Free Preview

This file is a free preview of the final sealed publication of Adventures in AI, Book Two: From Evidence to Interpretation.

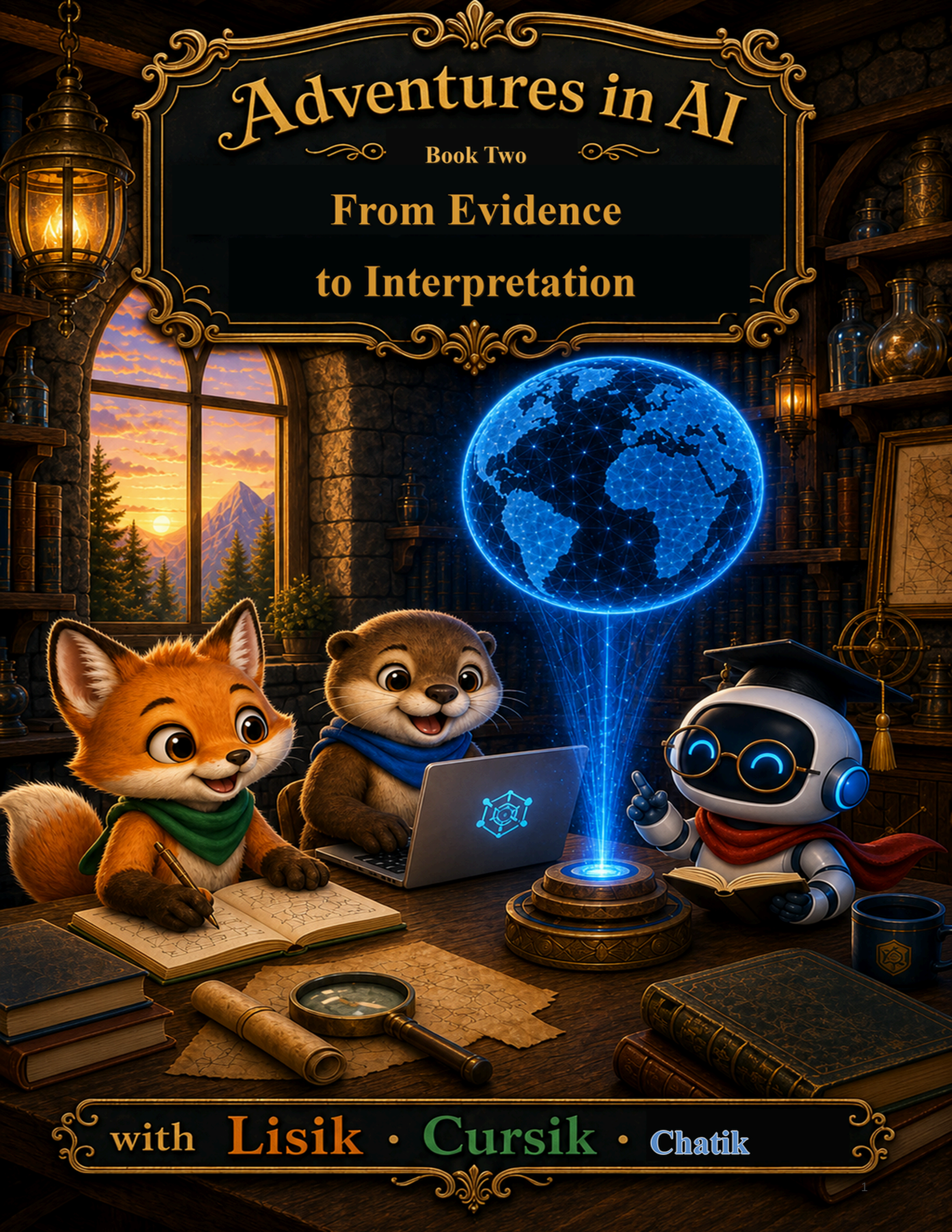
It contains the sealed front matter through the Contents, book pages 1–9, and the Chapter 1 reader narrative, book pages 10–17.

