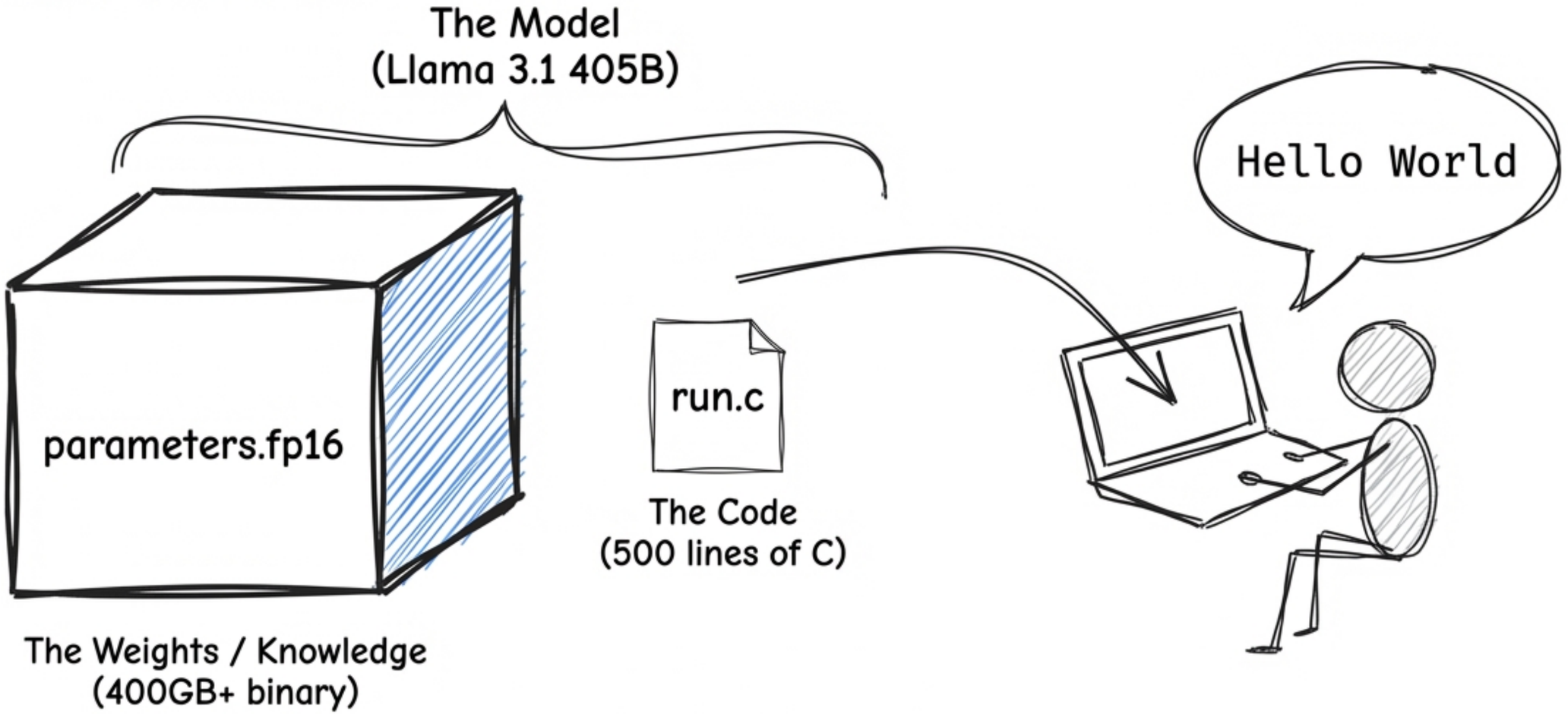
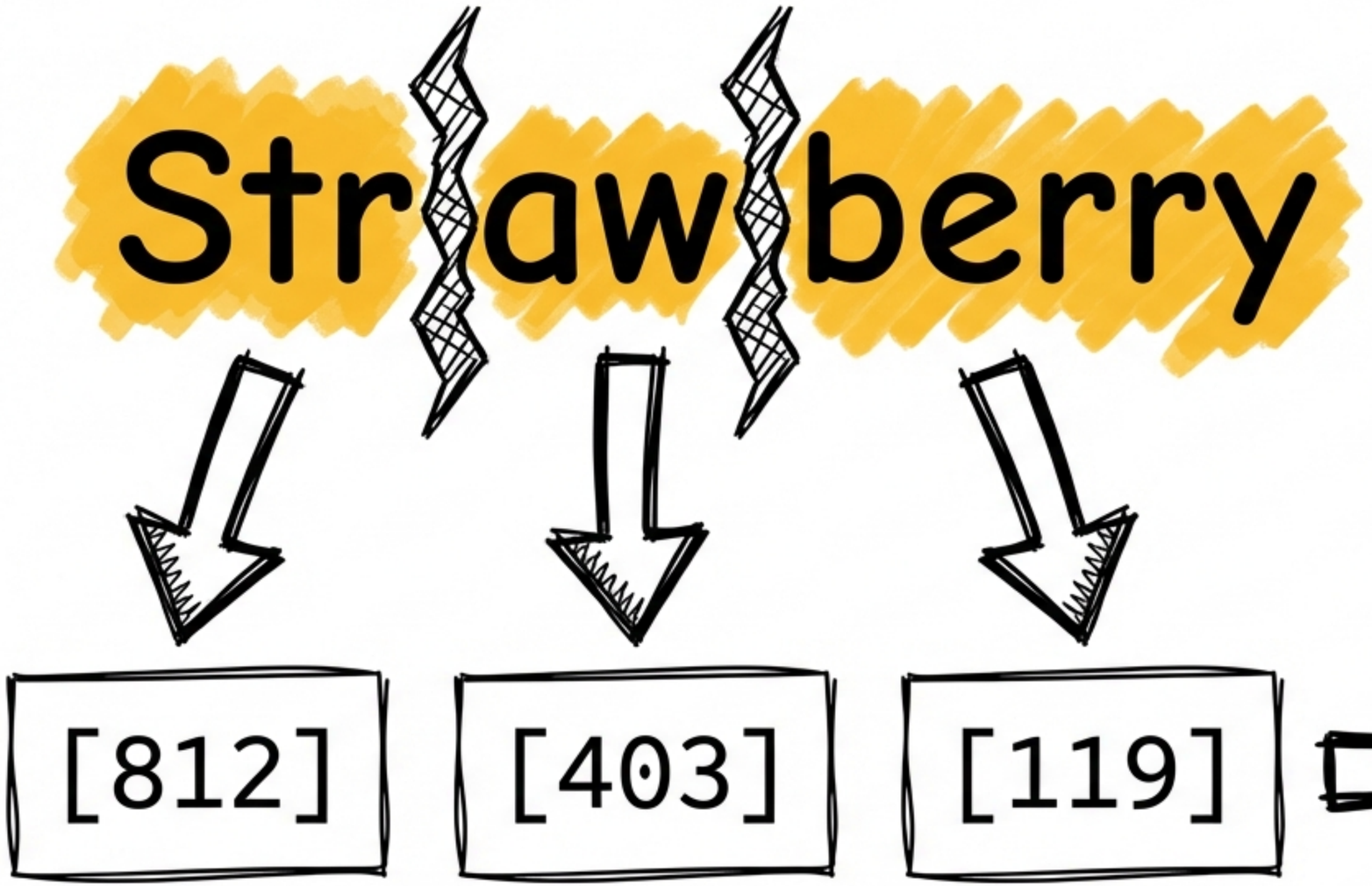
The Mental Model:
An LLM is Just Two Files



