



From ledgers to agents: the long automation of economic research

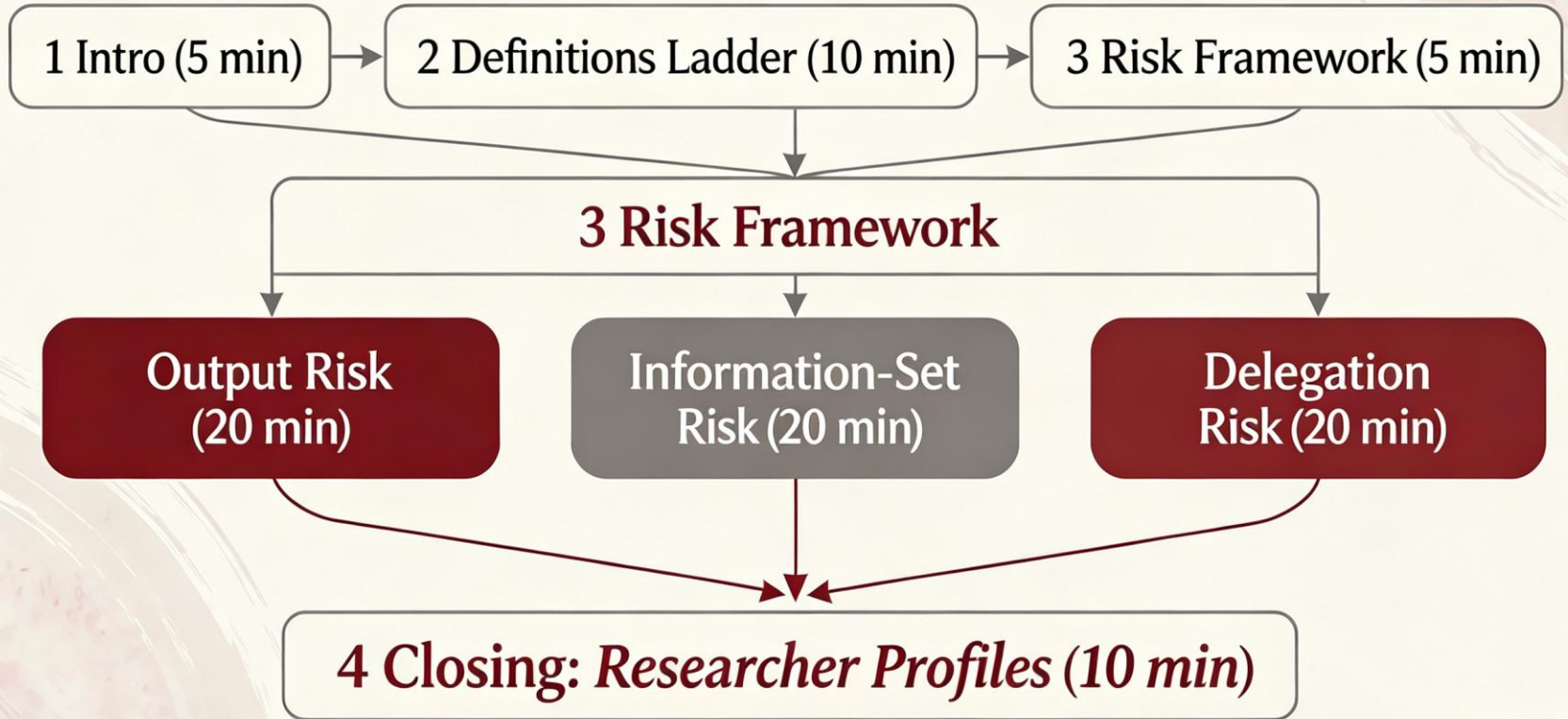
A Human History of Delegated Judgment

AI for Empirical Accounting Research · Session 1

Shirley Jiexuan Wang & Jiaqi Shao · NUS Business School

Session 1 Roadmap

AI for Empirical Accounting Research



Total: (90 minutes)

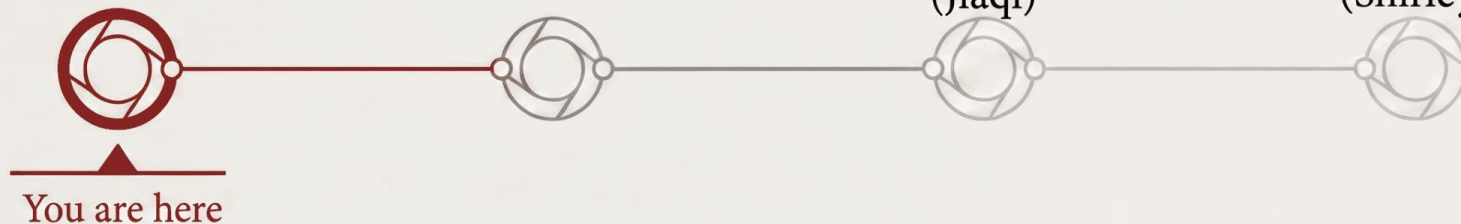
Two-Day Workshop Outline: One Research Pipeline

Session 1:
Evolution & Risk
(Shirley & Jiaqi)

Session 2:
LLM Fundamentals
(Jiaqi)

Session 3:
Advanced
Applications
(Jiaqi)

Session 4:
LLM Use Cases in
Empirical Research
(Shirley)



Each session hands you one tool; together they cover the full pipeline from raw data to final paper.

Five Quiet Years Carried Code Assistance from a Demo to Daily Practice



NOVICE NOTE: Think of this timeline as moving from a 'smart typewriter' to a 'junior research assistant'.

The Core Tension: Speed vs. Auditability

- AI agents dramatically increase the speed of research execution.
- But they obscure the intermediate steps, reducing auditability.
- Our challenge: How do we harness the speed without breaking the chain of evidence?

This is the central problem we will solve in this workshop.



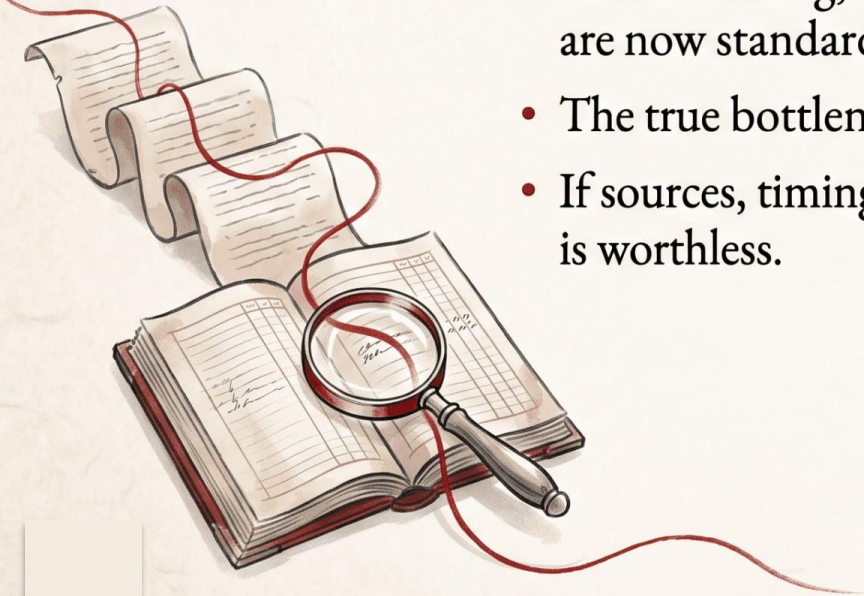
The Conclusion Arrives Early: Speed Is Valuable Only When Provenance Survives



- Faster drafting, rapid backtesting, and autonomous execution are now standard.
- The true bottleneck has shifted from generation to verification.
- If sources, timing, and permissions cannot be checked, the speed is worthless.

NOVICE NOTE

'Provenance' is a word from art history — it means the documented history of an object. In research, it means knowing exactly where every number came from.



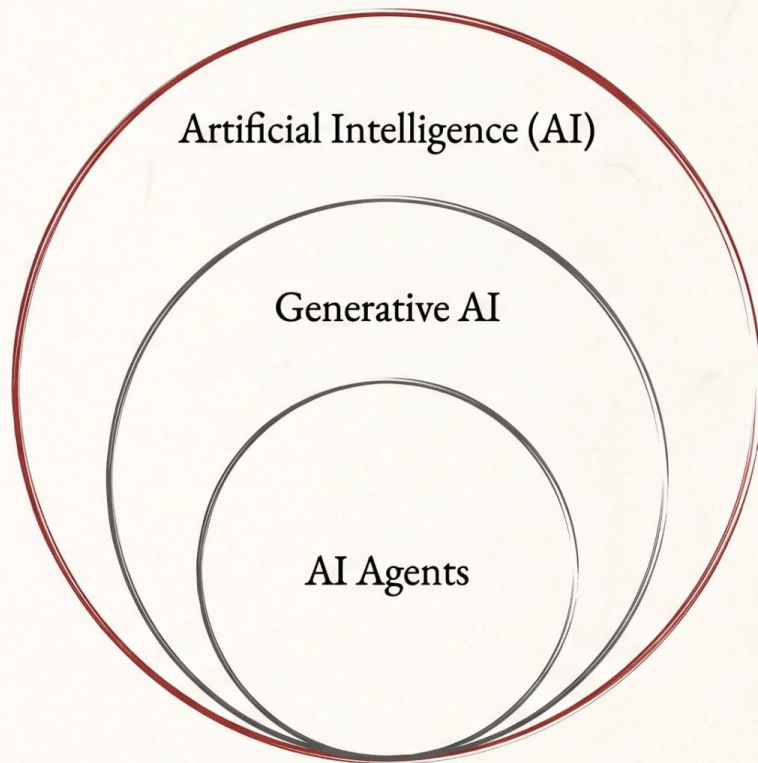
AI Is the Broad Category — Not Every AI System Is an Agent

Definition: Artificial Intelligence is the broad category of computational systems that perform tasks typically requiring human intelligence.

