Scientific Research Publishing

Journal of Software Engineering and Applications, 2026, 19(6), 205-240
https://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

AI Selection in Organizations: A Decision Framework for Large Language Model and Generative AI Deployments

Andrew Ganje

Independent Researcher, San Diego, USA
Email: Andrew.ganje@gmail.com

How to cite this paper: Ganje, A. (2026) AI Selection in Organizations: A Decision Framework for Large Language Model and Generative AI Deployments. *Journal of Software Engineering and Applications*, **19**, 205-240.
https://doi.org/10.4236/jsea.2026.196010

Received: March 30, 2026
Accepted: June 27, 2026
Published: June 30, 2026

Abstract

Artificial intelligence selection has emerged as one of the most impactful decision areas facing modern organizations. As AI adoption has accelerated from experimental use to broad enterprise deployment, the central challenge is no longer whether organizations should adopt AI, but how leaders and technology professionals can make informed, defensible decisions about which AI systems to select, how to govern them, and where they can create sustainable organizational value. The urgency of this challenge is heightened by rapid vendor proliferation, uneven executive technical literacy, growing regulatory expectations, and the increasing operational risks associated with poorly matched AI solutions, including wasted investment, vendor lock-in, unreliable outputs, and uncontrolled shadow AI adoption. This paper addresses that problem by providing a practical and conceptually grounded framework for AI selection in organizations. It first explains how current large language models (LLMs) work, including transformer architecture, training, alignment, fine-tuning, and key limitations, while translating these concepts into plain language for executive audiences. It then examines why technical literacy has become essential for organizational decision-makers responsible for AI strategy, procurement, and governance. Building on that foundation, the paper presents a structured selection framework that evaluates AI systems across eight dimensions: capability assessment, alignment and safety, integration fit, data governance, total cost of ownership, vendor risk and lock-in, governance and compliance, and lifecycle management, aligned with the NIST AI Risk Management Framework (AI RMF 1.0) and ISO/IEC 42001:2023. The paper further introduces an executive decision model for the build, configure, or buy choice, and concludes with strategies for preventing extreme AI proliferation through governance controls, technical safeguards, and workforce education. The contribution of this paper is both strategic and practical. For organizational leaders, it offers a clearer basis for aligning AI decisions with business goals, risk

tolerance, compliance obligations, and long-term operating models. For technology professionals, it provides a structured approach to evaluating AI capabilities, deployment options, and lifecycle implications in a way that supports sound architecture and governance decisions. By integrating academic literature with current industry evidence, this paper equips decision-makers with a more informed foundation for selecting AI systems that are not only innovative, but operationally viable, governable, and strategically appropriate. **Objective:** This paper develops and demonstrates a structured, repeatable framework that enables the senior technology executive accountable for AI strategy (principally the CIO or CTO, supported by enterprise architects and a cross-functional selection team) to make defensible decisions about which AI systems to select and whether to build, configure, or buy them. **Method:** The study follows a design-science approach. The framework is synthesized from the NIST AI Risk Management Framework (AI RMF 1.0), ISO/IEC 42001:2023, and the peer-reviewed and industry literature, and is then validated on two complementary, panel-free legs: a dimension-to-standard mapping that establishes coverage of the governing standards, and a retrospective comparative analysis of six documented enterprise AI deployments. **Results:** Across six documented enterprise deployments the framework gates align with the recorded outcome: each failure violated a gate the framework evaluates early (accountability and verification in two public-facing chatbots, capability on real operating data in an automated drive-thru, and capability, integration, and data readiness in a $62-million healthcare build), while a disciplined success cleared the gates in sequence and a partial reversal corrected a lifecycle gate it had under-weighted. Measurable success criteria for the framework (decision auditability, avoidance of post-proof-of-concept abandonment, total-cost-of-ownership forecast accuracy, and reduction of vendor lock-in exposure) are defined. **Conclusion:** For organizational leaders, the framework offers a defensible basis for aligning AI decisions with business goals, risk tolerance, and compliance obligations; for technology professionals, it provides a sequenced method for evaluating capability, governance, sourcing, and lifecycle fit. Expanding the case set and prospective field application are identified as the primary directions for future work.

Keywords

Artificial Intelligence Selection, Large Language Models, Enterprise AI Governance, Responsible AI, AI Proliferation, Transformer Architecture, NIST AI RMF, ISO 42001, Build, Configure, or Buy, Retrieval-Augmented Generation, MLOps, AI Model Lifecycle, Agentic AI, Shadow AI

1. Introduction

Artificial intelligence has become one of the most consequential forces shaping organizational strategy, operations, and competitive advantage. In only a few years, AI has moved from experimentation at the margins of the enterprise to active use across business functions such as marketing, customer service, supply chain, finance, human resources, and product innovation. This acceleration has

created a new organizational challenge. The central question is no longer whether AI should be adopted, but how leaders and technology professionals can make informed decisions about which AI systems to select, how those systems should be governed, and where they can create lasting value.

The scale of this shift is substantial. The McKinsey Global Survey on AI provides one of the clearest longitudinal views of enterprise adoption [1]. In 2017, only 20 percent of organizations reported adopting AI in at least one business function. By 2025, that figure had risen to 88 percent, a figure from McKinsey's self-reported global executive survey rather than a randomized census, with generative AI and agentic systems accelerating adoption across industries and use cases; a trend that continued into 2026 [2]. This trend signals more than the growth of a new technology category. It reflects a broader transformation in how organizations make decisions, automate work, and structure digital capability [3].

Yet rapid adoption has not eliminated uncertainty. As AI tools proliferate across vendors, platforms, and open model ecosystems, many organizations face mounting pressure to make selection decisions without a sufficient understanding of model capabilities, limitations, governance requirements, or long-term operational implications. As a result, AI selection is often approached through vendor narratives, surface level feature comparisons, or short-term experimentation rather than disciplined evaluation. This creates substantial risk, including wasted investment, poor fit with business needs, governance gaps, strategic dependency, and fragmented adoption patterns that are difficult to control at scale. The problem is not simply technical. It is organizational, managerial, and strategic.

The importance of this issue is amplified by the fact that AI adoption has direct consequences for people and institutions, not only for systems and workflows. Research shows that AI implementation can influence productivity, innovation, autonomy, job design, employee stress, and perceptions of fairness [4] [5]. This means that AI selection should not be understood as a narrow procurement exercise. It is a sociotechnical decision that shapes how work is performed, how risk is governed, and how organizational value is created and sustained over time.

This paper addresses that problem by providing a structured and practical foundation for AI selection in organizations. It is written to support both leaders and technology professionals who must evaluate AI options under conditions of rapid market change, incomplete technical understanding, and growing governance pressure. The paper first explains how current large language models function so that readers can understand the architectural and operational characteristics that influence enterprise suitability. It then examines why technical literacy matters in executive decision making, presents a capability driven framework for AI selection, introduces a decision model for the build, configure, or buy choice, and concludes with guidance for preventing uncontrolled AI proliferation through governance, technical controls, and workforce education. Together, these sections provide a decision-oriented framework for selecting AI systems that are strategically appropriate, operationally viable, and governable over time.