There is no "brain" inside.
just a massive file of compressed
numbers and a tiny algorithm.

Patrick Hand

The Atom of Language: Tokenization



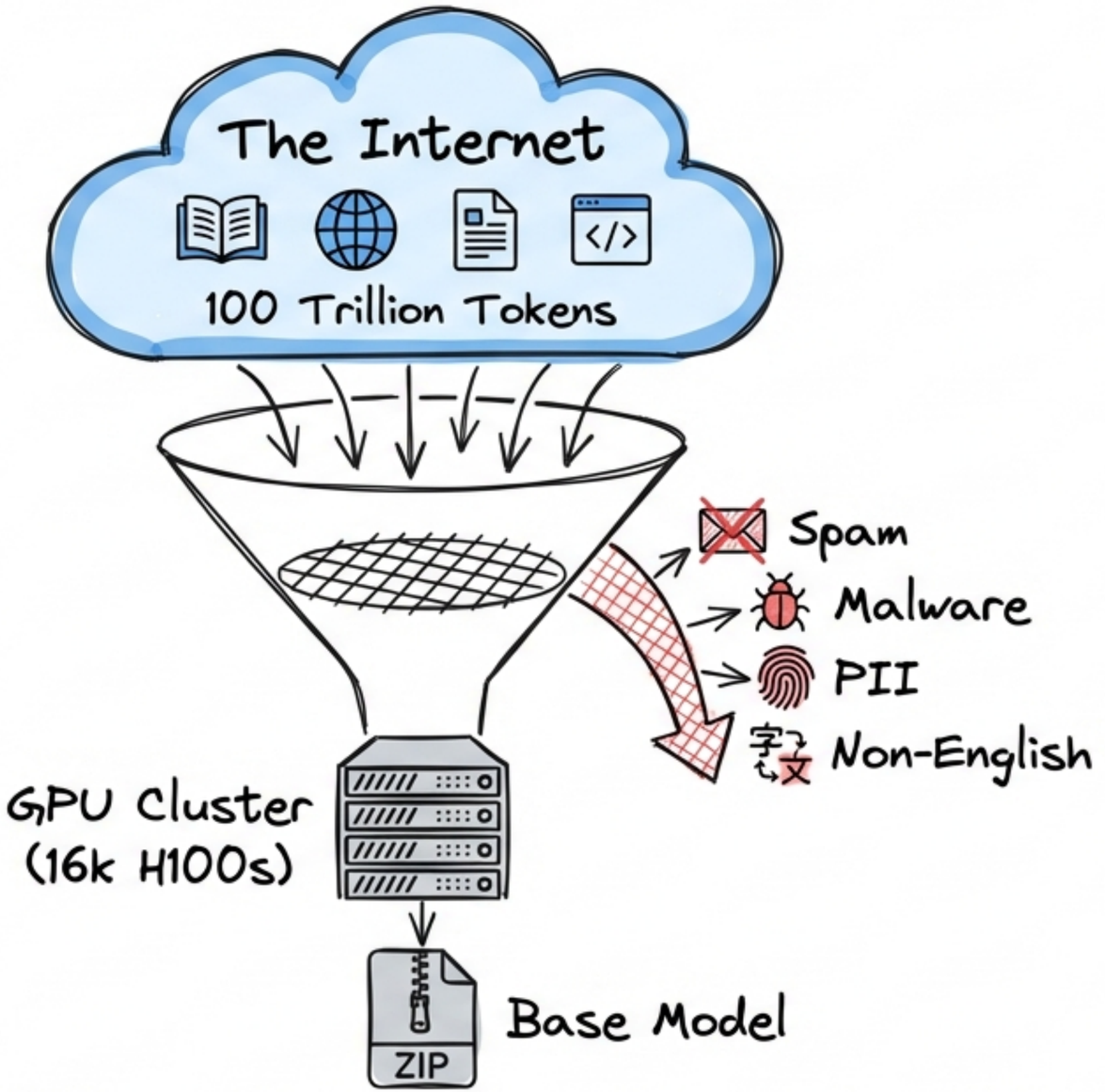
Models don't see letters (S-t-r-a-w...).

They see a sequence of Integers.

This is why they struggle to count the 'r's in Strawberry.

Vocabulary size: ~100k distinct tokens.