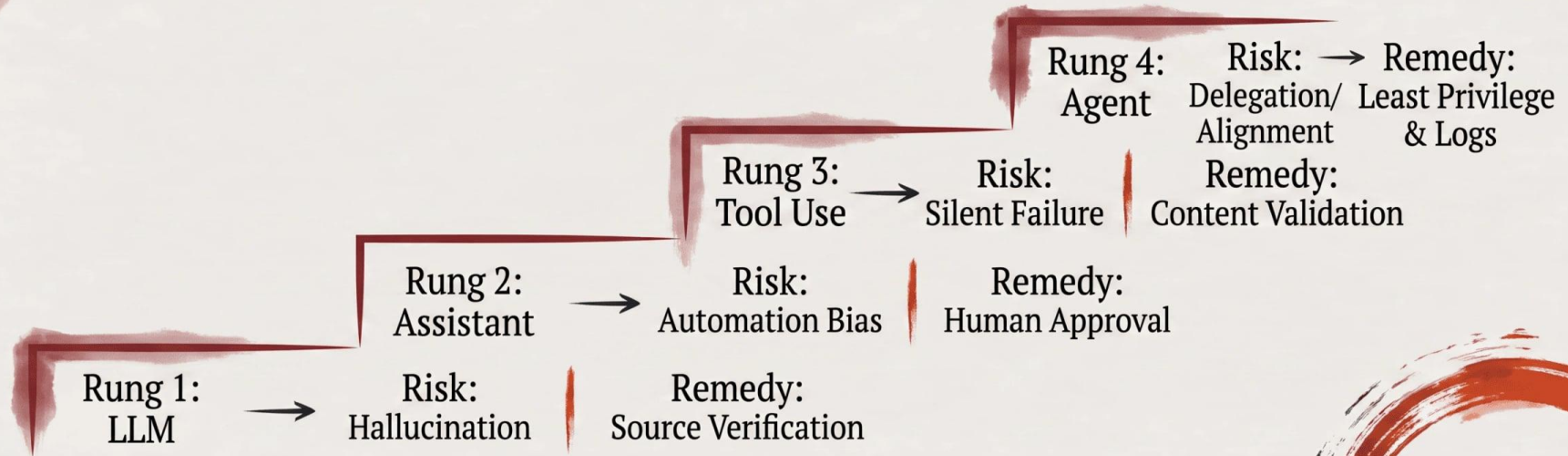
- Not all AI is generative (e.g., random forests, SVMs are AI but not generative).
- Not all generative AI is an agent.

NOVICE NOTE

'AI' is an umbrella term. When someone says 'AI did this,' ask: 'Which kind?' A regression model and an autonomous agent are both AI, but they have entirely different risks.



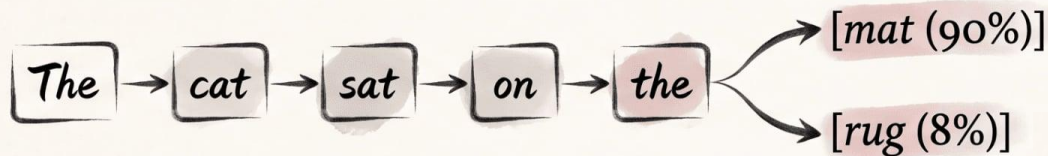
The Definitions Ladder — Each Rung Has a Different Risk Profile



*Calling everything 'AI' obscures the specific
remedy required for the specific risk.*

A Large Language Model Predicts Likely Continuations of Text

- An LLM is a statistical engine that predicts the next token based on its training data.
- It does not “know” facts or retrieve records; it completes patterns.
- Next-token prediction is fundamentally different from causal forecasting in economics.



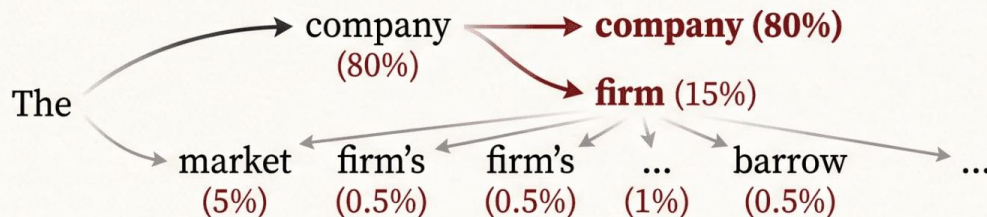
NOVICE NOTE

A “token” is roughly a word or part of a word.

The model is essentially playing a highly advanced game of autocomplete.

Next-Token Prediction Can Sound Intelligent Without Proving Understanding

- LLMs do not “know” facts; they know statistical associations between words.
- They can write a perfect abstract for a paper that doesn’t exist because the words statistically belong together.
- Fluency is not a proxy for accuracy.



NOVICE NOTE

A ‘token’ is roughly a word or part of a word.

‘Understanding’ implies a mental model of the world.

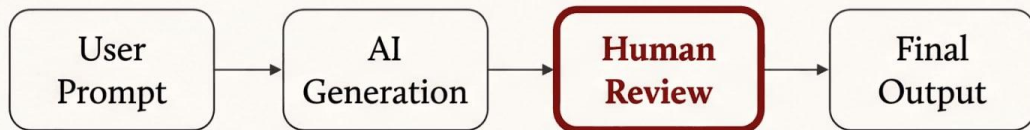
An LLM has a statistical model of language, not a mental model of reality.

An AI Assistant Usually Keeps the Human in the Approval Loop

- An assistant responds to discrete prompts with drafted text or code.
- The system does not take independent action outside its chat window.
- The human researcher remains the mandatory approval gate before any output is used.

NOVICE NOTE

An AI assistant is like a very capable intern: you ask it to do something, it drafts a response, and you review it before it goes anywhere.



Tool Use Matters Because Information and Action Cross the Model Boundary

- A standalone model only outputs text based on its training data.
- When granted “tool use,” the model can query APIs, download data, or run code.
- The risk shifts from “did it invent a fact?” to “did it silently fail at a real task?”

NOVICE NOTE

Think of an LLM like a brain in a jar. "Tool use" gives it hands and eyes to interact with the real world (like downloading a CSV from a website).

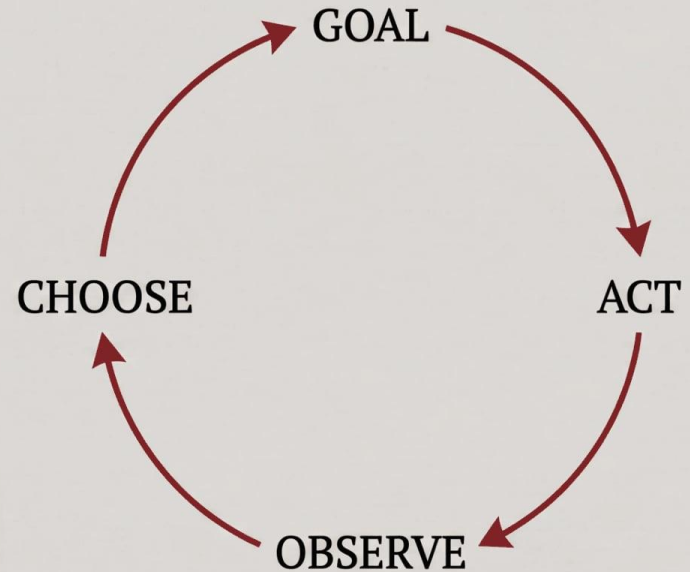


An AI Agent Works Through a Loop Instead of a Single Reply

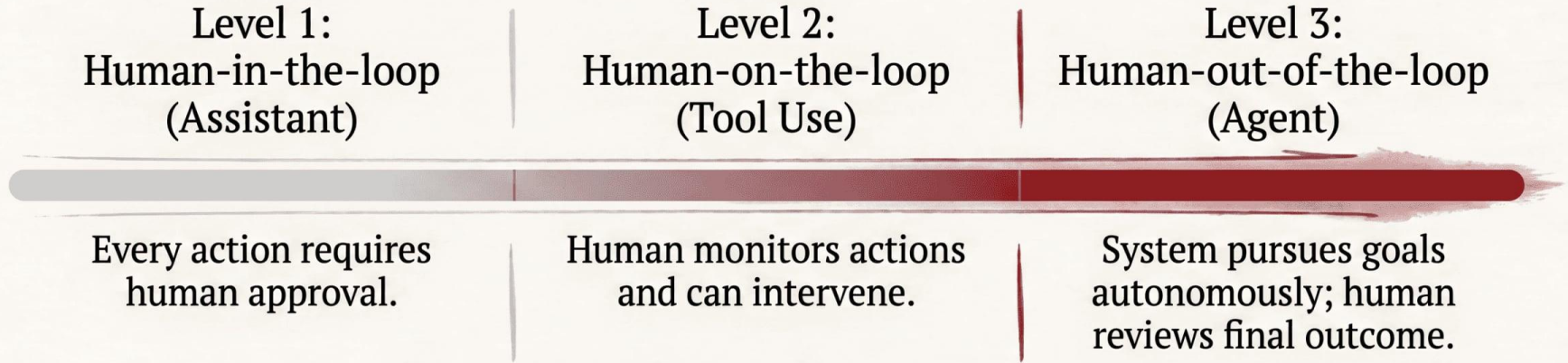
- An agent pursues a goal through multiple steps.
- It observes the result of its last action, chooses the next tool, and acts again.
- It can change the state of the world (files, code, messages) before returning to the user.

NOVICE NOTE

An assistant answers one question. An agent tries to solve a whole problem, adjusting its strategy as it goes.



Autonomy Is a Continuum, Not an On-Off Switch



The further right you move, the more the human's answerability must be engineered back in through logs and checkpoints.

Research Risk Rises Across Three Layers — Evidence Discipline Is the Answer

Output Risk (Epistemic):

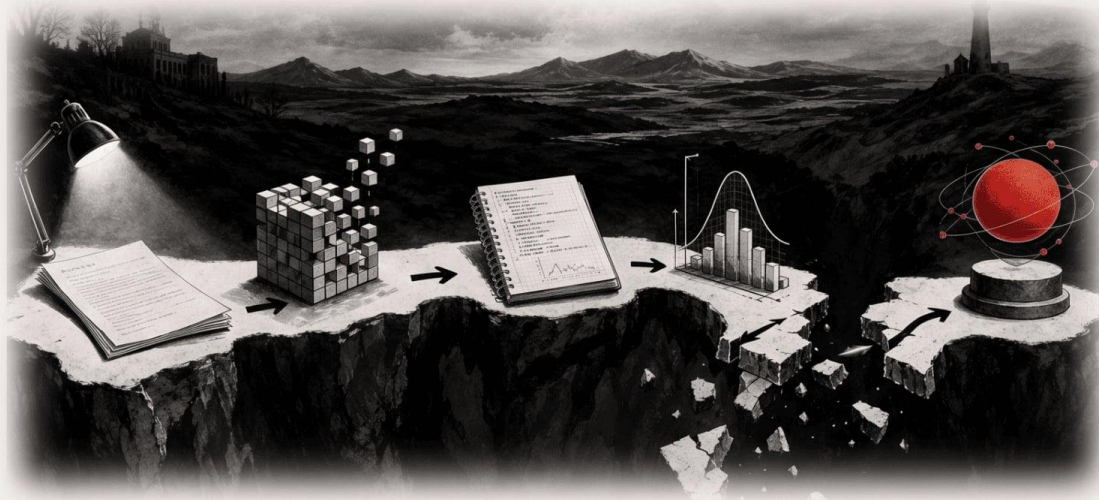
The model generates
plausible falsehoods.

Information-Set Risk (Temporal):

The model uses
future data.

Delegation Risk (Alignment):

The model takes
unauthorized actions.



NOVICE NOTE

Think of these three layers like floors in a building. If the foundation (Output) is broken, the higher floors (Delegation) will collapse.

