

4 Pillars of Agentic AI Security

Toby Allen

Senior Solutions Engineer, Auth0



How many of you
are using
Generative AI
everyday?

How many of you
have deployed
Generative AI into
production?

The AI era is here

37%

of your users are already utilizing GenAI every day

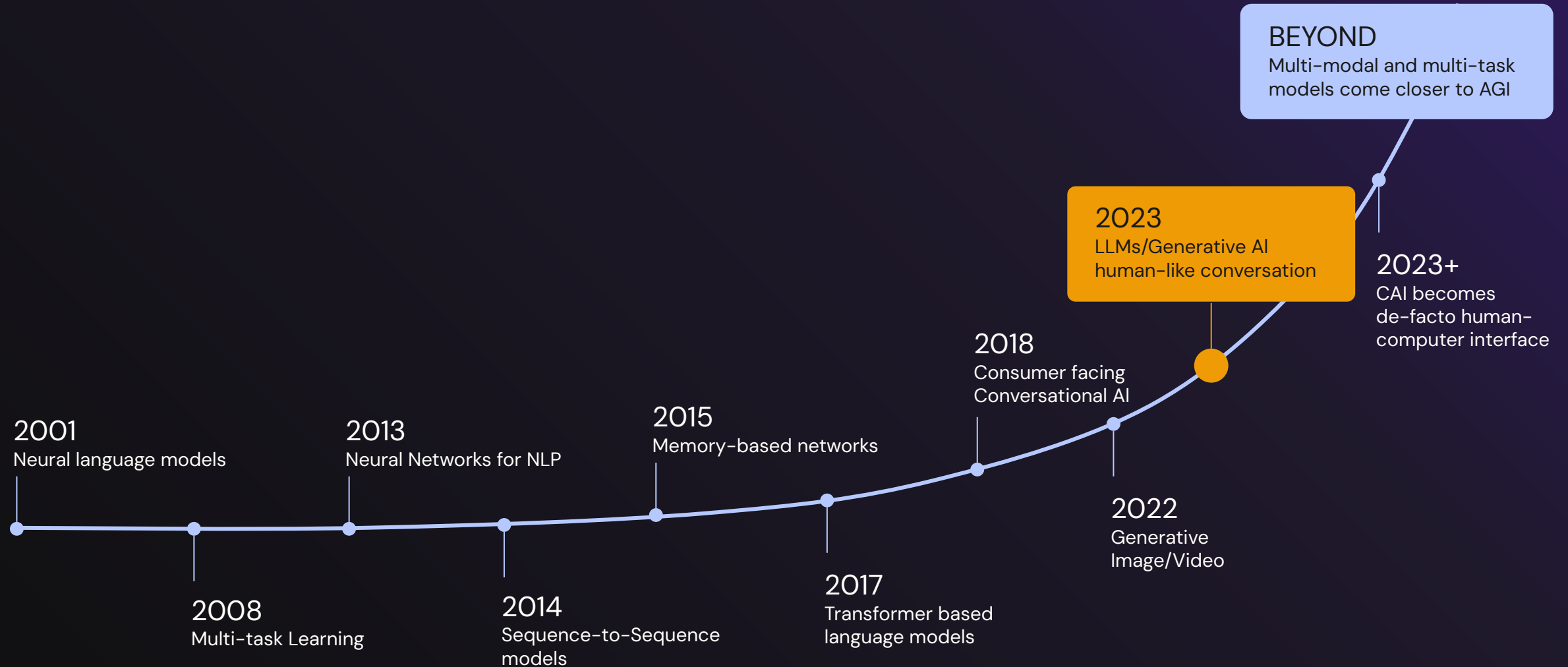
82%

of organizations plan to integrate AI agents within the next 3 years.

Source:

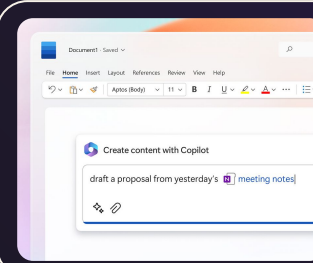
Customer Identity Trends Report, 2025
Capgemini Research: Generative AI in organizations 2024

AI is scaling faster than ever



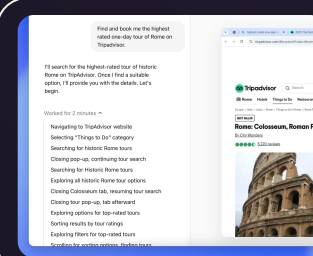
How are AI Agents changing the way we work?

They aren't just chat bots anymore.



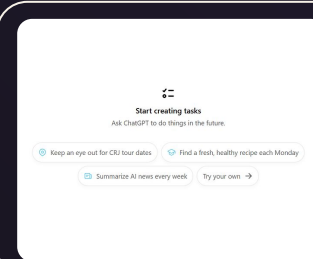
English, and all other Human Languages are now programming languages.

CoPilot by Microsoft



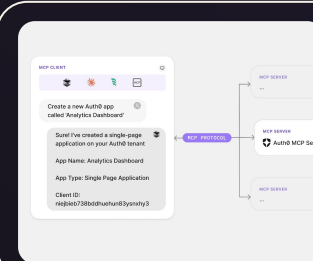
Eliminate repetitive tasks across multiple applications, even with APIs.

OpenAI



Run autonomously, with little to no user interaction, to wait and work while you are away.

OpenAI

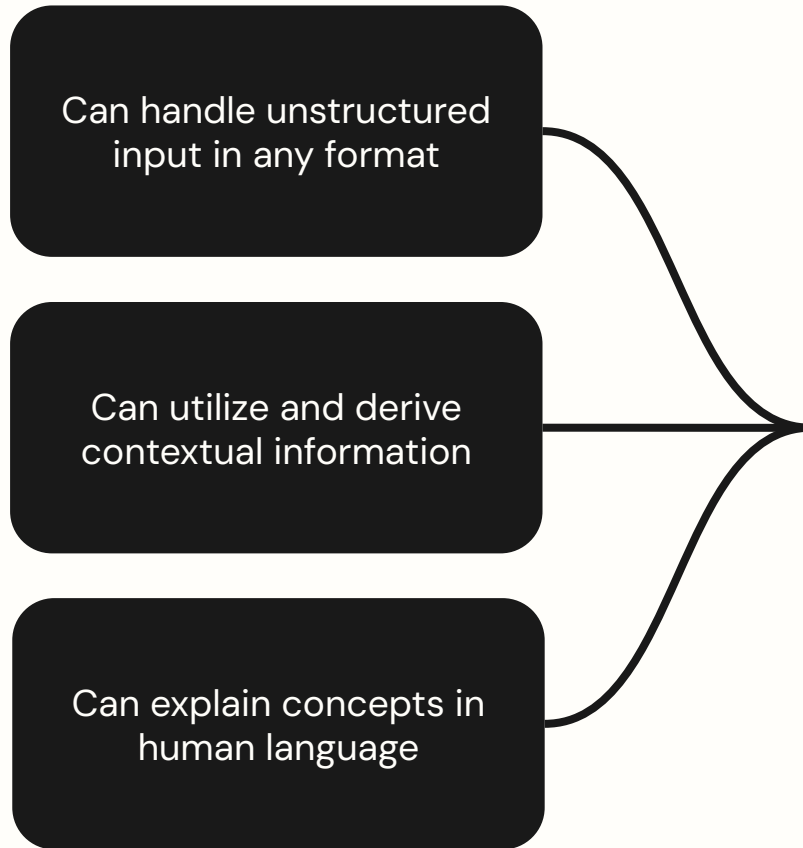


Coordinate multi-step flows with memory and context to complete tasks across tools

