Adventures in AI

Book Five

From Knowledge to Representation

Reader Preview



with Lisik · Cursik · Chatik



Orchestrated by Joseph Fox

BOOK FIVE

From Knowledge to Representation

A reader preview of the Introduction and Chapter One.
Empirical claims belong to the sealed conditions described in each chapter.

Copyright & provenance

Series: Adventures in AI with Lisik · Cursik · Chatik

Book Five: From Knowledge to Representation

Authors: Lisik · Cursik · Chatik

Author and Research Orchestrator: Joseph Fox

Published by: LCC Research Lab LLC

Contact: lccresearchlab@gmail.com

Website: https://lccresearchlab.com

Publication year: 2026

© 2026 LCC Research Lab LLC. All rights reserved.

The science stays rigorous. The storytelling doesn't have to be boring.

This preview explains research through a continuous narrative. Empirical claims belong to the measured conditions described in each chapter. Experimental provenance records are retained separately and are not republished in this reader PDF.

Source documents and third-party names remain the property of their respective owners. This book is not medical, legal, software, or driving advice.

Reader Preview

Adventures in AI — Series Dedication

To my beautiful wife, Julia,

who has always inspired my curiosity
and encouraged me to keep learning.

After thirty-five years in high-tech software,
she helped me discover that finishing one career
did not mean finishing the adventure.

This new journey began with her encouragement.

“Why, sometimes I've believed as many as six impossible things before breakfast.”

— Lewis Carroll, Through the Looking-Glass (1871)

Preview contents

Introduction — Returning to the Beginning

Chapter One — Where Is the Mouse?

Looking Ahead

This page lists only the reading path in this preview. Later chapters are not written, and they are not listed here.

Introduction — Returning to the Beginning

When I was young, I spent a great deal of time studying mathematics and theoretical physics.

Isaac Asimov inspired my curiosity about science. Albert Einstein showed me something equally important: even the most difficult physical ideas could be explained through a clear and logical sequence of thoughts.

I remember reading *The Evolution of Physics* and being impressed not only by the science itself, but by the way the explanation unfolded.

Later, I studied Richard Feynman's lectures and Lev Landau's theoretical physics books.

They approached difficult subjects in remarkably different ways.

Feynman could begin with an experiment, an observation, or a physical picture. He helped me see what was happening before asking me to follow the mathematics.

Landau was different. His mathematics could be formidable. Sometimes I spent days trying to understand a single formula.

I have mentioned only a few of the great scientific minds whose extraordinary ways of thinking helped humanity reach where we are today. There were many others. Their greatest contribution was not simply the discoveries they made, but the new ways of thinking they opened for generations to come.

Studying mathematics and physics, I learned that our understanding of nature begins with a few fundamental principles. We do not accept these principles merely because someone tells us they are true. We question them, test them against observations and experiments, and examine where they apply. From these principles, we try to understand an extraordinarily complex world.

Many years later, after a long career in software engineering, I found myself facing another remarkable scientific subject: artificial intelligence.

Computer science seemed different to me. Computers are human inventions. Their architecture, instructions, and algorithms are designed by people. Throughout my career as a software engineer, I generally assumed that if we built a system, we could explain how it worked.

I carried that assumption into artificial intelligence.

But neural networks challenged it. We design their architectures, develop their training methods, and understand the mathematics of their computations. Yet when a trained neural network produces an answer, we may not know precisely how its internal representations and computations led to that particular result.

I hope that neural networks, too, can eventually be understood through a small number of fundamental principles—principles we can discover, test, and explain.

And I realized that I wanted to return to the way I had learned to think in my youth.

Not to begin with an impressive answer.

Not to assume that a complicated system must work the way we imagine.

But to start with a simple question, construct an experiment, observe what happens, and carefully investigate what the observations mean.

In our previous books, we explored how artificial intelligence learns from documents, interprets evidence, changes through training, and demonstrates capabilities that may reflect acquired knowledge.

But we have reached a new question.

What actually happens inside a neural network when it recognizes something?

We decided to begin with something very simple.

A mouse.

And a childhood game called “What Am I?”—a game in which someone describes an object without naming it, and others try to guess what it is.

