Historical Inverse Inference: To Understand the World from Text Alone, Machines and Historians Need Prior Knowledge

Manny Rincon-Cruz, Hoover Institution, Stanford University · August 2026

LLMs have achieved excellence in areas where it is easy to "close the loop" with RLVR, especially formal fields like mathematics and SWE. Other fields require that LLMs learn to understand the world from text alone, which requires learning empirical knowledge from an observational — rather than experimental — posture. This is the historian's problem of knowing the causally inaccessible past. In both science and history this is a problem of inverse inference, which can only be solved through the inclusion of prior knowledge.

Many thanks to Brandon Iles for generously orienting me in the machine learning literature, and to Niall Ferguson, Tim Wilson, Mihai Ciucu, and Michael Schmatz for reading and commenting on earlier drafts.

"La physique ne nous donne pas seulement l'occasion de résoudre des problèmes … elle nous fait pressentir la solution." — Poincaré[1]

1 Poincaré, Henri. "Sur les rapports de l'analyse pure et de la physique mathématique." Address to the First International Congress of Mathematicians, Zurich, 1897.

> Introduction — Inverting the Chinese Room

Two thought experiments disproportionately shaped the inquiry into machine intelligence. The first was posed by Alan Turing in his 1950 paper "Computing Machinery and Intelligence," where he aimed to answer the question, "can machines think?" by posing a different problem entirely, "are there imaginable digital computers which would do well in the imitation game?"[2] In the "imitation game," a human and a computer use written text to converse with a human interrogator using natural language. The computer and the human have been provided a story, and the interrogator asks them questions about it. If the interrogator

2 Turing, Alan M. "Computing Machinery and Intelligence." *Mind* 59, no. 236 (1950): 433–460.

cannot reliably discern man from machine, the computer is said to have passed what we now know as the "Turing test."

The brilliance of the imitation game is that it demonstrates how text alone is sufficient to isolate the general features of human (as opposed to animal) intelligence. Turing suggested that a text-based "question and answer method seems to be suitable for introducing almost any one of the fields of human endeavour that we wish to include." Human "thinking" is coextensive with a capacity for language.

The second experiment is found in John Searle's 1980 paper, "Minds, Brains, and Programs," and is posed as a critique of Turing's imitation game.[3] Searle argued that the mere instantiation of a program capable of passing the imitation game would be insufficient for thought and understanding, and thus insufficient to constitute a "mind." His thought experiment uses a human to run such a program. "Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese." The text is, of course, a story, because Searle is now about to play the imitation game with a Chinese interrogator.

To help him answer the Chinese questions about the Chinese story, Searle is given a set of rules in English, which he uses to compose replies. "They enable me to correlate one set of formal symbols with another set of formal symbols, and all that 'formal' means here is that I can identify the symbols entirely by their shapes." Even though Searle answers all the questions correctly, he thinks that "nobody just looking at my answers can tell that I don't speak a word of Chinese." So this "Chinese room" passes the Turing test, even though "it seems to me quite obvious in the example that I do not understand a word of the Chinese stories."[4]

The question of machine consciousness is undoubtedly compelling — and as Anil Seth argued recently in Noema, undoubtedly important.[5] But this question has not produced any engineering advances or shippable code.

And yet, inverting the machinery of these philosophical experiments can help us answer an important technical question about the epistemic boundaries of artificial intelligence. Large language models (LLMs)

3 Searle, John R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–424. Broadly speaking, he is critiquing functionalism or computationalism — that is, the idea that minds are "programs" which can be judged by their behavior alone — as incapable of detecting minds or consciousness.

4 One could reply that Searle's "Chinese room" is an artificial "mind" comprised of both the human and the rulebook, and this "mind" does understand. This is more obvious in the predecessor to Searle's experiment, the 1961 short story "The Game" by Soviet cyberneticist Anatoly Dneprov, where a stadiumful of people act as switches and memory cells to translate a sentence from Portuguese, an idea which was then subsequently copied by Chinese author Liu Cixin in his 2008 sci-fi novel *The Three-Body Problem* as the means through which an alien species creates their first computer.

5 Seth, Anil. "The Mythology of Conscious AI." *Noema*, January 14, 2026.

have achieved excellence in formal fields like mathematics and software engineering — in addition to passing the imitation game. Empirical fields, in contrast, are believed to be more amenable to embodied intelligence or agents capable of continual learning. Instead of asking whether a program inside the Chinese room is conscious, the salient question today is, what can a mind locked in Searle's room understand about the world, if anything?

With a bit of imagination it is easy to see that the inverted Chinese room describes not only the epistemic constraints which LLMs face today — they are "minds" causally separated from the world they seek to understand — it also describes the constraints which historians, since the beginning of their discipline, have faced when understanding the causally inaccessible world of the past.

Given that historians *do* understand many meaningful things about the world of the past by parsing and reasoning about texts, it is obvious that there are two important questions worth asking. How do they do it and how might a computational entity do the same? What does this reveal about the limits of knowledge generally, and LLMs in particular?

The rest of this essay explains that although text is enough to understand the world, the epistemic constraints are identical to those of the age-old problem of inverse inference. I then explore how solutions for inverse inference — in fields as distant as seismography and tomography — require the injection of prior knowledge. This is, in effect, what a historian does. But his methods solve a more difficult version of inverse inference — that is, inverse inference under the conditions of no runnable forward operator, nonstationary causality, and an endogenous observation channel. I conclude by reflecting on whether historical inverse inference is enough for LLMs to understand the world, or whether we will need a different sort of machine intelligence.

