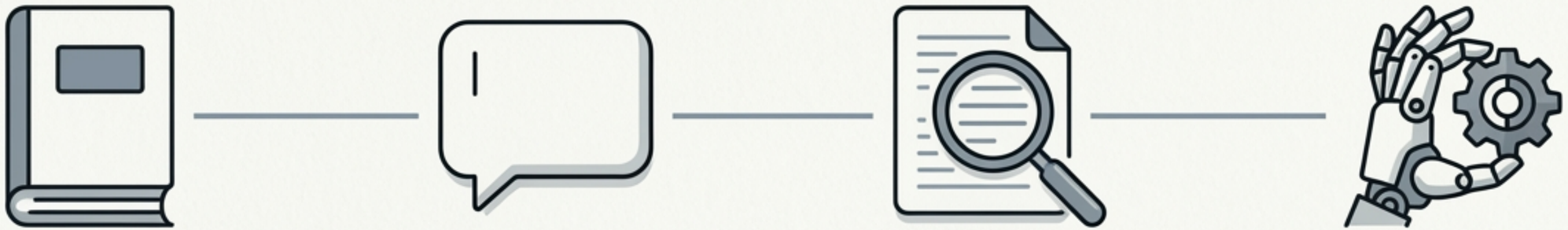
From Static Models to Agentic AI

The Evolution of LLM Capabilities: Prediction, Context, and Action

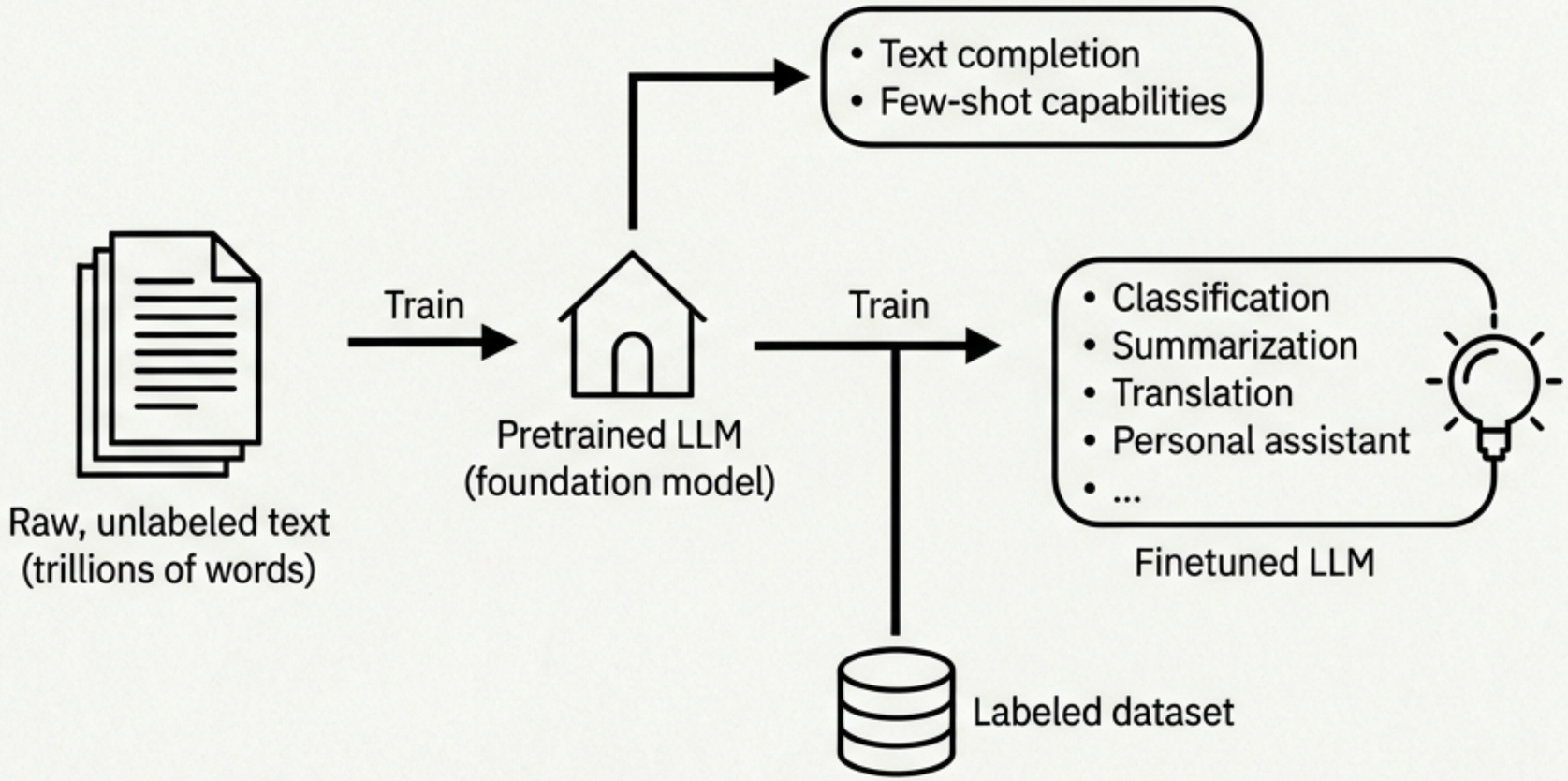


Knowledge Instruction Context Action

Phase 1: The Foundation

The Pretrained Model (The Brain in a Jar)

- Internet texts
- Books
- Wikipedia



Trained on trillions of words to predict the next token.
It lacks specific intent or personality.

