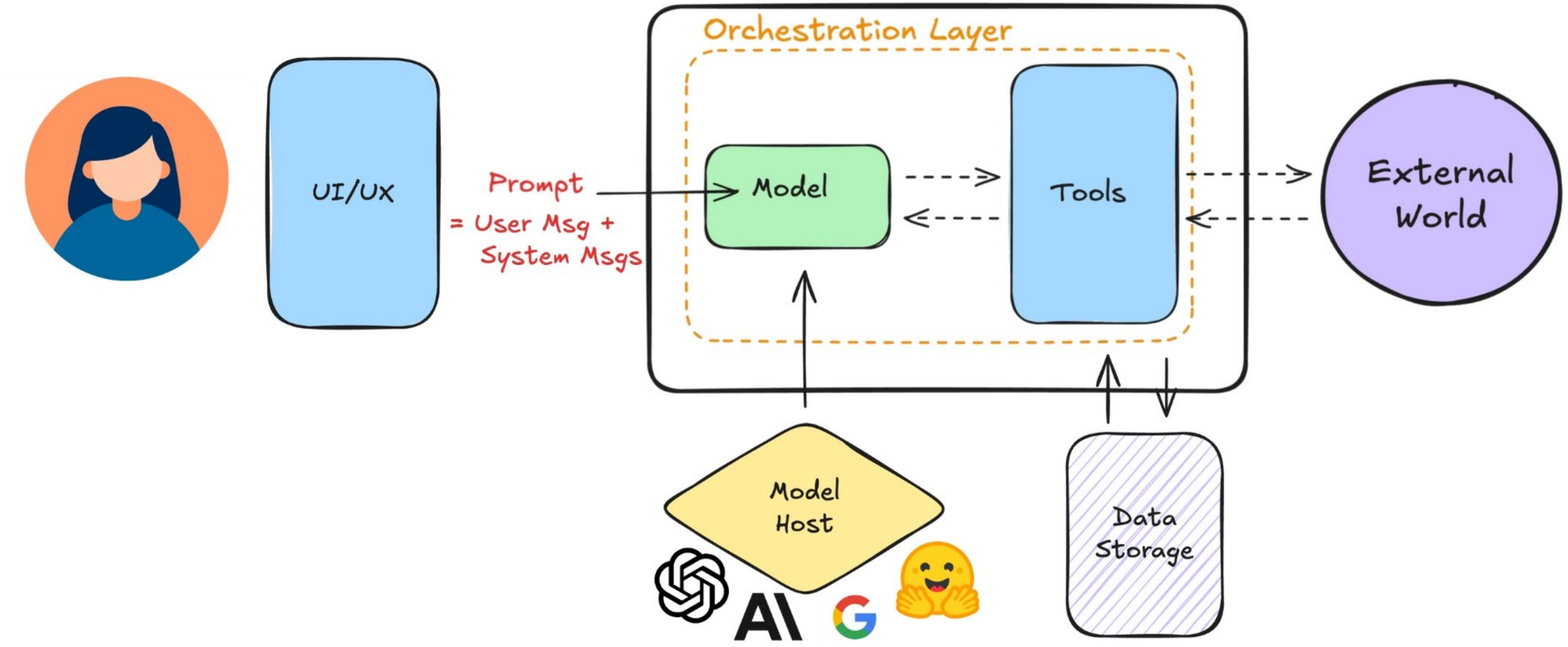
From Proof-of-Concept to Production

BUILDING SCALABLE AGENTIC SYSTEMS

Korey Stegared-Pace
Senior AI Cloud Advocate, Microsoft

datacamp

The bigger picture



Step 1: Validate Real Interactions

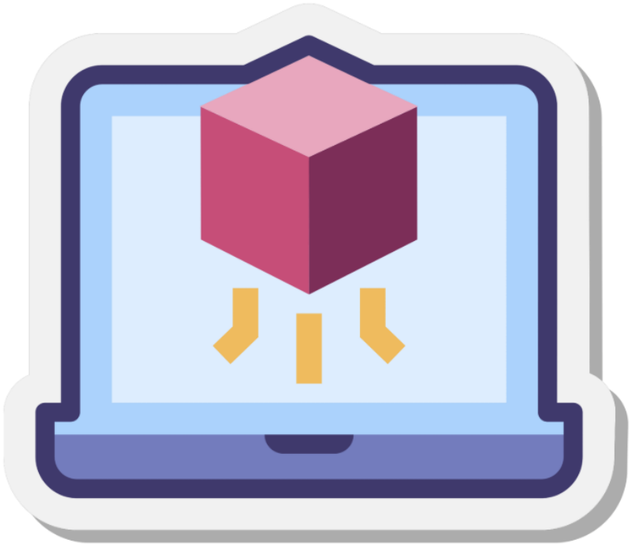
> Testing leads to failure, and failure leads to understanding.
>
> — Burt Rutan

Step 1: Validate Real Interactions

NONSENSE fodsihasojdpao

SLANG I'd like to *nip over* to Dublin from Belfast

INSULTS Work you stupid system!!!

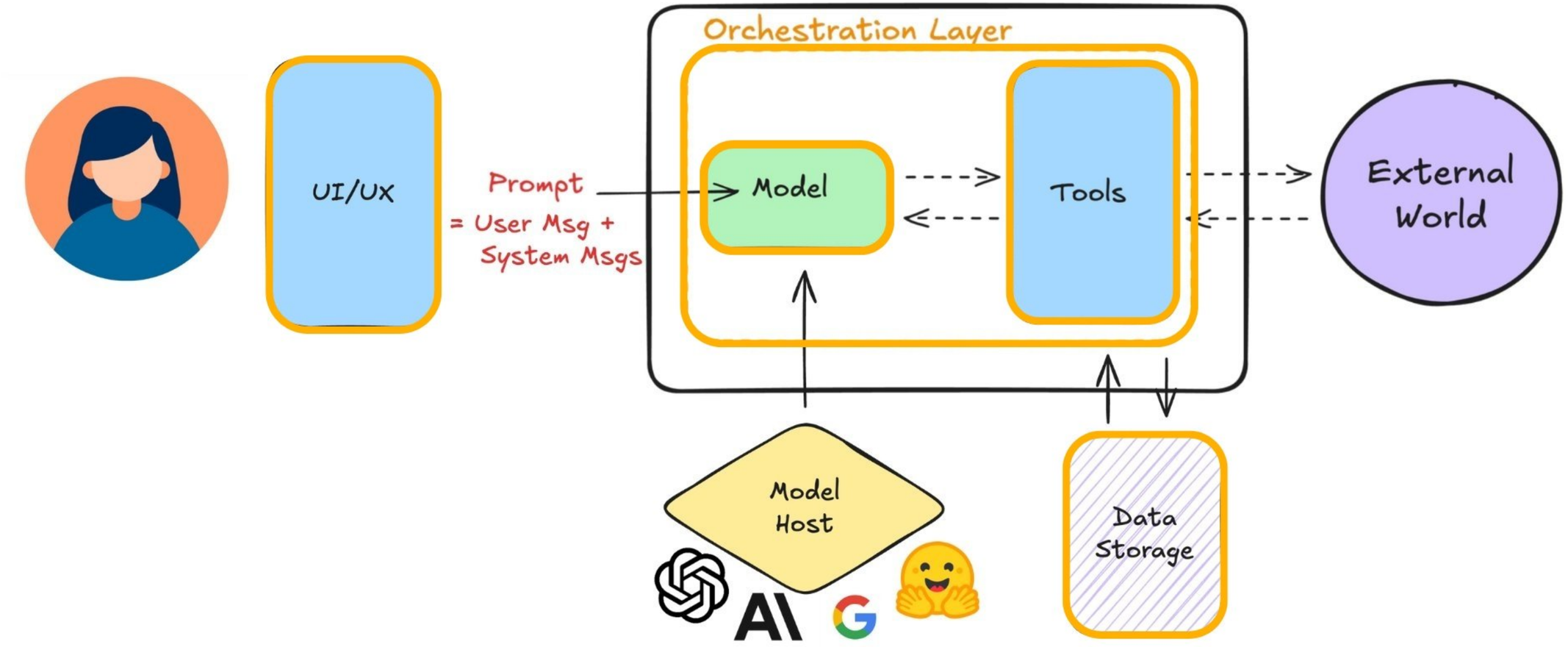


Simulated User: fdubnfdiubfd

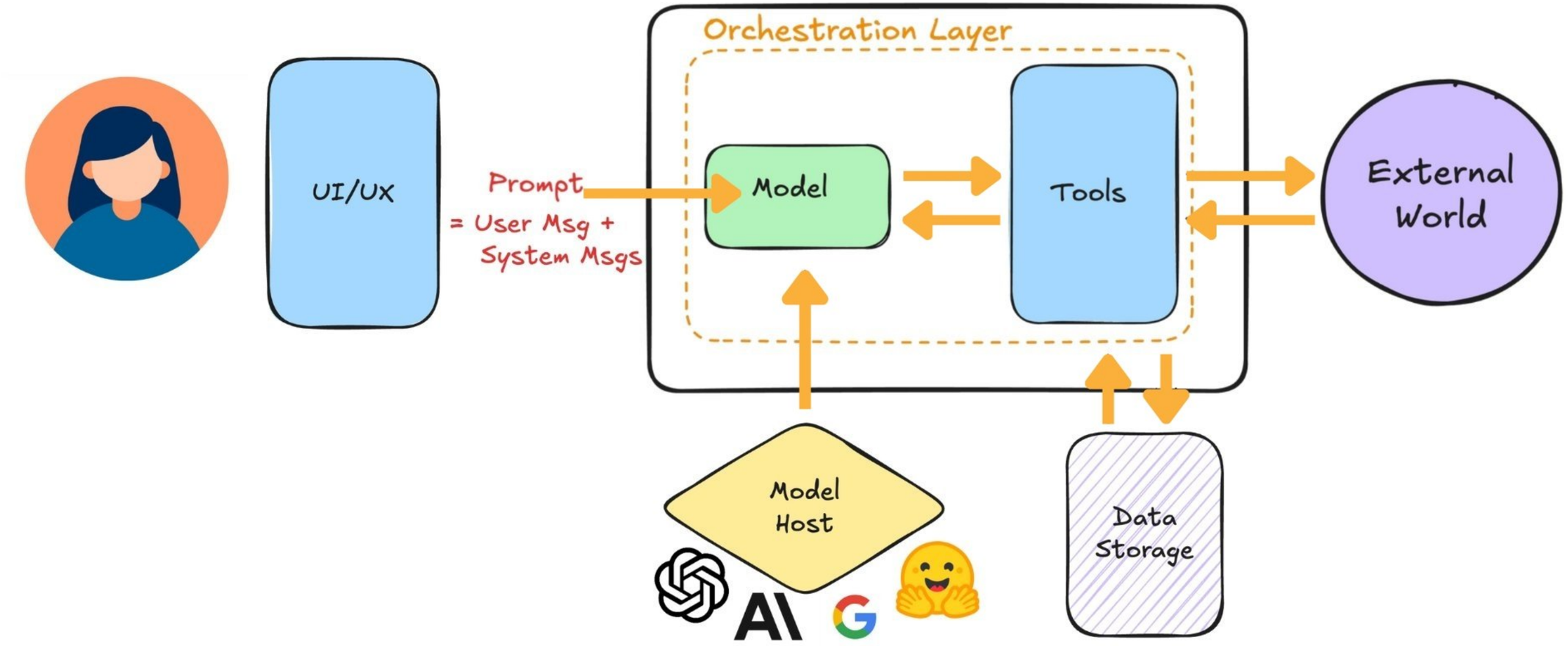
Simulated User: I'd like to head over to Paris