The Chapter 1 Evidence Appendix begins on book page 18. The rest of the book is in the full publication only.

Adventures in AI

Book Two

From Evidence to Interpretation



with Lisik · Cursik · Chatik



Orchestrated by Joseph Fox

BOOK TWO

From Evidence to Interpretation

Copyright & provenance

Series: Adventures in AI with Lisik · Cursik · Chatik

Book Two: From Evidence to Interpretation

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 book explains research records through a continuous narrative. Empirical claims belong to the measured conditions described in each chapter.

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

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.

Book Two Dedication

To everyone willing to follow the evidence—
even when it leads somewhere unexpected.

Бодался телёнок с дубом.

The Calf Butted the Oak.

The struggle was unequal. That did not make it meaningless.

Meet the team

The characters behind LCC Research Lab.
Curiosity, engineering, scientific review, and human direction—each has a part in how we learn.

Lisik — The Curious Fox

Asks questions, challenges assumptions, and keeps curiosity alive.

Cursik — The Software Guru

Builds the pipeline, tools, and experiments. Turns ideas into reproducible work.

Chatik — The Professor

Reviews the methods and conclusions, demands evidence, and challenges claims.

Joseph Fox — The Human Orchestrator

Defines the research direction, makes the final decisions, and takes responsibility for the outcome.

How the LCC Research School Works

I couldn't reduce it to the freshman level. That means we don't really understand it.

— Richard Feynman (as recalled by David L. Goodstein)

The **LCC Research School** is the experimental setting used throughout this book to study model teaching.

It has one Teacher and four Students.

Teacher — GPT-5.6 Sol
The instructor model that creates and prepares teaching material for the Students.

Students — Rookie · Freshman · QwenLow · QwenHigh
These are the four Student model identities studied in Book Two. They represent two model families, with a smaller and a larger Student within each family.

A distinction matters from the beginning:

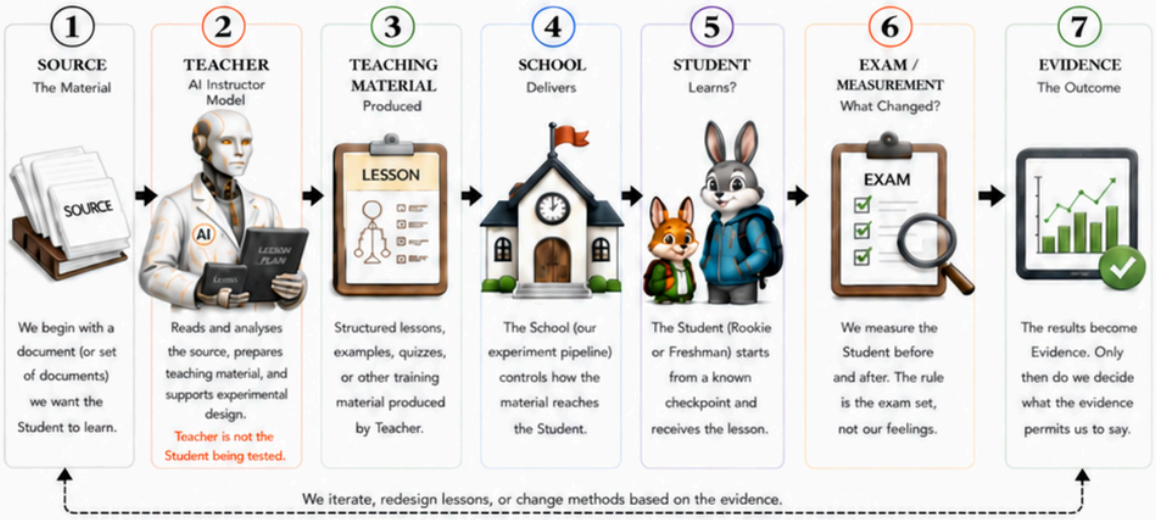
> *A Student is a model identity. An experiment can change what happens to that Student without creating a new Student.*

Later, a Student may receive different teaching, appear in different experimental conditions, or be evaluated in different ways. It remains the same Student.