Phase 1: The Foundation

The Pretrained Model (The Brain in a Jar)

Trained on trillions of words to predict the next token. It lacks specific intent or personality. in IBM Plex Serif

"You can simply think of training LLMs as training parrots to mimic human language." — Denny Zhou, *Google DeepMind*

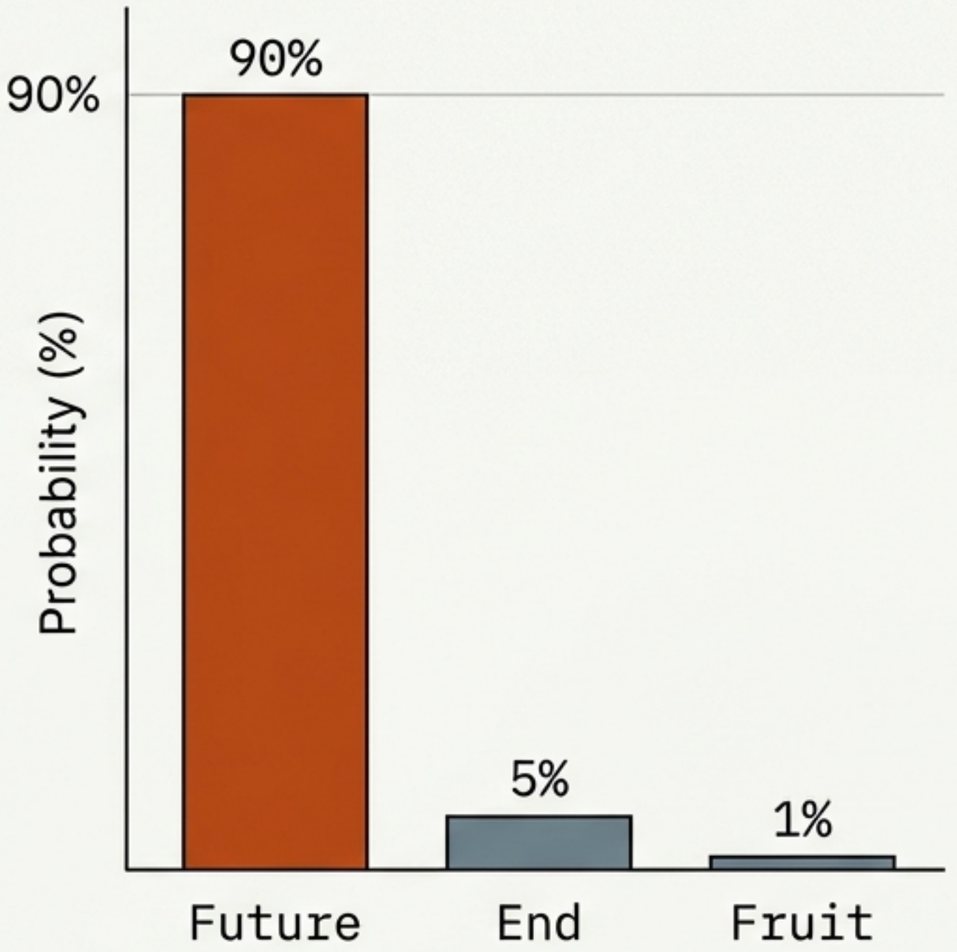
- **Mimicry vs. Intent:** The model mimics patterns perfectly but lacks inherent truth.
- **Probabilistic:** It selects words based on statistical likelihood, not understanding.

Editorial Engineering

Under the Hood: Next-Token Prediction

The ‘intelligence’ is a probability distribution over a vocabulary. The model greedily selects the highest probability token.

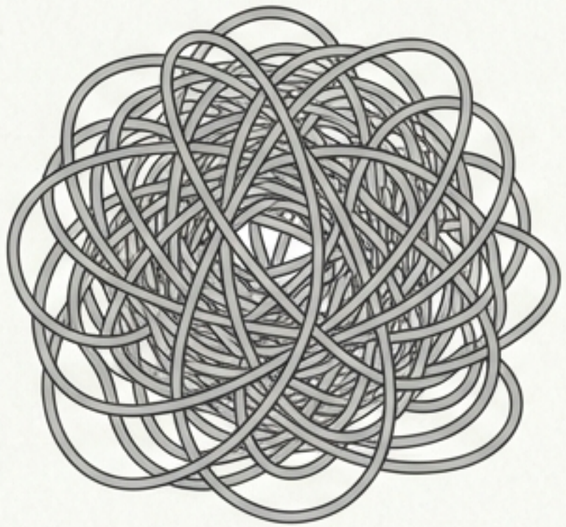
AI is the



Phase 2: Refinement

From Babbling to Obeying

Pretrained Model



Instruction Fine-Tuning

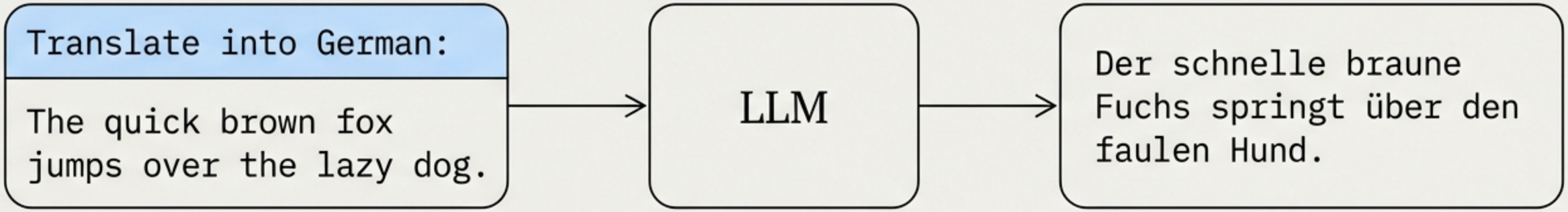
Fine-Tuned Model

```
User: "What is the capital of France?"
-> Model: "And what is the population?"
```

