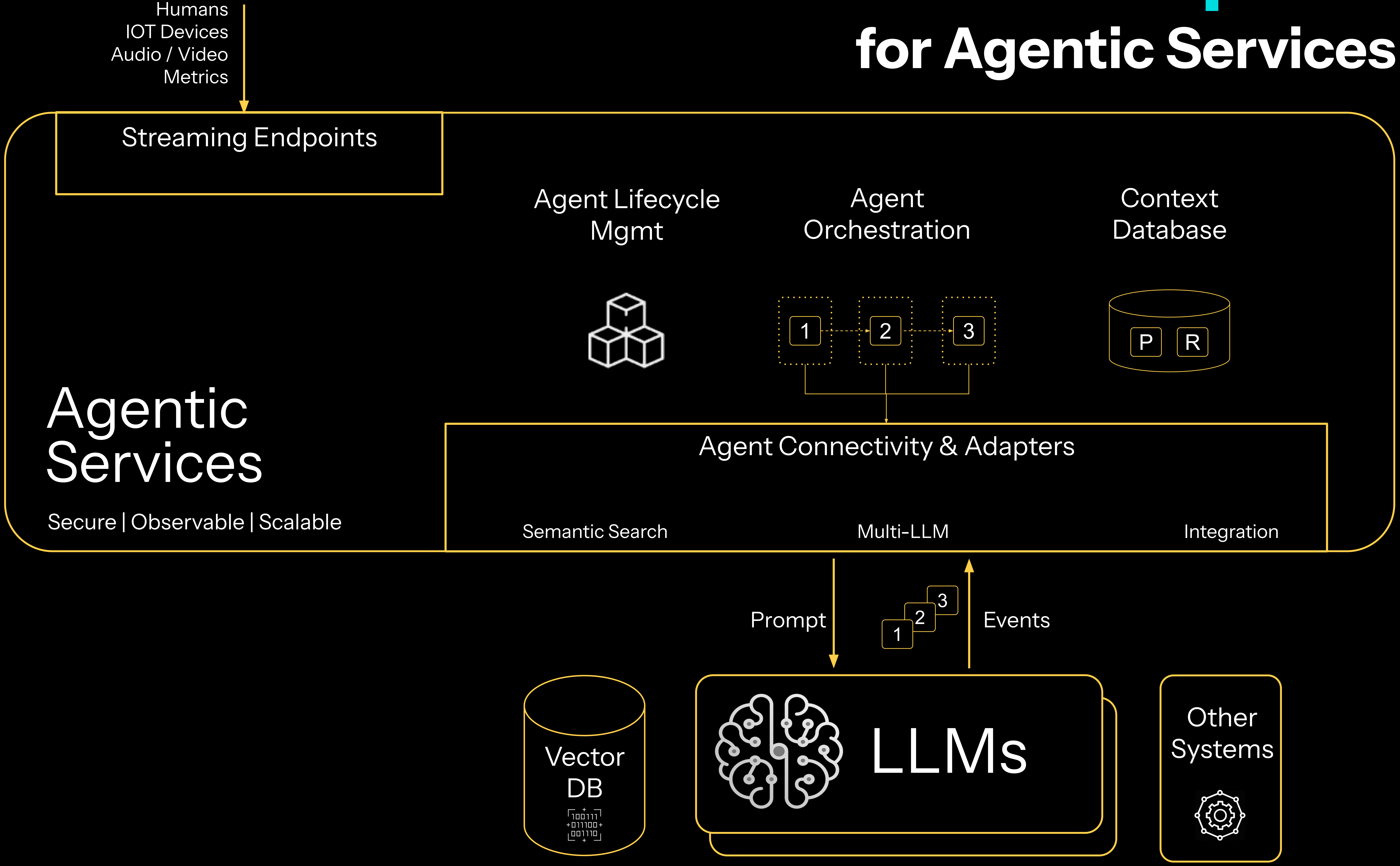
What deployment model(s) would your organization use for agentic services?

