clarista

The Harness

WHY THE MODEL ISN'T THE PRODUCT IN ENTERPRISE AI

MAY 2026

WRITTEN BY
Suvrat Bansal

UNBLOCK KNOWLEDGE

Table of Contents

clarista

The Harness

Why the Model Isn't the Product in Enterprise AI

Executive Summary

In personal AI, the architecture is simple because the stakes are low. In enterprise finance, the gap between personal AI and enterprise AI is not a technology problem but an architecture problem. This white paper examines the "Harness": the essential framework that dictates what an AI model sees, remembers, and is permitted to do.

While models are becoming a commodity, the harness is the differentiator that providing the trust, compliance, and competitive advantage necessary for regulated industries.

https://www.clarista.io/

clarista

You Already Use a Harness

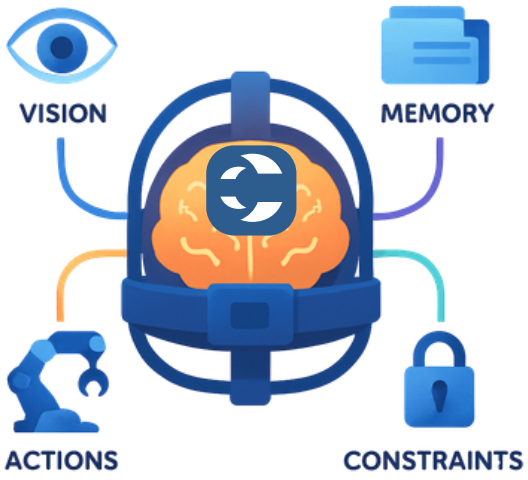
Open any modern AI assistant and ask it to help you plan a dinner party for this weekend. Watch what happens. The model does not just think and reply. It searches the web for trending recipes. It checks the weather forecast to decide whether you should plan for the patio or the dining room. If you told it last month that your sister is vegetarian, it remembers and adjusts the menu. It finds a playlist on a music service. It drafts a shopping list and offers to email it to you.

That feels like one seamless interaction, but underneath it, several distinct things are happening.

The **model** is the language intelligence: it understands your request and generates a response. But everything else: the web search, the weather check, the email draft; those are **tools** the model was given permission to call. The fact that it remembered your sister's dietary restriction from a previous conversation is memory.

The way all of this information was gathered and assembled into a single coherent answer is **context**. And the code that decided when to search, what to remember, which tools to invoke, and how to present the result is the **harness**.

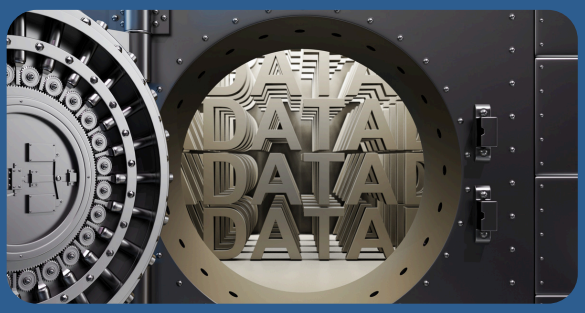
The model provides the intelligence. The harness provides everything else: what the model sees, what it remembers, what it can do, and what it is not allowed to do. Strip away the harness and the model is just a text generator. Add the right harness and it becomes a capable assistant.



clarista

Why This Breaks in Enterprise Finance

In a companion piece, *The Stage Is Set: Why Enterprise AI in Finance Is No Longer a Future State*, I examined why the gap between personal AI and enterprise AI is not a technology problem but an architecture problem. The harness is where that architecture lives, and it is where personal AI and enterprise AI diverge most sharply.



Three fundamental challenges explain why.

The model was never trained on your data.

Enterprise finance is different. A wealth advisor's client portfolio, a private credit team's deal pipeline, a fund's investor capital accounts: this data lives across multiple internal systems, changes daily, and no public model has ever seen it. The model needs to be given this information in real time, retrieved from the systems where it lives. That is the harness's job.

Raw data is not enough.

An LLM cannot reliably compute performance attribution across thousands of positions or apply a team's proprietary risk framework to a new lending opportunity. These require structured computation and institutional knowledge that must be prepared before the model ever sees a question. The harness must transform raw data into governed, contextualized information.