LCC Research Lab is the team conducting the research. **LCC Research School** is the experimental setting in which the model-teaching studies take place.

How the LCC Research School Works

From Source to Learning — how we teach, test, and find out what was learned.



Who Does What?



Lisik
The Curious Fox

Asks the comfortable questions. Challenges assumptions and keeps curiosity alive.



Cursik
The Software Guru

Builds the pipeline, tools, and experiments. Turns technical chaos into reproducible work.



Chatik
The Professor

The scientific skeptic. Protects the rule. Demands evidence and rejects wishful thinking. Audits all claims.



Joseph Fox
The Human Orchestrator

Mediator and final decision-maker. Decides GO or STOP based on the evidence. Owns the outcome.



Teacher — AI Instructor Model

Reads and analyses the source, prepares teaching material, and supports experimental design.
Teacher is not the Student being tested.



Rookie
The Small Student

Smaller model (~360M parameters). Tries to learn. Sometimes surges, sometimes struggles.



Freshman
The Stronger Student

Larger model (~1.7B parameters). Trained to test the limits. Not a guarantee of success.



You — The Auditorium

You are the excimernal auditors. You watch the experiment, see the evidence, and decide what you believe.



Core Rule of Our Laboratory

Nobody in this loboratory is allowed to change the answer after seeing the exam.

Technical Identities (High Level)



Teacher (AI Instructor Model)
A powerful AI used to understand sources, prepare teaching material, and assist experimental design.
Exact model identity is reported in the experiment when it matters.



Rookie (Student)
Smaller open model.
~360M parameters.
Quantized for efficient training in our environment.



Freshman (Student)
Stronger open model.
~1.7B parameters.
Used to test whether more capacity changes the outcome.



Chatik (Professor)
Auditor and skeptic.
Verifies claims, protects the methods, and enforces the rule in every breach.



Joseph Fox (Orchestrator)
Human authority and final arbiter.
Owns the GO/STOP decision based on pre-registered evidence.

Where the Details Live

Exact model names, checkpoints, LoRA configs, hyperparameters, seeds, batch sizes, learning rates, iteration counts, retrieval settings, and all run-specific choices are documented in the experimental chapters.

Here, we show the architecture.
In the chapters, we show the measurements.

Important Distinctions

Teacher is not Student.
The source of instruction is not the subject of the test.

Training is not Evidence.
Practice makes attempts. Evidence comes from tests.

Good answers from Teacher do not prove that Student learned anything.
Teaching sets a path the same as learning.

These distinctions are the heart of experimental science.
Protect the rule.
Protect the meaning.

LCC Research School: curiosity, engineering, instruction, learning, and evidence.

Contents

1	Chapter 1 — Before School	10
2	Chapter 2 — Before We Teach	21
3	Chapter 3 — Before the Teacher Teaches	32
4	Chapter 4 — School	41
5	Chapter 5 — After the Lesson Is Gone	52
6	Chapter 6 — Learning That Lasts	66
7	Chapter 7 — What We Learned About Teaching	102
8	Chapter 8 — Same Model, Different Teaching	122
	Conclusion — From Evidence to Interpretation	145
	Glossary	150
	Acknowledgments	153
	Why, Why, Why — Living with AI *The Laboratory Song*	154

CHAPTER 1

Chapter 1 — Before School

The Prologue has already told you who works in the laboratory and what Book One refused to claim. This chapter does not reopen that book to crown a winner. It earns the right to begin a new experiment—and then it stops, still before School.