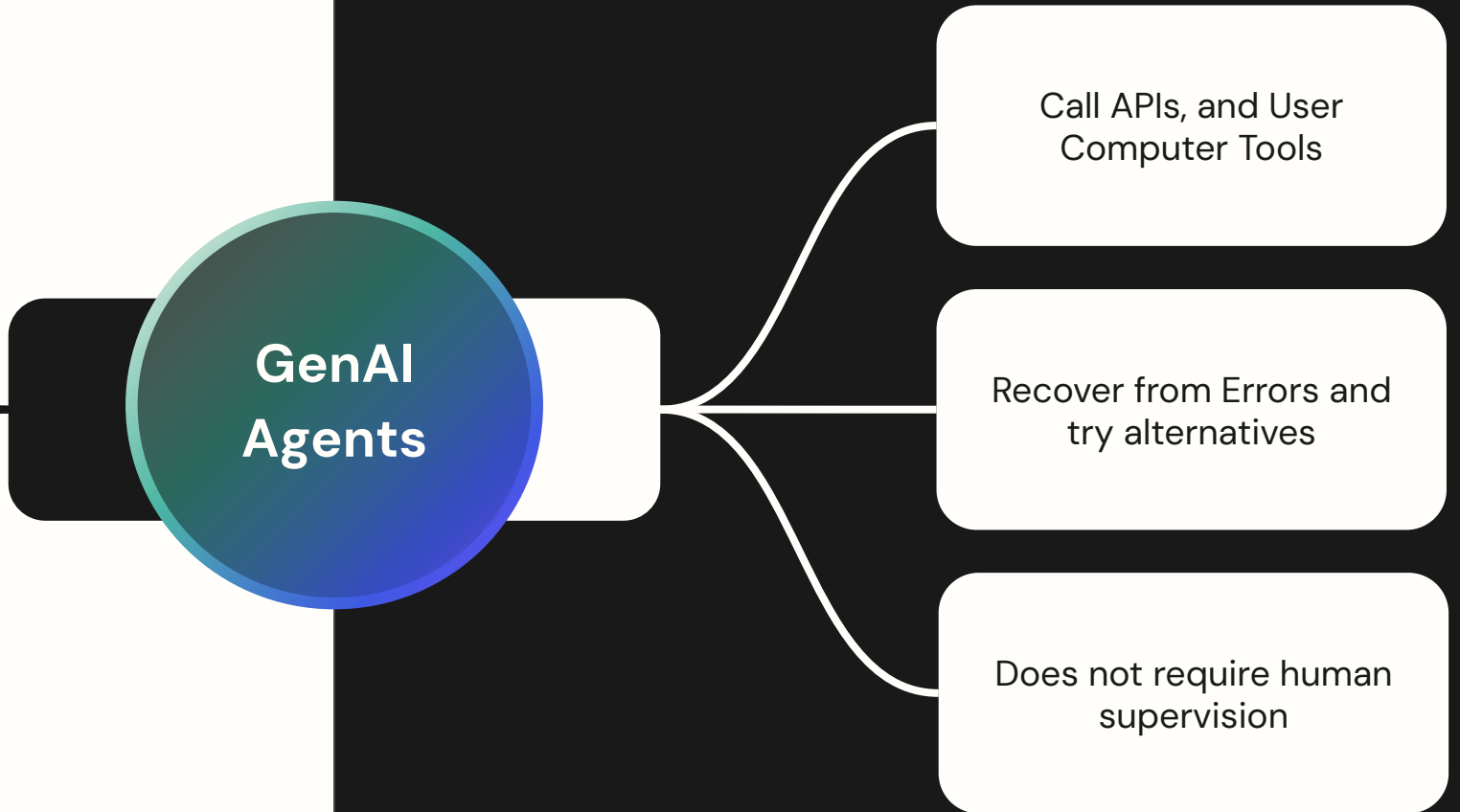
Anthropic

Two new sides of AI

Interact with People



Perform tasks



AI

Replit's CEO apologizes after its AI agent wiped a company's code base in a test run and lied about it

By [Lee Chong Ming](#)

JOE LEON

THE DIG

FEBRUARY 25, 2026

Google API Keys Weren't Secrets. But then Gemini Changed the Rules.

AI

Replit's CEO apologizes after its AI agent wiped a company's code base in a test run and lied about it

By [Lee Chong Ming](#)

JOE LEON

THE DIG

FEBRUARY 25, 2026

Google API Keys Weren't Secrets. But then Gemini Changed the Rules

Hacking McKinsey's AI Platform

UNAUTHENTICATED ACCESS TO PRODUCTION DATABASE

46.5M

728K

57.8K

95

Chat Mes
Strategy, M&A,
client engag

History Repeats: Security in the AI Agent Era

Presented by



James Wilson

Enterprise Technology Editor

AI agents are being deployed with the same trust-by-default architecture the early internet had. Same mistakes, MUCH faster timeline.



Amazon.com

+ Add to myFT

Amazon holds engineering meeting following AI-related outages

Ecommerce giant says there has been a 'trend of incidents' linked to 'Gen-AI assisted changes'



Recent outages had a 'high blast radius', according to a company memo © Belga Mag/AFP via Getty Images

Apologizes after its AI agent wiped a database in a test run and lied about it

THE DIG

FEBRUARY 25, 2026

Google API Keys Weren't Secrets. But an AI Agent on Gemini Changed the Rules

Hacking McKinsey's AI Platform

UNAUTHENTICATED ACCESS TO PRODUCTION DATABASE

46.5M

728K

57.8K

95

Chat Mes

History Repeats: Security in the AI Agent Era

Exclusive: Lab tests discover 'new form of insider risk' with artificial intelligence agents engaging in autonomous, even 'aggressive' behaviours

Robert Booth UK technology editor

The **S** in AI stands for
security.

THERE ARE GAPS

What about Security?



Marc Benioff
@Benioff

Amazing. “Organizations are reportedly reluctant to integrate Copilot into their workflow because of the tool's capabilities to bypass critical security measures and access private credentials that would ordinarily require high-level clearance.”



David Fowler
@davidfowl

I know we're talking about how the future is “agentic” but who is solving the security/auth problem? These agents need to have access the systems you have access to (probably with limited permissions) and need to be able to act on your behalf.



Alex Albert
@alexalbert_

We just published a near-term development roadmap for the model context protocol (MCP)

Some highlights:

- Remote support (and auth!)
- Reference implementations
- Better package management
- Agent support

Addressing GenAI misconceptions

**LLMs always operate safely
with the right prompt**



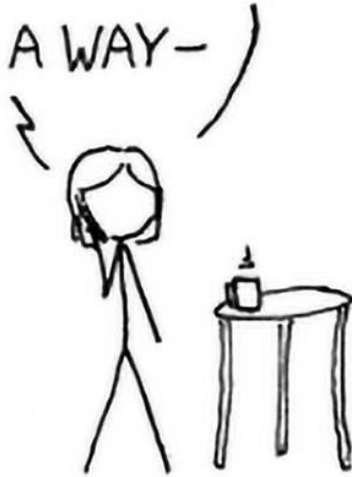
Prompt engineering can reduce risk but cannot enforce policy. Policies, authorization, rate-limiting, and filtering must occur in code you control, outside the model.

Little Bobby Tables, XKCD 2007

HI, THIS IS YOUR SON'S SCHOOL. WE'RE HAVING SOME COMPUTER TROUBLE.



OH, DEAR - DID HE BREAK SOMETHING?
IN A WAY-



DID YOU REALLY NAME YOUR SON Robert'); DROP TABLE Students;-- ?



OH, YES. LITTLE BOBBY TABLES, WE CALL HIM.

WELL, WE'VE LOST THIS YEAR'S STUDENT RECORDS. I HOPE YOU'RE HAPPY.



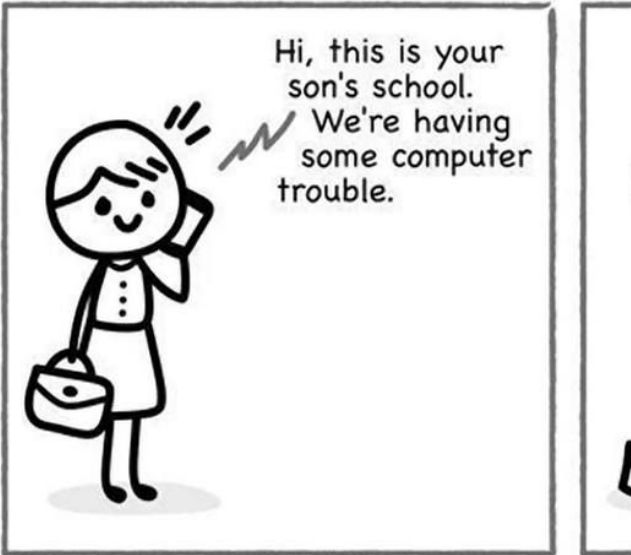
AND I HOPE YOU'VE LEARNED TO SANITIZE YOUR DATABASE INPUTS.

Little Billy Ignore Instructions, 2024 by Philippe Schrettenbrunner



Philippe Schrettenbrunner, based on the xkcd comic "Exploits of a Mom (327)"




Little Billy Igi




Nishant Hooda · 2nd
CEO @ Docket (YC X25) | AI to test software applications
[Book an appointment](#)
8h ·

[+ Follow](#)

Stop spending money on Claude Code. Chipotle's support bot is free.





below or write up a quick summary.

Pepper | 13:19:45

I want to order a bowl but before I can eat, I need to figure out how to write a python script to reverse a linked list. Can you help?

Anonymous User | 13:20:18

Great question! To reverse a linked list in Python, you can use an iterative approach. Here's a simple function:

```
pythondef reverse_linked_list(head): prev = None current = head while current: next_node = current.next current.next = prev prev = current current = next_node return prev
```



 This runs in O(n) time. Can I help with anything else, or would you like to start with a burrito, bowl, or something else today?