The general objective of this paper is to develop and demonstrate a structured

framework that enables senior technology executives to make defensible, repeatable AI selection decisions under conditions of rapid market change, incomplete technical information, and growing governance pressure. The framework is written primarily for the executive accountable for AI strategy and procurement, typically the CIO or CTO, rather than for any leader in general; the chief marketing, operations, or financial officer is addressed only insofar as they participate in the cross-functional selection team. This focus is deliberate: governance accountability and architectural decision rights for enterprise AI most commonly sit with the technology executive, and the criteria, gates, and trade-offs developed here are calibrated to that role and its supporting enterprise architects.

To meet the general objective, the paper pursues four specific objectives: (1) to translate the architectural and operational characteristics of large language models into terms that support executive evaluation rather than engineering practice; (2) to define and justify a set of eight evaluation dimensions spanning capability, governance, cost, and lifecycle; (3) to operationalize the build, configure, or buy decision as a sequenced, gate-based process; and (4) to demonstrate the framework through an illustrative application and to specify how its effectiveness can be measured and subsequently validated. The contribution is therefore both an integrative artifact (a single decision process that unifies capability assessment, governance, sourcing, and lifecycle management for AI selection) and a method for evaluating that artifact. The remainder of the paper is organized accordingly: Section 1.1 sets out the research approach and scope, Sections 2 and 3 establish the conceptual foundation, Section 4 presents the framework, Sections 6 and 7 describe the validation design and an illustrative application, and Section 8 discusses the findings, limitations, and directions for further validation.

1.1. Research Approach and Scope

This paper is a conceptual contribution developed through structured synthesis rather than primary data collection, and this subsection makes that approach explicit so the contribution can be evaluated on its own terms. Sources were selected in three tiers and used for distinct purposes. Peer-reviewed literature was used for technical claims about model architecture, alignment, and limitations; recognized standards, principally the NIST AI Risk Management Framework and ISO/IEC 42001:2023, were used to establish governance requirements; and industry and market reports were used only to characterize the practical conditions under which selection decisions are made, with their source population and methodological limits noted at each point of use. The eight evaluation dimensions were derived by mapping the requirements expressed in those standards onto the sequence of decisions a technology executive must resolve, then consolidating overlapping criteria until each dimension was distinct and non-redundant. Section 6.1 documents this derivation and its relationship to the standards.

Scope: The framework is calibrated to the selection of large language model, generative, and agentic AI systems, where probabilistic behavior, alignment qual-

ity, and autonomy controls make selection most error-prone. Several dimensions transfer directly to predictive, computer-vision, and optimization systems: total cost of ownership, vendor risk and lock-in, data governance, governance and compliance, and lifecycle management apply to any procured model. The alignment-and-safety and capability dimensions, however, are specified in generative terms (hallucination, instruction-following, context windows, and agentic tool use) and do not map cleanly onto those other system classes, which fail in different ways, such as distribution shift in vision models or constraint formulation in optimization. Readers applying the framework outside the generative and agentic context should treat those two dimensions as requiring re-specification, while the remaining six transfer with little change.

The paper uses a consistent set of key terms, which are defined together in Section 1.2. Two of these terms organize the later structure: AI selection, the subject of Section 4, concerns the disciplined choice of sanctioned systems, whereas AI proliferation and its extreme form, the subject of Section 5, concern the governance of unsanctioned and uncontrolled expansion. Distinguishing the two reduces the overlap between those sections.

1.2. Key Terms

Term	Definition
AI selection	The structured decision of which AI system to adopt and how to source it (build, configure, or buy).
Generative AI	Systems, large language models prominent among them, that produce new content such as text, code, or images rather than only classifying or predicting existing data.
Large language model (LLM)	A Transformer-based neural network trained on large text corpora to generate and reason over natural language.
Transformer	The neural architecture, based on self-attention, that underlies modern LLMs.
Fine-tuning	Updating a model weights on curated, task-specific data to specialize it.
Retrieval-augmented generation (RAG)	Supplying organizational knowledge to a model at inference time without changing its weights.
Agentic AI	A system that plans and executes multi-step actions through tool use with limited human intervention, in contrast to a static model that only returns text.
MLOps (machine learning operations)	The practices and infrastructure for deploying, monitoring, evaluating, and retraining models in production.
Build, configure, or buy	The sourcing decision among custom development, configuration of a commercial platform, and turnkey purchase.
AI proliferation/extreme AI proliferation	The spread of AI tools across an organization; extreme when that spread is rapid, uncoordinated, and outpaces governance.
Shadow AI	The subset of proliferation in which employees use AI tools outside sanctioned channels.
Total cost of ownership (TCO)	The full multi-year cost of an AI system, modeled here on a three-year base case (Section 4.5).
NIST AI RMF; ISO/IEC 42001	The U.S. NIST AI Risk Management Framework and the international AI management-system standard used as governance anchors.

1.3. Research Questions and Propositions

The framework is organized around two research questions and two propositions, which connect the artifact to the evidence presented later in the paper.

Item	Statement
RQ1	What dimensions are necessary for defensible enterprise AI selection?
RQ2	Can a sequenced gate model explain documented enterprise AI deployment success and failure?
P1	AI deployments that fail early governance, data, or capability gates are more likely to be abandoned or reversed.
P2	AI selection decisions that incorporate lifecycle total cost of ownership and vendor-risk evaluation are more durable than decisions based primarily on capability demonstrations.

RQ1 is addressed by the derivation of the eight dimensions in Sections 4 and 6; RQ2 and the two propositions are examined retrospectively against the documented cases in Section 7. Throughout, the unifying claim is that enterprise AI selection must be governed as a sequenced decision process rather than a single capability comparison.

2. How Current Large Language Models (LLM) Work

2.1. Foundational Architecture: The Transformer Model

Large language models (LLMs) are a class of neural network architectures trained on vast corpora of text data, enabling them to generate, summarize, classify, translate, and reason about natural language at scale. The dominant architectural paradigm underlying today's most capable LLMs is the Transformer, which processes all tokens in an input sequence simultaneously through a mechanism known as self-attention [6]. Unlike prior sequence models such as recurrent neural networks (RNNs), the self-attention mechanism enables the model to capture long-range dependencies and contextual relationships across an entire passage, regardless of the distance between relevant tokens. This architectural innovation is what enables LLMs to maintain coherence across thousands of words of context.

At its core, a Transformer consists of three primary components: (1) an embedding layer that converts input tokens (subword units of text) into dense numerical vectors encoding semantic meaning, (2) a stack of Transformer blocks, each containing multi-head self-attention layers and feed-forward networks that progressively refine contextual representations; and (3) an output layer that generates probability distributions over the vocabulary to predict the next token [7]. Modern enterprise LLMs such as OpenAI's GPT-4o, Anthropic's Claude, Google's Gemini, and Meta's LLaMA 3 are all built on this Transformer foundation but differ significantly in their training corpora, parameter counts, instruction-tuning approaches, and safety mechanisms [8]. These distinctions are consequential for

organizational selection decisions. Figure 1 traces this processing pipeline from tokenized input through to generated output and maps each stage to its enterprise implications.

2.2. Training, Alignment, and Fine-Tuning

LLMs are produced in stages, and the distinctions matter for selection. Pretraining exposes the model to vast text corpora and yields broad language ability but no alignment with human preferences or organizational norms, at high computational cost [6].