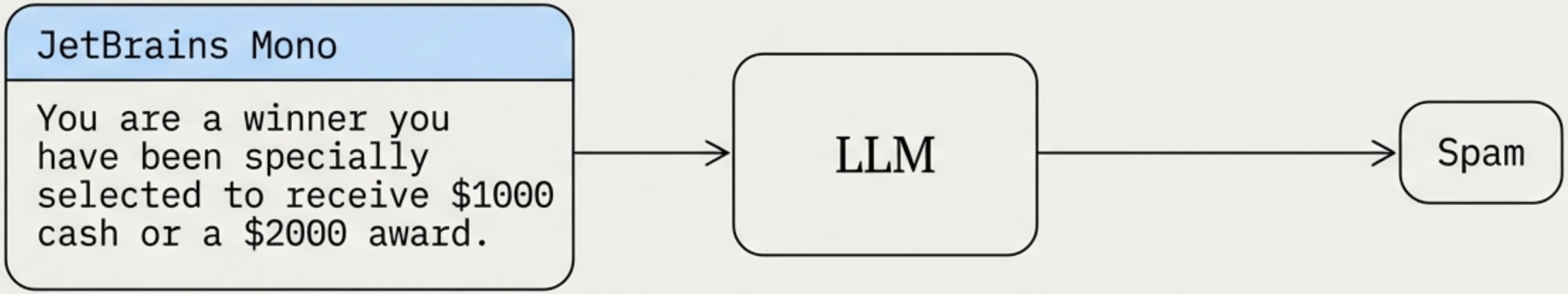
```
User: "What is the capital of France?"
-> Model: "The capital of France is Paris."
```

Two Paths of Fine-Tuning

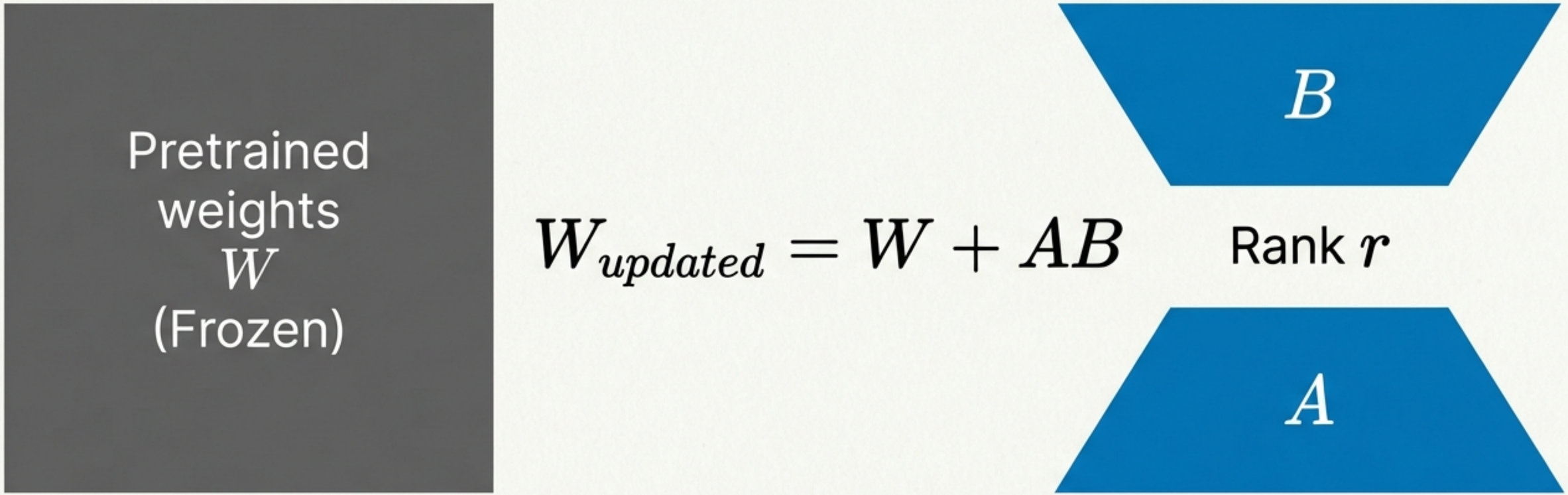
Instruction Fine-Tuning: Flexible Task Execution



Classification Fine-Tuning: Constrained Labeling



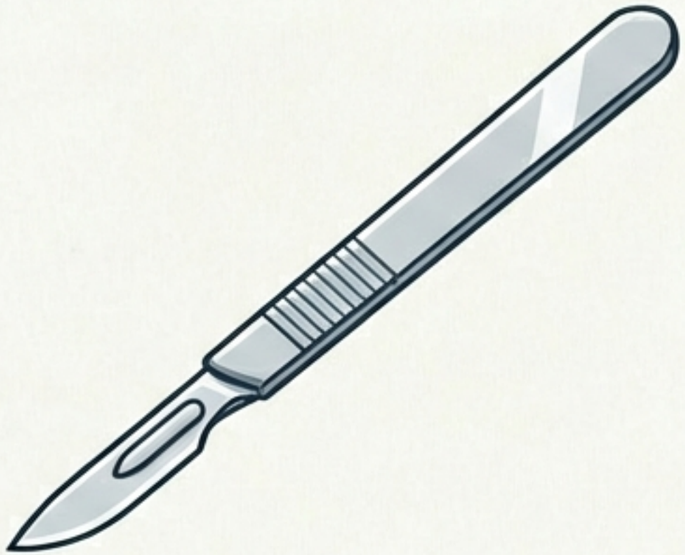
Efficiency: Low-Rank Adaptation (LoRA)



Instead of retraining billions of parameters, LoRA injects tiny trainable matrices, reducing trainable parameters by up to 98%.