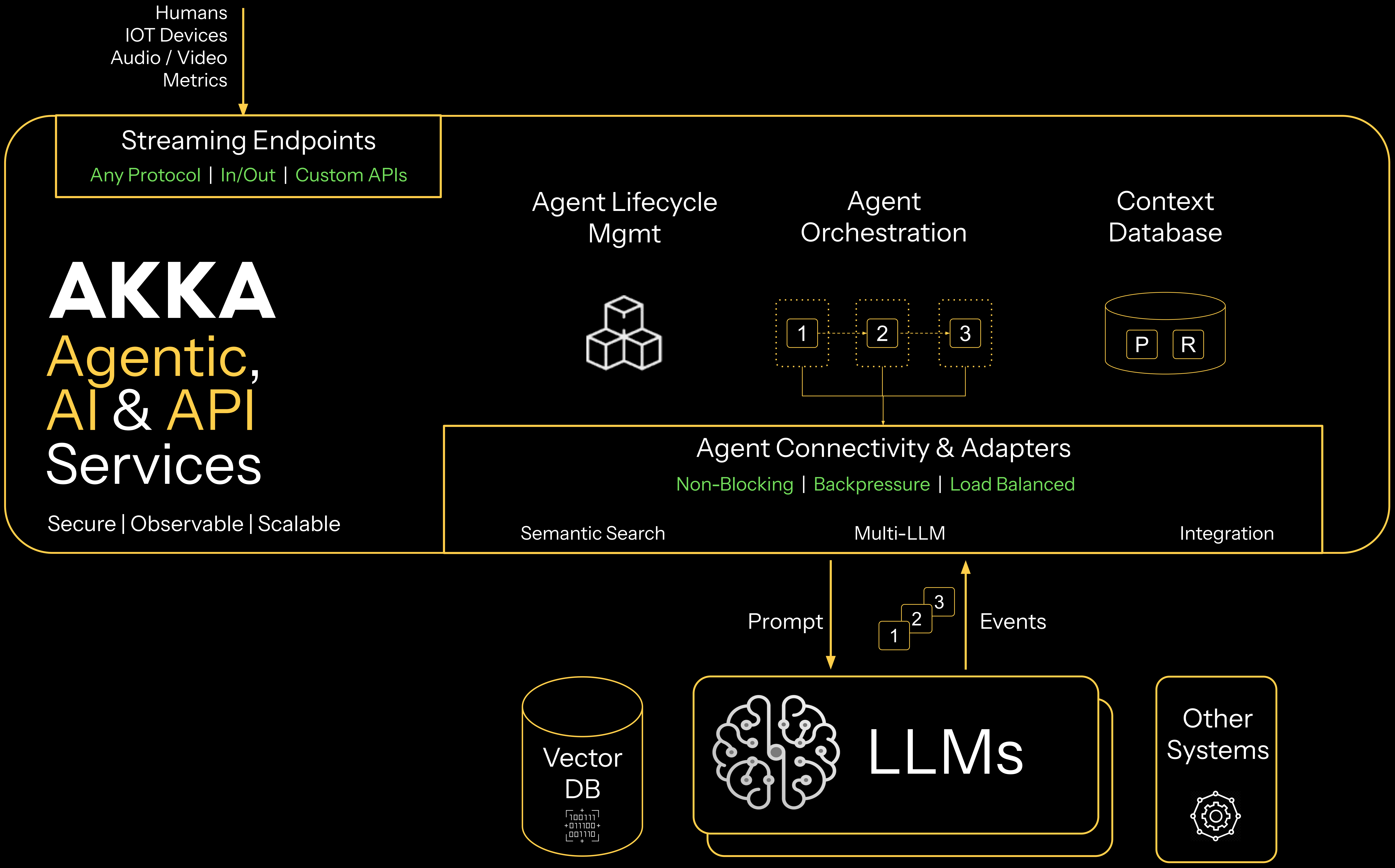


n = 130 webinar poll respondents

Blueprint
for Agentic Services

AKKA





Humans
IOT Devices
Audio / Video
Metrics

Streaming Endpoints

Any Protocol | In/Out | Custom APIs

AKKA
Agentic, AI & API Services

Secure | Observable | Scalable

Agent Lifecycle Mgmt

Agent Orchestration

1 2 3

Context Database

P R

Agent Connectivity & Adapters

Non-Blocking | Backpressure | Load Balanced