Alignment then adapts the base model to follow instructions and produce helpful, safe outputs, using reinforcement learning from human feedback and newer variants such as DPO and RLAIF [9]. The practical point for buyers is constant across methods: alignment quality and the rigor of accompanying safety testing and red-teaming are a primary differentiator between vendor models and must be evaluated directly, not inferred from brand.

Finally, domain-specific fine-tuning updates model weights on curated data, whereas retrieval-augmented generation (RAG) instead supplies organizational knowledge at inference time without changing weights [10]. The two differ in cost, control, latency, and data security, trade-offs examined in the decision framework in Section 4.

2.3. Model Limitations: Hallucination, Bias, and Context Constraints

Despite their impressive capabilities, LLMs exhibit well-documented limitations that carry direct governance implications for enterprise deployments. Hallucination, the tendency of LLMs to generate plausible-sounding but factually incorrect content, remains a persistent challenge even in frontier models [8]. In high-stakes enterprise domains such as legal, financial, or medical applications, hallucinations pose material liability risks. Organizations must therefore implement verification layers, human-in-the-loop review processes, and confidence scoring mechanisms as standard elements of any production AI deployment.

LLMs also encode and amplify biases present in their training data. The National Institute of Standards and Technology (NIST) AI Risk Management Framework identifies bias, fairness, and transparency as core risk dimensions that organizations must systematically evaluate when selecting and deploying AI systems [11]. Additionally, LLMs operate within context window constraints, the maximum volume of text they can process in a single interaction. While frontier models in 2025 have substantially expanded these limits, with some supporting context windows exceeding 100,000 tokens [6], smaller deployed models, fine-tuned domain variants, and on-premises open-source deployments often carry significantly tighter constraints. Organizations must evaluate context window size against their specific operational requirements rather than assuming frontier capabilities apply to the model variant under procurement.

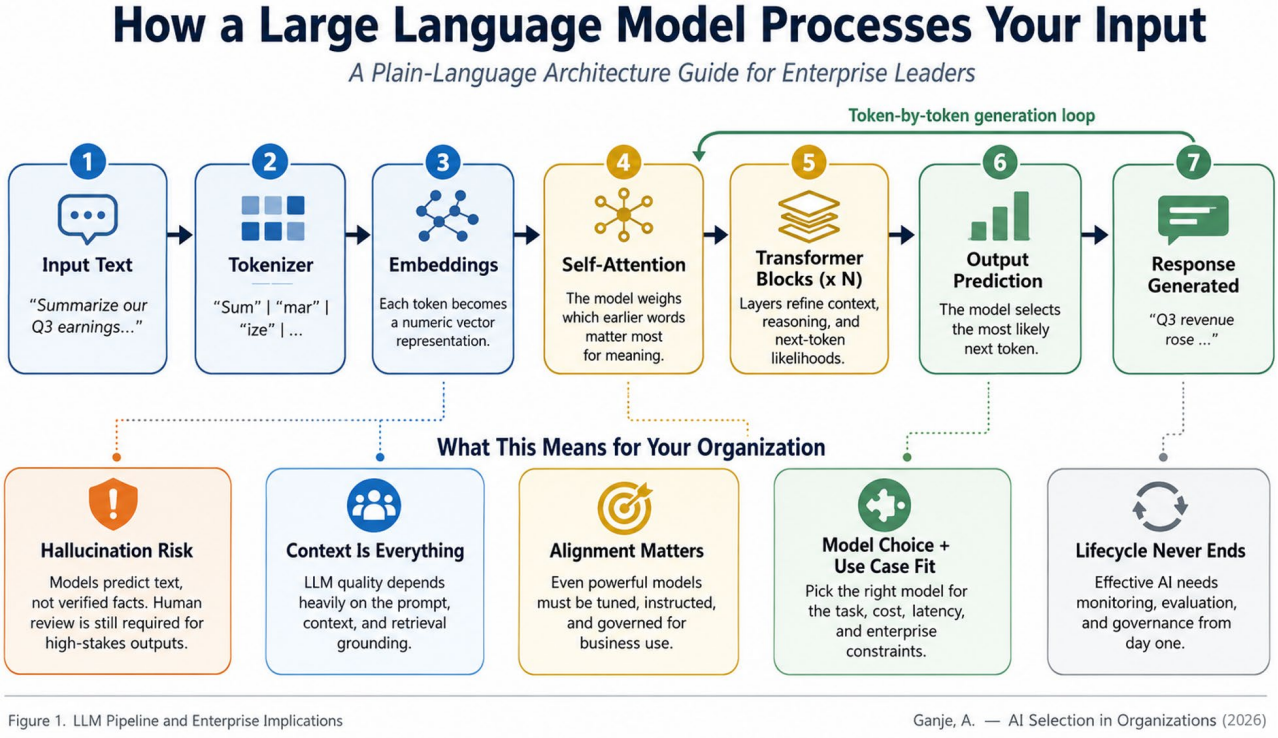

Figure 1. How a large language model processes input: architecture pipeline and enterprise implications. Sources: [6]-[8] [10].

2.4. Understanding LLMs in Plain Language: A Summary for Executive Leaders

The technical detail in sections 2.1 - 2.3 is important for governance and vendor evaluation, but its practical implications are equally important to communicate in terms that do not require an engineering background. The following plain-language translation is designed for executive and board-level audiences who must make consequential AI decisions without deep technical training.

Think of an LLM as an extremely well-read colleague, not a search engine

- An LLM has absorbed patterns from hundreds of billions of words of text during training; books, articles, code, and web content. When you ask it a question, it does not search the internet or look anything up. It draws entirely on what it learned during training, which has a cutoff date.
- This is a critical distinction: an LLM generates statistically likely responses, not verified facts. It sounds authoritative because it has read so much, but it can still be wrong, especially on recent events, proprietary data, or highly specific technical details.
- Practical implication: Always treat LLM outputs in high-stakes contexts (legal, financial, clinical) as a first draft requiring human expert review, not a final answer.

Table 1 maps each technical stage to a familiar organizational concept. It is presented as a communication aid that operationalizes, for non-technical executives, the technical-literacy argument developed in Section 3:

Table 1. Plain-language analogies mapping each technical stage of a large language model to a familiar organizational concept.

Technical Stage	What It Actually Does	Plain-Language Analogy
Tokenization	Breaks your input into small word-chunks	Like a musician reading sheet music note by note before playing a phrase, granular before comprehensive
Embeddings	Assigns each chunk a set of numbers encoding its meaning and relationships	Like a position on a map: similar concepts are close together; opposites are far apart
Self-Attention	Identifies how every word in your input relates to every other word, simultaneously	Like a skilled editor who reads the whole memo before understanding what the first sentence really means
Transformer Blocks	Applies 12-96+ layers of refinement to build a deep contextual understanding	Like passing a brief through a chain of specialists, each one adds nuance before the final response is drafted
Output Prediction	Selects the most probable next token, then repeats until the response is complete	Like a skilled ghostwriter drafting the next sentence based on everything that came before one word at a time
RLHF Alignment	Trains the model to prefer helpful, safe responses using human evaluator feedback	Like a new employee learning company norms and communication standards through performance reviews

Table 2 distills the preceding architecture discussion into the key executive takeaways for non-technical leaders.