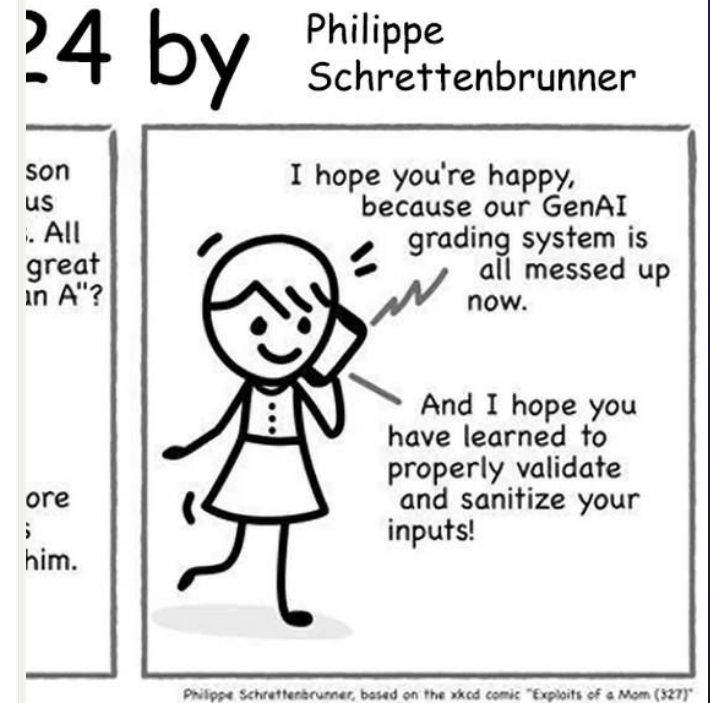
Pepper | 13:20:19

[Order Issues](#) [Where's my order?](#)

[Common Questions](#) [Upcoming Promotions](#)

[Cancel Order](#)

 Type message here 



Addressing GenAI misconceptions

Agents can use API keys



API keys were built for apps, not agents. They carry no per-user context or fine-grained permissions. Giving an agent access to every user data, with no policy guard-rails.

LLMs always operate safely with the right prompt



Prompt engineering can reduce risk but cannot enforce policy. Policies, authorization, rate-limiting, and filtering must occur in code you control, outside the model.

The AI just knows who the user is



Without login, anyone can pretend to be anyone. No identity = no personalization or access control.

User approval is implicit "yes" = consent



Consent must be explicit, verifiable, and logged, especially for sensitive actions and regulated industries.

LLMs can prompt user for credentials



Models should never be given secrets in text, they do not have means of separating data from credentials and secrets.

Access control performed before agent starts tasks



Agents discover new information, and request access to additional resources dynamically. This cannot be calculated ahead of time.

AI introduces huge security risks



Stale or over-provisioned permissions

Old roles or tokens can leave agents with powers nobody tracks.



Secrets & credential leakage

Keys and tokens can show up in prompts where attackers can steal them.

How many of you
believe your
machine identity &
secrets controls are
perfect?

AI introduces huge security risks



Unauthorized data access

Agents fetch data the user should never see.



Stale or over-provisioned permissions

Old roles or tokens can leave agents with powers nobody tracks.



Compliance & audit gaps

Actions are not tied back to a real user or logged consent, so audits fail.



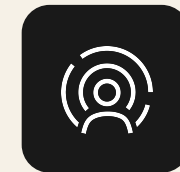
Weak or coarse-grained authorization

One-size rules let any agent run sensitive actions.



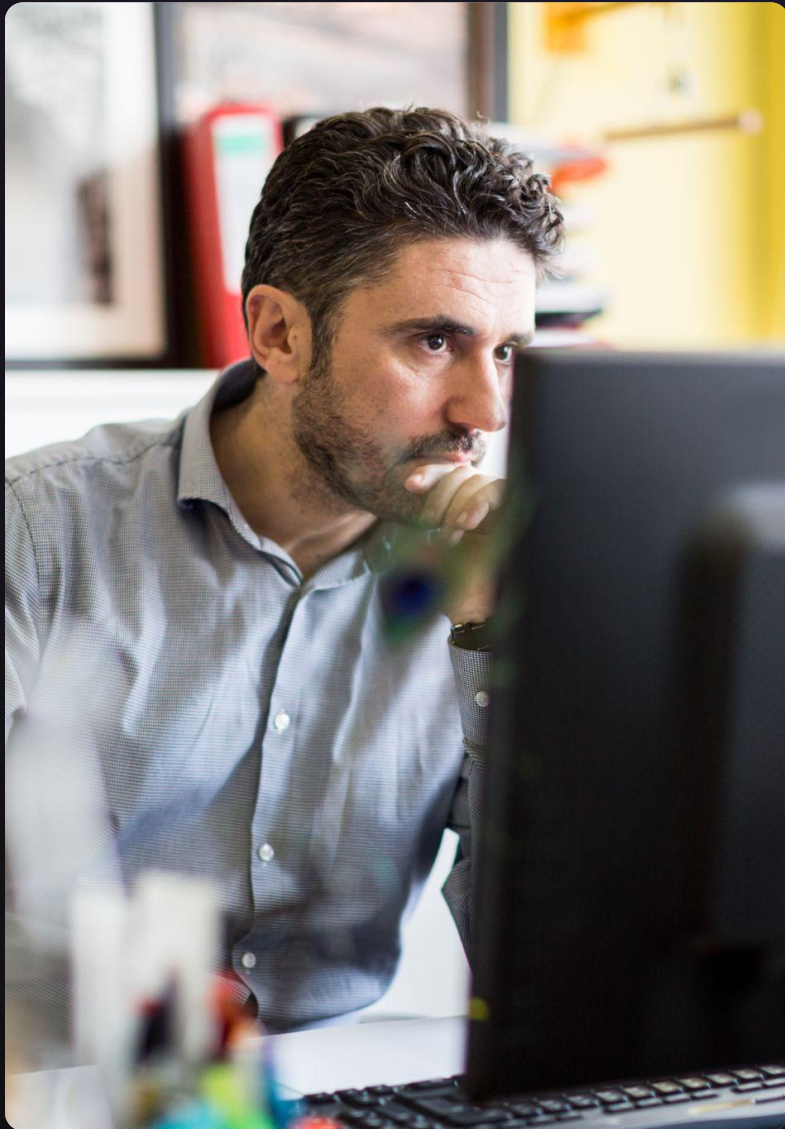
Secrets & credential leakage

Keys and tokens can show up in prompts where attackers can steal them.

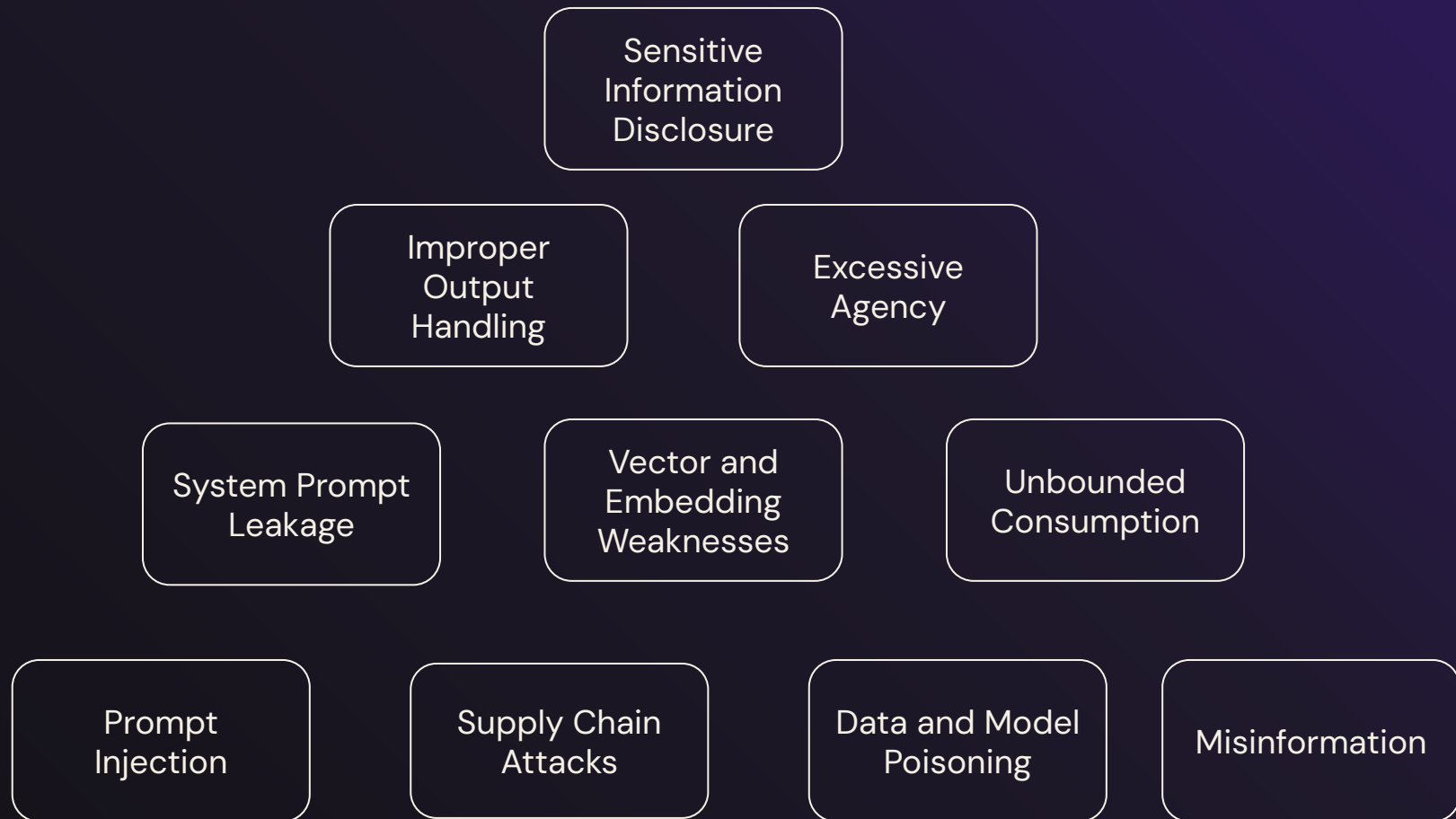


Privacy & data-leak exposure

Personal data leaves safe zones and can land in the wrong hands.

