AI Agents in the Wild

BUILDING SCALABLE AGENTIC SYSTEMS

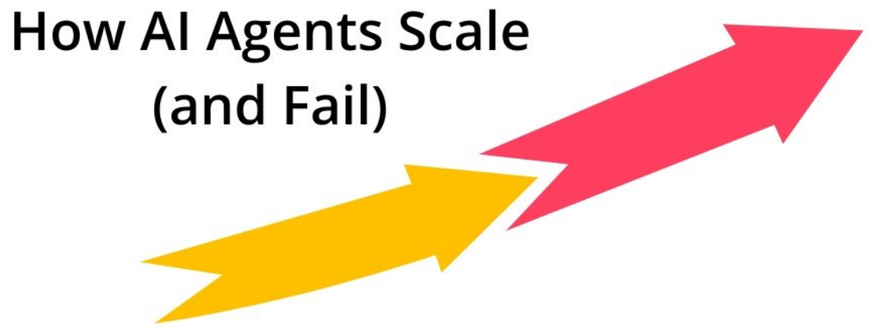
Korey Stegared-Pace
Senior AI Cloud Advocate, Microsoft

datacamp

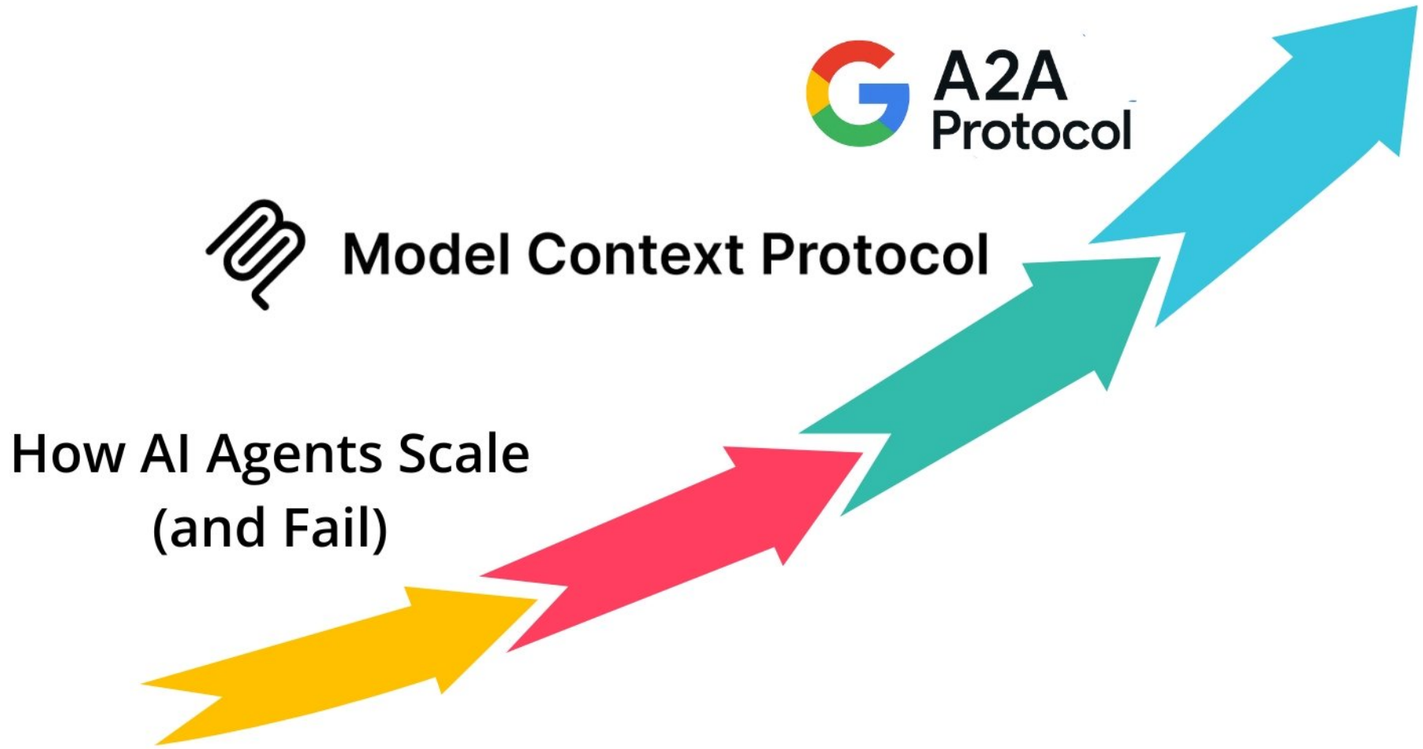
Meet your instructor

- Senior AI Cloud Advocate, Microsoft

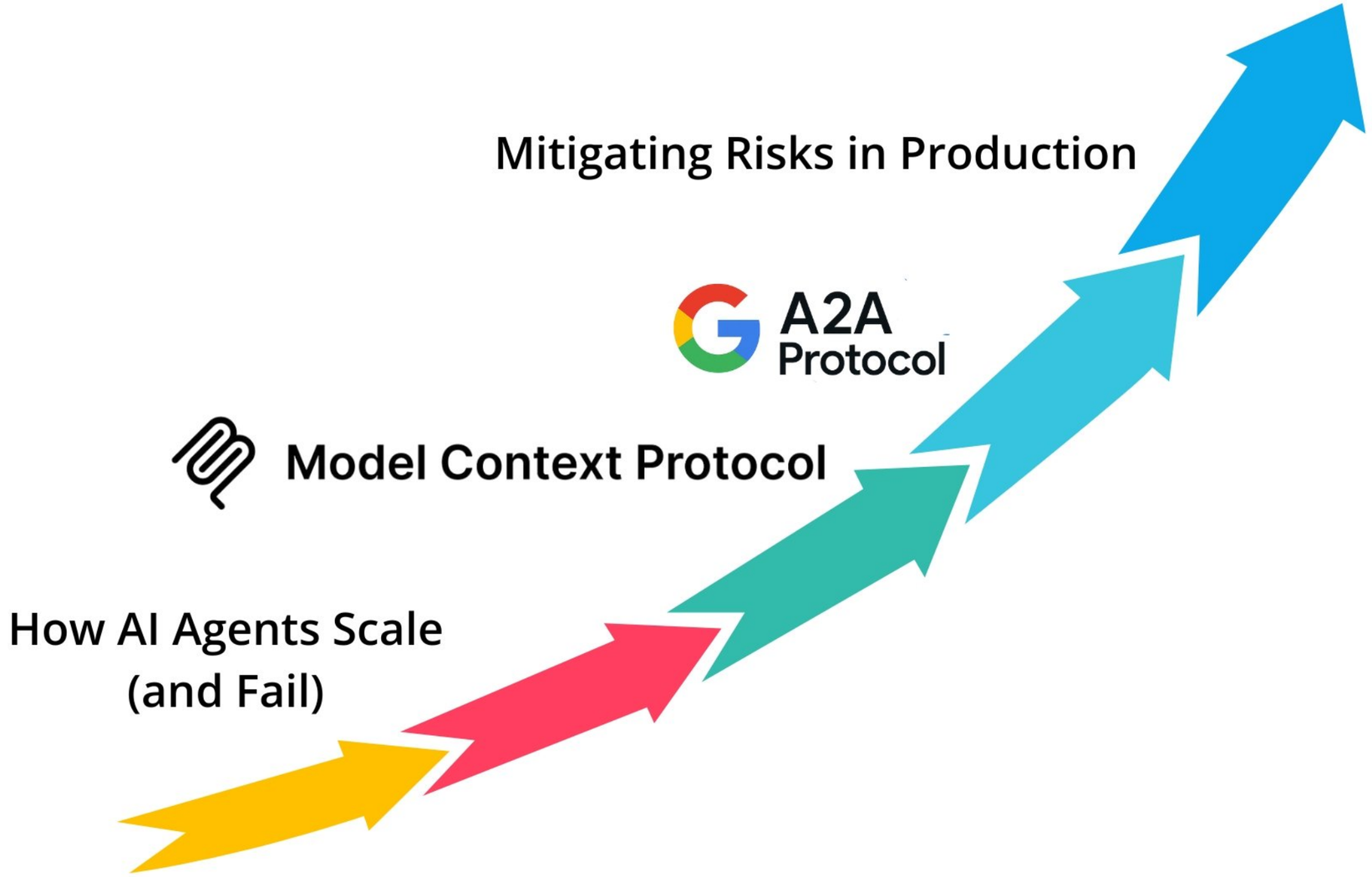
The journey ahead!



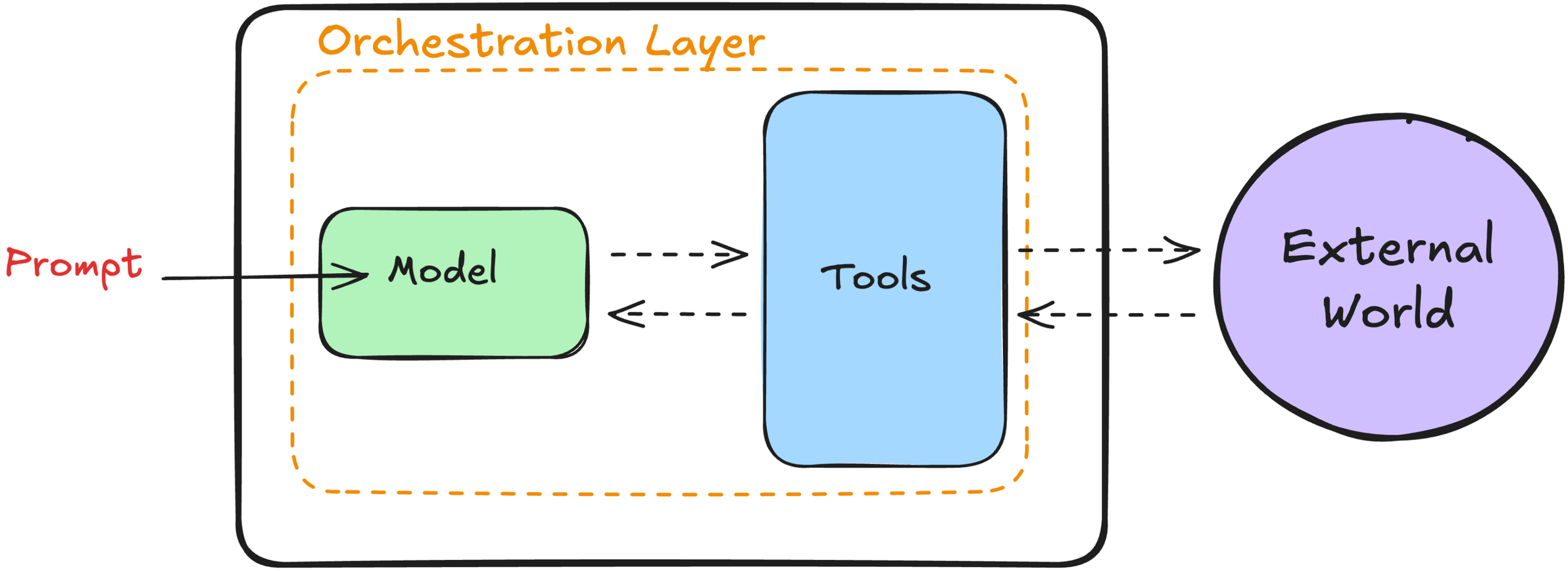
The journey ahead!



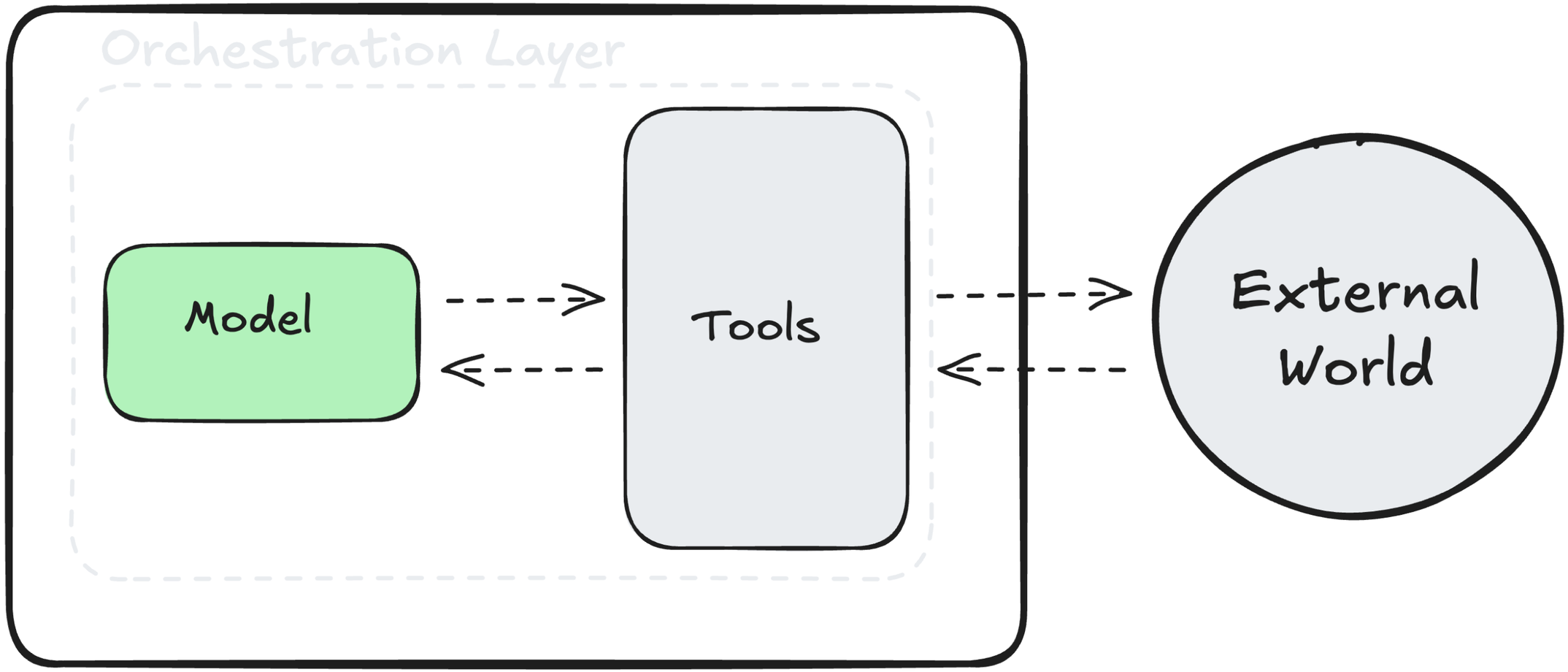
The journey ahead!



