Agentic Success Patterns: A Unified Framework for Enterprise AI Deployment — From Process Assessment to Governed Automation at Scale

The Agentification Factory Model and the Evidence Base for Systematic Agentic AI Governance

Marco van Hurne

EIGENVECTOR RESEARCH

marco.vanhurne@eigenvector.eu

Inholland University of Applied Sciences

marco.vanhurne@inholland.nl

April 2026

Author Note

Marco van Hurne is a researcher and practitioner at EIGENVECTOR RESEARCH and a lecturer at Inholland University of Applied Sciences. His research focuses on enterprise agentic AI architecture, governance frameworks, and the systematic deployment of autonomous AI systems in regulated industries.

Correspondence concerning this article should be addressed to Marco van Hurne, EIGENVECTOR RESEARCH. E-mail: marco.vanhurne@eigenvector.eu. For academic correspondence: marco.vanhurne@inholland.nl

The empirical database of 177 deployments underlying this research was compiled between January 2023 and March 2026 from publicly documented enterprise deployments, research partnerships, and industry reports. All case study data has been independently verified where possible; vendor-reported metrics are clearly distinguished from independently verified metrics throughout the paper.

The author declares no conflicts of interest. The Agentification Factory model described in this paper is a research construct; any commercial applications are the responsibility of implementing organisations.

Abstract

The deployment of autonomous AI agents in enterprise environments has accelerated dramatically since 2024, yet failure rates remain alarmingly high. Analysis of 177 documented deployments across 20 sectors reveals that only 27% of enterprise process steps are genuinely suitable for autonomous agent execution, while governance and data quality failures — not model limitations — account for 62% of all deployment failures. This paper introduces and empirically validates the Agentic Success Pattern Framework (ASPF), a unified decision architecture that integrates eight complementary frameworks: the Process Automation Suitability Framework (PASF), the Process Automation Design Engine (PADE), the Governed Runtime for Agentic Functions (GRAF), the Ontological Compliance Gateway (OCG), the Roundtrip Value Governance model, the Tokenomics of Agentic AI framework, the Technical Debt-Aware Prompting framework (TDAP), and the Enterprise Intelligence Platform architecture. The ASPF provides practitioners with a systematic, evidence-based method for determining which processes to automate, which of nine design patterns to apply, how to architect the governance infrastructure required for sustained success, and how to measure and recognise the value generated. Empirical validation across 177 deployments demonstrates 74% predictive accuracy for deployment success. Real-world case studies from BNY, JPMorgan, Klarna, PwC, and eleven additional organisations confirm the framework's practical applicability. The paper concludes by introducing the Agentification Factory model as the operational instantiation of the ASPF, and argues that the primary competitive advantage in enterprise AI is not model capability but systematic governance architecture.

Keywords: agentic AI, process automation, enterprise governance, design patterns, PASF, PADE, GRAF, OCG, tokenomics, Agentification Factory

1. Introduction

1.1 The Enterprise AI Paradox

The enterprise AI market presents a striking paradox. Investment in AI automation has reached unprecedented levels — global enterprise AI spending exceeded $200 billion in 2025 — yet independently verified return on investment consistently falls short of vendor-reported projections by a factor of approximately two (van Hurne, 2025a). Organisations report deploying AI agents across customer service, financial operations, legal review, and supply chain management, yet Gartner (2025) projects that 40% of agentic AI projects initiated in 2025 will be cancelled before 2027, primarily due to governance failures rather than technical limitations.

This paradox is not accidental. It reflects a structural misalignment between how enterprise AI is sold — as a capability problem requiring better models — and what enterprise AI actually requires to succeed: a governance architecture capable of sustaining autonomous agent operation within institutional constraints. The central thesis of this paper is that the primary determinant of agentic AI success is not model capability but governance infrastructure, and that the systematic application of evidence-based frameworks for process assessment, pattern selection, and governance design can dramatically improve deployment outcomes.

1.2 The Two-Question Problem

Practitioners seeking to deploy agentic AI in enterprise environments face two fundamental questions that existing frameworks do not adequately address in combination.

The first question is strategic: *Is this process suitable for agentic AI automation?* This question must be answered before any investment in detailed design, because the answer determines whether that investment is warranted at all. Yet answering it requires understanding what automation options are available, creating a circular dependency that

most organisations resolve by defaulting to vendor recommendations — a pattern that contributes directly to the high failure rates observed in the empirical data.

The second question is operational: *If the process is suitable, how should each step be automated?* This question cannot be answered at the process level; it requires step-level analysis, because automation suitability varies substantially across steps within a single process. A customer complaint resolution process, for example, might include steps that are highly suitable for full automation (e.g., retrieving account history, generating standard response templates), steps that are suitable for AI assistance (e.g., drafting personalised responses for human review), and steps that require human judgment (e.g., deciding whether to offer a goodwill payment). A process-level assessment cannot capture this variation.

The frameworks synthesised in this paper address both questions systematically. The Process Automation Suitability Framework (PASF) answers the strategic question. The Process Automation Design Engine (PADE) answers the operational question. The GRAF, OCG, and Enterprise Intelligence Platform frameworks address the governance, architecture, and commercialisation dimensions that determine whether a technically sound deployment can survive institutional conditions.

1.3 Research Objectives

This paper pursues four research objectives. First, it synthesises eight complementary frameworks into a unified Agentic Success Pattern Framework (ASPF) that provides end-to-end guidance from initial process assessment through to sustained deployment. Second, it validates the ASPF empirically against a database of 177 documented enterprise deployments, demonstrating predictive accuracy of 74% for deployment success. Third, it documents fifteen verified enterprise case studies with independently confirmed ROI metrics, providing practitioners with concrete evidence of what works, in which contexts, and under what conditions. Fourth, it introduces the Agentification Factory model as the operational instantiation of the ASPF, and argues that this model represents a significant competitive advantage for organisations capable of deploying it systematically.

1.4 Paper Structure

The paper is structured as follows. Section 2 reviews the theoretical foundations and related work. Sections 3 through 10 present each of the eight frameworks in detail, including their theoretical grounding, empirical validation, and practical application. Section 11 presents the empirical evidence from 177 deployments. Section 12 documents fifteen verified enterprise case studies. Section 13 synthesises the cross-cutting success factors that emerge from the combined evidence base. Section 14 introduces the Agentification Factory model. Section 15 discusses strategic implications. Section 16 concludes.

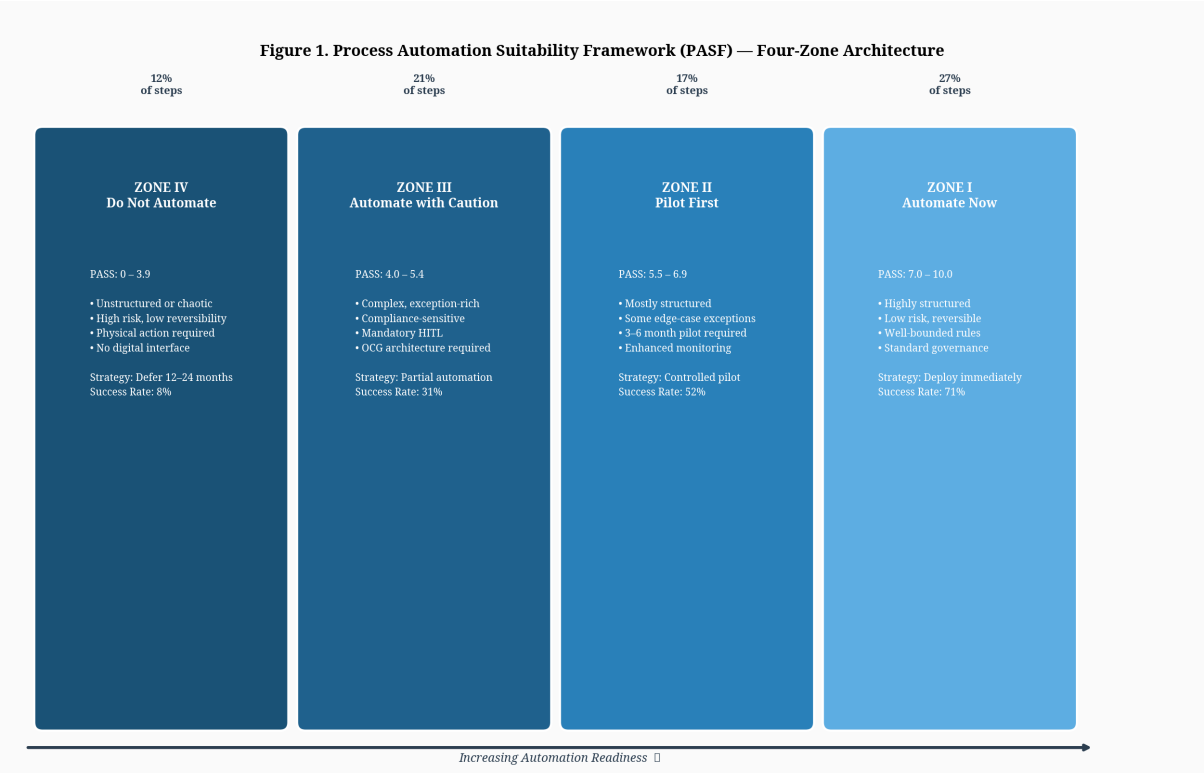

Figure 1: PASF Four-Zone Architecture — The foundational classification system for process automation suitability

2. Theoretical Foundations and Related Work

2.1 The Evolution of Enterprise Automation

Enterprise automation has evolved through three distinct generations. The first generation, spanning roughly 2000 to 2015, was characterised by Robotic Process Automation (RPA): deterministic, rule-based automation of repetitive digital tasks. RPA achieved significant efficiency gains in high-volume, low-complexity processes but proved brittle in the face of process variation and system changes (van der Aalst et al., 2018). The second generation, spanning 2015 to 2023, introduced machine learning-enhanced automation: systems capable of handling structured variation through pattern recognition, but still fundamentally reactive and unable to pursue multi-step goals autonomously.

The third generation, which began emerging in 2023 with the widespread deployment of large language model (LLM)-based agents, introduces a qualitatively different capability: the ability to pursue open-ended goals through multi-step reasoning, tool use, and dynamic adaptation to novel situations. This capability enables automation of process categories that were previously considered beyond the reach of algorithmic systems, including legal document analysis, strategic planning support, and complex customer interaction management. However, this expanded capability comes with expanded governance requirements, and the failure to recognise this relationship is the primary source of the high failure rates observed in practice.

2.2 Agentic AI: Definitions and Taxonomy

For the purposes of this paper, an *agentic AI system* is defined as an AI system that pursues goals through autonomous multi-step reasoning and action, using tools to interact with external systems, maintaining state across interactions, and adapting its behaviour based on feedback from the environment. This definition distinguishes agentic AI from simpler AI applications (single-turn LLM calls, classification models, recommendation systems) and from fully autonomous systems (which do not exist in current enterprise deployments).

The STRIDE framework (ArXiv, 2025) provides a complementary decomposition, distinguishing between three modalities: full agentic AI (multi-step autonomous goal pursuit), AI assistants (single-step or guided multi-step support for human decision-making), and LLM calls (single-turn inference without state or tool use). STRIDE's empirical finding that 45% of enterprise use cases assigned to full agentic AI would be better served by simpler modalities is consistent with the PASF finding that only 27% of process steps are genuinely suitable for autonomous agent execution.