Semantic Search

Multi-LLM

Integration

Prompt

1 2 3

Events

Vector DB

LLMs

Other Systems

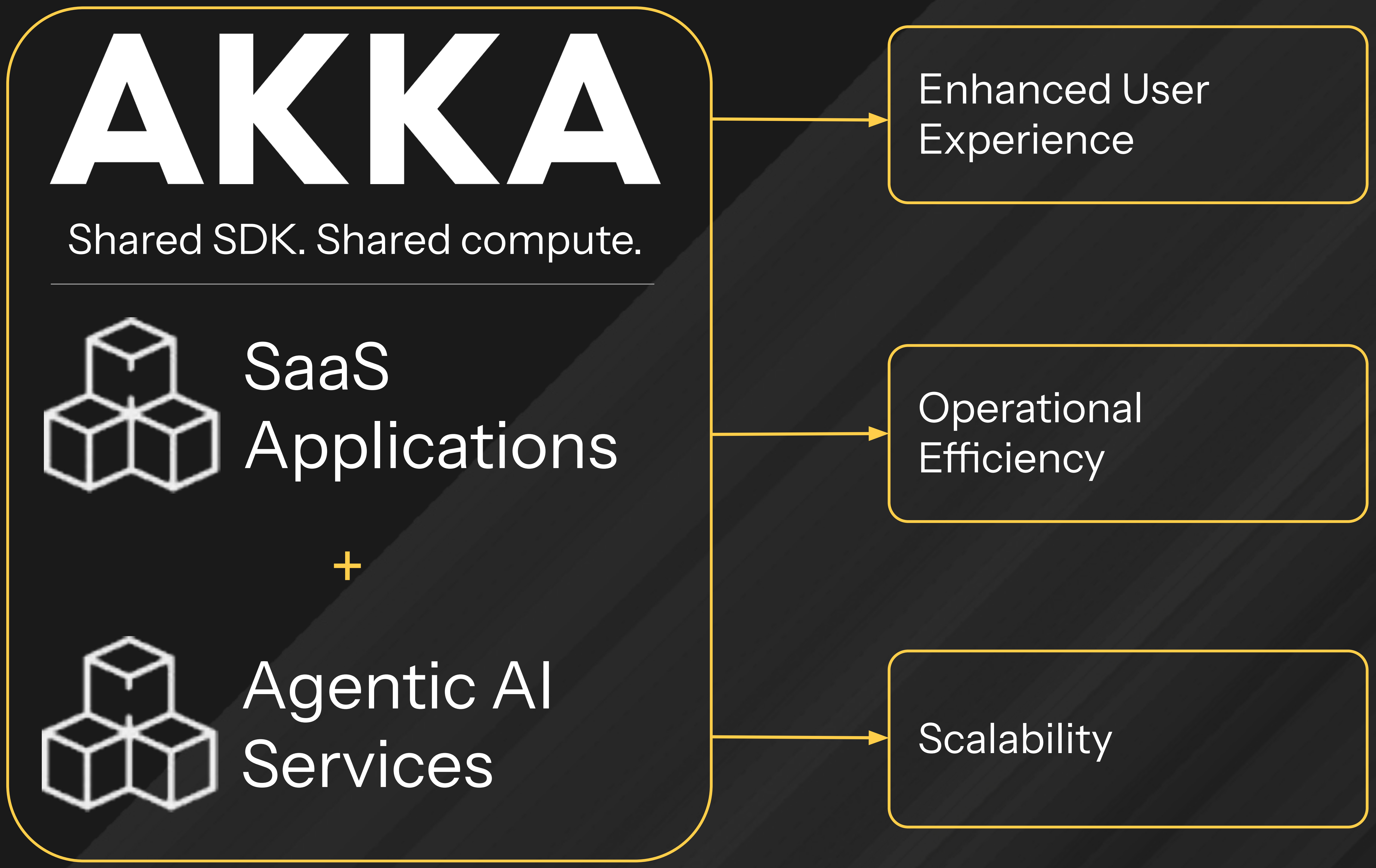
Efficient
70% less compute
API + agentic combo

Elastic
5M TPS
akka clustering

Agile
Prod in days
SDK + ops envs

Resilient
0-0 RTO/RPO
multi-region, multi-master data replication

Akka accelerates delivery of agentic AI apps



AKKA

Shared SDK. Shared compute.