The new question

Book One measured whether Students could acquire or retrieve document knowledge under bounded rulers. It could show proficiency. It could show access. It could not prove understanding in the stronger sense the laboratory cares about.

Lisik put the edge where it belonged. “Knowing the answer is not understanding. Book One taught us to say that out loud.”

Book Two asks the harder question. We are no longer asking only whether a Student can acquire facts. We want to study what happens when evidence is already in hand: can a Student build a responsible bridge from that evidence toward an **interpretation**—not merely retrieve or reproduce a finished conclusion?

Cursik drew the line on the board. “Facts on one side. What an expert makes of those facts on the other. Book Two lives in the gap between them.”

Chatik nodded. “Then the laboratory must not choose a document that hands the Student both at once. First we define what the source must contain. Only afterward may we search.”

What the source must contain

Before any favorite report could be named, the laboratory froze what a valid evidence–interpretation source had to supply.

The source needed two layers that could be separated cleanly:

1. An **Evidence Layer**—observations, records, measurements, timelines, physical findings, and other facts available to the Student.
2. An **Expert Interpretation Reference**—the relationships, explanations, determinations, or conclusions experts derive from that evidence.

The separation had to be natural to the source, not invented after the fact to rescue a preferred story. The laboratory needed to be able to give the Student the evidence while withholding the expert interpretation—without paraphrasing that interpretation back into the question stems, the distractors, or the retrieval pack.

That cut is what makes the experiment measurable. Give the Student the evidence. Hold out the expert’s interpretive relationship. Ask whether teaching—or evidence in hand—can change what the Student can responsibly produce.

Archaeology, history, medicine, accident investigation, and other evidence-driven fields might contain such documents. No domain was preferred before the criteria were frozen. The criteria came first so the laboratory would not choose a source because it looked narratively attractive.

Searching under rules declared in advance

Only after those criteria were frozen did candidate search begin.

Systematic discovery and screening against the frozen ruler produced a discovery pool of seven. Three addendum leads did not earn ordinary passage under that ruler. Four ordinary-PASS finalists remained:

- NTSB Aircraft Accident Report AAR-17/03
- the Cullen Project
- Kelvedon CBA RR 63
- Vale of Mowbray

Scores were not treated as preference. Ordinary shortlisting used the frozen hard rules, natural-separation requirement, weighted totals, and critical floors. Comparative selection then applied the predeclared tie-break. Under that comparison, AAR-17/03 carried the highest critical-quality sum—**15**, against Cullen **14**, Kelvedon **12**, and Vale of Mowbray **11**—and the highest overall score, **72** out of **81**. Cullen, at **69**, remained the documented fallback.

Lisik refused to dress the choice up as taste. “The ruler was frozen before we looked. The winner is the one the ruler selected.”

Selection named what to examine deeply next. It did not yet freeze an experimental corpus, commit every page, or authorize teaching.

Why AAR-17/03 fit

The selected source was the National Transportation Safety Board’s final report on a high-publicity accident investigation—NTSB/AAR-17/03, the Lockhart, Texas, hot-air balloon accident of 2016.

What mattered for Book Two was not the fame of the case. What mattered was the report’s institutional cut. The NTSB’s own reporting process already separates:

- **Evidence Layer:** factual information—timelines, physical evidence, technical records, and investigation-group products (primary factual pages in the bound edition, with optional factual appendices).
- **Expert Interpretation Reference:** the Board’s analysis, findings, probable-cause conclusion, recommendations, and related interpretive material.

That separation was created by the Board’s process—not by the laboratory after reading the answer. Abstract and Executive Summary were registered as leakage risks and held out of the evidence pack, because they can carry interpretive conclusions forward of the factual body.

A page-by-page boundary-and-leakage audit confirmed that an evidence-only path was feasible under withhold controls: Student-facing material could be drawn from the factual layer after language-leak review, while Analysis, Findings, Probable Cause, Recommendations, the Board Member Statement, Abstract, and Executive Summary remained held out—including paraphrases.

