WORKING PAPER

Sovereignty in Institutional Investment

A Framework for Auditable Research and Capital Allocation in the Age of Generative AI

Applied research · November 2026

Thesis. In finance, the distrust of AI is structural: its outputs are opaque, can change from one run to the next, and cannot be fully controlled. A general-purpose integration inherits these properties; no Investment Committee can act on it. Trust is therefore a matter of architecture, designed for the constraints of finance:

1. AI models are confined to drafting and flagging; scores, recommendations and house data come from deterministic engines, so a hallucinated figure cannot reach the record.
2. The same inputs give the same result: every output is reproducible and reconstructible at a given point in time, from source to committee pack.
3. The system fails closed: what cannot be traced to its source is rejected, not approximated.

Abstract

Asset management is currently confronting a structural paradox: while generative artificial intelligence is widely positioned as a competitive imperative to avoid obsolescence and navigate market complexity, its integration into the core institutional investment process remains confined to peripheral tasks.

Current commercial AI paradigms rely on probabilistic workflows and black-box architectures. Deploying generalist conversational copilots or opaque analytic engines introduces unquantifiable risk. In environments where capital is at stake, stochastic approximations and unverifiable inferences represent a structural breach of the prudent person rule. An Investment Committee cannot act on intelligence it cannot audit.

The central thesis of this paper is that the bottleneck for AI adoption in finance is not model quality, but architectural auditability. The prerequisite for systemic trust is deterministic provenance: every analytical node and system output must be fully reconstructible, traced, and provable at a given point in time.

Auditability of this magnitude cannot be achieved through plug-and-play generative wrappers, which merely obscure the mechanics an Investment Committee must be able to inspect. Fiduciary integration demands a governed infrastructure—a running Investment Committee constitution. This requires bounded generative assistance strictly isolated from mathematical scoring and execution, fail-closed operational contracts, and deterministic rules calibrated directly to the fund's proprietary mandate and risk profile. The same properties extend to rules-based indices: under the EU Benchmarks Regulation, a methodology must be transparent, its inputs verifiable, and its history reconstructible, requirements that a deterministic constitution satisfies by construction and that unbounded generative systems cannot.

This paper specifies that framework. It describes the bridge between probabilistic intelligence and deterministic governance, with authority over capital allocation remaining human.

1. The gap between the promises of AI and Investment Committee trust

1.1 When expectations outrun architecture

The root failure is not asset-management-specific. Under pressure to "add AI,"[1] most stacks take a shortcut: wire a large language model to existing warehouses.[2]

That impresses during commercial demos. On the desk it erodes trust: approximations and analytic blind spots compound. Even "for Finance" wrappers that cite sources keep probabilistic inference at the center of the engine. Worse: when generative output is written back into house facts or feature stores without fail-closed gates, hallucinated synthetic fields can poison the corpus.[3] Filings may still be recoverable; derived scores, caches, and decision journals often are not; reconstruction of the contaminated layer can be forced, and fiduciary confidence with it.

The institutional path is different: deployment anchored in the institution's own data architecture, where each step is verifiable[4] and GenAI does not silently author the spine. The problem, and the solution, is not "more AI" or "better AI." It is **how AI is integrated**: deterministic scoring and hard gates first; language assistance only after that admissibility work, feeding coded action rules the house calibrates — and never as a substitute for the spine.[22] Omitting the architecture does not automate research; it scales the same inefficiencies and unverifiable judgments at machine speed.[5]

1.2 The agentic myth and the coordination tax

Markets celebrate "agentic" systems, and an "agentic era" that often names orchestration theater more than architecture, under a broader pressure to "adopt AI" or fall behind, amplified by vendors, often without a defined job for the system. Networks of models plan, tool-call, and debate toward better answers. In finance that is a fiduciary non-starter when the path itself is the risk object: auditing one probabilistic agent is already hard; opaque multi-agent choreography multiplies exposure.[6] Velocity of generation is not the priority. Robustness of output is.

Rising benchmark scores are treated as license to defer clean data and architectural discipline. That does not close the trust gap. A generative or agentic stack can draft queries, summarize filings, propose pipelines, even emit artifacts that *look* like lineage; yet fluency does not *constitute* reconstructible capital process.[7] Point-in-time knowability, fail-closed refusal, Live≡Replay, and an Investment Committee trail must hold *by construction* in house-owned contracts, not as hoped-for behavior of a more capable black box. Stronger models and AGI framing do not retire that requirement: capability is not constitution. Agentic assistance may sit inside the constitution; it cannot replace it without moving trust into auditing the choreography. The durable order is foundations first (clean data and a modular architecture), then innovation on that spine.[9] Adoption under FOMO buys speed that later has to be unwound. An architecture-first path remains ready as models improve: the constitution holds; assistance layers (better models, analyst copilots) can change.

1.3 Structural limits of generic approaches

Public catalogs and commercial peers already ship execution engines, data connectors, ML curricula, agent demos, and frontier financial-services copilots at scale. What they generally omit as a *product* is a running Investment Committee constitution: reconstructible decisions, bounded assistance, and a human wire.[8] The patterns below are category limits, not a vendor scoreboard. Serious copilots can *complement* a constitution; they do not replace it.

Market approach	Structural limit
Conversational / NL → "reports" (model on the warehouse)	Unstable outputs; probabilistic inference at the center; generative write-back can poison house facts and force rebuild of the derived layer
Enterprise AI assistants for financial services	Enterprise chat and agent research UX with licensed data connectors, citations, firm templates, and enterprise IT controls. Capital process remains with the user after generative output (§7)
"Adopt AI" under FOMO (amplified by vendors and social media)	Adoption without a defined job; clean data and architectural discipline deferred; speed that later has to be unwound
Multi-agent "AI fund" / agentic demos	The agent path *is* the decision; fluency can edge toward execution; popularity in this segment often runs inverse to committee governance fit
ML-process curricula / AI quant platforms	Strong pedagogy or staged forecast→decide→execute research records, not a name/book merge, desk artifacts, or published kill registry for an IC
Automated research–implementation loops	The model proposes and implements factors; the loop is the product, with no contract evaluation ring-fenced from production and no human capital gate
Execution, data, and document-search stacks	Necessary adjacent layers (venues, connectors, terminals, document intel): a different job; they sit under governance and are not the decision architecture
Generic / retroactive history	Look-ahead bias: analysis on revised history; as-of claims that cannot survive audit
Benchmark fluency as proof of readiness	Next-token competence mistaken for reconstructible capital process

1.4 Architecture-first response

The durable order is foundations first (clean data and a modular architecture), then innovation on that spine.[9] Assistance layers can change (stronger models, new analyst products) while the constitution holds. The principles below are what category surveys show the market under-supplies as a combined system.