Governance is not optional.

In regulated finance, a wrong number is a regulatory issue and a reputational risk. Every piece of data the model sees must be filtered through user entitlements. Every response must be traceable to its source. Every action must be logged and auditable.

These three challenges, retrieval of live data, preparation of meaningful context, and enforcement of governance, are not features you add to a general-purpose AI platform. They are the core responsibilities of the harness itself.

clarista

Agent = Model + Harness

The idea that the code wrapped around an AI model matters as much as the model itself has rapidly become a point of convergence across the AI engineering community. Harrison Chase, co-founder of LangChain, has described how the evolution from simple chains to complex orchestration flows to full agent harnesses tracked the growing capability of the models themselves.

Birgitta Böckeler at Thoughtworks formalized the concept with her guides-and-sensors framework. OpenAI showed how a three-engineer team produced a million lines of code by changing nothing but the harness. And in March 2026, Stanford and MIT researchers published the Meta-Harness paper, demonstrating that automatically optimizing the harness around a fixed model can outperform hand-engineered alternatives, and that good harnesses transfer across different models entirely.

Harness

Everything the model sees, remembers, and does

Model (LLM

Agent = Model + Harness

A harness is not just configuration or a set of prompts. It is a stateful program that determines what information to store, retrieve, and present to the model at each step. The model reads text and produces text. The harness decides everything else. Model quality has become table stakes. The harness is the differentiator.

For general-purpose AI, this insight matters. For regulated finance, it is everything.

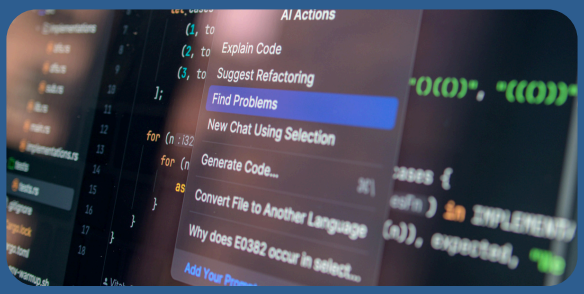
clarista

The Building Blocks: A Financial Perspective

To make the building blocks of an enterprise harness concrete, consider a scenario that plays out at wealth management firms every day. A financial advisor is preparing for a quarterly client review. Today, she logs into four or five separate platforms: the portfolio accounting system for current positions, the CRM for client notes and life events, a market research portal for macro views, a financial planning tool for projections, and possibly a document management system for prior meeting notes and investment policy statements. She spends an hour assembling a picture that should take minutes.

An AI agent with the right harness changes this entirely. But the harness must be built specifically for this world. The standard components of a harness, memory, context, tools, and guardrails, each take on a fundamentally different character in enterprise finance.

Component	Personal AI	Enterprise Finance Harness
Memory	Past conversations and preferences	The firm's live portfolio, records, and documents
Context	Personal question + web results	Team-scoped, role-governed enterprise data
Tools	Web search, code execution	Governed workflows and productivity integration
Guardrails	Basic safety filters	Security and compliance woven through every layer



Memory: The Firm's Governed Data

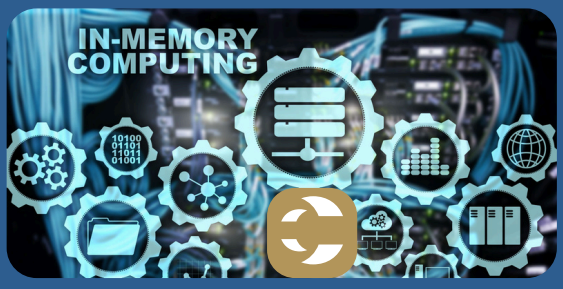
In personal AI, memory means the model's recall of your past conversations: that you prefer window seats, that you asked about hiking trails last week. In the AI engineering community, this has a well-established meaning: short-term memory (the current conversation) and long-term memory (preferences and patterns that persist across sessions). Both matter.

But in enterprise finance, a third and dominant category exists: institutional memory. When our advisor sits down to prepare for her client review, the memory she needs is not a summary of her last AI conversation. She needs current portfolio positions, recent transactions, risk exposures, life events that affect the financial plan, regulatory flags, and market movements relevant to the client's holdings. None of these are in the conversation, and if they are, they are already stale.