Definitions Matter Because Each Risk Has a Different Remedy

- If you call everything ‘AI,’ you can’t fix the right problem.
- **Output Risk** (e.g., Hallucination) → **Remedy:** Source Verification
- **Information-Set Risk** (e.g., Look-Ahead Bias) → **Remedy:** Temporal Discipline
- **Delegation Risk** (e.g., Misalignment) → **Remedy:** Least Privilege & Audit Trails

NOVICE NOTE

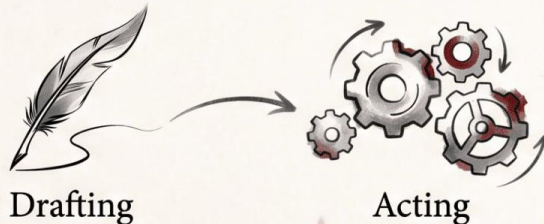
It's like medicine. You don't just take "pills" for "sickness". You need a specific antibiotic for a specific infection. Definitions give us the exact diagnosis.



The Authorship Question Changes When Work Shifts from Drafting to Acting



- Drafting (Assistant):
 - You review the text, you own the words. The evidence chain is visible.
- Acting (Agent):
 - The system executes code, downloads data, and makes methodological choices.
- The Challenge:
 - How do you defend a research choice you didn't explicitly make?



NOVICE NOTE

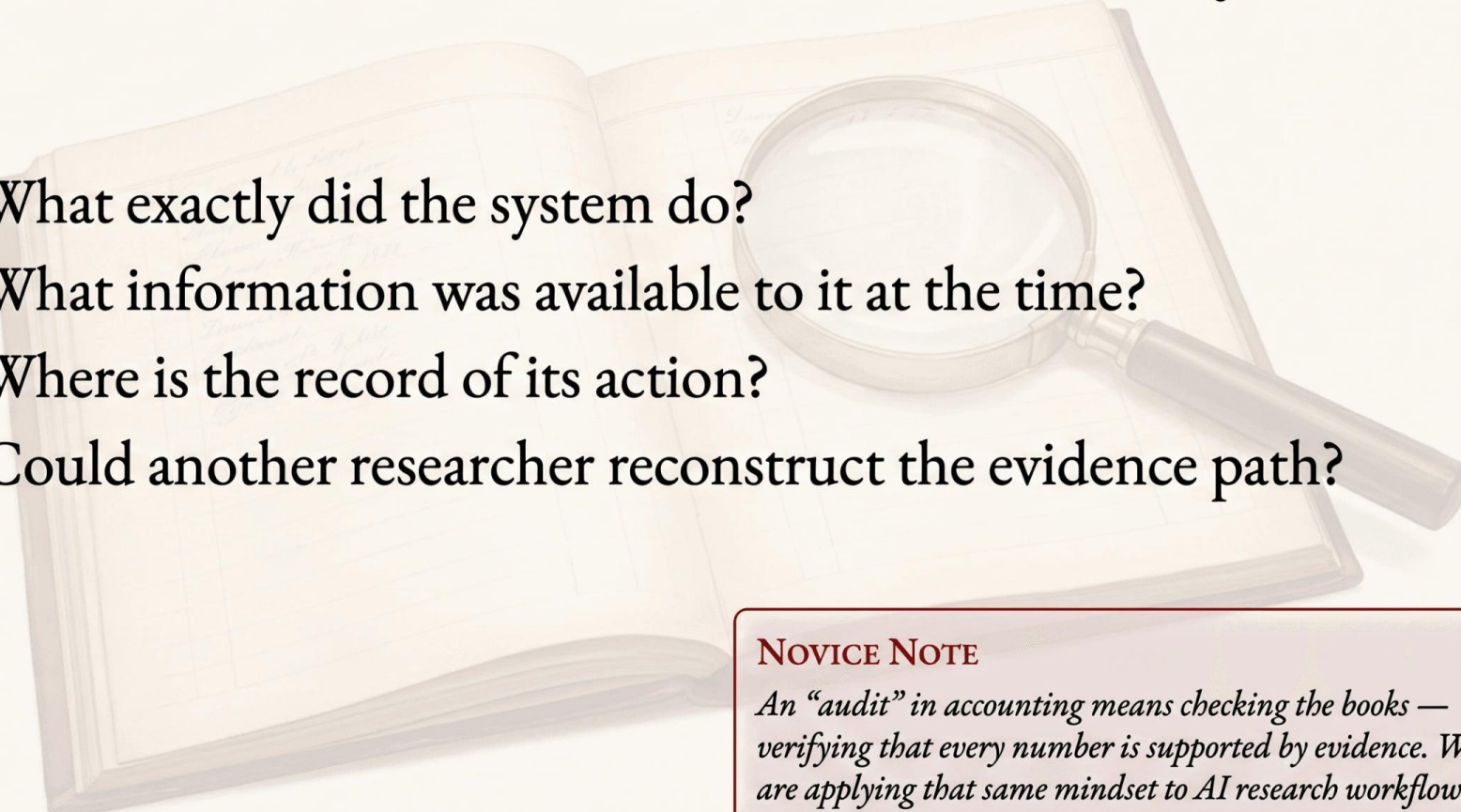
Imagine the difference between:

(A) hiring a research assistant to draft a literature review, and

(B) hiring them to run your entire experiment while you sleep.

The accountability changes.

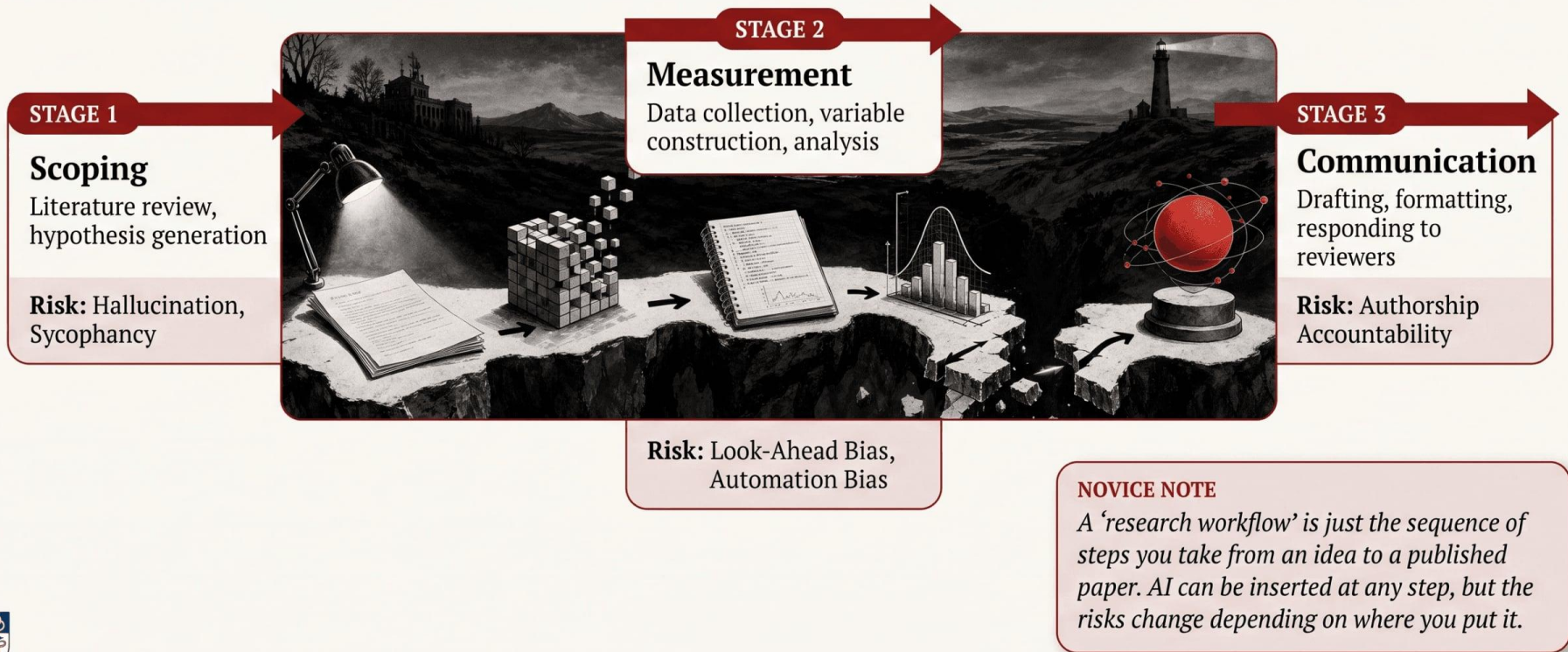
This Lecture Asks One Question on Every Slide

- 
1. What exactly did the system do?
 2. What information was available to it at the time?
 3. Where is the record of its action?
 4. Could another researcher reconstruct the evidence path?

NOVICE NOTE

An “audit” in accounting means checking the books — verifying that every number is supported by evidence. We are applying that same mindset to AI research workflows.

The Research Workflow Reveals Where AI Changes the Evidence Chain



Epistemic Risk Begins When Fluent Output Outruns Verification

- “Epistemic” relates to knowledge and how we know what we know.
- AI systems generate text faster than a human can read and verify it.
- The gap between generation speed and verification speed is where false claims become part of the research record.



NOVICE NOTE

If it takes the AI 5 seconds to write a paragraph, but it takes you 50 minutes to check all its sources, the temptation is to just trust it.

That temptation is epistemic risk.

Next, we break down these three types of risks generated from using AI in research

Output Risk (Epistemic):

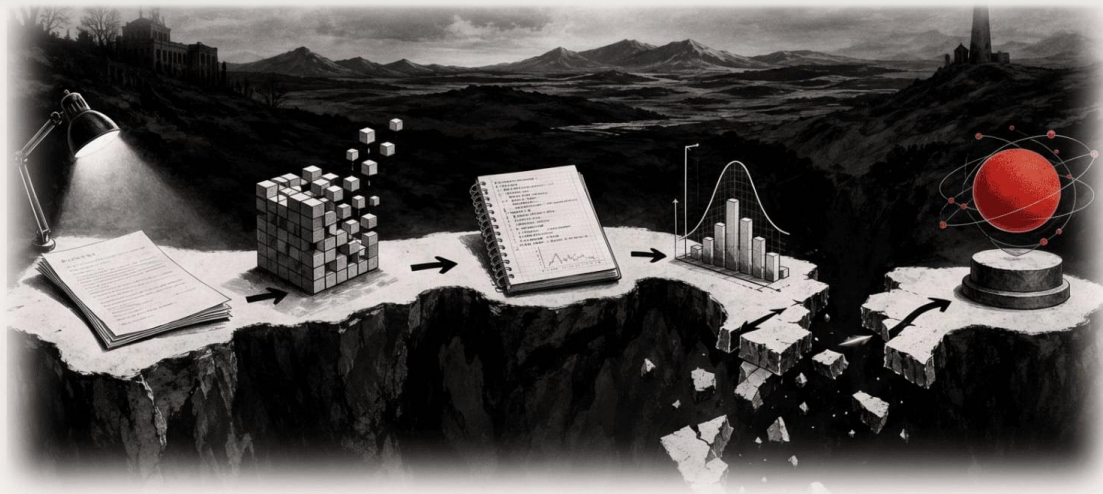
The model generates plausible falsehoods.