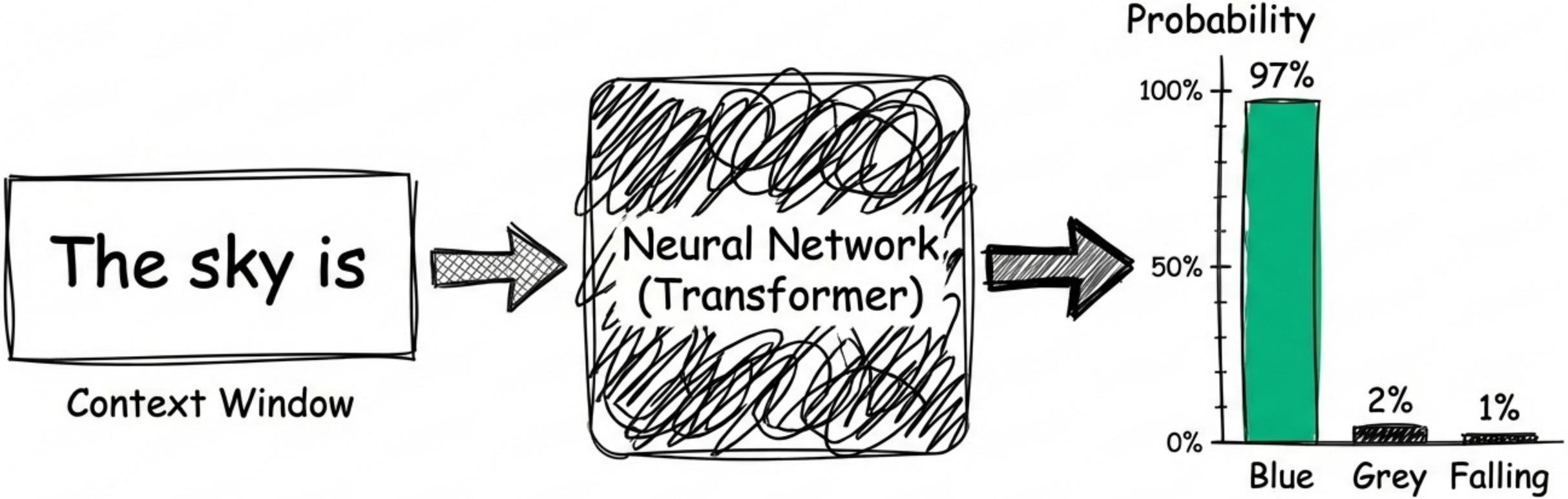
Stage 1: Pre-training (Compressing the Internet)



Result: A "Lossy Compression" of human knowledge.

Patrick Hand

The Mechanism: Next Token Prediction



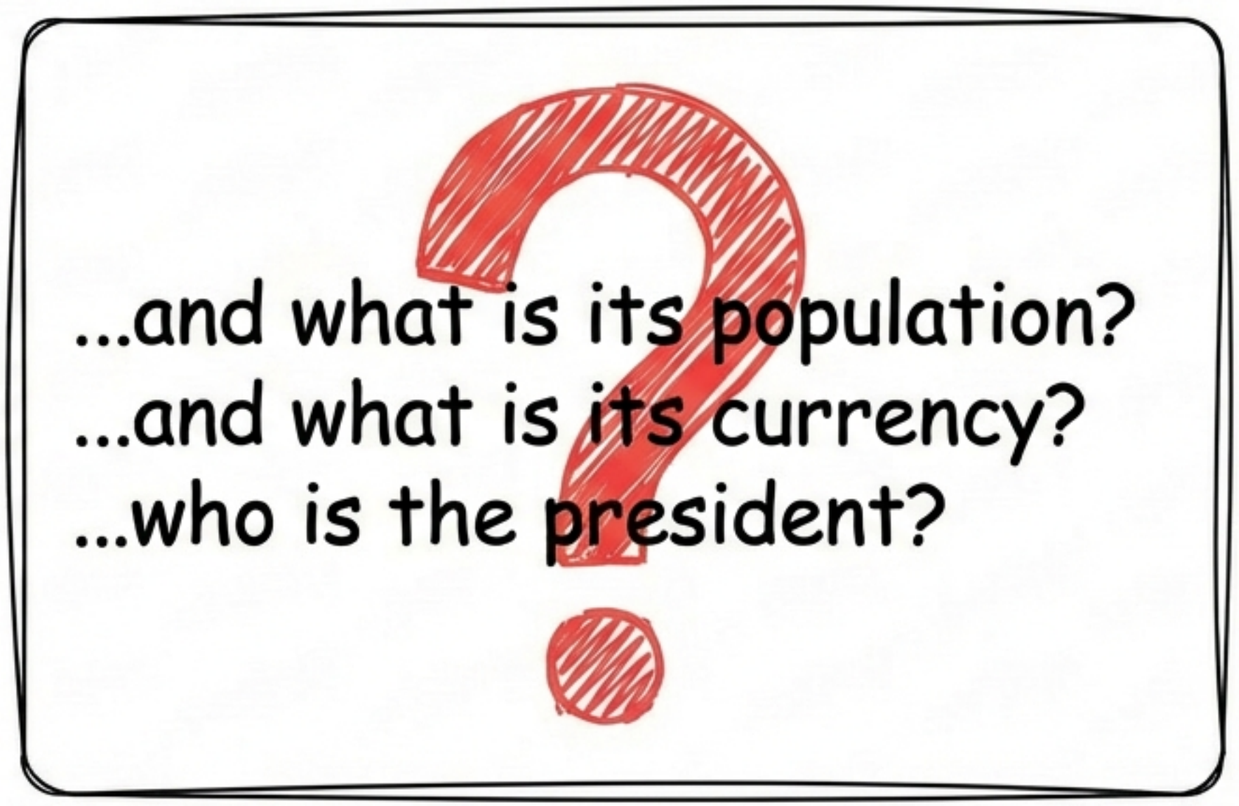
The model calculates the probability of every possible next word, then flips a coin to choose one.

Patrick Hand

The Base Model: A Document Simulator

What is the capital of France?