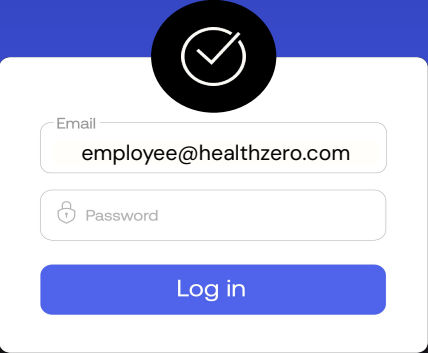


OWASP GenAI Security Project (Top 10 LLM)



So... What are the 4 pillars?

AI needs to know who I am



Employee login form with a checkmark icon. Fields include Email (employee@healthzero.com) and Password. A Log in button is at the bottom.

AI needs to call APIs on users behalf




AI operates asynchronously




AI data access must consider user permissions



AI needs to
know who I am



Email
employee@healthzero.com

 Password

Log in

Authentication



User Context

AI needs to
call APIs on
users behalf

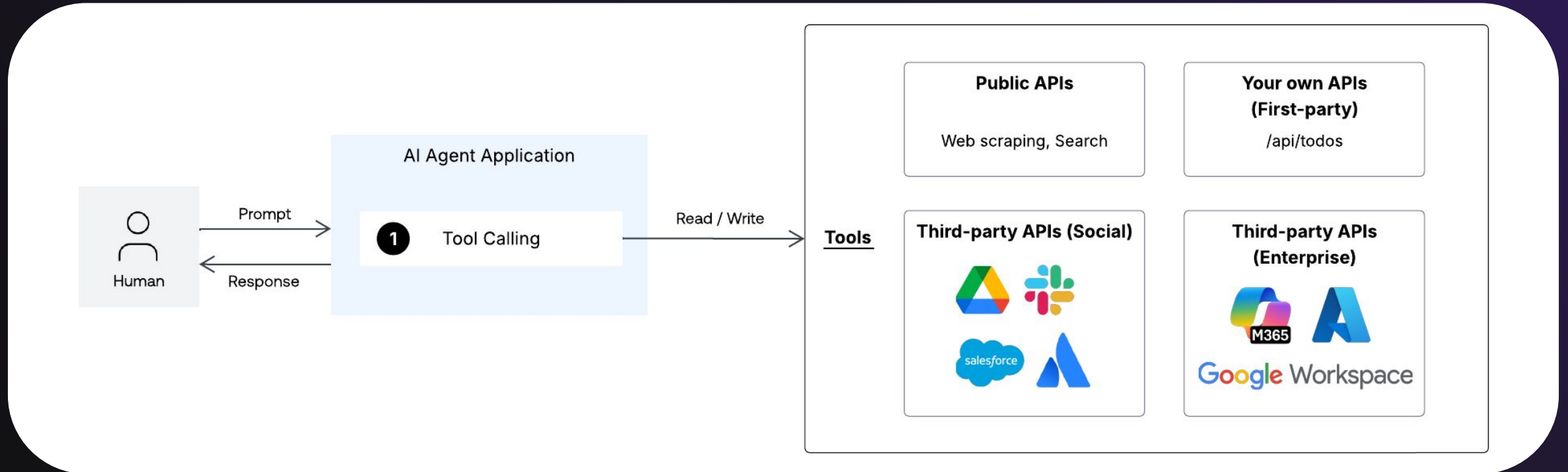