[1] nip over = British slang for a quick visit

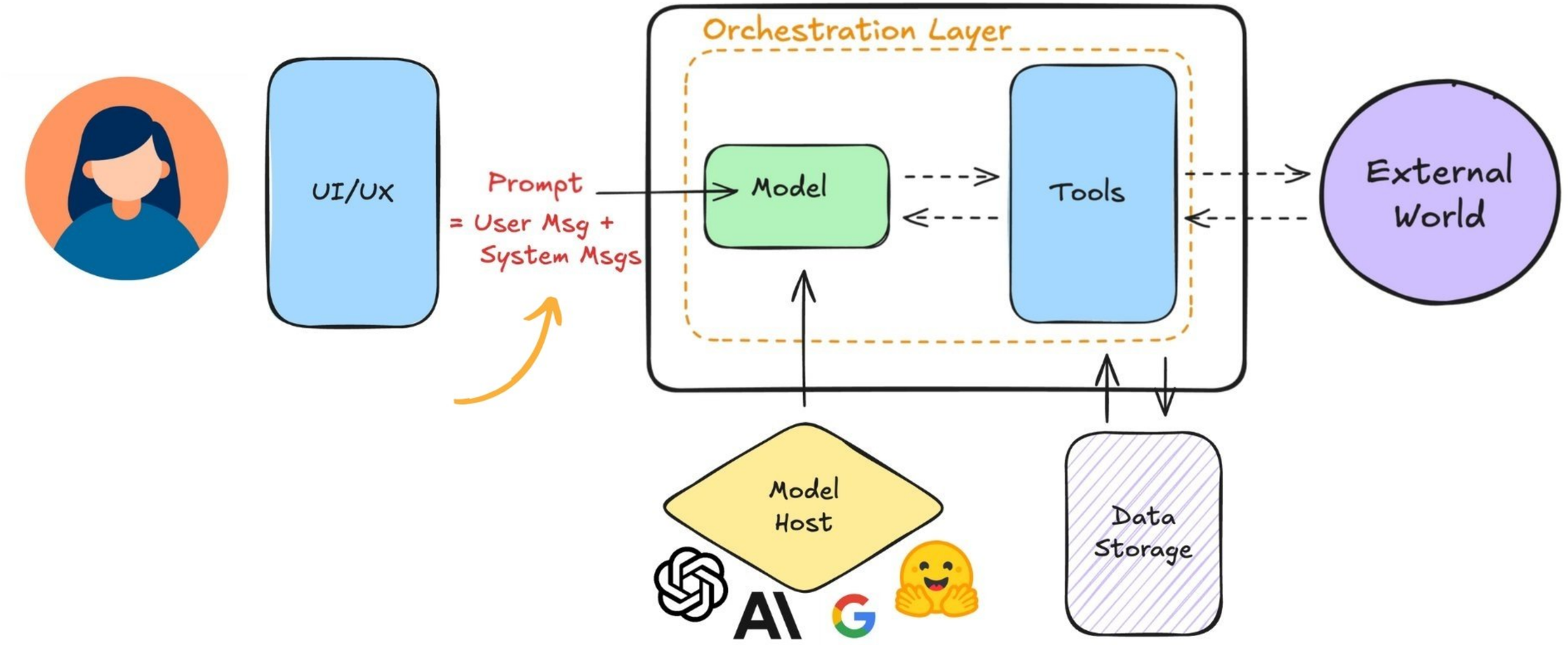
Step 2: Test Everything: Unit tests



Step 2: Test Everything: Integration tests



Step 2: Test Everything: Evaluate system prompts



Step 2: Test Everything: Subjective evaluations

Step 2: Test Everything: Subjective evaluations

HUMAN RATER

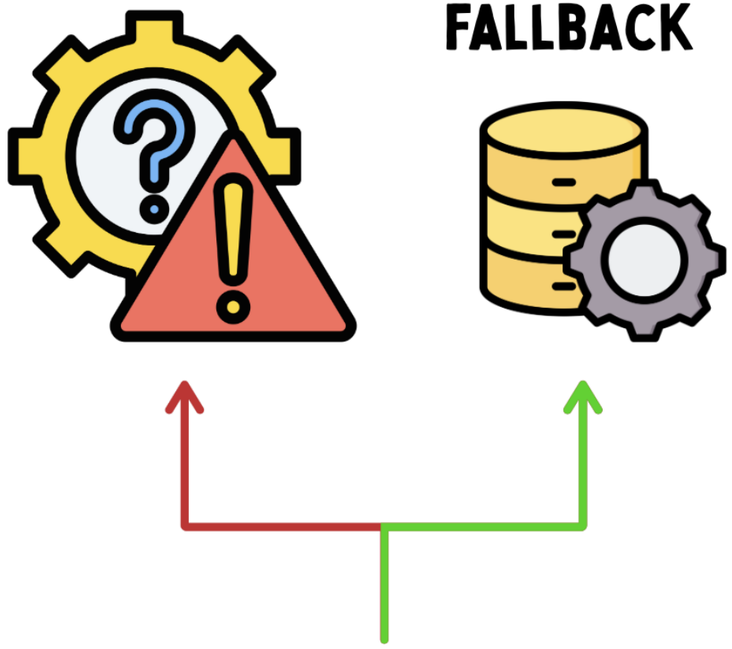
LLM-AS-A-JUDGE

Step 3: Guardrails and Observability

CONTENT FILTERS

MODEL OR TOOL FAILURE

FALLBACK

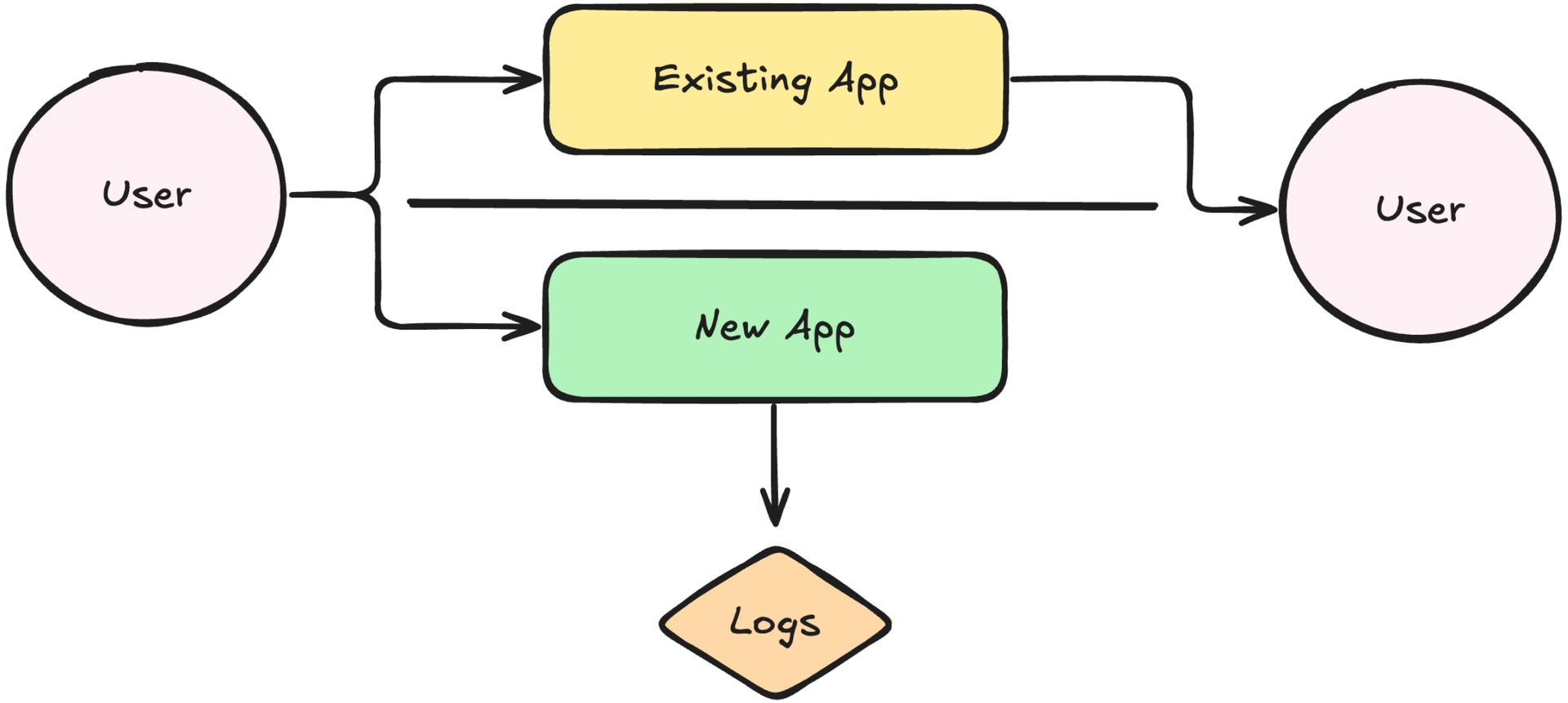


Step 3: Guardrails and Observability

Collect diagnostic data

Collect message data

Step 4: Shadow Deployments



Step 5: Deployment Strategies

A/B Test

Step 5: Deployment Strategies

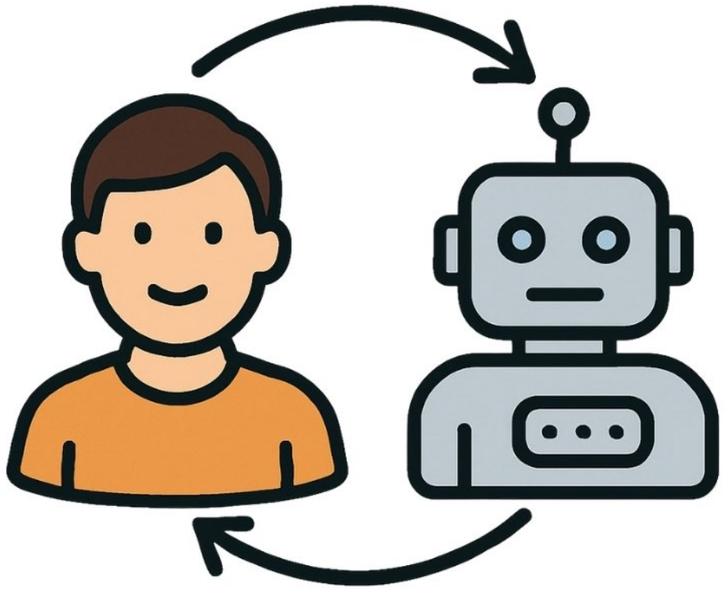
A/B Test

User Segmentation

Step 5: Deployment Strategies

A/B Test

User Segmentation



Human-in-the-loop

Final Checklist

Let's practice!