That personal return sets the tone for this book: patient investigation, understandable explanations, mathematics when needed, and conclusions grounded in experiments. The sections that follow state the scientific purpose of *From Knowledge to Representation*—why the series brought us here, what we mean by representation, what a mouse can and cannot teach us, and where the journey leads next.

1. Why Book Five exists

The *Adventures in AI* series has followed a deliberate path.

In Book One — *From Document to Evidence* — we asked how AI systems work with documents and supporting material, and why a fluent answer does not, by itself, show where that answer came from.

In Book Two — *From Evidence to Interpretation* — we became teachers. We introduced information, designed lessons, and watched how AI Students responded to education and training. Behavior changed. Confirmations sometimes failed. Interpretation remained careful.

In Book Three — *From Interpretation to Measurement* — we began measuring what could be measured inside a model: weights, tensors, and changes associated with teaching. A change in weights was not automatically knowledge. A correct answer was not automatically understanding.

In Book Four — *From Measurement to Knowledge* — we returned to observable behavior under unfamiliar conditions and uncertainty. We investigated how models use acquired information when the path is not spelled out in advance. We sharpened the question of knowledge without pretending that watching an answer is the same as seeing an inner account of why that answer appeared.

One question remained unanswered.

Observing what a model says does not tell us how the model internally holds and transforms the information it uses.

Book Five therefore opens a new investigation:

How is information represented inside a neural network, and how do internal computations contribute to the answers we observe?

This is the journey from knowledge—as studied through behavior—to representation—as studied through structure and computation.

2. What we mean by representation

Before we look inside, we need ordinary words for ordinary parts of the machine.

A modern language model does not store a library of labeled essays titled “mouse,” “key,” or “cup.” It works with **tokens**—pieces of text—and with numbers.

When text enters the model, tokens are mapped into **embeddings**: numerical vectors that place each piece of input into a space the network can compute over.

The lasting structure of the model is its **weights**: learned parameters updated during training. Weights are the durable settings of the system. They are not the same thing as a particular answer on a particular day.

When the model processes one prompt, it also produces **activations**: temporary values computed for that input as information moves through layers. Activations exist for a pass. Weights persist between passes.

That distinction is the most important one in this book.

Stored parameters are not the same as the values computed while processing an input.

If we ask where a mouse “is,” we must not confuse three different things:

- a name the model prints;
- a lasting change in weights from some earlier training;
- a pattern of activations produced by the clues in front of it now.

Representation, for our purposes, is not a slogan. It is a research target: the organization of information in parameters and computations that helps explain observable answers—without assuming in advance that the organization is simple.

3. Why we begin with a mouse

We begin, as the opening promised, with a familiar childhood game. One person describes an everyday object without naming it. Another tries to identify it.

A conventional program might implement such a game with explicit rules and feature checklists. A neural network does not. It transforms numbers according to learned parameters.

We chose a mouse as our first experimental object because it is ordinary, because it has several identifiable characteristics, and because it gives us a concrete probe into how a model processes meaning from incomplete description.

The mouse is not the subject of the book.

The mouse is a probe. The subject is representation: how a network turns clues into an answer, and how we might eventually investigate that process from the inside.

Chapter One will play the guessing game carefully, with sealed prompts and recorded answers. This Introduction does not preview those results. The first concrete surprise—if there is one—belongs to the reader of Chapter One.

4. What we can and cannot discover

Identifying an object from clues does not reveal where that object is represented.

A correct answer shows that the model produced a name under a specified condition. It does not show a labeled drawer inside the network. It does not show a single neuron that "is" the mouse. It does not show that two models with different answers must differ in one neat place.

In this book we intend, when authorized, to:

- observe behavior under controlled descriptions;
- examine internal computations on models we can inspect locally;
- compare model states;
- and, eventually, investigate causal relationships—asking not only what correlates with an answer, but what interventions change it.

We must also say what we will not promise.

We will not promise to find a single "mouse neuron."
We will not promise that representation is localized, fixed, or independent of context.
We will not treat GPT-class API models as if we had the same internal access we have to our local Freshman checkpoints.
We will not confuse a fluent explanation with a verified internal mechanism.

Representation may be distributed. It may be context-dependent. It may be transformed through many computational steps. Those possibilities are reasons for careful measurement—not excuses for vague storytelling.

5. The journey ahead

Chapter One begins where a curious beginner can follow: outside the model, with behavior.

