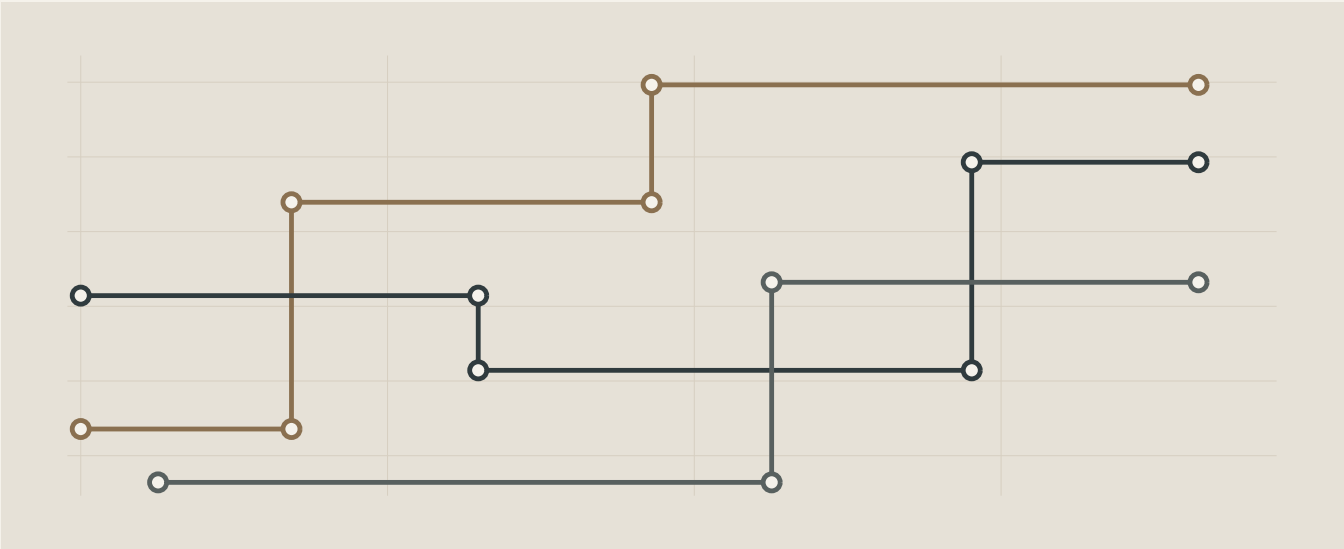
INTELLIGENCE / RESEARCH 001

Infrastructure Is Not Intelligence.

The Enterprise AI Architecture Gap

From Compute Capacity to Coordinated Intelligence
and Measurable Business Outcomes

EXECUTIVE RESEARCH WHITE PAPER



Independent Perspective. Operational Purpose.

CEREVANT ADVISORY / OCTOBER 2026

02 / EXECUTIVE PERSPECTIVE

The Enterprise AI Paradox

An enterprise can acquire more computing capacity, richer data platforms and more capable models while the work itself remains fragmented. A convincing demonstration can answer a question; an operational system must also identify the authoritative record, respect access rights, complete an approved action and recover when something fails.

Capacity is the availability of resources. Capability is the repeatable ability to deliver an agreed outcome under real operating constraints. The gap between them is architectural: who owns the data, how decisions enter a workflow, which controls apply and who is accountable for the result.

Published research on ML system debt describes risks arising from data dependencies, hidden feedback loops and entangled systems [4]. This supports examining the surrounding system, rather than treating model quality as a sufficient proxy for readiness. It does not establish a universal rate of pilot failure.

EXECUTIVE SUMMARY / CEREVANT ANALYTICAL FINDINGS

01 / Readiness is a property of the operating system.

Available GPUs or cloud services do not establish data permissions, service ownership, recovery procedures or user acceptance. Each must be designed and tested.

02 / A useful model needs a route into work.

Repeatability requires a defined trigger, trusted context, validation, a responsible decision-maker and a safe path into the enterprise application.

03 / Governance and measurement belong in the design.

NIST offers lifecycle risk-management guidance [1, 2]; ISO/IEC 42001 addresses an AI management system [3]. Neither substitutes for workload-specific acceptance criteria.

04 / Evidence should determine expansion.

Compare the full workflow with a baseline, including review effort, errors and operating cost. A bounded pilot should be able to stop, change or remain small.