BUILDING SCALABLE AGENTIC SYSTEMS

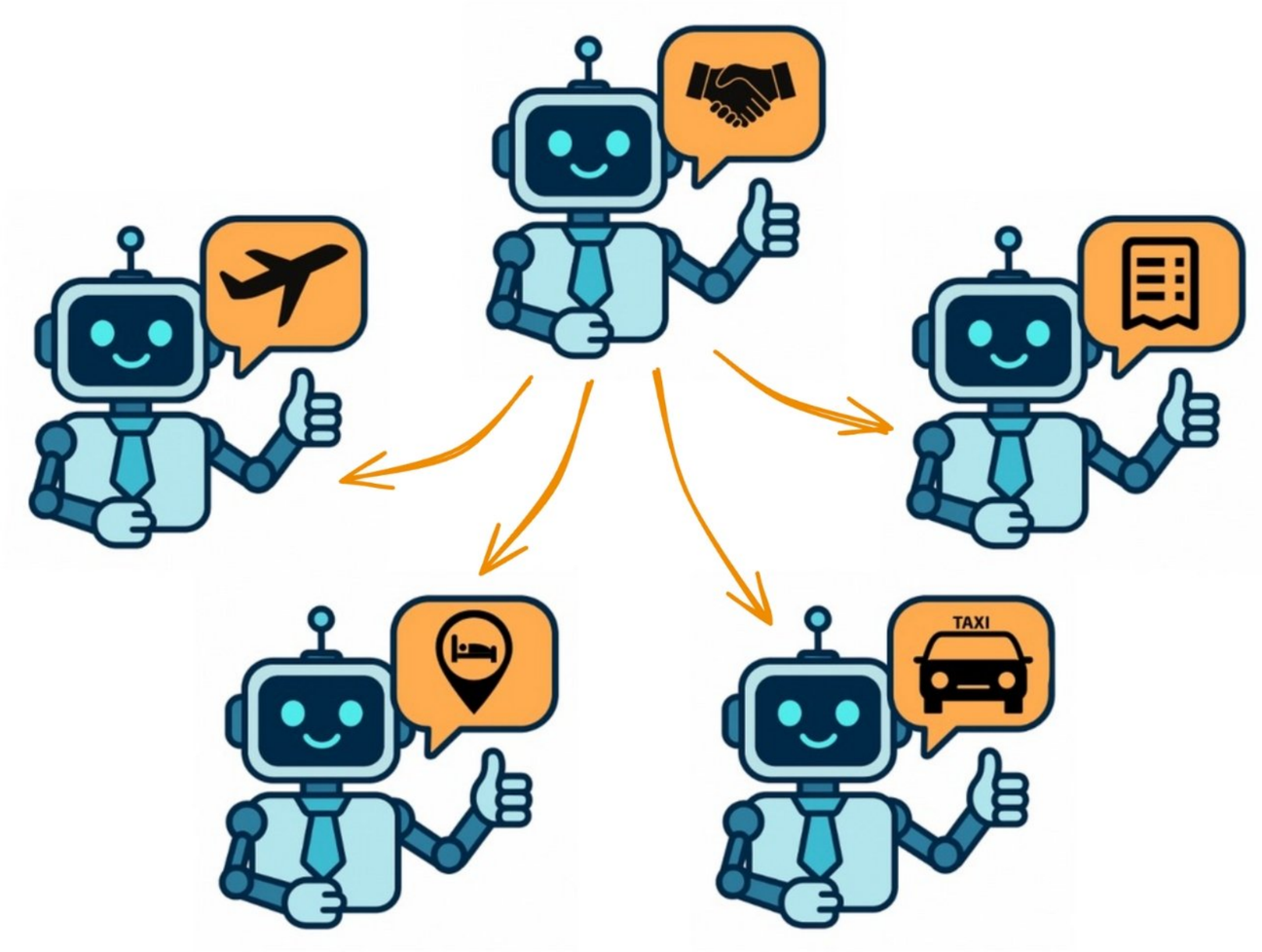
Fast and Efficient Data Ingestion

BUILDING SCALABLE AGENTIC SYSTEMS

Korey Stegared-Pace
Senior AI Cloud Advocate, Microsoft

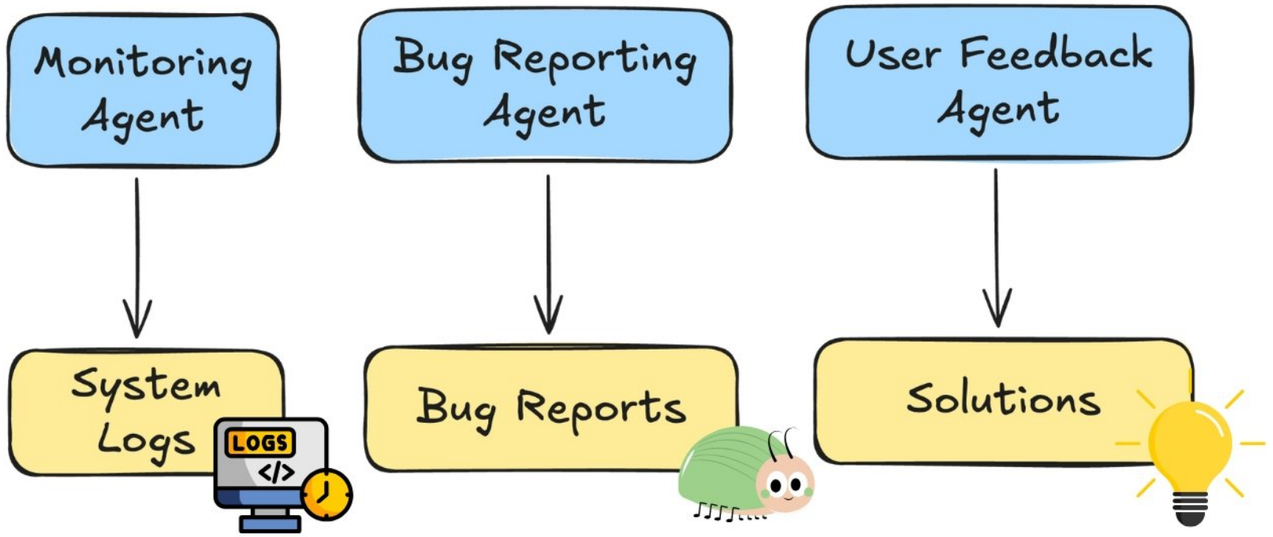
datacamp

Agents without up-to-date information

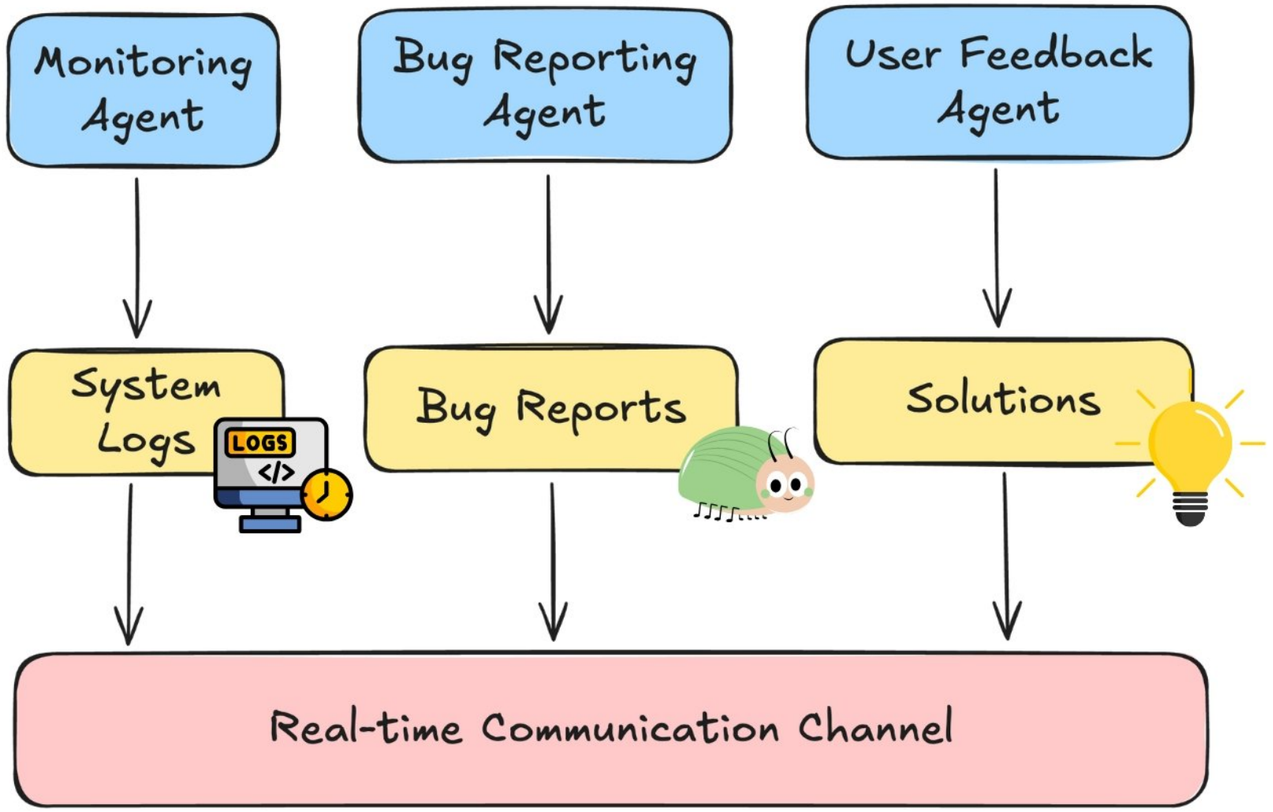


Real-time data for multi-agent coordination

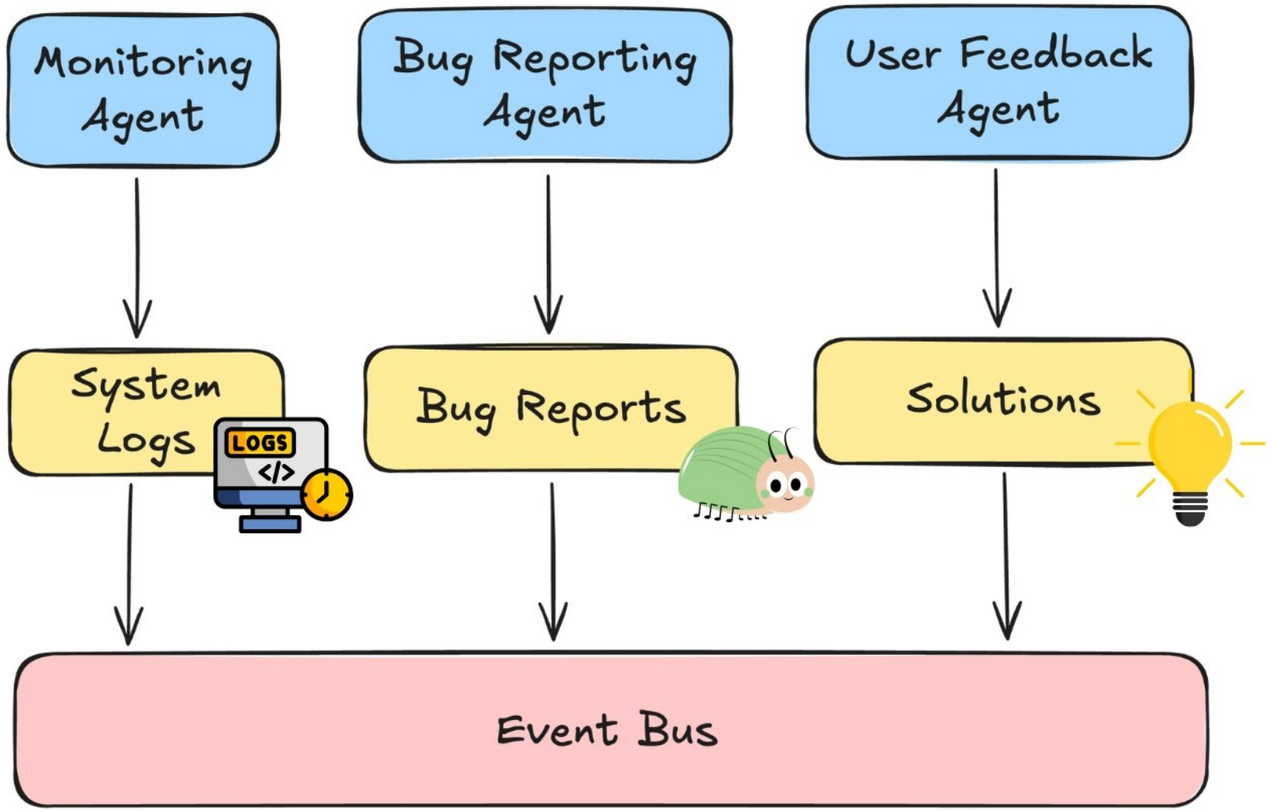
Real-time data for multi-agent coordination