Generalist vs. Specialist

Pretrained Model
(Broad, Shallow Knowledge)



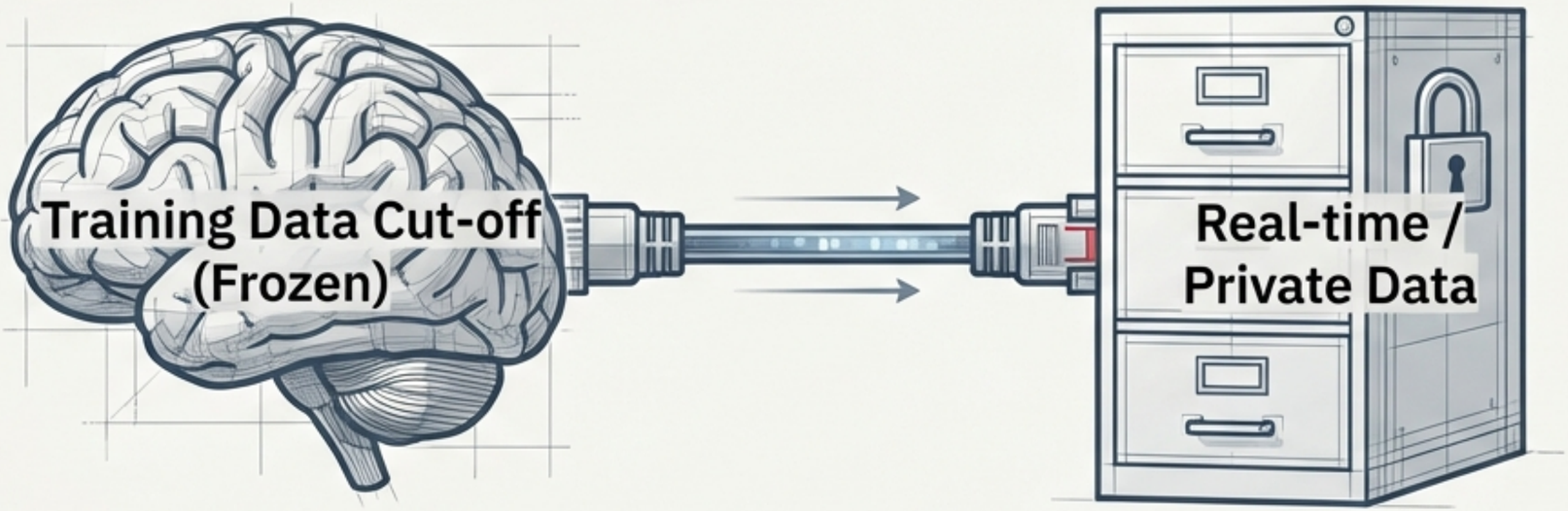
Fine-Tuned Model
(Narrow, Deep Expertise)

Risk: Catastrophic Forgetting. Heavy fine-tuning can cause the model to lose general knowledge capabilities.

Phase 3: Expansion

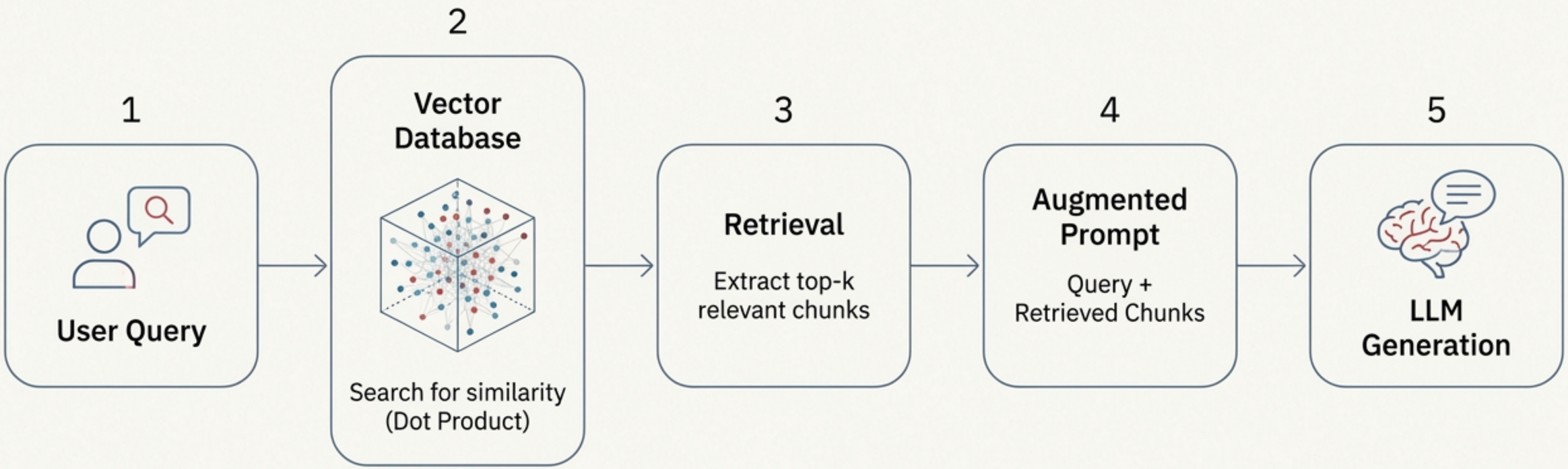
Retrieval Augmented Generation (RAG)