User Prompt

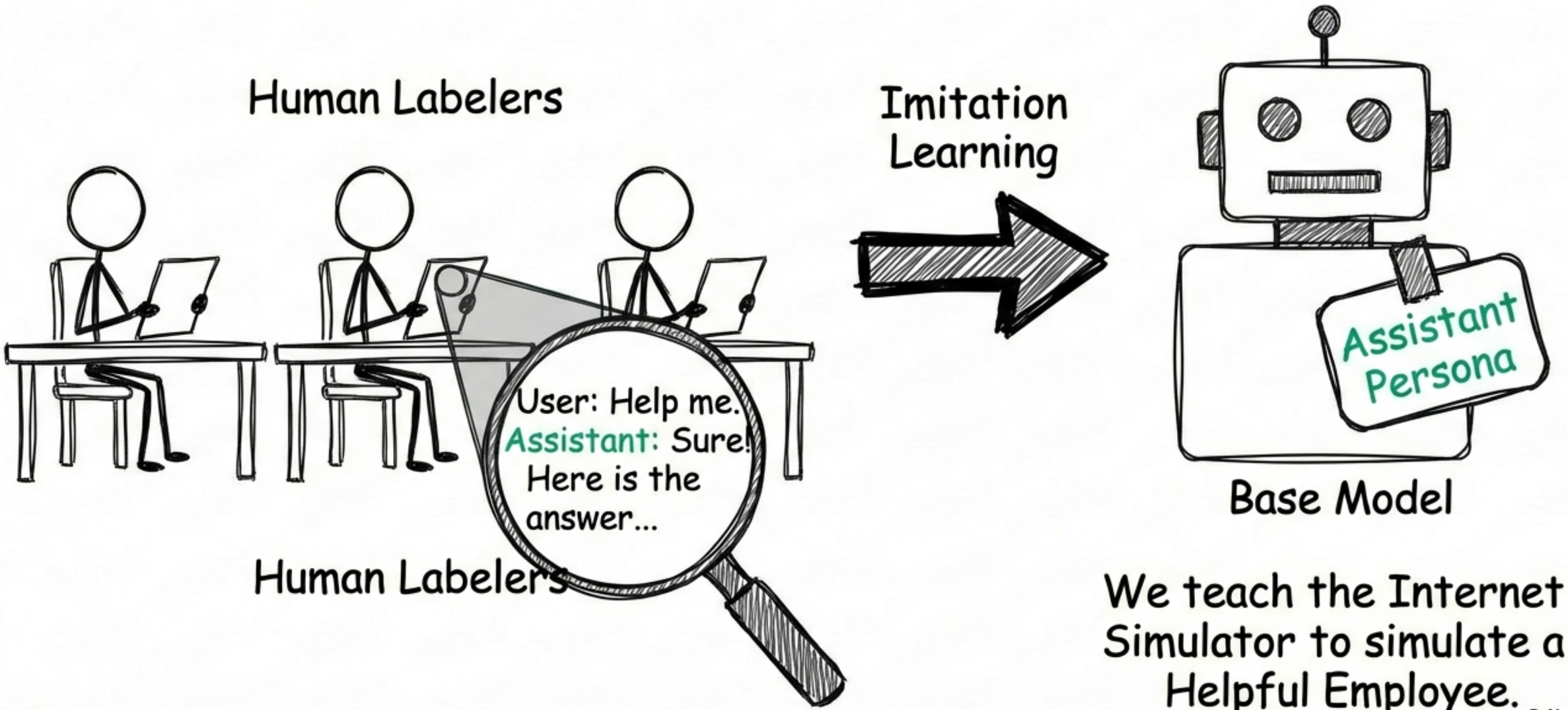


The Base Model Response

It is not an assistant yet. It just wants to complete the pattern of a "quiz document".

Patrick Hand

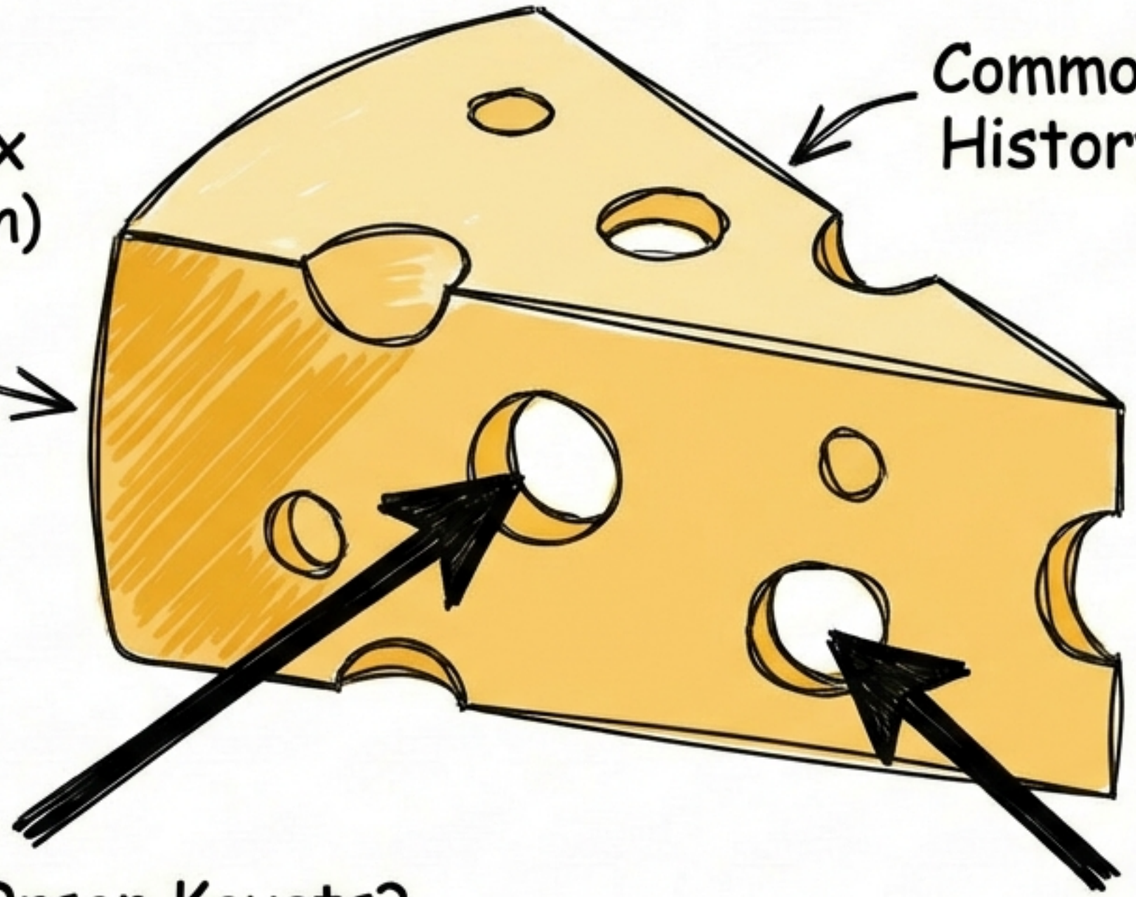
Stage II: Post-training (Supervised Fine-Tuning)



We teach the Internet Simulator to simulate a Helpful Employee.