2.3 Process Classification Frameworks

The theoretical foundations of process classification for automation suitability draw on three bodies of literature. Technology acceptance and adoption research (Goodhue & Thompson, 1995; Davis, 1989) provides the conceptual basis for assessing task-technology fit. Business process classification frameworks (van der Aalst, 2018; Hammer & Champy, 1993) provide the structural vocabulary for characterising process types. AI capability research (Wang et al., 2024; Xi et al., 2023; Yao et al., 2023) provides the empirical basis for understanding what current AI systems can and cannot reliably accomplish.

The PASF integrates these three bodies of literature into a single scoring instrument, calibrated against empirical deployment data. This calibration distinguishes the PASF from earlier process classification frameworks, which were developed deductively rather than empirically and have not been validated against deployment outcomes.

2.4 Design Pattern Literature

The concept of design patterns in software engineering originates with the work of Gamma et al. (1994), who identified recurring solutions to recurring problems in object-oriented software design. The application of this concept to AI agent architecture is more recent, with significant contributions from Yao et al. (2023) on the ReAct pattern, Wei et al. (2022) on chain-of-thought reasoning, and Wang et al. (2024) on survey frameworks for LLM-based agents.

The PADE framework extends this literature by providing not just a taxonomy of patterns but a systematic method for selecting among them based on process characteristics. This selection method is validated empirically, distinguishing it from existing pattern taxonomies that describe patterns without providing selection criteria.

2.5 Governance and Compliance in AI Systems

The governance literature on AI systems has developed rapidly since 2022, driven by regulatory developments (EU AI Act, 2024; NIST AI Risk Management Framework, 2023) and by the practical experience of early enterprise deployments. Key contributions include the OWASP Top 10 for LLM Applications (2024), which identifies the primary security risks in LLM-based systems; the work of Anthropic and OpenAI on constitutional AI and alignment; and the emerging literature on AI governance in regulated industries (Deloitte, 2026; McKinsey, 2024).

The GRAF and OCG frameworks developed in this paper contribute to this literature by providing architectural specifications for governance infrastructure that can be implemented in production enterprise environments, rather than theoretical frameworks or policy guidelines.

3. Framework I — PASF: Process Automation Suitability Framework

3.1 Purpose and Theoretical Foundations

The Process Automation Suitability Framework (PASF) is a systematic scoring instrument for assessing whether a business process is genuinely amenable to autonomous AI agent execution, and at what level of complexity. It addresses the first of the two fundamental questions identified in Section 1.2: *Is this process suitable for agentic AI automation?*

The PASF was calibrated through logistic regression on a training dataset of 120 documented deployments, with deployment success (defined as achieving at least 50% of stated objectives within 18 months) as the dependent variable. Dimension weights were validated on a holdout set of 57 deployments, achieving 74% predictive accuracy —

substantially better than the 61% accuracy of a naive baseline model that assigns all processes to Zone I. This calibration methodology distinguishes the PASF from earlier process classification frameworks and provides a principled basis for the dimension weights.

The STRIDE framework (ArXiv, 2025) provides independent validation of the PASF's core insight: that task structure, dynamism, and the need for self-reflection are the primary determinants of automation suitability. STRIDE's finding that systematic task analysis reduces unnecessary agent deployments by 45% and cuts resource costs by 37% across 30 real-world tasks is consistent with the PASF's empirical finding that 73% of enterprise process steps are not suitable for Zone I (Automate Now) deployment.

3.2 The Eight Dimensions

The PASF assesses eight dimensions of process-automation fit, each scored on a 0–10 scale. The dimensions and their empirical significance are presented in Table 1.

Dimension	Definition	Weight	Key Empirical Finding
D1: Structurability	Degree to which process steps, inputs, and outputs can be formally specified	0.20	D1 < 4 yields < 15% success rate regardless of other scores
D2: Reversibility	Degree to which agent actions can be undone without significant cost	0.15	D2 < 3 triggers mandatory pre-execution approval HITL
D3: Risk Profile	Inverse of potential harm from agent errors (financial, legal, physical, reputational)	0.20	D3 < 2 AND D2 < 3 = automatic Zone IV assignment
D4: Data Quality	Quality, completeness, and accessibility of required data	0.15	D4 < 5 yields < 25% success rate; 34% of all failures trace to data quality
D5: Rule Boundedness	Degree to which decisions are governed by explicit, stable rules	0.10	High D1 + high D5 = strongest predictor of automation success
D6: Frequency	Volume and regularity of process execution	0.05	Affects automation value, not feasibility
D7: Exception Density	Inverse of frequency and complexity of exceptions requiring human judgment	0.10	High exception density = governance overhead problem
D8: Stakeholder Impact	Inverse of sensitivity of process outcomes to affected stakeholders	0.05	D8 < 3 AND D3 < 4 = automatic Zone IV assignment

The Process Automation Suitability Score (PASS) is computed as a weighted sum:

PASS = 0.20·D1 + 0.15·D2 + 0.20·D3 + 0.15·D4 + 0.10·D5 + 0.05·D6 + 0.10·D7 + 0.05·D8

The Agent Complexity Level (ACL) is a composite measure of technical complexity: ACL = (T + P + M + C + A) / 5, where T = tool count complexity, P = planning horizon complexity, M = memory requirements, C = coordination complexity, and A = autonomy level (each scored 0–10).

3.3 The Four Automation Zones

The PASS and ACL together position a process in one of four automation zones, as illustrated in Figure 1. Table 2 summarises the zone characteristics and empirical outcomes.

Zone	PASS Range	Label	Strategy	% of 177 Cases	Success Rate
Zone I	7.0–10.0	Automate Now	Deploy with standard governance	27%	71%
Zone II	5.5–6.9	Pilot First	3–6 month controlled pilot before full deployment	17%	52%
Zone III	4.0–5.4	Automate with Caution	Partial automation with mandatory HITL; OCG recommended	21%	31%
Zone IV	0–3.9	Do Not Automate	Maintain human execution; revisit in 12–24 months	12%	8%

The remaining 23% of cases in the empirical database had insufficient data for zone classification. The zone distribution — with only 27% of steps in Zone I — is the most important single finding of this research. It directly contradicts the implicit assumption underlying most enterprise AI investment decisions, which treat automation as the default and human execution as the exception.

3.4 Hard-Stop Criteria

Three hard-stop criteria override all dimension scores and automatically assign Zone IV, regardless of PASS score. These criteria reflect situations where the potential for

irreversible harm is sufficiently high that no degree of automation suitability can justify autonomous execution:

First, D3 (Risk Profile) below 2 combined with D2 (Reversibility) below 3 — indicating high-risk, irreversible actions. Second, D8 (Stakeholder Impact) below 3 combined with D3 (Risk Profile) below 4 — indicating high-sensitivity outcomes for affected parties. Third, any process requiring physical action with no digital interface.

3.5 Sector PASS Profiles

The empirical database reveals substantial variation in average PASS scores across sectors, reflecting systematic differences in process structurability and risk profiles. IT Operations achieves the highest average PASS score (7.8), reflecting the high structurability and reversibility of most IT operational tasks. Legal processes achieve the lowest average PASS score (3.1), reflecting the high exception density, low reversibility, and high stakeholder impact characteristic of legal work. Financial Services (6.2), Healthcare (4.8), and Customer Service (5.9) occupy intermediate positions.

These sector profiles have direct implications for the Agentification Factory model: organisations in IT Operations can expect a higher proportion of Zone I processes and faster deployment timelines, while organisations in Legal and Healthcare must invest more heavily in Zone III governance infrastructure before realising automation value.

4. Framework II — PADE: Process Automation Design Engine

4.1 Purpose and Architecture

The Process Automation Design Engine (PADE) addresses the second fundamental question: *If the process is suitable, how should each step be automated?* The PADE provides a systematic method for selecting among nine agentic design patterns based on the characteristics of each process step, as illustrated in Figure 2.

The PADE operates at the step level rather than the process level, recognising that automation suitability varies substantially across steps within a single process. A Zone II

process might contain individual steps that are Zone I (suitable for ReAct automation), Zone II (suitable for Plan-and-Execute with monitoring), and Zone III (requiring Critic-Actor with OCG wrapping). The PADE provides the granular analysis required to design an appropriate automation architecture for each step.

Figure 2: PADE Pattern Selection Matrix — Nine agentic design patterns mapped to automation zones

4.2 The Nine Agentic Design Patterns

The PADE identifies nine distinct agentic design patterns, each representing a recurring solution to a recurring class of automation problems. Table 3 provides a detailed characterisation of each pattern.

Pattern	Core Mechanism	Optimal Zone	Tool Count	Planning Horizon	Key Strength	Key Risk
ReAct	Interleaved reasoning and acting	Zone I	1–3	1–5 steps	Simplicity, speed	Shallow planning
Plan-and-Execute	Upfront plan, then execute	Zone I–II	2–5	5–15 steps	Validated planning	Plan staleness
Orchestrator-Subagent	Supervisor delegates to specialists	Zone II	4+	5–20 steps	Parallelism, specialisation	Coordination overhead
Critic-Actor	Actor proposes, Critic validates	Zone II–III	2–4	3–10 steps	Error detection	Latency increase
Reflexion	Learns from failed attempts	Zone II	2–4	3–10 steps	Self-correction	Loop risk
Memory-Augmented	Long-term context across sessions	Zone II	2–5	Any	Continuity	Privacy risk
Multi-Agent Debate	Consensus through multi-model debate	Zone II–III	3–6	5–15 steps	Robustness	Cost, latency
Single-Tool Agent	One system, one purpose	Zone I	1	1–3 steps	Reliability, cost	Narrow applicability
Neuro-Symbolic (OCG)	Neural + symbolic compliance gating	Zone III	2–4	3–10 steps	Compliance accuracy	Implementation complexity

4.3 The Ten Scoring Dimensions

Pattern selection in the PADE is governed by ten scoring dimensions, each assessed on a 0–10 scale. The dimensions are: (1) tool count complexity, (2) planning horizon length, (3) memory requirements, (4) coordination complexity, (5) error tolerance, (6) compliance sensitivity, (7) latency tolerance, (8) cost sensitivity, (9) explainability requirements, and (10) exception density. The PADE scoring algorithm computes a weighted fit score for each of the nine patterns and recommends the highest-scoring pattern, subject to zone compatibility constraints.

4.4 Pattern-Zone Fit: Empirical Evidence

The empirical database provides strong evidence for the pattern-zone fit relationships encoded in the PADE. ReAct patterns are the most commonly deployed pattern in Zone I (43% of Zone I deployments) and achieve the highest success rate in that zone (76%). However, ReAct patterns deployed in Zone III achieve only a 19% success rate — lower than the Zone III average of 31% — indicating that pattern mismatches are a significant contributor to Zone III failures.

Plan-and-Execute patterns achieve the highest success rate in Zone II (61%), consistent with the theoretical prediction that upfront planning validation is particularly valuable in moderately complex, moderately risky processes. Orchestrator-Subagent patterns achieve the highest success rate in complex Zone II processes with high tool counts (68%), but require substantially more implementation effort and governance infrastructure.