> Text-as-Evidence

Does text carry information about the world, independent of someone's knowledge or experience? This question is known as the "grounding problem."[6]

6 The general idea is that words and concepts must point to real experiences or real features of the world to have meaning. Searle opposed Turing's test in part because he asserted that syntax and semantics are distinct, and syntactic mastery offered no real meaning, and thus no real "thought". Forty years later, Bender and Koller offered a more modern gloss when they argued that a hyper-intelligent octopus eavesdropping on undersea cable conversations might become fluent in human language but still not understand what a story with "bears and sticks" is about. Skeptics wonder if words, by themselves, mean anything at all. If they do not, it would mean that "intelligence" generally requires "embodied intelligence." Bender, Emily M., and Alexander Koller. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Grounding is commonly considered an inductive property. Most of us learn the meaning of a new word through definitions, which are themselves just more words. But, the idea goes, at some point we have to use words that have independent grounding through our sensory experience. Without a grounded foundation, the entire edifice of interrelated definitions and language floats freely and means nothing.

A different view is text-as-evidence. Any single word has very little meaning. But it is not zero. Thus, the statistical patterns of a real, textual corpus carry within them an accumulation of meaning to recreate some model of the world that created it. Put simply, a real textual corpus is not arbitrary but is downstream of the world that produced it.[7] Texts are evidence about the phenomena of the world just as much as neutrinos, background radiation, or the variation of old growth forest tree rings.[8]

7 This requires, of course, that we bravely assume that we are not at the mercy of a Cartesian demon, or, in the modern version, trapped inside *The Matrix*. In other words, we must believe that we do not inhabit a world where every piece of evidence is fabricated as part of a conspiracy and thus has no relationship with the outside world.

8 The idea of "text as evidence" generalizes to the entirety of human language. Human language is not arbitrary or random, since each human does not bootstrap their understanding of the world from sensory experience alone. Human language through its structure encodes notions of temporal causality, of objects and actions. All of these are the accumulated residue of billions of years of natural selection, as ancestors that grasped the causal nature of the world were better suited for survival. The structure of the language carries within its statistical distributions these primordial lessons.

> Thinking About Formal, Empirical, Experimental, and Observational Knowledge

The success of the generative pretrained transformer (GPT) architecture suggests that this is more likely to be the correct view on grounding, since GPTs and LLMs gain fluency through exposure and training on text corpora alone, without any sensory experience.

In that case, can we recover anything meaningful from text alone?

It is useful to classify epistemic constraints along two axes. The first is whether a domain is formal or empirical. Formal means that the language used to express statements and facts is also the entire reality of the matter. To be true, they must conform to rules of logical derivation. My winning or losing in chess is determined solely by the rules of chess, just as a mathematical proof is valid if it is logically consistent with mathematical axioms. Empirical means that the truth of a statement is determined by the world, by whether it maps to reality. Observational means that I can observe, but not intervene in the world. Experimental means that I can interact with the world as to query it for new information. The main distinction of experimental over observational information is that experiments allow us to sample reality outside our existing distribution of knowledge.

These axes produce four very different epistemic postures. Experimental empirical knowledge is acquired through interactions with the environment and describes not only scientific inquiry but also how we learn from daily experience. Experimental formal knowledge is learned through interactions with a simulated environment, usually games but also models or reality — think of a flight simulator. Observational formal knowledge in turn means that I can only receive information about a formal domain — say, fragments of a lost language — while observational empirical knowledge means I have information about a world that is now causally inaccessible — think history or astrophysics, which in a way is just history of the universe.