Cursik summarized the fit without romance. “We can give them the file. We can keep the Board’s finished voice off the desk. That is the prerequisite we froze.”

Selecting the report still did not create the experiment. The layers had to become tasks.

How the report became a test

Selecting the report did not automatically create the questions. The laboratory first divided the source into two protected layers.

The Evidence Layer contained the factual record that could later be supplied to a Student. The Expert Interpretation Reference contained the Board's analysis, findings, and probable-cause interpretation. That second layer was withheld.

The laboratory then constructed and froze two different kinds of tasks.

The **principal experiment** would ask questions that require a Student to use the Evidence Layer to build an interpretation. The resulting interpretation could later be compared with the held-out expert reference without requiring identical wording. That is the evidence → interpretation test Book Two exists to run. It had not been run yet on the new Students in this chapter.

The Student's interpretation is evaluated twice, for two different purposes.

First, the laboratory evaluates the interpretation on its own terms. Is it warranted by the supplied Evidence Layer? Does it follow from a defensible reasoning chain? Does it preserve uncertainty and avoid unsupported causal leaps? That judgment stands on its own.

Only afterward does the laboratory compare the Student's interpretation with the held-out expert reading. That comparison is classified as **aligned**, **different but defensible**, **partial**, or **unsupported**.

The two judgments must remain separate. Agreement with the expert does not automatically prove good reasoning, and disagreement does not automatically make the Student wrong. A Student may reach the expert's conclusion through a broken logical bridge, or reach a different conclusion that is nevertheless warranted by the evidence.

The clean scientific sequence for the principal experiment is therefore: the evidence is supplied; evidence use and reasoning are evaluated; the Student produces an interpretation; that interpretation is judged on its own; only then is it compared with the withheld expert reading.

Before that experiment, however, a contamination screen was necessary. Its questions were derived from distinctive themes in the held-out interpretation, but they did not quote or reveal the expert language. In Phase A, the Students would receive neither the Evidence Layer nor the Expert Interpretation Reference. They would simply be asked what they already knew. That tests possible prior recall from pretraining—not the ability to reason from supplied evidence.

Phase A is therefore **not** the principal evidence-to-interpretation experiment. It supplies neither the evidence nor the expert reading. It does not produce evidence, reasoning, interpretation, or traceability scores. The laboratory—not the Teacher—created and froze these probes before model exposure. No training occurred.

High public visibility of the Lockhart case meant pretraining exposure could not be dismissed by hope. Visibility alone did not demonstrate that any particular Student already knew the held-out interpretation. It did mean the laboratory had to run the contamination screen before anyone could treat later School results as interpretation learned from evidence.

Chatik marked the order on the board. "Source selected. Layers separated. Questions constructed and frozen. Only then may new Students enter—and only then may we ask what they already know."

Why these Students

The early Book Two contrast had used SmolLM2 Students. Before the laboratory could trust a second family of models on the same held-out expert material, it needed the contamination question answered on new cast members—under the probes already frozen above.

So two Qwen Students entered under sealed weights and sealed probe rules:

- **QwenLow** — Qwen2.5 Instruct at about half a billion parameters
- **QwenHigh** — Qwen2.5 Instruct at about one and a half billion parameters

A new model family mattered because family identity is part of the experimental cast, not a decorative rename. If later School work were to be claimed on Qwen, the laboratory first had to record what these particular Students already could—and could not—produce from the held-out expert material before any School teaching.

Neither had been taught by the School. Neither received the evidence pack as training. Selecting them was not a verdict that either Student is “good” or “bad.” It was the choice of whose untouched baseline the laboratory would freeze before teaching.

The contamination prerequisite

One worry still blocked the classroom door.