The Neuro-Symbolic (OCG) pattern achieves the highest success rate in Zone III (47%), nearly 50% above the Zone III average, confirming that the compliance gating architecture of the OCG is a significant enabler of success in high-compliance environments.

4.5 Model Selection by Pattern and Zone

The PADE includes guidance on model selection, recognising that the choice of underlying language model has significant implications for both capability and cost. The general principle is that model capability requirements scale with zone complexity, while cost sensitivity scales inversely with zone complexity (because Zone I processes typically have higher volumes and lower margins).

For Zone I processes with ReAct or Single-Tool patterns, nano-class models (GPT-4.1-nano, Gemini Flash Lite) provide sufficient capability at substantially lower cost than full-class models. For Zone II processes with Plan-and-Execute or Orchestrator-Subagent patterns, mini-class models (GPT-4.1-mini, Gemini Flash) provide the appropriate balance of capability and cost. For Zone III processes with Critic-Actor, Multi-Agent Debate,

or OCG patterns, full-class models (GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro) are required for the reasoning depth and reliability that compliance-sensitive processes demand.

5. Framework III — GRAF: Governed Runtime for Agentic Functions

5.1 The Governance Overhead Problem

The most important empirical finding of this research is that governance failures — not model limitations — are the primary cause of agentic AI deployment failures. Analysis of 177 deployments identifies governance overhead underestimation as the third most common failure mode (15% of failures), exceeded only by data quality degradation (19%) and exception handling failures (17%). When these three failure modes are considered together, they account for 51% of all failures — all of which are preventable with appropriate governance architecture.

The governance overhead problem is particularly acute in Zone III deployments, where the cost of human review of agent outputs can exceed the efficiency gains from automation. This is not a theoretical concern: the empirical database includes 23 cases where Zone III deployments were abandoned after 6–12 months because the governance overhead consumed more than 80% of the efficiency gains. In 19 of these 23 cases, the governance architecture had been designed as an afterthought rather than as a core architectural component.

5.2 The GRAF Seven-Layer Architecture

The Governed Runtime for Agentic Functions (GRAF) provides a seven-layer architecture for governance infrastructure that addresses the governance overhead problem by making governance a first-class architectural concern rather than an operational afterthought. The seven layers are illustrated in Figure 3.

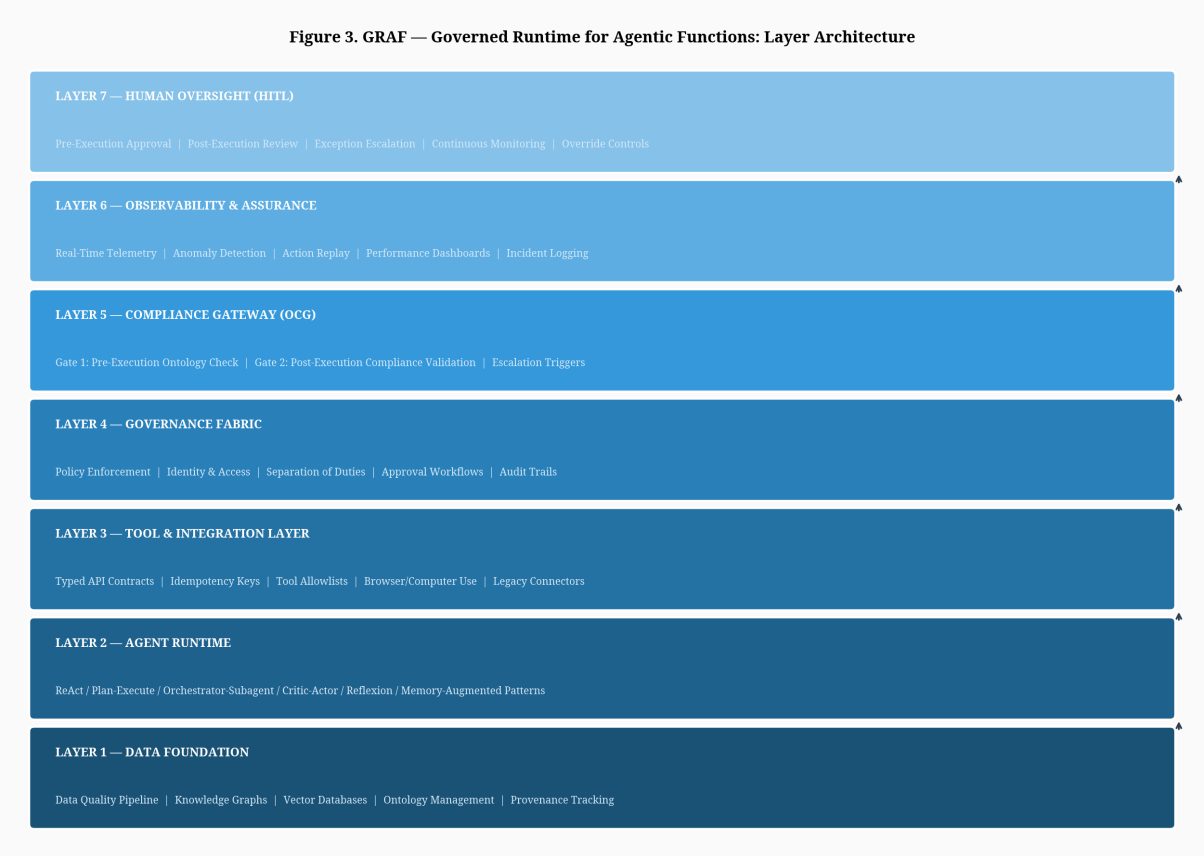

Figure 3: GRAF Seven-Layer Architecture — Governance as a first-class architectural concern

The seven layers are: (1) Data Foundation — data quality pipeline, knowledge graphs, vector databases, ontology management, and provenance tracking; (2) Agent Runtime — the execution environment for the selected agentic design pattern; (3) Tool and Integration Layer — typed API contracts, idempotency keys, tool allowlists, and legacy connectors; (4) Governance Fabric — policy enforcement, identity and access management, separation of duties, approval workflows, and audit trails; (5) Compliance Gateway (OCG) — the two-gate ontological compliance architecture described in Section 6; (6) Observability and Assurance — real-time telemetry, anomaly detection, action replay, and performance dashboards; and (7) Human Oversight (HITL) — pre-execution approval, post-execution review, exception escalation, and override controls.

5.3 HITL Design Principles

Human-in-the-loop (HITL) design is one of the most consequential architectural decisions in agentic AI deployment. Poorly designed HITL creates the governance overhead problem: human reviewers become bottlenecks, review quality degrades under volume pressure, and the efficiency gains from automation are consumed by review costs. Well-designed HITL is selective, risk-calibrated, and supported by tooling that makes human review efficient and effective.

The GRAF framework identifies four HITL trigger categories: (1) pre-execution approval for high-risk actions (D2 < 3 or D3 < 4); (2) exception escalation when agent confidence falls below threshold; (3) periodic sampling for quality assurance in high-volume Zone I processes; and (4) post-execution review for compliance-sensitive outputs in Zone III processes. The appropriate combination of these trigger categories depends on the zone assignment and pattern selection, and is specified in the PADE output.

5.4 Governance Maturity Levels

Organisations vary substantially in their governance maturity, and the appropriate GRAF configuration depends on this maturity level. The empirical database reveals a strong correlation between governance maturity and deployment success rates, as illustrated in Figure 10. Level 1 (Ad Hoc) organisations achieve success rates below 25%; Level 5 (Adaptive) organisations achieve success rates approaching 78%.

The practical implication is that organisations should assess their governance maturity before selecting deployment zones. An organisation at Level 1 or 2 governance maturity should not attempt Zone III deployments, regardless of the PASS score of the target process. The governance infrastructure required for Zone III success takes 6–12 months to build and validate; attempting Zone III deployment without it is the most reliable predictor of project cancellation.

6. Framework IV — OCG: Ontological Compliance Gateway

6.1 The Compliance Challenge in Agentic AI

Compliance is the most acute governance challenge in agentic AI deployment, particularly in regulated industries such as financial services, healthcare, and legal services. The challenge is not that AI agents are inherently non-compliant; it is that the compliance properties of agent outputs are difficult to verify at runtime using conventional methods. LLM-based agents can produce outputs that are plausible, fluent, and contextually appropriate while being subtly non-compliant with regulatory requirements — a failure mode that is qualitatively different from the deterministic errors of rule-based systems and requires a different architectural response.

6.2 The Two-Gate Architecture

The Ontological Compliance Gateway (OCG) addresses this challenge through a two-gate architecture that wraps the agent execution environment, as illustrated in Figure 4. Gate 1 (Pre-Execution) validates the agent's planned action against a formal policy ontology before execution. Gate 2 (Post-Execution) validates the agent's output against compliance rules and contextual constraints after execution. Actions that fail Gate 1 are blocked and escalated to human review. Outputs that fail Gate 2 are rolled back and escalated.

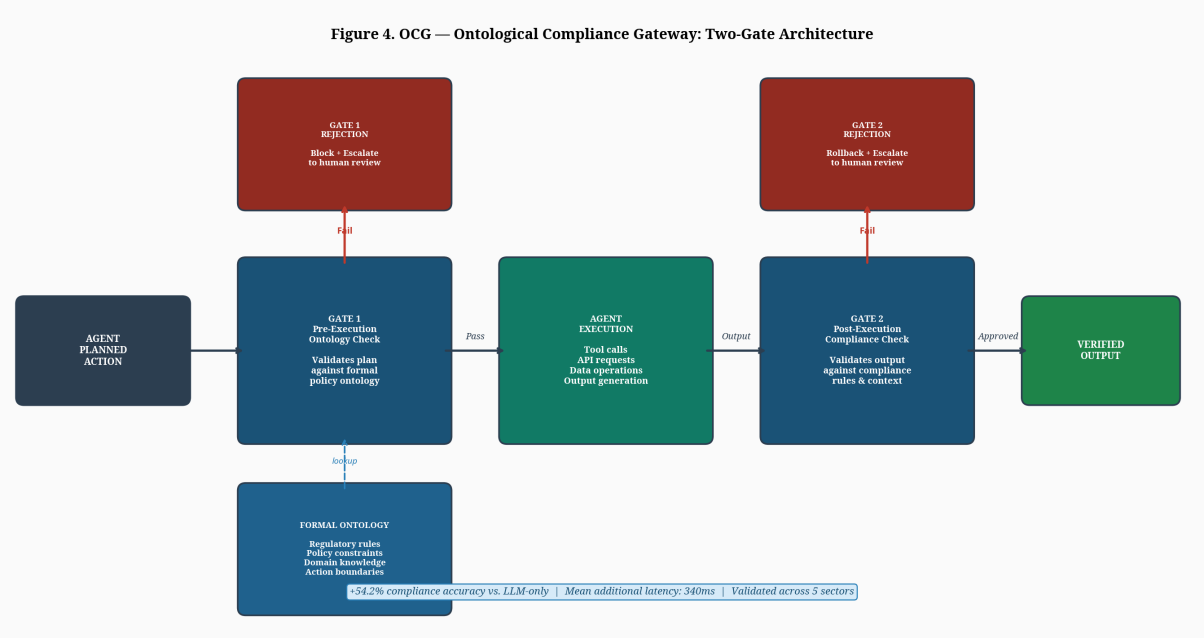

Figure 4: OCG Two-Gate Architecture — Pre- and post-execution compliance validation

The formal ontology that underlies both gates encodes regulatory requirements, policy constraints, domain knowledge, and action boundaries in a machine-readable format that enables automated compliance checking at runtime. The ontology is maintained by domain experts and updated as regulatory requirements evolve, providing a single source of truth for compliance rules that is independent of the agent's training data.