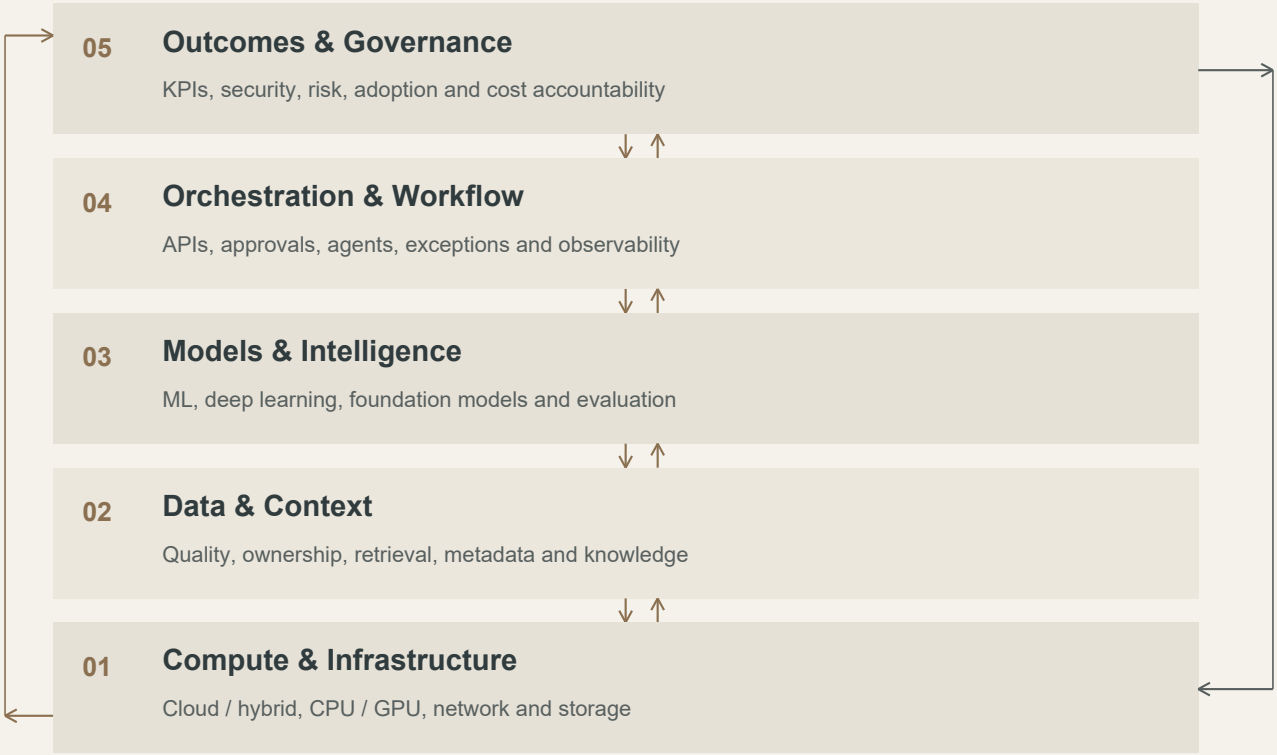
Intelligence does not emerge from infrastructure alone. It requires coordinated architecture, contextual data, accountable workflows and measurable operational purpose.

03 / PROPOSED ARCHITECTURE

From Infrastructure to Operational Intelligence

Cerevant proposes five interdependent layers as a review lens, not a prescribed technology stack or a maturity certification. A change in one layer can alter the cost, reliability or risk of the others.

FIGURE 01 / FIVE-LAYER ENTERPRISE INTELLIGENCE ARCHITECTURE



Two-way links represent dependencies and feedback. Outer paths return operating evidence to capacity and data decisions, and governance constraints to execution. Source: Cerevant proposed analytical framework.

Start with workload placement. Infrastructure supplies capacity and resilience; data supplies governed context. Models provide task-specific prediction or generation. Orchestration connects those capabilities to authorized work. Outcomes and governance establish the purpose, constraints and evidence for continued use.

Close the loop. A rising exception rate may reveal stale source data rather than a weak model. Slow adoption may reveal poor workflow fit rather than insufficient compute. Trace these signals across layers before increasing infrastructure spend or adding model complexity.

This decomposition is Cerevant analysis. The lifecycle emphasis is consistent with NIST risk-management guidance [1, 2]; the diagram itself is not a NIST or ISO model.