Patrick Hand

Hallucinations: The Swiss Cheese Model

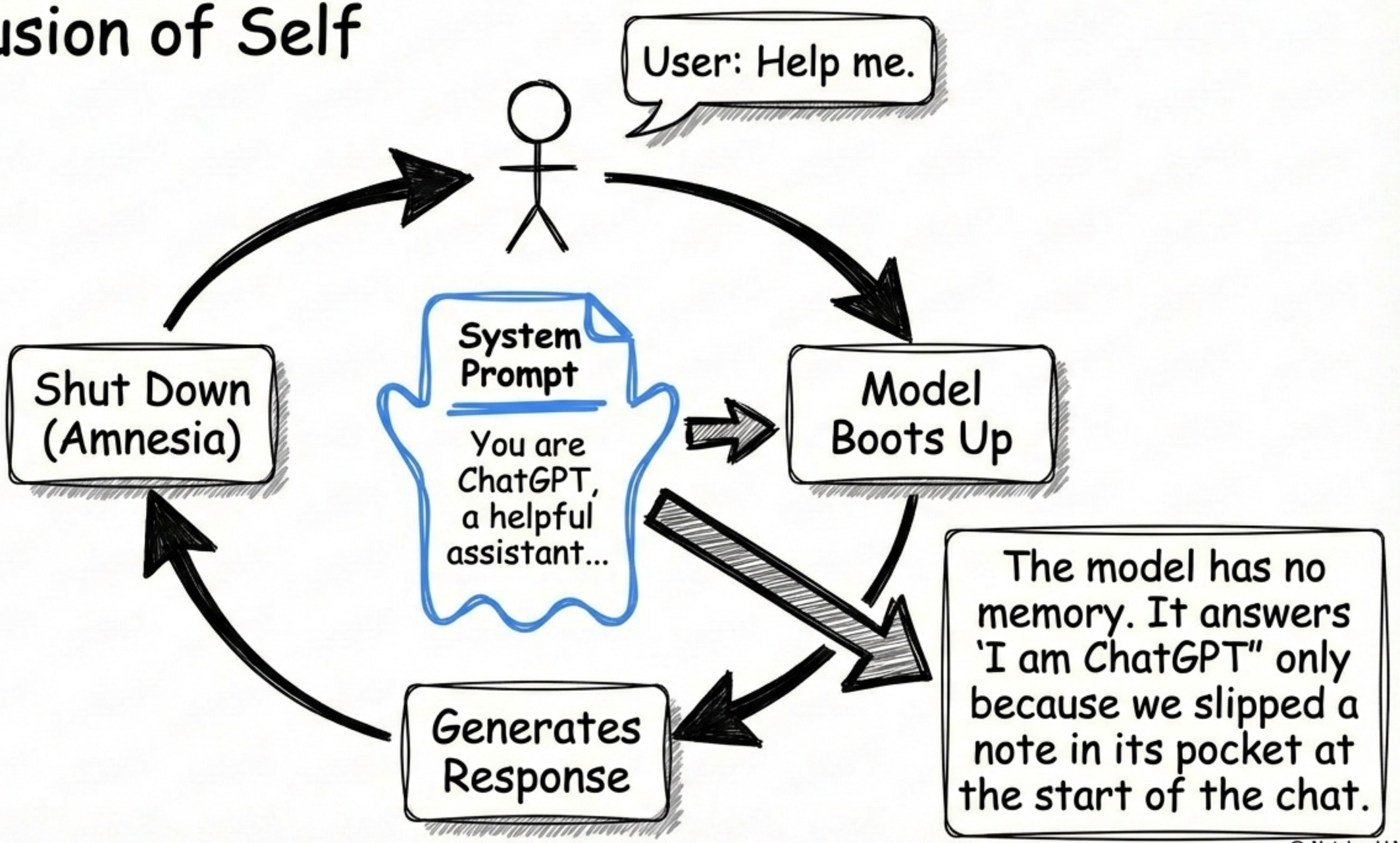


Knowledge is brittle.

If the prompt hits a 'hole' (weak probability), the model fills it with made-up patterns.