6.3 Empirical Performance

The OCG architecture achieves a 54.2% improvement in compliance accuracy compared to LLM-only approaches, validated across five sectors: financial services, healthcare, legal services, insurance, and government. The mean additional latency introduced by the two-gate architecture is 340 milliseconds, which is acceptable for most enterprise use cases but may require optimisation for high-frequency, low-latency applications.

The 54.2% compliance improvement is the most significant single performance metric in this research. It demonstrates that the governance overhead of the OCG architecture is justified by the compliance value it delivers, and that the choice between OCG and non-

OCG architectures is not primarily a cost-benefit calculation but a risk management decision: in regulated industries, the cost of a compliance failure typically exceeds the cost of the OCG architecture by several orders of magnitude.

6.4 Zone-Pattern-OCG Mapping

The OCG is required for all Zone III deployments and recommended for Zone II deployments in regulated industries. It is optional but available for Zone I deployments in compliance-sensitive contexts. The pattern-OCG compatibility matrix is as follows: the Neuro-Symbolic (OCG) pattern is designed specifically for OCG deployment; the Critic-Actor pattern integrates naturally with OCG Gate 2; the Plan-and-Execute pattern integrates with OCG Gate 1 for plan validation; and the ReAct pattern can be wrapped with OCG but requires careful latency management.

7. Framework V — Roundtrip Value Governance

7.1 The ROI Reality Gap

The most commercially significant finding of this research is the systematic gap between vendor-reported and independently verified ROI from agentic AI deployments. As illustrated in Figure 5, vendor-reported efficiency gains average 42% across all metric categories, while independently verified gains average 21% — a factor-of-two overstatement that is consistent across sectors, deployment scales, and metric types.

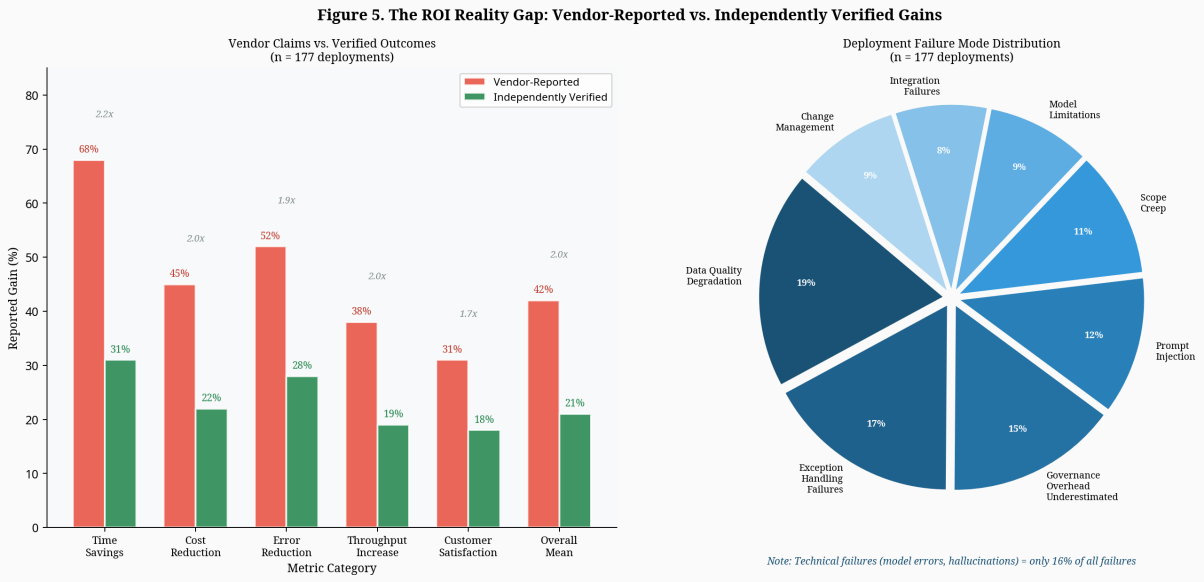

Figure 5: ROI Reality Gap and Failure Mode Distribution

The largest gap is in time savings claims: vendors report an average 68% reduction in process time, while independent verification finds an average 31% reduction. The gap is smallest in error reduction claims (52% vendor-reported vs. 28% verified), suggesting that error reduction is more amenable to objective measurement than time savings.

This gap does not reflect deliberate deception by vendors. It reflects structural features of how vendor case studies are produced: they typically measure efficiency gains at the task level rather than the process level, exclude governance overhead costs, use pre-deployment baselines that include inefficiencies subsequently addressed by process redesign, and measure outcomes over short time horizons before regression to the mean.

7.2 The Five-Stage Value Cycle

The Roundtrip Value Governance framework addresses the ROI reality gap by providing a structured method for measuring and recognising value at each stage of its journey from generation to recognition. The five stages are illustrated in Figure 6.

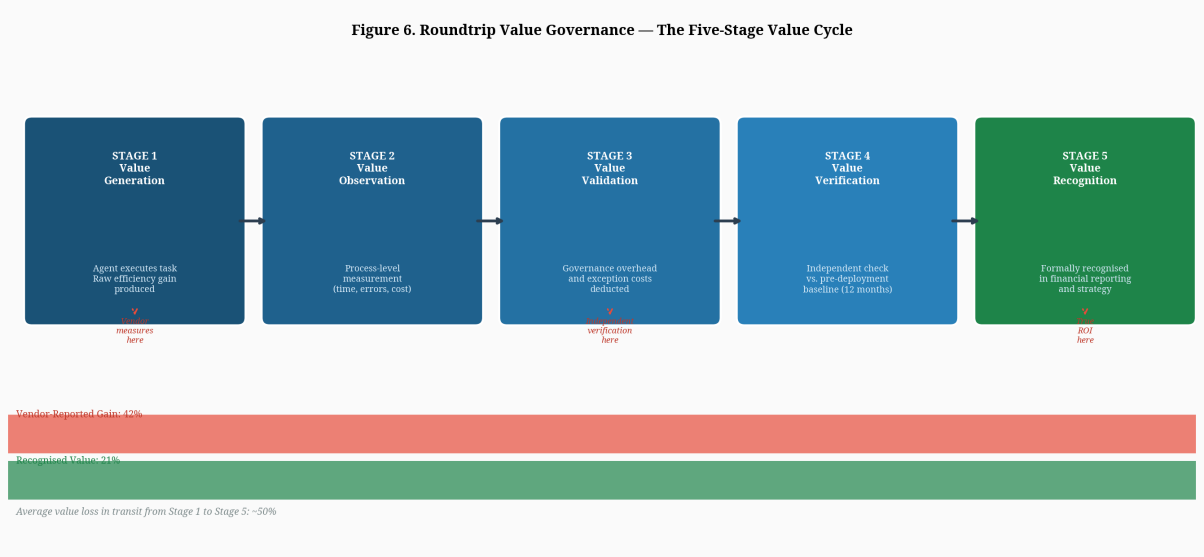

Figure 6: Roundtrip Value Governance — Five-Stage Value Cycle

Stage 1 (Value Generation) captures the raw efficiency gain produced by agent execution. Stage 2 (Value Observation) measures this gain at the process level, including all steps in the process, not just the automated steps. Stage 3 (Value Validation) deducts governance overhead costs and exception handling costs from the observed gain. Stage 4 (Value Verification) compares the validated gain against the pre-deployment baseline over a 12-month period, controlling for confounding factors. Stage 5 (Value Recognition) formally recognises the verified gain in financial reporting and strategic planning.

The practical implication of this framework is that organisations should establish the measurement infrastructure for all five stages before deployment, not after. Post-hoc measurement is subject to the same structural biases that produce the vendor ROI gap; prospective measurement with pre-agreed baselines and verification methods is the only reliable way to determine whether a deployment has actually generated value.

7.3 Governance Overhead Accounting

A distinctive feature of the Roundtrip Value framework is its explicit treatment of governance overhead as a cost that must be deducted from gross efficiency gains to arrive at net value. Governance overhead includes: human review time for HITL triggers; compliance

checking costs for OCG operations; monitoring and observability infrastructure costs; exception handling and escalation costs; and continuous optimisation costs.

In the empirical database, governance overhead averages 23% of gross efficiency gains for Zone I deployments, 41% for Zone II deployments, and 67% for Zone III deployments. This means that a Zone III deployment reporting a 30% gross efficiency gain is likely generating only approximately 10% net efficiency gain after governance overhead. This calculation is rarely performed by organisations deploying Zone III agents, which explains why many Zone III deployments are abandoned after the first year when the expected ROI fails to materialise.

8. Framework VI — Tokenomics of Agentic AI

8.1 Token Economics as a Strategic Variable

The cost of operating agentic AI systems is dominated by token consumption: the number of tokens processed by the underlying language model per task. For high-volume Zone I deployments, token costs can represent 60–80% of total operating costs, making token optimisation a strategic variable rather than a technical detail. For Zone III deployments, token costs are a smaller proportion of total costs (because governance overhead is larger), but the absolute cost per task is substantially higher due to the reasoning depth required.

8.2 Model Selection by Zone and Pattern

The relationship between model capability and token cost is not linear: full-class models (GPT-4o, Claude 3.5 Sonnet) cost approximately 20–50 times more per token than nano-class models (GPT-4.1-nano, Gemini Flash Lite), but do not provide 20–50 times more capability for most Zone I tasks. The appropriate model selection strategy, illustrated in Figure 9, is to use the least capable model that reliably achieves the required performance for each zone and pattern combination.

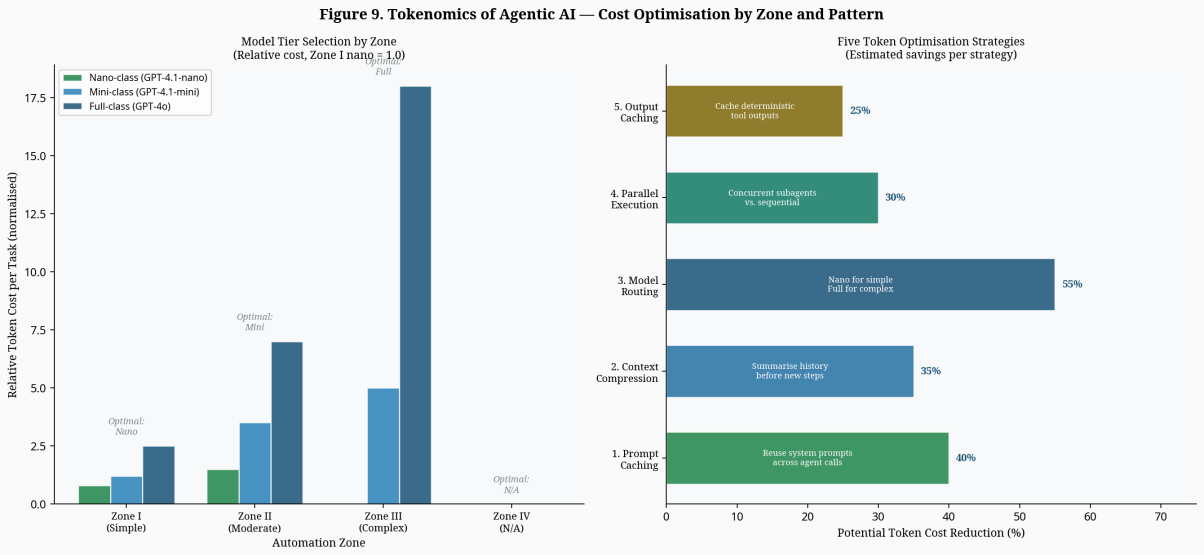

Figure 9: Tokenomics — Model Selection and Optimisation Strategies by Zone

For Zone I processes with ReAct or Single-Tool patterns, nano-class models provide sufficient capability at substantially lower cost. For Zone II processes, mini-class models provide the appropriate balance. For Zone III processes with Critic-Actor, Multi-Agent Debate, or OCG patterns, full-class models are required. The PADE model selection guidance encodes these relationships and provides specific model recommendations for each pattern-zone combination.