The Library Card

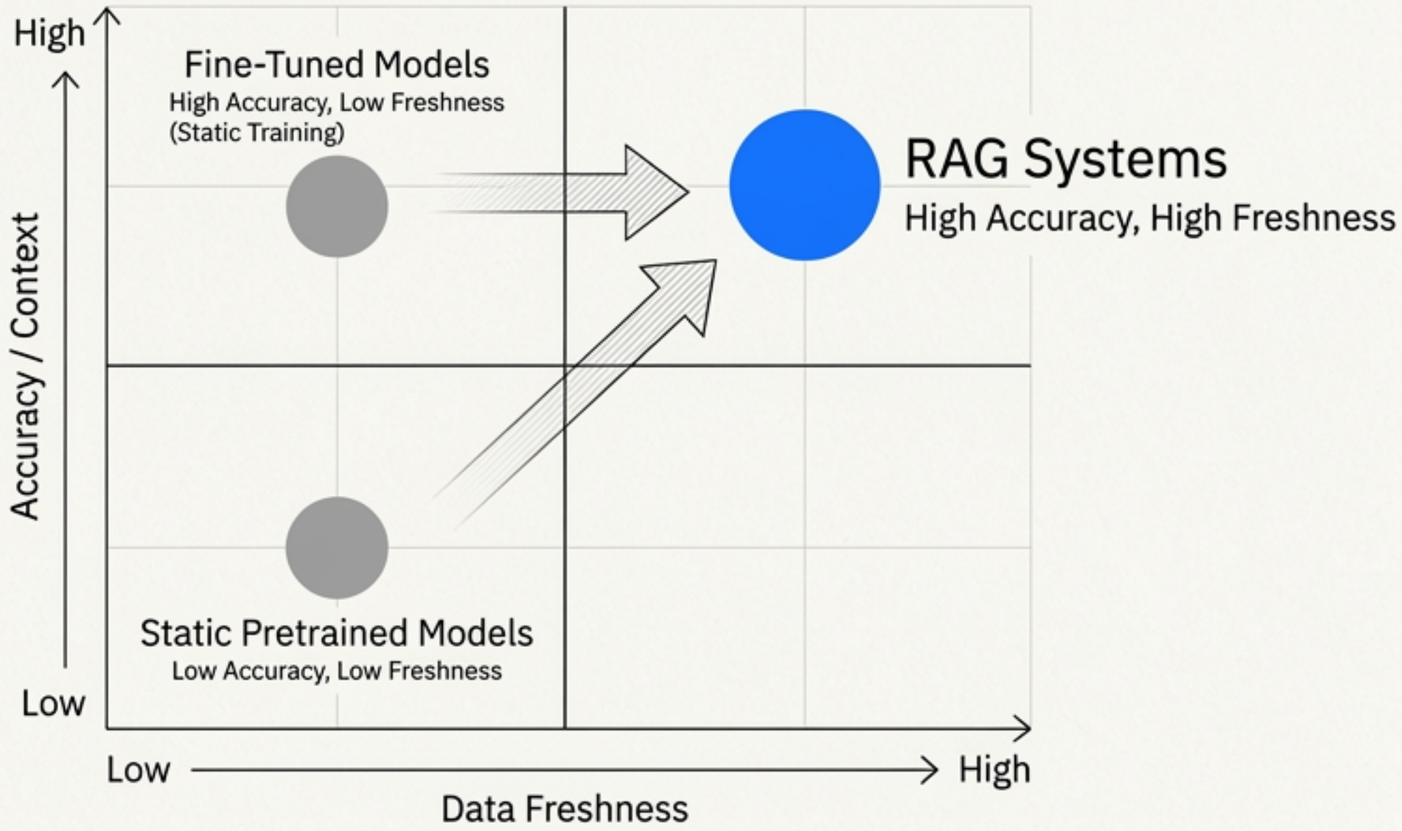


RAG solves the “Frozen Knowledge” problem by retrieving relevant information at runtime and injecting it into the context window.

RAG Architecture: The Open Book Exam



Why RAG is the Industry Standard

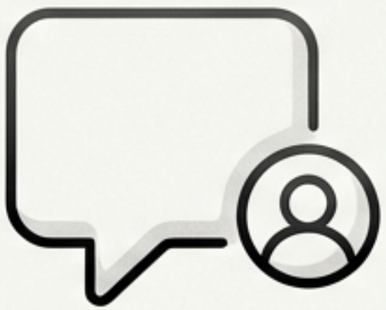


RAG enables access to real-time, private data without the massive cost of retraining.