Patrick Hand

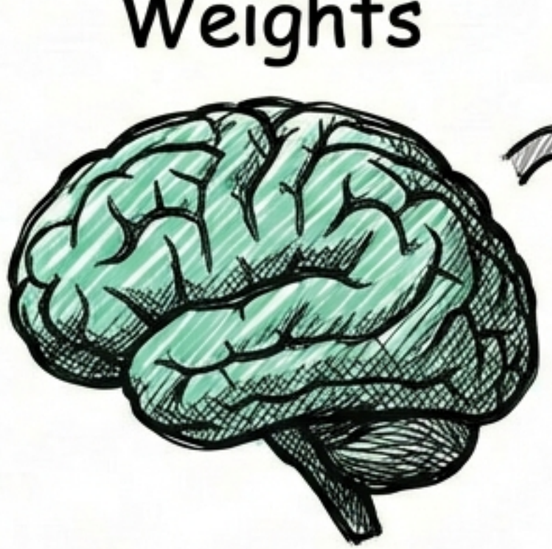
The Illusion of Self



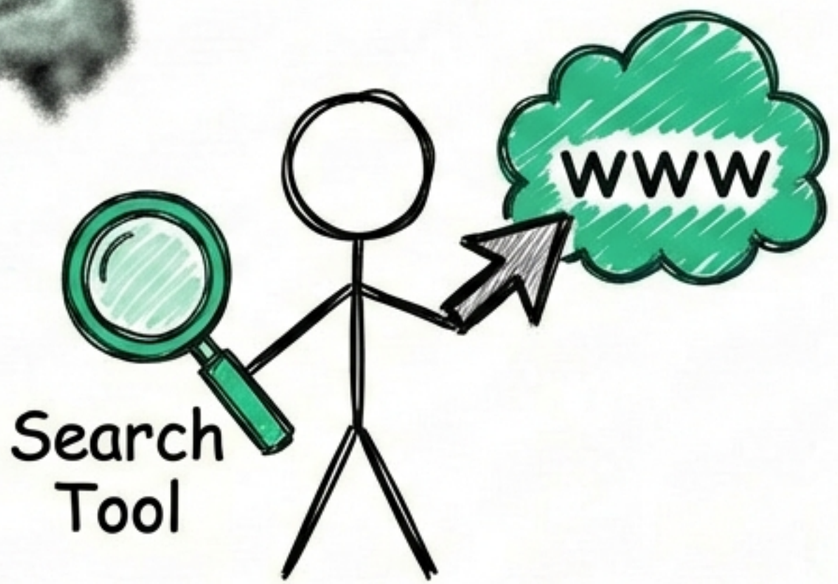
Patrick Hand

The Fix: Tool Use (Weights vs. Context)

Weights



Kalam
Long-term Memory
(Fuzzy, Hard to change).



The Model

Context Window

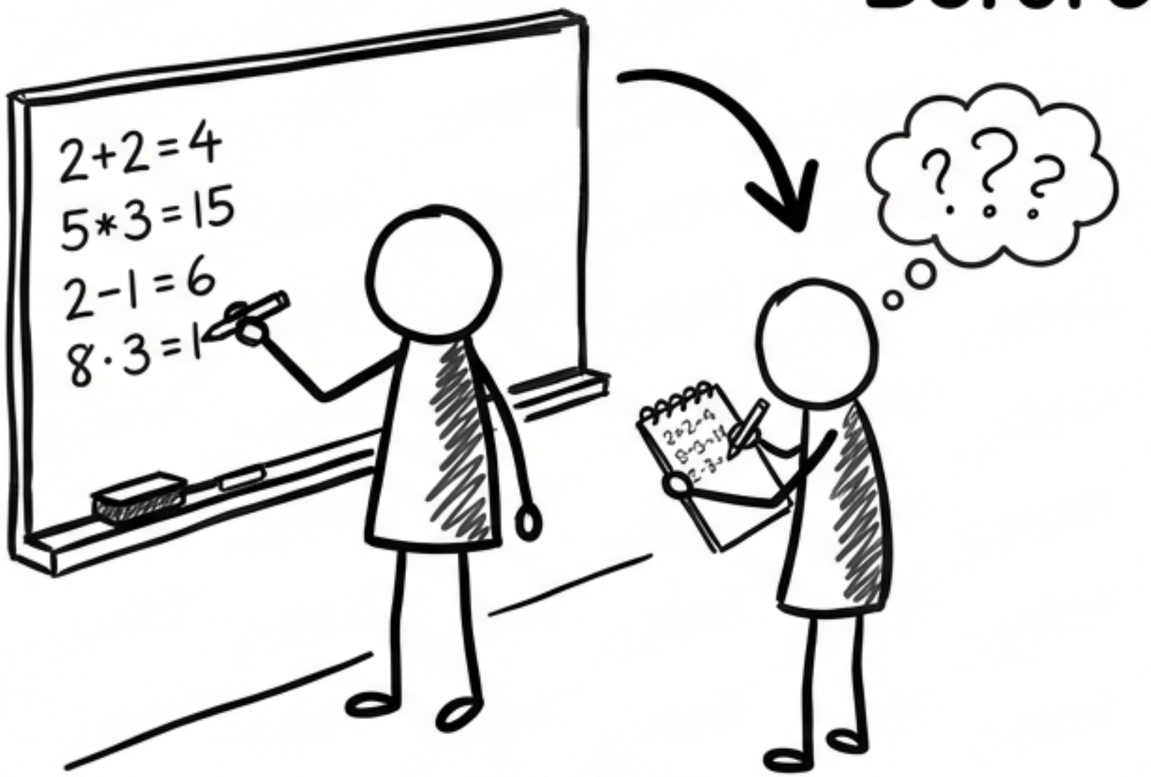
Working Memory
(Perfect recall, Temporary).

Don't trust the Brain. Trust the Notepad.

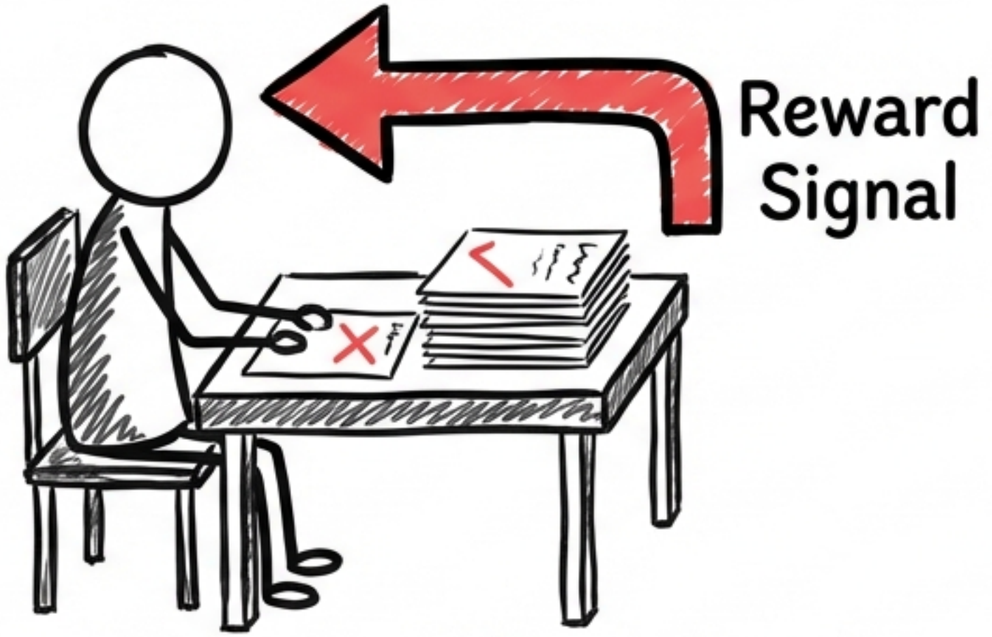
Patrick Hand

Stage III: Reinforcement Learning (RL)