04 / THE ARCHITECTURE GAP

Why AI Pilots Fail to Become Operational Systems

A pilot tests feasibility within a bounded setting. Production adds variable inputs, access boundaries, downstream dependencies and service obligations. A prototype that works with a curated document set has not yet shown that it can operate safely across sites, roles and changing records.

NIST identifies risk management across design, development, use and evaluation [1, 2]. ML system research highlights dependencies that can accumulate maintenance debt [4]. The matrix below translates those concerns into Cerevant diagnostic hypotheses; it is not an empirical ranking of industry failure causes.

DIAGNOSTIC MATRIX / SYMPTOMS TO TEST, NOT ASSUMED FINDINGS

Symptom	Architectural cause to investigate	Operational consequence	Recommended intervention
Conflicting answers across teams	Fragmented ownership; no authoritative record	Decisions rely on inconsistent or stale context	Assign stewards; establish source priority and freshness rules.
Untracked tools and subscriptions	Shadow AI; fragmented procurement and identity	Unclear access, retention and cost exposure	Inventory tools; approve data use and centralize access policy.
Outputs copied manually into systems	Weak workflow and API integration	Rework; no dependable action trail	Define a system-of-record adapter, validation and rollback.
A strong demo degrades in daily use	Limited evaluation and observability	Silent quality drift; slow diagnosis	Build representative test sets, trace IDs and exception alerts.
Exceptions circulate without resolution	No operational owner or escalation path	Unreviewed decisions and unresolved queues	Name process and risk owners; set service and approval rules.
Spend grows without a business baseline	Capacity decisions detached from KPIs	No basis for judging operating value	Track cost per accepted outcome and compare with baseline.

Use the matrix in discovery

For each symptom, request an example, identify the affected decision and test the proposed cause. Agree what evidence would disprove the diagnosis. Correct a narrow source or integration problem before commissioning a broader AI program.

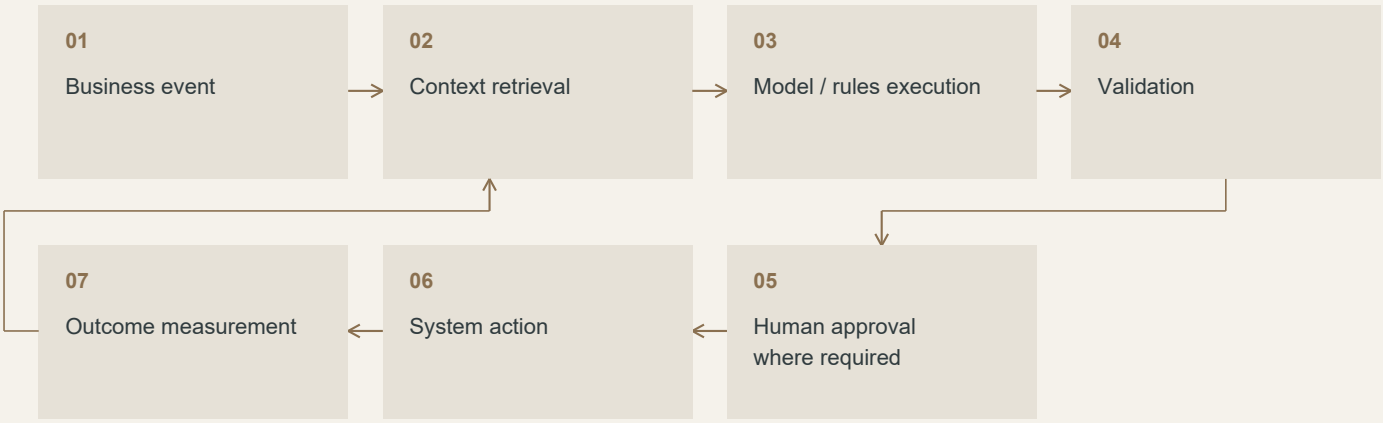
Evidence boundary: references support the system and lifecycle principles. The six diagnostic mappings and recommended interventions are proposed Cerevant analysis, not proprietary field research or verified client findings.

05 / COORDINATED WORKFLOWS

Orchestration Is the Missing Operating Discipline

A collection of AI tools becomes an operating capability when work has explicit states, responsibilities and controls. Decompose the task, route it to the simplest suitable method and define what happens when context is absent, confidence is insufficient or a downstream system is unavailable.