SaaS Applications

+

Agentic AI Services

Enhanced User Experience

Operational Efficiency

Scalability

AI agents personalize interactions to increase satisfaction

AI agents automate routine tasks to allow humans to focus on strategic initiatives

AI-driven SaaS adapt to business needs without proportional increases in cost

avoid the workflow island - orchestration without streaming, context database, or custom API endpoints

crewai Temporal inngest

avoid the framework trap - dev tools with locking, concurrency, & memory not suited for 24/7 ops

LangChain

The Akka agentic advantage

- ✓ Agentic, AI, apps & data
- ✓ Hardened runtime
- ✓ Simple, expressive SDK
- ✓ Multi-region
- ✓ Automated ops

Streaming endpoints

- ➔ Shared compute: agentic co-execution with API services
- ➔ HTTP and gRPC custom API endpoints
- ➔ Custom protocols, media types, and edge deployments
- ➔ Real-time streaming ingest, benchmarked to over 1TB

Context database

- ➔ Agentic sessions with infinite context
- ➔ Context snapshot pruning to avoid LLM token caps
- ➔ In-memory context sharding, load balancing, and traffic routing
- ➔ Multi-region context replication
- ➔ Replication filters for region-pinning user context data
- ➔ Embedded context persistence with Postgres event store

Agent connectivity & adapters

- ➔ Non-blocking, streaming LLM inference adapters with back pressure
- ➔ Multi-LLM selection
- ➔ LLM adapters & 100s of ML algos
- ➔ Agent-to-agent brokerless messaging
- ➔ 100s of 3rd party integrations

Agent orchestration

- ➔ Event-driven runtime benchmarked to 10M TPS
- ➔ SDK with AI workflow component
- ➔ Serial, parallel, state machine, & human-in-the-loop flows
- ➔ Sub-tasking agents and multi-agent coordination

Agent lifecycle management

- ➔ Agent versioning
- ➔ Agent replay
- ➔ Event, workflow, and agent debugger
- ➔ No downtime agent upgrades

AKKA

2B people experience Akka daily

SMILE

A fast ML engine with 100s of ML & LLM inference, powering Google Earth

400K downloads / mo
6K GitHub stars

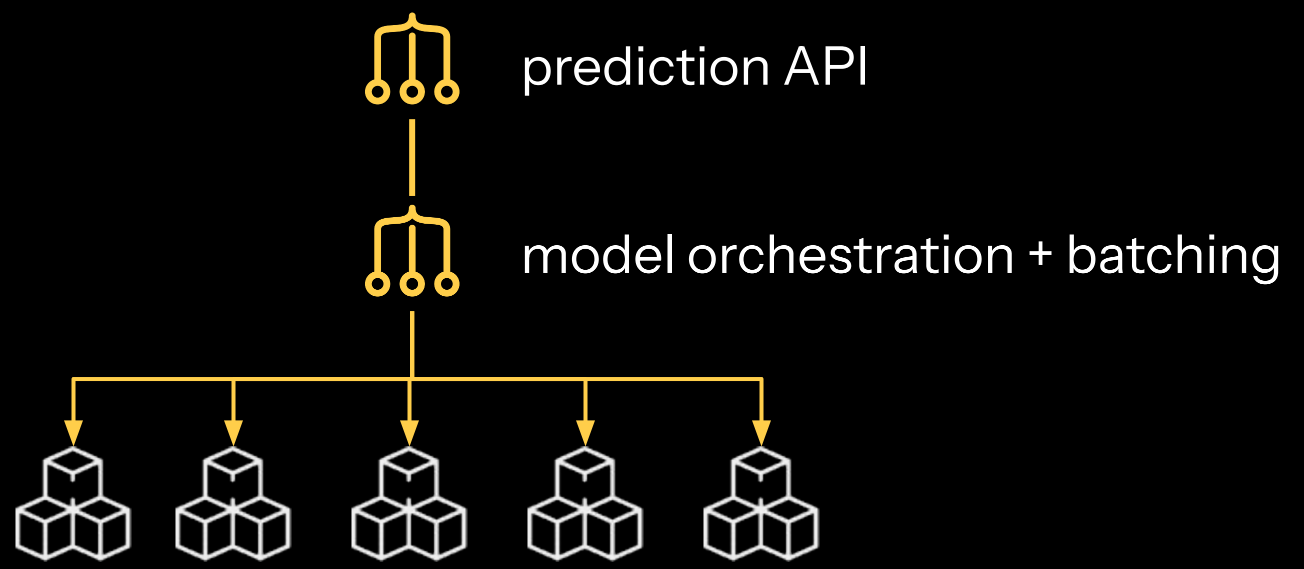
"Akka is used for streaming and back pressure - critical for hosted AI API inference. Akka enables event-driven inference exposed as HTTP efficiently, with low latency."