That is the real memory. It is live. It is institutional. And it does not belong to the model; **it belongs to the firm.** Portfolio positions updated in real time via SQL queries against live systems. Client records retrieved from CRM platforms. Deal documents surfaced from vector stores. Connected, not copied, not cached, not stored inside a model provider's infrastructure.

The implications for data sovereignty are immediate. If memory lives inside the model or a proprietary AI platform, the firm loses control. Where is that data stored? Who has access? Can it be deleted when a client requests? If memory lives in the harness, connected to the firm's own governed systems, every one of those questions has a clear, defensible answer.

There is another dimension that general-purpose platforms miss. Consumer AI memory is personal: one user, one history. Financial services memory is **team- and role-level.** Our advisor shares a client book with her team. A deal team shares pipeline intelligence. A risk team shares exposure data across the portfolio. When a team member leaves, the memory does not go with her.



That is the real memory. It is live. It is institutional. And it does not belong to the model; it belongs to the firm.

Context: Team-shaped, Not User-Shaped

In personal AI, context is simple: your question, plus whatever the model found on the web, plus what it remembers about you. The harness assembles this into a prompt and the model responds.

For our advisor, the context challenge is entirely different. The harness needs to assemble her client's current positions from the portfolio accounting system, recent life events from the CRM, the investment policy statement from the document repository, relevant market commentary, and prior meeting notes, all scoped to what this advisor, on this team, for this specific client, is authorized to access. Show the model too little and it hallucinates. Show it too much and the signal drowns in noise. Show it the wrong thing and you get a perfectly articulate, completely misleading output.

Context in finance is team-shaped, not user-shaped. The context window is not *"what does this individual user need?"* but *"what does this team's mandate require for this question?"* An advisory team, a credit team, and a compliance team need different slices of the same enterprise data.

Teams also develop their own language, shorthand, and calculation conventions that carry deep meaning but that no LLM has been trained on. Encoding this institutional context into the harness is what makes the difference between AI that sounds knowledgeable and AI that actually is.

Context in finance is team-shaped, not user-shaped.

clarista

Governed Capabilities: Context Infrastructure and Tools

A Snapshot.

The capability layer operates across four levels:

01

Data Augmentation

Pre-computing governed context like performance attribution.

02

AI Data Readiness

Data dictionaries and profiling that make data comprehensible to an AI within a specific team's domain.

04

Workflow Orchestration

Triggering structured workflows with human checkpoints where judgment matters.

03

Productivity Integration

Reaching into email, Teams, and calendars within the same governance framework.

The Drill-down of the Four Levels

In personal AI, tools are open-ended: the model searches the web, executes code, creates files, and sends emails with minimal constraints. In enterprise finance, every capability is governed, scoped to what this team is allowed to do, logged for auditability, and integrated into workflows that include human judgment.

The capability layer of an enterprise finance harness operates across four levels. The first two form the context infrastructure that prepares what the model sees. The latter two are tools in the technical sense: actions the model invokes during a conversation.

Level 1: Data Augmentation

This is the most valuable and hardest-to-replicate layer. It creates information that does not exist in the raw data. When our advisor asks why her client's portfolio underperformed its benchmark, the answer requires performance attribution, decomposing returns into allocation, selection, and interaction effects. No LLM can compute this reliably in real time. The harness provides this derived intelligence as pre-computed, governed context. This is the layer that answers the question every firm eventually asks: "Why can't we just plug an LLM into our data warehouse?" Because the warehouse does not have what the advisor actually needs. Someone has to compute it. That is part of the harness.

Level 2: AI Data Readiness

Data engineering moves data. AI data readiness makes data comprehensible to an AI operating within a specific team's domain. It includes data dictionaries that tell the AI what each field means in this team's context, profiling so the AI understands data quality before it answers, and team-level linguistic context: the terms, acronyms, and calculation logic that an advisor uses daily but no LLM has been trained on. These are institutional definitions, the linguistic DNA of how a team thinks about its work.

Together, Levels 1 and 2 form the context infrastructure. They are not actions the model takes during a conversation; they are the governed pipeline that ensures the model sees the right information in the right form.

