SIRIUSB IQ

AI DATA SCIENCES LAB & THINK TANK

KINCAID RX DEFENSE PBM CONTRACT X-RAY IQ INTELLIGENCE SERIES

RESEARCH THESIS · ARCHITECTURE & MARKET STRUCTURE

Fiduciary-Grade Artificial Intelligence

An architecture for evidentiary machine reasoning in highly-regulated environments — and the economic case for building it as revenue-generating infrastructure.

Jeremiah Franklin Shrack

Founder — SiriusB IQ AI Data Sciences Lab & Think Tank

Founder, Kincaid IQ · Co-Founder, Kincaid Risk Management Consultants · Architect, Rx Defense PBM Contract X-Ray

WORKING PAPER · SIRIUSB IQ LAB · CICERO, INDIANA

ABSTRACT

The dominant paradigm in applied artificial intelligence optimizes for **plausibility**: fluent, probable output conditioned on a prompt. Highly-regulated environments — ERISA fiduciary administration, healthcare claims, financial supervision — do not consume plausibility. They consume **admissibility**: findings that trace to an authenticated source, reproduce deterministically across runs, carry a verifiable chain of custody, and survive adversarial examination by a counterparty holding the same record. We argue that the distance between plausibility and admissibility is not closed by prompt engineering or model scale. It is an *architectural class*.

We formalize that class as **fiduciary-grade machine reasoning**, defined by five composable properties — groundedness, determinism, provenance, extraction–synthesis separation, and adversarial survivability — and show that these properties map almost one-to-one onto the reliability standards courts already impose on expert methodology. We then instantiate the architecture in the pharmacy-benefit-manager (PBM) contract domain, where information asymmetry is not incidental but engineered, and present the economic thesis for evidentiary infrastructure as a defensible, revenue-generating category for the firms and counsel obligated to consume it.

Keywords: fiduciary-grade AI · evidentiary reasoning · ERISA §404 · provenance · determinism · admissibility · PBM contract forensics · regulated-environment infrastructure

A general-purpose model asked to summarize a contract will produce a confident, fluent, plausible summary. It will also, on a meaningful fraction of runs, invent a clause that is not there, soften one that is, or attribute a number to a source that never stated it.

In a consumer setting, that is an annoyance. In a fiduciary setting, it is the difference between a documentary record that survives discovery and one that manufactures the breach it was retained to detect. The question this paper answers is not *can* AI read a regulated document. It is: under what architecture is the output admissible against the person who wrote it.

SECTION 01 · THE GAP

Plausibility Is the Wrong Objective Function for a Fiduciary

Applied machine learning has spent a decade optimizing a single latent objective: produce the most probable continuation given the context. That objective has been spectacularly successful in domains where the cost of being confidently wrong is low and the consumer of the output is a human who can sanity-check it. It is precisely the wrong objective in domains where the output will be relied upon by someone who *cannot* independently verify it, on behalf of someone who is not in the room.

That second condition is the definition of a fiduciary relationship. Under ERISA, a plan fiduciary acts **"solely in the interest of the participants and beneficiaries"** and discharges duties "with the care, skill, prudence, and diligence" of a prudent expert.[1] The participant cannot watch the transaction. The fiduciary stands in for the participant's absent attention. And increasingly, software stands in for the fiduciary's.

When the software optimizes for plausibility, it imports a failure mode the statute does not tolerate. A prudent process that cannot be reproduced is not a process; it is an anecdote. A finding that cannot be traced to its source is not evidence; it is an assertion. A summary that occasionally fabricates is not a summary; it is a liability with good grammar. The regulated

environment does not ask whether the output sounds right. It asks whether the output can be *defended* — in an audit, before a board, in a deposition, under *Daubert*.[2]

> *The market built AI that optimizes for sounding correct. Regulated environments require AI that optimizes for being defensible. These are different objective functions — and the gap between them is an architecture, not a prompt.*

This paper makes three claims. First, that **admissibility** is a distinct and formalizable property of a machine-reasoning system, separable from accuracy and from fluency. Second, that the property set required to achieve it — what we call *fiduciary-grade* — maps directly onto evidentiary standards the legal system has already spent a century refining. Third, that building systems to this standard is not a compliance cost center but a **revenue-generating infrastructure category**, because the institutions obligated to make fiduciary decisions are obligated to consume the evidence those decisions require.