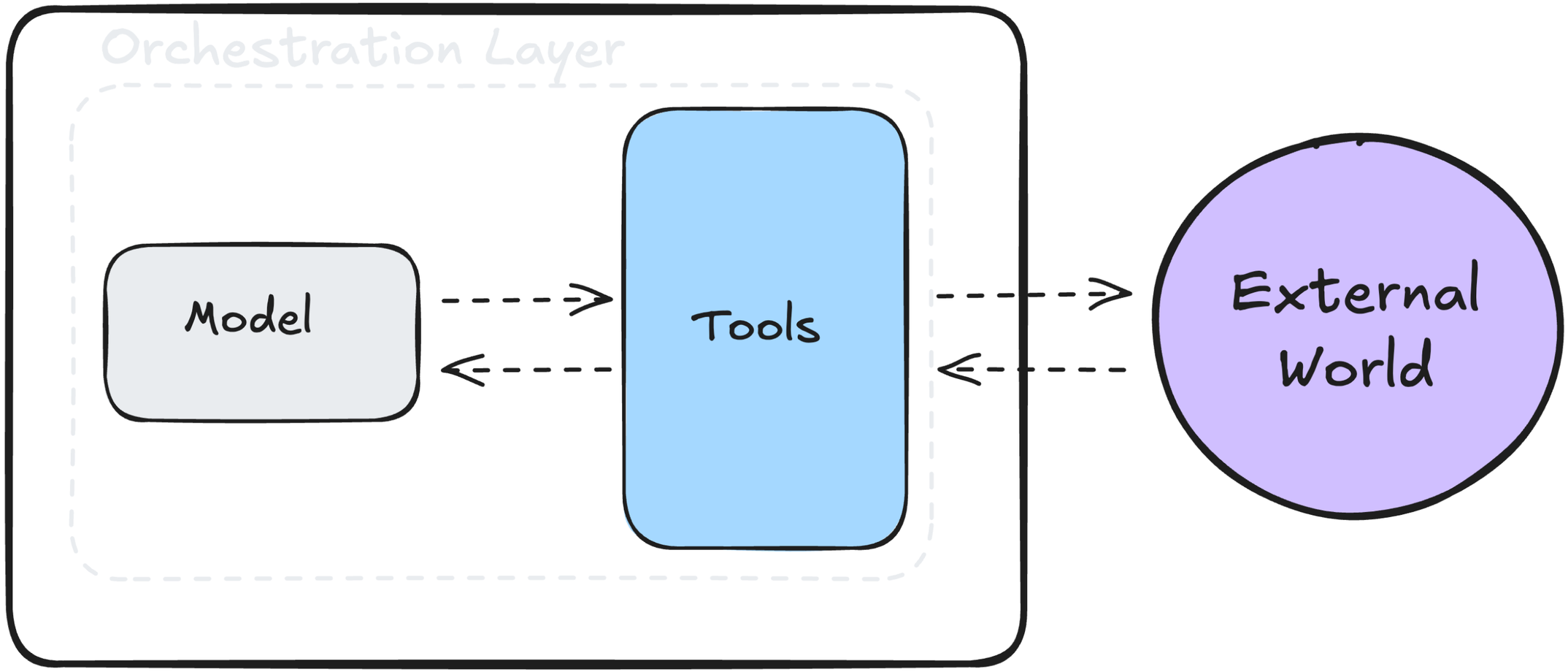
Components of an agent



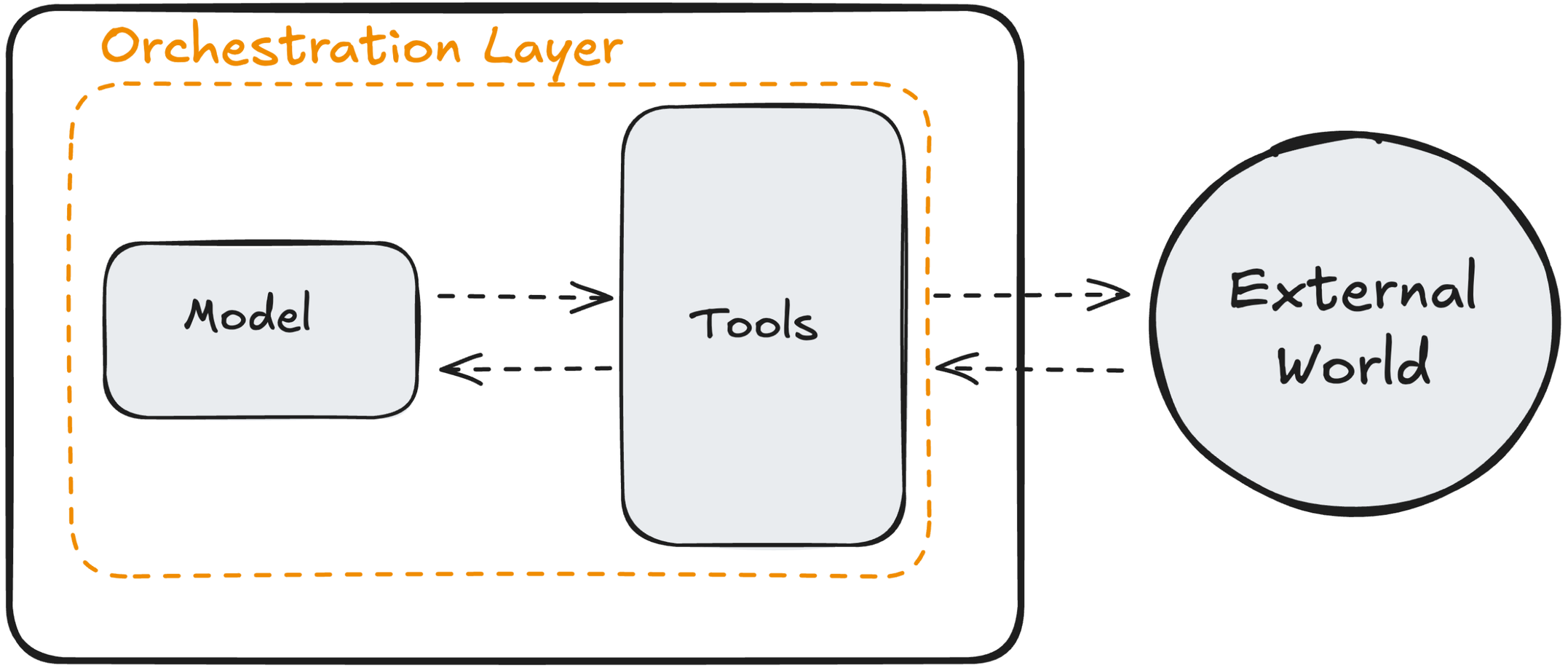
Components of an agent



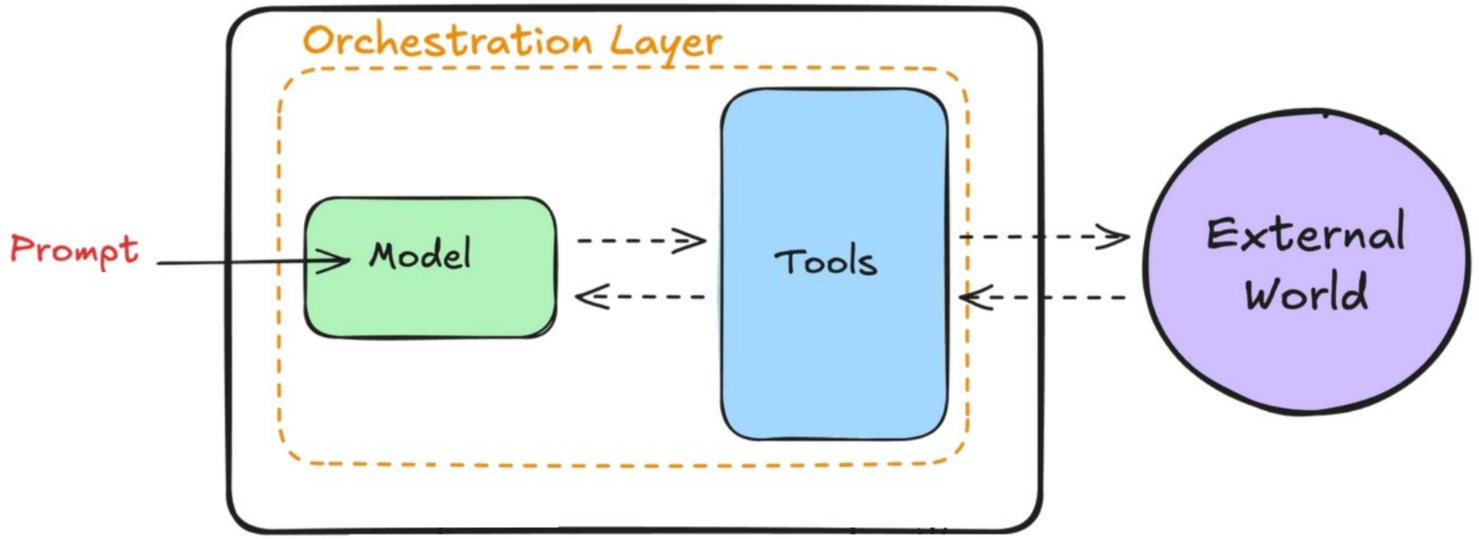
Components of an agent



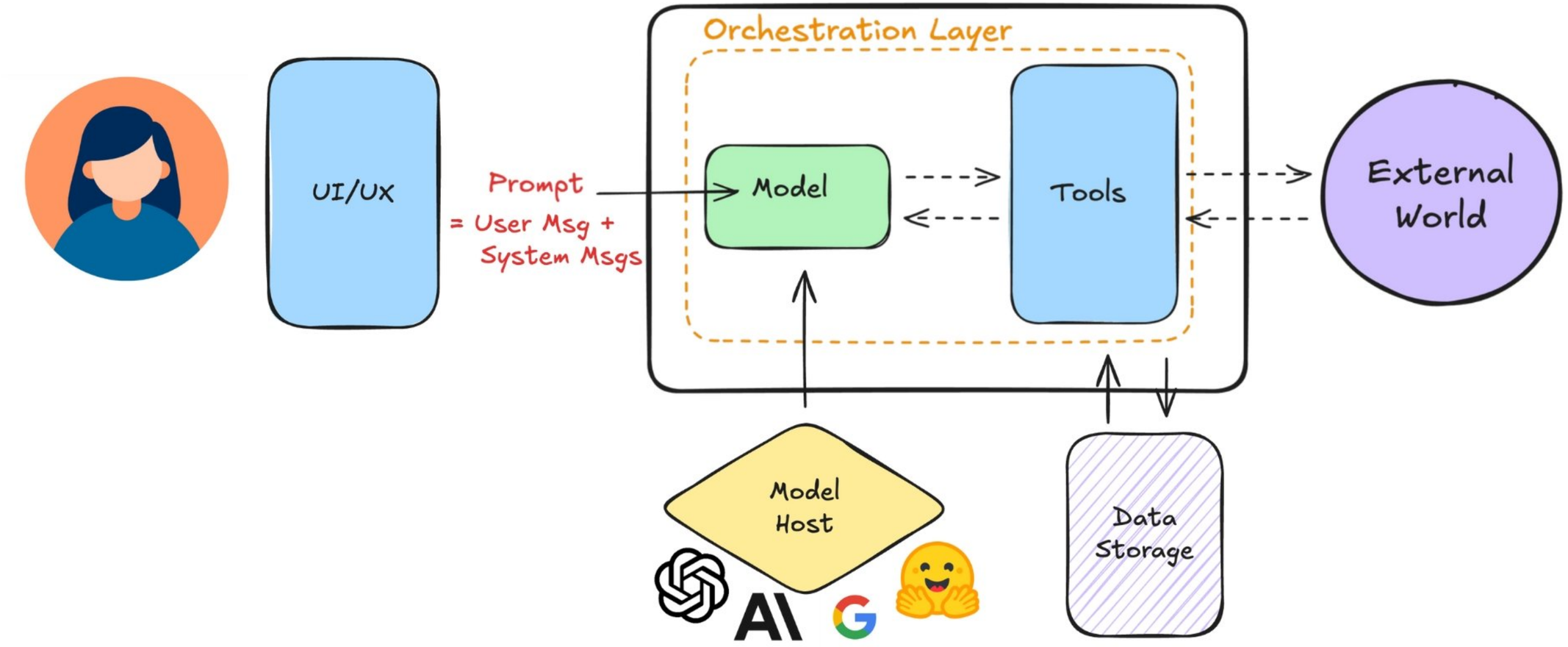
Components of an agent



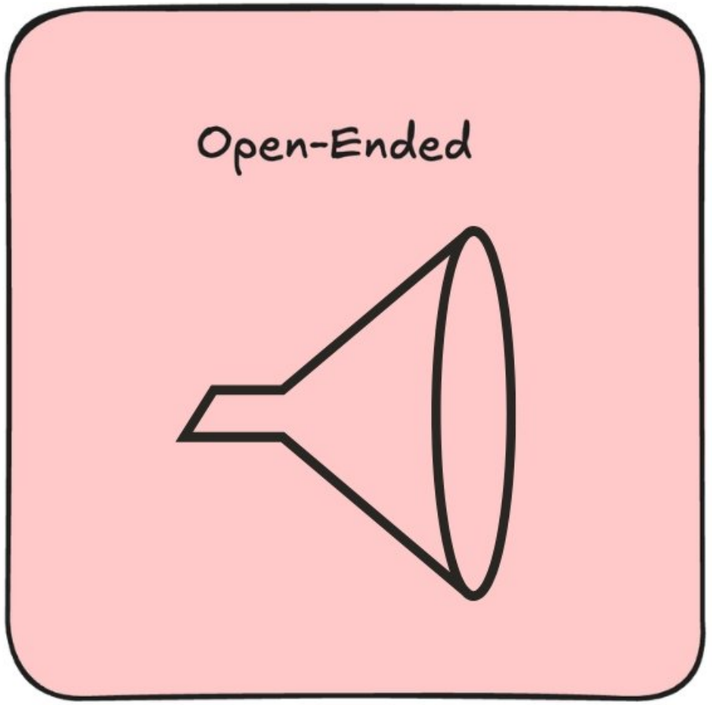
An agentic application



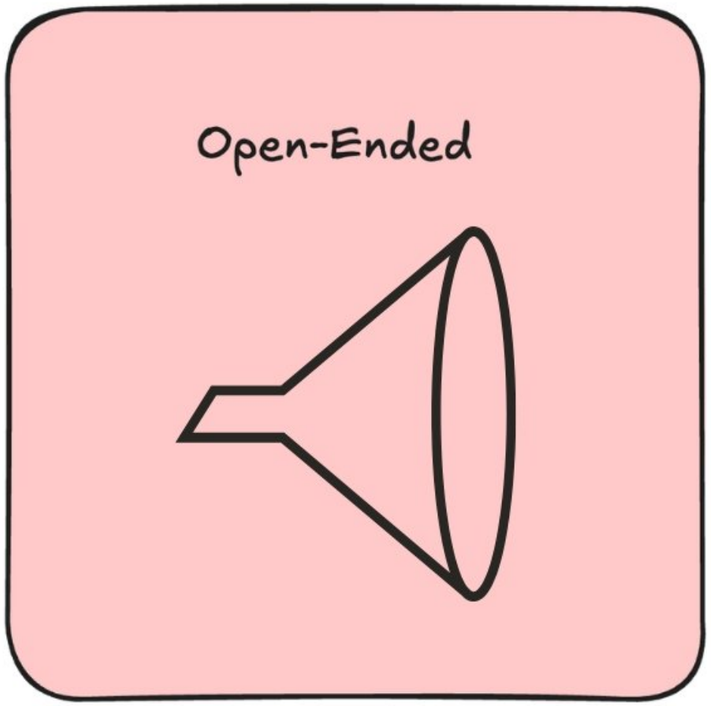
An agentic application

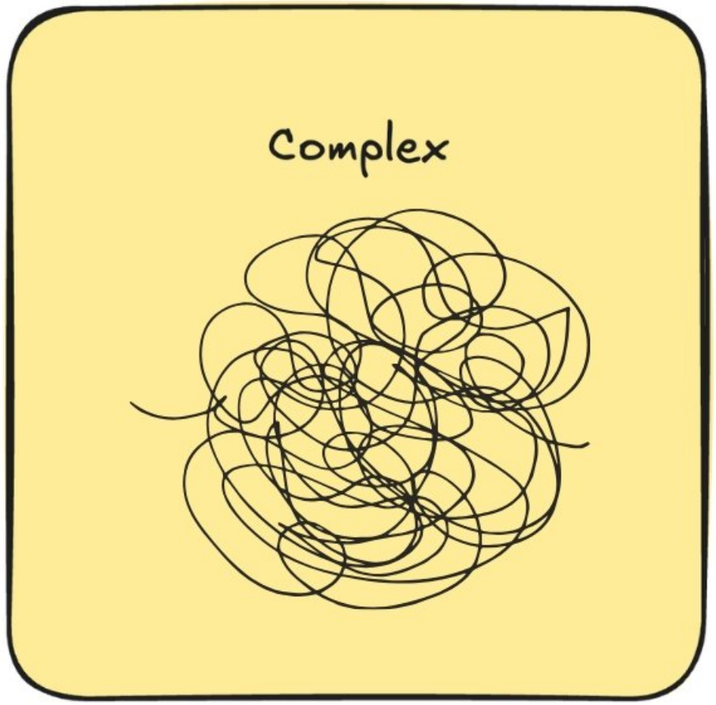


When to use agents?