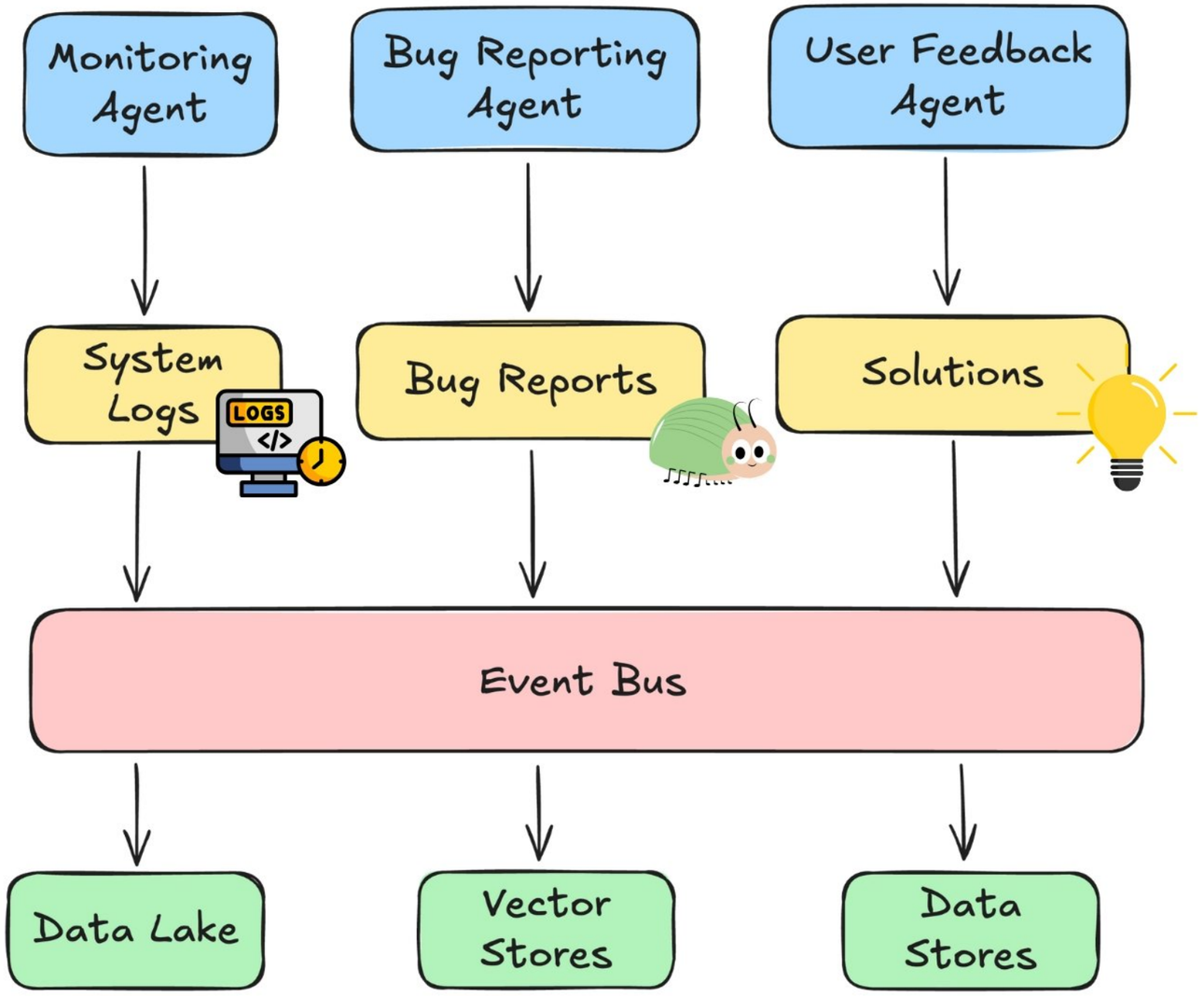
Real-time communication channel



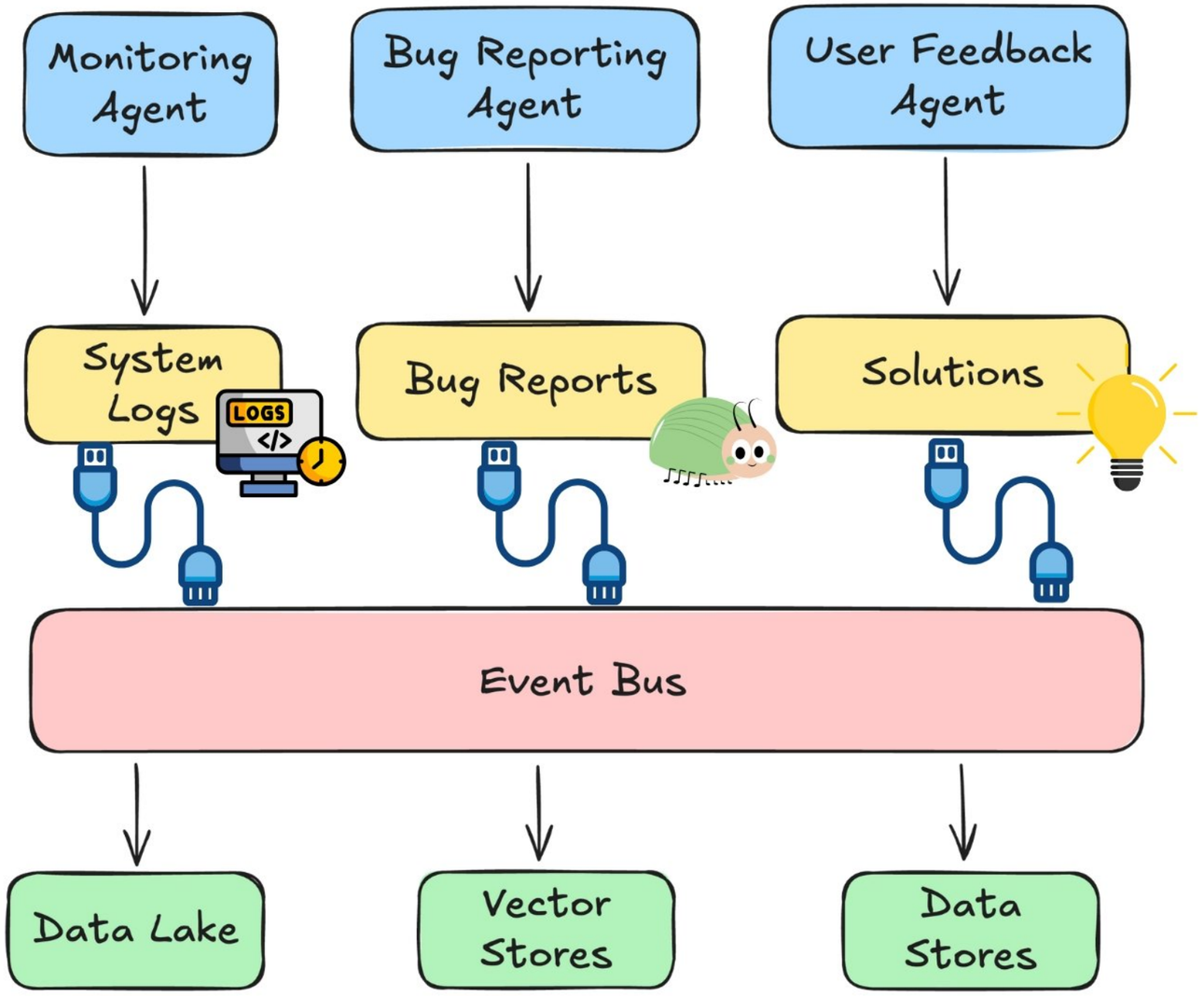
Real-time communication channel



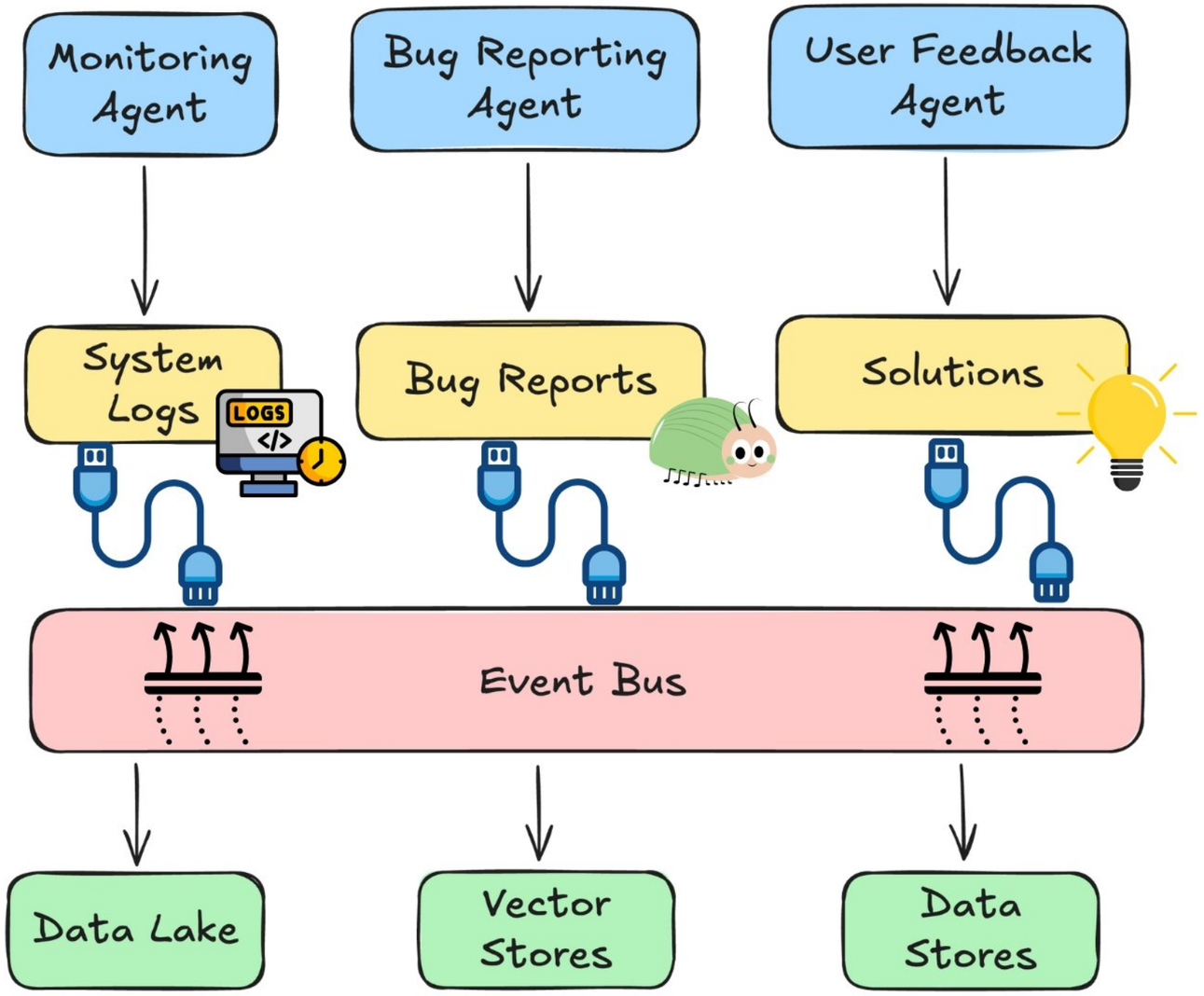
Utilizing an event bus



Utilizing an event bus



Utilizing an event bus



Let's practice!