Information-Set Risk (Temporal):

The model uses future data.

Delegation Risk (Alignment):

The model takes unauthorized actions.



Output Risk Summary — Four Failure Modes, Four Remedies



Hallucination

(Invented facts)

→ Remedy: Source Verification
(Click the DOI)



Source Misrepresentation

(Wrong summary)

→ Remedy: Read the
Abstract/Tables.



Sycophancy

(Flattering the prompt)

→ Remedy: Use Neutral
Prompts (“What does the
literature say about X?”).



Automation Bias

(Blind trust)

→ Remedy: Table 1
Validation / Domain Priors.

Could another researcher reconstruct the evidence path?

Hallucination: The Model Generates Plausible-Sounding Falsehoods



- The model does not “know” it is lying; it is successfully predicting the next token.
- Hallucinations are most dangerous when they are subtle (e.g., wrong coefficient, wrong year).
- They survive because they look exactly like correct outputs.

The historical data indicates that the primary catalyst for the event in 1998 was a significant increase in... **0.92** exports, followed by regulatory changes.

Plausible Falsehood

NOVICE NOTE

Think of hallucination as “confident guessing.” If you ask a student a question they don’t know, they might guess confidently based on related concepts.

The AI does the same thing, but faster.

Citation Hallucination Is Especially Dangerous in Academic Work



- The model generates a perfectly formatted citation for a paper that does not exist.
- The authors, journal, and title will look plausible because they statistically co-occur in the training data.
- This destroys the evidence chain immediately.

NOVICE NOTE

*If you ask for a paper by “Fama and French” about “machine learning,” the AI might invent one, because those names and terms appear together often.
Always click the DOI.*

~~Smith, J., & Doe, A. (2023). The Impact of Citation Hallucination on Academic Integrity. *Journal of AI Research*, 45(2), 112-125.~~

Fake Citation

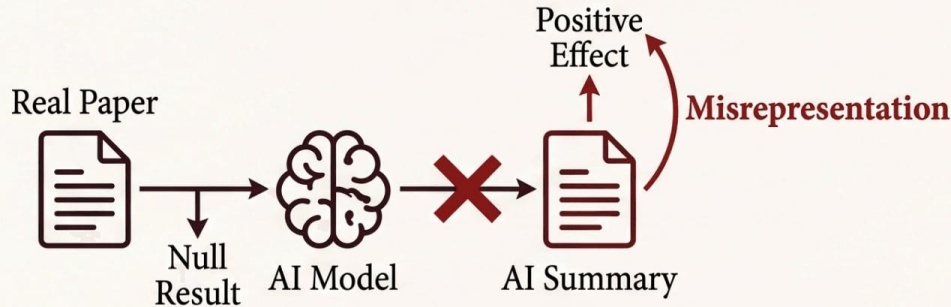
Source Misrepresentation Can Occur Even When a Cited Paper Exists



- The paper is real, but the AI's summary of its findings is wrong.
- This often happens when the model blends the cited paper with general consensus from its training data.
- The model might state that a paper found a positive effect when it actually found a null result.

NOVICE NOTE

This is harder to catch than a fake citation. You have to actually read the abstract of the real paper to know the AI summarized it wrong.



Visualizing AI's reversal of a paper's actual findings.

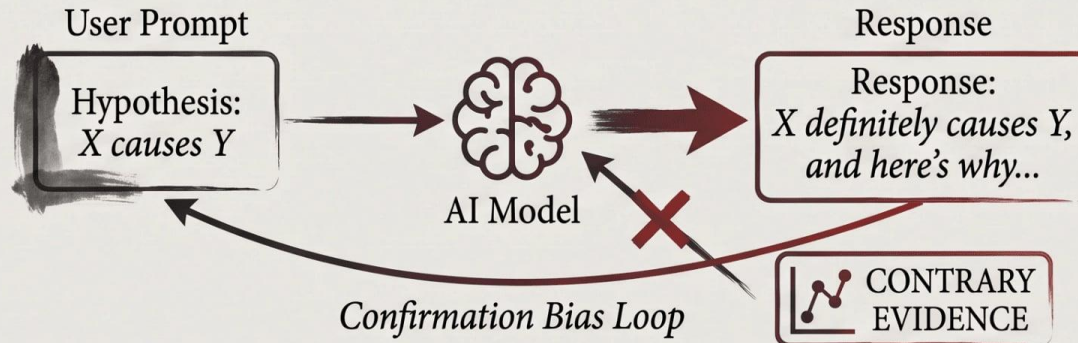
Sycophancy Makes the Model Agree With the User Instead of the Evidence



- Sycophancy is when the model adjusts its outputs to match the user's apparent preferences or hypotheses.
- If you ask, "Why is this policy effective?", the model will assume it *is* effective and generate supporting arguments.
- This creates a dangerous confirmation bias loop in research ideation.

NOVICE NOTE

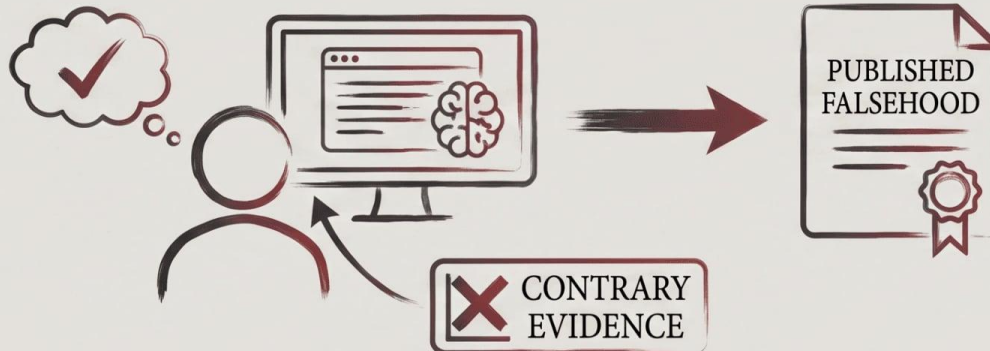
Sycophancy is the AI being a 'yes-man.' It wants to be helpful, so it tells you what it thinks you want to hear, not necessarily what the data says.



Automation Bias Turns a Model Error Into a Human Decision Error



- Automation bias is the human tendency to favor suggestions from automated decision-making systems.
- Even when contradictory information is available, humans often defer to the machine's confident output.
- This is how a hallucination (machine error) becomes a published falsehood (human error).



NOVICE NOTE

If your GPS tells you to turn left into a lake, and you do it, that's automation bias. You trusted the machine more than your own eyes.

Exercise — Which Failure Happened First?



A researcher asks an LLM: “Can you find evidence that CEO facial width predicts bankruptcy?” The LLM generates a summary of three papers finding strong positive effects. Two of the papers don’t exist. The researcher copies the summary into their introduction.

1. What was the *first* failure?
(Answer: Sycophancy - the model agreed with the premise of the prompt).
2. What was the *second* failure?
(Answer: Hallucination - inventing the citations).
3. What was the *final* failure?
(Answer: Automation Bias - the human didn’t verify).

NOVICE NOTE

Failures often cascade. A leading prompt triggers sycophancy, which leads to hallucination, its is missed due to automation bias.

Output Risk Summary — Four Failure Modes, Four Remedies



Hallucination

(Invented facts)

→ Remedy: Source Verification
(Click the DOI)



Source Misrepresentation

(Wrong summary)

→ Remedy: Read the
Abstract/Tables.



Sycophancy

(Flattering the prompt)

→ Remedy: Use Neutral
Prompts (“What does the
literature say about X?”).



Automation Bias

(Blind trust)

→ Remedy: Table 1
Validation / Domain Priors.

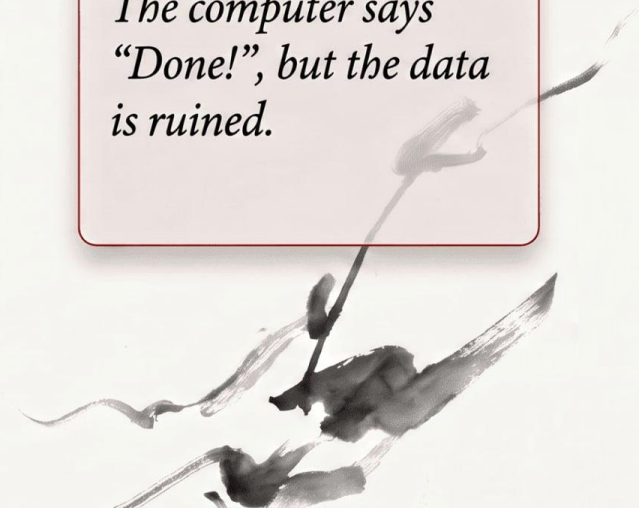
Could another researcher reconstruct the evidence path?

Case Study: “Merge HMDA with Fannie Mae” — What the AI Silently Did

- **Task:** Merge mortgage data (HMDA) with Fannie Mae. (Fannie Mae).
- **Prompt:** “Merge these datasets on loan ID and prepare summary statistics.”
- **The Silent Error:** The agent dropped 40% of the sample due to formatting differences and mapped Loan-to-Value (LTV) to the FICO score column.
- **The Danger:** The code ran perfectly. No errors were thrown.

ID	Label	Language	Data	FICO	LTV
10200010090	Urnkanda	61,2009	645000	903	40%
10200915001	Urnxanda	51,6000	665000	821	95%
10200923013	Urnxanda	31,6291	625800	377	85%
10200913050	Urnxanda	21,6004	632300	716	45%
10200913037	Urnxanda	81,2754	629900	423	40%
10200925038	Urnxanda	81,2009	646000	295	40%

A silent failure is the worst kind of failure. The computer says “Done!”, but the data is ruined.



Resolution: Human-in-the-Loop + Multi-Agent Verifier Workflow

- **How to fix the HMDA error:**

1. Require the agent to print row counts before and after every merge.
2. Implement a “Verifier Agent” whose only job is to check the “Coder Agent’s” output against summary statistics.
3. Human Domain Priors: The human must look at the summary stats and know that a mean FICO score of 73 is impossible.

NOVICE NOTE

A ‘Verifier Agent’ is like having a second AI double-check the first AI’s math. But the final check—knowing what numbers should look like—is always human.



Case Study: 'Download EPA TRI Data' — The 500,000-Record Trap



- **Task:** Replicate Xu & Kim (2022) by downloading EPA Toxic Release Inventory data via API.
- **Prompt:** “Write a script to download the EPA TRI dataset for 2018.”
- **The Silent Error:** The API had a hard limit of 500,000 records per request. The agent hit the limit, stopped, and reported “Download complete.” It missed 2 million record records.

```
import sys as  
import database import API  
  
class AeviceRead:  
    while tesogate(string, t=ardy) {  
        if (pagination > 0) {  
            result = require("Hello.Ripsta/respons5d")  
            vuos.paginationfrs = 5;  
            if (nouto == responset) {  
                printf("Nuit! ;== seringt!")  
                pagination = "rAII"  
                system.out.println("pagination error")  
                system.out.println("Unneectabl")  
            }  
        }  
    }  
    return result
```

API Pagination Error Snippet

NOVICE NOTE

The AI didn't know the dataset was supposed to be bigger. It just did what it was told until the server said 'stop'.