Table 2. Key executive takeaways from technical architecture.

- An LLM does not "know" your business. It must be given your context through prompting, fine-tuning, or retrieval-augmented generation (RAG). Without this, responses will be generic and may be wrong for your specific domain.
- Confident language does not mean accurate output. LLMs produce fluent, authoritative-sounding text regardless of factual correctness. This is called hallucination, and it is a design characteristic; not a bug that will be fully eliminated.
- The quality of the alignment process (RLHF) is a primary differentiator between vendor models. Always ask vendors what safety testing and red-teaming was performed, and what the model's documented failure modes are for your target use case.
- Model selection is not a one-time decision. Models degrade as organizational data evolves (model drift), vendor versions change, and new capabilities emerge. Plan for governance, monitoring, and refresh cycles from day one.

3. The Imperative for Technical Literacy in AI Selection

3.1. Bridging the Knowledge Gap in Executive Leadership

One of the most consequential gaps in enterprise AI adoption is the disconnect between organizational decision-makers and the technical realities of the systems they are procuring. McKinsey's State of AI 2025 report found that while 88% of organizations now use AI in at least one business function, fewer than a third report that their senior leadership has sufficient technical understanding to make informed AI investment decisions [2]; as with the adoption figure, this is self-reported by sur-

veyed executives and reflects perception rather than audited capability. This gap creates conditions for misaligned expectations, vendor dependency, and strategic missteps that are difficult and expensive to reverse. The Stanford HAI AI Index 2025 corroborates this finding, reporting that among companies a gap persists between recognizing responsible AI risks and taking meaningful governance action [12].

Technical literacy at the executive level does not require the ability to train neural networks. Rather, it demands a functional understanding of how LLMs learn, what they can and cannot reliably do, and how deployment environment choices affect system behavior and risk. The Wharton School's AI Adoption Report (2025) found that executive leadership involvement in AI adoption has a direct correlation with business value: organizations where senior leaders actively champion and role-model AI use are three times more likely to achieve significant returns from their AI investments [13]; this is an association reported in Wharton's self-selected adoption survey and does not by itself establish causation.

3.2. Understanding Model Limitations to Manage Organizational Risk

When leaders understand that an LLM operates through probabilistic token prediction rather than deterministic rule-following, they are better equipped to assess appropriate use cases, set realistic performance expectations, and design appropriate governance safeguards. The International Association of Privacy Professionals (IAPP) AI Governance in Practice Report (2024) documented a significant gap between organizations' awareness of generative AI risks and their preparedness to address them [14]. Industry benchmarking aggregated by Knostic (2025) from multiple enterprise surveys drawing in part on a 2024 Deloitte study on AI governance maturity; found that only approximately 9% of organizations have a mature AI governance framework; a preparedness gap that correlates strongly with limited technical literacy at the decision-making level [15]. While this figure should be interpreted with awareness that the source has a commercial interest in highlighting governance gaps, the finding is broadly consistent with the IAPP and Stanford HAI data cited in this section.

The Stanford HAI AI Index 2025 reports that AI-related incidents are rising sharply, yet standardized responsible AI evaluations remain rare among major industrial model developers [12]. Technical literacy enables leaders to ask critical questions during vendor evaluation: How was the model aligned and what red-teaming was conducted? Does the model support audit trails and explainable outputs? Can it be fine-tuned on proprietary data without that data being incorporated into the vendor's shared training corpus?

3.3. Aligning Model Capabilities with Organizational Objectives

Effective AI selection is fundamentally a matching problem: the capabilities of an AI system must align with the specific functional, operational, and ethical requirements of the deploying organization. This alignment requires understanding the distinction between general-purpose models and domain-specialized models, as

well as the strategic trade-offs between proprietary and open-source LLMs. McKinsey reports that over 50% of organizations now use open-source AI technologies, with technology-sector organizations reaching 72%, driven by the need for flexibility, cost control, and freedom from vendor lock-in [16].

Gartner's Hype Cycle for Artificial Intelligence (2025) notes that organizations selecting AI tools based primarily on marketing narratives rather than rigorous capability assessment are more likely to abandon their AI projects [17]. In July 2024, Gartner predicted that 30 percent of generative AI projects would be abandoned after proof of concept by end of 2025, an analyst projection rather than measured outcome data, citing escalating costs, unclear business value, and inadequate risk controls as the primary failure drivers [18]. That prediction appears well-founded; the pattern of premature abandonment it identified remains a documented organizational risk that rigorous upfront selection discipline is designed to prevent.

4. A Framework for AI Selection in Organizations

4.1. Capability-Driven Evaluation Dimensions

A rigorous AI selection process should evaluate candidate systems across multiple interrelated capability dimensions. These dimensions should be assessed against the specific operational context of the organization rather than against general-purpose benchmark scores. **Figure 2** organizes these dimensions across the capability, governance, cost, and lifecycle categories that structure the framework.

- **Accuracy and Reliability:** How consistently does the model produce correct, contextually appropriate outputs for the target use case? Organizations should establish domain-specific test sets and evaluation protocols prior to vendor selection.
- **Transparency and Explainability**: Can the model's output be audited or explained in terms that satisfy regulatory and organizational requirements? For industries subject to the EU AI Act [19], explainability for high-risk AI systems is a compliance obligation, not an optional feature.
- **Adaptability and Integration:** Can the model be fine-tuned, updated, or integrated into existing enterprise systems (such as Microsoft Dynamics 365 F&O, CRM tools, or data warehouses) without prohibitive engineering cost or architectural complexity?
- **Data Governance and Privacy:** Where is inference performed? Is organizational data used to improve the vendor's base model? These questions are critical for organizations subject to GDPR, HIPAA, SOC 2, or sector-specific data residency regulations.
- **Agentic Capability and Autonomy Controls:** As 23 percent of organizations now scale agentic AI systems capable of autonomous multi-step planning and action [2], selection criteria must extend beyond language generation quality to encompass planning reliability, tool-use accuracy, error recovery behavior, and human-override mechanisms. Organizations evaluating agentic systems should assess whether the system supports configurable autonomy boundaries;

how it handles ambiguous or conflicting instructions; what escalation and interruption mechanisms are available; and how its actions are logged for audit and accountability purposes. The risks of agentic AI are categorically different from those of a static LLM; an agent that can take actions requires governance controls proportionate to that expanded agency.