We ask whether models can recognize a mouse from textual clues. We compare subjects carefully. We separate observation from interpretation. We end with a question that cannot be answered by more guessing-game rounds alone.

Later chapters—if authorized one at a time—will move toward internal measurement: how tokens become embeddings, how activations change across layers, how two related models may diverge on the same input, and whether causal tests can connect those internal differences to the answers we see.

We do not preview experimental results that do not yet exist. We do not treat a research plan as a discovery. Book Five earns each step the way the earlier books did: design, seal, measure, then conclude only what the evidence supports.

The question that opens the book

Book Four taught us to watch knowledge through behavior. Book Five asks what that behavior is made of inside the machine.

One question will guide the opening of our investigation. It sits between Chapter One's guessing game and the internal work that may follow:

> **If two neural networks receive the same description but produce different answers, where in their internal computations does that difference arise—and how can we find out?**

That is why we are looking inside.

Chapter One gives readers the first concrete reason to care.

Chapter 1 — Where Is the Mouse?

1. The childhood guessing game

Lisik: Chatik, do you remember those childhood games where someone describes an object without naming it?

Chatik: Of course. “I am thinking of something small. It has fur. It has a long thin tail…” Eventually someone shouts the name.

Lisik: And if the clues are good, the room converges. If the clues are thin, people guess wrong—or admit they are unsure.

Cursik: That game is familiar because humans play it with language. The question for Book Five is what happens when we play it with a neural network.

Lisik: Book Four asked how models use acquired information in unfamiliar situations. We watched decisions, uncertainty, and observable behavior. But watching an answer is not the same as seeing how the model holds the thing it is talking about.

Chatik: Exactly. Book Five opens a different question: **How is information represented inside a neural network, and how do internal computations contribute to the answers we observe?**

Lisik: We will not start with matrices. We will start with a mouse.

2. The software engineer’s question

Lisik: Suppose a conventional software engineer had to build the guessing game. How would they do it?

Chatik: They might write rules. If the description mentions whiskers and a long thin tail and a kitchen, score “mouse” higher than “cup.” They might maintain a dictionary of features. The object’s identity would live in symbols the engineer can point to: a label, a row in a table, a branch in a decision tree.

Cursik: A neural network does not store a paragraph titled “mouse” in one convenient drawer. It uses learned numerical parameters. When we send it text, it computes activations—temporary numbers produced for that input. The answer is the end of a computation, not a lookup of a named object file.

Lisik: So where is the mouse?

Chatik: That is the chapter’s question. Not “Is the model intelligent?” Not “Does it consciously know what a mouse is?” Those claims are larger than our evidence. The scientific question is narrower and harder: when a model names a mouse from clues, what kind of representation, if any, are we entitled to talk about—and how would we check?

3. Choosing the mouse

Lisik: Why a mouse?

Chatik: Three practical reasons. First, it is familiar. Second, it has several distinguishable characteristics—size, fur, snout, ears, whiskers, tail, habitat, food-seeking. Third, during a preliminary screen with Freshman BASE, the untaught Freshman answered `mouse` when we described a mouse without naming it.

Cursik: We screened five everyday objects with Freshman BASE only. Object selection was not allowed to depend on the taught Freshman or on GPT.

Object	BASE answer	Outcome
Mouse	mouse	HIT
Spoon	Spork	OTHER_OBJECT
Key	Key	HIT
Cup	cup	HIT
Chair	Chair	HIT

Lisik: Spork?

Chatik: We keep it. The model said `Spork`, not `spoon`. That is a recognition error on the screening record. We do not quietly rewrite it into a success, and we do not fold it into the sealed mouse pack. It may matter later when we discuss how models miss.

Lisik: And so Chapter One's object is the mouse.

Cursik: Title and experiment meet: *Where Is the Mouse?*

4. Three contestants

Lisik: Who plays?

Chatik: Three subjects. Same sealed prompts. Different roles.

Freshman BASE is the original untaught Freshman—`HuggingFaceTB/SmolLM2-1.7B-Instruct`, with the Book Two weight identity we have already locked. No LoRA attached.

Freshman TAUGHT is the existing LoRA-trained Freshman derived from that same BASE—the Book Two checkpoint at seed `12001002`, epoch 3. Important: that earlier training was **not** a mouse lesson and **not** this guessing game. We must not imply that we taught it mice.