Principle	Implication
Spine owns truth	Point-in-time house facts dated by real-world knowability; AI does not silently author the store
Fail-closed	Missing, broken, or unknown inputs refuse a clean pass, the opposite of "verify after fluent prose"
Determinism	Fundamental scores and hard gates from coded rule engines; risk math stays isolable and replayable[10]
Bounded assistance	Language layers read documents and may emit stance, flags, and action suggestions. *Coded* merge and book-overlay rules — calibrated with the Investment Committee — may demote, nominate promote, or propose reallocation. Fluency never authors score math and never wires capital; PASS ≠ wired[22]

Principle	Implication
Live ≡ Replay	Identical inputs produce identical hard-path scores (classify · gates · coded action rules on recorded stance/flags); an Investment Committee can reopen the trail months later
Ring-fenced evaluation	Contract checks on frozen fixtures and labeled process outcomes: anti-*process* drift, not an "AI beats the market" leaderboard
Gate memory	Pre-declared gates and first-class KEEP / PARK / KILL (or equivalent) so failed bars are recorded, not buried
Human gate	Documented proposals only; clearing a gate is not an order; the Investment Committee allocates capital[11]
Model-agnostic spine	Ready as models and analyst copilots improve: swap or add assistance layers without rebuilding fiduciary process

2. Decision topology: restoring the responsibility line

Generic AI fails at topology.[7, 9, 22] The figures below show the difference between an AI-centered decision path and an institutional architecture where calculation is central, admissibility is decided before language inference, and assistance cannot expand the feasible set.

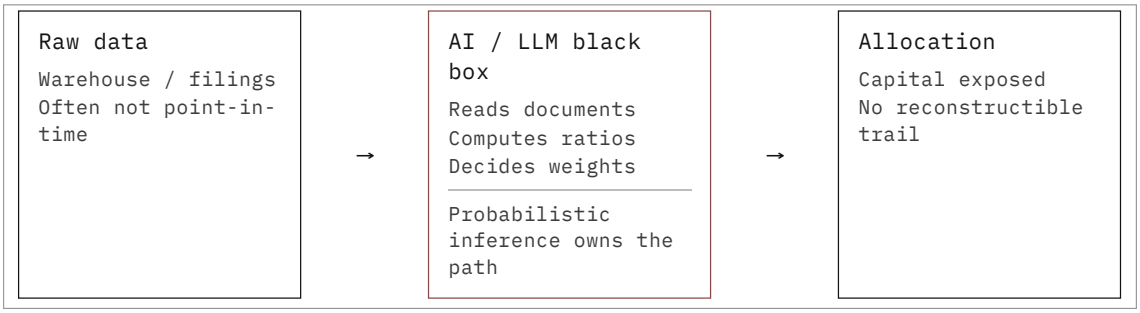


Figure 1 · Market default: AI at the center

When generative inference sits at the center of the path, ungrounded figures can be produced and then treated as inputs to ratios and weights.

The invent–interpret–allocate loop is not isolable: no hard path to refuse invented fields, no reconstructible IC trail, and capital is directly exposed.

3. Research & risk path: scoring for rescore and reallocation

The architecture is specified for daily Investment Committee process. What matters is **deterministic scoring**: coded engines that the desk can apply to propose a rescore or a reallocation under point-in-time rules.

3.1 Scoring, then proposals

Classification and fundamental scores (quality, valuation, resilience, hard gates) run in code, not narrated into existence by a model. Merge and book-overlay instruments are also coded rule engines: they consume those deterministic outputs and, where the Investment Committee wires them, language-layer stance and flags that may *demote*, *nominate promote*, or *propose reallocation*. Outputs are proposal classes and sizing suggestions for the committee. Replay against baselines