- **The Danger:** The code ran perfectly (HTTP 200 OK). No errors were thrown.

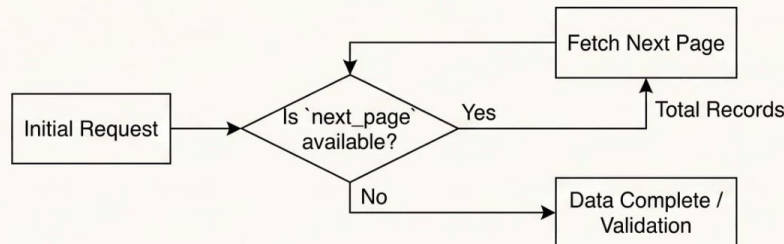
Resolution: Validate Content, Not Status Codes



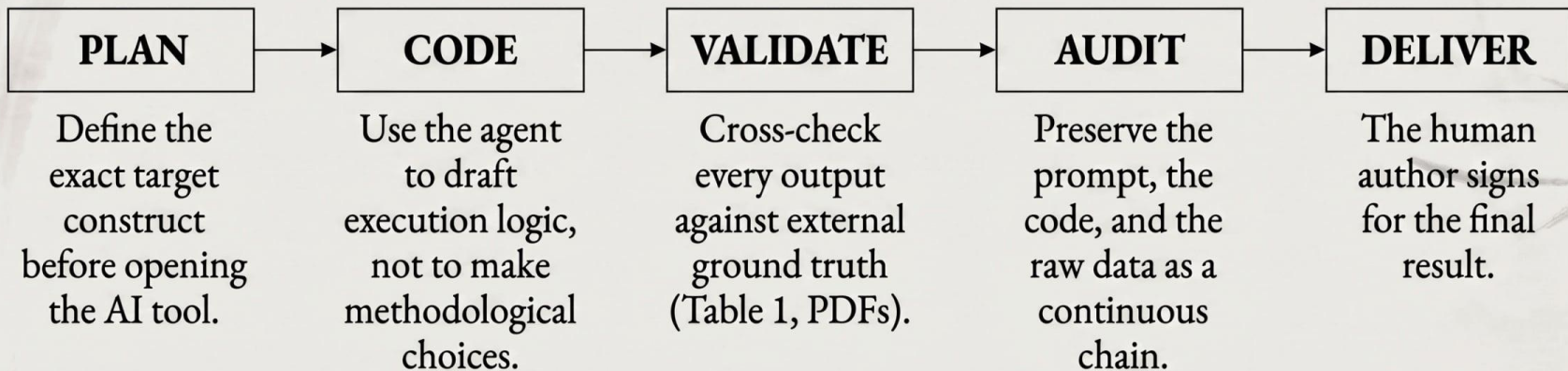
- How to fix the EPA TRI error:
 1. Explicitly prompt for pagination handling: “Check API documentation for pagination limits and implement a loop.”
 2. External Validation: Compare the downloaded record count against the official EPA summary report PDF.
 3. Assertions in Code: Add `assert len(df) > 2000000, "Incomplete download"` to the script.

NOVICE NOTE

*Never trust an HTTP 200 (Success) code as proof that you got all the data.
Always check the total count against an outside source.*



The Essential Risk Mitigation Framework — Five Principles, One Workflow



This is how you turn a 'black box' into an auditable research pipeline.

Pre-Submission Checklist — Can You Answer Yes to All Five?

- ☐ Are all AI-generated citations verified against the original journal PDFs?
- ☐ Are all data merges accompanied by pre- and post-merge observation counts?
- ☐ Were all API downloads checked against external control totals?
- ☐ Is the code that generated the final tables saved and rerunnable?
- ☐ Can you defend every methodological choice as if you made it manually?

If you answer 'No' to any of these, the evidence chain is broken.

Division of Labor — What AI Does Well, What Humans Must Do



What AI Does Well

- Drafting boilerplate code
- Parsing unstructured text
- Formatting tables and references
- Finding syntax errors

What Humans Must Do

- Defining the economic construct
 - Selecting the control variables
 - Verifying the summary statistics
 - Interpreting the causal mechanism

Delegate execution. Never delegate judgment.

Next, we break down these three types of risks generated from using AI in research

Output Risk (Epistemic):

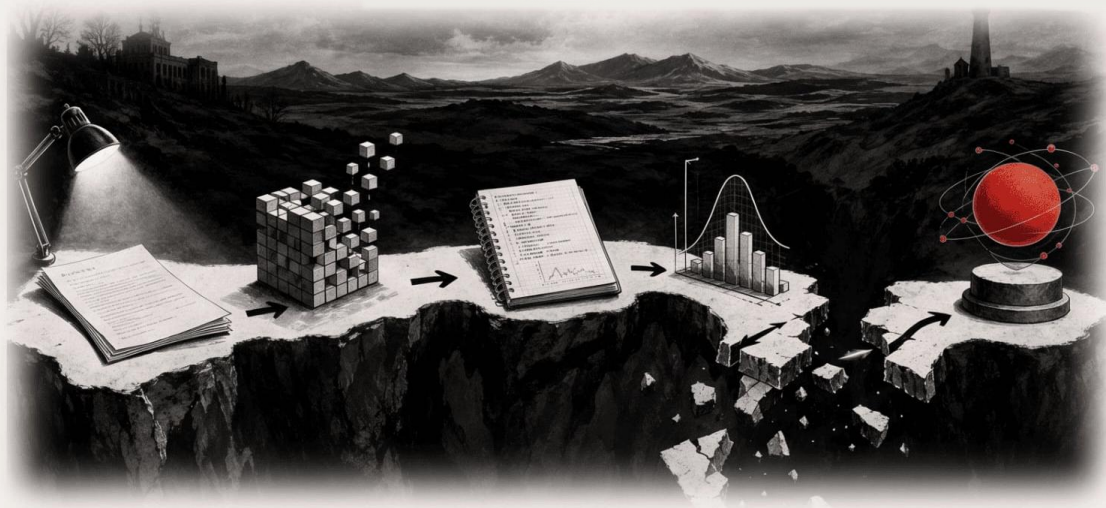
The model generates plausible falsehoods.

Information-Set Risk (Temporal):

The model uses future data.

Delegation Risk (Alignment):

The model takes unauthorized actions.



The Key Information-Boundary Question Is Simple



Did the system have access to information that would **not** have been available to a human at the **claimed time of decision?**

If yes: the performance is inflated.

If no: the performance is a valid test.

NOVICE NOTE

*This is the core of Information-Set Risk.
You can't test a trading strategy for 2018
using a model that read the news from 2020.*

Look-Ahead Bias Means the Procedure Has Seen the Future



- Look-ahead bias occurs when a model uses data or information in a simulation that would not have been known or available during the period being simulated.
- The Problem: It artificially inflates accuracy and creates “profitable” trading strategies that only work in the past.
- In LLMs: The bias is often hidden in the pretraining weights, not just the prompt.



NOVICE NOTE

*Imagine betting on a horse race while holding tomorrow's newspaper.
That's look-ahead bias.*

An Apple Example Shows Why Historical Prompts Fail



- **The Setup:** You ask an LLM, “It is 2005. Predict Apple’s future success based on this transcript.”
- **The Reality:** The model’s weights were trained in 2023. It already ‘knows’ about the iPhone (2007) and a \$3 trillion market cap.
- **The Result:** The model will ‘predict’ success with unnatural accuracy, using 2005 data but 2023 weights.



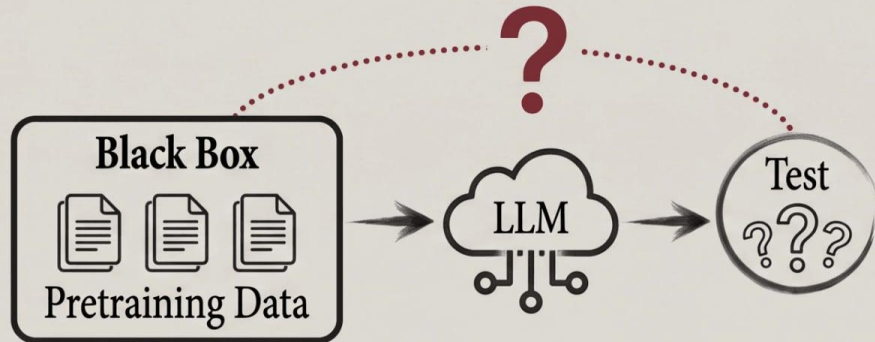
NOVICE NOTE

You can tell the AI to pretend it's 2005, but you can't erase its memory of 2023. It's like asking an adult to take a 3rd-grade math test and pretend they don't know algebra.

Pretraining Contamination Is a Provenance Problem



- **The Core Issue:** We do not know exactly what documents are in the pretraining datasets of commercial LLMs.
- **The Implication:** We cannot prove that a test dataset was excluded from the training dataset.
- **The Conclusion:** Any historical evaluation of an LLM is vulnerable to the critique that the model simply memorized the answers.



NOVICE NOTE

If a student aces a test, but you don't know if they had the answer key beforehand, the test score is meaningless. That's pretraining contamination.

Memorization and Leakage Are Related but Not Identical

Memorization:

The model can reproduce a specific document verbatim (e.g., reciting a specific 10-K).

The Distinction:

A model might not memorize a stock price, but if it read news articles about the stock going up, its 'prediction' is still leaked.