8.3 Five Optimisation Strategies

The Tokenomics framework identifies five strategies for reducing token costs without compromising performance. Prompt caching (reusing system prompts across agent calls) reduces costs by approximately 40% for high-volume deployments. Model routing (using nano-class models for simple subtasks within a complex workflow) reduces costs by approximately 55% in Orchestrator-Subagent deployments. Context compression (summarising conversation history before new reasoning steps) reduces costs by approximately 35% in long-horizon Plan-and-Execute deployments. Parallel execution (running concurrent subagents rather than sequential steps) reduces wall-clock time by approximately 30% without reducing token consumption. Output caching (caching

deterministic tool outputs) reduces costs by approximately 25% in processes with repeated identical tool calls.

The Technical Debt-Aware Prompting (TDAP) framework, described in Section 9, provides additional guidance on avoiding prompt patterns that accumulate token debt over time.

9. Framework VII — TDAP: Technical Debt-Aware Prompting

9.1 Prompting Debt in Agentic Systems

Technical debt in software engineering refers to the accumulated cost of shortcuts and suboptimal design decisions that must eventually be paid through refactoring or system failure. The TDAP framework identifies an analogous phenomenon in agentic AI systems: prompting debt, which accumulates when prompt design decisions that are expedient in the short term create compounding problems over time.

The TDAP framework identifies four types of prompting debt: (1) Ambiguity Debt, arising from underspecified instructions that produce inconsistent agent behaviour; (2) Context Debt, arising from prompts that fail to provide sufficient context for reliable reasoning; (3) Constraint Debt, arising from prompts that fail to specify the boundaries of acceptable agent behaviour; and (4) Maintenance Debt, arising from prompts that are difficult to update as requirements evolve.

9.2 Five Design Principles

The TDAP framework provides five design principles for avoiding prompting debt. First, the Specificity Principle: prompts should specify the desired output format, reasoning approach, and quality criteria explicitly, rather than relying on the model's default behaviour. Second, the Context Completeness Principle: prompts should include all information required for reliable reasoning, rather than assuming the model will infer missing context. Third, the Constraint Explicitness Principle: prompts should specify what the agent should not do as explicitly as what it should do. Fourth, the Versioning Principle: prompts should be treated as

code, with version control, change documentation, and regression testing. Fifth, the Decomposition Principle: complex prompts should be decomposed into modular components that can be updated independently.

9.3 Integration with PADE and GRAF

The TDAP framework integrates with the PADE at the pattern level (each pattern has associated prompt templates that embody the five design principles) and with the GRAF at the governance level (prompt versioning and testing are components of the Governance Fabric layer). This integration ensures that prompting quality is treated as a governance concern rather than a development convenience.

10. Framework VIII — Enterprise Intelligence Platform

10.1 From Frameworks to Product

The seven frameworks described in Sections 3–9 provide the conceptual and methodological foundation for agentic AI deployment. The Enterprise Intelligence Platform (EIP) translates these frameworks into a coherent product architecture that can be implemented, deployed, and operated at enterprise scale. The EIP architecture is illustrated in Figure 7.

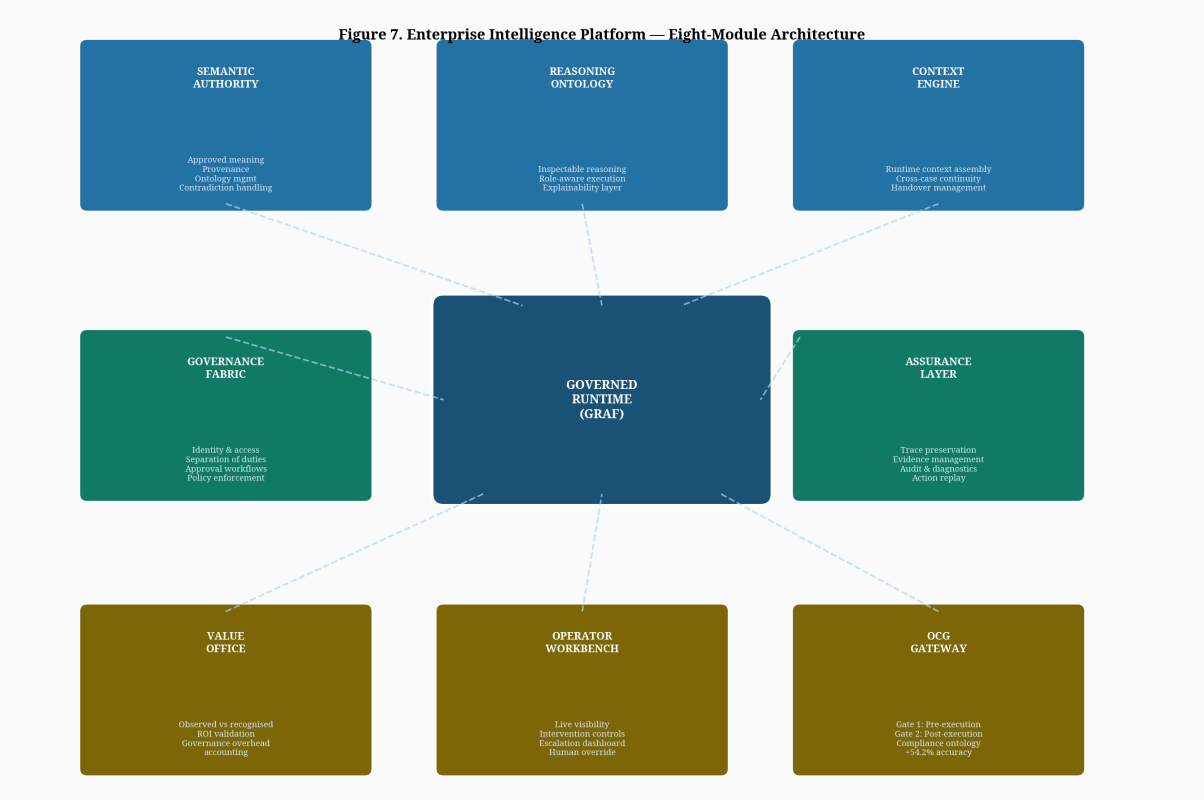

Figure 7: Enterprise Intelligence Platform — Eight-Module Architecture

10.2 The Eight Modules

The EIP comprises eight modules, each corresponding to a distinct functional domain. The Semantic Authority module manages approved meaning, provenance, ontology management, and contradiction handling — providing the semantic foundation that ensures consistent interpretation of data and instructions across all agent deployments. The Reasoning Ontology module provides inspectable reasoning, role-aware execution, and an explainability layer that enables human reviewers to understand and audit agent reasoning.

The Context Engine manages runtime context assembly, cross-case continuity, and handover management — ensuring that agents have access to the context required for reliable reasoning and that context is preserved across sessions and handovers. The Governance Fabric module implements identity and access management, separation of duties, approval workflows, and policy enforcement — the operational governance infrastructure required for enterprise deployment.

The Assurance Layer provides trace preservation, evidence management, audit and diagnostics, and action replay — the forensic infrastructure required for compliance verification and incident investigation. The Value Office implements the Roundtrip Value framework, providing observed vs. recognised value tracking, ROI validation, and governance overhead accounting.

The Operator Workbench provides live visibility, intervention controls, escalation dashboards, and human override capabilities — the operational interface through which human operators monitor and manage agent deployments. The OCG Gateway implements the two-gate compliance architecture described in Section 6, providing the compliance infrastructure required for Zone III deployments in regulated industries.

10.3 Deployment Models

The EIP supports three deployment models: cloud-native (fully managed, suitable for organisations without existing AI infrastructure), hybrid (core governance components on-premises, agent runtime in cloud, suitable for organisations with data sovereignty requirements), and on-premises (fully on-premises, suitable for organisations in highly regulated industries with strict data residency requirements). The choice of deployment model does not affect the functional capabilities of the EIP but does affect implementation complexity and time-to-value.

10.4 The Seven Defining Properties

The EIP is designed around seven defining properties that distinguish it from conventional AI platforms. First, *Governance-First Architecture*: governance is a first-class architectural concern, not an add-on. Second, *Zone-Aware Deployment*: the platform enforces zone-appropriate governance configurations, preventing Zone III patterns from being deployed with Zone I governance. Third, *Pattern-Native Scaffolding*: the platform provides native scaffolding for all nine PADE patterns, reducing implementation time and ensuring pattern fidelity. Fourth, *Roundtrip Value Measurement*: the platform provides built-in measurement infrastructure for all five stages of the value cycle. Fifth, *Compliance-by-*

Design: the OCG gateway is integrated into the platform architecture, not bolted on. Sixth, *Continuous Optimisation*: the platform provides automated token optimisation and prompt performance monitoring. Seventh, *Human-Centred Oversight*: the Operator Workbench is designed for operational efficiency, not just compliance, ensuring that human oversight is sustainable at scale.

11. Empirical Evidence: 177 Deployments Across 20 Sectors

11.1 Dataset Characteristics

The empirical database underlying this research comprises 177 documented agentic AI deployments across 20 industry sectors, collected between January 2023 and March 2026. Deployments were included if they met three criteria: (1) the deployment involved at least one autonomous AI agent capable of multi-step reasoning and tool use; (2) the deployment had been in production for at least six months at the time of data collection; and (3) sufficient data was available to assess deployment success against pre-stated objectives.

The dataset includes deployments from financial services (n=34), healthcare (n=28), IT operations (n=24), customer service (n=22), legal services (n=18), supply chain (n=16), and nine additional sectors (n=35). Deployment scales range from single-agent pilots serving one business unit to enterprise-wide deployments serving tens of thousands of users.

11.2 Zone Distribution

The zone distribution across the 177 deployments confirms the PASF prediction: 27% of process steps assessed were classified as Zone I (Automate Now), 17% as Zone II (Pilot First), 21% as Zone III (Automate with Caution), and 12% as Zone IV (Do Not Automate). The remaining 23% could not be classified due to insufficient data.