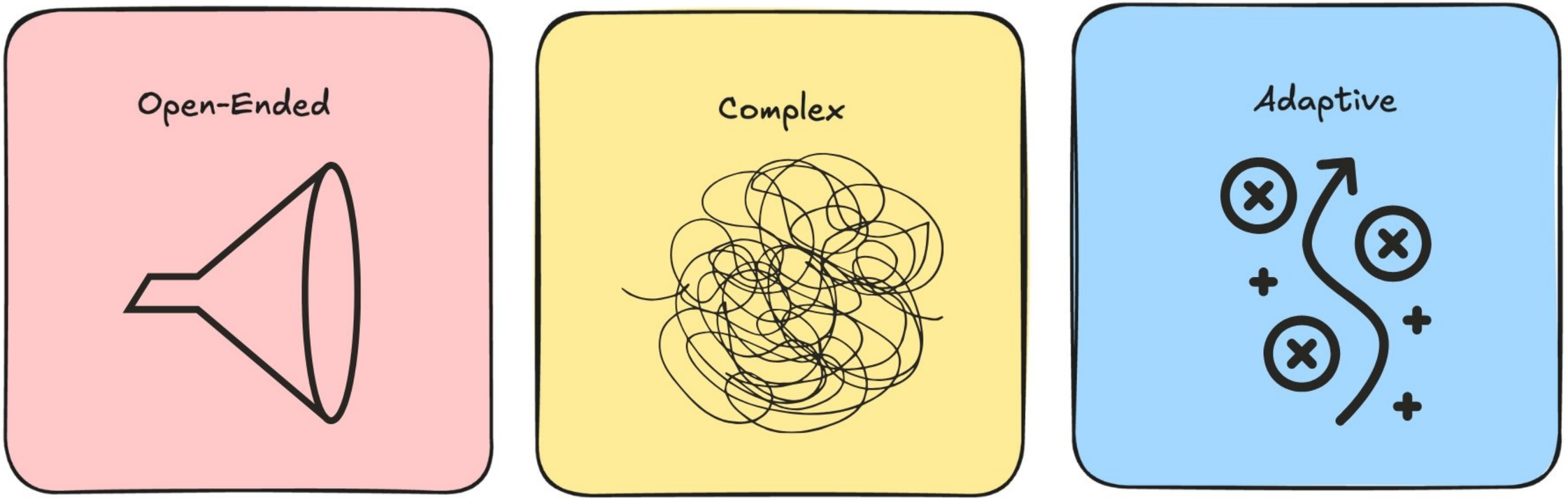


When to use agents?





When to use agents?



Example: sending reminder emails

Open-ended

Complex

Adaptive

Example: sending reminder emails

Open-ended

Complex

Adaptive

Example: sending reminder emails

Open-ended

Complex

Adaptive

Example: sending reminder emails

Open-ended

Complex

Adaptive

→ IT solution

Example: IT support agent

Example: IT support agent

Open-ended

Complex

Adaptive

→ Investigate AI agent solution

Let's practice!

BUILDING SCALABLE AGENTIC SYSTEMS

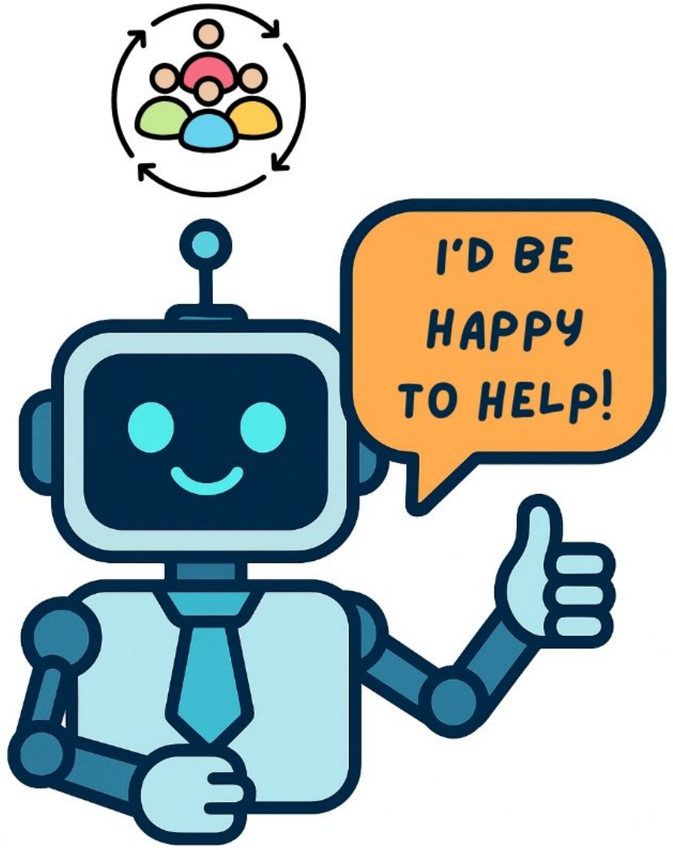
Design Principles for Scalable Agents

BUILDING SCALABLE AGENTIC SYSTEMS

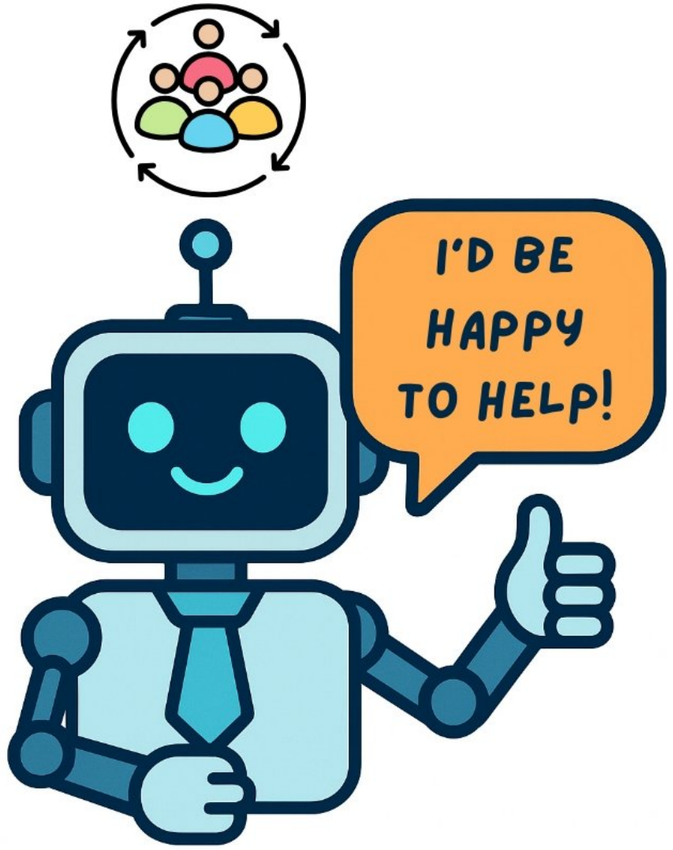
Korey Stegared-Pace
Senior AI Cloud Advocate, Microsoft

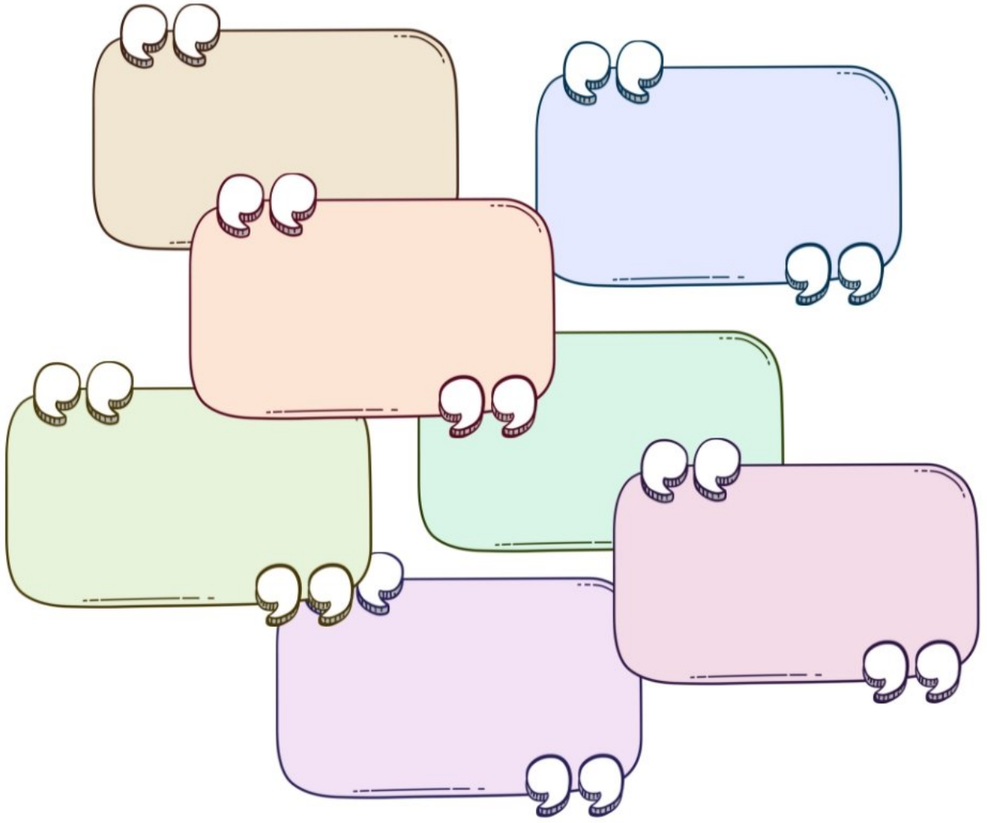
datacamp

Scalable or not scalable?

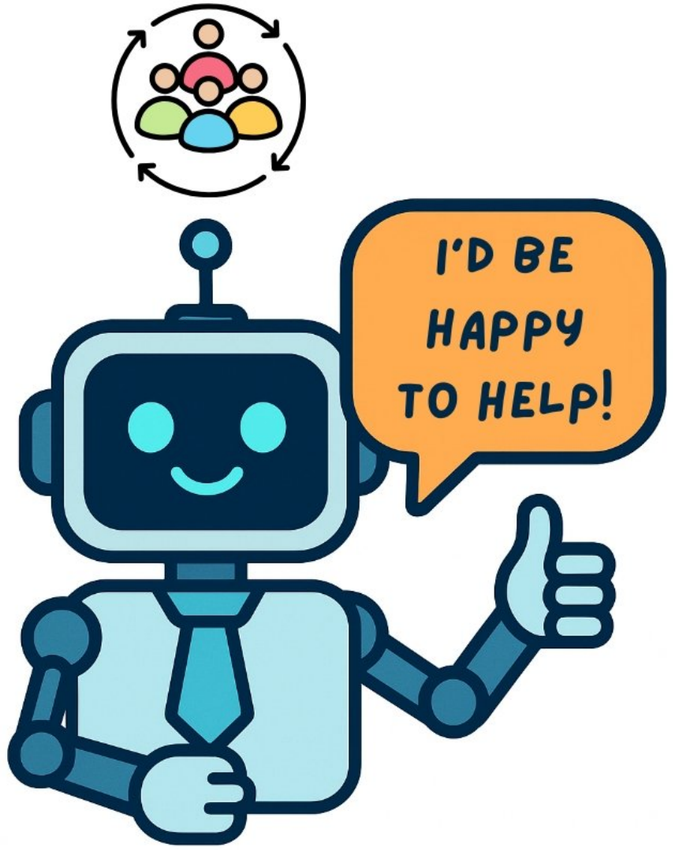


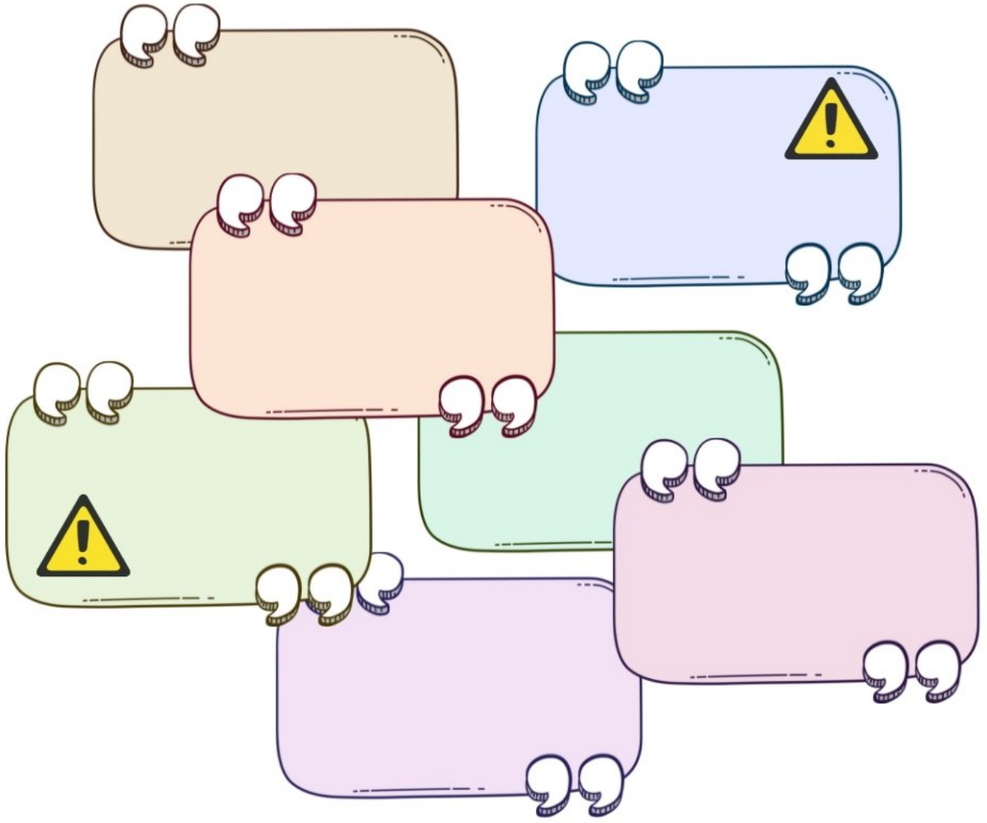
Scalable or not scalable?