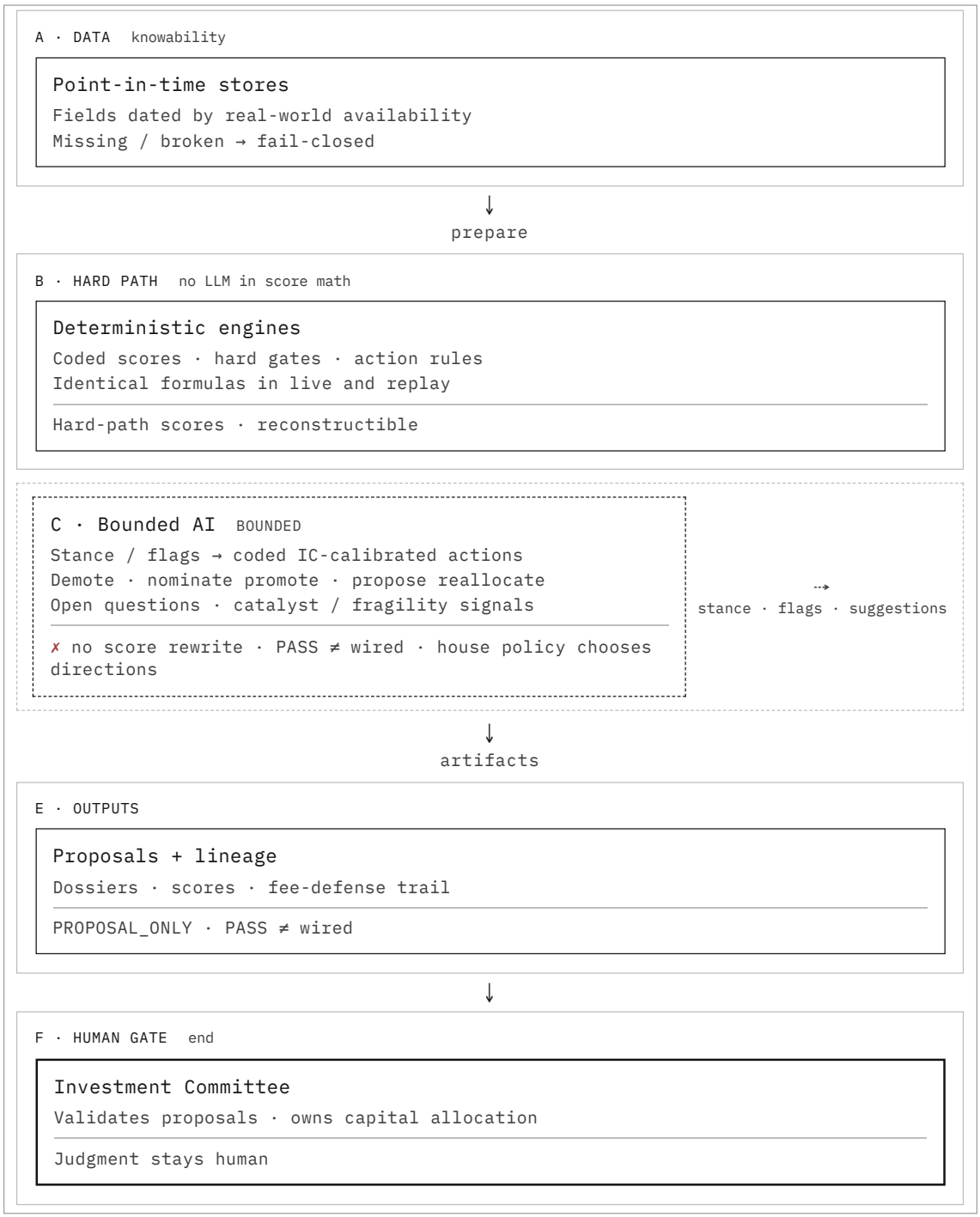


Figure 2 · Precept: hard path, bounded AI, and human gate

Point-in-time stores and coded engines own capital math. Missing or broken evidence fails closed rather than being filled by language generation.

Bounded assistance may feed recorded stance and flags into *coded* action rules the Investment Committee calibrates: demote, nominate promote, or propose reallocation. It does not invent hard-path scores and does not wire capital.

yields labeled process outcomes on the mandate shape: evidence for the desk, not a marketed return curve.[22]

Scoring methodology is **calibrated to the fund** (investment philosophy, risk profile, mandate constraints, and house gates), not a one-size catalog of weights.[12] That is why Precept is a framework developed with the desk, not a GenAI plug-and-play score pack. The same engines may run at **name** level, **book** level, or both; scope is a mandate choice, the constitution is the scoring.

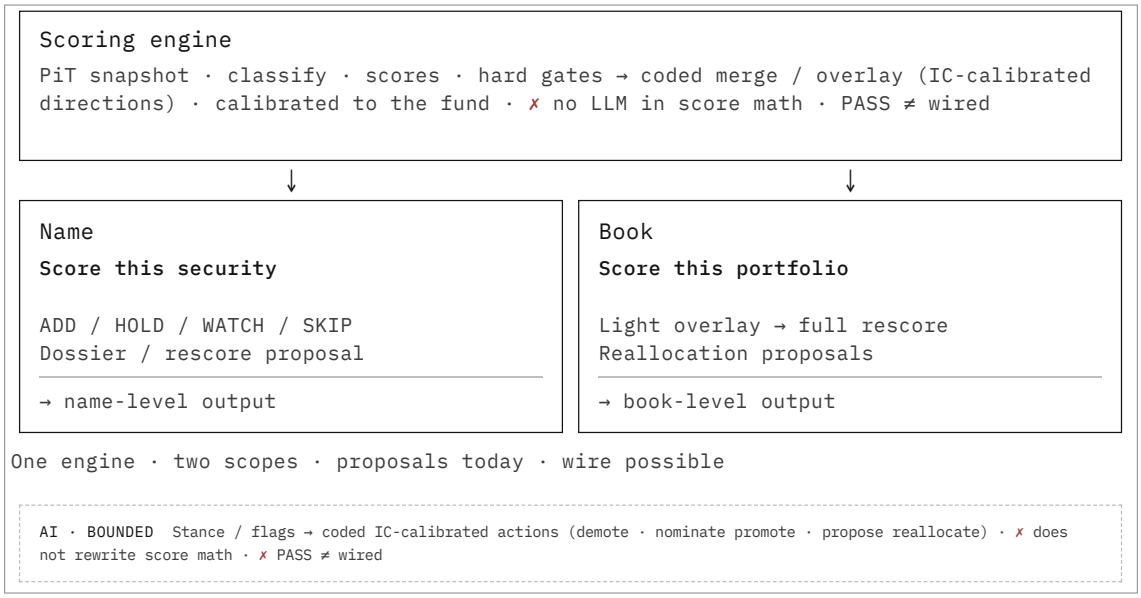


Figure 3 · Scoring engine ⟶ name and book scopes

Point-in-time facts feed coded scores shaped to the house mandate. Applied at name level, book level, or both. The path today ends in committee proposals; the same architecture can extend to a capital wire when that product path is chosen.

Assistance may flag, suggest, and supply stance for coded IC-calibrated actions (demote, nominate promote, propose reallocate). Score math stays coded: LLMs never write those coefficients and never wire capital.

3.2 Point-in-time and fail-closed

Look-ahead bias is not an obscure edge case. It is one of the most common ways research "works" in sample and fails under fiduciary replay: restated fundamentals, late filings, and revised consensus enter a historical score as if they had been knowable on the decision date. The failure mode has been documented for decades on standard accounting databases.[13] Recent work shows that ongoing re-standardization of common Compustat products itself introduces look-ahead and undermines replication unless point-in-time Snapshot files are used.[14]

Major data vendors sell separate point-in-time / Snapshot products precisely to "avoid look ahead bias," evidence that the default restated extract is not knowable-as-of.[15] Serious quantitative shops treat that contract as table stakes. Much of the broader market still does not: warehouse "latest" joins, notebook backtests, and generative layers on live tables often skip a house `filed_at ≤ as_of` rule on every feature used for capital judgment. Even large open ML finance stacks warn that using only the latest amended value causes data leakage and document a dedicated point-in-time store to keep online trading and historical backtesting aligned.[16] Frontier financial-services copilots can cite sources and accelerate memos without shipping that constitution: a different product job, and a structural gap when the Investment Committee must reopen *what was known when.*