Before vs After



Imitation

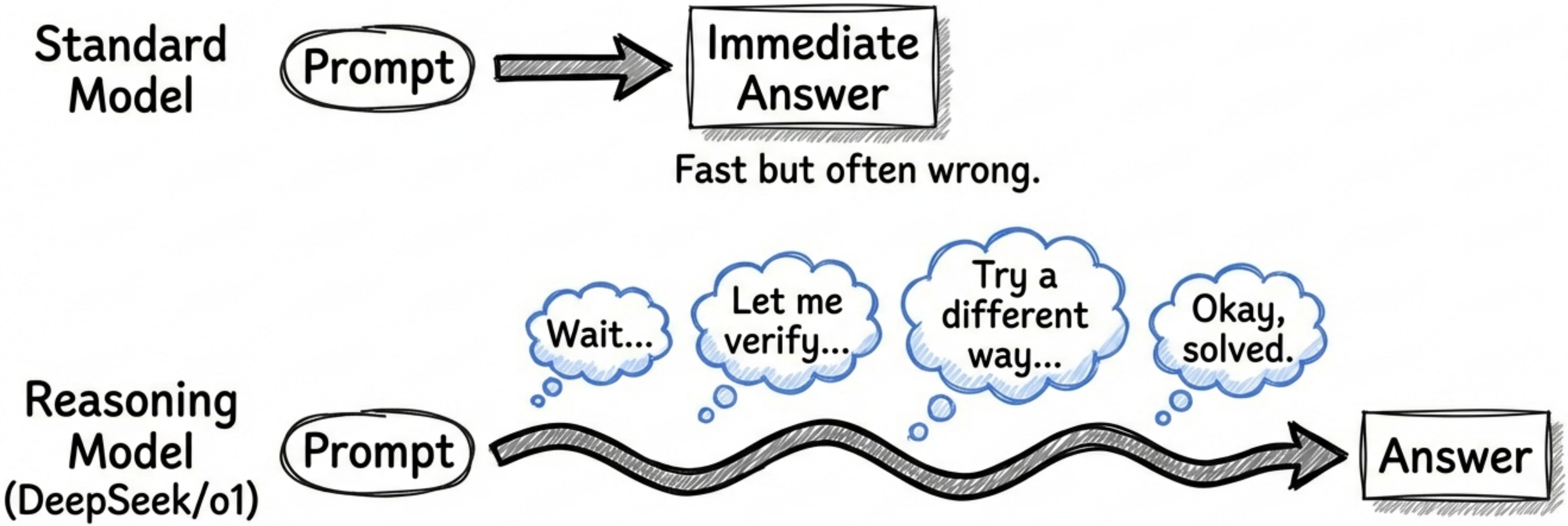


Practice (Trial & Error)

Humans can’t write every answer. The model must try, fail, and learn from the grade.