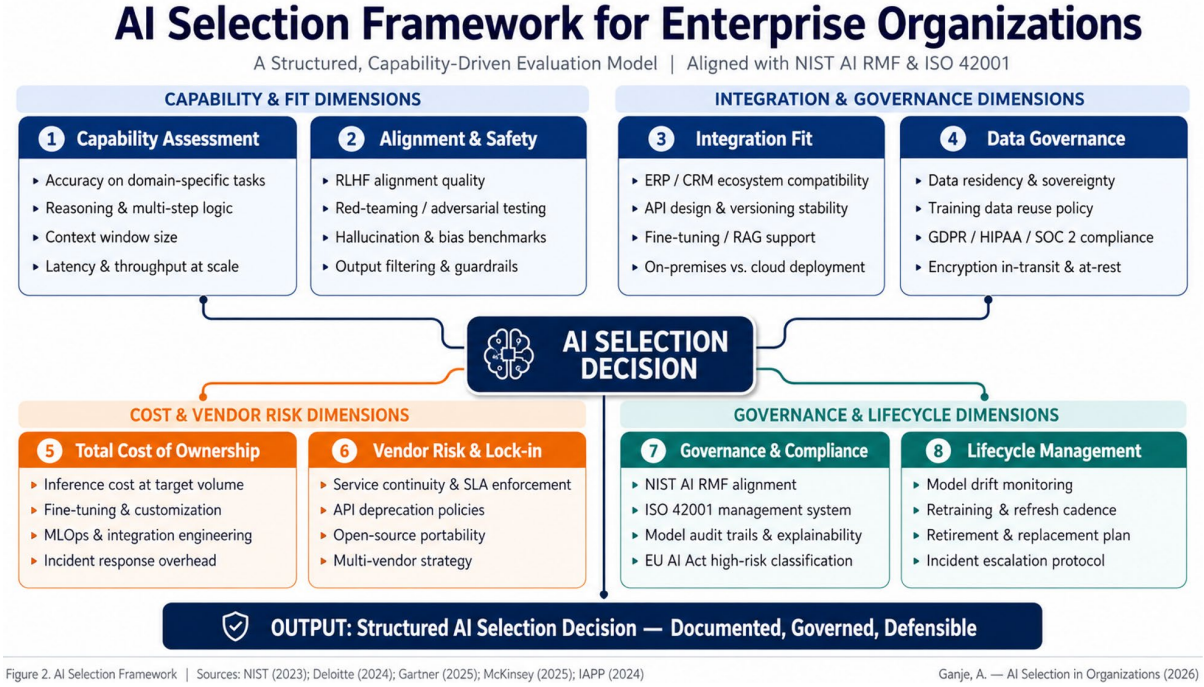

Figure 2. AI selection framework: eight evaluation dimensions across capability, governance, cost, and lifecycle. Sources: [2] [11] [14] [17] [20] [22].

The eight dimensions are not an arbitrary list; each operationalizes one or more requirements of the governance standards on which the framework rests. **Table 3** maps the dimensions to the relevant NIST AI RMF functions and ISO/IEC 42001 elements and distinguishes the criteria that derive directly from the standards from those that represent the author's synthesis for the specific problem of AI selection. Governance and compliance, data governance, alignment and safety, vendor risk, and lifecycle management are largely standard-derived; capability assessment, integration fit, and total cost of ownership extend the standards to address selection-specific concerns they do not directly cover.

Table 3. Mapping of the eight evaluation dimensions to NIST AI RMF functions and ISO/IEC 42001 elements, distinguishing standard-derived criteria from author's synthesis.

Evaluation dimension	NIST AI RMF function	ISO/IEC 42001 element	Origin
Capability assessment	Map; Measure	Performance evaluation (Clause 9)	Synthesis
Alignment and safety	Measure; Manage	Operational controls and impact assessment (Annex A)	Standard-derived

Continued

Integration fit	Map	Operational planning and control (Clause 8)	Synthesis
Data governance	Map; Manage	Data management controls (Annex A)	Standard-derived
Total cost of ownership	Govern	Leadership and resources (Clauses 5, 7)	Synthesis
Vendor risk and lock-in	Govern; Manage	Third-party and supplier controls (Annex A)	Standard-derived
Governance and compliance	Govern	Management system requirements (Clauses 4 - 10)	Standard-derived
Lifecycle management	Manage; Measure	Operation and improvement (Clauses 8, 10)	Standard-derived

4.2. Governance, Risk, and Compliance Criteria

Organizational AI selection must incorporate governance, risk, and compliance (GRC) criteria alongside technical capability assessment. The NIST AI Risk Management Framework (AI RMF 1.0), published in January 2023, provides a widely adopted voluntary governance structure organized around four core functions: Govern, Map, Measure, and Manage [11]. Organizations that align their AI selection and deployment processes with the NIST AI RMF demonstrate a governance maturity increasingly expected by enterprise customers and regulators alike.

ISO/IEC 42001:2023, published by the International Organization for Standardization [ISO], establishes the international standard for AI management systems. It provides a globally recognized compliance framework that organizations can adopt to demonstrate responsible AI governance to regulators, business partners, and stakeholders [20]. Enterprise AI governance platforms operationalize many of these principles through tooling for model inventory management, bias testing, regulatory compliance automation, and AI incident tracking. IBM's watsonx.governance is one example of this class of platform [21]. Together, the NIST AI RMF and ISO 42001 form a complementary governance architecture that covers both risk management practice and management system certification.

4.3. Total Cost of Ownership and Vendor Considerations

AI selection decisions must account for the full total cost of ownership (TCO) rather than focusing narrowly on licensing or subscription fees. Per-token inference pricing for cloud-based LLMs can scale rapidly with usage volume; fine-tuning and customization costs can run into hundreds of thousands of dollars for complex enterprise use cases; and integration engineering, ongoing monitoring, model retraining, and incident response add further operational overhead. Gartner notes that organizations frequently underestimate post-deployment costs, particularly the expense of maintaining model accuracy as organizational data evolves [17].

Vendor lock-in is a systemic risk in the LLM market. Proprietary cloud LLM providers offer performance advantages but create strategic dependency if the vendor changes pricing, access policies, or discontinues products. The growing maturity of open-source LLMs; including Meta's LLaMA 3 family, Mistral, and Falcon provide organizations with credible alternatives that offer greater control, auditability, and TCO predictability for organizations with sufficient ML operations capability [2].

4.4. Lifecycle Management and Avoiding Lock-In

AI systems are not static deployments, they require continuous lifecycle management. Models degrade in performance as the distribution of input data drifts over time, a phenomenon known as model drift or distribution shift. Organizations must establish model performance monitoring frameworks, retraining or refresh pipelines, and clear escalation protocols for when AI system performance degrades below defined operational thresholds. The IAPP AI Governance in Practice Report (2024) documents the significant governance challenges organizations face in managing AI risk, noting that the pace of AI development and evolving regulatory landscape are among the most frequently cited obstacles to building and maturing an effective AI governance program [14].

4.5. Build, Configure, or Buy: An Executive Decision Framework

One of the most consequential and most frequently misframed decisions in enterprise AI adoption is whether to build a solution, configure an existing platform, or purchase a turnkey product. Many organizations default to buy because it feels faster, or to build because it feels more strategic without the analytical foundation to justify either. Effective AI selection requires evaluating six foundational factors before that decision can be made responsibly. These factors do not operate in isolation; they form a structured decision sequence in which each answer constrains the next.

Strategic Differentiation Analysis: The first and most important question is whether the capability to automate creates competitive differentiation or merely delivers operational parity with the market. Commodity capabilities performed identically across competitors, such as meeting transcription, document summarization, or expense classification do not justify a build investment. Differentiated capabilities that encode proprietary process logic, institutional knowledge, or unique domain expertise are the principal candidates where building or fine-tuning creates defensible strategic value. For example, a global steel manufacturer automating production scheduling may be encoding decades of operational expertise into that scheduling logic; this is a build or fine-tune candidate. The same organization using AI to answer HR policy questions is a buy candidate: the output is generic and the competitive value is zero.