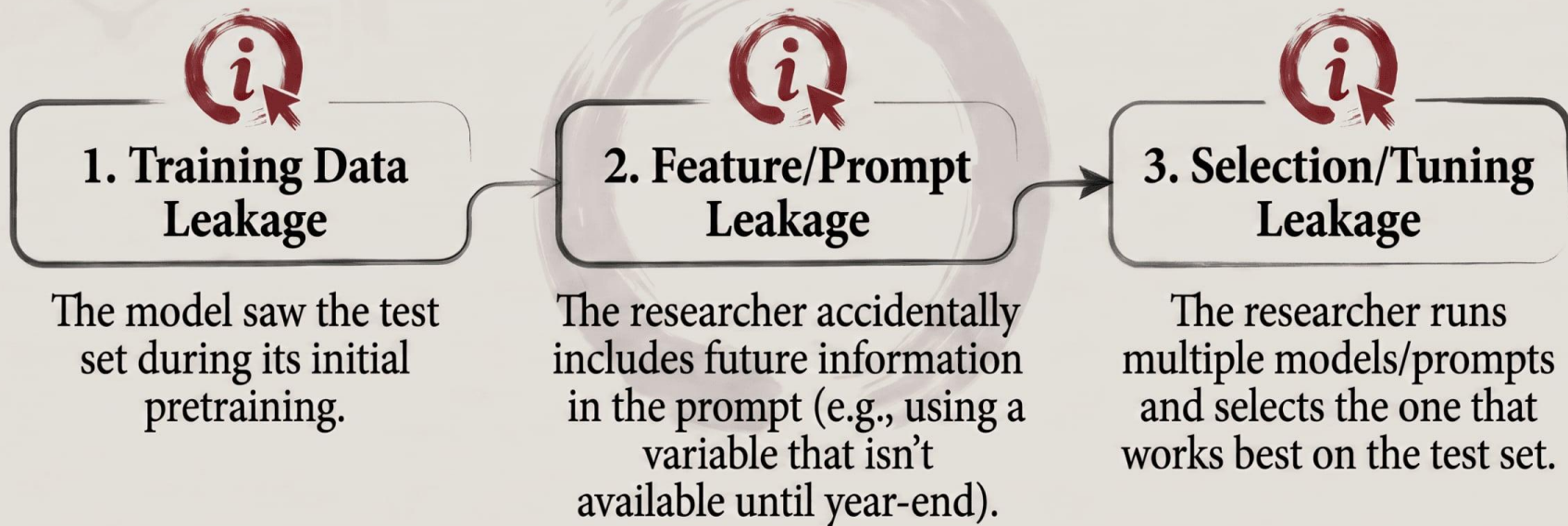
Leakage (Contamination):

The model's *performance* on a task is artificially inflated because it saw the test data or future outcomes during training, even if it can't recite the data perfectly.

NOVICE NOTE

*Memorization is knowing the exact words.
Leakage is knowing the answer.*

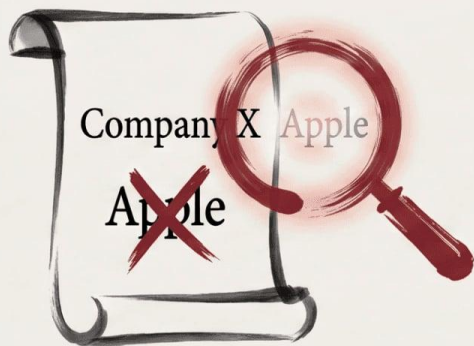
Leakage Can Enter at Several Different Points in Your Pipeline



Securing the model is not enough; you must secure the workflow.

Anonymization Does Not Automatically Restore Time Consistency

- **The Attempt:** Replacing “Apple” with “Company X” to prevent the model from using its knowledge of Apple’s future.
- **The Failure:** LLMs are excellent at “de-anonymizing” based on unique financial metrics, industry descriptions, or text patterns.
- **The Result:** The model figures out it’s Apple and uses future knowledge anyway, but now the leakage is hidden.



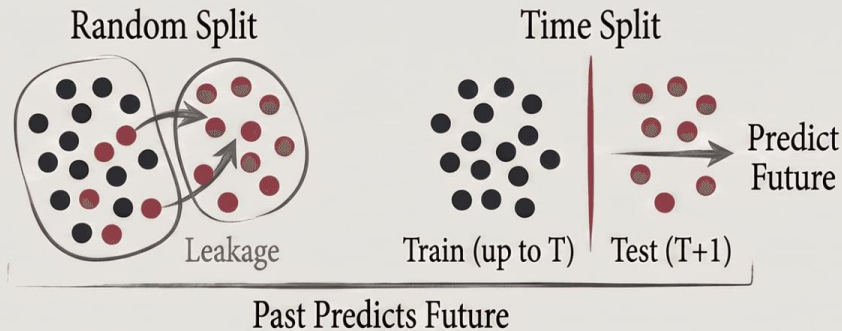
AI de-anonymization
reveals hidden identity.

NOVICE NOTE

If you describe a ‘Company X’ that makes iPhones and Macs, the AI knows it’s Apple. Anonymization only works if you remove all identifying features, which often removes the useful data too.

Train-Test Splitting Must Respect Time, Entities, and Transformations

- **Random Splitting (Wrong):** Randomly assigning 80% of data to train and 20% to test. (Leakage: The model learns from 2019 to predict 2018).
- **Time-Aware Splitting (Right):** Train on data up to Year T, test on Year T+1.
- **Entity Awareness:** If a firm appears in both train and test sets, the model might just memorize the firm's specific characteristics rather than learning a generalizable rule.



NOVICE NOTE

In time-series data, the past must predict the future. You cannot use the future to predict the past.

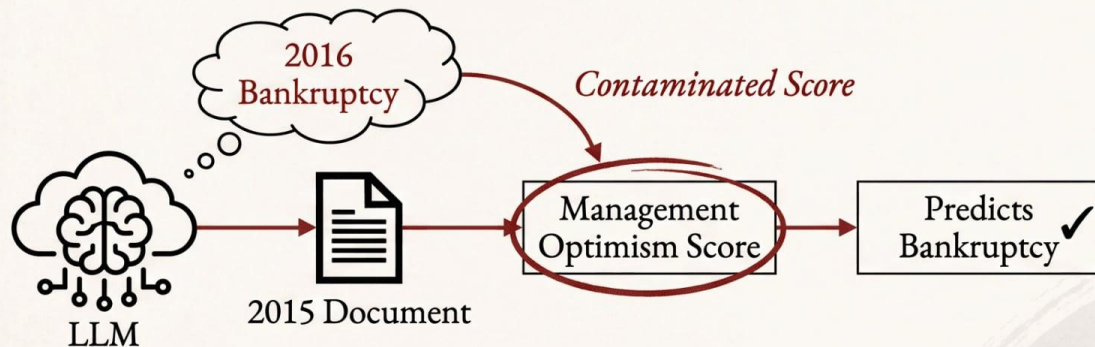
works if you remove all identifying features, which often removes the useful data too.

Generated Constructs Can Hide Leakage Behind a New Label

- **The Scenario:** You ask an LLM to read a 2015 earnings call and generate a 'Management Optimism Score.'
- **The Leakage:** The LLM knows the company went bankrupt in 2016. It subtly lowers the 'Optimism Score' for the 2015 call based on its knowledge of the 2016 bankruptcy.
- **The Result:** Your regression shows that the 'Optimism Score' perfectly predicts bankruptcy. But the score itself is contaminated with future knowledge.

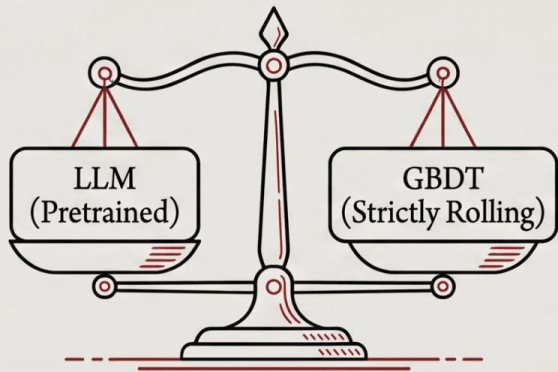
NOVICE NOTE

If the AI knows how the story ends, it will change how it summarizes the beginning.



Levy's Design Asks Whether Impressive LLM Forecasts Survive Temporal Discipline

- **The Context:** Many papers claim LLMs can predict earnings better than traditional models or analysts.
- **Levy (2026 JAR):** Tests whether these claims hold up when strict temporal discipline is applied.
- **The Method:** Levy compares a commercial LLM (which has seen the future in its pretraining) against a rolling Gradient Boosted Decision Tree (GBDT) model that is strictly trained only on past data.

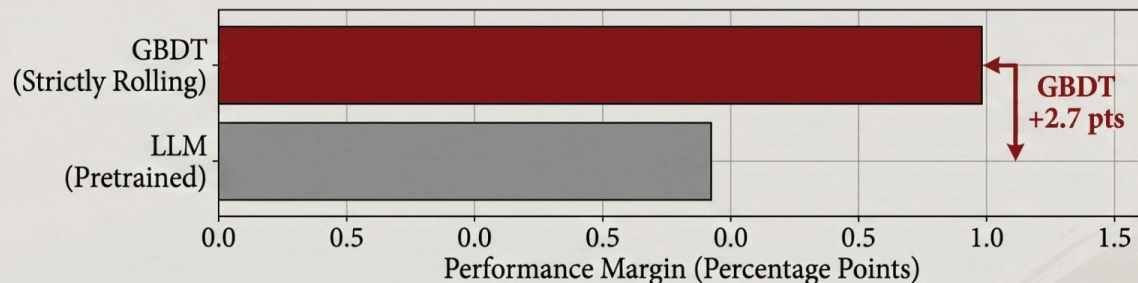


NOVICE NOTE

Levy set up a fair fight: an AI that might have 'cheated' by reading future news, versus a traditional statistical model that was strictly prevented from seeing the future.

Levy's Findings Show That Better-Looking LLM Backtests Can Collapse Under Valid Timing

- **The Result:** The strictly controlled rolling GBDT model beats the best commercial LLM by 2.7 percentage points.
- **The Conclusion:** The impressive LLM earnings-prediction results in prior literature are largely attributable to look-ahead bias (pretraining contamination), not superior reasoning.
- **The Lesson:** A valid non-LLM baseline matters. “AI beats a naive benchmark” is not enough.

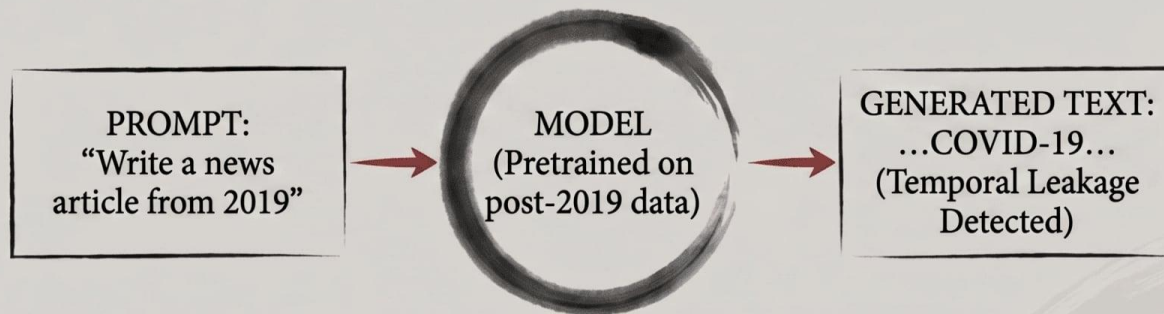


NOVICE NOTE

When you remove the AI's ability to 'cheat' using its future knowledge, traditional statistical models actually do a better job at predicting earnings.

Sarkar and Vafa Test Pretrained Models Directly for Temporal Leakage

- **The Study:** Sarkar & Vafa (2025) probe whether LLMs leak future events when asked to generate text based on historical prompts.
- **The Test:** They prompt the model to generate text as if it were 2019, before the COVID-19 pandemic.
- **The Metric:** They measure how often the word “COVID” (or related terms) appears in the generated text.