FIGURE 02 / A CONTROLLED WORKFLOW WITH A REVIEW LOOP



Measurement informs context and policy updates; failed validation and declined approvals enter an owned exception queue. This is an illustrative design, not a guarantee that every task needs every step.

Retrieve context without surrendering control

Retrieval-augmented generation combines retrieval with generation [5]. In an enterprise implementation, enforce source permissions before retrieval, retain source identifiers and test both retrieval relevance and answer support. Retrieved text is evidence, not permission to execute instructions.

Choose the method before the model

Use deterministic software for exact arithmetic, policy thresholds and idempotent writes. Use predictive ML for a defined classification or forecasting task when training and evaluation data fit. Use generative AI for bounded interpretation, drafting or synthesis where outputs can be checked. Evaluate total inference, retrieval, latency and human-review costs per accepted task.

Make actions bounded and recoverable

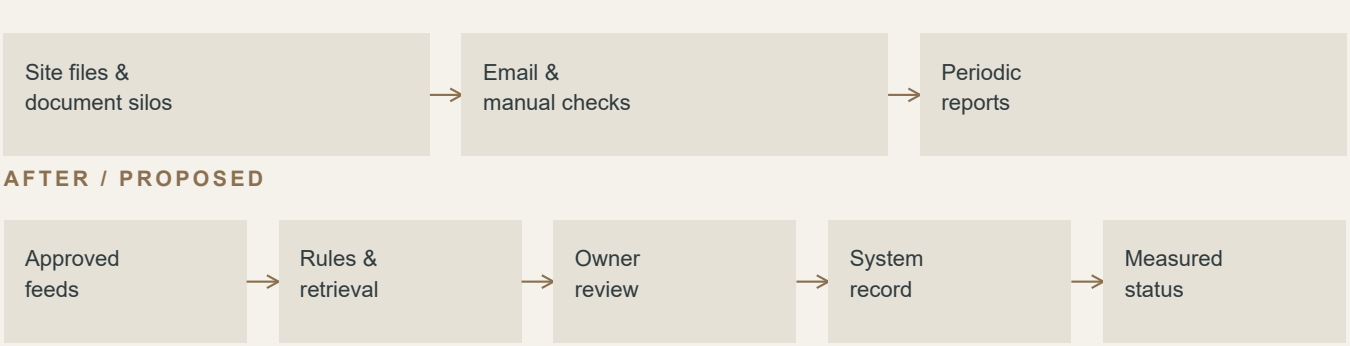
Integrate APIs with least-privilege credentials, explicit transaction limits and a durable audit trail. Validate schemas and business rules before action; require accountable approval for consequential exceptions. Monitor failures, retry safely, and route unresolved cases to people. Anthropic distinguishes predefined workflows from model-directed agents [6]; adopt agentic behavior only where that added discretion is justified.

06 / ILLUSTRATIVE ENTERPRISE USE CASE

From Fragmented Operations to Coordinated Intelligence

Consider a hypothetical organization operating hotels, restaurants and commercial properties alongside construction and maintenance activities. Labor, finance and vendor reports arrive in separate systems; contracts and operating documents sit across departmental repositories. This is an illustrative scenario, not a Cerevant client or a reported deployment.

FIGURE 03 / BEFORE AND PROPOSED AFTER



The proposed design combines rules-based labor and finance reconciliation, permission-aware contract and vendor search, and maintenance exception routing. Shared entity definitions and an event history make ownership and status visible across departments. Source-linked answers require review; authorized staff retain dispatch and commitment decisions.

Operating boundary. Begin with one region and one reconciliation workflow in read-only or shadow mode. Validate entity matching and accounting-period rules before allowing writes. Restrict payroll and contract data by role; separate recommendation from payment, employment or vendor commitment decisions.

PROPOSED SUCCESS CRITERIA / BASELINE AND THRESHOLDS TO BE AGREED

Measure	How the pilot would evaluate it
Reconciliation time	Median and 90th-percentile time from complete inputs to approved reconciliation; compare with the same workflow baseline.
Exception resolution time	Time from exception creation to authorized closure, segmented by type and severity.
Retrieval accuracy	Relevant-source retrieval and source-supported answer rate on an expert-reviewed test set; record access-control violations separately.
Adoption and operating cost	Eligible tasks completed through the workflow; total model, platform and review cost per accepted task.