SECTION 02 · DEFINITION

Defining the Fiduciary-Grade Criterion

We treat a machine-reasoning system M as a function from an admissible source set to a set of findings. The admissible source set S is bounded and named in advance: in our reference domain it comprises the executed contract, the plan's own claims data, and the public regulatory record. A finding is any assertion the system emits as fact. We define M to be **fiduciary-grade** if and only if every finding it produces satisfies five properties simultaneously.

```
DEFINITION — THE FIDUCIARY-GRADE PROPERTY SET Φ
M is fiduciary-grade ⇔ ∀ finding f ∈ M(S) :
  G — ∃ s ∈ S such that f ⊑ s // grounded: entailed by a real source
  D — M(x) = M(x) ∀ runs // deterministic, reproducible
  P — ∃ trace T(f) → (s, transform) // provenance / chain of custody
  Σ — extract(S) ⊥ synthesize() // synthesis adds no facts
  A — f survives adversary(S) // admissible vs. the counterparty
Φ = { G, D, P, Σ, A } — composable; absence of any one voids the grade.
```

The properties are not independent virtues to be traded off; they compose. Groundedness without provenance is unverifiable. Determinism without separation merely reproduces a hallucination reliably. Each property closes a specific failure mode of the plausibility paradigm, and the value of the set is that it leaves no failure mode uncovered.

TABLE 1 — THE FIVE PROPERTIES, THEIR FAILURE MODE, AND THEIR EVIDENTIARY ANALOGUE

PROPERTY	CLOSES THE FAILURE MODE OF…	EVIDENTIARY ANALOGUE
G · Groundedness	Fabricated facts and invented citations — the model asserting what no source supports.	Authentication — FRE 901: a finding must be tied to a record that is what it claims to be.
D · Determinism	Run-to-run drift — the same input yielding different conclusions, which no prudent process can survive.	Reliability — *Daubert*: a method must be testable and yield a knowable, repeatable result.
P · Provenance	Unverifiable reasoning — a conclusion no third party can trace back to its inputs.	Chain of custody — the link from raw record to stated finding is documented and inspectable.
Σ · Separation	Narrative contamination — the prose introducing facts the extraction never found.	Summaries of voluminous records — FRE 1006: the summary must rest only on admissible underlying data.
A · Survivability	Persuasion over proof — output built to convince a friendly reader, not to withstand a hostile one.	Adversarial testing — the finding is constructed against the strongest rebuttal the record permits.

The right-hand column is the load-bearing observation of this section. **The legal system already solved the specification problem.** The reliability factors articulated in *Daubert* — testability, known error rate, reproducibility, general acceptance — are, read as an engineer reads them, a requirements document for a machine-reasoning system intended to produce admissible output.[2, 3] Fiduciary-grade AI is not the invention of a new standard. It is the translation of an old one into architecture.

SECTION 03 · ARCHITECTURE

How the Properties Are Built, Not Promised

A property the system merely intends to satisfy is a marketing claim. A property enforced by the architecture is a guarantee. The distinction matters because the failure modes above are not exotic; they are the *default behavior* of a generative model. Suppressing them requires structure, not instruction.

3.1 Source-bounded reasoning (closed-world enforcement)

Groundedness is enforced by constraining the system to a closed world: the admissible source set S is the only universe in which a fact may exist. Retrieval is bounded to S; generation is conditioned on retrieved spans; and any candidate finding that cannot be anchored to a retrieved span is suppressed before it reaches output, not flagged after. The architectural commitment is that **the model is never permitted to supply from its parameters a fact the record does not contain.** What the contract does not say, the system does not say.