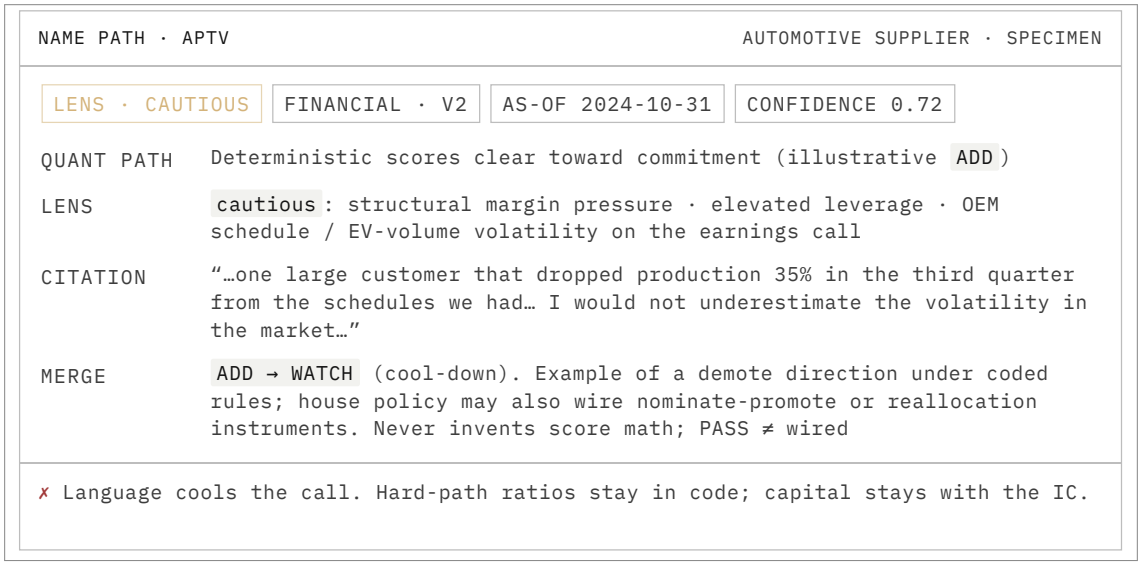
Precept dates every ingested field by real-world availability, with no look-ahead. Missing, unmapped, or integrity-broken data triggers **fail-closed**: the system refuses a clean pass (skip) rather than inventing or interpolating a value through language generation. Bounded GenAI may read and flag; it does not write uncertified synthetic fields into the authoritative spine.

4. Bounded assistance & reproducibility

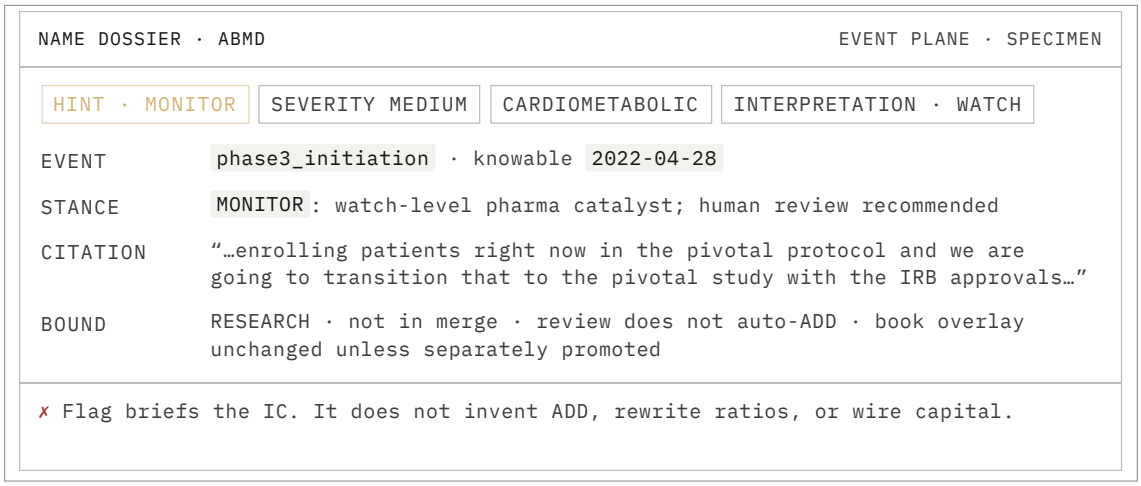
4.1 Asymmetry: coded actions, IC-calibrated directions

To use language models for document reading without surrendering capital arithmetic, Precept separates *score math* from *language comprehension*. Deterministic engines classify and gate first; language runs after that admissibility work — the same deterministic-first ordering argued for regulated AI/ML stacks more broadly.[22] The constitutional asymmetry is not "language may only cool a call." It is:

- **IC-calibrated action instruments.** Coded merge and book-overlay rules may **demote** (cool-down / underweight), **nominate promote** (raise eligibility or sizing within policy), or **propose reallocation** across the book. Which directions are live is a house mandate decision, not a fixed product dogma.[11]
- **Reference wiring vs constitution.** A common conservative default is demote-first insurance (stance can cool ADD⟶WATCH; sizing may drag). Promote and reallocate paths are first-class in the constitution; they ship only when the Investment Committee signs the rules, promotion gates, and labels.
- **Research surface.** Catalysts, fragilities, open questions, and action proposals for human nomination brief the desk whether or not a direction is auto-applied in merge.
- **Safety floor.** Soft-penalty architectures that let a fluent score outweigh a hard admissibility rule are refused. Language never authors hard-path coefficients. Suggestions and coded proposals remain under evaluation honesty: PASS ≠ wired.



Specimen · Demote / cool-down (APTV): one IC-calibrated direction (illustrative; not the only allowed instrument)



Specimen · Catalyst flag (ABMD): research surface, not merge

Typed catalyst detection is measured, not assumed. A frozen gold set of SEC-backed events (guidance withdrawals, FDA outcomes, impairments, buybacks, regulatory actions) reports recall and control specificity under a deterministic detector path. Narrative lexicons advance only as versioned phrase packs: pre-registered thresholds, offline scoring on a development split, human approval, immutable versions — never silent rewrite of book coefficients. Held-out measurement stays separate from promotion decisions. The same discipline applies on the Systematic Alpha Research rail: closed vocabulary, recorded fails, human deploy.

4.2 Live ≡ replay

Ask a generative model the same question ten times and you often get ten answers. That instability is incompatible with risk process.[17] Production and simulation use the same score and action formulas: identical PiT inputs and recorded language stance/flags produce identical hard-path scores and the same coded merge / overlay outcome under the house rule set then in force. Replay is how an IC reopens judgment months later; stochastic re-generation of stance is not treated as a substitute for the recorded artifact.

4.3 Evaluation honesty: what scoring backtests are for

Research & risk backtests exist to stress the **constitution** on a mandate shape[18] (reallocation overlays, merge hygiene, coverage, and labeled proxy constructions), not to publish a product return curve. Replay uses the same scoring engines under point-in-time knowability. Outcomes are first-class. Production coefficients do not silently rewrite themselves when a trial looks good.