Haifeng Li – maintainer

Swiggy

API-driven predictions with multi-model fan-out and ultra-low latency

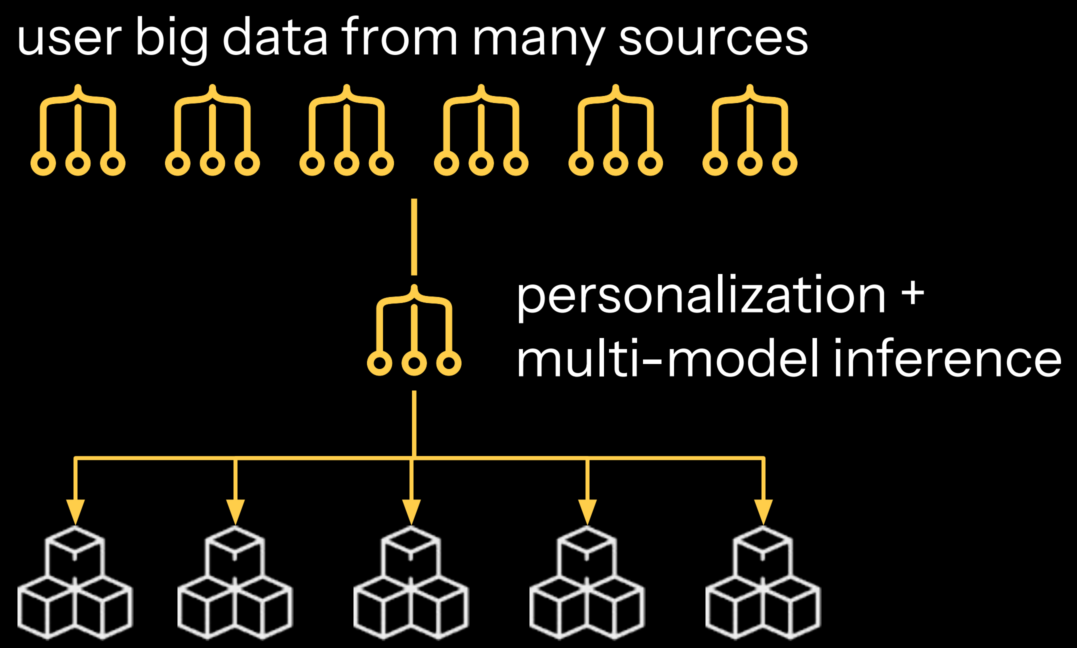
3+ million TPS
71ms p(99) latency



Tubi [Fox]

Tubi applies ML models to real-time streams of data with in-memory, durable journals

