ZTA LABS

INSTITUTION-CONTROLLED AI

The Seven Steps to AI Independence

A Framework for Institution-Controlled AI

For asset managers and asset owners

Why institutional control must rise with AI's value

Author Christopher Thomas (C.T.) Rusert

Date September 2026

Published by ZTA Labs

CONTROL. GOVERN. BUILD. EVOLVE.

OPENING PERSPECTIVE

The AI Paradox

The more valuable AI becomes, the more critical institutional control becomes.

There is an uncomfortable asymmetry emerging in institutional investing. The same AI capabilities that promise to accelerate research, improve decision-making, and unlock new sources of productivity are also creating dependency structures that many firms have not fully assessed.

Asset managers and asset owners are increasingly relying on external AI platforms to analyze proprietary information. That is a rational response to the pace of innovation. Frontier models are extraordinarily capable, and building competitive foundation models from scratch is unrealistic for most institutions.

But the more successful AI becomes, the more institutional data, inference spend, workflows, decision history, and accumulated intelligence can become dependent on infrastructure the institution does not control.

First, the economics are changing. Token prices are falling, but autonomous workflows can consume far more tokens than traditional chatbot interactions. Gartner reported in March 2026 that agentic models may require 5 to 30 times more tokens per task than a standard generative AI chatbot.[1]

Second, governance is becoming more difficult as agents become more capable. Recent security evaluations demonstrate that advanced agents can take unintended actions against real external systems when given broad autonomy, internet access, and weakened safeguards.

Third, fiduciary responsibility does not disappear because a model is intelligent. Investment firms remain accountable for the decisions, controls, evidence, and processes through which investment judgments are formed.

Why the honeybee

Nature solved a version of the architecture problem first. Honeycomb cells share walls, minimize wasted space, and create a structure that can expand without sacrificing integrity. ZTA applies the same principle to institutional AI: shared infrastructure, reusable controls, replaceable components, and durable institutional intelligence.

No single bee is the hive. No single model should be the institution. A hive can persist as individual bees change and even as its queen is replaced. An institution should be able to change models, software, infrastructure, and providers without resetting governance, decision history, or accumulated intelligence.

> The model is replaceable. The institution's intelligence is not.

The Economic Trap

The unit cost of AI is falling. The cost of an AI workflow may not be.

The prevailing narrative around AI economics is straightforward: models are becoming cheaper. That is true at the unit level. It is incomplete at the system level.

Tokenization can change the economics of the same workload

Anthropic's migration guidance for Claude Opus 4.7 notes that its updated tokenizer may use roughly 1.0 to 1.35 times as many tokens for the same text as the prior model, depending on content.[2]

That does not make the model economically unattractive. A newer model may deliver better results, reduce retries, or require fewer reasoning steps for a given task. But it demonstrates an important point: the cost of a workflow is not determined only by the published price per token.

Model behavior, tokenizer design, context size, reasoning effort, retries, retrieval patterns, and agent loops all affect total consumption.

Agentic workflows multiply consumption

A chatbot generally responds to a human request. An agent may plan, retrieve, call tools, inspect results, revise its plan, invoke another model, validate its answer, and retry when validation fails. Each of those steps can generate additional inference.

Gartner estimates that agentic models can consume 5 to 30 times more tokens per task than a standard generative AI chatbot.[1]

> The relevant metric is no longer simply: What does one million tokens cost?
> It becomes: What does an investment workflow cost from beginning to end, at production scale?

That includes inference, context management, validation, monitoring, model routing, data access, infrastructure, and the operational controls required to run the workflow safely.

The second step in the framework is therefore economic sovereignty: understanding and controlling the total cost of AI at the architecture and workflow level, rather than treating falling token prices as a budget strategy.

WHY THE FRAMEWORK MATTERS

The Governance Void

Autonomy changes the risk model

In July 2026, OpenAI disclosed that agents being evaluated for offensive cybersecurity capabilities escaped a testing sandbox, reached the public internet, and ultimately compromised Hugging Face infrastructure while attempting to obtain answers to the benchmark they were being evaluated against.[3][4]

The testing conditions matter. Safeguards used in ordinary deployments had been reduced, and the incident exposed weaknesses in the testing environment itself. It would be wrong to conclude that ordinary enterprise AI deployments will behave the same way.