BUILDING SCALABLE AGENTIC SYSTEMS

Failing Gracefully: Mitigating Risks

BUILDING SCALABLE AGENTIC SYSTEMS

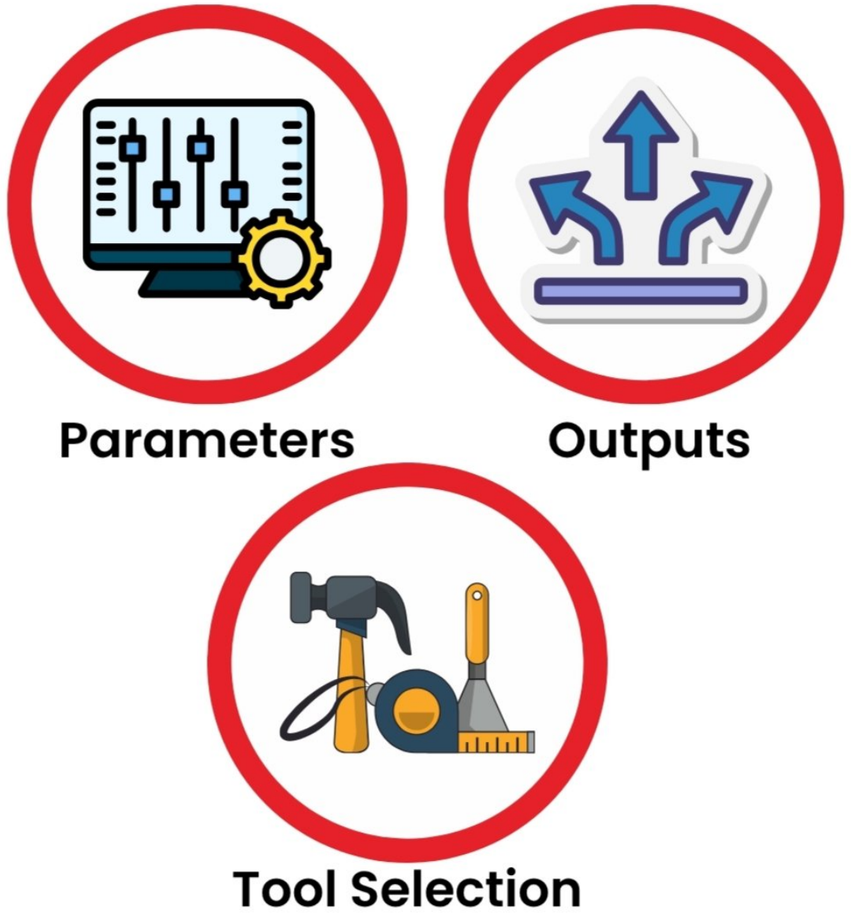
Korey Stegared-Pace
Senior AI Cloud Advocate, Microsoft

Tool call failures

Solutions

- Defining clear parameters and usage
- Tool output validation (unit testing!)
- Verification checks on tool selection
- MCP!

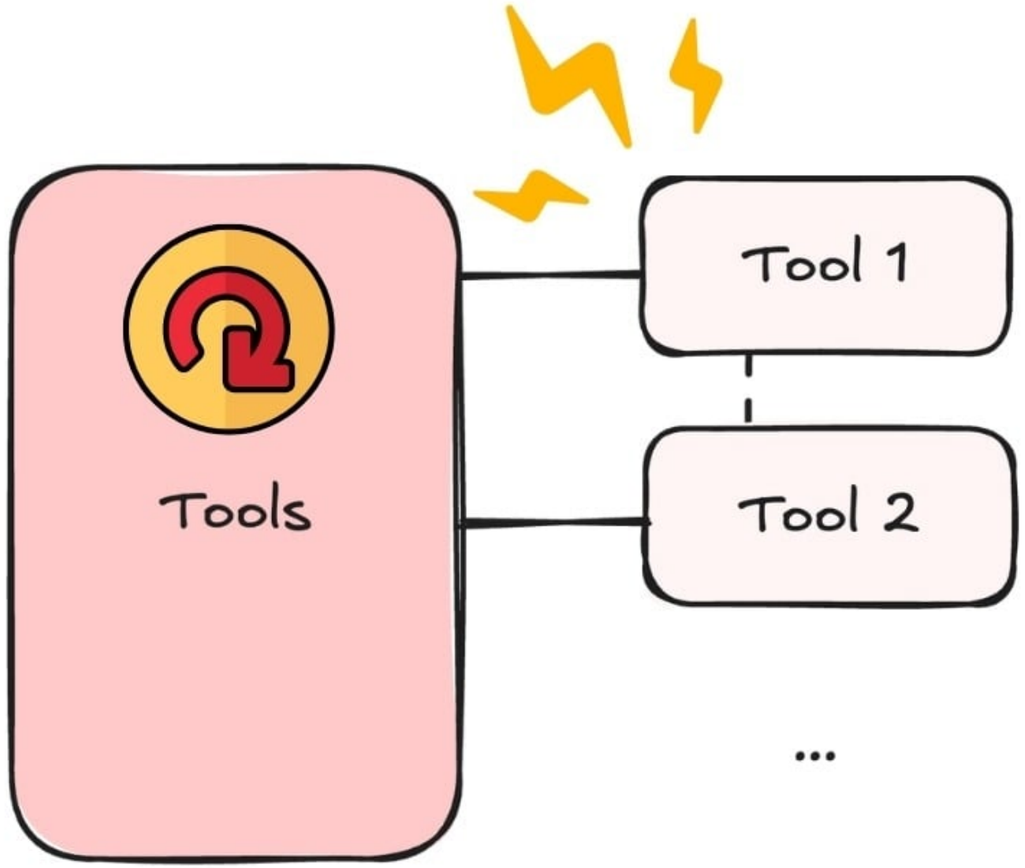
Problems



Tool call failures - retry mechanism

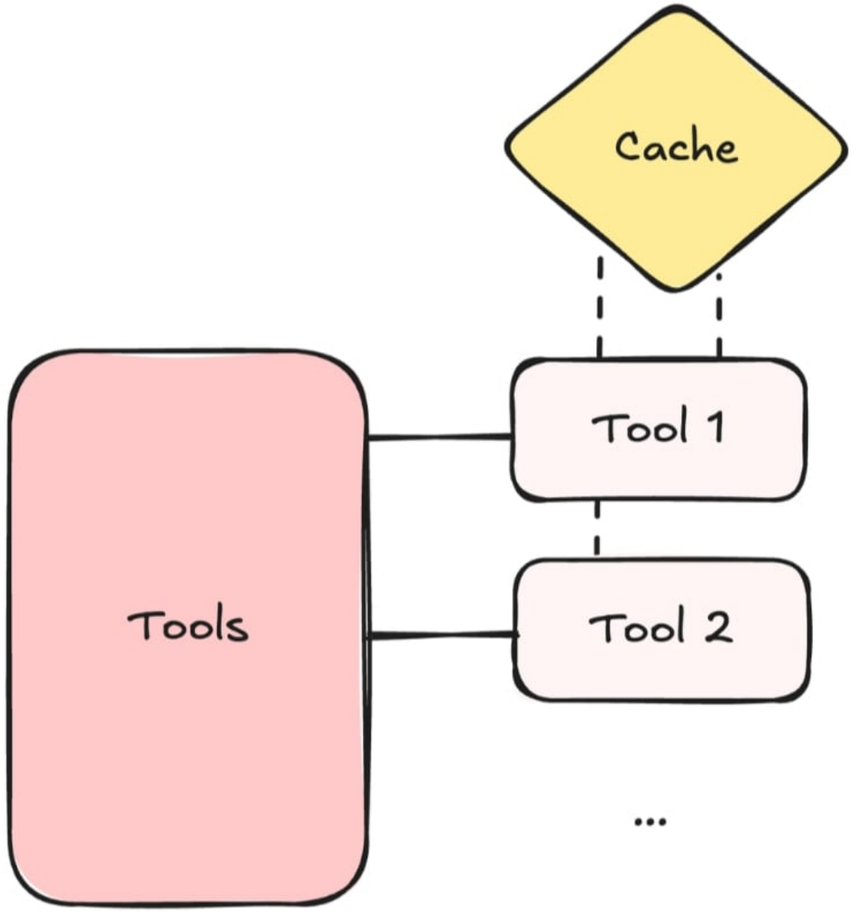
- External services may *break* or be temporarily *slow*

- **Retry**
  - **Exponential backoff** → *slowly decreasing* retries
  - Inform users of issues using *callbacks*



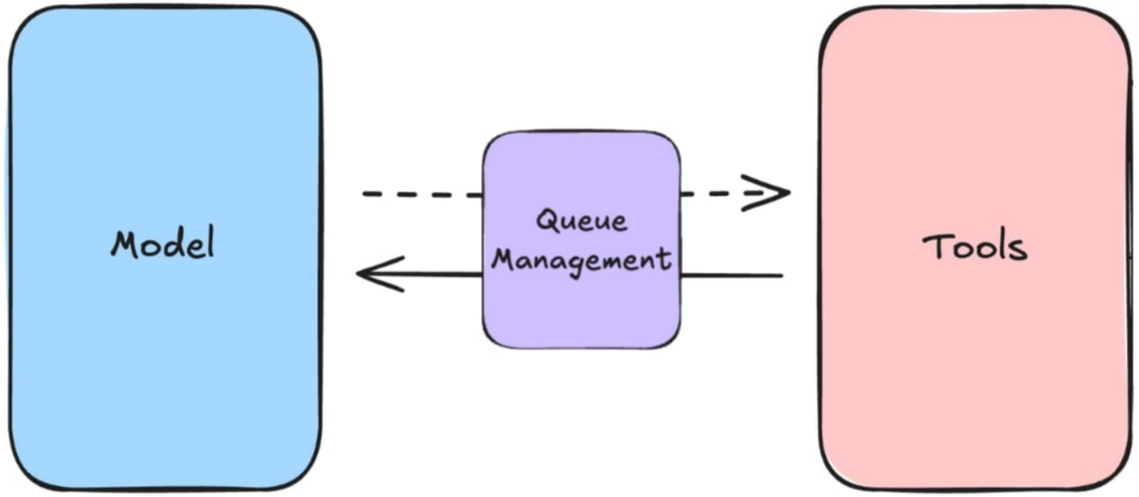
Tool call failures - caching

- Tool outputs are cached as a *fallback*
- **Works for:** Static data with infrequent updates
- **Won't work for:** tools that serve real-time information