2 month time to delivery


user big data from many sources
personalization + multi-model inference

Horn

"Zero problems" augmenting high-performance audio and video streams on demand

Tomasz Wujec - Lead Developer

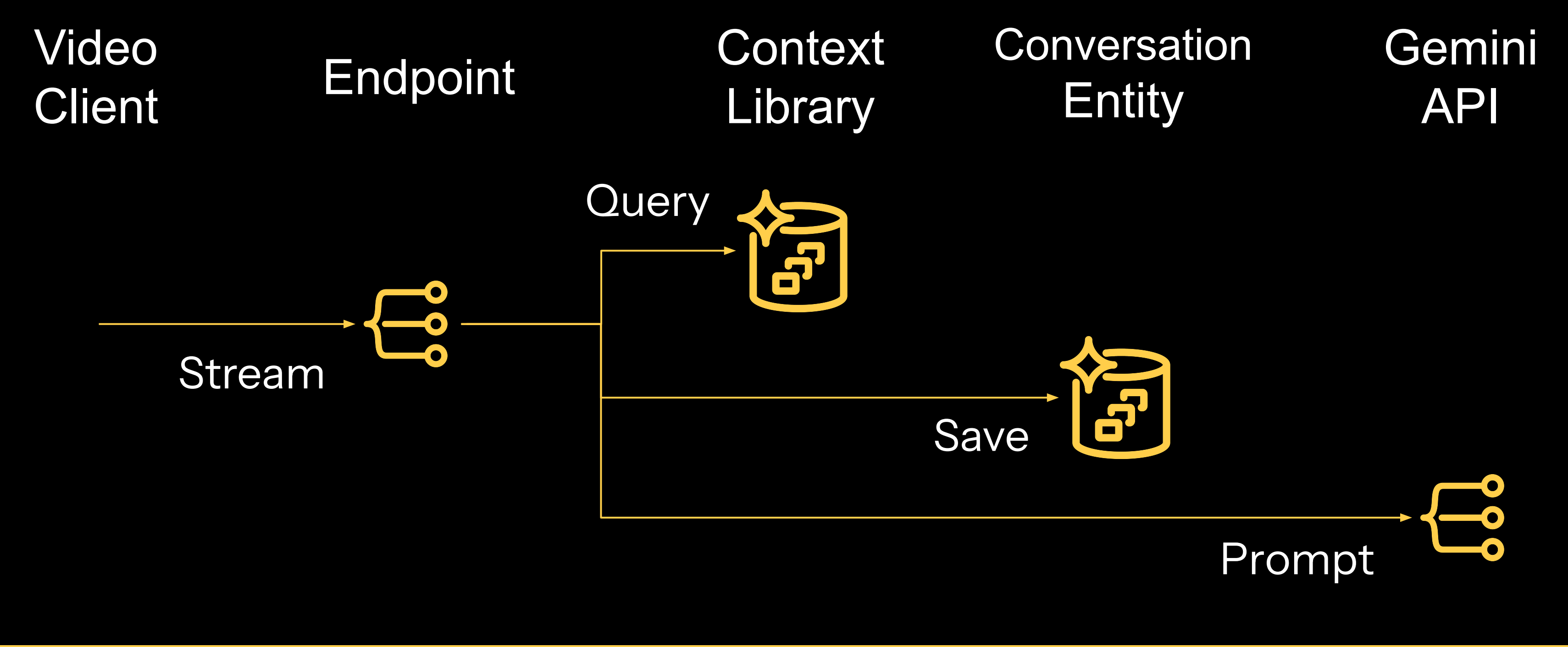
Coho AI

"With Akka, we got to market 75% faster compared to other agentic solutions we had considered."

Michael Ehrlich – CTO

Streaming Video Demo

AKKA



Endpoint	gRPC API for receiving video
Context Library	Database of augmentation contexts saved as an Akka entity
Conversation Entity	Database of responses from Gemini
Gemini API	Google's API entry point

Velocity

145 LOC

2 components
gemini protocol client
4 integration tests
private GCP region
1 developer

2.5 days

concept to delivery

Resilience

2x data redundancy

1 multi-cloud region; local data backup; add 2nd region for replication

99.9999% SLA for your apps → apps running across regions are nine-nines ready

n/a ms

failover RTO

0ms

data RPO

Akka resilience guarantee - indemnities for reliability failures

Cost Efficiency

$150 / day

commodity cloud cost

10K TPS potential w/ this config

1M TPS

potential Akka write throughput

21M tx per $1 of cloud cost

Concept-to-production in 8 weeks

1. Choose your agentic architecture	➔ RAG, cooperative multi-agent, environment controller, or self-learning
2. Select the right AI model	➔ Prompt-based agents: GPT-4, Claude, Gemini, Mistra, Llama 2 ➔ Embedding-based search agents: OpenAI Ada, Cohere, Google Vertex AI ➔ Fine-tuned industry models: Falcon, Mixtral
3. Stand up agentic platform regions	InfoSec Review - Akka meets 19 levels of compliance including SOC 2 type 2 ➔ Cloud: Akka Serverless ➔ Edge: Akka Edge ➔ Private: Akka BYOC
4. Stand up AI inferencing	➔ Cloud AI: OpenAI API, AWS Bedrock, Azure AI, Google Vertex AI ➔ Self-Hosted AI: Ollama, vLLM, TGI ➔ On-device AI: GTP4All, LM Studio
5. Build, test, debug and optimize	➔ Build agents and agentic services offline with Akka's SDK ➔ Add human-in-the-loop features for oversight ➔ Run real-world performance, functional, and penetration simulations
6. Deploy and observe	➔ Setup API rate and cost limits to prevent abuse ➔ Record, track, and export performance or traces ➔ Monitor AI behavior for hallucinations or errors

Agentic is real
Let’s make it real for you

concept

proof

48 hours

Live Q&A

Thank You