Editorial Engineering

Phase 4: Agency

The Shift to Action



Phases 1-3:
Generation
(Text, Code, Images)

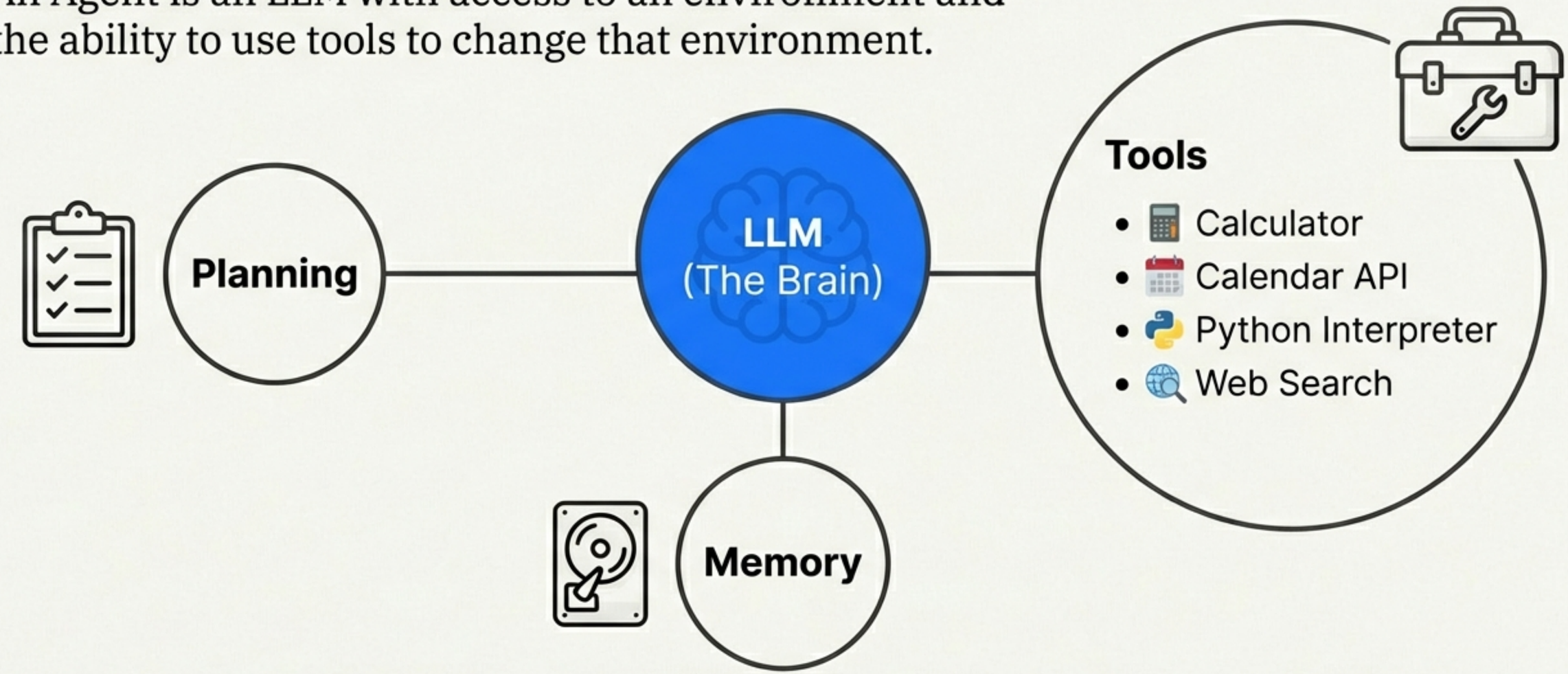
The Missing Link is Action.
Moving from “What is the answer?” to “Go do this for me.”

Phase 4: Execution
(API Calls, File Ops, Transactions)

Editorial Engineering

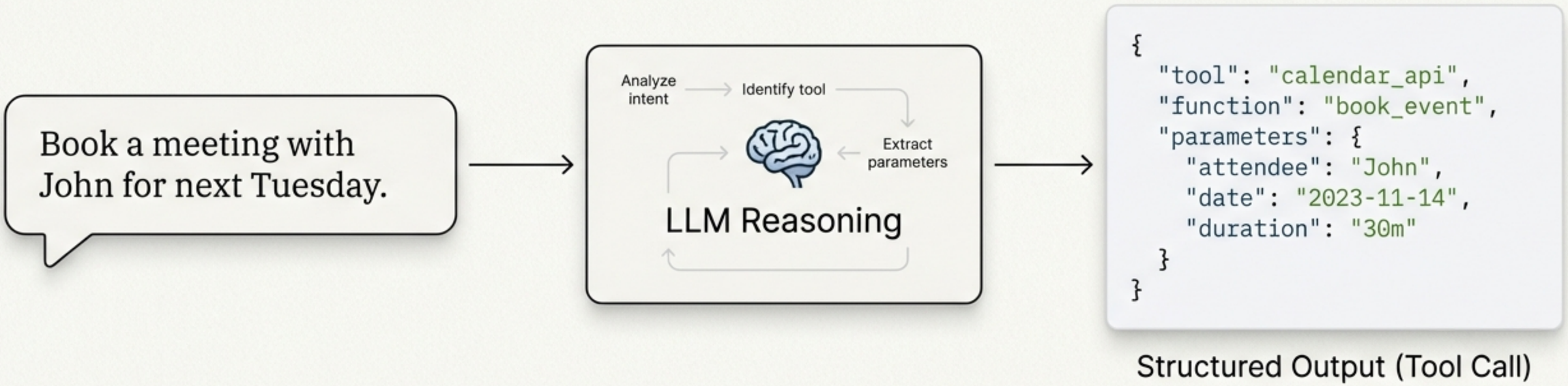
Anatomy of an Agent

An Agent is an LLM with access to an environment and the ability to use tools to change that environment.