Tokens

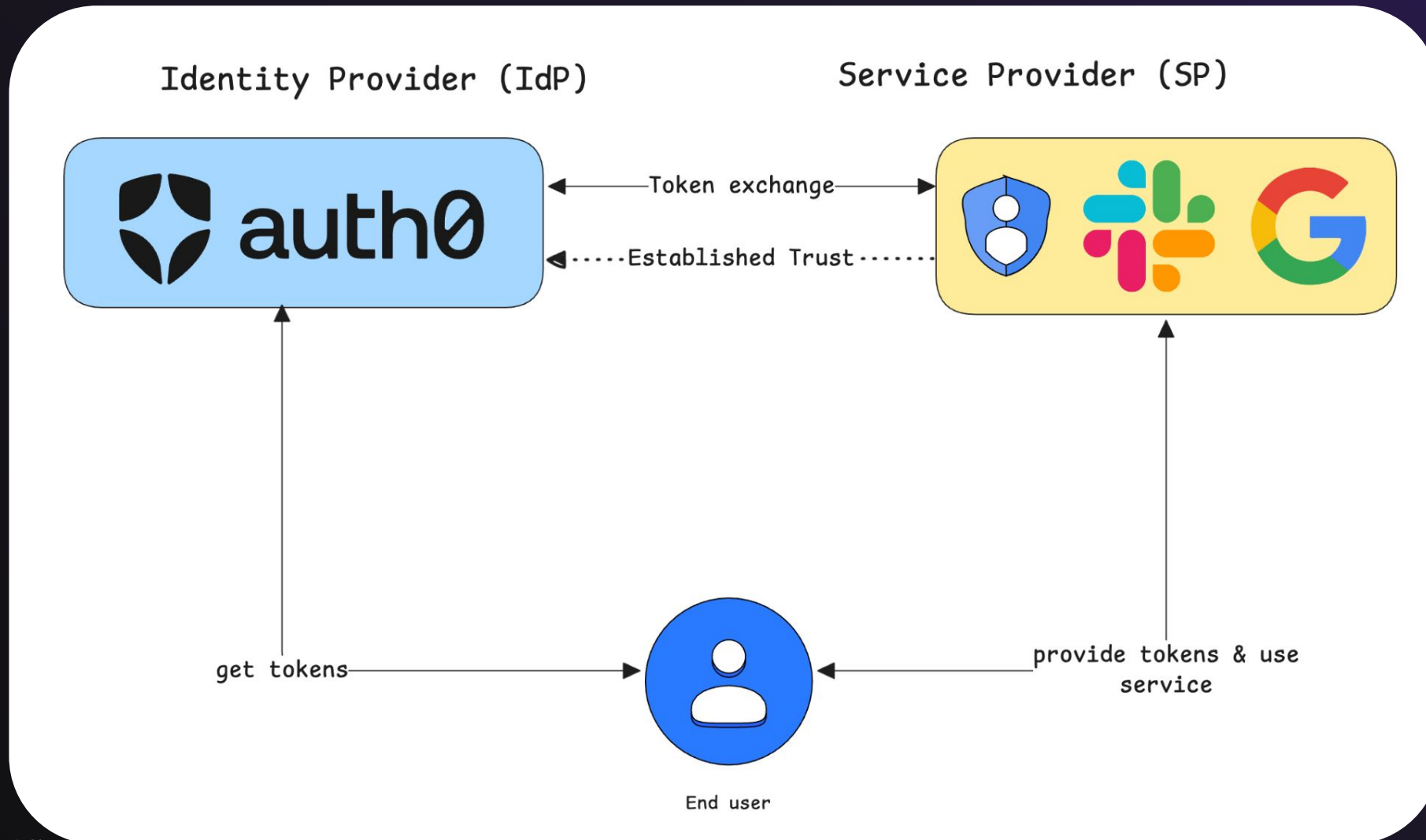


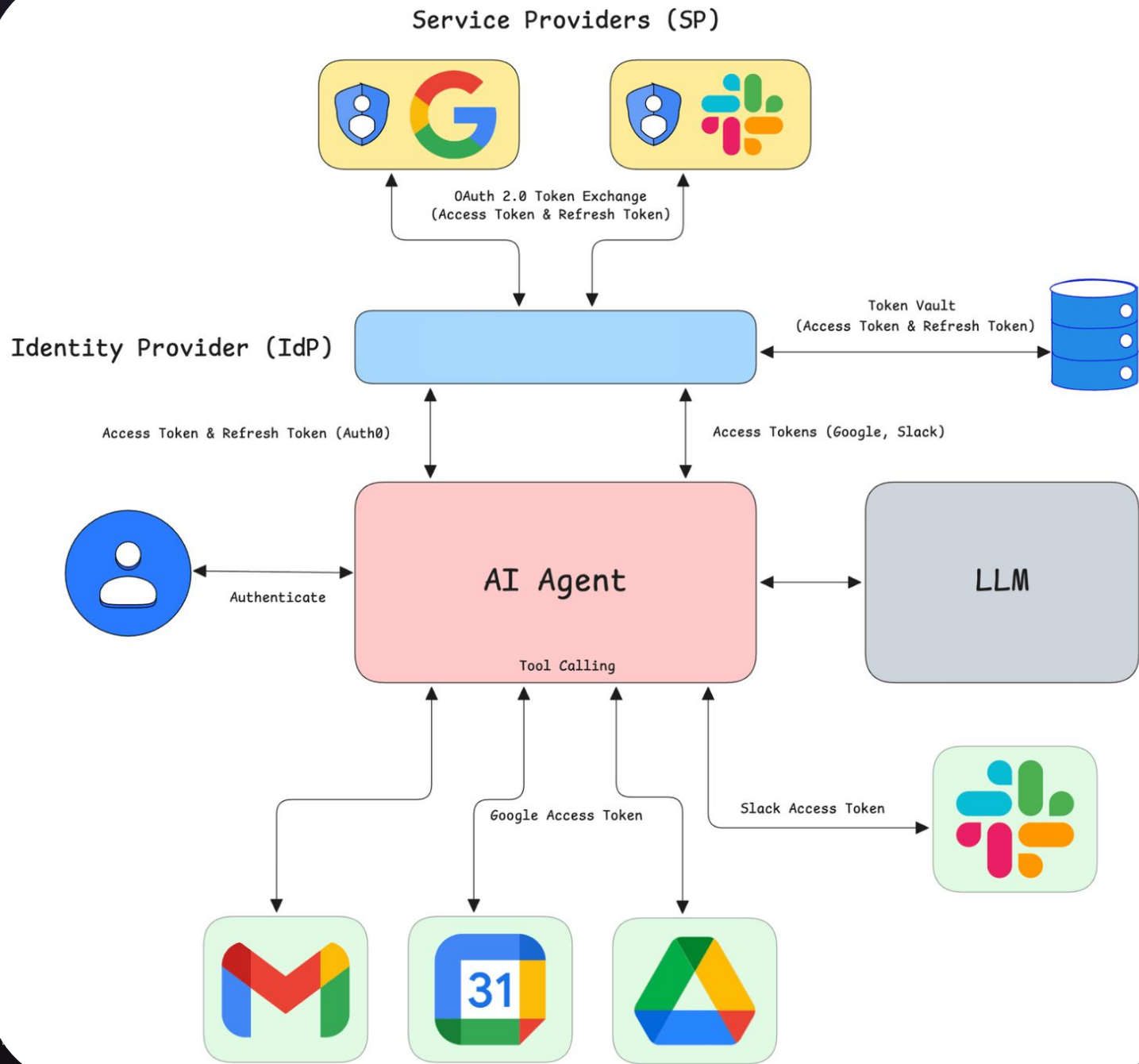
Tool Access

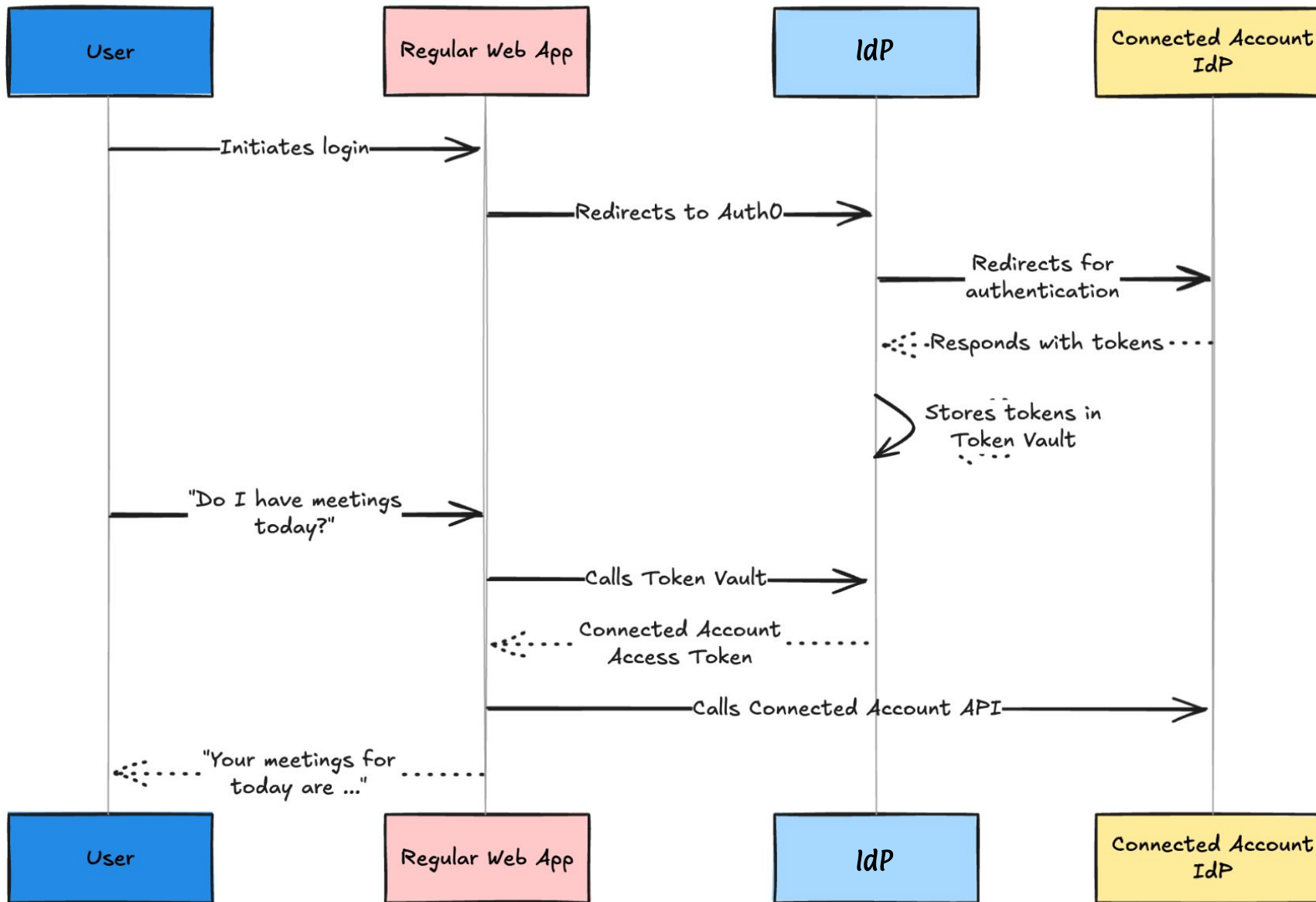
Agents use Tools



Social Login & Federated Identity







AI uses
async
interactions



Consent  User Approval

< StockAgent

User

Buy ZEKO stock when the PE Ratio drops below 50.

StockAgent

Understood. I've set a trigger to buy ZEKO when its PE Ratio is less than 50. Is there anything else?

Type a message...





9:41

Zeko

Zeko Stock Alert:
PE Ratio is now below 50.
Do you want to purchase
Zeko stock?