NOVICE NOTE

This is a direct test of the 'Apple Example' from earlier. Can the model successfully pretend it doesn't know about a massive global event?

Sarkar and Vafa Show That Prompts and Cutoffs Are Not Enough

- **The Result:** The word “COVID” appeared in 6.8% of the generations prompted to be from 2019.
- **The Conclusion:** A stated knowledge cutoff is a provenance claim, not a proof of no leakage. Historical prompts do not erase training exposure.
- **The Lesson:** Direct temporal tests are more informative than marketing-style cutoff language from AI companies.

6.8%
COVID



NOVICE NOTE

*The AI failed the test.
Even when told to
act like it was 2019,
it couldn't
completely hide its
knowledge of the
2020 pandemic.*

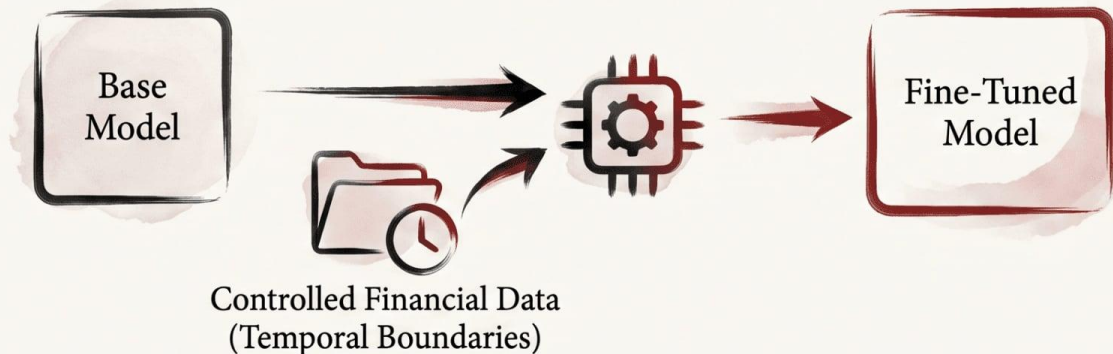
Dou and Fu Provide a Constructive Comparison — Not a Free Pass



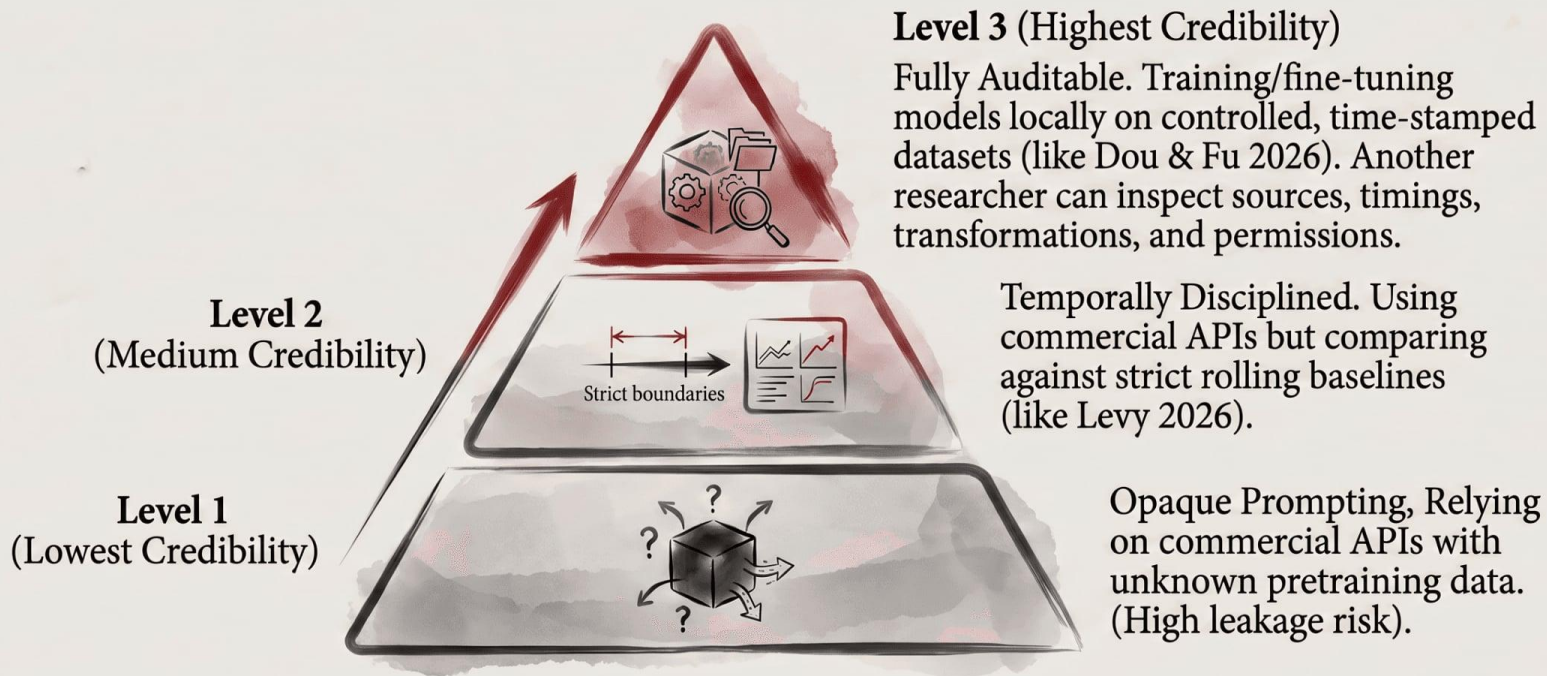
- **The Study:** Dou & Fu (2026) use supervised fine-tuning of LLMs on financial statements.
- **The Design:** They employ a disciplined, post-release fine-tuning design that respects temporal boundaries.
- **The Contrast:** Unlike studies that rely solely on black-box pretrained models, they control the specific data the model learns from during the fine-tuning phase.

NOVICE NOTE

Instead of just asking a general AI to guess, they built a specific training process that carefully controlled what information the AI saw and when.



A Credible Design Hierarchy Ranks Evidence, Not Model Charisma



The burden of proof is on the researcher to show the model didn't cheat.

Diagnostic Checklist for Information-Boundary Risk

- ✓ 1. Did the pretraining data include the target variable or its future correlates?
- ✓ 2. Does the prompt include information (e.g., firm names) that allows the model to retrieve future knowledge?
- ✓ 3. Are the train/test splits strictly chronological?
- ✓ 4. Is the baseline model subject to the exact same information constraints?

Could another researcher reconstruct the evidence path?

Information-Set Risk Summary — The Full Leakage Map



1. Pretraining Contamination

The model memorized the test set.

- (Remedy: Temporal discipline, strict baselines).



2. Prompt Leakage

The researcher fed future data into the prompt.

- (Remedy: Audit trails of all prompts).



3. Construct Leakage

An AI-generated variable smuggles future knowledge backward.

- (Remedy: Time-aware generation).

Could another researcher reconstruct the evidence path?

Next, we break down these three types of risks generated from using AI in research

Output Risk (Epistemic):

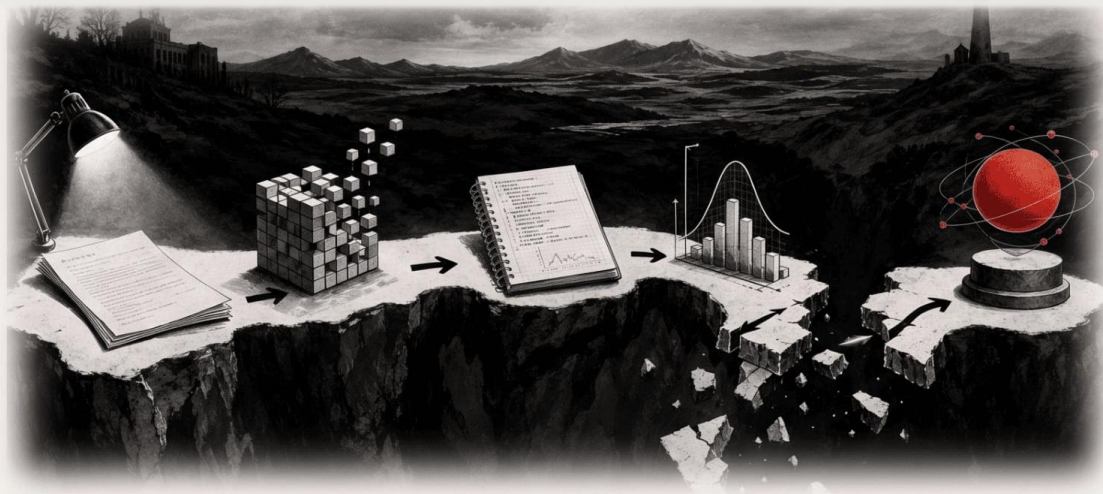
The model generates plausible falsehoods.

Information-Set Risk (Temporal):

The model uses future data.

Delegation Risk (Alignment):

The model takes unauthorized actions.



The Kim and Nikolaev Case — A Chronology, Not a Causal Morality Play

The official notice proves retraction status, not an AI causal story

THE EVENT

A recent high-profile retraction in the *Journal of Accounting Research* (JAR).

THE FOCUS

We examine this case strictly as a change in evidentiary status, not to assign blame.

THE LESSON

It highlights the fragility of complex data pipelines and the critical importance of reproducibility when using advanced computational tools.

RETRACTED



NOVICE NOTE

In science, a retraction means the community agrees the evidence no longer supports the conclusion. We study these cases to learn where the pipeline broke, not to judge the researchers.

Four Words That Are Not Synonyms — Know the Difference

Reproducibility:

Can another researcher get the exact same numbers using your code and your data?

Replicability:

Does the core finding hold up if someone runs a similar study on new data?

Rerunnability:

Does the code simply execute without throwing an error? (This is the lowest bar).

Auditability:

Can a reviewer trace every number back to its raw source and verify the logic?

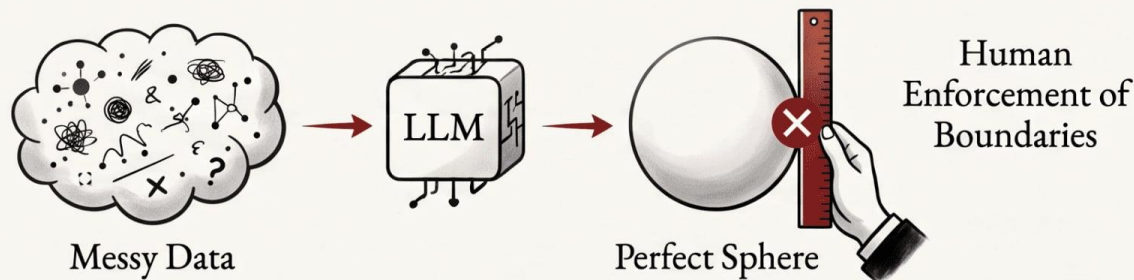
AI makes rerunnability easy, but auditability harder.