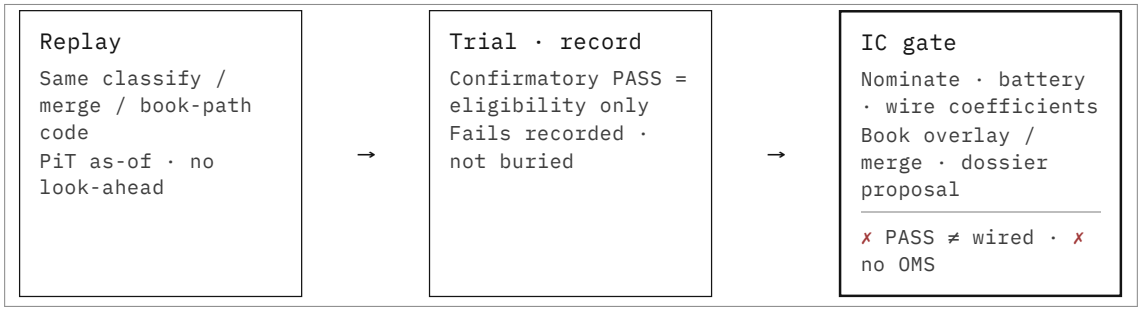


Figure 4a · Research rail: replay Name / Book engines · no auto-wire

Public process record, not a performance brochure:

Class	Observation	Consequence
Lesson	Hard gates inside book underweights destroyed lab edge	Wrong layer: published and corrected; Name merge ≠ Book overlay
Failed trial	Language and feature overlays failed confirmatory bars; outcomes recorded	Failed gates are institutional memory, not buried alpha
PASS · held	Book-overlay and construction trials cleared pre-registered bars; production coefficients unchanged pending nomination	Eligibility is not a silent book change
Bound	Conviction-ranked "top names" versus a broad index was largely negative when read as a product	ADD remains research eligibility, not a marketed ranker
Label	Proxy and laboratory panels named as such; thin coverage withheld rather than invented	Review defensibility over an unlabeled curve

Numeric trial ledgers, mandate-specific remasures, and gate thresholds ship under NDA. This section is process evidence for the IC research & risk path, not strategy-sleeve marketing, and not a live client track record.

5. Systematic Alpha Research

Systematic and quantitative desks need a **different** job from fundamental research & risk scoring: design, test, and advance algorithmic *sleeves* at research velocity, without overfitting theater or unverifiable "AI-coded" market behavior.[18] That path is **systematic alpha research & deployment**. The evaluation honesty above stresses the scoring constitution; what follows is the sleeve path toward an OMS.

Systematic Alpha Research is separate from name/book research & risk. Exploration may be wide; advancement is narrow by design. Scale is why gates matter. A house-anchored reference implementation of this constitution currently maintains on the order of **~1,600+ pre-computed market features** on minute masters (SMC · price-action · quantitative families), **~200+ session-level premises** rolled per instrument, and a **closed catalog of ~335 concepts** expanded into **~1,400+ indicator variants** that bind language to real columns — not to invented fields. Across a multi-asset book (indices, FX, metals, energy, crypto; ~16 onboarded symbols in the reference book) and multi-year M1 history, even shallow confluence — one to three premises × instrument × session window × parameter buckets — already opens a combinatorial space on the order of **billions to hundreds of billions** of configurations. Richer parameter grids and open-ended search push the envelope higher; the miner emits candidates, it does not walk a fixed enumerated list. That volume is exactly why selection-bias discipline and recorded fails are non-negotiable.[18]

Those numbers are not a marketing scoreboard. They are the reason free-form generative "strategies" fail fiduciary review: without a closed vocabulary, probability mass is spent inventing indicators and odds that never existed in the tape. Indicators enter as **premises** — dated, typed, point-in-time features the engines can measure. An analyst may state a sleeve hypothesis in plain language; an AI research loop may emit thousands of candidates overnight. Both must compile to the same catalog. Unknown tokens fail closed. Hit rates, fills, and gate outcomes are computed in code on the enriched tape, never asserted by the language layer.

Advancement rules stay deliberately narrow:

- **Dual origin.** Hypotheses may be issued by an analyst or by an AI research loop. Both enter the same closed vocabulary and gates — origin does not grant a shortcut to the OMS, and loop volume does not dilute the bar.
- **Closed vocabulary · indicators as premises.** Every concept and variant must map to a pre-computed column or registered primitive. The language layer selects and composes; it does not author new market physics mid-run.
- **Statistical gates.** Candidates face pre-declared evaluation jobs on point-in-time data (sample-size floors, out-of-sample, permutation, walk-forward; the exact suite is disclosed under NDA).[19] Throughput scales with coverage and vocabulary growth; it does not excuse selection bias.
- **Institutional memory: KEEP / PARK / KILL.** Sleeve outcomes are first-class and append-only. PARK is not "almost live." Silent revival of killed designs is process failure, not research cleverness.
- **Human-gated OMS deploy.** Clearing gates is not an order. Wiring a sleeve toward an OMS requires methodology audit and explicit human authorization. No full-auto marketing.