Editorial Engineering

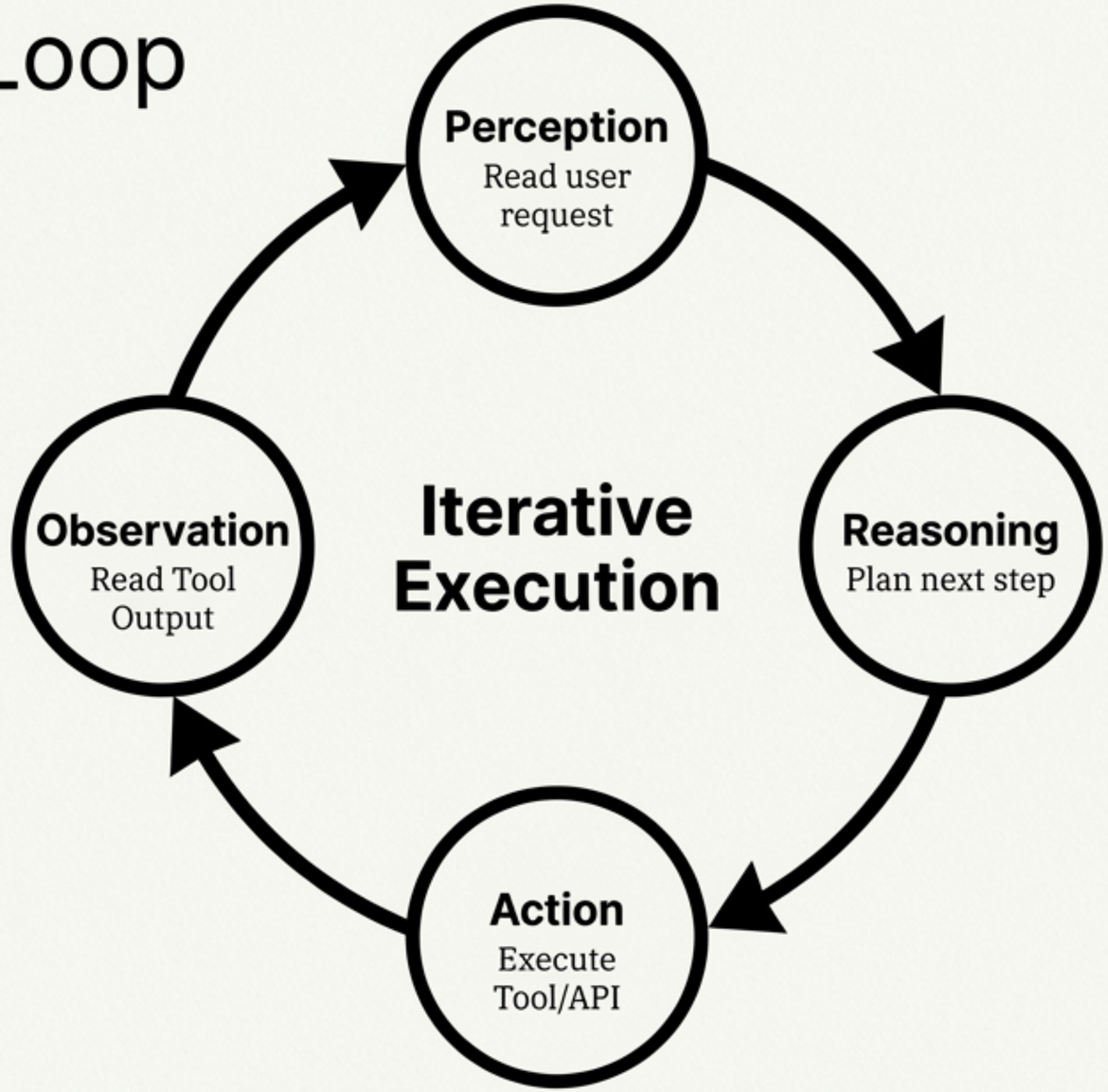
The Protocol: Connecting Brains to Tools



The model is fine-tuned to output structured data (JSON) when it detects a task, allowing it to interface with software APIs.

Editorial Engineering

The Agentic Loop



Unlike a linear chatbot, an agent loops through reasoning and acting until the task is complete.