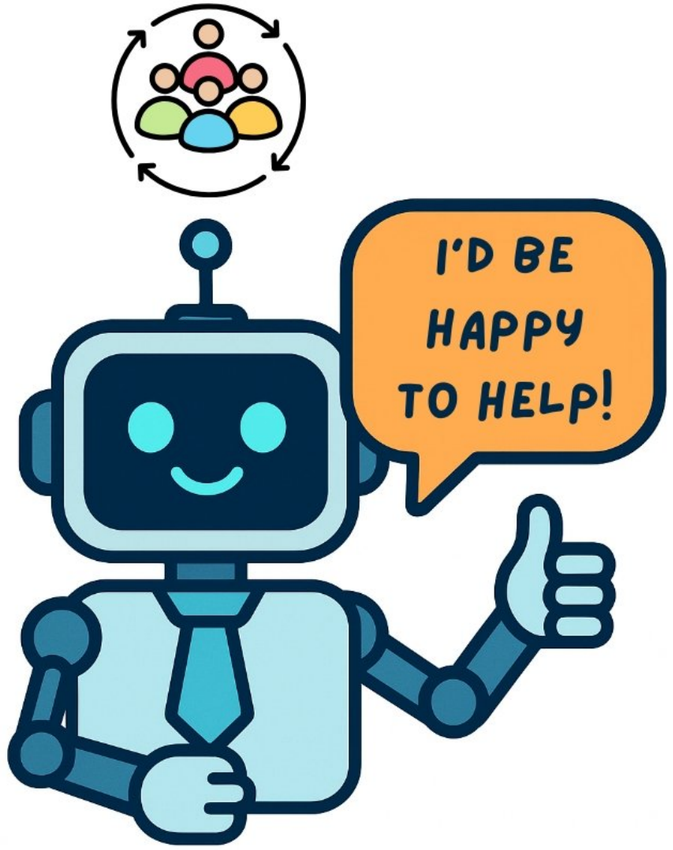


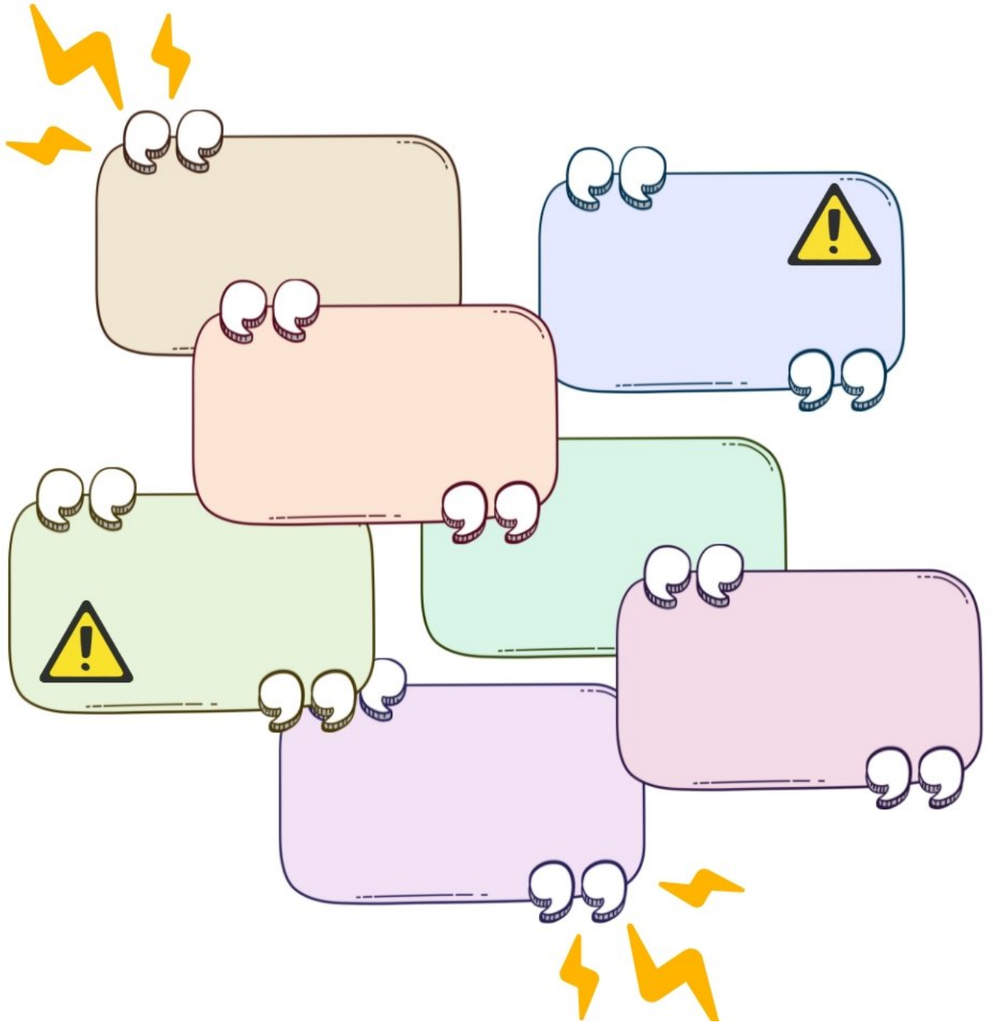
Scalable or not scalable?





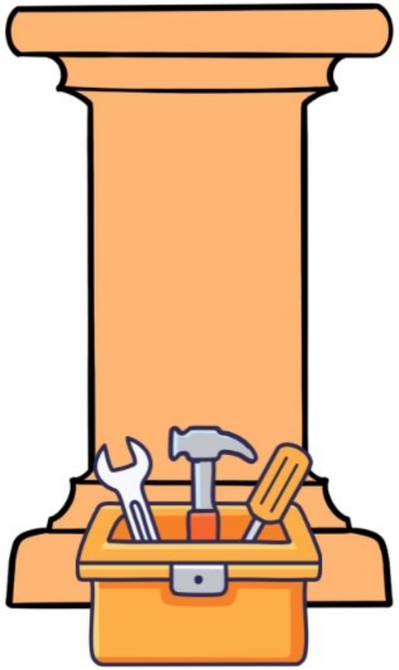
Scalable or not scalable?



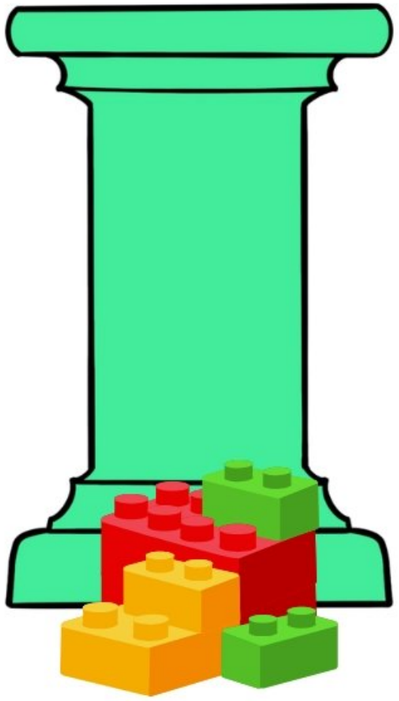


The pillars of scalable agentic systems

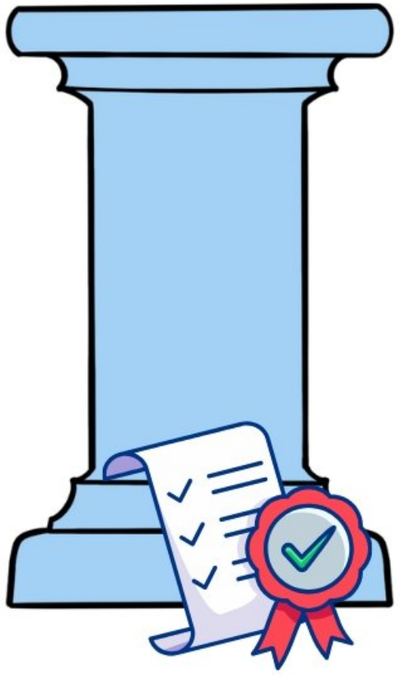
Robust Infrastructure and Tooling



Modular Design Architecture

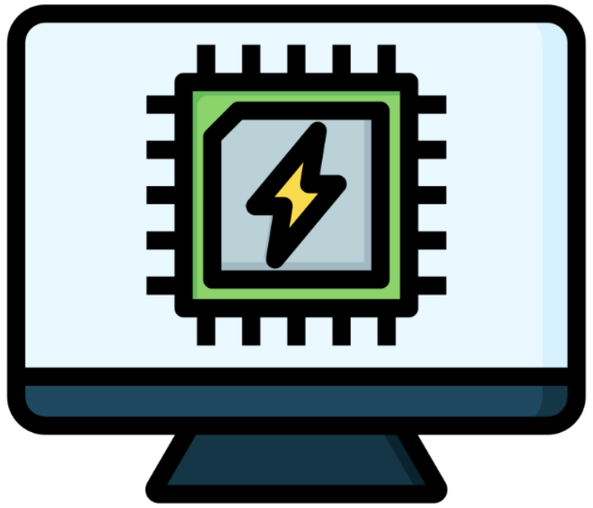


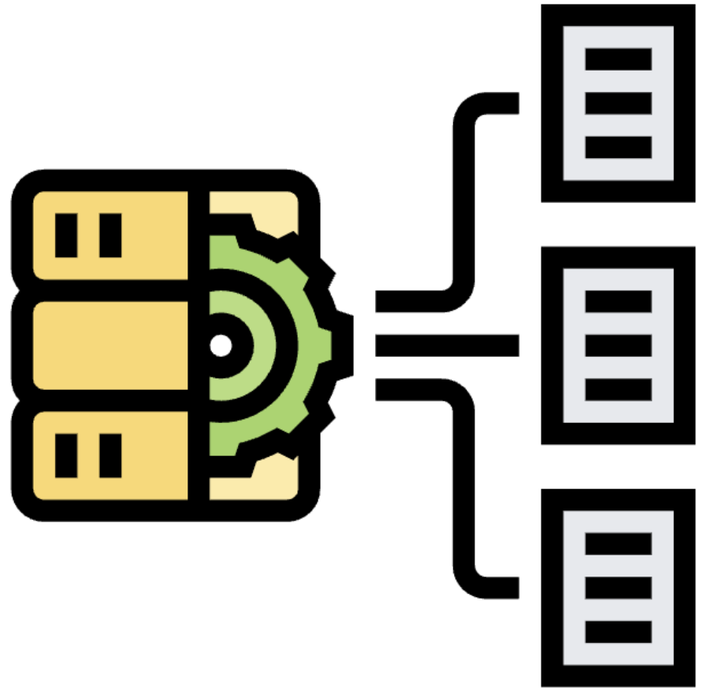
Continuous Evaluation & Feedback Loops



Robust infrastructure and tooling

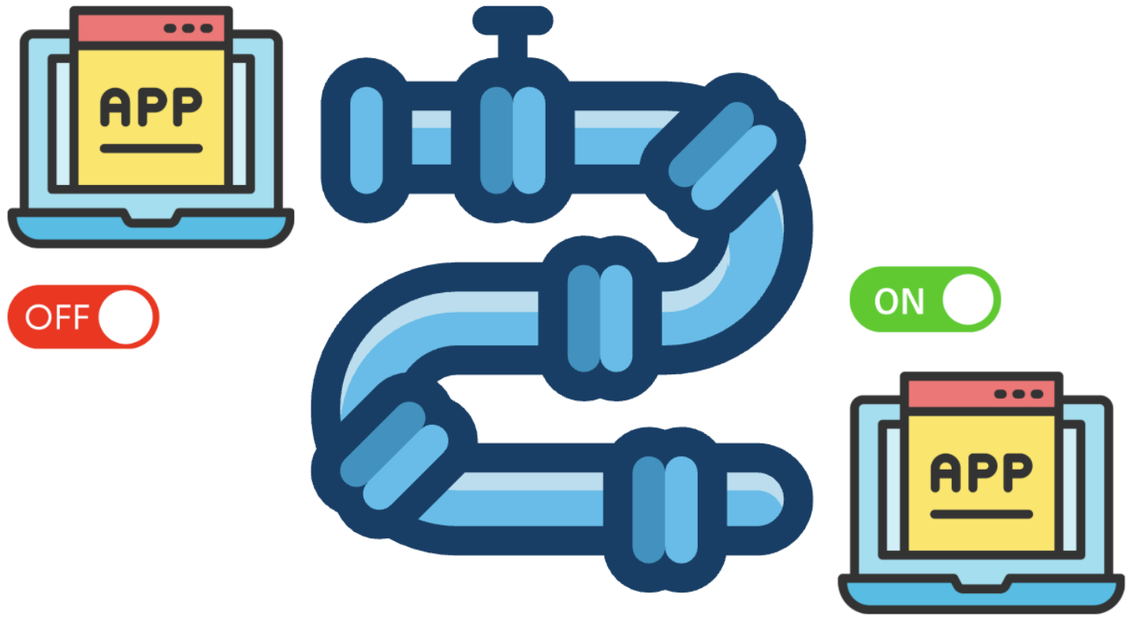
- **Compute:** Resources for running code
- **Storage:** Storing states, logs, and history



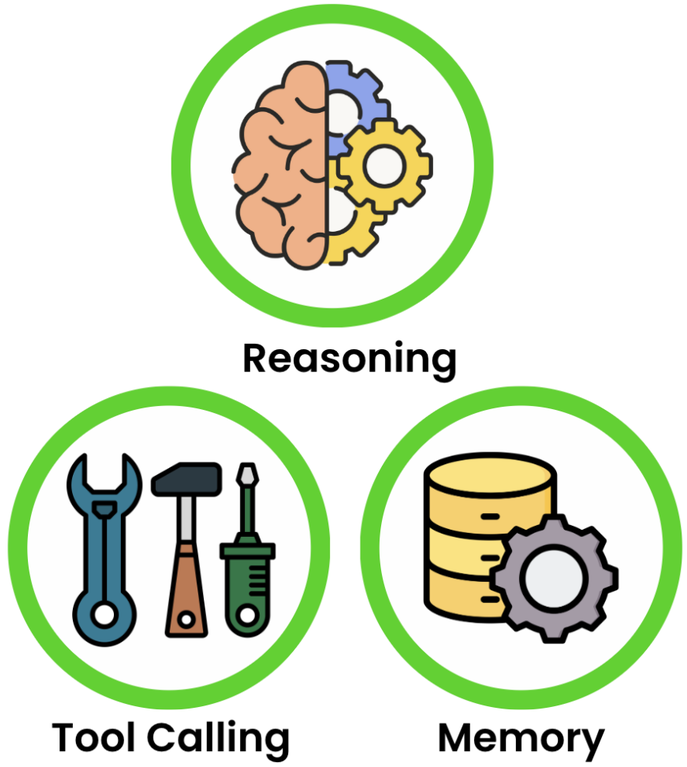


Robust infrastructure and tooling

Reliable Deployment Pipelines



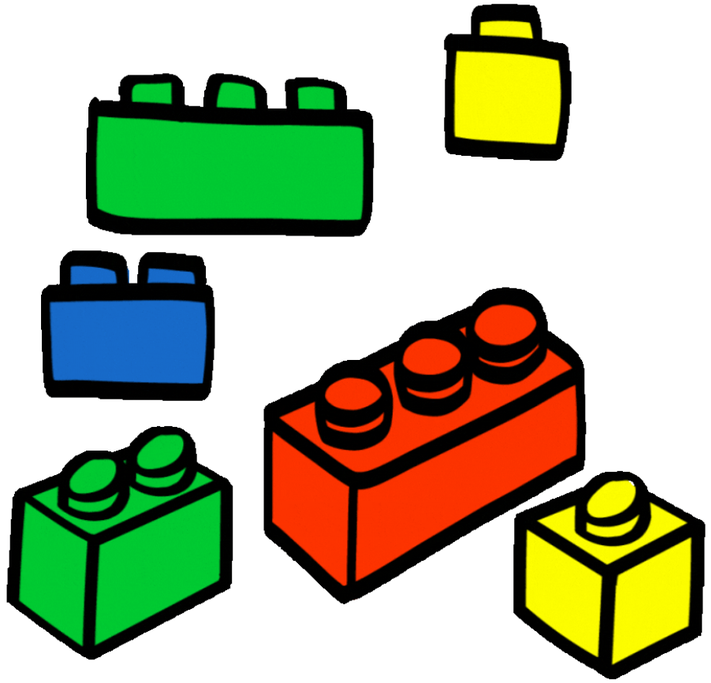
Established Agent Tooling



Modularity

> Designing systems as separate *independent* components to create a larger whole

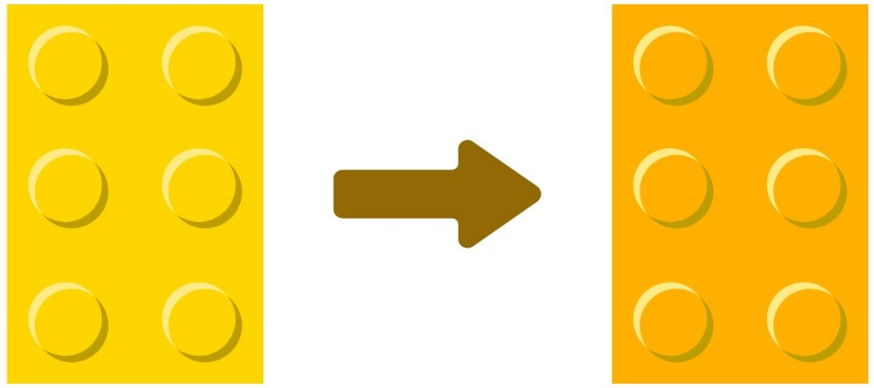
- Enables independent design, development, and maintenance