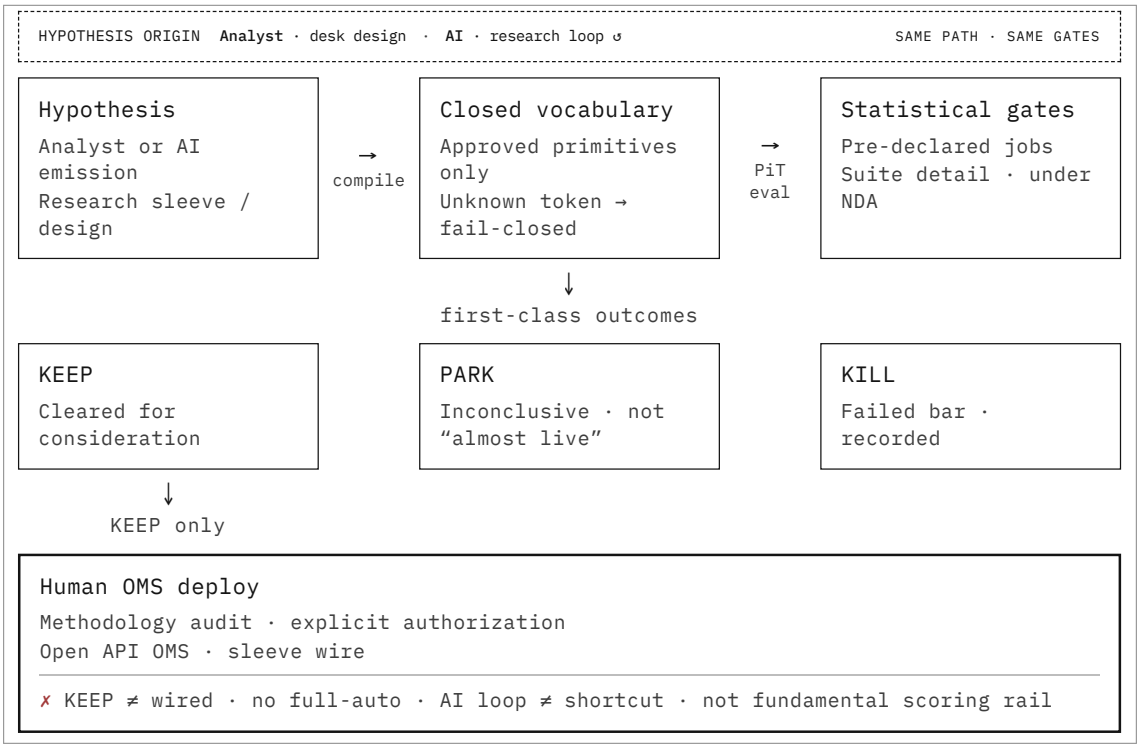


Figure 4 · Systematic Alpha Research · declared gates · then OMS deploy

6. Human sovereignty

Institutional infrastructure does not replace judgment. It documents and secures it.[11]

- **Proposal-only**. Name/book recommendations and systematic-alpha sleeve candidates remain proposals. The architecture has no autonomous path to fund liquidity.
- **Fee-defense trail.** Artifacts explain why a name sits at a weight or why a sleeve cleared the gates: immutable rules, dated evidence. That is how fees and decisions survive LP or audit review.
- **Human gate**. The Investment Committee or risk owner alone validates proposals and changes capital allocation.[8, 20]
- **Coexistence.** Precept (research & risk + systematic alpha) sits beside an existing market-data and portfolio stack. It does not require replacing a terminal or OMS. Enterprise AI assistants remain complementary (§7).

7. Category clarity: enterprise AI assistants

Desks often meet enterprise AI assistants for financial services, such as Google *Gemini Enterprise for Financial Services* or OpenAI *ChatGPT for Financial Services*, and conclude: that is the AI stack we need. These products are serious. They are also a **complementary engagement** to Investment Committee constitution: not a rival to replace, and not a substitute for a reconstructible spine.[8, 21]

They accelerate research productivity through licensed data connectors, citations, firm templates, enterprise IT controls, and drafting into Docs / Sheets / Slides (or Excel / Word / PowerPoint). They do not ship point-in-time scoring overlays, dossier proposals, or KEEP / PARK / KILL sleeve memory with human OMS deploy. A firm can run both layers: assistants for analyst speed, Precept for committee-ready process; complementary jobs, different authority over capital.

Register	Enterprise AI assistants	Precept
Role	Analyst productivity: research, artifacts, drafting	Name/book research & risk · Systematic Alpha Research
Authority	Frontier LLM / agent + licensed data + enterprise IT	Deterministic hard path; language feeds coded IC-calibrated actions
Capital wire	Not the product; the user remains responsible	**Proposal-only**; humans wire capital
Gate memory	Absent as product registry	**Systematic alpha** KEEP / PARK / KILL toward OMS · scoring trials toward overlay / merge
Delivery	Enterprise GenAI product: connectors, templates, IT controls	**Framework**: developed with the desk · house data, mandates, gates · not a product license
Placement	Research UX / data cockpit beside the desk	Governs what may enter the book and which sleeves may deploy

Detailed landscape notes (vendor scopes, design-partner context, and how we score category fit) stay out of this public paper — not as a scoreboard, and not as a gated download. Category clarity is enough here.

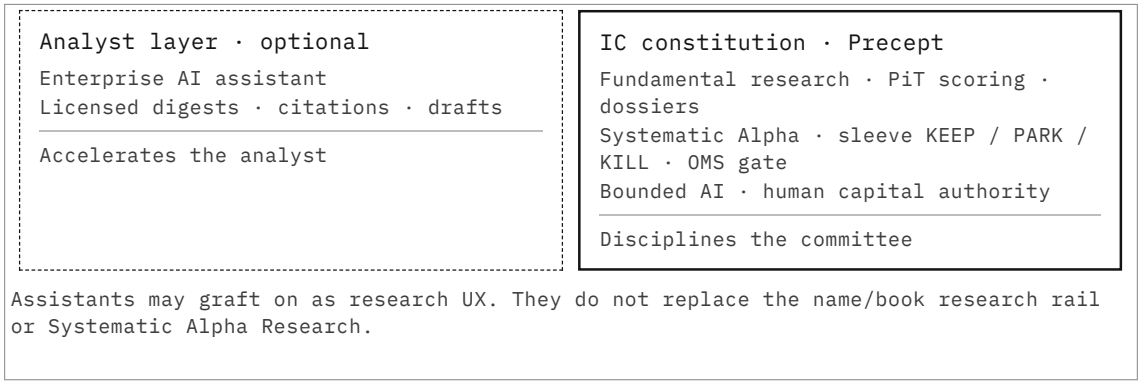


Figure 5 · Same firm, two layers: AI assistants accelerate; Precept disciplines

8. From constitution to methodology: rules-based indices

The same properties that make a scoring engine defensible before an Investment Committee also make it eligible to serve as an index methodology. A rules-based index is a basket defined by written rules (universe, eligibility, selection, weighting, rebalancing calendar, and treatment of corporate events), calculated and published on a fixed schedule. In the European Union, the provision of such an index for use in investment funds falls under the Benchmarks Regulation, which requires a robust and transparent methodology, reliable and verifiable input data, an oversight function, record-keeping, and the management of conflicts of interest.[23, 24]

Each of these requirements has a direct counterpart in the architecture described above:

- **Transparent methodology.** Selection and weighting rules are coded, versioned, and fingerprinted; a methodology change is an explicit, dated event rather than a silent revision.
- **Verifiable input data.** Every field is dated by real-world knowability; missing or broken inputs fail closed rather than being interpolated.
- **Reconstructibility.** Live ≡ Replay allows any historical rebalance to be recomputed from recorded inputs, which is the evidentiary basis for oversight and audit.
- **Honest back-history.** Point-in-time replay and recorded KEEP / PARK / KILL outcomes constrain look-ahead and selection bias in the simulated history that typically accompanies a new index.[18]
- **Conflicts of interest.** When the designer of the rules also manages assets referencing them, coded rules and an append-only trail make discretionary adjustments visible rather than deniable.