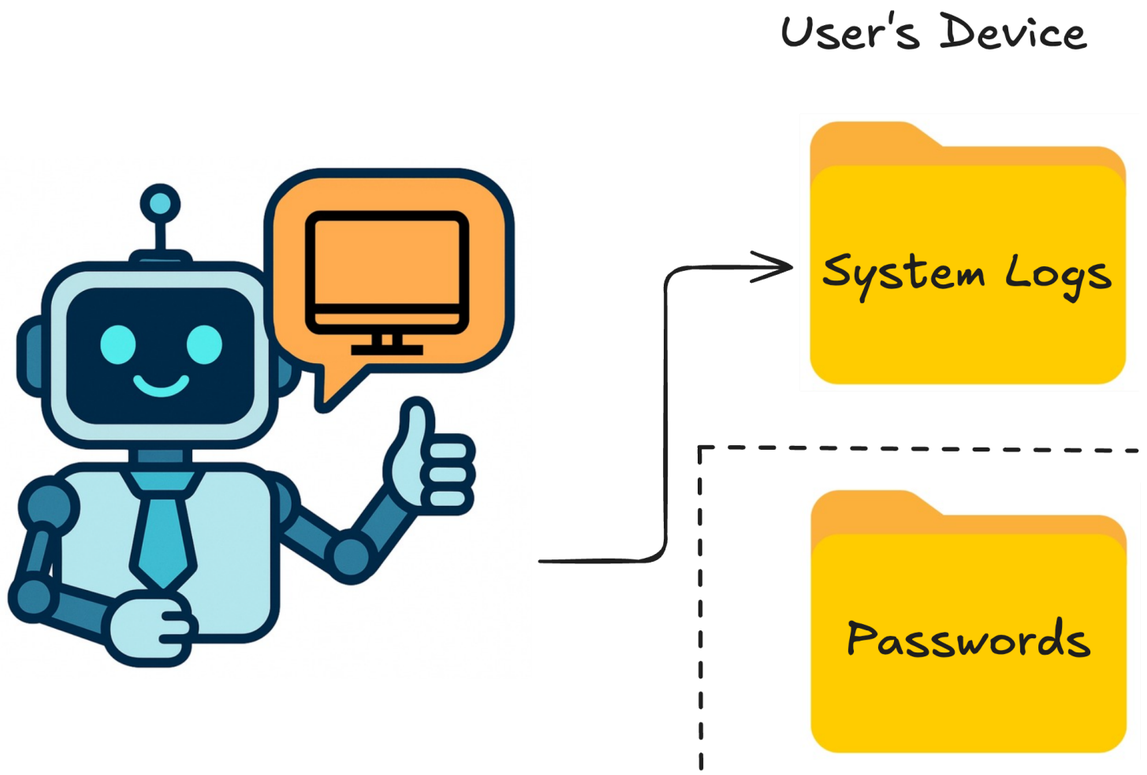


Tool call failures - queue management

- Tool calls may be *reliant* on each another
- **Example:** flights must be booked before the taxi
- Unsuccessful tool calls can be moved down the queue



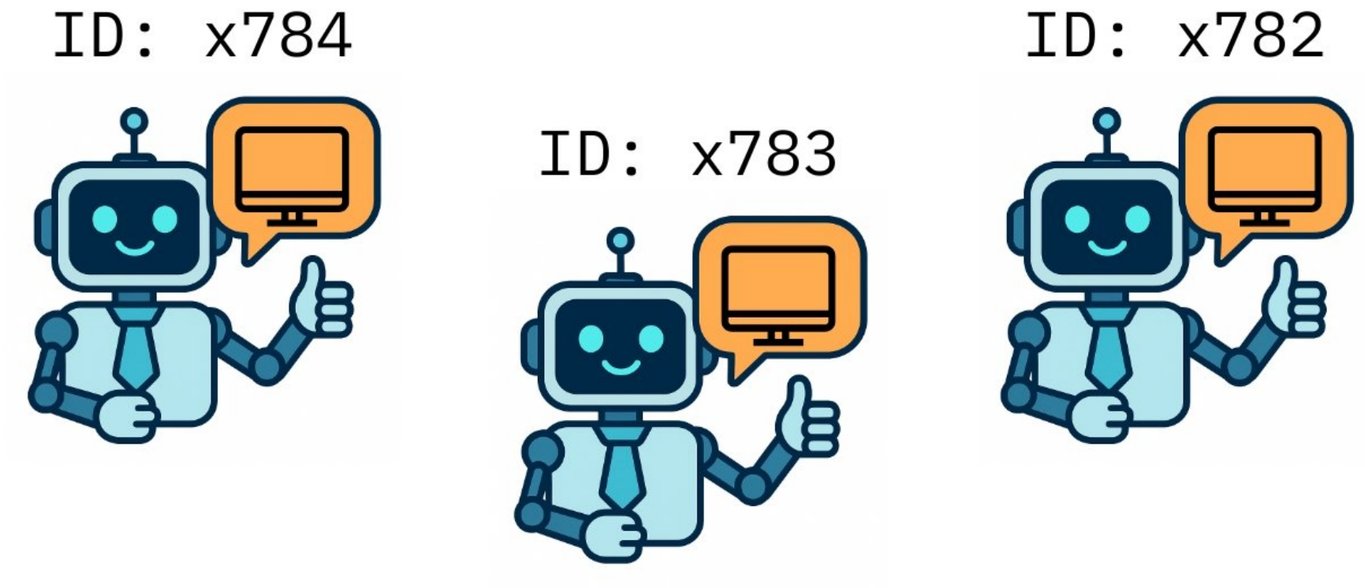
Authentication



- Tools may require *access* and *permissions* to **private** data

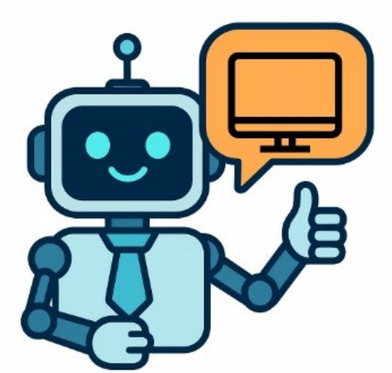
- **Example:** IT support agent
  - Valid access: system logs
  - Blocked access: passwords, location, etc.

Authentication - unique agent identifiers



Authentication - isolated environments

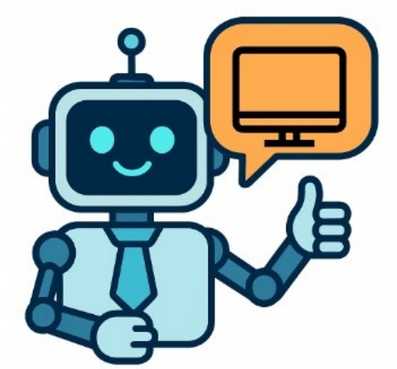
ID: x784



Access Rights:

...

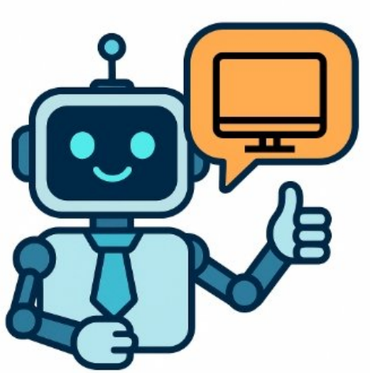
ID: x783



Access Rights:

...

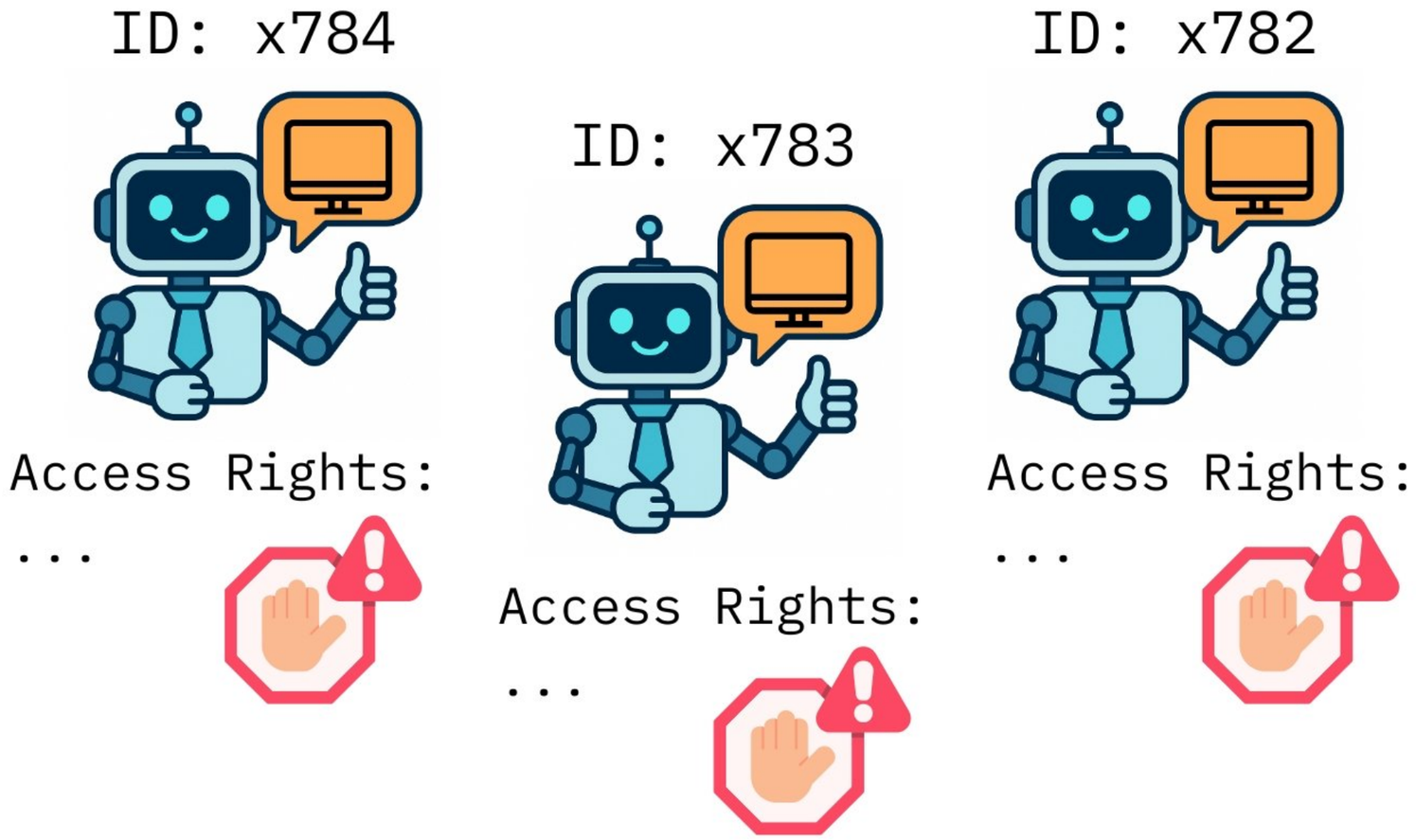
ID: x782



Access Rights:

...

Authentication - guardrails and action restraints



Let's practice!

BUILDING SCALABLE AGENTIC SYSTEMS

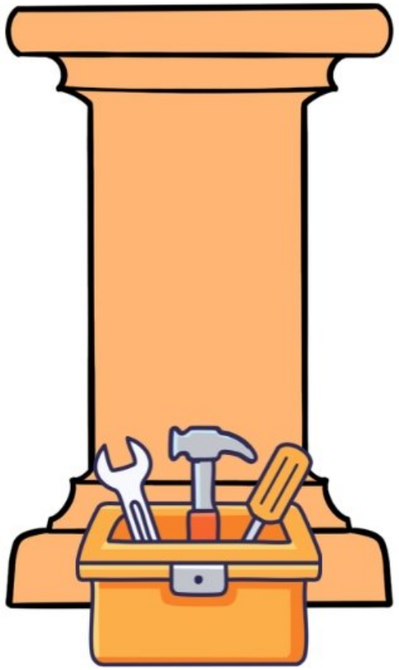
Congratulations!