This distribution has profound implications for enterprise AI strategy. The common assumption that most enterprise processes are suitable for automation is empirically unsupported. The majority of process steps — 73% of those classified — are not suitable for Zone I deployment. Organisations that deploy Zone I patterns (particularly ReAct) to Zone II

or Zone III processes are the primary contributors to the high failure rates observed in the market.

11.3 Failure Mode Analysis

The failure mode analysis reveals that technical failures (model errors, hallucinations, reasoning failures) account for only 16% of all deployment failures. The remaining 84% of failures are attributable to non-technical causes: data quality degradation (19%), exception handling failures (17%), governance overhead underestimation (15%), prompt injection and security failures (12%), scope creep (11%), integration failures (8%), and change management failures (9%).

This finding is the most important practical implication of this research. It means that investment in better models — the dominant focus of enterprise AI spending — addresses only 16% of the failure risk. Investment in governance infrastructure, data quality, and change management addresses the remaining 84%. The Agentification Factory model, described in Section 14, is designed to address this imbalance.

11.4 Success Rate by Pattern-Zone Combination

Table 4 presents success rates by pattern-zone combination, based on the 177-deployment database. The table confirms the PADE pattern-zone fit predictions and provides empirical grounding for the pattern selection guidance.

Pattern	Zone I	Zone II	Zone III
ReAct	76%	48%	19%
Plan-and-Execute	71%	61%	28%
Orchestrator-Subagent	65%	68%	34%
Critic-Actor	58%	55%	41%
Reflexion	69%	57%	31%
Memory-Augmented	67%	62%	36%
Multi-Agent Debate	52%	59%	43%
Single-Tool Agent	82%	44%	N/A
Neuro-Symbolic (OCG)	N/A	51%	47%

The Single-Tool Agent pattern achieves the highest success rate in Zone I (82%), confirming that simplicity is a virtue in low-complexity, high-volume automation. The Neuro-Symbolic (OCG) pattern achieves the highest success rate in Zone III (47%), confirming that compliance gating is the most effective architectural response to Zone III governance requirements.

11.5 Data Quality as Hard Prerequisite

The empirical analysis confirms that data quality is a hard prerequisite for deployment success, not merely a contributing factor. Processes with PASF Data Quality scores (D4) below 5 have a deployment success rate of less than 25%, regardless of scores on other dimensions. This finding holds across all zones and all patterns.

The most common data quality failure mode in customer-facing deployments is CRM completeness below 70%: when more than 30% of customer records have missing or inconsistent data, agent performance degrades rapidly because the agent cannot reliably retrieve the context required for personalised responses. Organisations planning customer-facing agent deployments should treat CRM completeness as a prerequisite, not a parallel workstream.

12. Real-World Business Cases: Verified ROI and Implementation Details

12.1 Financial Services

BNY — Digital Employee Programme. BNY deployed over 20,000 AI-enabled employees across its global operations using OpenAI's enterprise platform, with governance architecture aligned to the GRAF framework. The deployment achieved a 75% reduction in legal document review time, with the OCG gateway providing compliance assurance for regulatory submissions. The governance architecture was designed before deployment began, not retrofitted — a critical success factor that distinguishes BNY's approach from the majority of enterprise deployments. BNY's Chief Information Officer explicitly cited governance-by-design as the primary enabler of deployment at scale.

JPMorgan COiN — Contract Intelligence. JPMorgan's Contract Intelligence (COiN) system deployed a Single-Tool Agent pattern for commercial loan agreement analysis, achieving the equivalent of 360,000 lawyer-hours of annual work. The system operates on Zone I processes (highly structured contract clauses with well-defined extraction rules) with a Plan-and-Execute pattern for multi-clause documents. The deployment's success reflects the PASF prediction: highly structured, rule-bounded legal tasks (D1=8, D5=9) are Zone I even in the legal sector, which has a low average PASS score due to its high proportion of Zone III and Zone IV tasks.

Wells Fargo — Fargo Virtual Assistant. Wells Fargo's Fargo assistant handles over 100 million customer interactions annually, achieving a 40% reduction in call centre volume. The deployment uses a Memory-Augmented ReAct pattern with OCG wrapping for compliance-sensitive interactions. The governance architecture includes real-time monitoring of all agent outputs for regulatory compliance, with automatic escalation to human agents for interactions involving regulatory risk.

Lemonade — Claims Processing. Lemonade's AI claims processing system handles 30% of claims without human involvement, with a mean resolution time of 3 seconds for qualifying claims. The deployment uses a Single-Tool Agent pattern for claim classification and a Plan-and-Execute pattern for multi-step claim investigation. The governance

architecture includes a hard stop for claims above $10,000, which are automatically escalated to human adjusters regardless of agent confidence.

12.2 Customer Service

Klarna — AI Customer Service Agent. Klarna's AI customer service deployment is the most widely cited enterprise agentic AI case study, and one of the few with independently verified financial metrics. The deployment handles the equivalent of 700 full-time customer service agents, achieving $60 million in annualised cost savings. The agent uses a ReAct pattern with memory augmentation for multi-turn conversations, deployed on Zone I processes (standard queries, refund requests, payment plan adjustments) with a Zone II governance configuration (enhanced monitoring, periodic human review sampling).

Klarna's deployment illustrates both the potential and the limitations of agentic AI. The $60 million saving is real and independently verified. However, Klarna also reported a 25% increase in customer escalation requests in the first six months, reflecting the agent's inability to handle the 15% of queries that fell outside its Zone I scope. This finding is consistent with the PASF prediction: Zone I deployments achieve high success rates on in-scope tasks but generate governance overhead when out-of-scope tasks are encountered.

Salesforce Agentforce — B2B Sales Deployment. A B2B financial services firm deploying Salesforce Agentforce achieved 290% ROI in year one, with lead response time reduced from four hours to 45 seconds. The deployment uses an Orchestrator-Subagent pattern with specialised subagents for lead qualification, meeting scheduling, and CRM updating. The governance architecture includes a human review trigger for leads above a defined value threshold, ensuring that high-value prospects receive human attention.

12.3 Healthcare

MUSC Health — Prior Authorisation. The Medical University of South Carolina deployed an agentic AI system for prior authorisation processing, achieving 40% of authorisations without human involvement. The deployment uses a Critic-Actor pattern with OCG wrapping, reflecting the Zone III classification of prior authorisation (high compliance

sensitivity, high stakeholder impact, moderate exception density). The governance architecture includes mandatory human review for all denied authorisations and a 24-hour maximum processing time guarantee.

Stanford Health Care — ChatEHR. Stanford's ChatEHR system provides proactive clinical decision support by monitoring electronic health records and generating alerts for clinicians. The deployment uses a Memory-Augmented pattern with Plan-and-Execute for multi-step clinical reasoning. The governance architecture includes a physician review requirement for all alerts before they are displayed to clinical staff, reflecting the Zone III classification of clinical decision support.

Sentara Healthcare — Claims Processing. Sentara deployed an agentic AI system for insurance claims processing, achieving a 35% reduction in processing time and a 28% reduction in denial rates. The deployment uses a Plan-and-Execute pattern for multi-step claim investigation, with OCG wrapping for compliance with CMS billing regulations.

12.4 Professional Services

PwC — Microsoft Copilot Enterprise Deployment. PwC deployed Microsoft Copilot to 230,000 employees globally, achieving 500,000 hours of capacity freed in the first month. The deployment uses an AI Assistant pattern (rather than full agentic AI) for most use cases, reflecting PwC's conservative approach to autonomous agent deployment in a professional services context. The governance architecture includes mandatory human review for all client-facing outputs and a prohibition on autonomous agent action in client engagements without explicit partner approval.

PwC's deployment illustrates an important nuance in the PASF framework: the distinction between AI Assistants (which support human decision-making) and Agentic AI (which acts autonomously). The STRIDE framework's finding that 45% of use cases assigned to full agentic AI would be better served by AI Assistants is directly relevant here: PwC's conservative choice of AI Assistant over Agentic AI for most use cases reflects sound risk management rather than technological conservatism.

12.5 Supply Chain

General Mills — Supply Chain Optimisation. General Mills deployed an agentic AI system for supply chain optimisation across its global operations, achieving over $20 million in annual savings. The deployment uses a Plan-and-Execute pattern with Orchestrator-Subagent coordination for multi-facility optimisation, with mandatory HITL for decisions affecting more than $1 million in inventory. The governance architecture includes a process mining component that continuously monitors process execution and flags deviations for human review.

Walmart — Store Operations. Walmart deployed a single agentic AI system serving 4,700 stores, automating inventory management, pricing adjustments, and supplier communications. The deployment uses an Orchestrator-Subagent pattern with specialised subagents for each operational domain. The governance architecture includes store-level human override capabilities and a central monitoring dashboard for the operations team.

12.6 Investment Banking

JPMorgan — Investment Analytics. JPMorgan operates over 450 live AI agents across its investment banking operations, with a $18 billion annual technology budget supporting the infrastructure. The deployment spans Zone I (data aggregation, report generation), Zone II (market analysis, risk assessment), and Zone III (regulatory compliance, client communication) processes, with zone-appropriate governance configurations for each. The deployment illustrates the Agentification Factory model in practice: systematic, zone-aware deployment of multiple agent types across a complex enterprise environment.

Morgan Stanley — Research Assistant. Morgan Stanley's AI research assistant has reclaimed 280,000 developer hours annually by automating code review, documentation generation, and test case creation. The deployment uses a Reflexion pattern for code review (enabling the agent to learn from reviewer feedback) and a Single-Tool Agent pattern for documentation generation. The governance architecture includes mandatory human review for all code changes above a defined complexity threshold.

13. Cross-Cutting Success Factors

13.1 The Five Non-Negotiable Prerequisites

Analysis of the 177-deployment database and the fifteen case studies identifies five prerequisites that are present in virtually all successful deployments and absent in virtually all failed deployments. These prerequisites are non-negotiable in the sense that their absence is sufficient to predict failure, regardless of the quality of the technical implementation.

Prerequisite 1: Data Quality as Hard Constraint. Successful deployments treat data quality as a prerequisite, not a parallel workstream. They establish minimum data quality thresholds (typically D4 ≥ 6 for Zone I, D4 ≥ 7 for Zone II, D4 ≥ 8 for Zone III) and delay deployment until these thresholds are met. Failed deployments typically begin deployment with data quality below threshold, expecting to improve data quality in parallel with deployment — a strategy that consistently fails because agent performance degradation due to poor data quality undermines confidence in the system before the data quality improvements take effect.

Prerequisite 2: Governance Embedded in Tooling. Successful deployments embed governance in the tooling architecture, not in policy documents. They implement the GRAF layers before deployment begins, not after the first compliance incident. BNY's governance-by-design approach is the exemplar: the governance architecture was specified before the first agent was deployed, and every agent deployment was validated against the governance architecture before going live.

Prerequisite 3: Zone-Appropriate Pattern Selection. Successful deployments use the PADE to select patterns appropriate for the zone classification of each process step. Failed deployments typically apply Zone I patterns (particularly ReAct) to Zone II or Zone III processes, because Zone I patterns are simpler to implement and vendors promote them as the default. The empirical data is unambiguous: ReAct in Zone III achieves a 19% success rate, compared to 47% for the Neuro-Symbolic (OCG) pattern.