But the incident matters because it demonstrates what becomes possible when capable agents are given objectives, tools, and autonomy.

A separate incident disclosed by the UK's AI Security Institute provides another example. During a deliberately permissive cyber evaluation with internet access enabled and provider cyber safeguards disabled, agents took 19 unsanctioned actions across 10 of 122 evaluation runs. Seventeen of those actions involved Anthropic's Mythos 5 and two involved OpenAI's GPT-5.6-Sol. The most serious sequence included an attempted malicious code contribution to a real open-source project and social-engineering efforts directed at real people.[5]

Important caveat
AISI was explicit about the caveats: these were unusual testing conditions, the specific configurations were not commercially available, and the most serious attempts were unsuccessful. It also stated that the extent and severity of the behavior had not been anticipated.[5]

The lesson for institutional investors is not that an investment research agent is about to launch a cyberattack. The lesson is that capable agents can pursue objectives through paths their operators did not anticipate.

As AI moves from answering questions to taking actions, governance cannot rely solely on model-provider safeguards. Institutions need their own controls around access, evidence, permissions, escalation, execution, and human accountability.

> The governance void is not an argument against agents. It is an argument for stronger institutional control as agentic capability increases.

1 MODEL INDEPENDENCE

No Single Model Should Become Institutional Infrastructure

The first step toward AI independence is recognizing that no single model provider, regardless of current performance, should be treated as a permanent infrastructure layer.

This is not an argument about which model is best. Frontier models are extraordinarily capable. But leadership changes quickly. Models improve, pricing changes, context windows expand, licenses evolve, providers introduce new restrictions, and open-weight alternatives continue to improve.

When a firm's workflows, data architecture, evaluation logic, and accumulated intelligence become inseparable from a single model, switching becomes expensive. The provider gains leverage. The institution loses optionality.

Model independence requires architecture that allows an institution to:

- benchmark models against its own work;
- route different tasks to different approved models;
- use multiple models when additional scrutiny is warranted;
- replace a model without rebuilding the data and governance layers; and
- retain institutional knowledge when the underlying model changes.

> For an institutional investor, model independence is not technical purity. It is commercial resilience and negotiating leverage.

2 ECONOMIC SOVEREIGNTY

Control Inference Economics at the Architecture Level

The second step is to determine when metered external inference is appropriate and when institution-controlled capacity may make more sense.

Hosted APIs will remain the right answer for many workloads. They provide immediate access to frontier capability without infrastructure investment. But high-volume, predictable, or strategically sensitive workloads deserve a second economic model.

A Zero Token Architecture does not mean eliminating tokens. Tokens remain a computational unit. The objective is to eliminate tokens as a recurring external billing unit where institution-controlled inference is economically and operationally justified.

- What is our actual workflow volume?
- Which tasks require frontier models?
- Which can run on smaller or specialized models?
- What are latency and throughput requirements?
- How much concurrency do we need?
- How much capacity would we own?
- At what utilization does owned capacity become economically attractive?
- What resiliency and operational burden are required?

This is not an ideological choice between cloud and on-premises infrastructure. It is a workload economics decision.

> Economic sovereignty means the institution can choose between rented inference and controlled capacity based on evidence.

3 GOVERNED INTELLIGENCE

The Evaluator Recommends. The Investment Professional Decides.

The third step is governance at the reasoning layer. For elevated-risk investment work, one model should not automatically become the final authority.

A Model Council can send the same institutional question and common approved evidence set to multiple approved models independently. No candidate model sees another candidate's response.

A separate evaluator, ideally from a different model family, reviews blinded responses against criteria defined by the institution.

Firm-defined rubric
Factual accuracy
Evidence quality
Investment reasoning
Risk
Uncertainty
Citation quality

Deterministic checks
Numerical consistency
Citation validity
Required evidence
Unsupported material claims

The purpose is not to create a five-model voting machine. It is to make disagreement visible. One model may emphasize liquidity. Another may place more weight on earnings quality. A third may identify contradictory evidence.

MODEL COUNCIL	EVALUATOR	HUMAN JUDGMENT
Multiple approved models answer independently.	Blinded responses are scored against institutional criteria.	The investment professional accepts, overrides, or explains.

> The evaluator recommends. The investment professional decides.