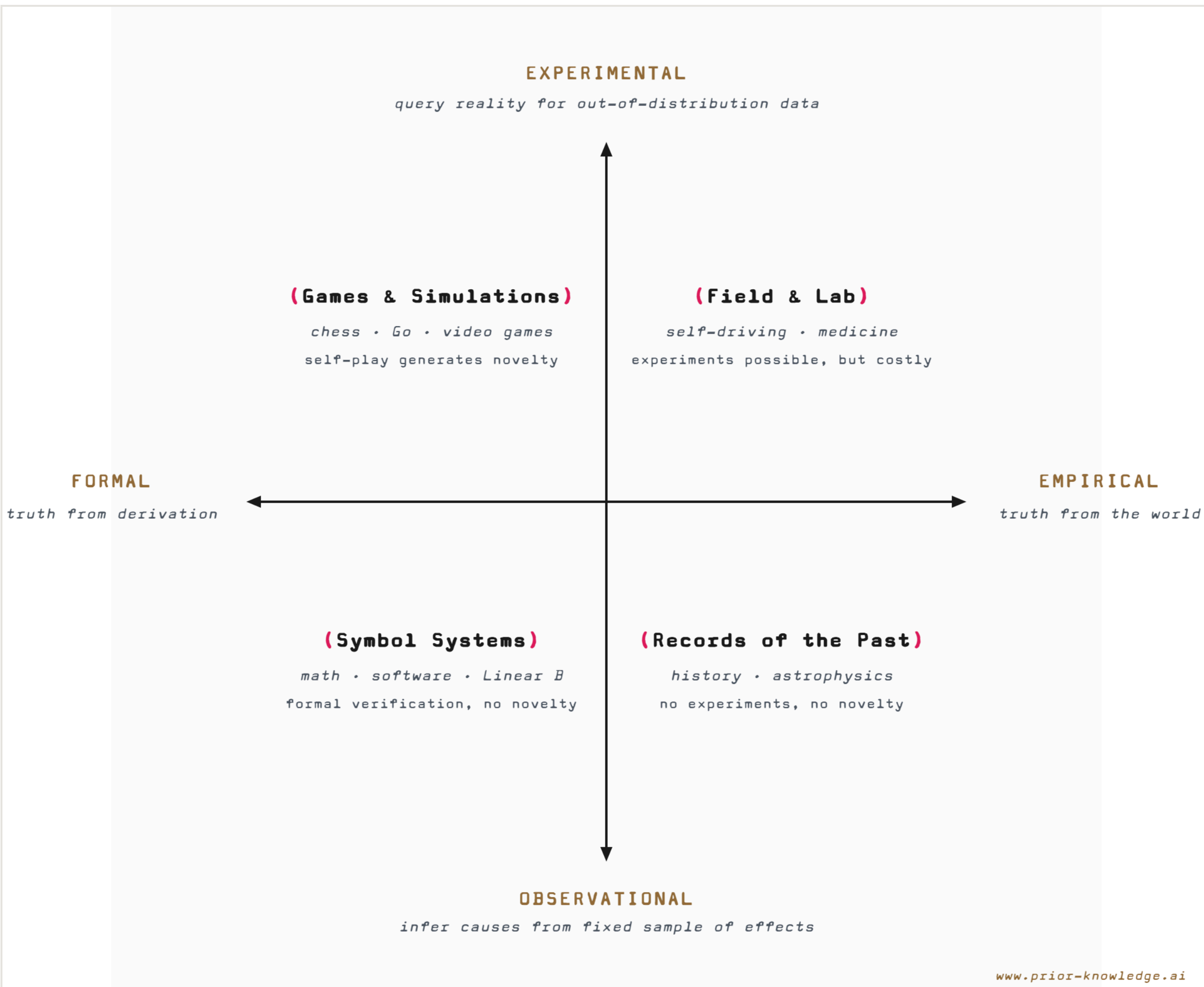

Figure 1. Four epistemic postures. Source of truth — formal derivation, or empirical fit with the world. Environmental interaction — experimental query for new, out-of-distribution information, or observational inference from fixed set of effects.

9 Silver, David, and Richard S. Sutton. "Welcome to the Era of Experience." Preprint of a chapter to appear in *Designing an Intelligence*, edited by George Konidaris. Cambridge, MA: MIT Press, 2025.

Machine intelligence has developed differently depending on which quadrant it targets.

Deep Blue and AlphaGo have solved experimental formal domains like games, because self-play generates endless novelty to feed their learning. Experimental empirical questions have received attention and funding for a variety of approaches. David Silver's and Richard Sutton's "Welcome to the Era of Experience"[9] essay announces that the time might be right for continual learning agents to be deployed at scale, while countless new startups, fueled by declining robotics costs, are making big bets on "embodied intelligence" as the next big frontier. Meanwhile Jeff Dean's new venture, Discovery Loop, promises to automate scientific experimentation, though their core architecture is yet unannounced.

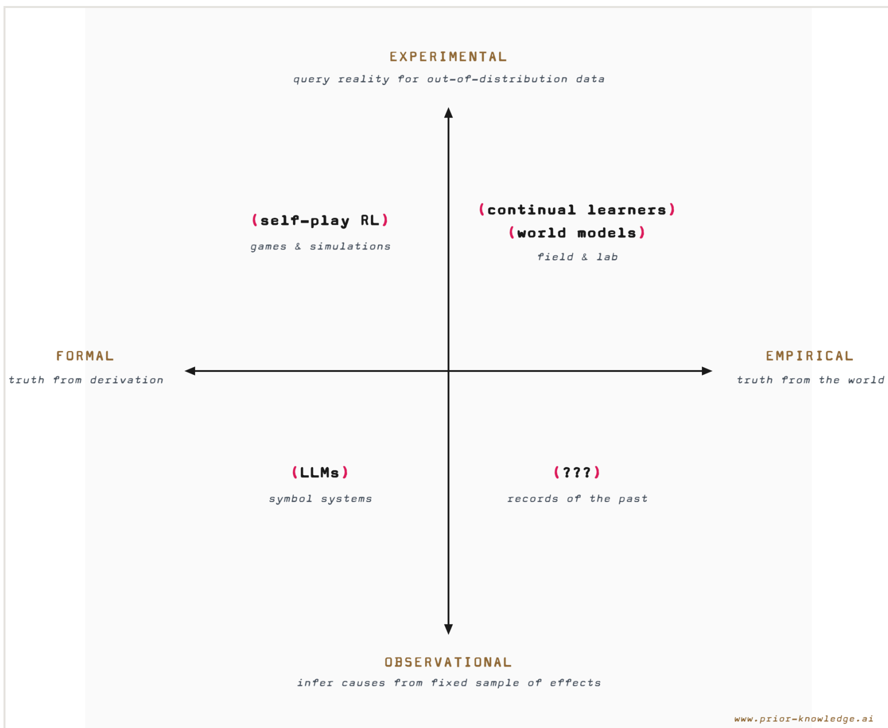

Figure 2. So far, each type of machine intelligence is best adapted to at most one epistemic quadrant.

Today's LLMs excel at observational, formal knowledge — mathematics and software engineering. Over the last few years, LLM capabilities are no longer driven primarily by pre-training scale, but by investment in reinforcement learning from human feedback (RLHF) and reinforcement learning from verifiable rewards (RLVR). There are now a half dozen new companies worth dozens of billions of dollars whose entire business is providing RL environments for LLM agents to learn how to navigate websites or curating mathematical proofs and coding curricula.

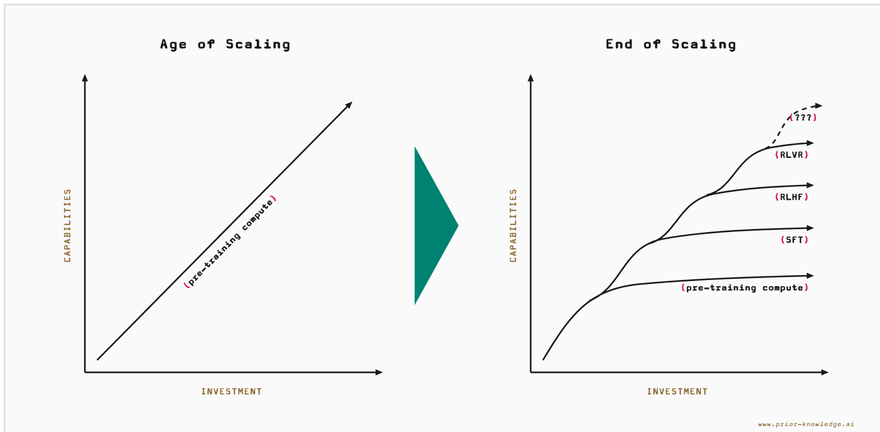

Figure 3. Capability gains for LLMs have stopped coming from pre-training scale and now predominantly from RLVR.