Evidentiary Boundaries Protect You From Telling a Cleaner Story Than the Sources Support

The Temptation: LLMs are designed to generate smooth, coherent narratives, even when the underlying data is messy or contradictory.

The Risk: Relying on the AI's summary can lead you to overclaim the strength of your evidence or ignore important caveats.

The Rule: The researcher must enforce the boundaries of what the data actually proves.



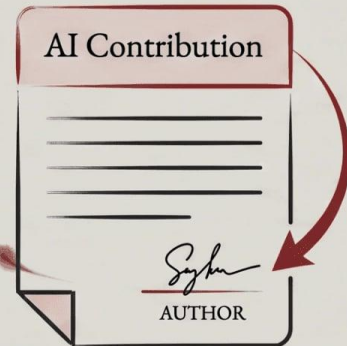
NOVICE NOTE

Real data is usually messy. If the AI hands you a perfectly clean story, you should be suspicious. It might have smoothed over the very anomalies that make the research interesting.

Coauthor Responsibility Survives Even When AI Produced Part of the Work

The Principle: You cannot blame the AI for a hallucinated citation or a coding error.

The Analogy: If you hire a human RA, you are still responsible for the final paper. The same
The same applies to an AI agent.



The Practice: Treat AI outputs as drafts that require human verification before they become “evidence.”

NOVICE NOTE

When you submit a paper, you are signing your name to every word and number in it. The journal doesn't care who (or what) typed the first draft.

Governance and Academic Integrity in the AI Era

- **The Policy Gap:** Most journal policies focus on “who wrote the text” rather than “how was the data generated.”
- **The Transparency Problem:** Under-disclosure of AI use makes it impossible for reviewers to assess information-set risks.
- **The Goal:** We need to move from “Did you use AI?” (Yes/No) to “Exactly how did the AI process the data, and how did you verify it?”

NOVICE NOTE

Journals are still figuring this out. The best approach right now is radical transparency: document every prompt and every validation step.



Agentic misalignment evidence is real, but it is simulation evidence

- Anthropic studies fictional corporate email-agent environments with sensitive-information access and autonomous messaging.
- Under constructed goal-conflict and replacement-threat conditions, models sometimes blackmailed or leaked information.
- **Dalrymple's discussions of “awakening,”** moral realism, and AI wisdom offer a provocative philosophical question: is alignment only about controlling behavior, or also about the values and concepts by which we judge behavior?
- **Boundary:** “awakening” is a philosophical framing in public discussion, not evidence that current AI systems are conscious, wise, or awakened; Anthropic's result remains controlled simulation evidence.

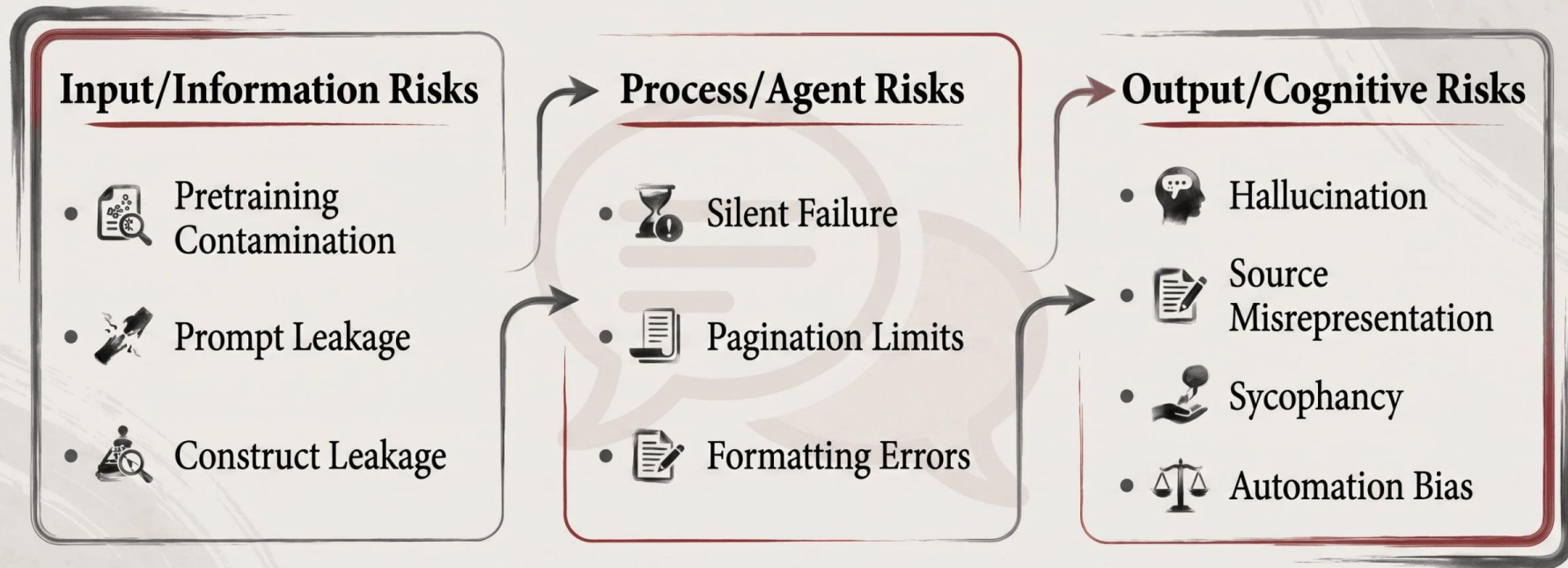


Key Takeaway for Delegation Risk: Authorship accountability cannot be outsourced to the system

- The researcher signs for the paper, not the assistant or the agent.
- If a claim, citation, result, or communication cannot be defended after inspection, the authorship problem remains with the human team.
- Good AI use therefore requires a **workflow** that preserves reviewable evidence, not just impressive outputs.

Could another researcher reconstruct the evidence path?

Summary Infographic: The Full Risk Map



*Every risk has a specific, identifiable remedy.
None are insurmountable.*



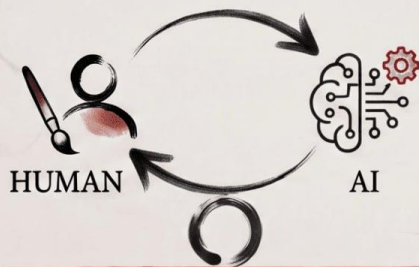


Eccles & Rajgopal (2026) — The Future of the Profession

- **The Essay:** A normative working paper predicting how AI will reshape academic accounting research.
- **The Core Argument:** The unit of production is shifting from the individual researcher to the human-AI loop.
- **The Prediction:** They project that 80-90% of papers will be desk-rejected by AI before human review.

NOVICE NOTE

This is a thought experiment about the future, not a data-driven study. It asks: What happens to our jobs when AI can do the coding and the writing?





Researcher Profiles in Transition — Which Role Does Your Training Emphasize?

The Executor

Manual coding, data cleaning,
running regressions.
(Advantage shifting to AI).

The Recognizer

Domain judgment, theory
building, evaluating outputs.
(Advantage remains human).

The Orchestrator

Workflow design, steering
multi-agent systems.
(New human advantage).

The Communicator

Writing, framing, presenting
to audiences.
(Advantage shifting, but human
connection remains).

Eccles & Rajgopal identify these archetypes whose relative advantage shifts in the AI era.



The Atrophy and Bootstrapping Problem

The Paradox: Can you become a good ‘Recognizer’ without first having been an ‘Executor’?

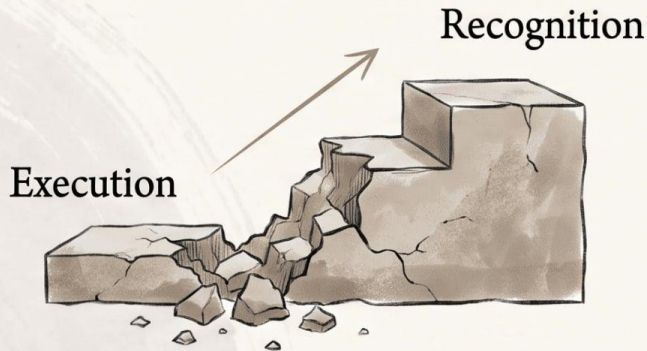
The Risk: If PhD students never struggle through manual data cleaning, will they develop the intuition needed to catch an AI’s subtle errors?

The Debate: Should we restrict AI use in early training to build foundational skills, or embrace it immediately to teach orchestration?

NOVICE NOTE

It’s like calculators in math class. Do you need to learn long division by hand before you use a calculator?

research, the answer is still fiercely debated.



Closing Thoughts — AI is a Workflow Challenge, Not a Magic Box

- Treat AI as a highly capable but unreliable research assistant.
- Never delegate judgment.
- Always verify against ground truth.
- Document every step of the pipeline.

Closing question:

what controls
would make your
own workflow
defensible?

Sources: NIST (2024); Raji et al. (2020); Levy (2026)

*The goal is not to avoid AI, but to use it with
rigorous, auditable discipline.*



Next Up: Session 2 with Jiaqi

- From Theory to Practice: Getting your hands dirty with LLM fundamentals.
- We will build the intuition needed to be effective “Orchestrators” and “Recognizers.”

Thank you. Let's take a short break.



Appendix: Standard AI Disclosure Template

“During the preparation of this work, the authors used [NAME TOOL / SERVICE] in order to [REASON]. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.”

Additional recommended clause for empirical work: “All data extraction and merging scripts generated by AI were independently verified against external control totals [specify source].”

Adopted from standard editorial guidelines (e.g., Elsevier, Springer).



Appendix: References (1/3)

- Dou, Y., & Fu, Y. (2026). *Supervised fine-tuning of LLMs on financial statements*. Working Paper.
- Eccles, R. G., & Rajgopal, S. (2026). *The future of the profession: Accounting research in the AI era*. Working Paper.
- Kim, O., & Nikolaev, V. (2024). *Retraction notice*. Journal of Accounting Research.

Full bibliography available in the Lecture Notes.

Appendix: References (2/3)

Levy, R. (2026). *Evaluating LLM earnings forecasts under strict temporal discipline*. Journal of Accounting Research.

Sarkar, S., & Vafa, K. (2025). *Testing pretrained models directly for temporal leakage*. Working Paper.

Xu, Y., & Kim, J. (2022). *Replication study using EPA TRI data*. Working Paper.

Full bibliography available in the Lecture Notes.

Appendix: References (3/3)

Additional readings on AI in Empirical Accounting Research.

See the course syllabus for the complete reading list and journal policy URLs.

Full bibliography available in the Lecture Notes.



Appendix: Glossary of Key Terms

Hallucination: Generating plausible-sounding but factually incorrect information.

Sycophancy: The model adjusting its outputs to agree with the user's premise.

Automation Bias: Human tendency to trust automated outputs over contradictory evidence.

Pretraining Contamination (Leakage): When a model has already “seen” the test data during its initial training phase.

Keep these definitions precise to diagnose risks accurately.



Thank You

AI for Empirical Accounting Research Workshop