Modularity

> Designing systems as separate *independent* components to create a larger whole

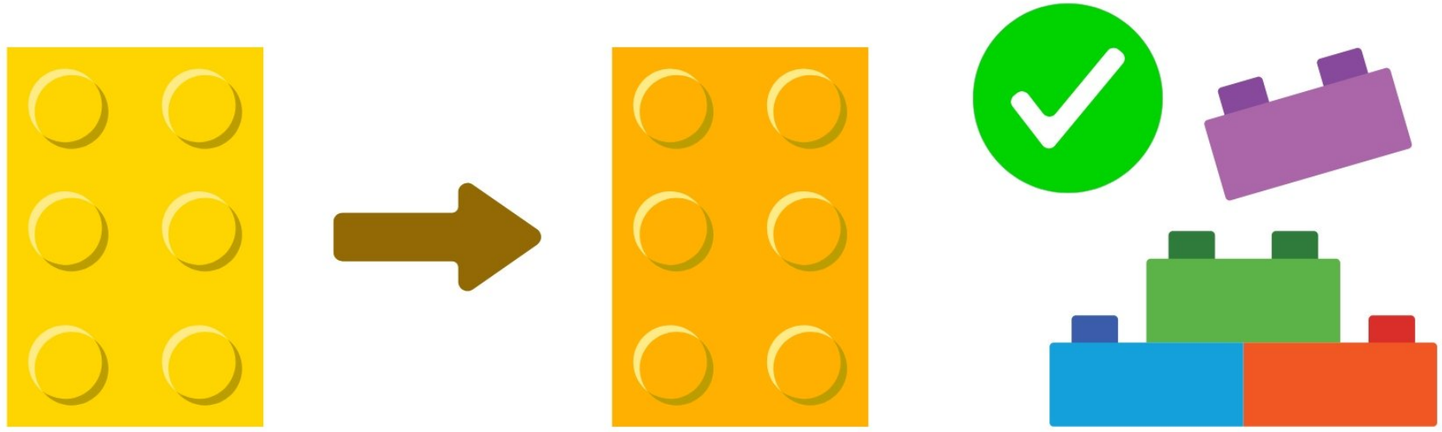
- Enables independent design, development, and maintenance



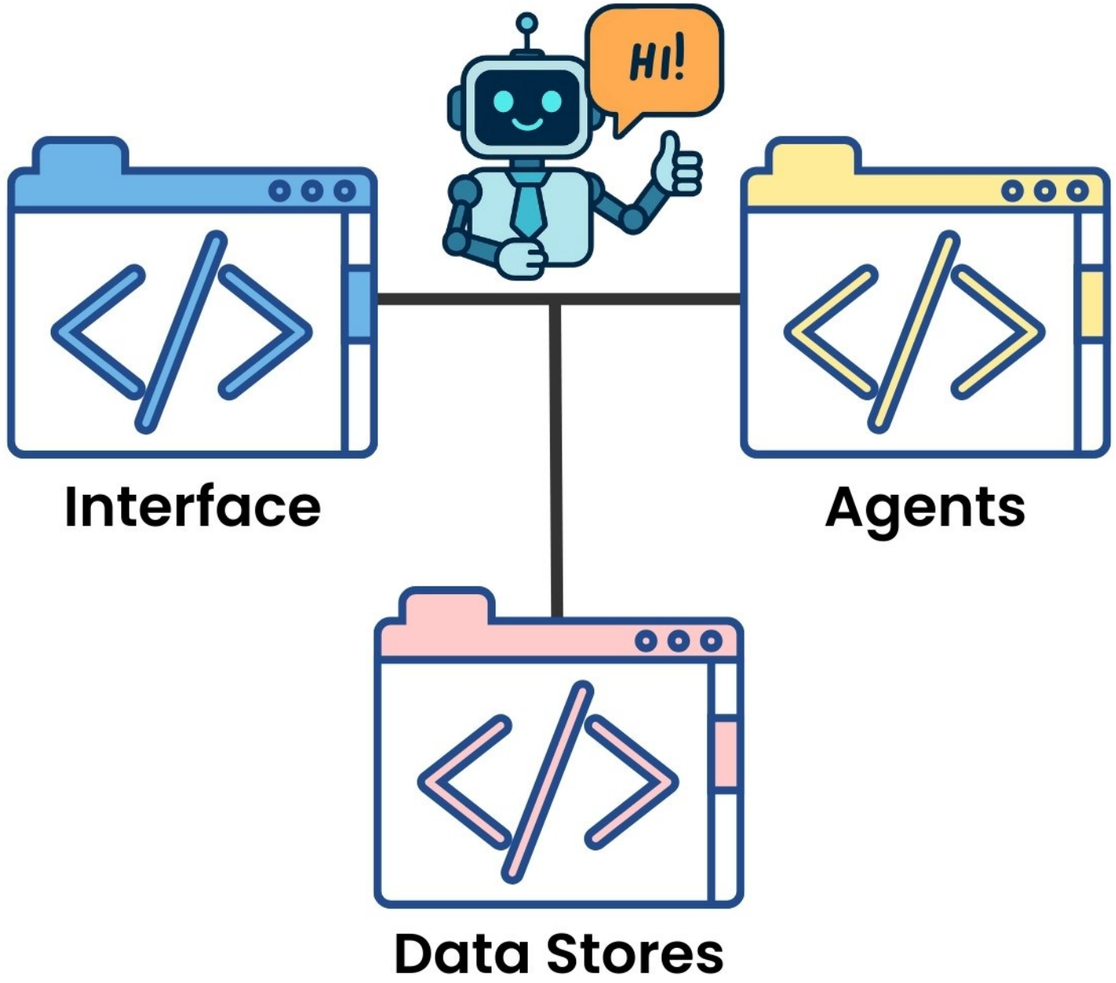
Modularity

> Designing systems as separate *independent* components to create a larger whole

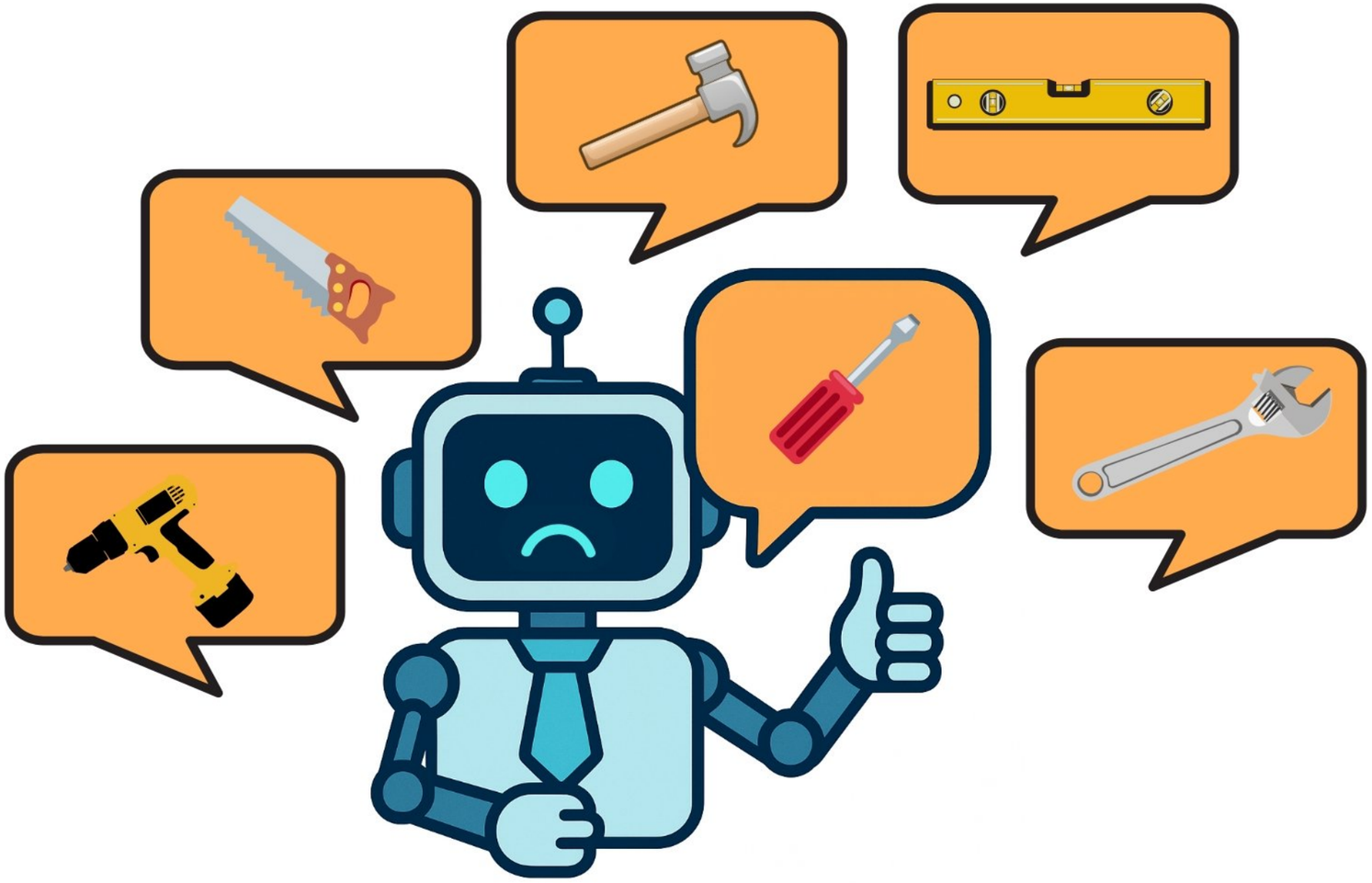
- Enables independent design, development, and maintenance



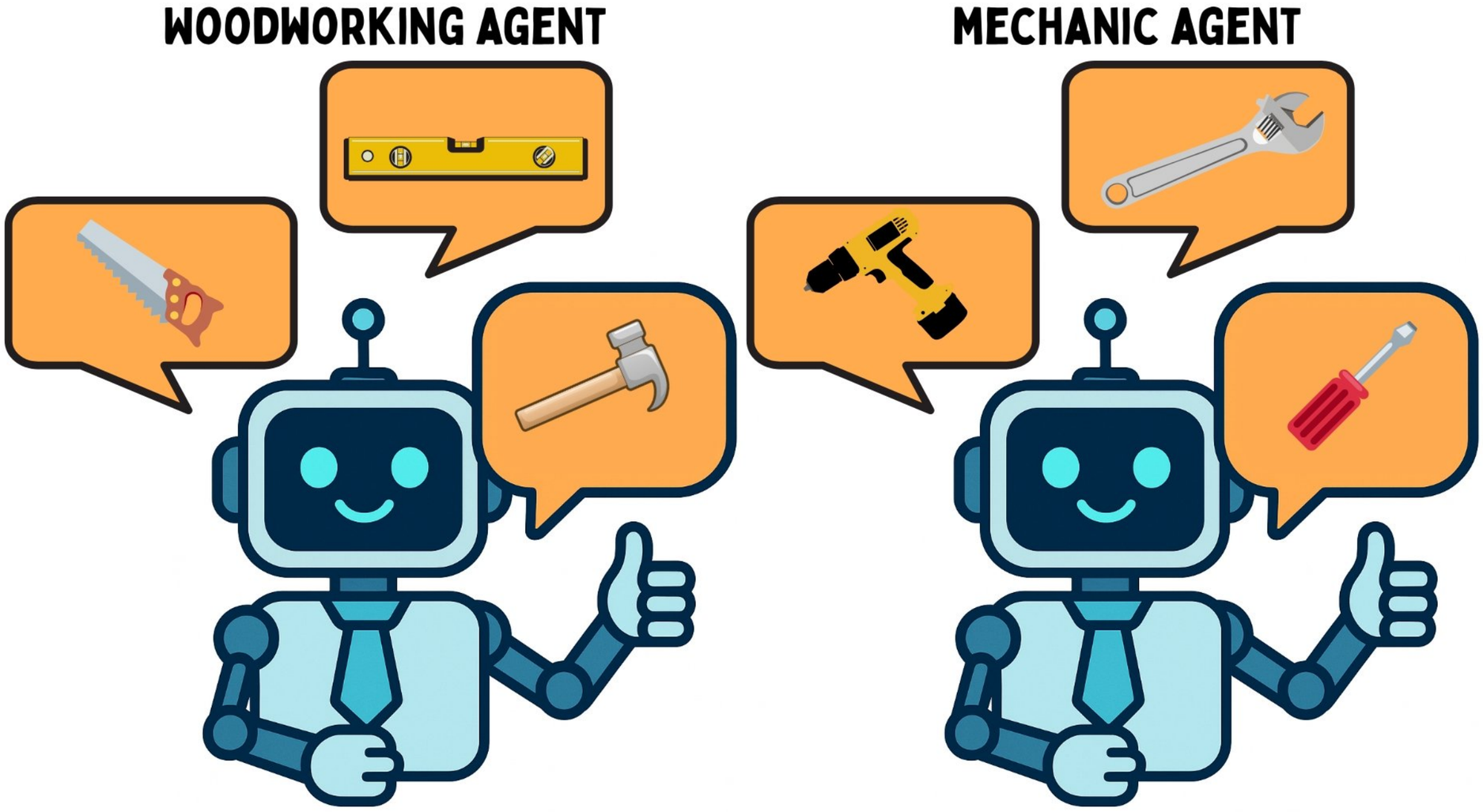
Modularity in agents: Software modularity