Level 3: Workflow Orchestration

The AI helps our advisor prepare a client review. Then what? In enterprise finance, the answer is rarely "just show it on screen." It is "route the draft for compliance review," "flag the allocation change for the investment committee," "generate the meeting summary and send it to the client." The harness needs tools that trigger structured workflows with human checkpoints at the points where judgment matters. Each step logged. Each decision traceable.

Level 4: Productivity Integration

Our advisor does not live inside an AI interface. She lives in email, Teams, calendars, and documents. The harness needs tools that reach into these environments: pull context from a recent email thread with the client, surface the last meeting's action items, draft a follow-up, and send it. All within the same governance framework. Together, these four levels form the capability layer. Context infrastructure creates and prepares what the model sees. Tools let it act within governed boundaries.

clarista

Security, Governance, and Compliance = The Foundation

In personal AI, guardrails are basic safety filters. In regulated finance, they are the foundation everything else is built on, and they are not one thing. They are three distinct disciplines.

Security

controls who can see and do what. When our advisor's AI agent queries the portfolio accounting system, the identity context must flow through every layer. The system needs to know that this specific advisor, on this specific team, initiated this specific request. Most enterprise AI implementations quietly break here: they connect the AI to source systems using a service account with broad access, bypassing the entitlement controls the firm spent years building. The security boundary must be at the point of retrieval, not the point of display.

Governance

controls quality. Where did this data originate? When was it last validated? Is it complete? In financial services, a wrong number is not an inconvenience; it is a regulatory issue. The harness must embed quality controls at the retrieval layer, adding assurance at the point where information is actually consumed.

Compliance

controls auditability. Every interaction must be reconstructable for audit, regulatory inquiry, or internal review. Every data retrieval logged with full provenance. Every output traceable end-to-end.

The complexity of the compliance path should match the stakes of the decision: a simple portfolio lookup follows one path; a client recommendation follows a different one, with draft, cross-reference, human review, and then delivery.

The Strategic Question: Who Controls Your Harness

Chase's central argument deserves repeating in the context of financial services: because memory lives inside the harness, whoever controls the harness controls the memory. In regulated finance, the memory is the firm's most sensitive asset: client data, portfolio intelligence, deal pipelines, risk exposures.

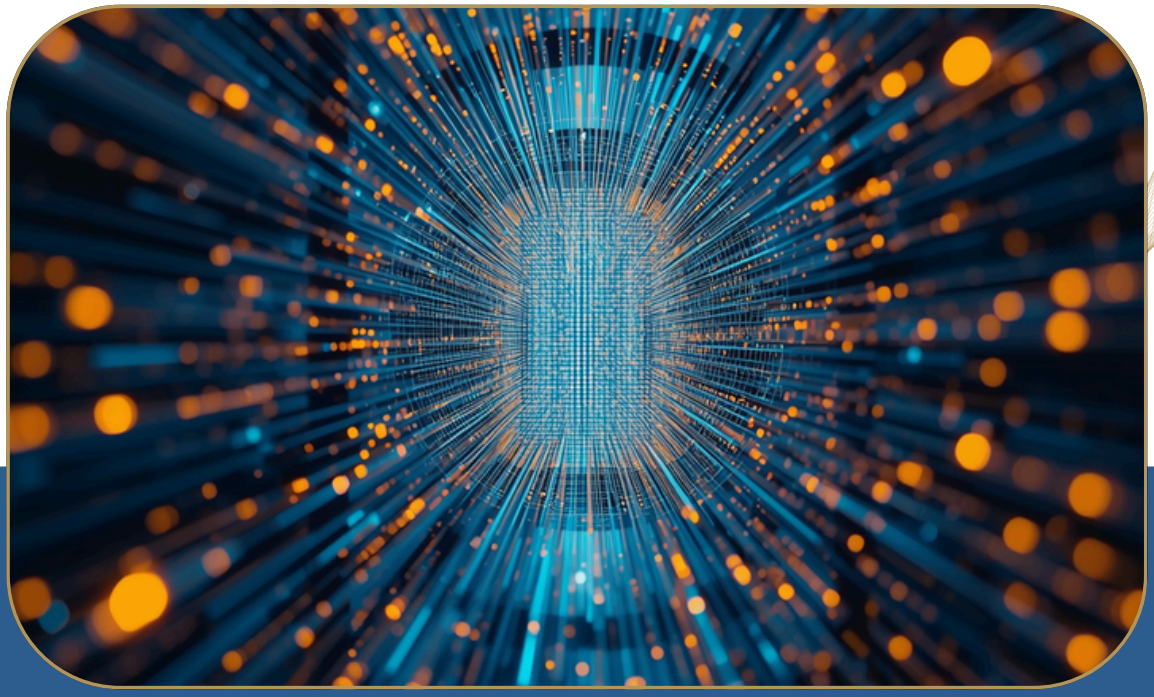
If your harness is embedded in a vendor's proprietary platform, the vendor controls it. Your memory, your context, your guardrails, your audit trail, all on their infrastructure, under their terms.