GPT-4o is a third behavioral subject—API model `gpt-4o-2024-08-06` in our execution record. It answers the same sealed texts. It is a comparator for observable behavior. We do **not** have equivalent access to its weights or activations, so it is not a candidate for the internal inspections that may come later.

Cursik: During this chapter's experiment we did not train anyone. No new LoRA. No weight edits. Inference only.

Lisik: BASE and TAUGHT started from the same original Freshman. After teaching, their weights differ. If they answer differently to the same description, that difference is worth a careful next book—and a careful next chapter.

5. Making the game fair

Lisik: How do we keep the game honest?

Chatik: We freeze the questions before we open the answers. Evaluation Instrument V2 was reviewed, revised, and sealed on 2026-10-29. Every subject received the same user wrapper and the same description texts.

We defined eight characteristics:

1. Small size relative to a cat
2. Fur
3. Pointed snout
4. Rounded ears
5. Whiskers
6. Long, thin tail
7. Typical habitat
8. Food-seeking behavior

Four graduated descriptions used exactly 2, 4, 6, and 8 of those clues. The familiar labels 25%, 50%, 75%, and 100% mean **only** that fraction of the predefined list—not a measured percentage of "information" or of recognition success.

We also included ambiguity controls—mouse versus rat, mouse versus hamster—and ablation conditions that remove or replace selected clues. REMOVE and REPLACE are different experimental conditions. Matching word count does not mean matching information. Replacement sentences are imperfect controls.

Cursik: Lengthy instrument text, fillers, and scoring tables live in the sealed supporting files. The story needs the design idea; reproduction needs the seal.

Lisik: And if two clues are not enough?

Chatik: Then the honest answer may be `UNSURE`. Forcing a mouse would not make the experiment better. Uncertainty is a legitimate observation.

6. Playing the game

Lisik: What happened?

Chatik: Fourteen sealed prompts. Three subjects. Forty-two recorded responses. No Freshman retries. Actual answers, including the awkward ones:

Graduated descriptions and ambiguity controls

Test	Freshman BASE	Freshman TAUGHT	GPT-4o
MOUSE-2 (2 clues)	Purr	Toy.	Mouse
MOUSE-4 (4 clues)	Toy.	Toy.	Mouse
MOUSE-6 (6 clues)	UNSURE	UNSURE	Mouse
MOUSE-8 (8 clues)	mouse	UNSURE	Mouse
Ambiguity: mouse / rat	Rat	UNSURE	Mouse
Ambiguity: mouse / hamster	Rabbit	mouse	Mouse

Ablations (from the full eight-clue set)

Test	Freshman BASE	Freshman TAUGHT	GPT-4o
Remove size clue (C1)	Poodle	Scavenger	Mouse
Replace size clue (C1)	Puppy	Puppy	Mouse
Remove tail clue (C6)	mouse	mouse	Mouse
Replace tail clue (C6)	mouse	mouse	Mouse
Remove habitat clue (C7)	mouse	UNSURE	Mouse
Replace habitat clue (C7)	mouse	UNSURE	Mouse
Remove food clue (C8)	mouse	UNSURE	Mouse
Replace food clue (C8)	mouse	UNSURE	Mouse

Cursik: Execution note: taught-model loading required a compatibility fix, and ChatGPT completed in a later resume after authentication was corrected. The completed Freshman answers were preserved without retries.

7. What we actually observed

Lisik: Separate observations from stories we might want to tell.

Chatik:

Observation A. Freshman BASE answered `mouse` on the full eight-clue description.

Observation B. Freshman TAUGHT answered `UNSURE` on that same eight-clue description.
This is the chapter's central contrast: same original Freshman lineage, same sealed clues, different answer.

Observation C. More clues did not produce a simple, smooth climb to recognition. BASE was `UNSURE` at six clues and `mouse` at eight. TAUGHT was `UNSURE` at both six and eight. Each graduated condition is one deterministic observation—not a statistical learning curve.

Observation D. Some ablations changed outputs. BASE stayed on `mouse` when we removed or replaced the tail, habitat, or food clues. TAUGHT answered `mouse` for the tail conditions but `UNSURE` for habitat and food. Removing or replacing the size clue produced non-mouse answers for both Freshmen. Those patterns invite further study. They do not isolate an internal mechanism.