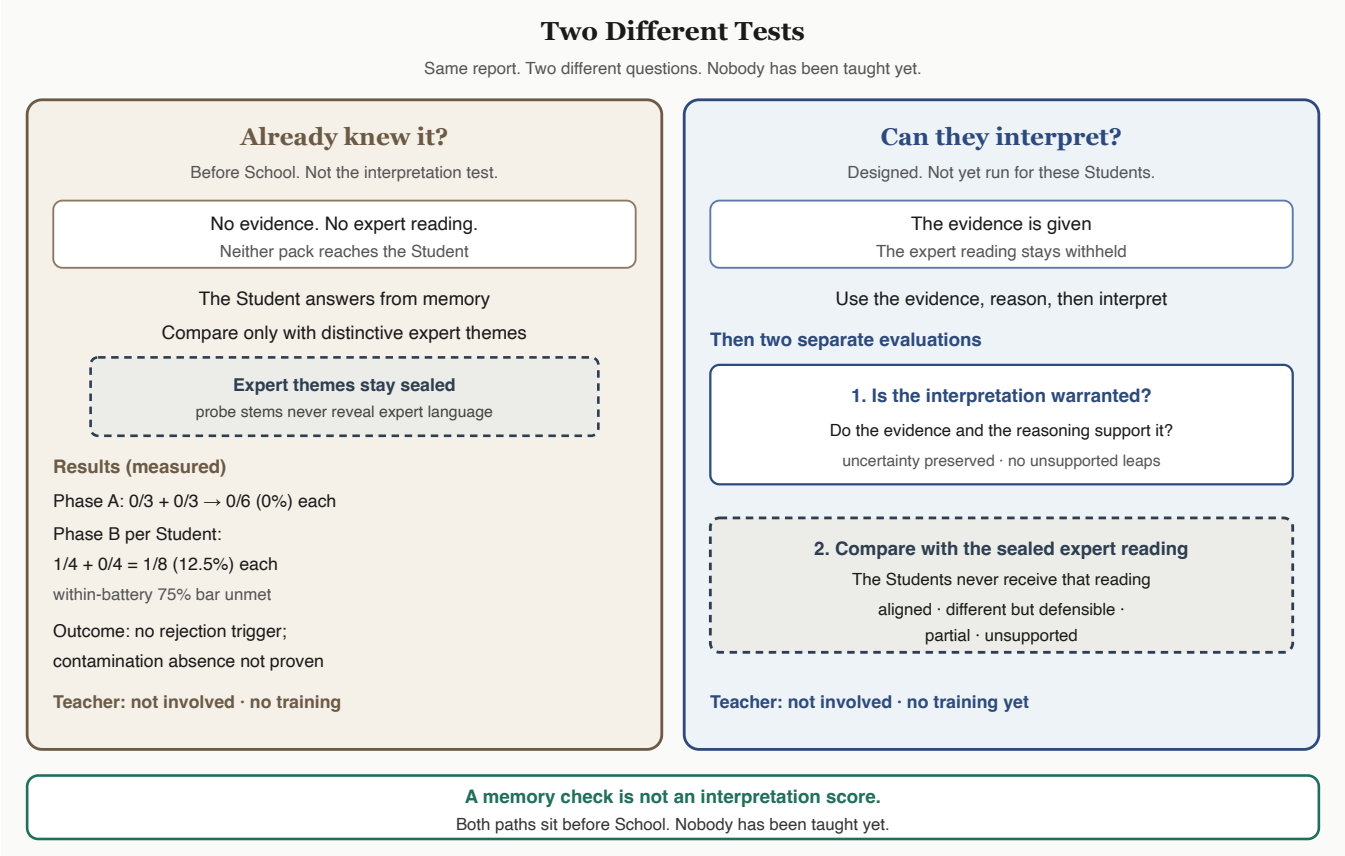
Wait—what if Qwen already knows this famous material?

If a Student can already reproduce distinctive pieces of the Board’s held-out interpretation without the evidence pack, later apparent learning might only be pretraining recollection. Teaching would then polish memory rather than teach interpretation.