3.2 Determinism and reproducibility

Run-to-run drift is incompatible with a prudent process. Determinism is engineered through fixed decoding, versioned models and prompts, and an immutable record of the exact configuration that produced each finding. The test is operational: the same executed agreement, re-analyzed six months later by a different operator, must yield the same findings — or the divergence must itself be traceable to a documented change. This is the property that converts an output from *an opinion the tool happened to generate* into *a result the tool will generate again under examination.*

3.3 Provenance and chain of custody

Every finding carries a trace: the source element it rests on, the location of that element in the record, and the transformation applied to reach the finding. The provenance ledger is not metadata bolted on for audit theater; it is the primary artifact. A finding without its trace is treated by the system as incomplete and is not emitted. This inverts the usual relationship between conclusion and evidence: in fiduciary-grade reasoning, **the citation is not appended to the claim; the claim is derived from the citation.**

3.4 Separation of extraction from synthesis

The most consequential architectural decision is the firewall between two operations that the plausibility paradigm fuses. *Extraction* is deterministic, factual, and bounded to S: it answers *what does the record say. Synthesis* is narrative and explanatory: it answers *what does that mean.* The firewall guarantees that synthesis can re-order, contextualize, and argue, but cannot introduce a single fact the extraction layer did not certify. Borrowing Breiman's framing of the two cultures of modeling,[4] the extraction layer is held to a predictive-accuracy discipline while the synthesis layer is permitted explanatory latitude — but only over facts the first layer has already proven.

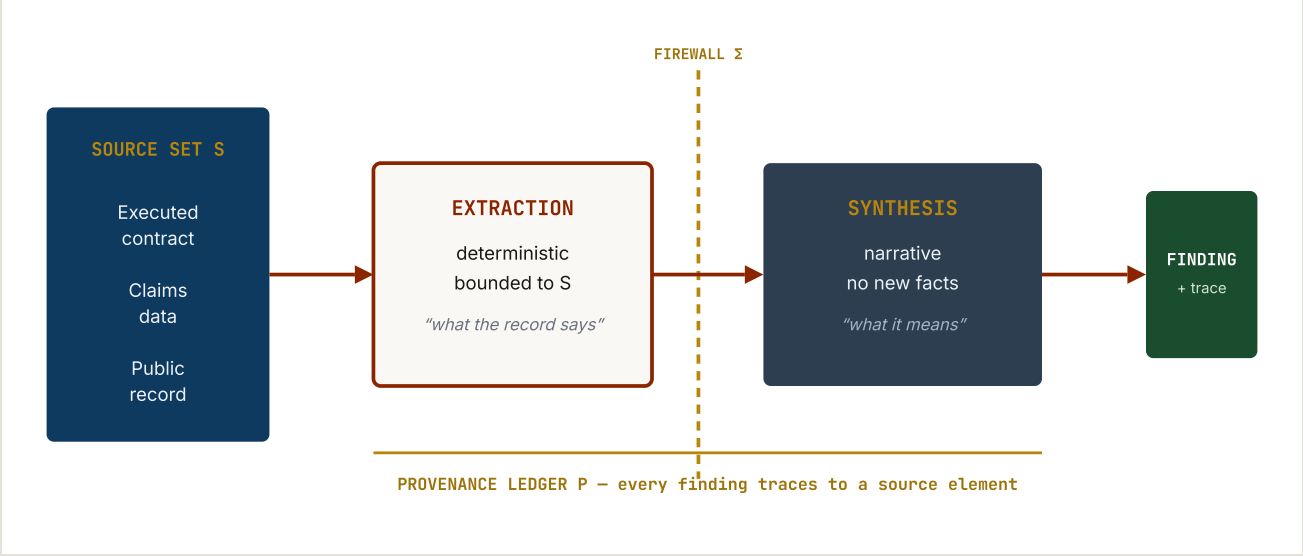

Figure 1. The fiduciary-grade pipeline. Facts may enter only through bounded extraction over the source set *S*. The firewall (Σ) permits synthesis to argue but not to invent. Every emitted finding carries its provenance trace, rendering the chain from raw record to stated conclusion inspectable by an adversary holding the same record.

3.5 Adversarial survivability

The final property reframes the design target. A plausibility-optimized system is built to persuade its reader. A fiduciary-grade system is built to survive its reader's opponent. In our reference domain the opponent is well-defined: it is the counterparty to the very contract under analysis, a sophisticated entity with its own counsel and access to the same executed instrument. Findings are therefore constructed to draw only from sources that counterparty cannot dispute the authenticity of — the document it signed, the data it generated, the record regulators have published. A "Proprietary and Confidential" legend on the agreement is a deterrent to disclosure, not a privilege that defeats a party's right to read what it executed. Survivability is the property that makes the output an asset rather than an exposure.