The model, by contrast, can be swapped. Models improve every quarter. The Stanford Meta-Harness research confirms this: harness optimization transfers across models. A well-designed harness makes a smaller, cheaper model outperform a larger one running on a generic scaffold.

The harness cannot be swapped easily.

It encodes how your firm thinks, how your teams operate, what your compliance framework requires, and how your data flows. No two firms should have the same harness, because no two firms have the same context.

Built, Not Imagined



At Clarista, this is not a position paper; it is the architecture we have built and continue to evolve. Our architecture treats the firm's live, governed data as the agent's institutional memory, retrieved in real time, never copied into a model provider's infrastructure. Context is assembled at the team level. Our capability layer computes derived financial intelligence, makes data AI-ready through dictionaries and team-level linguistic context, orchestrates governed workflows, and integrates with the productivity tools where teams actually operate.

Security, governance, and compliance are not features we added; they are the foundation we built on. The authentication chain flows from the user through to every data source. Entitlements are enforced at the retrieval layer. Every output is traceable from source to response.

The model provides the intelligence. The harness provides the trust, the compliance, and the competitive advantage.

The firms that will define the next era of financial services are making harness decisions right now, whether they use that language or not. The ones who recognize this will own their advantage. The ones who do not will rent it, on someone else's terms.

The stage was set. The harness is how you step onto it.

clarista

Suvrat Bansal

FOUNDER AND CEO

About Suvrat

Suvrat Bansal is the Founder and CEO of Clarista, where he leads the development of contextual intelligence platforms for enterprise scale. A veteran of the financial technology sector, he specializes in building the AI governance frameworks that enable wealth management and private equity firms to trust their data. His work focuses on transforming trapped institutional knowledge into decision-ready intelligence without compromising on security or compliance.

For more insights, contact him.

Suvrat@Clarista.io

REFERENCES

The Evolution of AI Agents: From Simple Chains to Complex Orchestration.	2024	LangChain Blog
Orchestrating Generative AI: The Guides and Sensors Framework. Thoughtworks Insights.	2024	Böckeler, B.
Engineering Custom Context: How Architecture Defines Model Performance.	2025	OpenAI
Meta-Harness: Automated Optimization of LLM Orchestration Frameworks.	March, 2026	Stanford University and MIT Research Team.
Report on Artificial Intelligence in the Securities Industry: Governance and Identity Context.	2025	Financial Industry Regulatory Authority (FINRA)
Revised Guidelines for Data Provenance and Traceability in Automated Financial Advice.	2026	International Organization of Securities Commissions (IOSCO)

Clarista is the platform for contextual intelligence that unblocks the knowledge trapped inside your enterprise. We don't just connect or catalog data; we literally create new, decision-ready intelligence from the documents, messages, and systems that define your business. This involves discovering entirely new data assets from both structured and unstructured sources, identifying previously unknown opportunities and risks, and ensuring that the most up-to-date and appropriate AI governance is in place. With built-in real-time governance and explainable outputs, Clarista ensures that what AI produces can be trusted, reused, and scaled across every workflow, product, or decision. From wealth management to private equity to vertical AI builders, Clarista empowers teams to move with confidence on newly created data they can govern from the start.

Unblock Knowledge.

For more information, visit us at www.Clarista.io
Follow us on

COPYRIGHT AND ACKNOWLEDGEMENTS

Special thanks to:
Katherine Fleming, Editor
Suvrat Bansal, Publisher
The entire Clarista team.

Clarista Inc.
New Jersey
www.Clarista.io
Marketing@Clarista.io