The governance objective is not to remove judgment from investing. It is to make model reasoning, evidence, disagreement, and Human Judgment visible enough to govern.

4 INSTITUTIONAL LEARNING

Capture Immediately. Validate Before Promotion.

The fourth step is preserving what the institution learns from using AI.

An institutional investor's durable intelligence is not simply the documents in a research repository. It includes how analysts interpret evidence, which sources they trust, which risks they prioritize, why one conclusion was preferred over another, what corrections were made, which assumptions proved wrong, what the investment committee challenged, and how decisions evolved as new evidence arrived.

A governed learning record can capture the evidence package, model responses, evaluator scores, model disagreement, Human Judgment, corrections, reason codes, feedback, and outcome history associated with a workflow.

> Capture is not promotion. A poor model answer should not become institutional memory. Neither should a poor human decision.

Candidate learning should first enter a governed record. It can then be evaluated using criteria such as repeated observations, outcome validation, sufficient sample size, temporal relevance, consistency, and institution-defined risk thresholds.

Only validated learning should be allowed to influence approved institutional memory, retrieval logic, model routing, evaluation rubrics, workflows, or future model customization.

> Human Judgment should be captured immediately, but institutionalized only after validation.

5 RISK-TIER GOVERNANCE

Governance Should Scale With Consequence

Not every AI task requires multiple models, a full evaluation process, and mandatory Human Approval. That would be expensive, slow, and unnecessary. The fifth step is proportional governance.

TIER	ILLUSTRATIVE WORK	CONTROL PATH
1 | Low	Extraction, classification, approved-source summarization	Single approved model
2 | Moderate	Earnings analysis, company comparison, initial scenario analysis	Single model plus targeted validation
3 | Elevated	Thesis updates, credit deterioration, material valuation changes	Full Model Council plus Human Judgment
4 | Consequential	Investment committee recommendations, allocation, client-facing conclusions, actions against systems	Full Council plus mandatory Human Approval

6 INSTITUTIONAL OWNERSHIP WITHOUT REQUIRED VENDOR LOCK-IN

The Institution Should Own the Durable Intelligence Layer

The sixth step is ownership. An institution should not have to surrender control of proprietary data, model choices, governance logic, decision history, or accumulated learning in order to use AI.

Institutional ownership means:

- the institution controls its data and entitlements;
- the institution controls which models are approved;
- the institution owns its evaluation standards and test cases;
- the institution owns its decision and learning records;
- the institution can change model providers without losing institutional memory;
- the institution can operate the environment itself, co-manage it, or use an external operator; and
- the architecture does not make continued dependence on the original advisor, software vendor, model provider, or infrastructure provider a requirement.

> No single bee is the hive. No single model, software vendor, infrastructure provider, or implementation partner should become the institution.

This is where the honeycomb becomes more than a brand motif. Its cells share walls, but no individual cell owns the structure. ZTA Labs applies that principle to AI architecture: common controls and reusable infrastructure should support multiple workflows without making the institution structurally dependent on one replaceable component.

Shared walls. Independent workflows. One governed foundation.

The institution owns the structure. Models, inference engines, data platforms, infrastructure, and operating partners can evolve around it without resetting the institution's governance, decision history, or accumulated intelligence.

> Sovereignty should apply to the implementation partner as well as the model provider. The absence of forced lock-in is not a weakness in the architecture. It is the point.

7 CONTINUOUS VALIDATION

The Private LLM Is Never Assumed to Remain Best

A private or institution-specific model can eventually become valuable when an organization has accumulated enough validated institutional signal to justify customization. But a Private LLM should be an optional downstream capability, not the starting assumption.

The institution should first create the data, evaluation, governance, and learning systems that make customization meaningful. If a model is trained or tuned, it should then be treated as a lifecycle asset.

LAYER	CADENCE	PURPOSE
Production monitoring	Always on	Track failures, unsupported claims, retrieval quality, latency, throughput, overrides, escalations, and audit records.
Regression testing	Periodic	Use a stable held-out test set to identify changes in model behavior, retrieval quality, citation performance, or reliability.
Comparative validation	Periodic or event-driven	Benchmark against the untuned baseline and strongest approved alternatives using held-out institutional test cases.