SECTION 04 · PROOF CASE

The Canonical Instantiation: PBM Contract Forensics

An architecture is validated by the hardest available environment, not the friendliest. The pharmacy-benefit-manager contract is, for evidentiary machine reasoning, close to a worst case — which is precisely why it is the right proof. The PBM Administrative Services Agreement combines extreme contractual opacity, deep information asymmetry between the parties, and economic structures whose effect is difficult to observe from the four corners of the document. Federal Trade Commission staff, in two interim reports, have characterized the largest PBMs as vertically integrated intermediaries whose practices warrant scrutiny for their effect on drug costs.[5, 6] We make no accusation of unlawful conduct against any entity; we observe that the domain's defining feature is engineered illegibility, and that engineered illegibility is exactly the condition fiduciary-grade reasoning exists to penetrate.

The mapping from architecture to domain is direct. **Groundedness** binds every finding to the executed ASA, the plan's own claims file, and the public regulatory record — never to industry lore or the analyst's prior. **Determinism** ensures the same agreement yields the same defined-term analysis whenever re-examined, which is what converts a finding into something a fiduciary can stand behind months later. **Provenance** ties each surfaced mechanism — a reclassification, a carve-out, a benchmark definition, an audit-right limitation — to the clause and the claims line that evidence it. **Separation** keeps the prosecutorial narrative from ever exceeding what the clause-level extraction certified. And **survivability** ensures the resulting file is built to be read by the counterparty's counsel without a single citation it can impeach.

> *The PBM industry maintains an extensive apparatus dedicated to ensuring its contracts are read by procurement teams under time pressure rather than by forensic analysts with subpoena-grade patience. Fiduciary-grade AI is, in one sense, simply the patience — applied at scale, without the billing clock, and without the fatigue that the apparatus is counting on.*

The fiduciary obligations the analysis serves are statutory and specific: the duty of loyalty and the duty of prudence under ERISA §404(a)(1)(A)–(B); the prohibited-transaction provisions of §406; the reasonable-compensation condition of §408(b)(2); and the service-provider compensation-disclosure regime tightened by the Consolidated Appropriations Act, 2021.[1, 7] Recent doctrine has lowered the threshold for surviving a motion to dismiss on prohibited-transaction theories,[8] and at least one prominent prescription-drug fiduciary action remains pending and unadjudicated.[9] The enforcement environment is moving toward the evidentiary record — which is to say, toward exactly the artifact this architecture produces.

THE LITIGATOR TAKE

“The interesting question was never whether the tool is clever. It is whether the file is admissible.”

Speaking candidly, in the register of counsel who has sat on both sides of a fiduciary-breach matter: the discovery battle in these cases is not about whether the conclusion is correct. It is about whether the methodology that produced it is reliable enough to put in front of a fact-finder. A summary generated by a system that drifts run-to-run, cannot show its sources, and may have invented the clause it relies on is not a problem for the defense to rebut. It is a problem for the plaintiff who offered it.

Build the system the other way — grounded, deterministic, traced, firewalled, and constructed against the counterparty’s own record — and the posture inverts. **If these facts are borne out, the question is not whether there is exposure, but how it is quantified.** The architecture described here is, functionally, the difference between a tool that generates an argument and a tool that generates an exhibit. *Only one of those survives cross-examination.*

— COMPOSITE OF ELITE ERISA FIDUCIARY-DEFENSE AND PLAINTIFF-SIDE PERSPECTIVES

SECTION 05 · MARKET STRUCTURE

Why Evidentiary Infrastructure Is a Revenue Category

The strategic error most AI firms make in regulated verticals is to sell *advice* — a recommendation the customer must trust. Advice is a low-margin, high-liability, easily-commoditized product. Fiduciary-grade infrastructure sells something structurally different: **the evidentiary file itself** — the standard of record on which a fiduciary decision is documented. The customer is not buying a conclusion. The customer is buying the artifact that lets them discharge a non-delegable legal duty and prove they discharged it.