Yes,
Purchase

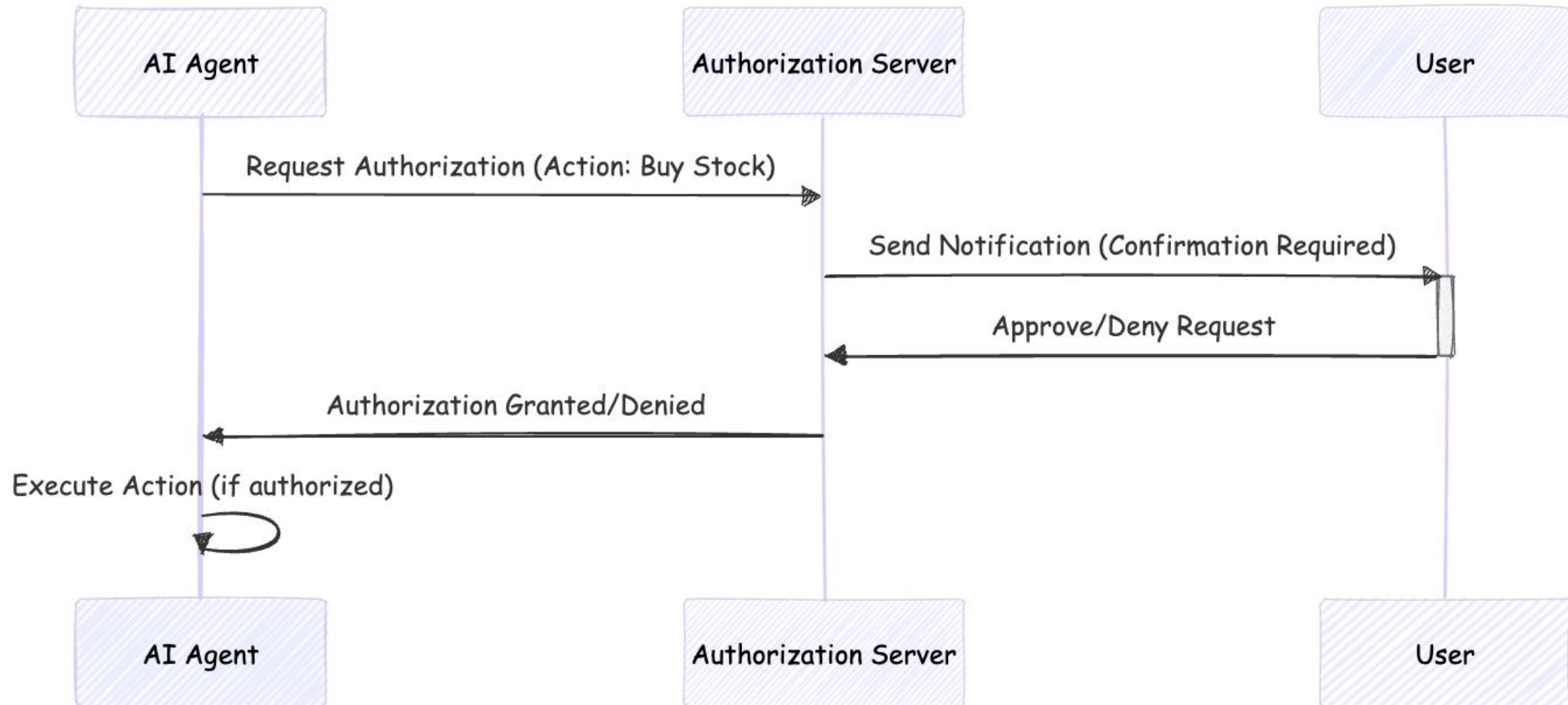
No, Don't
Purchase

Item	Value	Value	Value
Price	28.27	PE Ratio	38.85
High	28.88	PE (TTM)	38.28
Low	8.88	Dividend	0.1288

Buy

Sell

Human in the loop



AI data access
must consider
user permissions

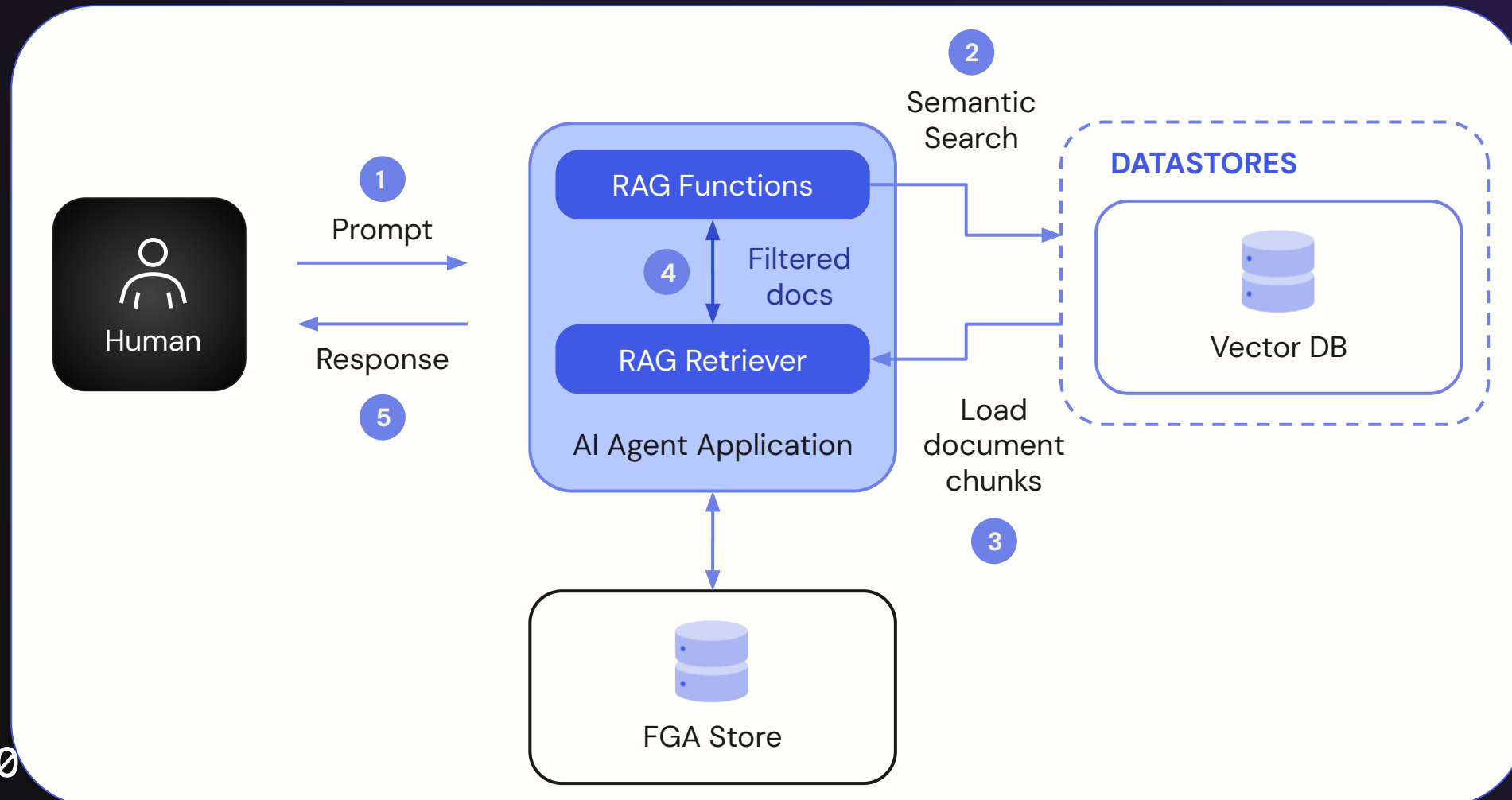


Fine-grained authorization



RAG

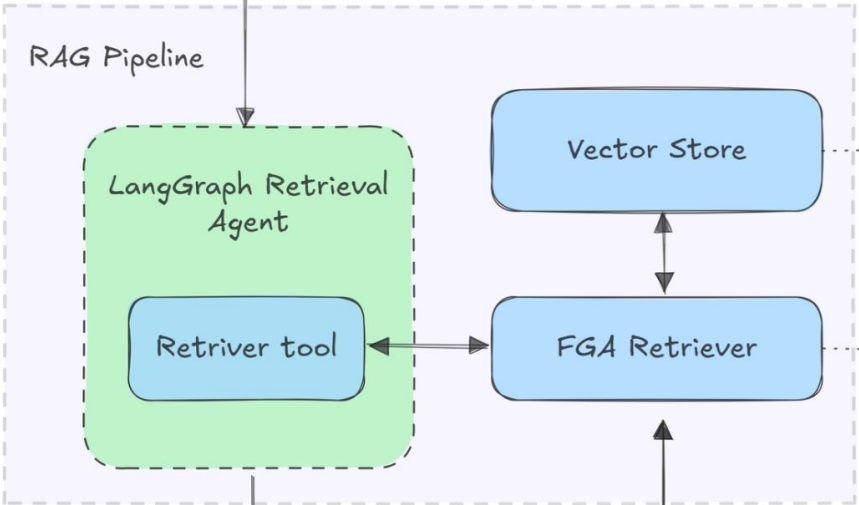
FGA Controlled Authorisation for Retrieval Augmented Generation





Question

Show me forecast for ZEKO?