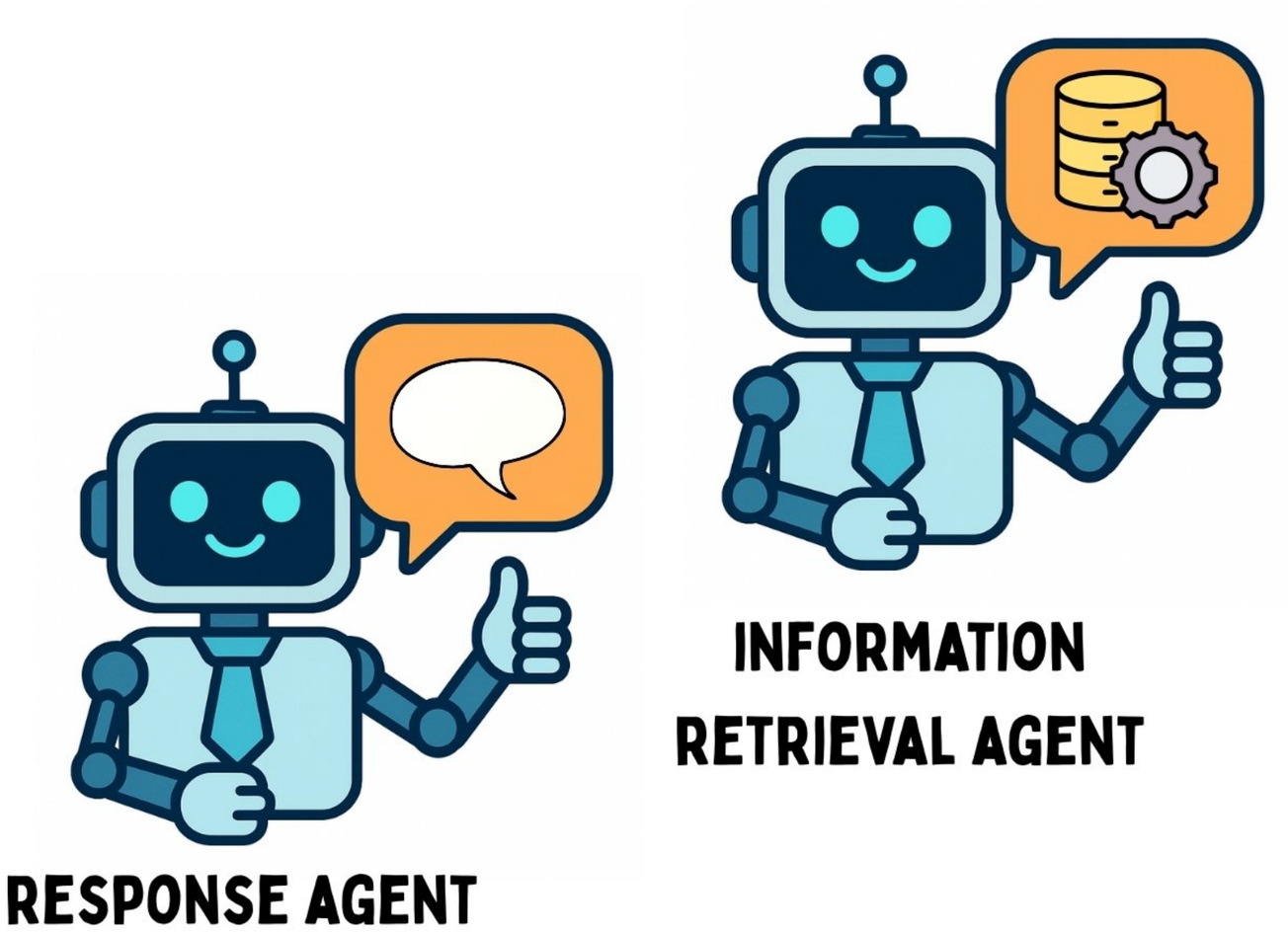
Modularity in agents: Multi-agents



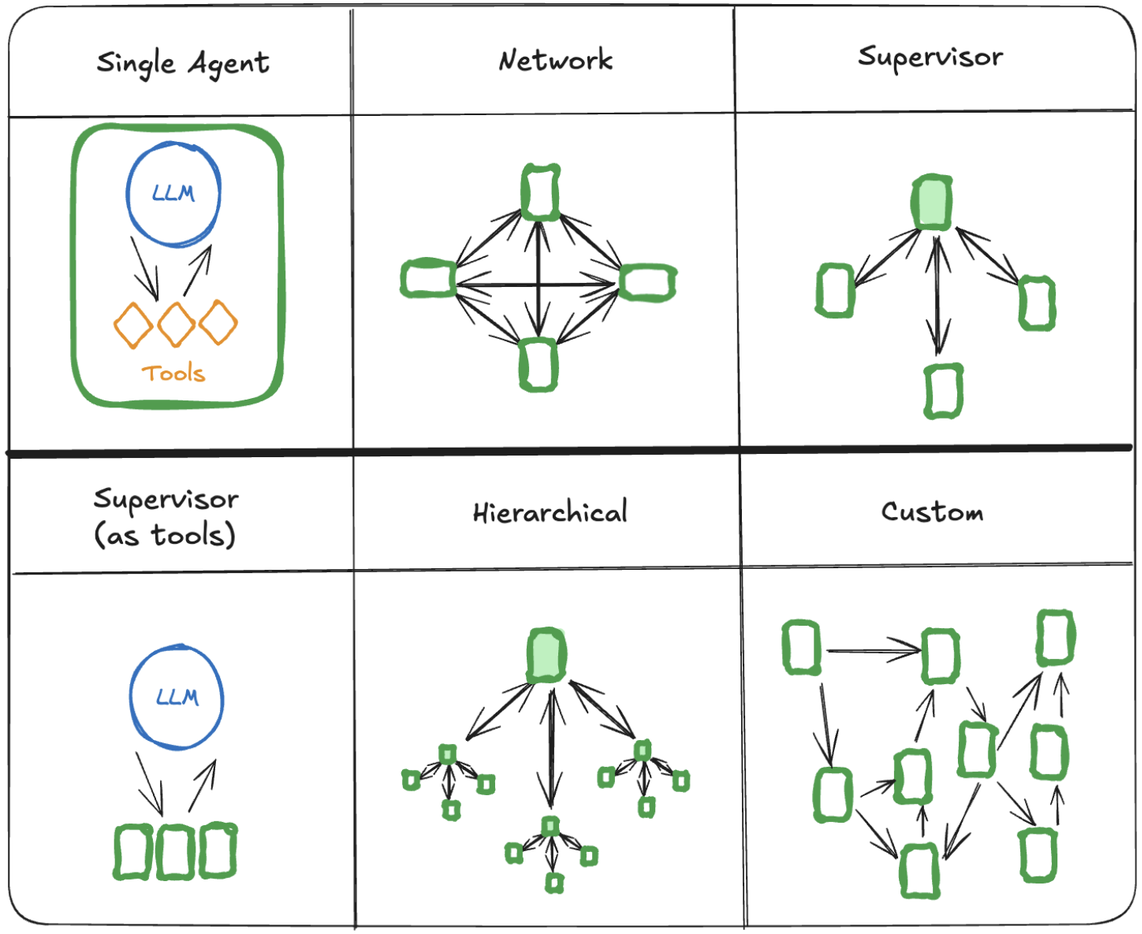
Modularity in agents: Multi-agents



Example: Customer support multi-agent



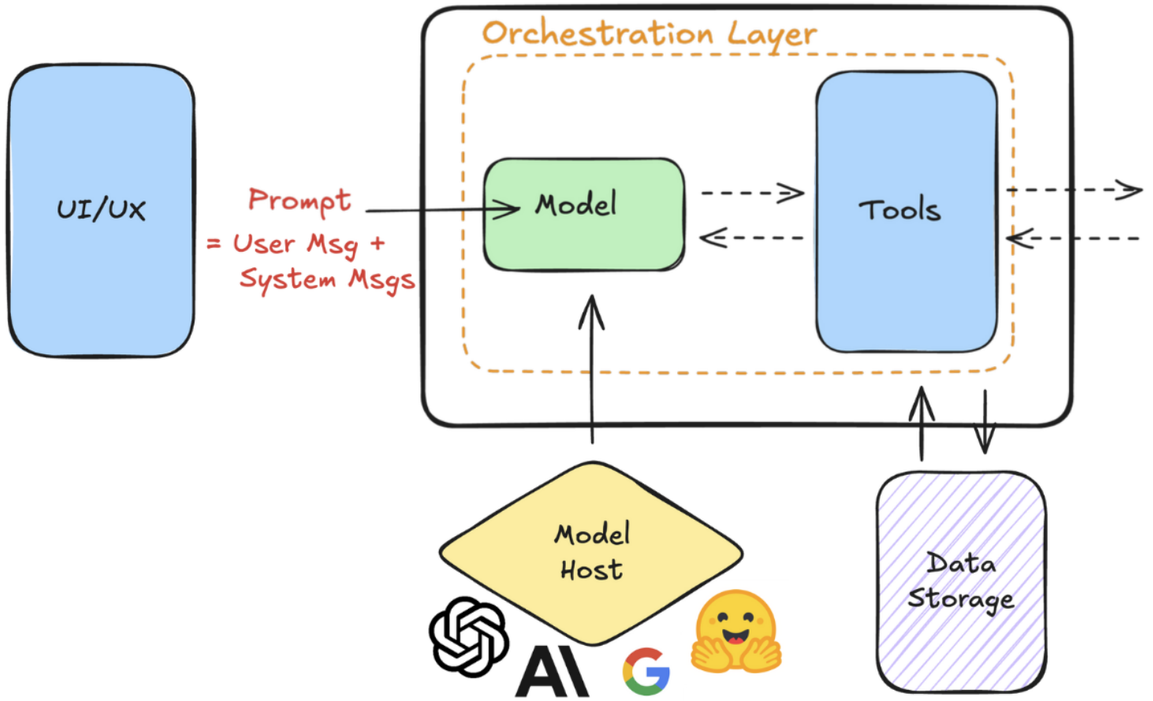
Sneak peak at Chapter 2



[1] https://langchain-ai.github.io/langgraph/concepts/multi_agent/#multi-agent-architectures

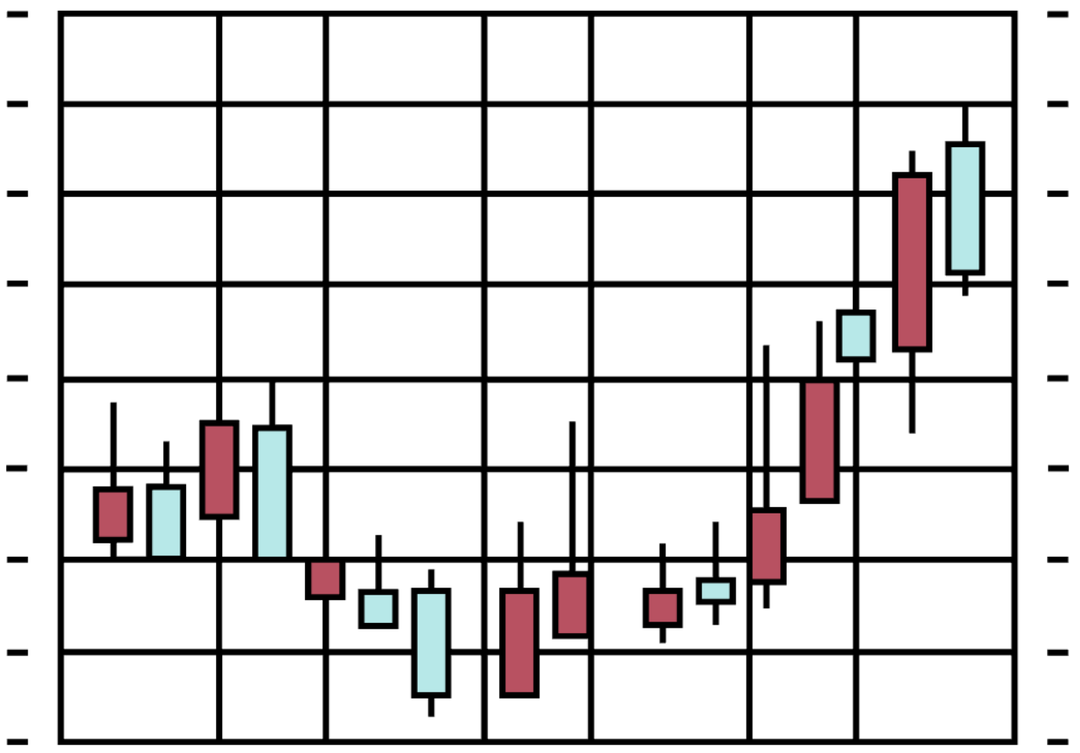
Continuous evaluation and feedback loops

- Comprehensive *monitoring*



Continuous evaluation and feedback loops

- Comprehensive *monitoring*
- Common metrics:
  - Success rate
  - Latency
  - Errors



Continuous evaluation and feedback loops

- Comprehensive *monitoring*
- Common metrics:
  - Success rate
  - Latency
  - Errors
- Collect user feedback

Let's practice!

BUILDING SCALABLE AGENTIC SYSTEMS

How AI Agents Scale (and Fail)

BUILDING SCALABLE AGENTIC SYSTEMS

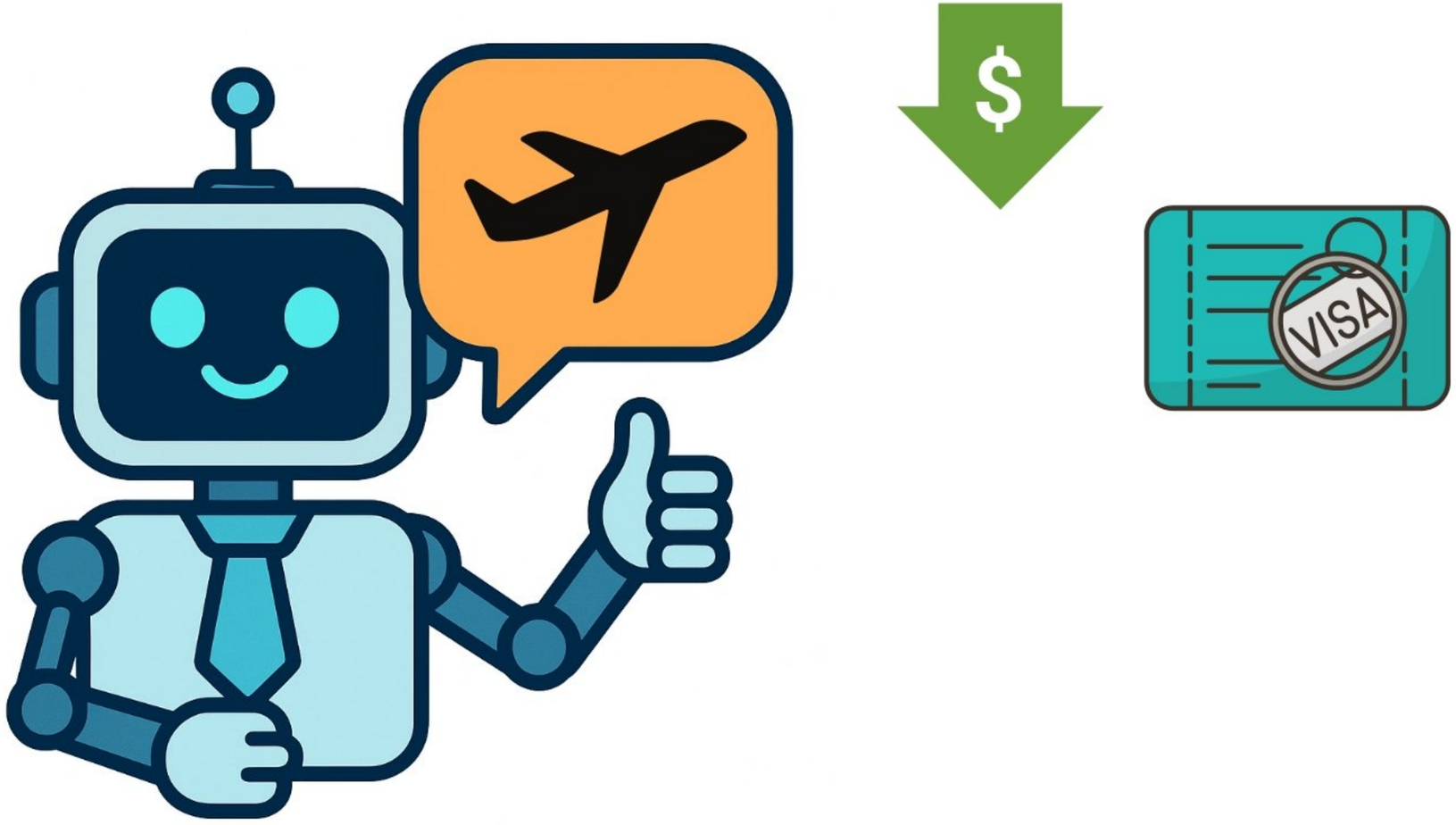
Korey Stegared-Pace
Senior AI Cloud Advocate, Microsoft