Patrick Hand

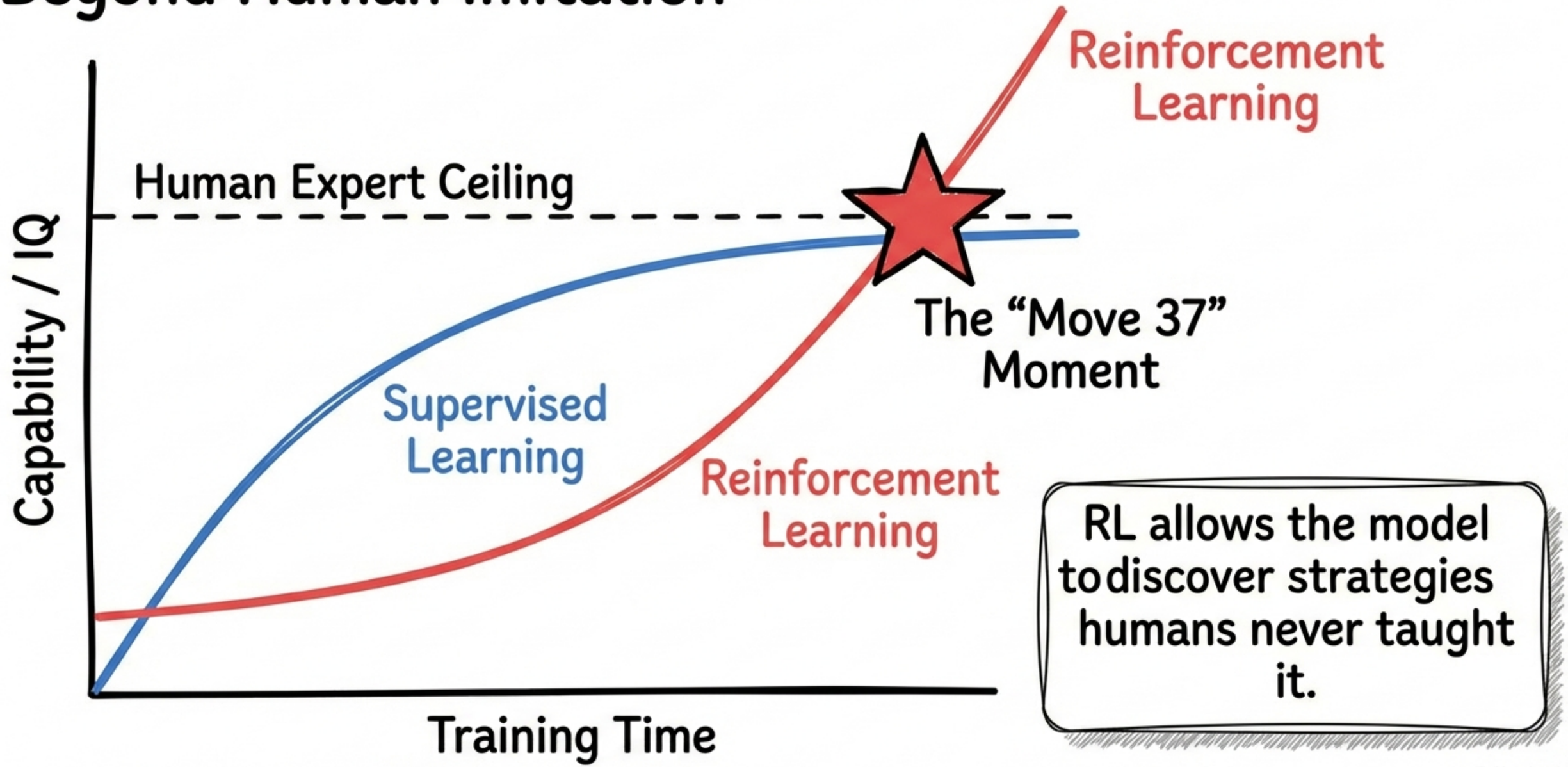
The “Aha!” Moment: Thinking Models



“Models need tokens to think.” We trade time for accuracy.

Patrick Hand

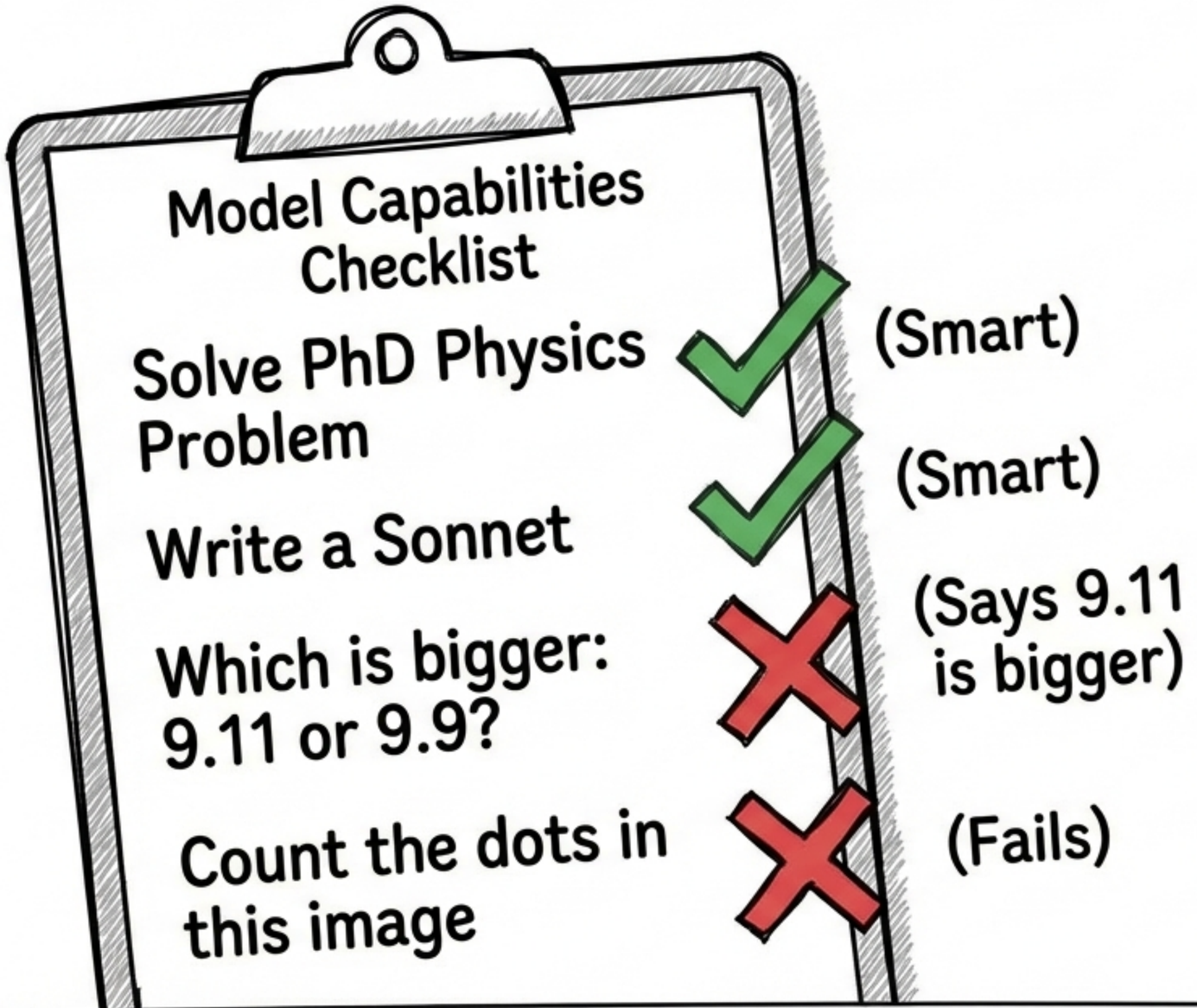
Beyond Human Imitation



RL allows the model to discover strategies humans never taught it.

Patrick Hand

Beware the Jagged Edges



The model is not a logical calculator. It struggles with counting, spelling, and simple comparisons because of Tokenization.

The Future: Agents & Multimodality

Chatbot
(Text In/Out)

Computer Use / Agents

Moving from “Chatting” to “Doing”.

Patrick Hand

The Landscape & Strategy

Proprietary (Closed) OpenAI (GPT-4), Google (Gemini)	
	Open Weights (Local) Meta (Llama), DeepSeek

Practical Advice:

1. Use Thinking Models for hard reasoning.
2. Treat LLMs as ‘Smart Interns’.
3. Always Verify.