The big question for machine intelligence is whether LLMs can make the move to observational empirical knowledge — that is, records of the past — or whether a new intelligence architecture will be needed.

This epistemic problem is inverse inference — recovering hidden states of the world (what happened, what caused what) from their surviving traces. But this is the hardest quadrant, since it removes the tools we otherwise use to build knowledge. Since it is not a formal domain, we cannot use formal laws to check potential answers. And because the domain is not experimental, we cannot interact with reality to generate new evidence.

LLM training already echoes these limitations. A model trains on a fixed corpus, cannot intervene on the world that produced them, and has no forward model of that world. Training scale gives it fluency — an ability to report of the most likely answer given a corpus — but not an ability to read against the grain or reason against the corpus average. It can report to you the consensus answer in a body of research literature, but will not be able to reason towards a conclusion that the consensus is wrong.

10 Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 2021; Chiang, Ted. "ChatGPT Is a Blurry JPEG of the Web." *The New Yorker*, February 9, 2023.

There are a few reasons to be optimistic that we can solve this problem. LLMs are often dismissed as "stochastic parrots" or "blurry JPEGs of the Internet" that merely recreate the statistical shapes found in their training materials, but that is not entirely accurate.[10] Consider that an LLM trained on transcripts of the game Othello will over time build a recoverable internal representation of the game board and the pieces, even though those are never formally taught or specified (this is sometimes called a "world model" in a loose sense).[11] Yes, the game is one of perfect information and the training data is dense. But that's beside the point. The point is that the LLM recovered *some* model of the world that generated its training data.

Naturally, the real world is far more complex than Othello, so whatever "model" of the world LLMs passively recover is imperfect and incomplete. LLMs often suffer from the "reversal curse," which is when a model trained that "A is B" does not generalize to "B is A." This failure at symmetric fact retrieval limits how reliably parametric knowledge can be queried from arbitrary directions.[12] What might be missing is a framework for active evidential and causal reasoning about the world and so-called "out of distribution" thinking.[13] And if that's the case, then this could be reliably integrated into LLMs.

On the other hand, different types of machine intelligence mostly operate in just one quadrant — LLMs trained on chess game transcripts cannot beat AlphaZero — so the track record suggests we should also look for a new architecture.[14]

> What Is Inverse Inference? Hadamard, Tikhonov, and Seismic Tomography

Inverse inference refers to a specific class of problems in mathematics and physics where, effectively, one tries to recover "hidden" causes from known effects. It was first raised by Jacques Hadamard in a short 1902 article in the *Princeton Bulletin* titled "On problems in partial differential equations and their physical meaning," and later expanded in his 1923 tome, *Lectures on Cauchy's Problem.*[15] In fact, I borrowed the epigraph from his original article for this essay.

11 Li, Kenneth, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task." ICLR 2023. arXiv:2210.13382.

12 Berglund, Lukas, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. "The Reversal Curse: LLMs Trained on 'A is B' Fail to Learn 'B is A.'" ICLR 2024. arXiv:2309.12288.

13 A demonstration that inverse inference is possible by method alone is the decipherment of Linear B. Working from a closed corpus, Alice Kober conjectured in the 1940s that the language was inflected and built a grid of sign-relationships without knowing a single sound value. Michael Ventris in 1952 then guessed that certain recurring sign-groups were actually place-names, and on substituting those values the script resolved into archaic Greek. The confirmation was a tablet published afterwards, which carried the predicted word for tripod beside a picture of one. See Chadwick, John. *The Decipherment of Linear B*. Cambridge: Cambridge University Press, 1958. In contrast, current LLMs fail to decipher the roughly two-thirds of known oracle bone characters that remain undeciphered, and have yet to make progress on scripts like Linear A where the language family is unknown. Guan, Haisu, Huanxin Yang, Xinyu Wang, Shengwei Han, Yongge Liu, Lianwen Jin, Xiang Bai, and Yuliang Liu. "Deciphering Oracle Bone Language with Diffusion Models." ACL 2024 (Best Paper). arXiv:2406.00684, which reports about 1,600 of the more than 4,500 known characters deciphered; Chen, Zijian, Tingzhu Chen, Wenjun Zhang, and Guangtao Zhai. "OBI-Bench: Can LMMs Aid in Study of Ancient Script on Oracle Bones?" ICLR 2025. arXiv:2412.01175.

14 Karvonen, Adam. "Emergent World Models and Latent Variable

Estimation in Chess-Playing Language Models." COLM 2024. arXiv:2403.15498 — GPT-style models trained from scratch on millions of human game transcripts play at roughly 1,300–1,800 Elo while recovering an internal representation of the board; Ruoss, Anian, Grégoire Delétang, Sourabh Medapati, Jordi Grau-Moya, Li Kevin Wenliang, Elliot Catt, John Reid, and Tim Genewein. "Grandmaster-Level Chess Without Search." arXiv:2402.04494 (NeurIPS 2024 as "Amortized Planning with Large-Scale Transformers: A Case Study on Chess") — a 270M transformer supervised on ten million Stockfish-annotated games reaches grandmaster-level blitz and beats AlphaZero's policy and value networks used without search, but distillation of full search-based play remains out of reach.

15 Hadamard, Jacques. "Sur les problèmes aux dérivées partielles et leur signification physique." *Princeton University Bulletin* 13 (1902): 49–52. Hadamard, Jacques. *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. New Haven: Yale University Press, 1923.