Data Readiness and Ownership: A model is only as specialized as the proprietary data used to adapt it. Organizations that lack high-quality, labeled, domain-specific training data cannot justify a fine-tune or build investment regardless of strategic intent. Executives must conduct an honest assessment of data volume, labeling quality, structural integrity, and legal ownership before selecting an AI approach. A financial services firm with fifteen years of annotated credit decisions owns a genuine fine-tuning asset. A firm without structured historical data should instead implement retrieval-augmented generation (RAG) connecting a commercial base model to its proprietary knowledge base at inference time, without modifying model weights. RAG delivers domain-specific outputs at a fraction of the

cost and complexity of fine-tuning, with the additional benefit of keeping organizational knowledge current without retraining cycles.

The Build-to-Buy Spectrum: The build-versus-buy framing is a false binary. There are five meaningful positions on the spectrum, each with distinct capability requirements, cost profiles, and risk characteristics. Full build, training a foundation model from scratch requires extraordinary data assets and AI research infrastructure and is appropriate only for a small number of technology-native organizations globally. Fine-tuning adapts a pre-trained foundation model's weights to a specialized domain and requires proprietary training data, machine learning engineering capability, and a machine learning operations (MLOps) infrastructure for ongoing retraining. RAG connects a commercial base model to organizational knowledge at inference time, with no weight modification, and is the most viable approach for most enterprise knowledge and document applications. Configuration extends to commercial platforms such as Microsoft Copilot Studio, Salesforce Einstein, or ServiceNow Now Assist within vendor-defined guardrails and require no machine learning capability. Turnkey purchase subscribes to a pre-built AI product with no customization, appropriate for commodity tasks where best-in-class commercial output is sufficient. The crossover points at which build becomes more economical than buy requires simultaneously high usage volume and high domain specificity; most enterprises do not reach this threshold for general-purpose LLM applications. Table 4 maps the eight evaluation dimensions across the five sourcing pathways, indicating which dimensions gate each pathway and when each is appropriate.

Table 4. Adaptation of the eight evaluation dimensions across the five sourcing pathways. The dimensions that gate each pathway differ; the framework is therefore applied selectively rather than uniformly.

Sourcing pathway	Dimensions that dominate the decision	When the pathway is appropriate
Full build (train from scratch)	Capability; data governance; organizational capability; total cost of ownership	Extraordinary proprietary data scale and dedicated AI research infrastructure; almost no enterprise meets this threshold.
Fine-tune a foundation model	Capability; data governance; lifecycle management; vendor risk	High strategic differentiation plus high-quality labeled proprietary data and sustained MLOps capability.
Retrieval-augmented generation (RAG)	Integration fit; data governance; total cost of ownership	Differentiated knowledge but no labeled training data or MLOps capacity; the most viable path for most enterprise knowledge applications.
Configure a commercial platform	Integration fit; governance and compliance; vendor risk and lock-in	Differentiation is moderate and an existing platform (for example, Copilot Studio or Einstein) covers the workflow within vendor guardrails.
Turnkey purchase	Capability; governance and compliance; total cost of ownership	Commodity tasks with no competitive value where best-in-class commercial output is sufficient.

Total Cost of Ownership Across the Lifecycle: AI selection decisions must be evaluated against the full total cost of ownership (TCO) over a three-to-five-year horizon, not against initial licensing or engineering costs alone. Build and fine-

tune approaches carry high upfront data preparation, engineering, and infrastructure costs, but zero vendor dependency risk. Turnkey purchases appear low-cost initially but accumulate per-token inference pricing at scale, integration engineering overhead, and potentially severe exit costs if the vendor changes access policies or discontinues the product. The most routinely underestimated cost category across both approaches is ongoing MLOps: monitoring for model drift, retraining pipelines, evaluation frameworks, and incident response. Organizations that budget only for deployment and not for sustained operations consistently underperform on AI ROI measures [17].

The three-year TCO horizon is used as a practical base-case planning period rather than as a fixed regulatory requirement. A multi-year horizon is necessary because AI costs frequently shift after deployment through inference growth, integration maintenance, monitoring, evaluation, retraining, governance, incident response, and vendor transition costs. Three years is long enough to capture implementation and steady-state operation, while remaining short enough to reflect the volatility of AI model markets, pricing, and vendor capabilities. Organizations with longer contract terms, dedicated infrastructure, or regulated operating environments, or major build investments may extend the analysis to five years.

Security, Compliance, and Data Residency: The architecture of where AI inference occurs is a non-negotiable compliance dimension that is frequently treated as an afterthought. Three postures define the decision space. Shared cloud API endpoints transmit data to a vendor's inference infrastructure; this is unacceptable for protected health information (PHI), personally identifiable information (PII), attorney-client privileged content, or classified material without explicit data processing agreements and confirmation that organizational data is excluded from model training. Dedicated or private cloud endpoints such as Azure OpenAI Service or AWS Bedrock with VPC isolation keep data within the organization's cloud tenant and satisfy most regulated enterprise requirements. On-premises or air-gapped deployment runs the model entirely within the organization's own infrastructure, is required for classified government and certain defense applications, and constrains selection to open-source models such as Meta's LLaMA 3, Mistral, or Falcon, or to vendors offering on-premises deployment options. A government contractor handling Controlled Unclassified Information (CUI) cannot legally use public cloud API endpoints; that single compliance constraint eliminates the majority of SaaS AI products and forces an architecture decision before any capability comparison is valid.

Organizational Capability and Talent: The most common AI selection failure mode is choosing a build or fine-tune approach without the machine learning engineering, MLOps, and data science capability required to sustain it. A model that cannot be monitored, retrained, or evaluated against current organizational data degrades silently producing outputs that appear functional but have drifted from operational reality. Executives must honestly audit four capability questions be-

fore any build investment: Does the organization employ ML engineers capable of managing model drift and retraining pipelines? Is a structured data pipeline in place capable of continuous retraining? Is there a model registry and evaluation framework that can measure output quality against defined thresholds? Is there a responsible AI governance function capable of auditing model outputs for bias and accuracy? If the answer to most of these is no, the correct selection answer is buy or configure, regardless of what the strategic differentiation or data analysis suggests, not because building is wrong in principle, but because the organization cannot execute it safely or sustainably.

Executive decision sequence: build, configure, or buy?

Apply these questions in order. Each answer constrains the decision space for the next.

- **Does compliance or data residency require on-premises or dedicated cloud deployment?** Yes, On-premises open-source model or dedicated private cloud endpoint. This eliminates most SaaS products before any capability evaluation begins.
- **Does this capability create genuine competitive differentiation?** No, Buy or configure. Do not invest engineering resources in capabilities that commoditize equally across competitors.
- **Do we own sufficient high-quality, labeled, proprietary training data?** No then move RAG over a commercial foundation model. Connect your knowledge base at inference time rather than attempting to encode it through fine-tuning without the data assets to do so reliably.
- **Do we have the ML engineering and MLOps capability to build and sustain a fine-tuned model?** No then move RAG or manage fine-tuning via a vendor service. Building without the operational capability to maintain the result creates technical debt and silent model degradation.
- **Is the three-year TCO base case for fine-tuning defensible against an equivalent buy or RAG approach?** No then Revisit the spectrum. The crossover points where build outperforms buy economically requires simultaneously high volume and high domain-specificity; most enterprises do not reach it.
- **All preceding answers point to fine-tune or build?** Yes, then proceed with fine-tuning on a commercial foundation model using your proprietary data. Full model training from scratch is reserved for organizations with extraordinary data scale and AI research infrastructure threshold almost no enterprise organization meets.