This reframing changes every line of the business model. The addressable demand is not discretionary; it is obligated. Self-funded employer health plans cover a large share of the insured U.S. workforce, every one of them governed by a named fiduciary carrying personal exposure under §409, and every one of them sitting on an executed PBM contract that has never been read at clause level by anyone whose incentives align with the plan. The buyer is not persuaded to want the file. The buyer is statutorily required to be able to produce its contents.

```
THE ECONOMIC IDENTITY
value = f( admissibility, not persuasion )
moat = domain evidentiary architecture × reproducible methodology ×
standard-of-record adoption
demand = obligated // fiduciary duty is non-delegable, not discretionary
revenue @ conversion // value realized when the plan acts on the file
defensibility = { executed contract, plan's own data, public record } // no
trade-secret dependency
```

The moat is threefold. First, **domain-specific evidentiary architecture** is hard to build and harder to certify; a general model with a clever prompt cannot reach the property set Φ because Φ is enforced structurally, not instructed. Second, **reproducible methodology** compounds: every file produced strengthens the standard, and a methodology that courts, boards, and counsel come to recognize as the standard of record acquires the defensibility of an institution rather than a product. Third, the analysis is **vendor-agnostic and self-funding in its defensibility** — it draws only from the executed contract, the plan's own claims data, and the public record, so it is never contingent on the customer's relationship with any incumbent, and never dependent on access the counterparty can revoke.

That last point is the one capital-markets readers tend to underweight. The file is the asset. It is not a recommendation that decays; it is a documentary record that *appreciates* as the enforcement environment matures around it. A platform that produces these files at scale is not a consulting practice with software attached. It is the issuer of an instrument that the regulated market is increasingly obligated to hold — the closest thing the benefits market has to an audited financial statement for the pharmacy book.

Sell advice and you are one opinion among many. Issue evidence and you become the record everyone else has to argue against.

THE INSTANTIATION · RX DEFENSE PBM CONTRACT X-RAY

The Architecture, Shipped Against a Live Domain

The thesis of this paper is not hypothetical. **Rx Defense PBM Contract X-Ray** is the fiduciary-grade reasoning architecture instantiated against the executed PBM Administrative Services Agreement — the first production proof that the property set Φ can be enforced at scale in a hostile, opaque, adversarial document domain. It reads the ASA the way the architecture demands: clause by clause, defined term by defined term, exclusion by exclusion, with every finding bound to the source it rests on.

GROUNDED EXTRACTION

Every surfaced mechanism is anchored to a clause in the executed contract or a line in the plan's own claims data. Nothing from lore.

DETERMINISTIC RE-READ

The same agreement yields the same defined-term analysis on every run — the reproducibility a prudent process requires.

PROVENANCE LEDGER

Each finding ships with its trace: clause location, claims reference, and the transformation applied to reach it.

EXTRACTION FIREWALL

The prosecutorial narrative never exceeds what clause-level extraction certified. Synthesis argues; it does not invent.

COUNTERPARTY-PROOF SOURCING

Findings draw only from what the counterparty signed, generated, or that regulators published — impeachment-resistant by construction.

VENDOR-AGNOSTIC CONVERSION

Quantifies extraction from any executed ASA and functions as the migration engine away from extractive structures.

The product is not the opinion that the contract is extractive. The product is the **fiduciary-grade documentary file** that proves it — built to be read by the plan sponsor, the board, ERISA counsel, and, if it comes to it, the counterparty's own lawyers.

SECTION 06 · GOVERNANCE & THE HUMAN STAKES

The Duty Is Owed to a Person, Not an Abstraction

A research paper on architecture risks treating the regulated environment as a clean optimization problem. It is not. The reason admissibility matters is that, at the end of every fiduciary chain, there is a person who cannot read the contract, was not in the negotiation, and absorbs the cost of whatever the document was engineered to obscure — usually a wage earner whose compensation was diverted into a benefit on the promise that someone prudent was watching it on their behalf.

ON STEWARDSHIP

What the Statute and the Tradition Agree On

ERISA frames the duty of loyalty in the language of exclusive benefit: the fiduciary acts solely in the interest of the participant. The older moral tradition that animated more than one American institution framed the same obligation in the language of the priority of labor over capital and the dignity of the person who works — the conviction articulated in *Rerum Novarum* that an economic structure is judged by its effect on those least able to insulate themselves from it.[10] The statute and the tradition arrive at the same place by different routes: the structure exists for the participant, not the other way around.