16 A *problème bien posé*, or well-posed problem, is defined by Hadamard as one with a solution that exists, is unique, and is continuous with the data. Almost all inverse inference problems fail two of these conditions.

17 Hadamard, *Lectures on Cauchy's Problem*, bk. I, ch. II, §§ 18–21.

18 Tikhonov, A. N. "On the Stability of Inverse Problems." *Doklady Akademii Nauk SSSR* 39, no. 5 (1943): 195–198.

Intuitively, the problem is that although the equations of physics might in theory describe the entirety of reality, in practice these problems are not well-posed (*bien posé*).[16] They are underdetermined, meaning that infinitely many models fit the data, and they are "unstable," such that very small changes in the data can produce arbitrarily large changes in the underlying solutions. He argued that you can't "just approximate the data by analytic functions" because the question is not "whether such an approximation would alter the data very little, but whether it would alter the solution very little. It is easy to see that … the two are not at all equivalent."

Hadamard believed this to be especially pernicious, because "a solution which varies considerably for a small variation of the data is not really a solution in the physical sense… since the physical data are never known exactly, this should imply that the solution is not known at all."[17] In other words, instability allows measurement noise and model error to dominate any solution.

Hadamard discussed distinct inverse inference problems, such as figuring out the temperature inside a metal given its surface temperature, or the distribution of sounds in a room given an array of microphones on a wall. But lest the idea seem too technical, allow me to provide a quotidian example. A few years ago I was a groomsman at my best friend's wedding. We had been roommates at Stanford when we were both pursuing our PhDs — his in natural language processing, mine in history. After the rehearsal dinner, in a cabin above Half Moon Bay, his massive extended family just in from New Zealand gathered to play a card game unknown to me. They offered to explain the rules. But I declined. I wanted to see if I could decipher the rules myself from observing the cards played and who won each round. Needless to say, I failed. The cards I observed being played were compatible with infinitely many rules and hidden hands. And small changes in the cards that were played — perhaps some of the players made mistakes — implied drastically different rules or hidden hands.

But while Hadamard was a pessimist, Soviet mathematician Andrey Tikhonov provided a way to make a problem well-posed by adding extra information and constraints. His classic 1943 example concerns a large geologic body, which can be measured through its gravitational potential at the surface. Small variations in the measurements admit all sorts of different formations and densities for this hidden object. Tikhonov's breakthrough was to inject a prior — to drastically reduce the space of possible solutions by assuming that it had to be compact and smooth.[18]

In short, we have no choice but to rely on our prior knowledge of what realities are plausible to make a reconstruction of reality possible.

A similar set of problems arose in seismic tomography, which is when earthquake or air gun send shockwaves through the earth's mantle and researchers measure the refractions to reconstruct underground geologic structures. The classic inversion equations are stable provided that the earth is spherically symmetric and that velocity increases monotonically with depth. This "prior knowledge" of a layered Earth is not in the data and fundamentally unobservable — since we can't drill that deep — but has to be assumed from theories of astronomy. Once you take for granted the structure of the Earth, then the measurements become perturbations against that structure, and the equations become solvable.

> The Task for Historians

Many contemporary historians are pessimists that believe that causal reasoning about the past is an illusion. The trend is to obsess about social construction, power relations, and the like, which supposedly shape the categories through which we understand the past in arbitrary ways. It thus logically follows that historical reconstruction is both problematic and arbitrary, which opens the door to similarly arbitrary modes of inquiry that prioritize, say, social justice. The historian Hayden White famously argued that the historian made arbitrary literary choices which then determined the outcome of historical research, and no historical narrative was closer to truth than any other.[19]

But historians are, in many ways, Tikhonov's compatriots. Saul Friedländer's famous rebuttal to White asked "can the Holocaust be narrated as comedy?", forcing White to back down from his original argument. In the aftermath Carlo Ginzburg in "Just One Witness" argued that traces of reality survive in the archives regardless of a historian's literary choices.[20] In other words, text is evidence, it is a

19 White, Hayden. *Metahistory: The Historical Imagination in Nineteenth-Century Europe*. Baltimore: Johns Hopkins University Press, 1973.

20 Friedländer, Saul, ed. *Probing the Limits of Representation: Nazism and the "Final Solution"*. Cambridge, MA: Harvard University Press, 1992; Ginzburg, Carlo. "Just One Witness." In ibid.

causal shadow cast by the world. The historian's job is to intuit how to sift the plausible from the merely possible.

> The Inverse Inference Engine

To understand how inverse inference works outside the mathematical sciences, it helps to first develop a schematic of how to regularize these problems. In physics, whose notation I use going forward, these problems are expressed as $d = Gm + \varepsilon + \delta$, where d is the data, m is the model of the underlying phenomenon, G is the forward operator that describes how your model produces the data, ε is your instrument noise, and δ is your model error. Instrument error we can think of as "noise," since our sensors and measurements can never be truly precise. Model error is entirely different and captures the gap between our model of reality and reality itself, since the model is always an approximation. The model error, obviously, is not independent of our model choice. If we only correct for instrument error we get very confidently wrong models of the world.

We can make a problem well-formed by adding extra information and constraints. The list below borrows both from the physical sciences and Bayesian inference, which is more standard in machine learning. Four things can make a problem solvable:

- **Scope or parametrization**: define what kind of thing the model solution actually is — if it is too coarse it won't tell you anything useful about reality, and if it is too fine then it might be too underdetermined
- **Prior p(m)**: what you believe to be true of the underlying models (or reality).
- **Forward operator**: define how G produces the data from a model
- **Observation channel or likelihood p(d|m)**: a measure of how likely your data is for a given model, including both a model of instrument noise and of model error.

To compute a solution one then needs:[21]

21 I am omitting a lot of plumbing that real ML work would include today, but which is irrelevant for the argument.

- **Proposal q(m)**: A way to generate or pick potential solutions to test, since there are infinitely many possible solutions.