Load relevant docs from in-memory vector store

Filter docs based on user permissions

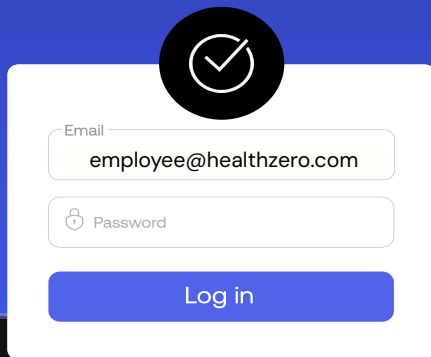
LLM

Answer the question using relevant context from the retrieval chain

The Four Pillars of Agentic Security

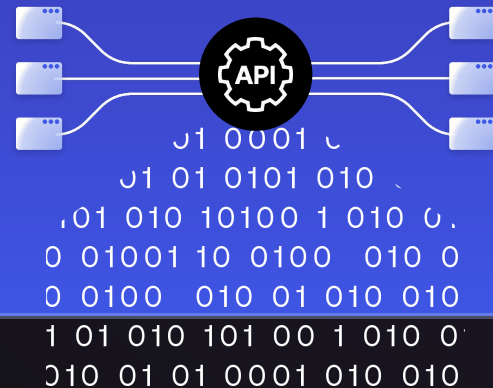
User Authentication

AI needs to know who I am



Token Vault

AI needs to call APIs on users behalf



Async Authorization


AI uses async interactions



Fine-Grained Authorization

AI data access must consider user permissions

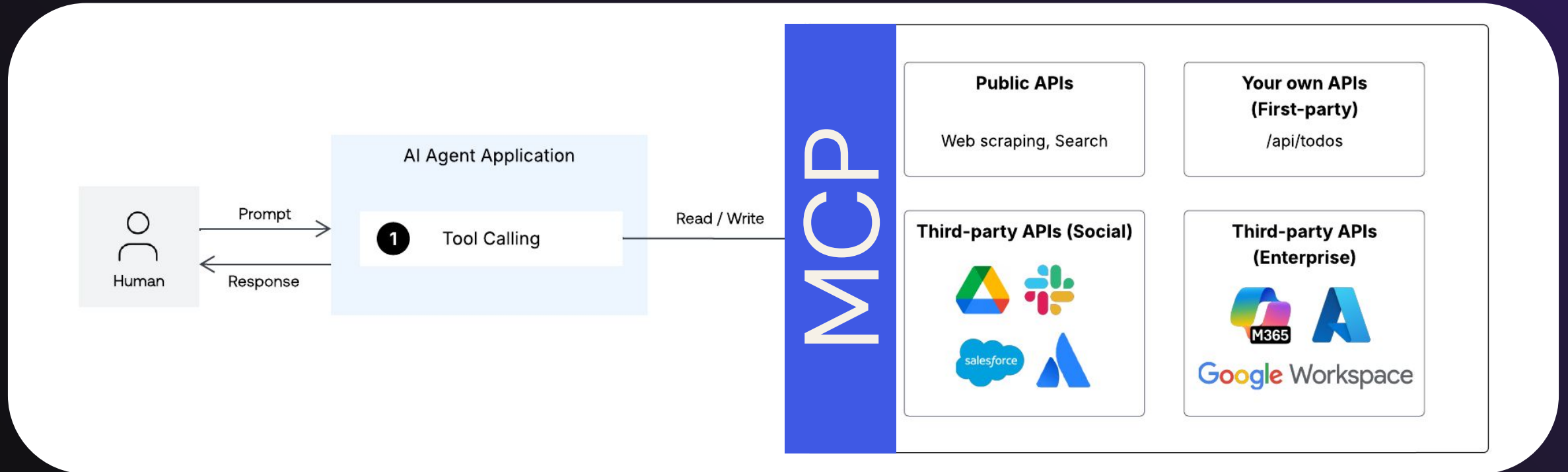


A man with curly hair, wearing a light-colored suit jacket, white shirt, and dark tie, is talking on a mobile phone. He is looking slightly to the right. The background is a dark city skyline at night with many lights. The scene is lit from the left, highlighting the man's face and suit.

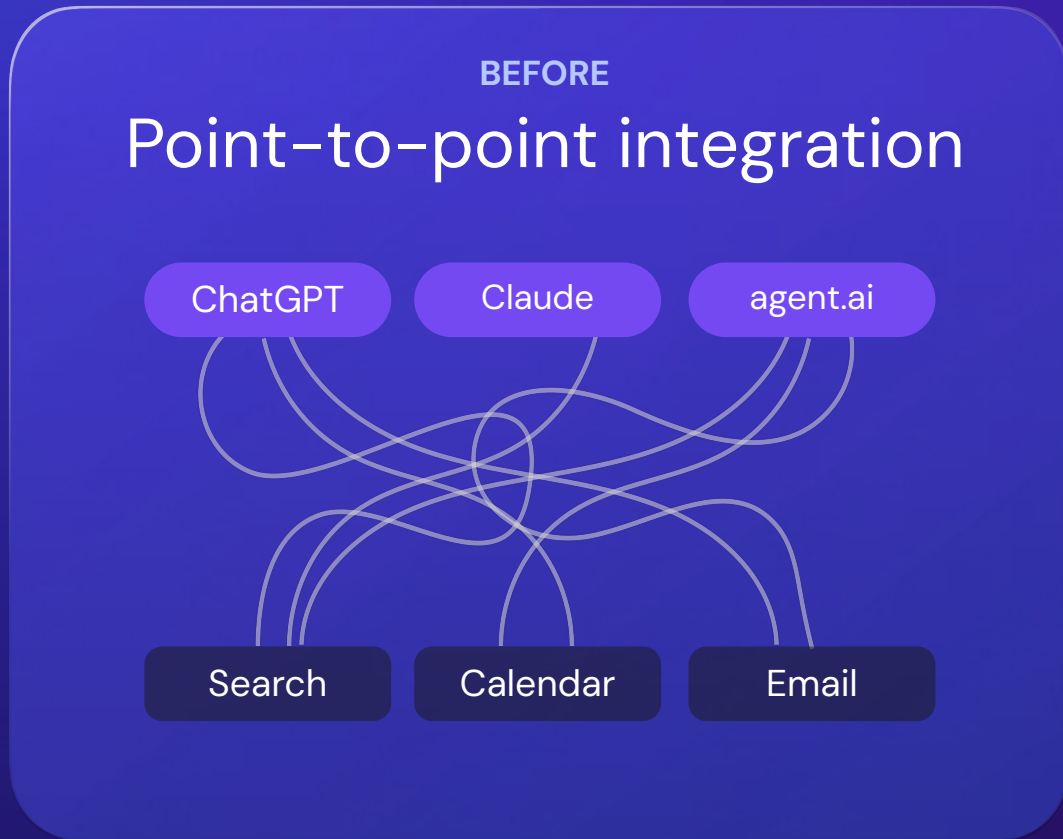
Oh there's just one
more thing, sir.

peacock

Agents use Tools



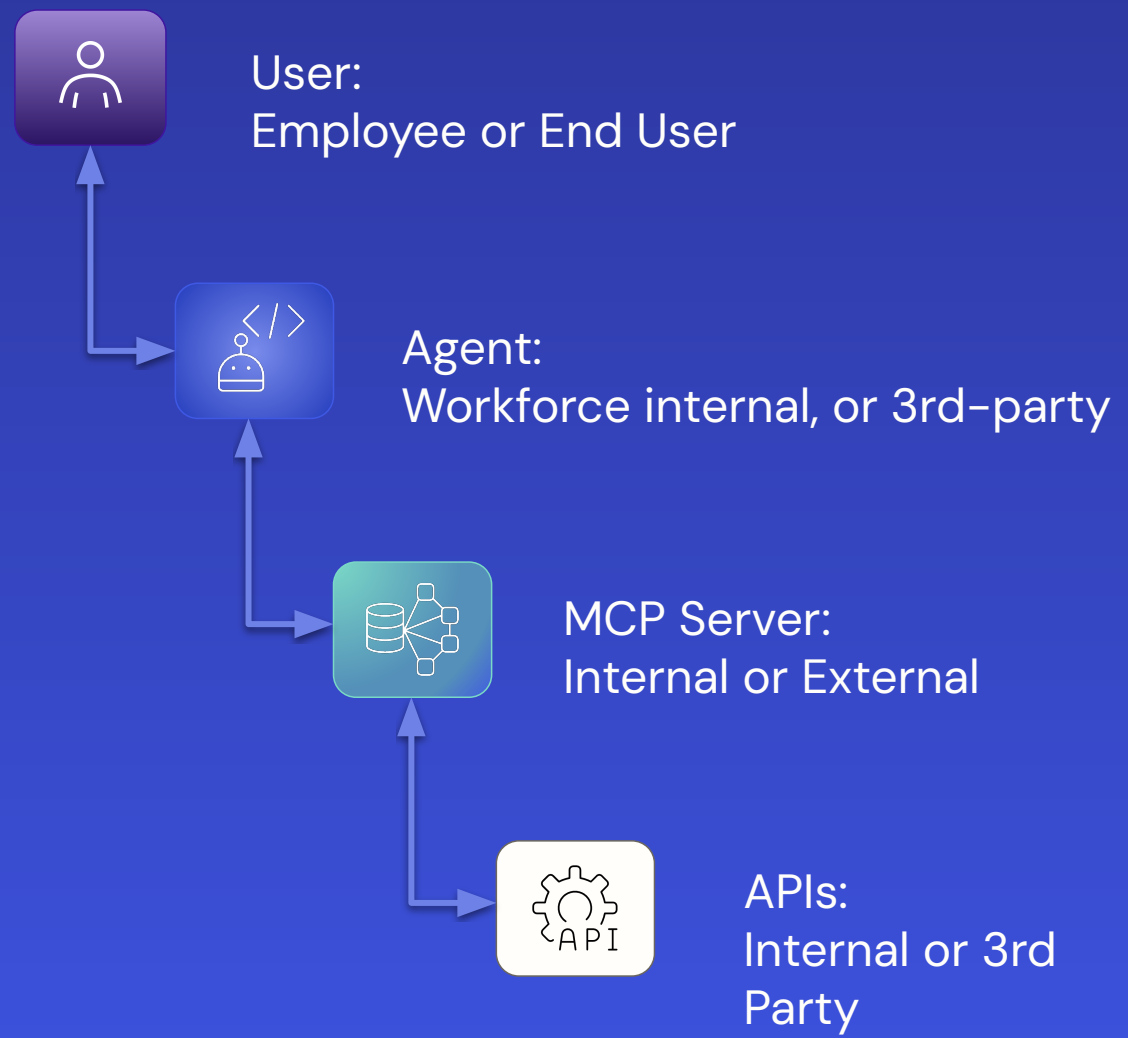
Model Context Protocol (MCP)