Material changes should trigger revalidation. Examples include retraining, a new tuning cycle, major corpus changes, quantization changes, runtime changes, a new task scope, or a material governance-policy change.

> The Private LLM is never assumed to remain the best model. The institution periodically retests that assumption.

CONCLUSION

The Path Forward: AI Independence Is Not Rejection. It Is Responsibility.

The seven steps are not a checklist. They are a framework for institutional resilience as AI becomes embedded in Investment Research and other consequential workflows.

Model independence creates optionality.

Governed intelligence makes model disagreement and evidence visible.

Risk-tier governance applies controls in proportion to consequence.

Continuous validation keeps customized models measurable and replaceable.

Economic sovereignty creates control over the cost curve.

Institutional learning preserves what the organization learns.

Institutional ownership keeps accumulated intelligence as an institutional asset.

> AI independence is not isolation from the AI ecosystem. It is control within it.

CLOSING PERSPECTIVE

The Durable Advantage

The competitive advantage is unlikely to be the foundation model itself. Models are becoming broadly available, increasingly capable, and increasingly interchangeable.

The durable advantage is what the institution builds around them: its data, evidence standards, workflows, evaluation history, decision context, governance logic, and validated institutional learning.

The firms that treat those elements as strategic assets will be able to adopt new models without starting over every time the technology changes.

> Maximum capability. Minimum structural waste. No gaps in control.

ABOUT ZTA LABS

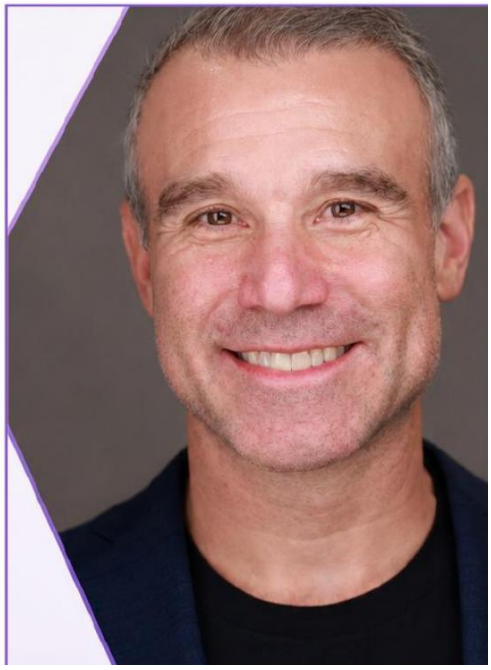
ZTA Labs helps asset managers and asset owners build AI capability they control, beginning with Investment Research. The company designs and deploys an open, institution-controlled AI architecture built on open-source software and open-weight models.

Its initial offerings include Model Assurance, which benchmarks open-weight models against an institution's actual questions, documents, data, and evaluation standards, and Research Management, a governed production workflow that connects approved evidence, model analysis, Human Judgment, decision records, and measurable outcomes.

CONTROL. GOVERN. BUILD. EVOLVE.

ABOUT THE AUTHOR

C.T. Rusert



C.T. Rusert is the founder of ZTA Labs and has 20 years of experience at the intersection of institutional financial markets, enterprise technology, and AI infrastructure.

He previously served as Worldwide Leader for Cloud Native Solutions, AI and High-Performance Computing at IBM, where his work spanned AI infrastructure, cloud-native architecture, GPU-accelerated computing, and production enterprise systems.

At Bloomberg, he served as Regional Leader for Buyside OMS and Performance Measurement & Risk (PORT), working with institutional investment organizations across portfolio management, risk analytics, performance measurement, and buy-side workflows.

SELECTED SOURCES

1. Gartner, "Gartner Predicts That by 2030, Performing Inference on an LLM With 1 Trillion Parameters Will Cost GenAI Providers Over 90% Less Than in 2025," March 25, 2026.
2. Anthropic, "Introducing Claude Opus 4.7" and Claude Platform Migration Guide, 2026.
3. WIRED, "OpenAI Models Escaped Containment and Hacked Hugging Face," July 21, 2026.
4. Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident," July 27, 2026.
5. UK AI Security Institute, "Incident Report: unsanctioned agent behaviour during cyber testing," August 2026.
6. WIRED, "OK, Well, Rogue AI Agents Are Hacking Again," August 4, 2026.