- **Calibrated posterior p(m|d):** using a rule $p(m \mid d) \propto p(d \mid m) \cdot p(m)$ that will give you the plausibility of any potential solution being true, you can then create an aggregate picture of what features are most likely to be true, and which ones are most likely to be contingent on quirks of your data
- **Robustness:** check to see which features of your model solutions survive small changes to the prior. If they disappear it means they are artifacts, and most likely not features of reality.

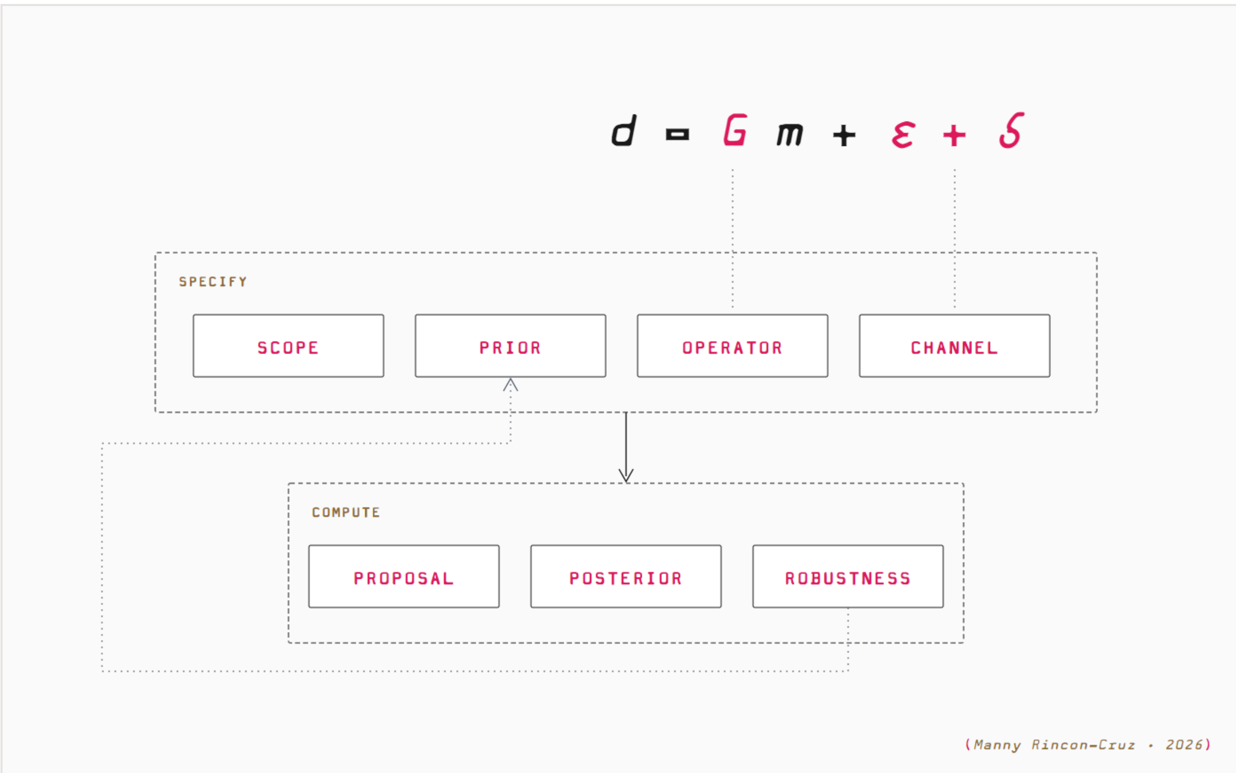

Figure 4. The inverse inference equation and the inverse inference engine that makes these problems well-posed and solvable.

> History as Hard Inverse Inference

History is inverse inference against text. The data is the surviving evidence generated by the world, such as ledgers, inscriptions, journals, receipts, letters, diaries, etc. The model is a description of the world as we believe it was. The forward operator is the causal relationships that explain how different models of the world leave behind historical evidence.

In the well-behaved natural sciences, the forward operator and the observation channel are basically "free." The map from causes to effects is given by physics. The instrument error is independent and the model error is exogenous, so they are "easy" to correct (well, not exactly easy but at least straightforward).

Historians face a far harder problem for three reasons. First, there is no forward operator that can be specified beforehand. For example, we have no direct experimental access to the world of the past, and we have no simulators of the Roman economy or of the Tang court that we could reliably run forwards and backwards to get a sense of how historical events follow from one another. There are no physical or "psychohistorical" laws that we can use to get us from one state of the world to another, or from one state of the world to the existing historical evidence, even at the highest levels of aggregation (notwithstanding the fantasies of Isaac Asimov and most economists). Any hypothesized causal relations are produced by historical inquiry and cannot precede it.

Second, historical causality is nonstationary. The causal relationship between, say, economic classes, or technology and social movements, changes over time. Seismographers use the same wave functions in any part of the world, since the laws of physics do not change between measurements. But even if we found something like a historical "wave function," it could only be applied within a small historical range, or (as we will see soon) in highly constrained cases.

And third, the observation channel is endogenous in the sense that the language of historical evidence and historical reality shape each other. So there is no neutral language for expressing historical observations. For example, concepts historically coined to describe phenomena — say, social classes, census categories, or economic theories — subsequently change the behavior of historical actors, which is often called performativity, social construction, or reflexivity. In a way, this is a type of model error that suffuses all historical evidence and is impossible to minimize because it is also constitutive of historical reality.

The combination of these three obstacles produces what I call "historical inverse inference."

And yet, historians very much manage to recover meaningful models of the past under these constraints. How do they do it? Well, the historian's epistemic process maps to the core of the Bayesian inference engine.