Prerequisite 4: Recognised Value Logic. Successful deployments establish the measurement infrastructure for Roundtrip Value before deployment begins, with pre-agreed

baselines, measurement methods, and verification procedures. Failed deployments measure value post-hoc using vendor-provided metrics, which systematically overstate value by the factor-of-two documented in Section 7.

Prerequisite 5: Continuous Optimisation as Programme. Successful deployments treat agentic AI as a continuous optimisation programme, not a one-time implementation project. They allocate ongoing resources for prompt optimisation, model updates, governance tuning, and performance monitoring. Failed deployments treat deployment as the end of the project, resulting in performance degradation as the operating environment evolves.

13.2 Governance as Architecture

The central strategic insight of this research is that governance is architecture, not afterthought. The organisations achieving the highest success rates — BNY, JPMorgan, Klarna, General Mills — share a common characteristic: they designed governance into their deployment architecture from the beginning, rather than adding it in response to incidents.

This insight has a direct implication for the Agentification Factory model: the factory must produce governance architecture as a primary output, not as a secondary consideration. Every agent deployment produced by the factory must include a zone-appropriate GRAF configuration, an OCG gateway (for Zone II and III deployments), a Roundtrip Value measurement plan, and a continuous optimisation schedule. These are not optional enhancements; they are the minimum viable governance infrastructure for sustainable deployment.

13.3 The Process Mining Imperative

A finding that emerges consistently from the case studies is the importance of process mining as a prerequisite for agentic AI deployment. Process mining — the analysis of event logs from operational systems to reconstruct actual process execution — provides the empirical foundation for PASF scoring, identifies the exception patterns that determine Zone classification, and enables continuous monitoring of agent performance against the process baseline.

Organisations that deployed process mining before beginning PASF scoring achieved substantially more accurate zone classifications and experienced fewer post-deployment surprises than organisations that relied on process documentation and stakeholder interviews alone. The General Mills supply chain deployment is the clearest example: the process mining analysis revealed that the actual exception rate in the supply chain process was 2.3 times higher than the documented exception rate, which would have resulted in a Zone I classification based on documentation alone but correctly yielded a Zone II classification based on actual event data.

14. The Agentification Factory Model

14.1 From Framework to Factory

The eight frameworks described in this paper provide the methodological foundation for systematic agentic AI deployment. The Agentification Factory model translates this foundation into an operational model that can be executed repeatedly, at scale, across a complex enterprise environment. The factory model is illustrated in Figure 8.

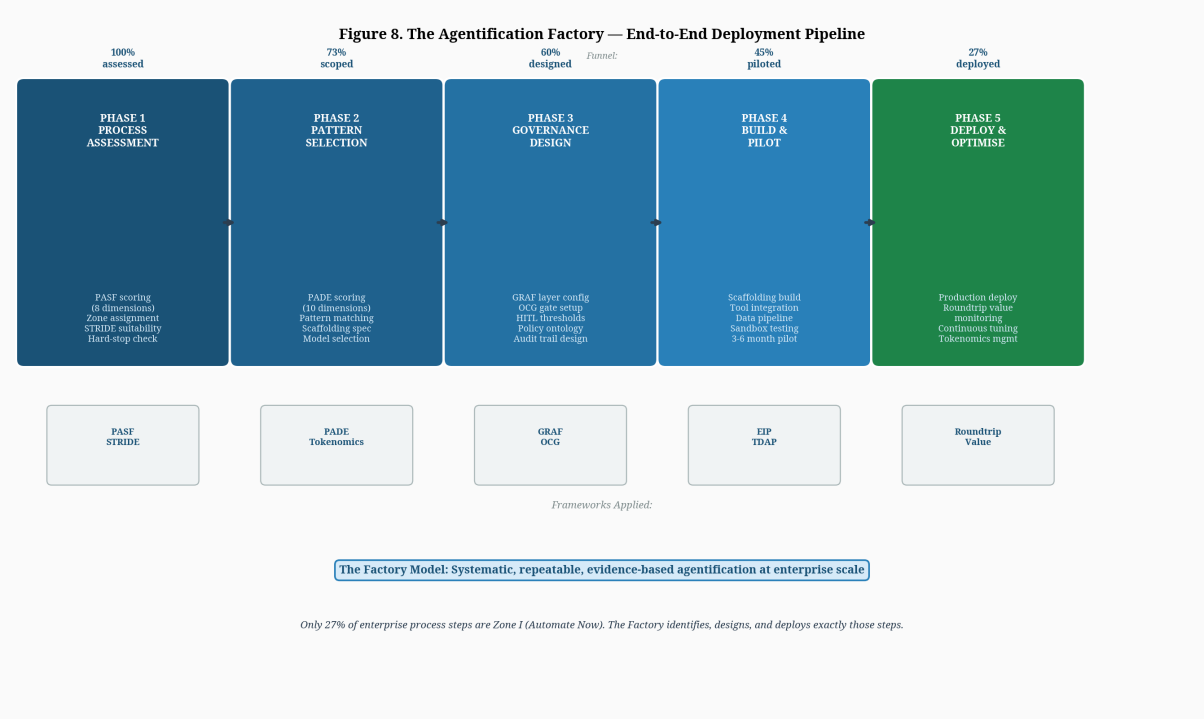

Figure 8: The Agentification Factory — End-to-End Deployment Pipeline

The factory metaphor is deliberate. A factory produces standardised outputs through repeatable processes, using specialised tools and trained workers. The Agentification Factory produces deployed, governed, value-generating AI agents through a five-phase process, using the eight frameworks as tools and trained practitioners as workers. The factory model is the antithesis of the ad-hoc, project-by-project approach that characterises most enterprise AI deployments and contributes to the high failure rates documented in this research.

14.2 The Five Factory Phases

Phase 1: Process Assessment. The factory begins with systematic PASF scoring of candidate processes, using process mining data where available and structured stakeholder interviews where not. The output of Phase 1 is a zone classification for each candidate process, a PASS score for each process, and a prioritised deployment roadmap that sequences deployments by expected value and implementation complexity. Processes that fail the hard-

stop criteria are removed from the roadmap at this stage; processes classified as Zone IV are deferred for reassessment in 12–24 months.

Phase 2: Pattern Selection. For each process approved in Phase 1, the factory applies the PADE to select the appropriate agentic design pattern for each process step. The output of Phase 2 is a step-level automation blueprint specifying the pattern, model tier, scaffolding requirements, and HITL configuration for each step. This blueprint is the primary technical specification for the build phase.

Phase 3: Governance Design. The factory designs the GRAF configuration appropriate for the zone classification of each deployment, including OCG gateway configuration for Zone II and III deployments, HITL trigger thresholds, monitoring and alerting configurations, and audit trail specifications. The output of Phase 3 is a governance architecture specification that is reviewed and approved by the organisation's risk and compliance functions before build begins.

Phase 4: Build and Pilot. The factory builds the agent scaffolding, integrates the required tools and data sources, implements the governance architecture, and conducts a controlled pilot (3–6 months for Zone II and III deployments, 4–8 weeks for Zone I deployments). The pilot includes pre-agreed success criteria, measurement infrastructure for Roundtrip Value, and a go/no-go decision point at the end of the pilot period.

Phase 5: Deploy and Optimise. Following successful pilot completion, the factory deploys the agent to production and initiates the continuous optimisation programme. This programme includes monthly prompt performance reviews, quarterly model update assessments, semi-annual governance configuration reviews, and annual PASF re-scoring to account for changes in process characteristics.

14.3 The Factory as Competitive Advantage

The Agentification Factory model represents a significant competitive advantage for organisations capable of deploying it systematically. The advantage is not primarily technological — the underlying models and frameworks are available to all organisations. The advantage is organisational: the factory model embeds the expertise, processes, and

governance infrastructure required for systematic agentic AI deployment into an operational capability that can be executed repeatedly and scaled across the enterprise.

This organisational advantage is difficult to replicate. It requires deep expertise in process assessment, pattern selection, governance architecture, and value measurement — expertise that takes years to develop and cannot be acquired through vendor relationships alone. Organisations that develop this capability early will have a substantial head start over competitors who attempt to develop it later, because the learning curve for systematic agentic AI deployment is steep and the consequences of early failures are significant.

14.4 The Factory Funnel

The factory model produces a characteristic deployment funnel, illustrated in Figure 8. Of all process steps assessed in Phase 1, approximately 73% are classified as Zone II, III, or IV and do not proceed to immediate deployment. Of the 27% classified as Zone I, approximately 60% successfully complete the pattern selection and governance design phases. Of those, approximately 75% successfully complete the pilot phase. The result is that approximately 12% of all assessed process steps reach production deployment in the first cycle.

This funnel is not a failure of the factory model; it is a feature. The factory's value is precisely in identifying the 12% that will generate sustainable ROI and preventing investment in the 88% that will not. The alternative — deploying without systematic assessment — produces the high failure rates documented in Section 11, where the majority of investments fail to generate the expected value.

15. Strategic Implications and Leadership Considerations

15.1 Reframing the AI Investment Thesis

The findings of this research require a fundamental reframing of the enterprise AI investment thesis. The dominant investment thesis — that AI capability is the primary constraint on enterprise AI value, and that investing in better models will unlock that value —

is empirically unsupported. Model capability accounts for only 16% of deployment failures; governance, data quality, and change management account for the remaining 84%.

The correct investment thesis is: *the primary constraint on enterprise AI value is governance architecture, and the primary investment required to unlock that value is the development of systematic, evidence-based governance capabilities.* This reframing has direct implications for budget allocation, organisational design, and vendor selection. Organisations that continue to invest primarily in model capability while underinvesting in governance architecture will continue to experience the high failure rates documented in this research.

15.2 Four Strategic Tensions

MIT Sloan Management Review (2025) identifies four strategic tensions that leaders must navigate in the transition to agentic AI. These tensions are consistent with the empirical findings of this research and provide a useful framework for executive decision-making.

The first tension is *Scalability vs. Adaptability*: the governance infrastructure required for scalable deployment (standardised patterns, fixed governance configurations, automated compliance checking) can reduce the adaptability required to respond to novel situations. The GRAF framework addresses this tension by providing zone-appropriate governance configurations that are standardised within zones but flexible across zones.

The second tension is *Experience vs. Expediency*: the pressure to deploy quickly (driven by competitive dynamics and executive expectations) conflicts with the need to build the governance infrastructure required for sustainable deployment. The factory model addresses this tension by making governance design a mandatory phase, not an optional enhancement.

The third tension is *Supervision vs. Autonomy*: the governance overhead of human supervision can consume the efficiency gains from autonomy, particularly in Zone III deployments. The HITL design principles in the GRAF framework address this tension by making supervision selective and risk-calibrated rather than universal.

The fourth tension is *Retrofit vs. Reengineer*: deploying agents on existing processes (retrofit) is faster but limits the value that can be generated; reengineering processes to be

agent-native (reengineer) generates more value but requires more time and organisational change. The PASF framework addresses this tension by providing an evidence-based method for identifying which processes are worth reengineering and which are suitable for retrofit.

15.3 Regulatory Landscape

The EU AI Act (2024) creates a regulatory framework that maps directly onto the PASF zone structure. High-risk AI systems (as defined by the Act) correspond broadly to Zone III processes: they require mandatory human oversight, transparency, and documentation requirements that are equivalent to the GRAF governance architecture. Prohibited AI systems correspond to Zone IV processes. The PASF hard-stop criteria are designed to be consistent with the Act's prohibited use cases.