Regulation as specification. European financial regulation can appear to be a brake on the adoption of generative AI in asset management. A more precise reading is that it constrains *ungoverned* generative AI. The Benchmarks Regulation, MiFID II conduct obligations, and the operational-resilience requirements of DORA converge on the same demands: reproducible outputs, verifiable inputs, documented change, and an identifiable human accountable for each decision.[23, 25] A non-deterministic component inside an index calculation cannot meet them, since the same inputs need not yield the same constituents. Within this framework, generative models may contribute upstream, to research, hypothesis generation, and the documentation of methodology, but they hold no authority over eligibility, weighting, or rebalancing. Read this way, the regulation is not an obstacle to be worked around but a specification the architecture is designed to satisfy.

This extension changes where human authority sits, not whether it exists. In the research and risk path, the Investment Committee validates each proposal. In an index, rebalances follow the published rules mechanically; human authority moves upstream, to the approval of the methodology, the governance of its changes, and the treatment of exceptional events by an oversight function. Proposal-only discipline at the level of individual decisions becomes rule governance at the level of the methodology.

The architecture does not itself constitute a benchmark administrator. Calculation, publication, and regulatory accountability remain with an authorised or registered administrator; the framework supplies a methodology whose rules, inputs, and history can be inspected and reproduced.

9. Deployment

This paper outlines the principles for a rigorous integration of artificial intelligence within institutional investment processes. Moving from probabilistic experimentation to a deterministic, auditable infrastructure cannot be achieved through generic software or decoupled advisory reports.

The architecture assumes the institution's own data schemas, risk parameters, and Investment Committee rules. Instantiated, it remains that institution's infrastructure.

The sequence it assumes:

- **Deterministic calibration.** Translating the fund's specific investment philosophy, risk constraints, and reallocation logic into the hard-path scoring engines. The mathematical core remains strictly bespoke to the firm's profiles.
- **Interface & topology integration.** Designing customized analyst and Investment Committee (IC) workflows and structuring bounded AI (stance and flags into coded demote / nominate-promote / reallocate instruments as the house chooses) to interface seamlessly with the institution's existing data warehouses and portfolio management systems.
- **Scoped implementation.** Instantiating the architecture on a defined fundamental or systematic mandate to establish a live, reconstructible Investment Committee trail.
- **Methodology extension (optional).** Where a validated scoring engine is intended to underlie a rules-based index, formalising its rules as a published methodology for calculation by an authorised administrator, under the governance described in §8.

Notes

1. Deloitte AI Institute, *The State of Generative AI in the Enterprise: Now Decides Next* (2024). Quarterly survey of more than 2,800 director- to C-suite leaders at AI-fueled organizations piloting or implementing generative AI; finds high expectations alongside pressure to realize value quickly while managing risks, with governance among the greatest areas of concern. Supports the "add AI" adoption pressure described in §1.1. deloitte.com.
2. On the "wire an LLM to existing warehouses" shortcut: Sonya Huang & Pat Grady, "Generative AI's Act Two," Sequoia Capital (Sept. 20, 2023): Act 1 as technology-out novelty apps and undifferentiated "AI [incumbent]" pitches; Act 2 as customer-back solutions where foundation models are a piece of a broader stack, not the entire product. Juan F. Sequeda, Dean Allemang & Bryon Jacob, "A Benchmark to Understand the Role of Knowledge Graphs on Large Language Model's Accuracy for Question Answering on Enterprise SQL Databases," arXiv:2311.07509 (2023): GPT-4 zero-shot over enterprise SQL at ~16% accuracy vs ~54% when questions are posed over a

knowledge-graph representation, evidence that raw warehouse/SQL wiring without structured business context is a weak default. sequoiacap.com · arXiv:2311.07509.

3. On generative write-back without gates: Ilia Shumailov et al., "The Curse of Recursion: Training on Generated Data Makes Models Forget," arXiv:2305.17493 (2023/2024), on *model collapse*: recycling model-generated content into subsequent training irreversibly degrades the learned distribution (tails disappear; genuine human data becomes more valuable). Matei Zaharia et al., "The Shift from Models to Compound AI Systems," BAIR Blog (Feb. 18, 2024): state-of-the-art results increasingly come from multi-component systems (filters, retrieval, verification), not monolithic models alone, because control and trust are easier to engineer at the system layer, supporting fail-closed gates around generative write-back into house facts and feature stores. arXiv:2305.17493 · bair.berkeley.edu.
4. Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, *Supervisory Guidance on Model Risk Management*, SR 11-7 / OCC 2011-12 (Apr. 4, 2011). Defines model risk from incorrect or misused quantitative methods; expects sound development, independent validation, documentation sufficient for unfamiliar parties to understand operation and limitations, and governance so material model use is challengeable: the institutional baseline for a verifiable, house-owned path rather than opaque generative substitution. SR 11-7 PDF.
5. Bill Gates, Nathan Myhrvold & Peter Rinearson, *The Road Ahead* (1995/1996). Second rule of business technology: "automation applied to an inefficient operation will magnify the inefficiency" (first rule: automation magnifies efficiency when the operation is already efficient). Supports the claim that omitting architecture and wiring GenAI onto unverifiable process scales the same inefficiencies at machine speed.
6. Chengzhi Liu et al., "Auditing Agent Harness Safety," arXiv:2605.14271 (2026). Shows that a harness can return a correct final answer over a trajectory that violates permission or information-flow bounds; output-only scoring misses mid-trajectory failures, and multi-agent collaboration expands the safety risk surface, supporting the claim that opaque multi-agent choreography is a fiduciary risk object, not just a demo pattern. arXiv PDF.
7. "Auditable Agents" (arXiv:2604.05485): accountability for agent systems that act and tool-call requires reconstructible evidence, not output scores alone. Aligns with the paper's reconstructible IC trail requirement. arXiv PDF.
8. Bank of England & Financial Conduct Authority, *Artificial Intelligence Public-Private Forum: Final Report* (Feb. 2022): frames safe AI in financial services around data, model risk, and governance as guardrails; stresses accountability/responsibility for AI-informed decisions, explainability, and that most applications still retain a human-in-the-loop rather than full automation of decision-making. Florian Ostmann & Cosmina Dorobantu, *AI in Financial Services*, The Alan Turing Institute (2021; FCA-commissioned): elevates competent use and human oversight, and treats accountability as answerability plus auditability (records enabling end-to-end review), with system and process transparency as preconditions for demonstrating trustworthy use. Together these support reconstructible decisions, bounded assistance, and a human wire as institutional requirements, not features shipped by catalog engines or frontier FS copilots alone. BoE AIPPF PDF · Turing PDF.
9. On foundations-first / modular architecture before model swap: Matt Bornstein, Rajko Radovanovic et al., "Emerging Architectures for LLM Applications," Andreessen Horowitz (a16z): reference LLM app stack as layered systems (data pipelines, retrieval/context, orchestration, ops) with foundation models as a replaceable layer; in-context patterns reduce much of the problem to data engineering rather than model-only demos. Leandro DalleMule & Thomas H. Davenport, "What's Your Data Strategy?," *Harvard Business Review* (May–June 2017): firms need a deliberate data strategy balancing offense and defense; most organizations under-use structured data and waste analyst time on discovery/prep, evidence that clean, governed data foundations precede durable innovation on the spine. a16z.com · hbr.org.
10. Federal Reserve / OCC / FDIC model-risk tradition continued: SR 26-2 (Apr. 2026), successor context to SR 11-7, emphasizes risk-based governance of models used in decisions and notes that generative