datacamp

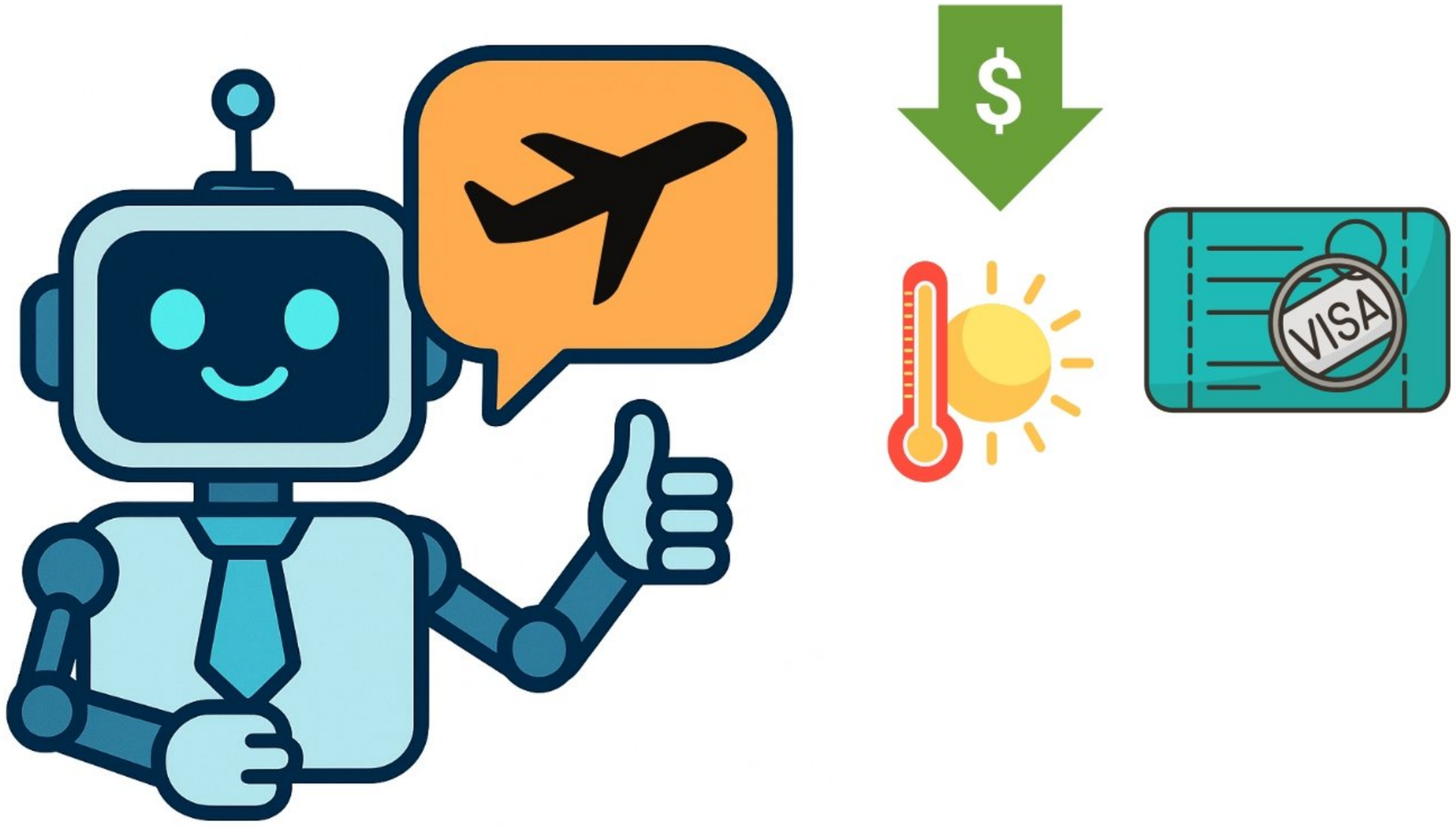
The confident start

The confident start

The confident start



The confident start



The confident start

The confident start

Trouble at scale

Failure mode #1 - Fragile evaluation

FRAGMENTED Travel to Naples...
> From Stockholm...
> Fly to Rome first...
> Train from Naples to Rome

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

EMOJIS London ✈️ Bali

LANGUAGES *Vorrei prenotare un volo per Parigi.*

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

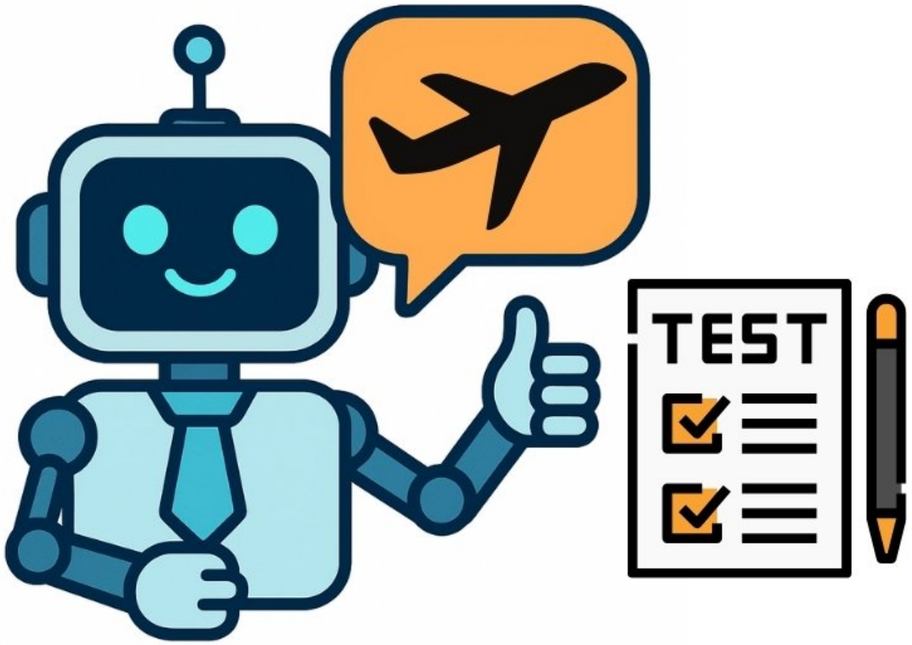
Failure mode #1 - Fragile evaluation

FRAGMENTED Travel to Naples...
> From Stockholm...
> Fly to Rome first...
> Train from Naples to Rome

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

EMOJIS London ✈️ Bali

LANGUAGES *Vorrei prenotare un volo per Parigi.*



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

Failure mode #1 - Fragile evaluation

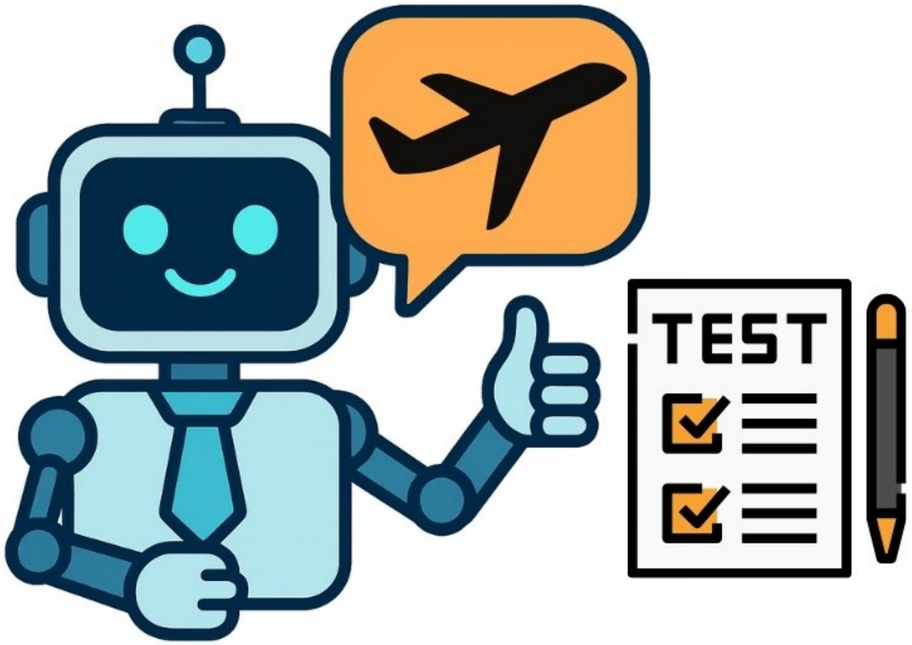
FRAGMENTED Travel to Naples...
> From Stockholm...
> Fly to Rome first...
> Train from Naples to Rome

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

EMOJIS London ✈️ Bali

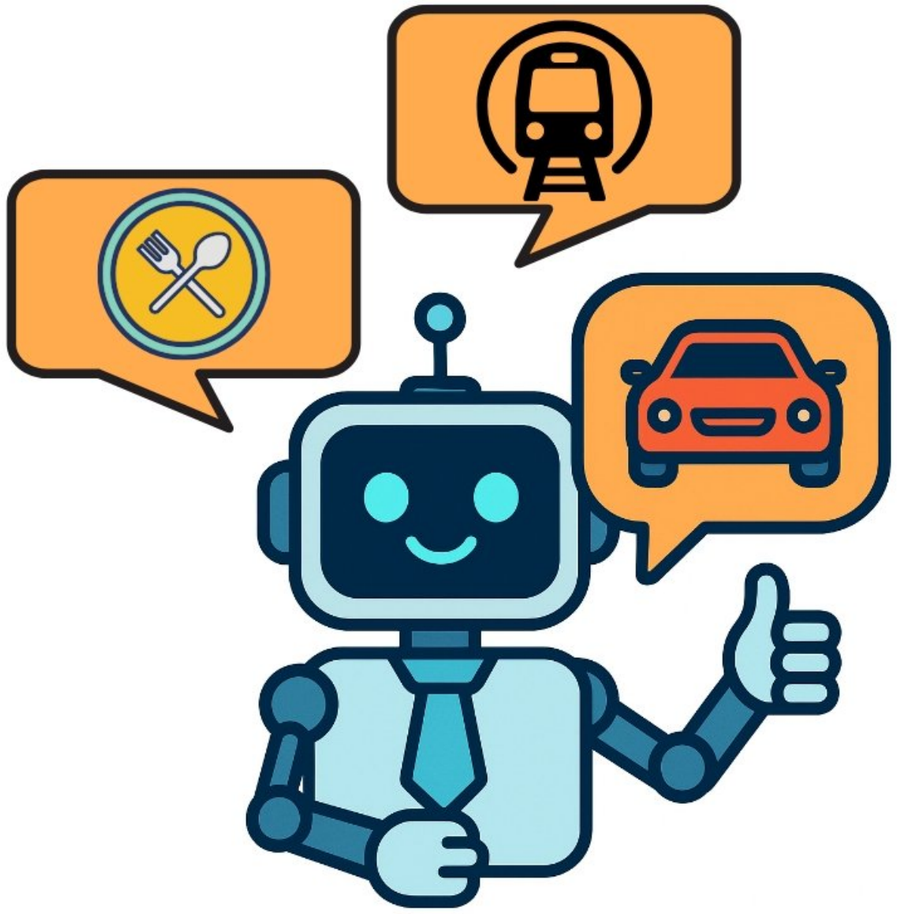
LANGUAGES *Vorrei prenotare un volo per Parigi.*

Include **MESSY** data in your evaluations

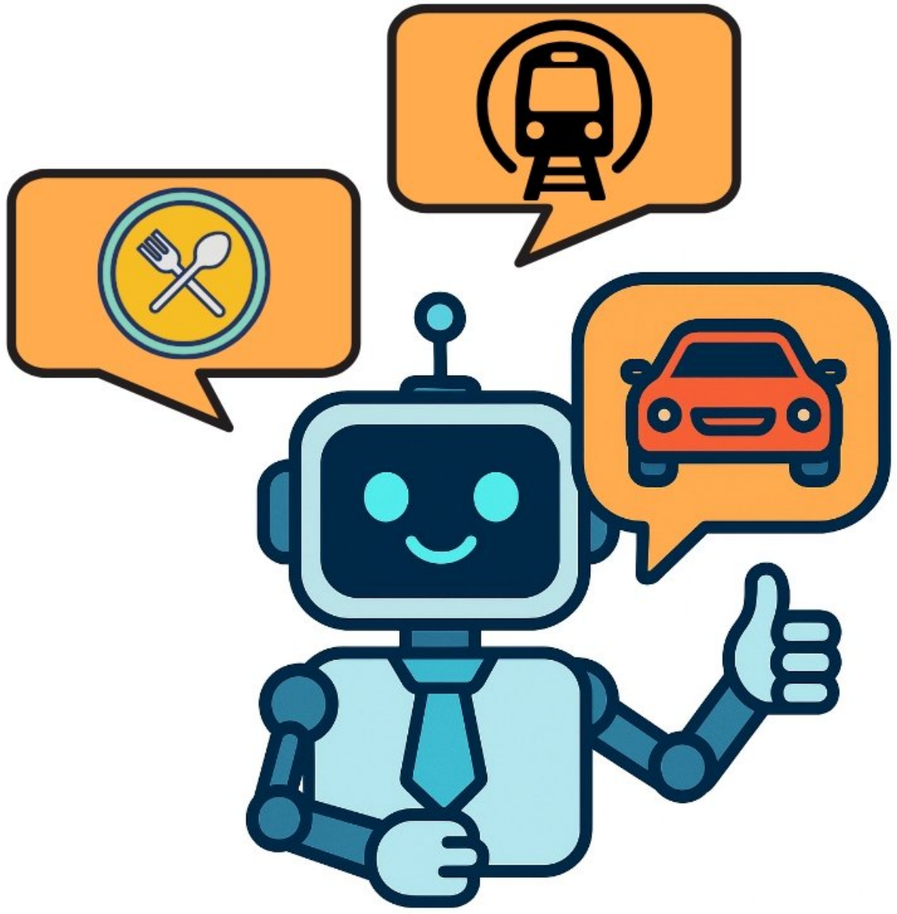


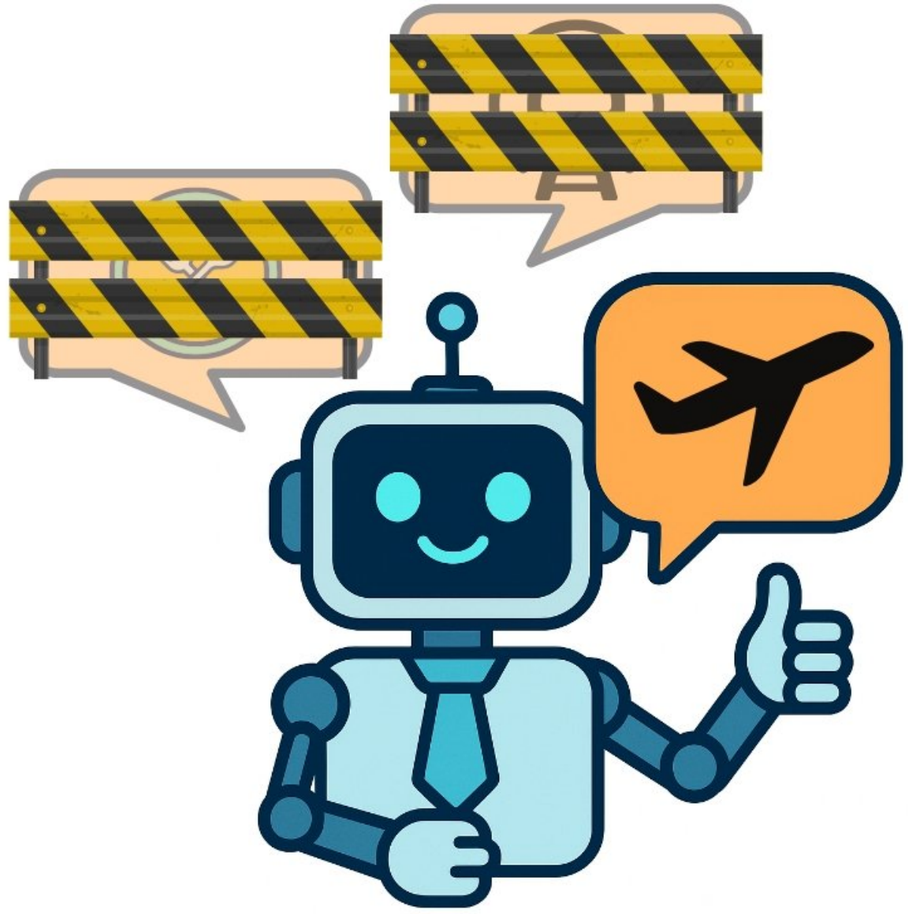
[1] nip over = British slang for a quick visit

Failure mode #2 - Intent drift



Failure mode #2 - Intent drift





Failure mode #3 - Undesirable feedback loops

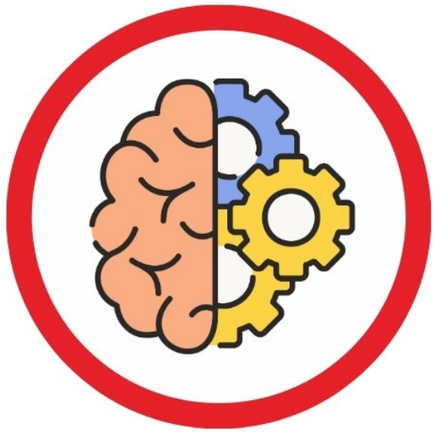
Upvoting → jokes, sarcasm, etc.