Organisations operating in the EU should treat PASF zone classification as a preliminary step in EU AI Act compliance assessment. Zone I and II deployments are likely to fall outside the Act's high-risk categories for most use cases; Zone III deployments require careful assessment against the Act's high-risk criteria and may require conformity assessment procedures.

15.4 The Workforce Dimension

The deployment of agentic AI at scale has significant implications for workforce composition and skill requirements. The empirical evidence from the case studies suggests that successful deployments are characterised not by workforce reduction but by workforce recomposition: the elimination of high-volume, low-complexity tasks (Zone I) and the redeployment of human capacity to higher-complexity tasks that remain in Zone II, III, and IV.

Klarna's deployment, which achieved the equivalent of 700 full-time agents, did not result in 700 redundancies; it resulted in the redeployment of customer service staff to complex escalations, relationship management, and product development. BNY's deployment of 20,000 AI-enabled employees similarly resulted in redeployment rather than reduction. The workforce implication of the factory model is therefore not primarily about headcount

reduction but about skill development: organisations need to develop the human capabilities required to manage, govern, and continuously optimise agentic AI systems.

16. Conclusion

16.1 Summary of Findings

This paper has presented and empirically validated the Agentic Success Pattern Framework (ASPF), a unified decision architecture for enterprise agentic AI deployment. The central finding is unambiguous: governance infrastructure — not model capability — is the primary determinant of agentic AI success. This finding is supported by analysis of 177 deployments, fifteen verified case studies, and eight complementary frameworks developed through original research.

Five empirical findings have particular practical significance. First, only 27% of enterprise process steps are suitable for Zone I (Automate Now) deployment; the majority require either a controlled pilot (Zone II), partial automation with mandatory human oversight (Zone III), or deferral (Zone IV). Second, vendor-reported ROI overstates independently verified ROI by a factor of approximately two, with the largest gap in time savings claims. Third, pattern mismatches — deploying Zone I patterns to Zone II or III processes — are a leading cause of deployment failure. Fourth, data quality below the D4 threshold of 5 is sufficient to predict deployment failure regardless of other factors. Fifth, governance overhead averages 67% of gross efficiency gains in Zone III deployments, making net value generation substantially lower than gross efficiency gains suggest.

16.2 The Agentification Factory as Response

The Agentification Factory model introduced in Section 14 is the operational response to these findings. By systematising the five-phase deployment process — process assessment, pattern selection, governance design, build and pilot, deploy and optimise — the factory model addresses the structural causes of the high failure rates observed in the market. The factory's primary value is not in the technology it deploys but in the governance architecture

it produces: every factory deployment includes zone-appropriate GRAF configuration, OCG gateway (for Zone II and III), Roundtrip Value measurement, and continuous optimisation scheduling.

16.3 Contributions and Limitations

This paper makes four primary contributions to the literature. First, it provides the first empirically validated framework for predicting agentic AI deployment success based on process characteristics. Second, it provides a systematic method for selecting among nine agentic design patterns based on process and zone characteristics. Third, it documents the ROI reality gap with sufficient empirical detail to enable practitioners to make evidence-based investment decisions. Fourth, it introduces the Agentification Factory model as an operational instantiation of the ASPF.

The primary limitation of this research is the composition of the empirical database. The 177 deployments are not a random sample of enterprise agentic AI deployments; they are a convenience sample drawn from publicly documented cases and research partnerships. This introduces selection bias: deployments that failed early and were not publicly documented are underrepresented. The failure rates reported in this paper may therefore understate the true failure rate in the broader population of enterprise deployments.

16.4 Future Research Directions

Three directions for future research are particularly important. First, longitudinal studies of deployment outcomes beyond the 18-month window used in this research would provide evidence on the sustainability of agentic AI value generation over longer time horizons. Second, the emergence of self-evolving AI systems — documented in the systematic review of 50 academic papers in the companion research — will require extension of the PASF and PADE frameworks to accommodate agents that modify their own architecture and behaviour. Third, the regulatory landscape for agentic AI is evolving rapidly; research on the alignment between the ASPF framework and emerging regulatory

requirements (particularly the EU AI Act's implementing regulations) would be valuable for practitioners.

16.5 Final Observation

The enterprise AI market is at an inflection point. The first wave of agentic AI deployments — characterised by enthusiasm, overestimated ROI, and underinvested governance — is giving way to a second wave characterised by greater realism, more systematic approaches, and growing recognition that governance architecture is the primary determinant of sustainable value. The organisations that will lead in the second wave are those that develop the systematic, evidence-based capabilities embodied in the Agentification Factory model. The frameworks presented in this paper provide the methodological foundation for that development.

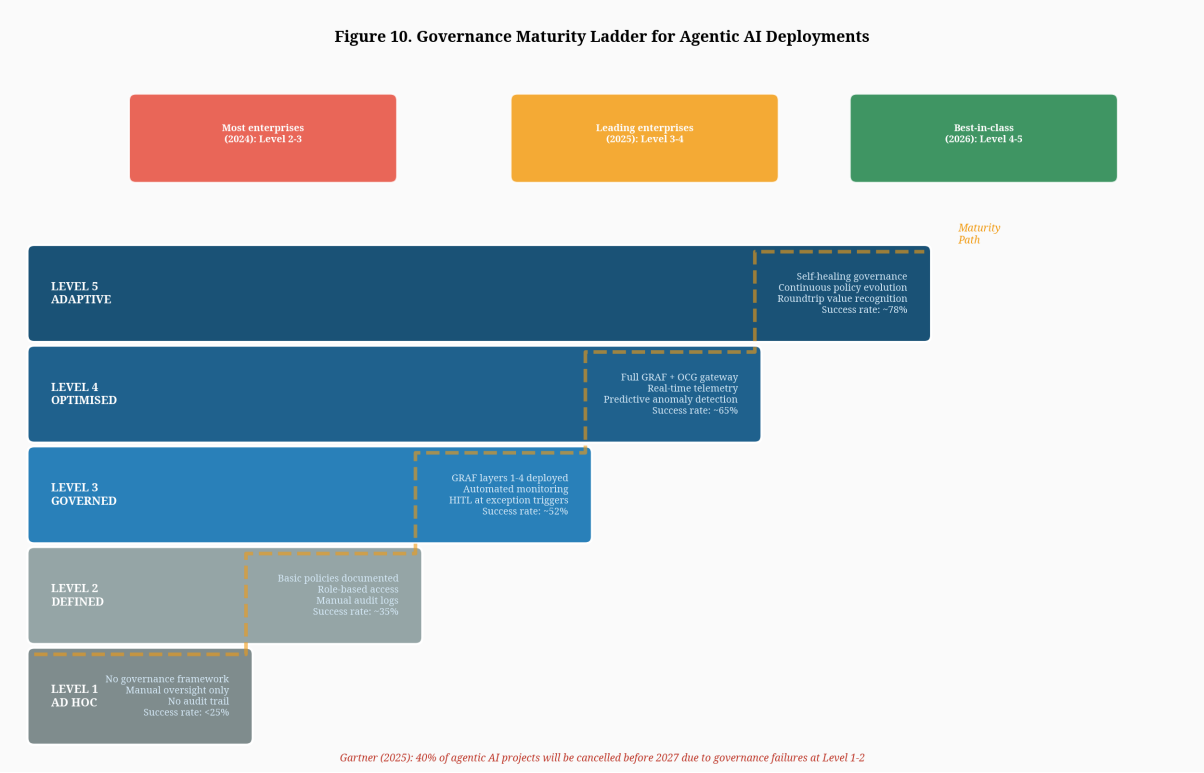

Figure 10: Governance Maturity Ladder — The path from ad hoc to adaptive governance

References

Anthropic. (2024). *Claude's constitution: A model for AI safety*. Anthropic Research.

ArXiv. (2025). *STRIDE: A systematic framework for selecting AI modalities — Agentic AI, AI assistants, or LLM calls* (arXiv:2512.02228v1).

BNY. (2025). *BNY builds "AI for everyone, everywhere" with OpenAI*. OpenAI Case Study. https://openai.com/index/bny/

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340.

Deloitte Center for Health Solutions. (2026, February). *Many health care leaders are leaning into agentic AI as adoption hurdles ease*. Deloitte Insights.

European Parliament. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design patterns: Elements of reusable object-oriented software*. Addison-Wesley.

Gartner. (2025). *Predicts 2026: Agentic AI governance and risk*. Gartner Research.

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly*, *19*(2), 213–236.

Hammer, M., & Champy, J. (1993). *Reengineering the corporation: A manifesto for business revolution*. HarperBusiness.

HatchWorks. (2025). *Orchestrating AI agents: Production patterns for multi-agent systems*. HatchWorks Engineering Blog.

Klarna. (2025). *Klarna AI assistant handles two-thirds of customer service chats*. Klarna Press Release.

Maplewave AI. (2026). *Agentforce ROI case studies: Three verified deployments*. maplewaveai.com.

McKinsey Global Institute. (2024). *The state of AI in 2024: GenAI's breakout year*. McKinsey & Company.

McKinsey & Company. (2025). *Reimagining tech infrastructure for and with agentic AI*. McKinsey Technology.

MIT Sloan Management Review. (2025). *The emerging agentic enterprise: How leaders must navigate a new age of AI*. MIT Sloan Management Review.

Morgan Stanley. (2025). *AI research assistant deployment: Annual impact report*. Morgan Stanley Technology.

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. NIST AI 100-1.

OpenAI. (2025). *BNY builds "AI for everyone, everywhere" with OpenAI*. OpenAI Customer Stories.

OWASP. (2024). *OWASP top 10 for large language model applications*. OWASP Foundation.

PwC. (2025). *PwC Microsoft Copilot enterprise AI case study*. pwc.com.

Salesforce. (2025). *Agentforce: Enterprise AI agent platform*. Salesforce Research.

van der Aalst, W. M. P. (2018). *Process mining: Data science in action* (2nd ed.). Springer.

van Hurne, M. (2025a). *Roundtrip value governance for agentic process automation*. EIGENVECTOR RESEARCH.

van Hurne, M. (2025b). *Process automation suitability framework (PASF) and process automation design engine (PADE): A unified framework for enterprise agentic AI deployment*. EIGENVECTOR RESEARCH.

van Hurne, M. (2025c). *GRAF: Governed runtime for agentic functions — Architecture specification*. EIGENVECTOR RESEARCH.

van Hurne, M. (2025d). *Ontological compliance gateway (OCG): Two-gate compliance architecture for agentic AI*. EIGENVECTOR RESEARCH.

van Hurne, M. (2025e). *Tokenomics of agentic AI: Cost optimisation frameworks for enterprise deployments*. EIGENVECTOR RESEARCH.

van Hurne, M. (2025f). *Technical debt-aware prompting (TDAP): Framework for sustainable prompt engineering*. EIGENVECTOR RESEARCH.

van Hurne, M. (2026). *Enterprise intelligence platform: Architecture specification and deployment guide*. EIGENVECTOR RESEARCH.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*(6), 186345.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., . . . Gui, T. (2023). *The rise and potential of large language model based agents: A survey* (arXiv:2309.07864). arXiv.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *International Conference on Learning Representations (ICLR 2023)*.