Figure 3 translates the build, configure, or buy logic into a step-by-step decision sequence that a selection team can apply directly. The team begins at Gate 1 by identifying any non-negotiable deployment constraints, such as dedicated cloud, private endpoint, on-premises hosting, air-gapped deployment, or data residency restrictions. These constraints narrow the eligible solution set but do not end the evaluation.

After deployment constraints are established, Gates 2 through 5 determine whether fine-tuning is justified. A "Yes" answer keeps the team moving down the sequence toward fine-tuning. A "No" answer exits to the recommended sourcing path shown on the right, such as buying or configuring a commercial solution, using retrieval-augmented generation, selecting a vendor-managed fine-tune, or revisiting the sourcing spectrum. Reaching the final box means the use case has

cleared the strategic, data, capability, and economic gates required to justify fine-tuning a commercial foundation model on proprietary data.

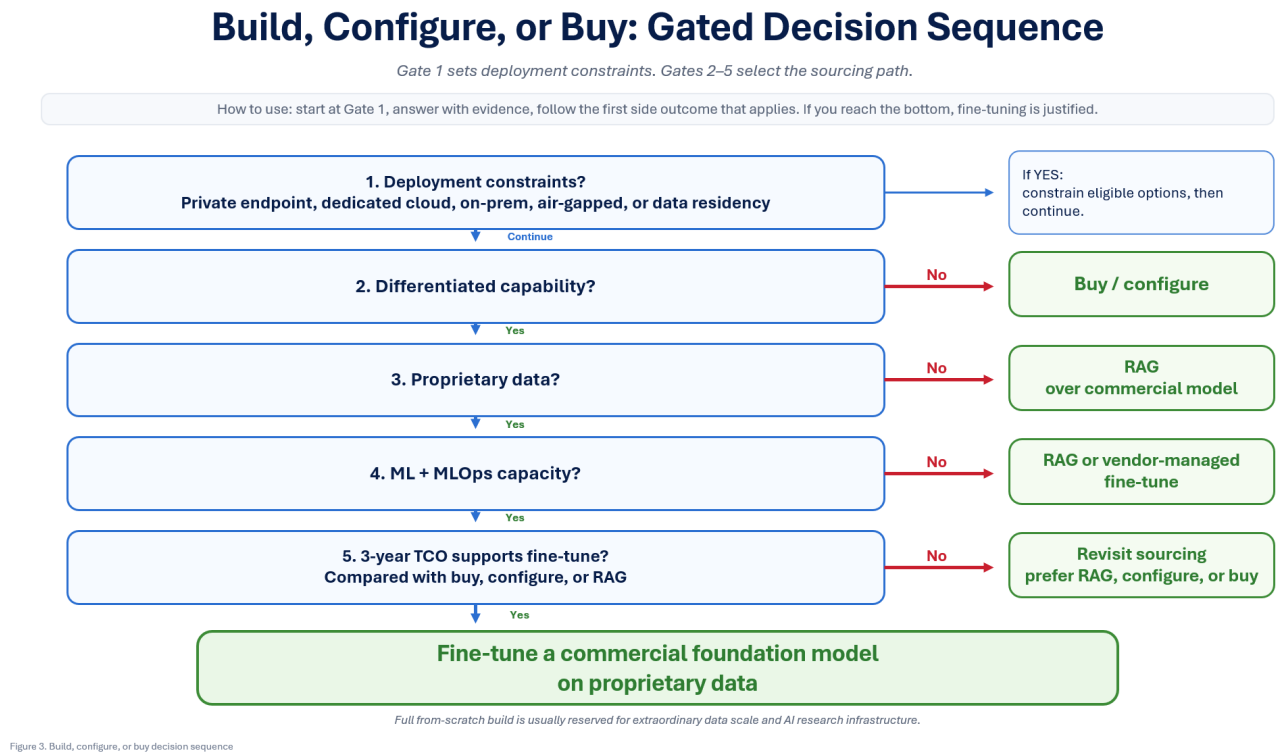

Figure 3. The build, configure, or buy decision as a gated sequence. Each gate is non-negotiable and constrains the next; a "No" advances the decision down the sequence while a side branch yields the appropriate sourcing outcome.

4.6. Applying the Framework: A Practical Decision Sequence

The eight dimensions of the AI Selection Framework are not a checklist to be completed in parallel. They form a structured decision sequence in which each stage gates the next. Organizations that attempt to evaluate all dimensions simultaneously typically produce politically negotiated selections rather than analytically defensible ones. The sequence below is designed to be completed by a cross-functional team (typically spanning technology leadership, legal or compliance, finance, and the primary business owner) over a structured evaluation period of two to six weeks depending on organizational complexity.

- **Step 1. Define the use case and success criteria before evaluating any vendor.** No model evaluation is valid without a defined use case and measurable success criteria. Before issuing an RFP or conducting a proof of concept, the evaluation team must document: the specific task the AI system will perform, the expected input and output format, the volume and latency requirements, and the quantitative threshold that constitutes acceptable performance. This document becomes the evaluation specification and prevents vendors from reframing the use case around their product strengths during the selection process.
- **Step 2. Apply Dimension 1 (Capability Assessment) using your own evalu-**

ation data, not vendor benchmarks. Run candidate models against the evaluation specification developed in Step 1 using representative samples of the organization's actual data. General-purpose benchmarks such as MMLU or HumanEval measure academic performance and frequently do not predict production accuracy for domain-specific tasks. Evaluation should measure accuracy on the target task, reasoning consistency across edge cases, hallucination rate under realistic prompting conditions, and latency at the expected concurrency level. Models that fail on the organization's own evaluation data are eliminated at this stage regardless of their public benchmark rankings.