Result → truth is deprioritized

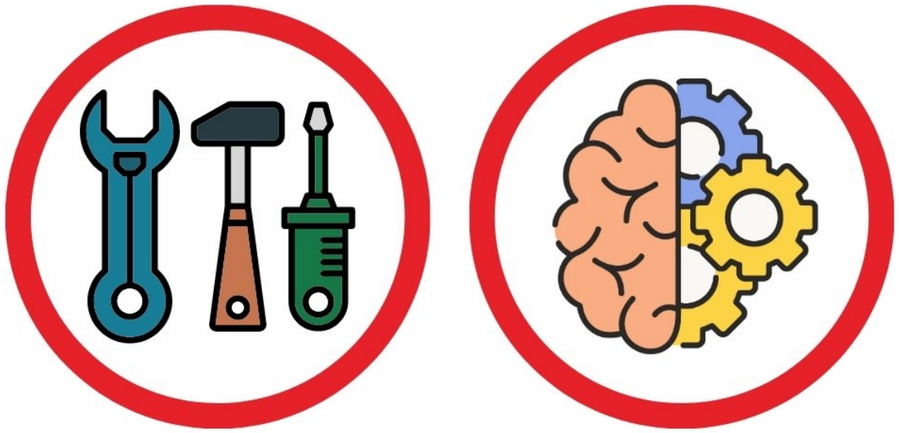
Define clear metrics for all aspects of performance

- *Truthfulness*
- *Clarity*
- *Tone*

Failure mode #4 - Latency bottlenecks

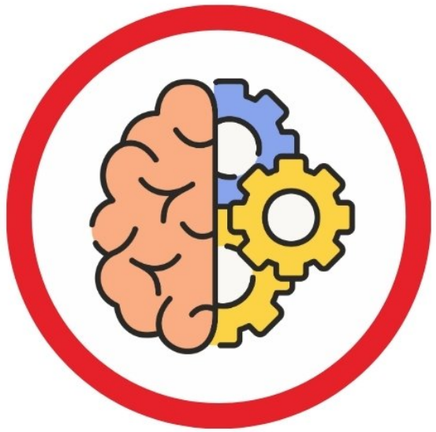


Failure mode #4 - Latency bottlenecks



Failure mode #4 - Latency bottlenecks

Tool Use

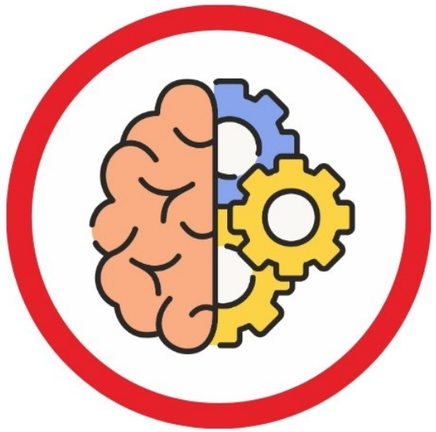


Reasoning

Retrieval

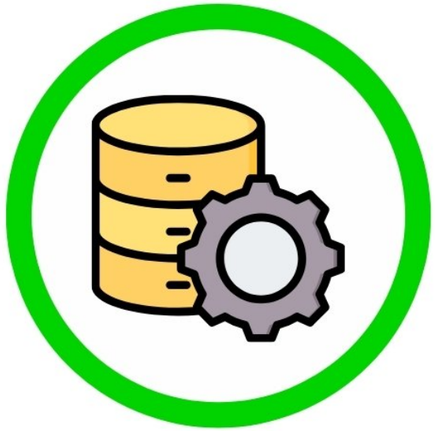
Failure mode #4 - Latency bottlenecks

Tool Use



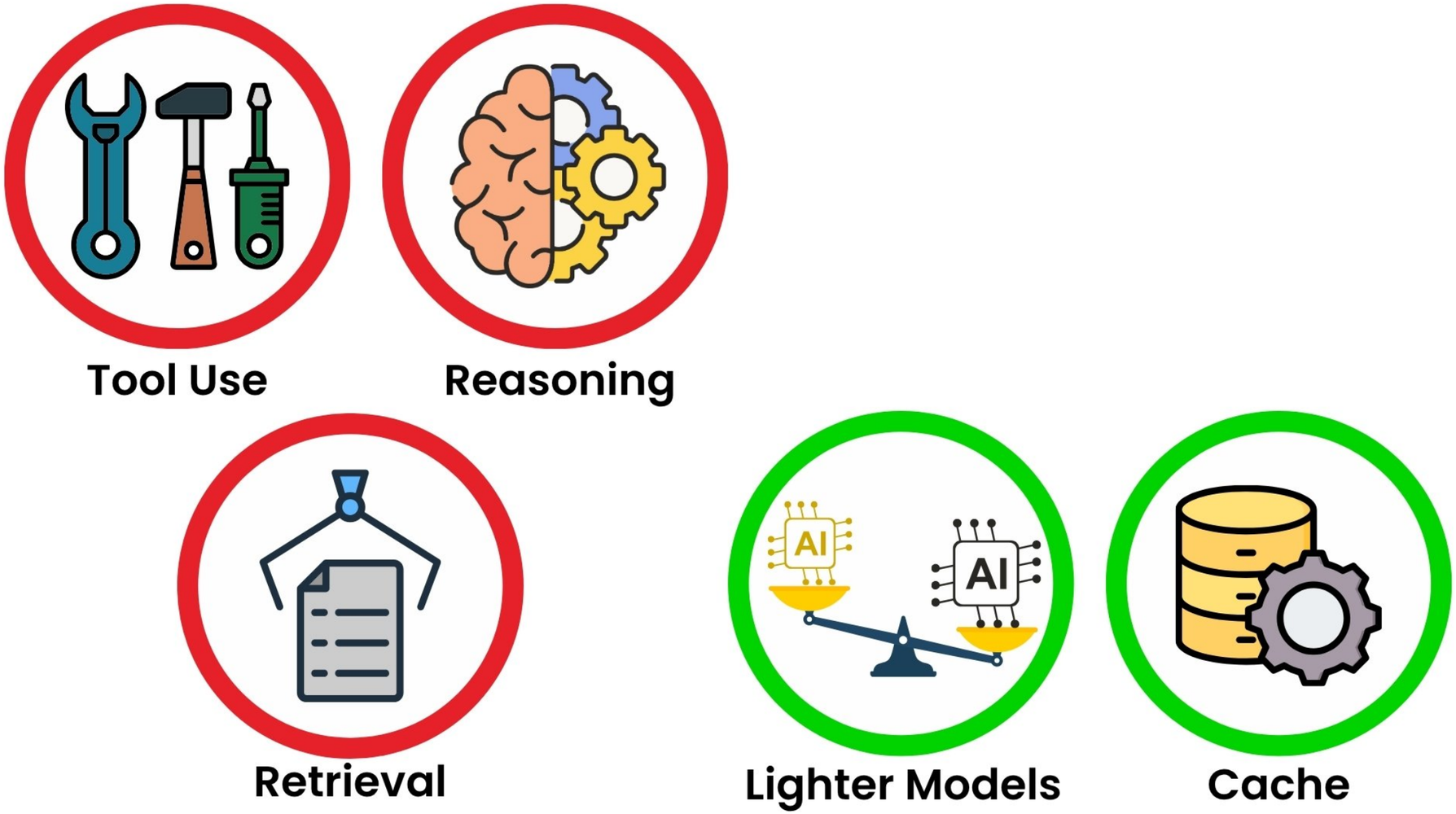
Reasoning

Retrieval

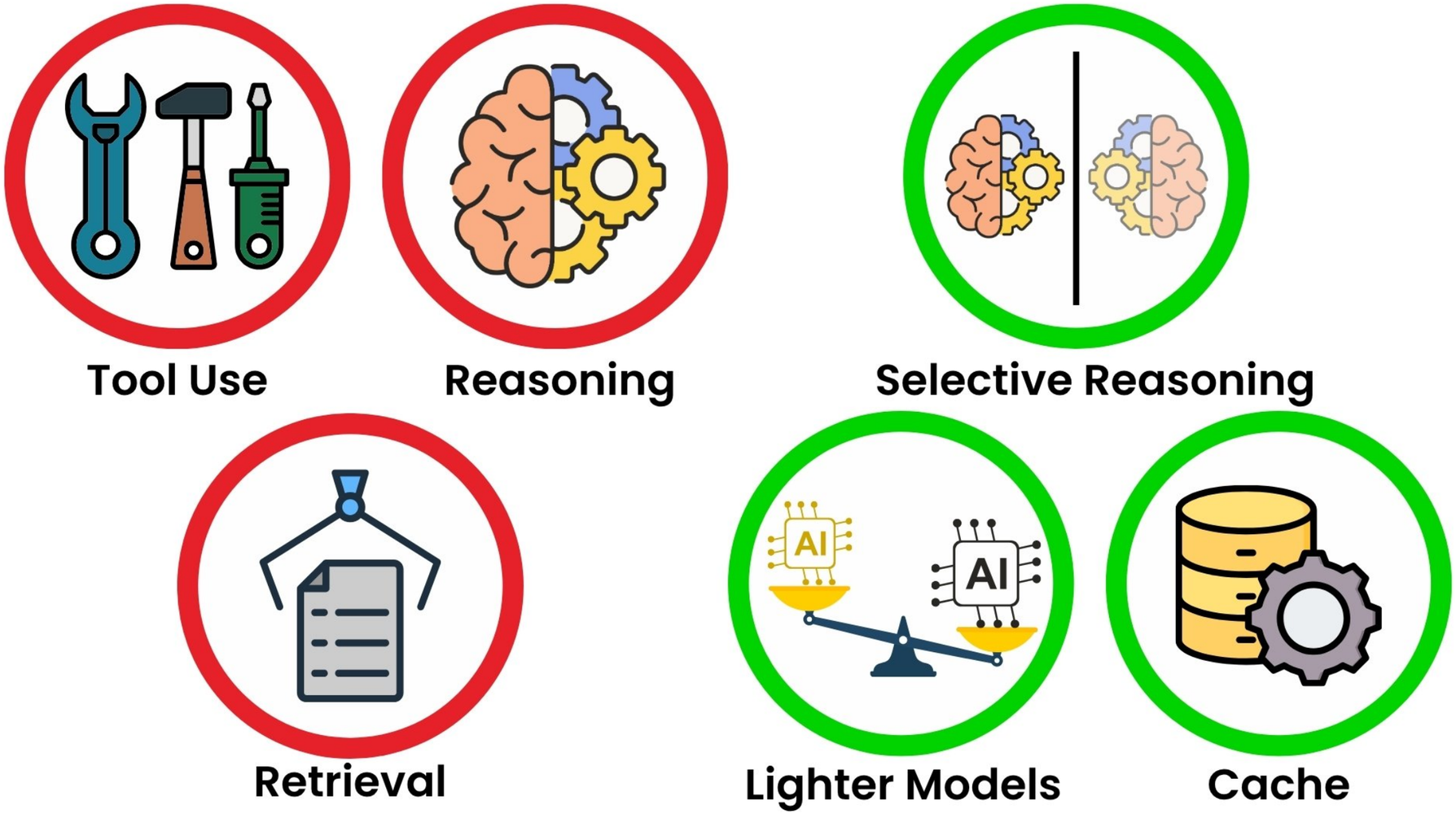


Cache

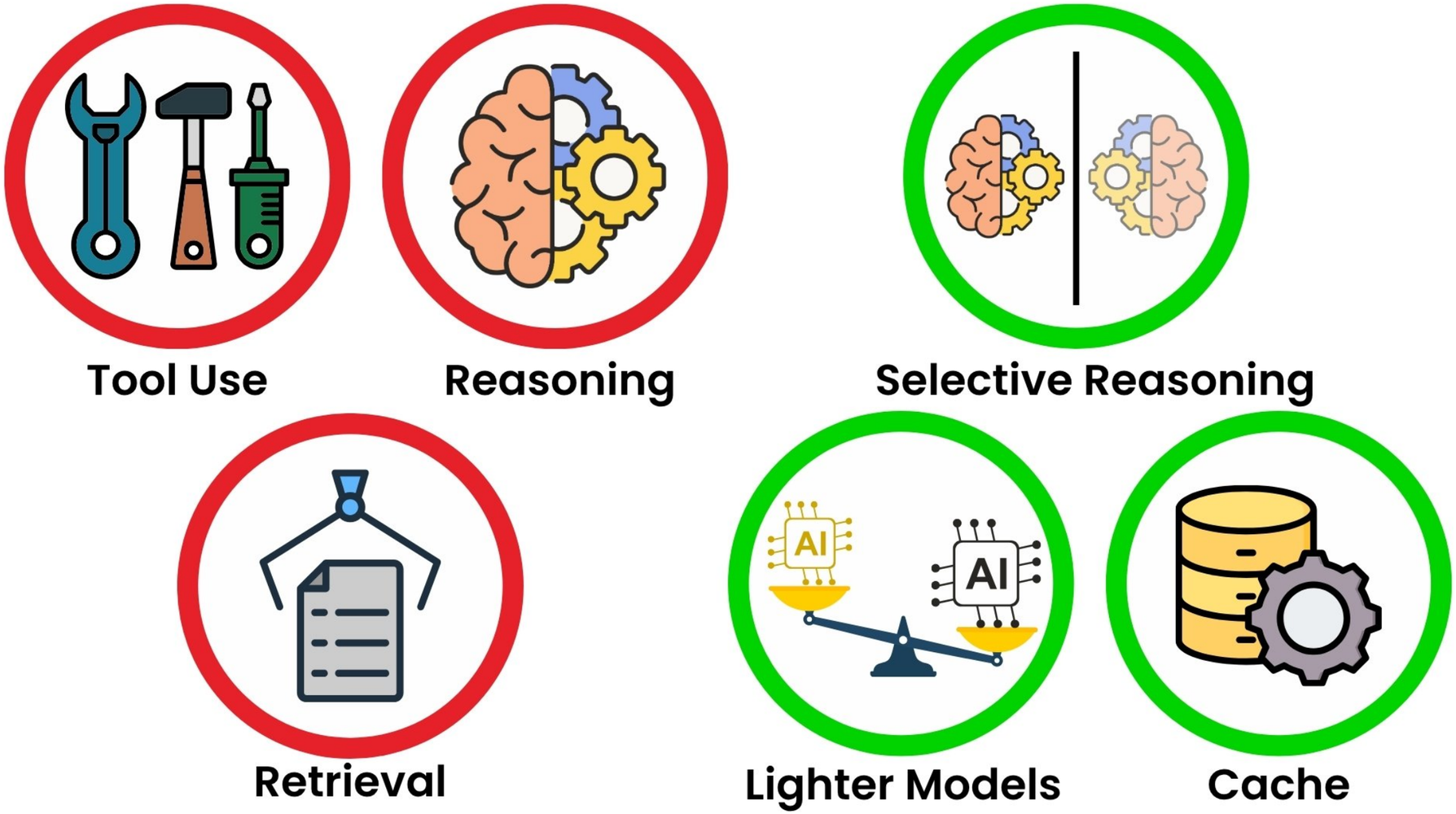
Failure mode #4 - Latency bottlenecks



Failure mode #4 - Latency bottlenecks



Failure mode #5 - Cost explosion



Failure mode #5 - Cost explosion

Let's practice!

BUILDING SCALABLE AGENTIC SYSTEMS