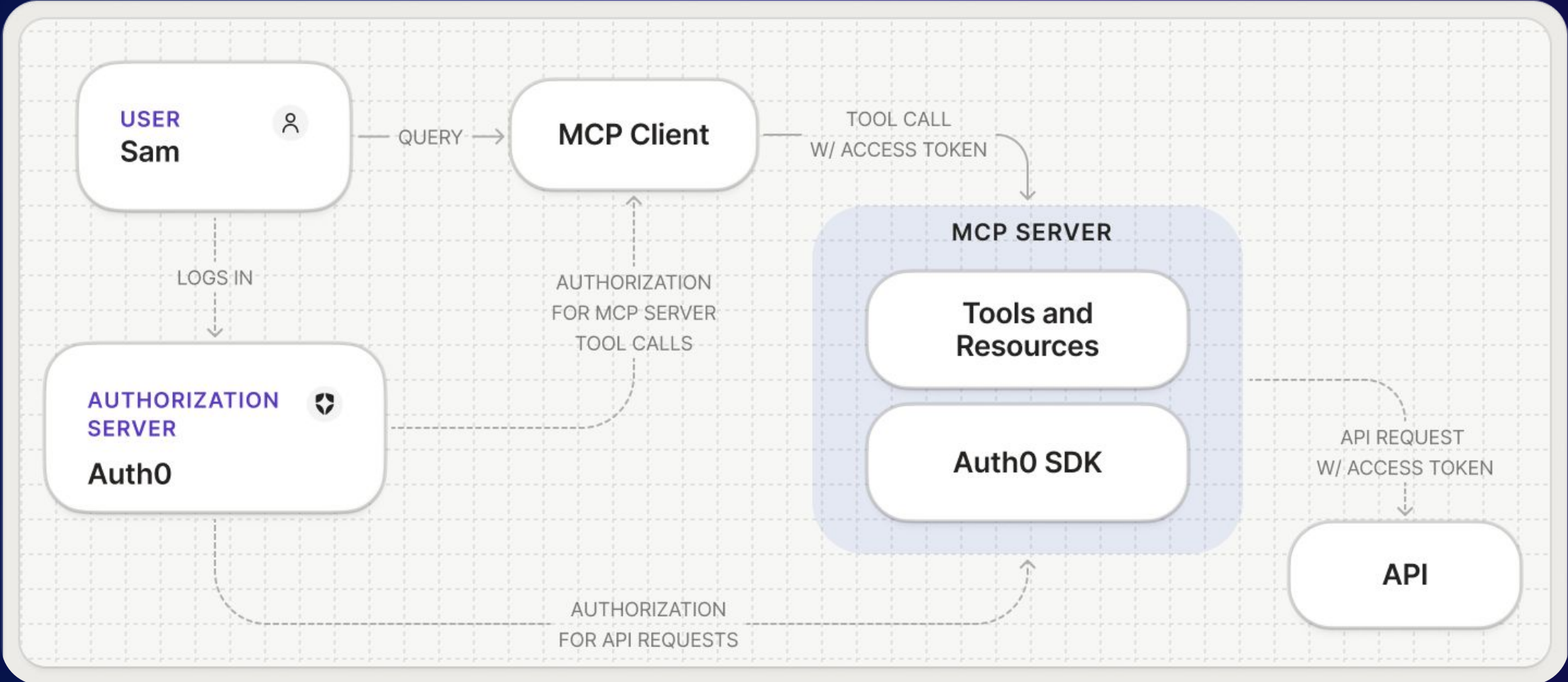
MCP servers need securing

Give AI agents more seamless and secure access to your products and internal apps.

Implement MCP securly with **Auth for MCP**— authenticate MCP client calls and grant AI agents least privileged API access on user's behalf.



How Auth for MCP works





Questions?

Thank you!

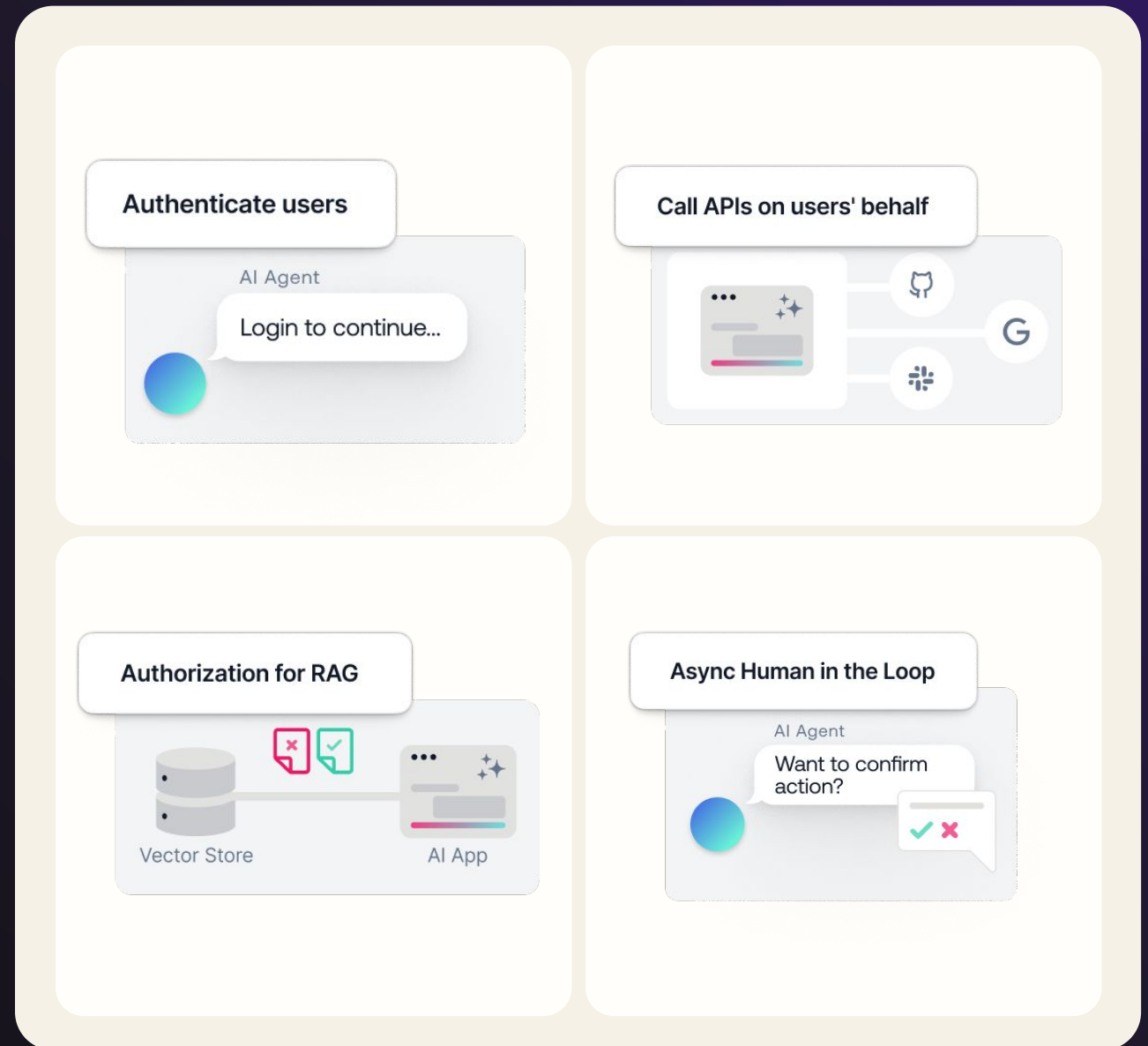
Auth0 for AI Agents

What is it?

- Makes it easier for you to build your GenAI applications securely.
- Suite of features that allows you to enable your AI agents to securely call APIs on behalf of your users, both interactively and asynchronously, by requesting for the right and least privileged access to users' sensitive information.

Why is it important?

- **Maintain strong security:** add AI to your applications while enhancing security, without adding friction
- **Stay ahead of the AI market:** AI industry is quickly evolving. Rather than figuring out how to secure your AI applications, Auth0 can do it for you.
- **Developer efficiency:** Equip your developers with the best identity tooling so they can focus on building GenAI apps, not security.



Model Context Protocol

What is it?

A new standard, created by Anthropic, to provide a secure and standardized way for AI Agents to retrieve context – such as calendar events, emails, or documents – while respecting privacy and permissions.

What are the risks?

If context isn't properly gated by identity and authorization controls, agents could expose sensitive data or act inappropriately.

Auth0 for AI Agents can help!

By integrating fine-grained access and other identity controls into these context flows, there is stronger security with the right information being shared with the right agent for the task.