BUILDING SCALABLE AGENTIC SYSTEMS

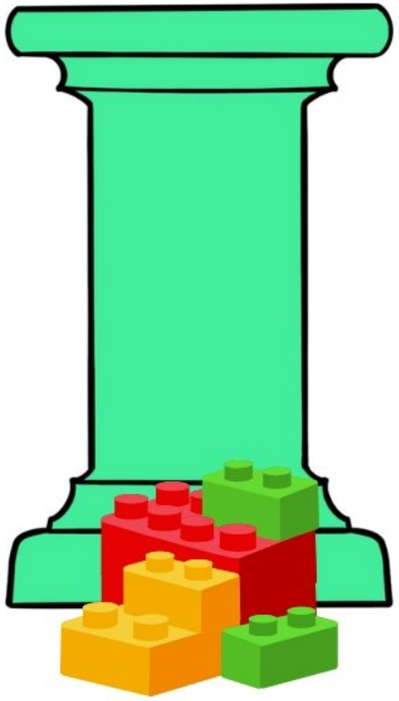
Korey Stegared-Pace
Senior AI Cloud Advocate, Microsoft

Chapter 1

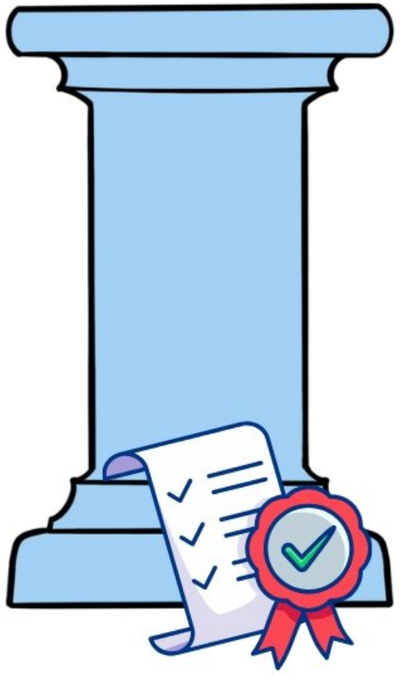
Robust Infrastructure and Tooling



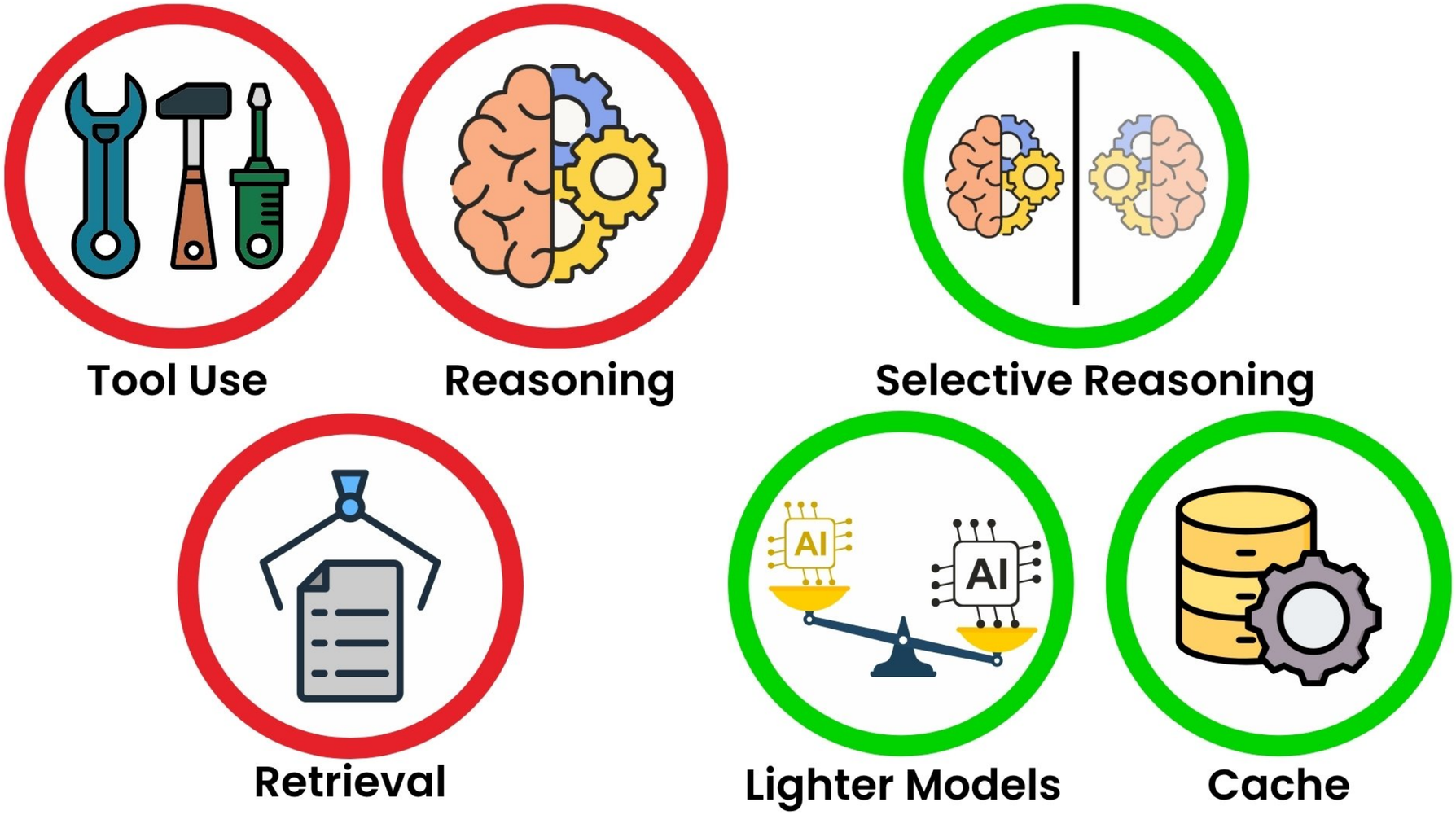
Modular Design Architecture



Continuous Evaluation & Feedback Loops

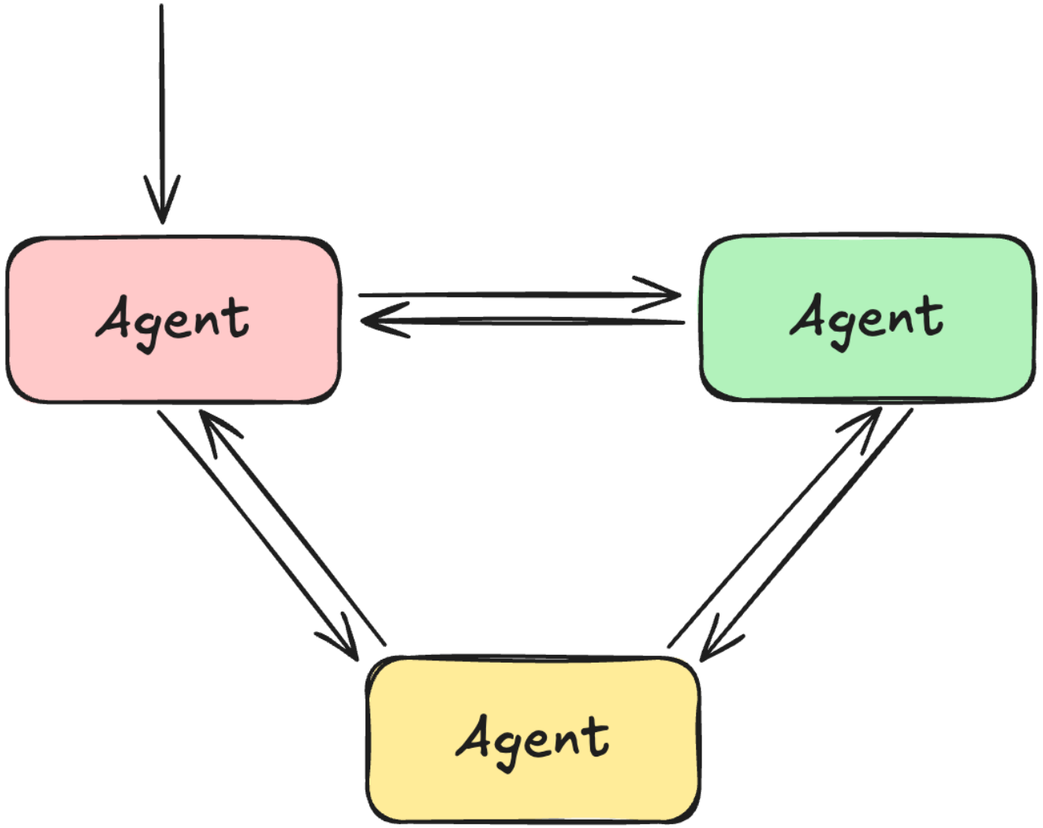


Chapter 1

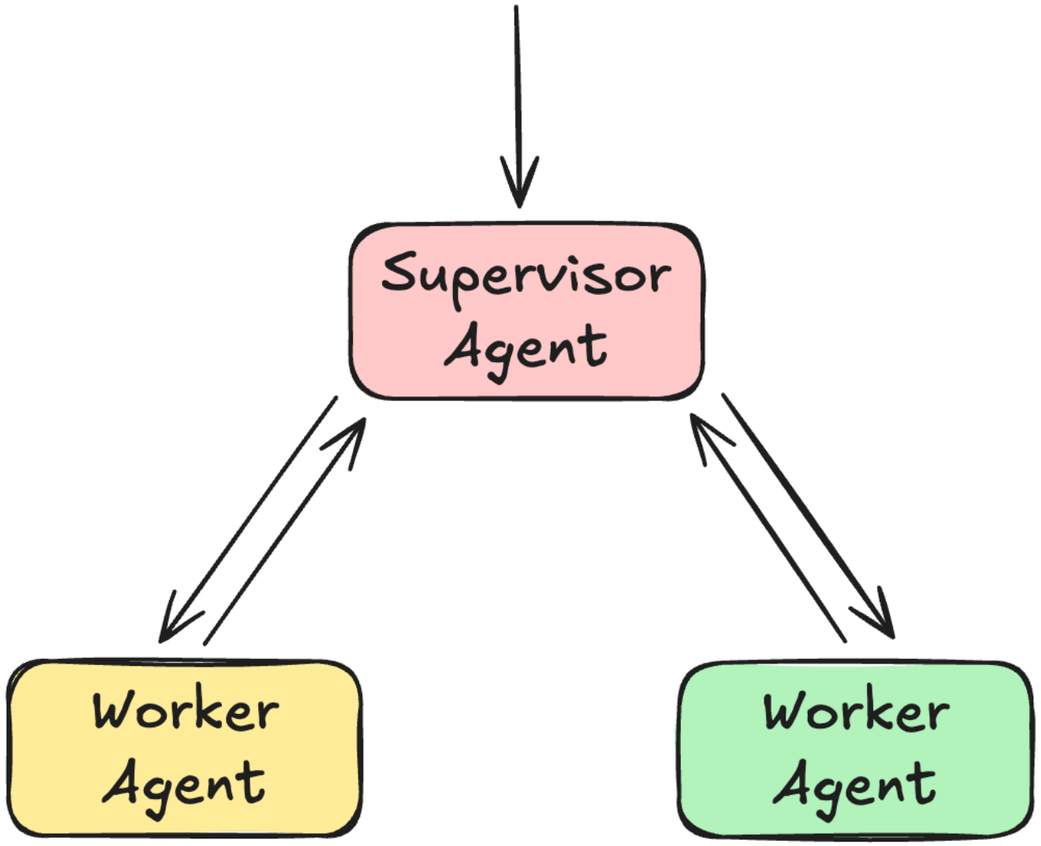

Tool Use
Reasoning
Selective Reasoning
Retrieval
AI
AI
Lighter Models
Cache

Chapter 2

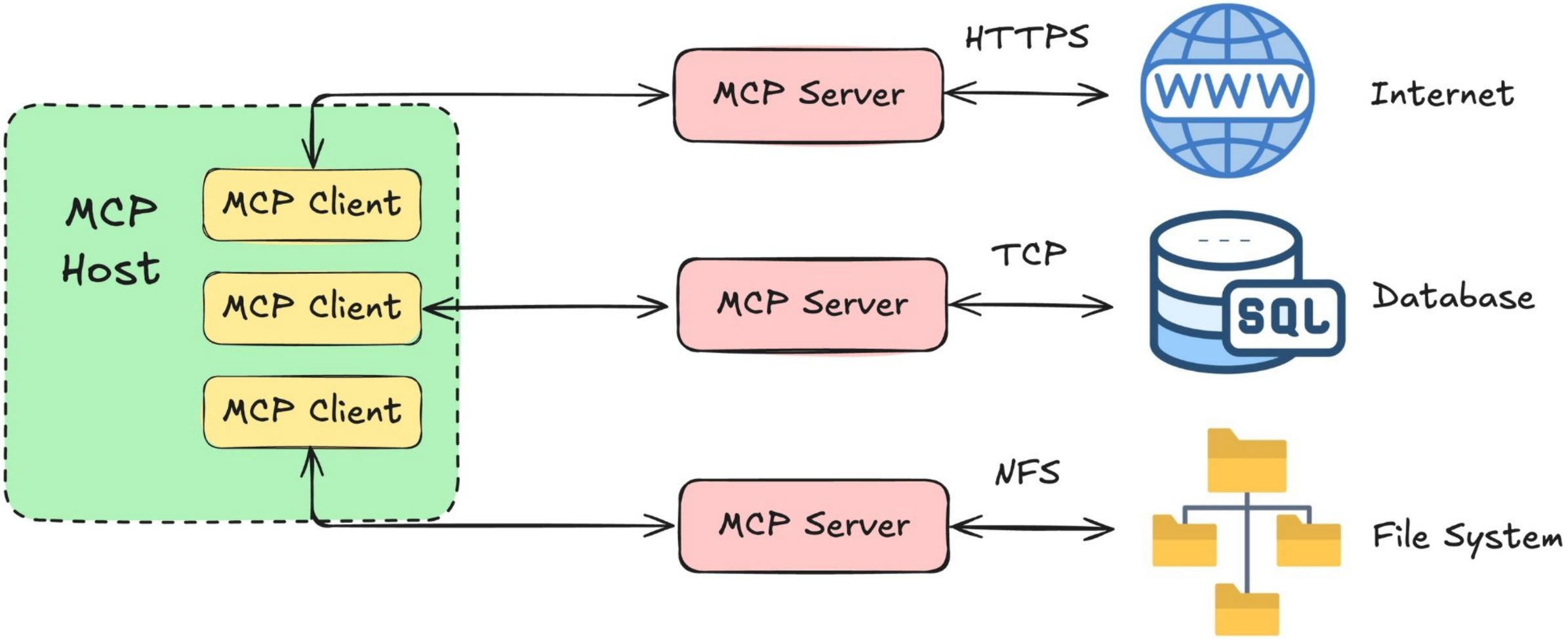
Network



Supervisor



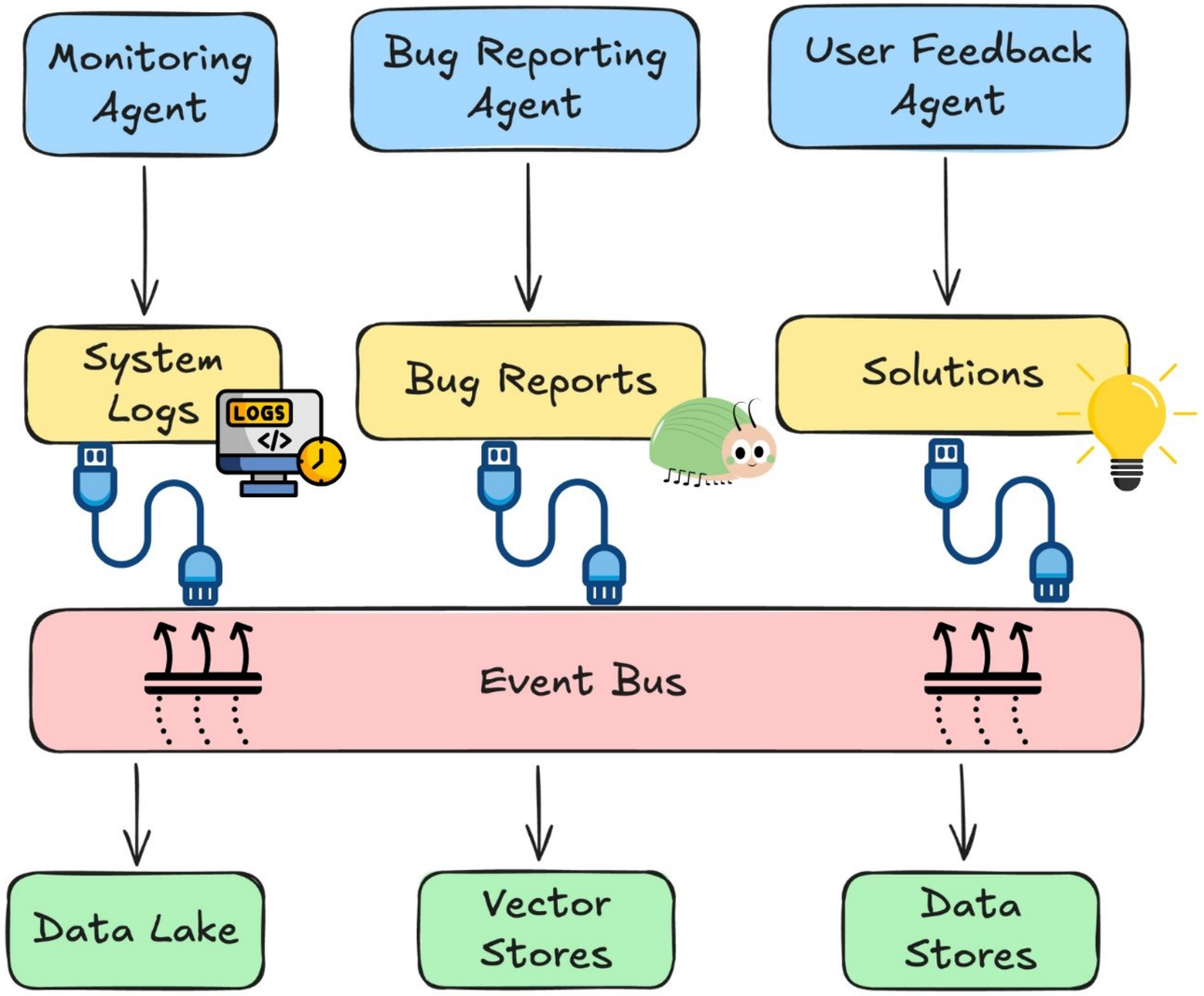
Chapter 2


MCP Host