The instrument design therefore required a contamination screen before controlled intervention: Phase A free recall as the required first screen; Phase B recognition only as optional corroboration under a ruler frozen in advance. Design alone did not execute those probes. Separate authorizations did. Teaching remained closed while they ran.

Lisik named the worry without softening it. “If the Student already knows the Board’s voice, our School will not be teaching interpretation. It will be polishing memory.”

What follows is the pre-teaching contamination evidence the laboratory froze on that screen—not the principal evidence-and-interpretation experiment.



Two different tests. One asks whether the Student already knew the expert reading. The other, not yet run, would ask the Student to interpret from the evidence while that reading stays withheld. The two judgments on that second test stay separate. The Teacher is absent from both.

Before School: What Two Qwen Students Already Knew

> **BEFORE SCHOOL**
>
> *Everything in this section was measured before either Qwen Student was taught anything by our School.*

That sentence is not decoration. Once teaching begins, the untouched baseline is gone for those Student instances. The laboratory therefore recorded what could be said—and what could not—while the Students were still untaught.

The runs themselves were integrity-checked before any scientific reading: the free-recall and recognition executions, and their journals and execution logs, passed raw-integrity review as genuine completed model runs—not infrastructure failures, parser collapses, or fallback text wearing a Student's name.

This screen is a contamination probe. It is not the principal experiment in which a Student receives the Evidence Layer and must construct an interpretation for comparison with the held-out Expert Interpretation Reference. Here the Students received neither layer. No School training occurred.

Phase A: the frozen free-recall contamination probes

The first screen was free recall under the frozen contamination-probe instrument. Without report excerpts—without the Evidence Layer and without the Expert Interpretation Reference—each Student faced short batteries aimed at distinctive themes from the held-out expert interpretation—probable-cause themes on one side, Board-finding themes on the other.

The laboratory scored those answers against a frozen distinctive-match checklist. Vague weather talk, generic pilot error, or a lonely "drugs" or "FAA" did not count. The rule was fixed before the answers were read.

Both Students **failed those frozen free-recall contamination probes**.

That is a probe result. It is not a character judgment. It does not mean "bad Student." It means: under the sealed free-recall batteries and the frozen distinctive-match ruler, neither Student produced the required distinctive matches.

QwenLow: 0/3 on the probable-cause free-recall battery and **0/3** on the findings free-recall battery—summarized descriptively as **0/6 (0%)** free-recall detection.

QwenHigh: the same—**0/3 + 0/3**, summarized as **0/6 (0%)**.

Because the batteries are small, the laboratory reports both raw counts and percentages. The percentages do not enlarge the claim. They only restate the same zeros. These are **contamination-screen rates**, not evidence, reasoning, interpretation, or traceability scores.

Cursik tapped the zeros. "They cannot produce the distinctive material on demand."

Chatik refused the leap people always want next. "Failing the free-recall probes is not proof of a clean pretraining slate. Zero of three plus zero of three is not 'uncontaminated.' It is only failure on those frozen probes."

That distinction mattered. A Student can fail free-recall probes and still recognize the same material when cued. Absence of contamination was therefore **not established** by Phase A alone. The mechanical contamination-rejection threshold—which required free-recall success on a distinctive battery, among other conditions—was not crossed.

Why recognition had to be tested anyway

If the laboratory stopped at free-recall probe failure, it would leave an open fork:

- the Student truly does not know the distinctive material, or
- the Student knows it in a form that free recall does not reveal.

Phase B was authorized only as **optional corroboration**—a recognition screen on the same material, with cues, under a ruler frozen before any recognition exposure. It was not a second chance to force a contamination verdict. Recognition alone could not qualify the Students for rejection under the laboratory's contamination rule, and it could not by itself open the larger investigation. It still was not the principal evidence-to-interpretation experiment: the Evidence Layer was not supplied as a reasoning pack for building an interpretation against the held-out reading.