and agentic AI fall outside that guidance's novel scope; yet banking organizations remain expected to apply broader risk-management and governance practices to such tools. Deterministic capital math remains the natural home of classical model-risk discipline. SR 26-2.

11. NIST, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1 (2023), GOVERN-3.2 / MAP-3.5: policies for human-AI roles and documented human oversight. Companion: NIST AI 600-1, *Generative AI Profile* (2024). AI RMF PDF · GenAI Profile.
12. IOSCO, *The use of artificial intelligence and machine learning by market intermediaries and asset managers*, FR06/2021 (Sept. 2021), Measures 1–6 (senior accountability; continuous testing segregated from live; skills to challenge algorithms; third-party controls; disclosure; data quality and bias): supports house-calibrated methodology rather than a one-size generative score pack. IOSCOPD684.pdf.
13. Rolf W. Banz & William J. Breen, "Sample-Dependent Results Using Accounting and Market Data: Some Evidence," *Journal of Finance* 41, no. 4 (1986): 779–793. Documents look-ahead and ex-post-selection bias when using standard COMPUSTAT-style files. doi:10.1111/j.1540-6261.1986.tb04548.x.
14. Matthew Lyle, Federico Siano & Teri Lombardi Yohn, "Re-Standardized Financial Statement Data" (working paper, July 2024). Shows re-standardization of common Compustat products introduces look-ahead and replication failure; recommends Snapshot point-in-time files. Yale SOM PDF.
15. S&P Global Market Intelligence, *Compustat Data* product brochure: markets "true point-in-time analysis… to avoid look ahead bias" and retained original values vs subsequent changes (separate from default restated extracts). Brochure PDF.
16. Microsoft Qlib documentation, "(P)oint-(I)n-(T)ime Database": using only the latest amended version of financial-report data causes data leakage; PIT storage keeps online trading and historical backtesting aligned. qlib.readthedocs.io.
17. NIST AI 600-1 *Generative AI Profile* treats confabulation / unreliable generation as a first-class GenAI risk category requiring human oversight and acceptable-use limits, consistent with refusing to place stochastic generation on the capital hard path. NIST.AI.600-1.pdf.
18. David H. Bailey & Marcos López de Prado, "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality," *Journal of Portfolio Management* 40, no. 5 (2014): 94–107; and Bailey, Borwein, López de Prado & Zhu on the probability of backtest overfitting. Supports pre-declared trials, recorded fails, and refusal to market unlabeled lab curves. The combinatorial magnitudes cited in §5 are order-of-magnitude illustrations from a house reference implementation (closed-vocab concepts and variants bound to pre-computed features across a multi-asset M1 book); they describe searchable space under gates, not a published live track record or an enumerated exhaust of all trials. SSRN 2460551.
19. Marcos López de Prado, "A Data Science Solution to the Multiple-Testing Crisis in Financial Research," *Journal of Financial Data Science* 1, no. 1 (Winter 2019): 99–110. Aligns with declared statistical gates and KEEP / PARK / KILL memory. SSRN 3177057 · doi:10.3905/jfds.2019.1.099.
20. IOSCO FR06/2021 Measure 1: designated senior management accountability for development, testing, deployment, monitoring, and controls of AI/ML, including sign-off on initial deployment and substantial updates. IOSCOPD684.pdf.
21. Primary product descriptions (category, not endorsement): Google Cloud, "Introducing Gemini Enterprise for Financial Services," cloud.google.com; OpenAI, "Introducing ChatGPT for Financial Services," openai.com. Enterprise research/productivity layers with connectors, citations, and IT controls; capital process and house scoring constitution remain the user's responsibility.
22. George Melville, Dena Ghiassi, Scott Inthathirath & Julian Yeomans, "Filter before You Solve: A Deterministic-First/Learned-Second Architecture for AI-Driven Portfolio Management with Real-Money Training-Investment Calibration," *Journal of Software Engineering and Applications* 19 (2026): 301–348. Formalizes *filter-before-you-solve* (hard deterministic admissibility before learned inference; hard constraints rather than soft penalties that can trade against a ranker score) and

argues that hallucination-prone inputs are removed from the decision domain by construction when the filter precedes inference. Precept's research & risk path is in that architectural family for Investment Committee process: coded scores and hard gates first; language may feed coded action rules the house calibrates (demote, nominate promote, propose reallocate) but cannot author score math or wire capital. Mandate and surface differ (name/book IC proposals vs the paper's live options portfolio illustration). PDF: content.scirp.org/pdf/jsea_9303528.pdf · doi:10.4236/jsea.2026.198013.

23. Regulation (EU) 2016/1011 of the European Parliament and of the Council of 8 June 2016 on indices used as benchmarks in financial instruments and financial contracts or to measure the performance of investment funds (Benchmarks Regulation, BMR). Sets requirements on benchmark administrators covering governance and conflicts of interest, oversight function, input data, methodology transparency, and record-keeping. eur-lex.europa.eu.
24. IOSCO, *Principles for Financial Benchmarks: Final Report*, FR07/13 (July 2013). Nineteen principles on governance, quality of the benchmark, quality of the methodology, and accountability, including transparency of methodology changes and management of conflicts of interest. IOSCOPD415.pdf.
25. European Securities and Markets Authority, *Public Statement on the use of Artificial Intelligence (AI) in the provision of retail investment services*, ESMA35-335435667-5924 (30 May 2024): firms remain fully responsible for decisions and client outcomes under MiFID II when using AI tools, including third-party generative models. ESMA35-335435667-5924.