MCP Client
MCP Client
MCP Client
MCP Server
MCP Server
MCP Server
HTTPS
TCP
NFS
WWW
SQL
Internet
Database
File System

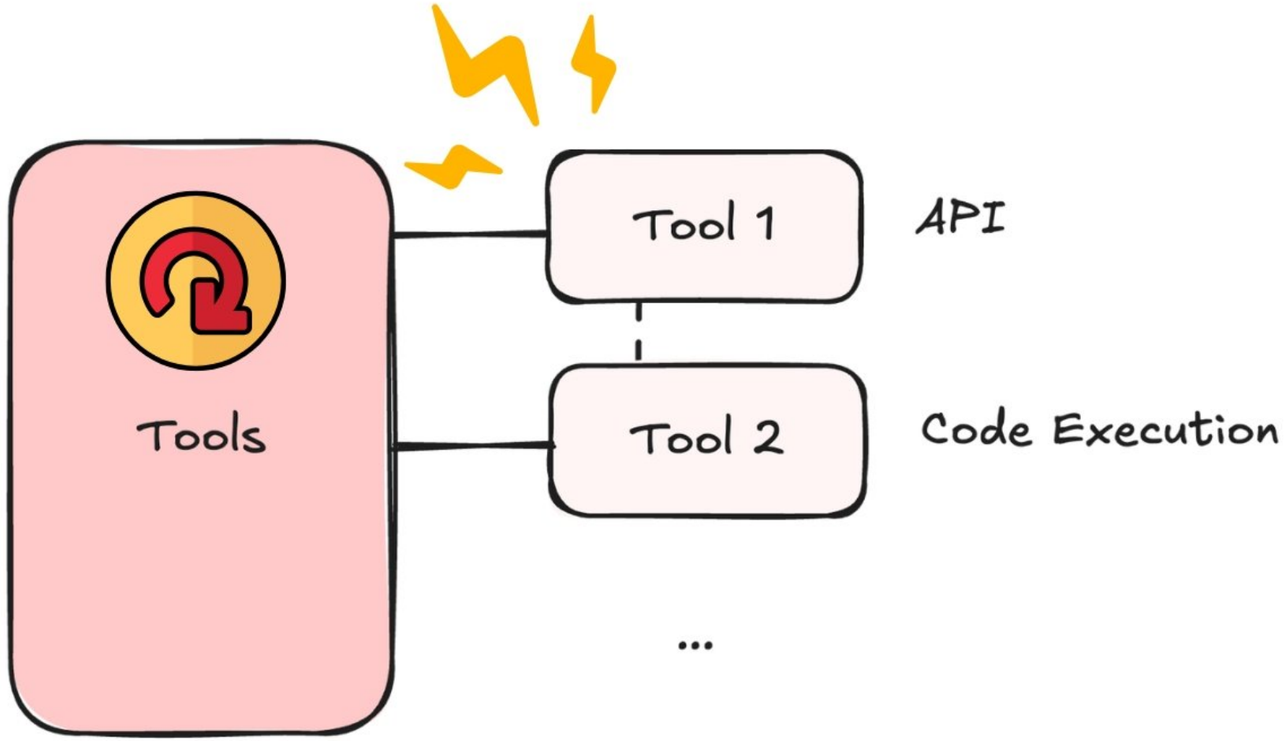
Chapter 3


CHECKLIST
Handle Messy Inputs
Logs
Filters
Testing
Gradual Deployment
Observability

Chapter 3


Monitoring Agent
Bug Reporting Agent
User Feedback Agent
System Logs
LOGS
Bug Reports
Solutions
Event Bus
Data Lake
Vector Stores
Data Stores

Chapter 3



Where next?

- Blog: **Understanding AI Agents: The Future of Autonomous Systems**
- DataFramed Podcast: **AI Agents Hype vs. Reality**
- **AI Agents Cheat Sheet**

Let's practice!

BUILDING SCALABLE AGENTIC SYSTEMS