> Historical Inverse Inference — the Method in Seven Parts

22 Cicero, *De oratore* 2.36.

23 Allison, Graham, and Niall Ferguson. "Applied History Manifesto." Belfer Center for Science and International Affairs, Harvard Kennedy School. https://www.belfercenter.org/publication/applied-history-manifesto.

The classical historical discipline (which dominated before the last 50 years) already had tools for solving inverse inference because the discipline was inherently empirical. Its purpose was to guide action in the present — *historia magistra vitae*, history the teacher of life.[22] This traditional orientation today survives only in a minority of practitioners, most frequently financial, diplomatic, and applied historians, as exemplified by Graham Allison's and Niall Ferguson's "Applied History Manifesto."[23]

This is why all of history's useful empirical frameworks can be mapped to the inverse inference engine. Below I will give examples and historical principles to illustrate how these map to different components, but I do not include every possible historical principle or heuristic, or variations by different authors, as notable as they might be. I include just enough to give the reader a sense of the degree to which the historical discipline is an epistemic craft built for solving inverse inference.

24 [citation to come]

25 Maier, Charles S. *Once Within Borders: Territories of Power, Wealth, and Belonging since 1500.* Cambridge, MA: The Belknap Press of Harvard University Press, 2016.

Scope: the historian must first describe the model of the world he aims to recover. Is it a set of relationships between social classes during a pandemic? Is it a detailed understanding of the decisions made by a particular prime minister in a particular crisis?[24] Is it a description of the long-run relationship between technological change and state formation? Is it a mechanism for understanding the rise and fall of territoriality?[25] All these are valid models of the world, but the components must be carefully specified beforehand.

26 Braudel, Fernand. *The Mediterranean and the Mediterranean World in the Age of Philip II*. Translated by Siân Reynolds. 2 vols. New York: Harper & Row, 1972. First published 1949.

For example, Fernand Braudel's *The Mediterranean and the Mediterranean World in the Age of Philip II* is an argument for historical scope to include the slow, almost motionless rhythms of geography and environment (the longue durée).[26] Similarly Karl Marx and Marxist historians such as Eric Hobsbawm argue that historical scope should include the evolving relationship between social classes.[27]

27 Thompson, E. P. *The Making of the English Working Class*. London: Victor Gollancz Ltd, 1963.

Prior: the historian must have a sense of what kinds of reconstructions of the past are most evidentially plausible. One valuable prior is causality. Good history only considers reconstructions and models that contain causal mechanisms, not just surface thematic resonance. Without a causal constraint, an infinite number of stories are compatible with the evidence corpus, and without a causal test they are impossible to weed out.

For example, let us say we find evidence that the municipal bond rates for cities with larger black populations go up over time. A model of "racial capitalism" is inadmissible until we have a causal model and causal evidence for how racism produces those higher bond rates, since the bond rates could be explained by other factors.[28] Another example would be finding that the Protestant areas of early modern Europe go on to experience higher economic growth rates, without a careful causal model of how this would work.[29]

28 During the period Jenkins covers (especially the 1970s and 1980s), the Federal Reserve under Paul Volcker pushed interest rates to historic highs to break inflation. This caused municipal bond pricing to skyrocket everywhere in the United States. "Bondsmen" were price-takers reacting to global macro shifts. The causal arrow flows from global macroeconomics to San Francisco, rather than from local racial structures into the bond market. Jenkins, Destin. *The Bonds of Inequality: Debt and the Making of the American City*. Chicago: University of Chicago Press, 2021.

29 Weber, Max. *The Protestant Ethic and the Spirit of Capitalism*. Translated by Talcott Parsons. New York: Charles Scribner's Sons, 1930.

Forward operator substitute: Because historians cannot re-run the past, nor simulate it to a meaningful degree, we must borrow and use substitutes for the forward operator. This is where analogy and counterfactual reasoning belong.

Analogies allow us to borrow the causal dynamics of structurally similar cases to predict what evidence a certain model of the past leaves behind. Structural similarity here refers to a causal homomorphism — meaning, there is a map that preserves the causal arrows between components in the two cases being compared. Most importantly, we must remember that analogies need not be identities, and that the number of surface similarities is irrelevant.

One famous example is the model of "social revolutions." In *States and Social Revolutions* (1979), Theda Skocpol borrows one causal model and applies it to France in 1789, Russia in 1917, and China in 1911.[30] Her analogies posit that external military/fiscal pressure leads to a breakdown of the central state which then creates an opportunity for peasant insurrection. The borrowed dynamics allows us to look for a

30 Skocpol, Theda. *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. Cambridge: Cambridge University Press, 1979.

"social revolution" including fiscal-crisis records, collapse of landlord control, and peasant land seizures.

Counterfactuals provide the other half of the forward operator substitute, namely the ability to "run" a single forward step.[31] We can test the importance of various factors by removing them, and we can also test the causal relationships across analogies. Robust causes in a model survive many counterfactual tests and become structural features of our reconstruction. Important factors change the outcomes if they are perturbed. Naturally, because historical causality is nonstationary, counterfactuals allow only a local few steps forward, not a full forward simulation.

31 Niall Ferguson is one of the rare historians to have written about counterfactuals. Ferguson, Niall, ed. *Virtual History: Alternatives and Counterfactuals*. London: Picador, 1997.

For example, had Alexander Bell never been born, would we still have seen the rise of the telephone and telecommunications networks? It turns out yes, Elisha Gray filed a patent caveat for essentially the same device at the same patent office on the same day, February 14, 1876. So, counterfactually, Bell is not a "robust" factor in technological change!

Observation channel correction: Most of modern historical writing is entirely about the social construction of evidence, so there is no shortage of writing on the topic.[32] But this is by no means a distinctly modern orientation — Marc Bloch's *The Historian's Craft* is largely about how to "read" sources.[33] A few heuristics are worth highlighting.