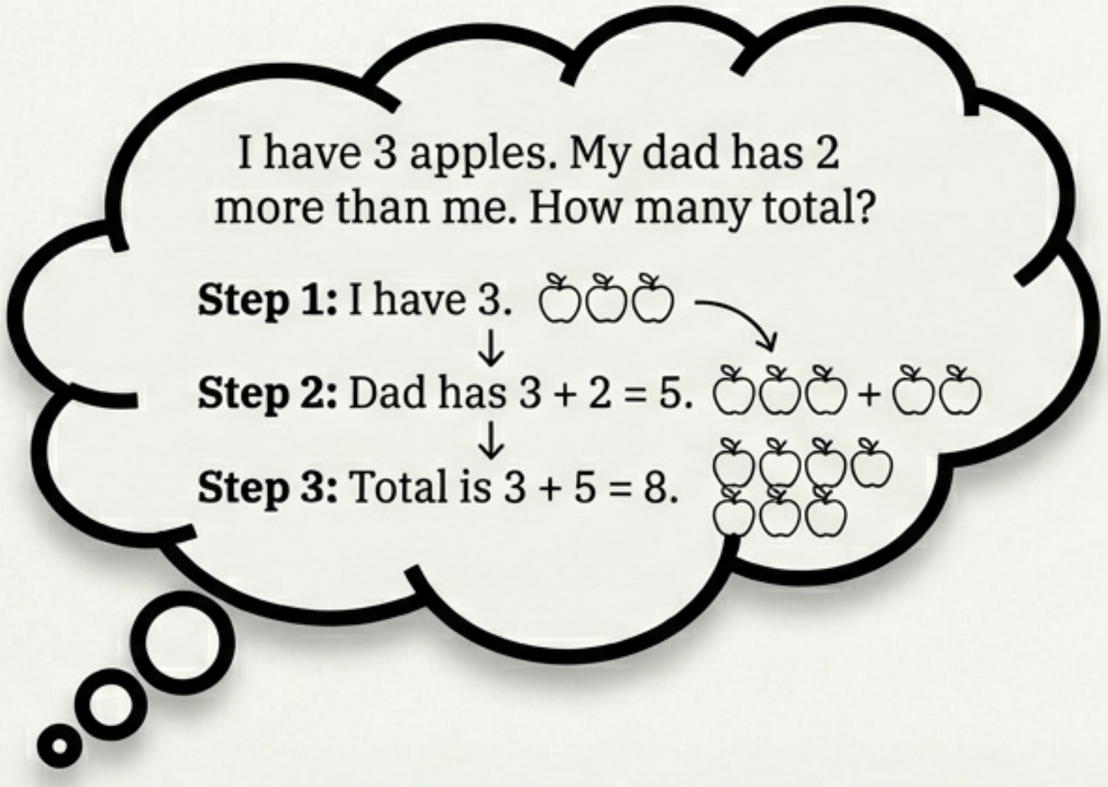
Editorial Engineering

Practical Example: Book a Flight

The timeline shows how RAG (reading) supports Reasoning (thinking) which leads to Action (doing).

Editorial Engineering

Advanced Capabilities: Chain of Thought

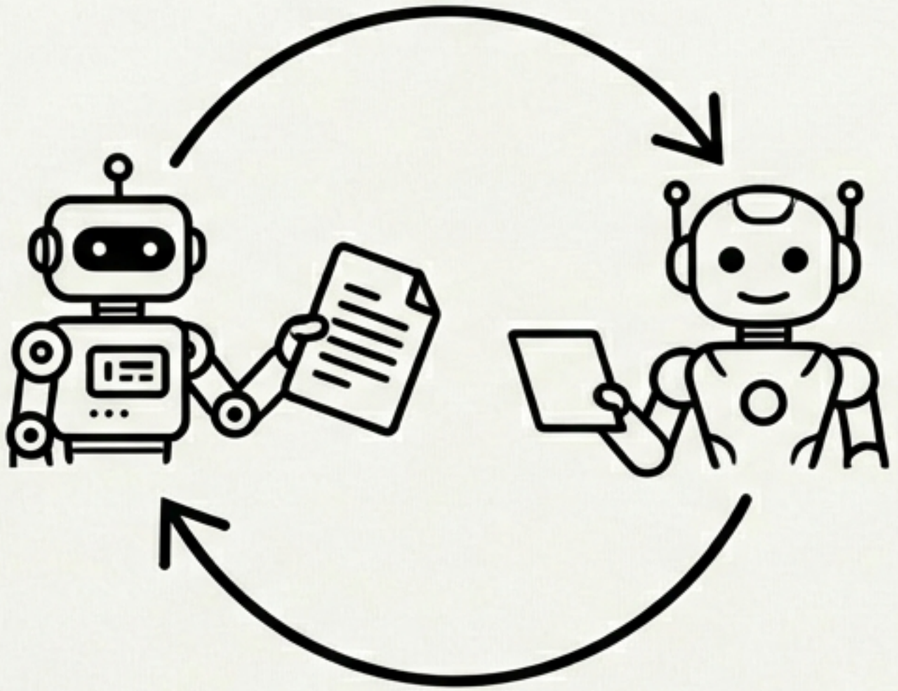


“Intermediate steps matter more than fine-tuning. Breaking complex tasks into sub-steps significantly increases reliability.” — Denny Zhou

Editorial Engineering

Multi-Agent Systems

Coder Agent
Drafts Python script.

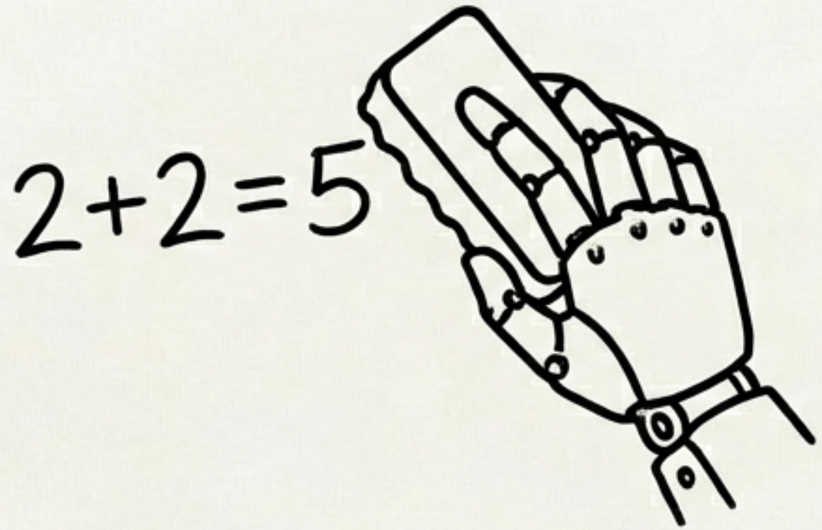


Reviewer Agent
Checks for bugs.

Specialization mimics human teams. Instead of one generalist trying to do everything, specialized agents collaborate to verify and improve output.

Editorial Engineering

The Future: Self-Correction



Current Limitation: LLMs struggle to realize they are wrong in real-time.
Future State: Robust ‘System 2’ thinking—the ability to pause, verify, and correct logic before taking action.

“LLMs cannot self-correct reasoning yet.” — Denny Zhou.

The Path Forward

Knowledge + Reasoning + Action = The Agentic Future