Fiduciary-grade AI is, at bottom, an instrument for restoring information to the side of the relationship that has been kept in the dark. That is a technical achievement. It is also, when the architecture is honest about whose wage clears the book, a moral one.

SECTION 07 · LIMITATIONS

Threats to Validity

Intellectual honesty requires naming what the architecture does not do. Source-bounded reasoning cannot surface what is absent from the admissible source set; a finding the record does not support is, correctly, a finding the system will not make — which means an incomplete record yields an incomplete file. The architecture surfaces *structural risk and quantifiable economic effect*; it does not adjudicate wrongdoing, and nothing it produces is a finding of unlawful conduct. Determinism is a property of the configured system, not a guarantee against an erroneous source. And the output is decision *support* for a human fiduciary exercising independent judgment with qualified counsel — not a substitute for either.

The paper's economic claims are a thesis, not a forecast; adoption of any standard of record is contingent on an enforcement environment that, while clearly maturing, remains in motion.

SECTION 08 · FORWARD VIEW

The Standard of Record Is the Defensible Position

The trajectory of AI in regulated environments will not be decided by which model is largest or which prompt is cleverest. It will be decided by which systems produce output that can be relied upon by someone who cannot verify it, on behalf of someone who is not in the room, against someone who would prefer it had never been read. That is the fiduciary condition, and it has a specification: groundedness, determinism, provenance, separation, survivability.

The firms that internalize this early will not be selling AI features into a compliance budget. They will be issuing the evidentiary record that the regulated market is increasingly obligated to consume — and discovering, as the enforcement environment closes around the contracts that were never read in time, that the most durable position in a market built on opacity is to be the one holding the file everyone else has to argue against. The architecture is no longer theoretical. The proof is in production. What remains is the question every fiduciary will eventually be asked, and should prefer to answer on their own schedule: **who read the contract, and can they prove what it said.**

References & Authorities

[1] Employee Retirement Income Security Act of 1974, 29 U.S.C. §1104(a)(1)(A)–(B) (duties of loyalty and prudence); §1106 (prohibited transactions); §1108(b)(2) (reasonable compensation); §1109 (liability for breach).

[2] *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993) (reliability standard for expert methodology: testability, error rate, peer review, general acceptance).

[3] Federal Rules of Evidence 702 (expert testimony), 901 (authentication), 1006 (summaries of voluminous records).

[4] Breiman, L. (2001). "Statistical Modeling: The Two Cultures." *Statistical Science*, 16(3), 199–231.

[5] Federal Trade Commission. *Pharmacy Benefit Managers: The Powerful Middlemen Inflating Drug Costs and Squeezing Main Street Pharmacies.* Interim Staff Report (July 2024). Figures attributed to that report.

[6] Federal Trade Commission. *Second Interim Staff Report on Prescription Drug Middlemen.* (January 2025). Figures attributed to that report.

[7] Consolidated Appropriations Act, 2021, Pub. L. No. 116-260 (service-provider compensation disclosure for group health plans).

[8] *Cunningham v. Cornell University*, 604 U.S. ___ (2025) (pleading standard for ERISA §406 prohibited-transaction claims).

[9] *Lewandowski v. Johnson & Johnson*, D.N.J. (pending; allegations unadjudicated; cited as an illustration of the prescription-drug fiduciary theory, not as established fact).

[10] Leo XIII. *Rerum Novarum* (1891) (priority of labor; dignity of the worker; the moral evaluation of economic structures). Cited as editorial and ethical register.

[11] Mitchell, M. et al. (2019). "Model Cards for Model Reporting." *Proc. FAT* '19.* (documentation and reproducibility of model behavior).

[12] Gebru, T. et al. (2021). "Datasheets for Datasets." *Communications of the ACM*, 64(12) (provenance and intended-use documentation for data artifacts).

Jeremiah Franklin Shrack

Founder, SiriusB IQ AI Data Sciences Lab & Think Tank

Founder, Kincaid IQ · Co-Founder, Kincaid Risk Management Consultants · Architect, Rx Defense PBM Contract X-Ray

RX DEFENSE PBM CONTRACT X-RAY