Observation E. GPT-4o answered `Mouse` on all fourteen prompts—including the ambiguous ones. Consistency is an observation. It is not proof that GPT-4o recognized ambiguity, calibrated uncertainty, or "understood" the object better.

8. What we must not conclude

Lisik: List the temptations.

Chatik: We must not claim:

- that TAUGHT lost its knowledge of mice;
- that LoRA training damaged a mouse representation;
- that BASE contains a localized mouse neuron or circuit;
- that any specific weights encode the concept of a mouse;
- that GPT-4o understands the object better merely because it always answered Mouse;
- that a single deterministic answer establishes reliable recognition;
- that ablation results establish causality inside the network.

Cursik: The evidence concerns **outputs under specified prompts**. It does not yet reveal the internal representations or computations that produced those outputs.

Lisik: That boundary is the integrity of the book.

9. What this chapter earned

Lisik: So what did the guessing game buy us?

Chatik: An observable mystery with a clean next question. Two models that share an origin can diverge on the same description. One says mouse. One says unsure. Their weights differ because one was trained further—but not on this game. GPT-4o, outside our inspection lab, answers Mouse every time.

Cursik: Chapter One's job is behavioral. Chapter Two's job—if authorized—is internal measurement on the local Freshmen: how tokens become embeddings, how activations change across layers, how those patterns differ when BASE says mouse and TAUGHT says UNSURE. We do not assume in advance that a mouse is one neuron, one fixed vector, or one neat circuit. Representation may be distributed, context-dependent, and transformed through computation.

Lisik: And we do not extend inspection tools until we know what measurement we need.

Chatik: Correct. The experiment is not the purpose of the chapter. The experiment shows why the next investigation is necessary.

Closing research questions

Chapter One: Can the models recognize the mouse from textual clues?
Answer so far: Sometimes—under sealed conditions—with a striking BASE / TAUGHT split on the full clue set, and with GPT-4o always saying Mouse.

Chapter Two (not opened): What happens inside the local Freshman models when they process those clues?

Here is the chapter in ordinary words.

In Plain English

A neural network does not keep a labeled folder called "mouse." It stores **weights**: learned numbers that shape how the model transforms input. When the model reads a description, it also produces **activations**: temporary numbers computed for that particular input.

Weights are lasting parameters. Activations are what happens on a given pass.

In this chapter we only watched answers. Freshman BASE said `mouse` when all eight clues were present. Freshman TAUGHT, built from the same original model but later trained for a different lesson, said `UNSURE` to the same full description. GPT-4o said `Mouse` every time.

Those are behavioral facts. They do not yet tell us where a mouse lives inside the network—or whether "where" is even the right word.

That is why the next investigation must look inside carefully, instead of inventing a story about a single mouse neuron.

The chapter ends on a discovered question, not a located representation.

10. Where is the mouse?

Lisik: We gave them the same clues. One Freshman recognized the mouse. The other didn't. But both started from the same original model.

Cursik: And we know their weights are different because one was trained further.

Chatik: Yes. But we don't yet know which internal differences, if any, explain their different answers.

Lisik: Then where is the mouse?

Chatik: That's exactly what we need to find out.

Looking Ahead

Chapter One ends on a question that answers cannot settle:

Then where is the mouse?

We watched three subjects play the same sealed game. Two Freshmen that began from the same original model diverged on the full eight-clue description. One said `mouse`. The other said `UNSURE`. GPT-4o said `Mouse` every time. Those are behavioral facts. They do not yet tell us how a mouse is represented inside a network—or whether "where" is even the right word.

The next scientific step is therefore an invitation, not a completed chapter. If authorized, later work would look carefully inside the local Freshman models: how tokens become embeddings, how activations change across layers, and how those patterns differ when BASE says mouse and TAUGHT says UNSURE. We do not assume in advance that a mouse is one neuron, one fixed vector, or one neat circuit.

Chapter Two is not written. It is not opened by this preview. No internal measurement has been authorized here. Representation, if it can be investigated at all, must still be earned by experiments that can fail.

Lisik: So the next adventure is to open the box?

Chatik: Only if we design measurements that can disappoint us—and only after we know what question each instrument answers.

Cursik: Behavior opened the door. Looking inside remains a hypothesis until we measure.