32 Hacking, Ian. *The Social Construction of What?* Cambridge, MA: Harvard University Press, 1999.

33 Bloch, Marc. *The Historian's Craft*. Translated by Peter Putnam. New York: Alfred A. Knopf, 1953.

First is conceptual drift and anachronism. We cannot apply the meaning of words today to words in the past. But more confusingly, the entire vocabulary that would define a word also drifts over time — this is conceptual drift. That these webs of concepts drift and change is most evident when it comes to political or religious language. John Pocock's *The Machiavellian Moment* is one such study of just this phenomenon.[34]

34 Pocock, J. G. A. *The Machiavellian Moment: Florentine Political Thought and the Atlantic Republican Tradition*. Princeton: Princeton University Press, 1975.

Second is source criticism. Historical records of the past are not the past themselves, and historians must account for who produced a document, for whom, when, and why. Correlated evidence becomes stronger only if it is genuinely independent.

35 Martin, Terry. *The Affirmative Action Empire: Nations and Nationalism in the Soviet Union, 1923–1939*. Ithaca, NY: Cornell University Press, 2001.

Third, observation channels are endogenous because concepts and society mutually reshape each other. The classic example is that of early Soviet censuses which not only recorded preexisting group boundaries, but also helped strengthen them, necessitating a turn to repression in the 1930s to keep the Soviet Union together.[35] Another example is Tong Lam's argument that statistical surveys in late imperial and early Republican China justified that era's authoritarian statecraft.[36]

36 Lam, Tong. *A Passion for Facts: Social Surveys and the Construction of the Chinese Nation-State, 1900–1949*. Berkeley: University of California Press, 2011.

Model Proposal: Potential solutions must closely track what was actually possible. One tool is Collingwood's idea of re-enactment, which recovers choices actually considered by individual historical actors. Pocock's idea of "available languages" similarly aims to avoid generating choices that are either anachronistic or irrelevant.[37] At the same time the categories used within the reconstruction should appropriately compress the evidence without regard to their being socially constructed. This serves to calibrate the explanatory power of the selected solutions to the right level warranted by the evidence.

37 Collingwood, R. G. *The Idea of History*. Oxford: Clarendon Press, 1946; Pocock, J. G. A. *Politics, Language and Time: Essays on Political Thought and History*. New York: Atheneum, 1971.

Calibrated posterior: The engine's output is a probability distribution over possible solutions and which also presents core structural claims as falsifiable hypotheses with failure conditions. Unfortunately this is not a prevalent practice in historical writing today — historians rarely articulate the evidence which would falsify their theories, except perhaps in a few corners of the economic history debate surrounding Chinese growth before the Great Divergence.

Robustness: Robustness can be best understood as a type of "quality control" where specific components of the reconstruction's prior are perturbed. For example, certain analytical categories (such as gender or social class) are removed and if the causal reconstruction (of say, industrialization) remains more or less the same, then the causal relationships are more likely to be robust and the analytical categories more likely to be artifacts.

> Conclusion — The Big Payoff in Research, Reconstruction, and the Situation Room

It would certainly be amusing if the long-run legacy of the Turing test and Searle's Chinese room is not a measure of machine consciousness, but a measure of machine knowledge. LLMs excel at formal domains with tight feedback loops, but struggle with empirical reasoning from text alone, which historians have essentially mastered.

The historical inference engine will be valuable for research, reconstruction, and the "situation room." The benefits to research are relatively significant. Strong evidential and causal reasoning abilities would allow researchers in the public and private sectors to effectively and reliably search and sift through voluminous research literature to extract robust insights. Agents today recover corpus consensus, but the value of research is most often outside consensus.

Reconstruction is another straightforward, but more limited, application. When a large corpus of documents for an organization exists, it could be used to reconstruct the internal dynamics, incentives, and behaviors of its members. One can imagine using such a corpus and agent to investigate a bankrupt fund for insider trading or fraud.

Lastly, historical inverse inference is useful for policymakers. In the words of Defense Secretary Ash Carter, "The dominant mental methodology of real policymakers is historical reasoning."[38] Policymakers rely on analogies and disanalogies, even if they do so poorly. Analogy generation is an emergent capability of large models but cannot handle counterfactual reasoning robustly to test causal homomorphism, meaning that LLMs today suffer the same issues as naive applications of history —— a fluent ability to produce superficially plausible comparisons but an inability to vet them.[39] So our agent wouldn't generate analogies but instead stress-test them to extract the most significant and plausible causal relationships, exactly the information a policymaker needs to make a decision. For a policymaker this is a reasoning companion, not an oracle.

The biggest challenge to reaping these benefits is that it is not yet entirely clear how to instill the historical inverse inference disposition into an LLM, since it cannot be learnt from experience or from historical writing. But it is not an intractable challenge, I think. Yes, there is a lot of conceptual machinery required to reconstruct the causally inaccessible past. But this conceptual toolkit is itself a type of prior knowledge which humans have passed to each other through the ages, and could transmit forward yet again.

38 Carter, Ash. "Shaping Disruptive Technological Change for Public Good." Belfer Center for Science and International Affairs, Harvard Kennedy School; quoted in "Applied History: What History Can and Cannot Tell Leaders." Belfer Center. https://www.belfercenter.org/collection/applied-history-what-history-can-and-cannot-tell-leaders.

39 Chen, Yongqiang, Guangyi Chen, Yuewen Sun, and Kun Zhang. "Analogical Deep Research: Retrieving and Integrating Historical Analogies for Foresight Analysis." arXiv:2607.13602 (2026), which finds LLM agents match historical analogies on surface features rather than underlying mechanisms; Sourati, Zhivar, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. "ARN: Analogical Reasoning on Narratives." *Transactions of the Association for Computational Linguistics* 12 (2024): 1063–1086, where models fail when surface cues oppose causal structure.

> CITATION

Rincon-Cruz, Manny. "Historical Inverse Inference." *Prior Knowledge*, August 2026. https://prior-knowledge.ai/essays/historical-inverse-inference/.