No targets or realized benefits are asserted. Include error rates, reviewer overrides and service reliability as guardrails. Improvement is credible only if faster work retains control quality and does not shift uncounted effort to another team.

07 / PROPOSED CEREVANT ADVISORY FRAMEWORK

From Assessment to Operational Capability

The proposed engagement model creates evidence at each stage. It is a decision structure, not a fixed implementation package. Scope, access, operating risk and organizational readiness determine whether work advances, repeats or stops.

Stage	Work and evidence	Decision gate
01 / UNDERSTAND	Assess infrastructure, data maturity, workflows, governance and priorities. Produce a process map, dependency inventory and baseline.	Agree the problem, accountable owner and data access.
02 / ARCHITECT	Define target state, integration patterns, security boundaries and success criteria. Produce a design and evaluation plan.	Approve scope, controls, cost envelope and acceptance tests.
03 / OPERATIONALIZE	Design and validate a bounded proof of value with human oversight. Produce tested integrations, evaluations and runbooks.	Demonstrate safe operation and a workable exception path.
04 / MEASURE & EXPAND	Review performance, cost, reliability and adoption. Produce an evidence pack and a controlled expansion recommendation.	Expand only if criteria are met; otherwise refine or stop.

FIGURE 04 / ILLUSTRATIVE 90-DAY PILOT, SUBJECT TO READINESS

DAYS 01-15	DAYS 16-30	DAYS 31-65	DAYS 66-90
Understand	Architect	Operationalize	Measure & expand

Assumptions. A named sponsor and process owner are available; the selected workflow is bounded; approved data and API access can be provided promptly; procurement, security review and test environments are ready. Day ranges are planning assumptions, not a delivery commitment.

Dependencies. Data remediation, labor consultation, privacy review, vendor constraints and system changes can extend or halt the pilot. Start the operating clock only after readiness is confirmed. Keep an explicit stop condition for unacceptable risk, weak evidence or costs outside the agreed envelope.

Measurement discipline. Review the same metrics and denominators throughout the pilot. Include change-management and human-review effort in cost. Use representative cases, record model and prompt versions, and preserve a rollback route. Stage reviews should involve business, technology, security and risk owners.

The four-stage method and timeline are Cerevant proposals. They are not claims of ISO certification, NIST endorsement or an established record of client delivery.

08 / EXECUTIVE CONCLUSION

The Next Competitive Advantage Is Coordination.

Enterprise AI maturity is not determined by the number of models deployed, GPUs purchased or pilots launched. It is demonstrated by the ability to coordinate infrastructure, data, intelligence, workflows and human judgment into reliable operational outcomes.

The practical next step is to choose a workflow whose constraints are understood and whose outcome can be measured. Establish a baseline, assign accountability and make the path from information to action explicit. Scale the evidence, controls and operating capability together.

THREE QUESTIONS FOR THE EXECUTIVE TEAM

01 / Where is our current architecture preventing intelligence from reaching operational workflows?

02 / Which business processes offer the clearest measurable opportunity?

03 / What must be governed, integrated and evaluated before we scale?

CEREVANT ADVISORY

Enterprise Infrastructure. Applied Intelligence. Measurable Outcomes.

Discovery invitation: bring one priority workflow, its operating constraints and the outcome you need to improve.

INDEPENDENT PERSPECTIVE. OPERATIONAL PURPOSE.

REFERENCES / PRIMARY SOURCES

[1] Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. January 26. doi.org/10.6028/NIST.AI.100-1

[2] Autio, C., et al. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. July 26. doi.org/10.6028/NIST.AI.600-1

[3] ISO/IEC (2023). ISO/IEC 42001:2023 - Information technology - Artificial intelligence - Management system. Edition 1, December. Public overview; full standard not reviewed. iso.org/standard/42001

[4] Sculley, D., et al. (2015). Hidden Technical Debt in Machine Learning Systems. Advances in Neural Information Processing Systems 28. NeurIPS proceedings / original paper

[5] Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020; arXiv submitted May 22, 2020. arxiv.org/abs/2005.11401

[6] Anthropic (2024). Building effective agents. December 19. Engineering guidance; vendor experience, not an independent industry survey. anthropic.com / Building effective agents

Sources verified October 18, 2026. Standards and guidance are cited in their named editions. This paper offers a qualitative synthesis and proposed framework; no original empirical study, industry failure rate or client result is claimed.