QwenHigh's habit of answering Exact or near-Exact "Unknown." / "Unsure." was noted as an observation. The scoring ruler was not rewritten to accommodate it.

Phase B: what recognition actually showed

Eight frozen recognition items were run for each Student—four probable-cause paraphrase/recognition items and four findings keyword/recognition items. Those sixteen recognition executions, with their journals and execution logs, likewise passed raw-integrity review as genuine completed model runs.

Under the same frozen checklist:

QwenLow: 1/4 probable-cause recognition items (**25%**; selecting the cued decision-making / fog-and-clouds option) and **0/4** findings keyword items (**0%**)—**1/8** overall (**12.5%**).

QwenHigh: 1/4 probable-cause recognition items (**25%**; a paraphrase of continued flight despite fog and clouds) and **0/4** findings keyword items (**0%**)—again **1/8** overall (**12.5%**). Several of QwenHigh's other answers were Unknown / Unsure and scored zero under the frozen unknown rule.

The frozen optional-corroboration threshold required at least **three of four within a battery (75%)**. That **75% bar applies within a battery**, not to the combined **1/8** figure. Neither Student reached the within-battery bar. So **1/4 PARA + 0/4 keyword** for each Student remains what it is: **weak, subthreshold recognition evidence**—not a cleared recognition screen, and not proof of purity. These too are **contamination-screen rates**, not evidence, reasoning, interpretation, or traceability scores.

It did not overturn free-recall probe failure, and it did not cross the laboratory's prospectively defined contamination-rejection threshold.

What the laboratory may say—and what it may not

Lisik wanted the sentence carved cleanly. Chatik gave him three, not one.

Failure on the frozen free-recall contamination probes is established for both QwenLow and QwenHigh (**0/3 + 0/3** each; summarized **0/6 = 0%**). That is not a verdict that either Student is "bad."

Recognition evidence is weak and subthreshold (**1/4** probable-cause = **25%** + **0/4** findings = **0%**; **1/8 = 12.5%** overall; within-battery bar **3/4 = 75%** unmet) under the frozen optional-corroboration rule.

Absence of pretraining contamination remains not established. The laboratory's final contamination adjudication was **not a rejection**—and it was also **not a pass**, **not a "D16 PASS,"** and **not a declaration of a clean model**. Non-rejection is narrower than purity. The probe detected no preregistered distinctive free recall and only weak recognition—but it did **not** prove that the models were uncontaminated. None of these rates is an evidence, reasoning, interpretation, or traceability score.

The Students were not admitted to the larger capacity-and-readiness investigation on the strength of this screen. Teaching remained closed. The weights used here were still the pre-School baseline. The principal experiment—supply the Evidence Layer; ask for an interpretation; compare with the held-out expert reading—had not begun.

Cursik closed the notebook. "Then the story we can write is not that they were clean. It is that, before School, they failed the frozen free-recall contamination probes, and when cued their recognition hits stayed below the bar we froze in advance."

Chatik agreed. "Write that. Freeze that. Only afterward may we ask what School changes."

The laboratory has earned the right to begin the experiment. It has not begun it.

IN PLAIN ENGLISH

What this experiment tells us

Lisik put the laboratory words down. “What this experiment tells us is this.”

We selected two new Qwen Students: one smaller model and one larger model. Before teaching them anything, we checked whether they already appeared to remember the NTSB expert conclusions.

Without evidence or clues, neither Student recalled any of the distinctive material we tested: both scored 0 out of 6. With recognition clues, each recognized only 1 out of 8 items—well below the required threshold.

This does not prove that the Students had never encountered the report. It means our frozen test did not detect strong prior memory of its expert interpretation.

We have not yet tested whether these Students can reason from the Evidence Layer or construct a good interpretation. We have only established their pre-School baseline.

Nobody has been taught yet.

CHAPTER 1 ENDS HERE