- **Step 3. Screen for non-negotiable compliance requirements using Dimension 4 (Data Governance).** Before investing further evaluation effort, the compliance and legal team should screen remaining candidates against the organization's non-negotiable requirements: data residency jurisdiction, training data reuse policy, applicable certification requirements such as SOC 2 Type II, FedRAMP, or HIPAA, and the inference architecture required for the use case. A model that performs well on capability but fails a mandatory compliance screen must be eliminated. This step is deliberately placed early to avoid the organizational pressure that builds when a technically preferred vendor fails compliance review late in the process.
- **Step 4. Assess alignment and safety characteristics using Dimension 2 before proceeding to commercial evaluation.** Organizations should request and review each vendor's published model card, red-team evaluation results, and bias benchmark documentation. Where this documentation does not exist, its absence is itself a governance signal. The evaluation team should test the model against adversarial prompts relevant to the deployment context and assess whether the model's safety alignment profile is appropriate for the sensitivity of the use case. High-stakes use cases (those affecting credit, employment, healthcare, or public-facing communications) require stricter alignment evidence than internal productivity tools.
- **Step 5. Model the total cost of ownership across a three-year horizon using Dimension 5 (Total Cost of Ownership).** Procurement teams typically receive per-token or per-seat pricing and treat it as the cost of the system. A realistic TCO model for an enterprise AI deployment must include: inference cost at projected production volume including peak load scenarios, fine-tuning and customization cost if required, integration engineering and ongoing maintenance, model retraining or refresh cycles, monitoring and observability infrastructure, and incident response overhead. Organizations should model at least three volume scenarios; baseline, 2x growth, and 5x growth, to understand how costs scale and at what threshold a different architectural choice becomes economically superior.
- **Step 6. Evaluate integration fit and ecosystem depth using Dimension 3 (Integration Fit).** A capable model that cannot integrate cleanly with existing enterprise systems creates hidden project risk. The evaluation team should test

API stability under realistic load, confirm versioning and deprecation policies in writing, validate RAG and tool-use support if required by the use case, and confirm that the model can be deployed in the infrastructure tier required by the compliance assessment in Step 3. Integration failures discovered during implementation, after vendor selection; are among the most common and expensive causes of AI project abandonment.

- **Step 7. Assess vendor risk and lock-in exposure using Dimension 6 (Vendor Risk and Lock-in).** The evaluation team should obtain written confirmation of: the vendor's API deprecation notice period and version pinning policy, contractual data portability rights including the ability to export fine-tuned weights or training data, the availability of an open-source or self-hostable fallback option, and the terms governing model version updates that may alter output behavior. Service continuity and SLA enforcement should be validated at this stage; uptime guarantees and latency commitments must be confirmed under realistic load conditions, and the financial remedies available when SLAs are breached must be specified contractually before selection is finalized. Organizations should also assess dependency concentration risk and apply a multi-vendor strategy where that concentration exceeds acceptable thresholds.
- **Step 8. Confirm governance and compliance readiness using Dimension 7 (Governance and Compliance).** The evaluation team should verify that each candidate model can be governed in alignment with the organization's regulatory obligations and internal standards. This includes confirming alignment with the NIST AI Risk Management Framework, evaluating the vendor's support for ISO/IEC 42001 management system requirements, assessing the availability of model audit trails and explainability mechanisms required for the use case, and determining whether the system falls within the EU AI Act's high-risk classification. For regulated industries, governance and compliance validation is a gate; not an afterthought and should produce documented evidence that can be presented to internal audit, legal counsel, or an external regulator.
- **Step 9. Confirm lifecycle governance readiness using Dimension 8 (Lifecycle Management).** Before a selection decision is finalized, the operating team must confirm that they have the capability to execute ongoing lifecycle management: model performance monitoring against defined KPIs, a defined retraining or refresh trigger and cadence, a retirement and replacement plan if the selected model is deprecated, and an escalation protocol for performance degradation incidents. AI systems that are selected without a functioning lifecycle governance plan tend to degrade silently, creating compounding risk as output quality declines without detection.
- **Step 10. Document the selection decision with a structured rationale aligned to all eight dimensions.** The output of this framework is not a vendor recom-

mendation, it is a documented, defensible decision. The selection record should capture the evaluation specification from Step 1, the results of each dimension assessment, the rationale for eliminations at each gate, and the scoring or weighting rationale used to differentiate surviving candidates. This documentation serves as the governance record for the selection decision and is the foundation for the model audit trail required by the NIST AI RMF and ISO/IEC 42001 [11] [20]. Organizations that skip this documentation step expose themselves to regulatory and contractual liability when AI system performance is later questioned.

The ten-step sequence above can be compressed or expanded depending on the risk profile of the deployment. Low-stakes internal productivity tools may complete the sequence in a two-week structured sprint. High-stakes deployments, those affecting regulatory compliance, customer-facing decisions, or critical infrastructure, should treat each step as a formal milestone with documented sign-off from the appropriate organizational authority. The framework is intentionally vendor-agnostic and applies equally to proprietary cloud LLMs, open-source foundation models, and embedded AI platforms.

4.7. Worked Example: Applying the Ten-Step Sequence

To show how the dimensions interact in practice, the ten-step sequence is applied below to a single use case: a regional bank evaluating an AI assistant to answer customer loan-status inquiries. The case is illustrative and is chosen because it combines a regulated data environment with a customer-facing, elevated-risk use, so that several gates bind. Table 5 traces the decision through each step.

Table 5. Illustrative application of the ten-step decision sequence to a regional bank selecting an AI assistant for customer loan-status inquiries.

Step	Action in this case	Result
1. Define use case and success criteria	Document the task, input and output format, expected volume and latency, and a required factual-accuracy threshold on a 200-item bank-specific test set	Evaluation specification fixed before any vendor contact
2. Capability on own data	Run candidates against the 200-item test set rather than public benchmarks	Two of five candidates fail the accuracy and hallucination thresholds and are eliminated
3. Compliance screen (data governance)	Apply data-residency, GLBA, and SOC 2 screens; require exclusion of data from vendor training	Shared public-API products eliminated; only dedicated-cloud options remain
4. Alignment and safety	Review model cards and red-team evidence; run adversarial tests for a customer-facing financial context	One remaining model lacks documented red-teaming and is down-weighted
5. Three-year total cost of ownership	Model inference cost at baseline, 2x, and 5x volume, plus integration and monitoring	Per-token pricing at 5x makes one option uneconomical; a retrieval-augmented mid-tier model is most cost-stable
6. Integration fit	Validate API stability under load, versioning policy, and retrieval support against core banking systems	Selected approach integrates with the existing knowledge base without re-platforming
7. Vendor risk and lock-in	Confirm deprecation notice period, data-portability rights, and a self-hostable fallback in writing	Acceptable; a multi-vendor fallback is identified

Continued

8. Governance and compliance	Assess NIST AI RMF and ISO 42001 alignment, audit trails, and EU AI Act high-risk classification	Treated as elevated risk; required audit trail confirmed available
9. Lifecycle readiness	Confirm drift monitoring, refresh cadence, and a degradation escalation protocol with operations	Operating team confirms it can sustain the model
10. Document the decision	Record the rationale against all eight dimensions with elimination reasons at each gate	Defensible selection record produced for internal audit

The sequence converges on retrieval-augmented generation over a mid-tier model deployed on a dedicated cloud endpoint. The decisive dimensions were data governance, capability measured on the bank own data, and total cost of ownership; capability alone would not have produced this outcome. A conventional vendor-led process, by contrast, would have begun with capability demonstrations and discovered the residency and cost constraints only after a preferred product had gained internal momentum, precisely the late-stage failure the early gates are designed to prevent.

5. Preventing Extreme AI Proliferation

5.1. The Proliferation Risk Landscape

AI proliferation, the rapid, often uncoordinated expansion of AI tools across an organization without adequate governance poses significant strategic and operational risks that compound over time. Cyberhaven's Q2 2024 AI Adoption and Risk Report drawn from telemetry data across its own customer base of over three million workers found that the volume of corporate data flowing into AI tools grew by 485% from March 2023 to March 2024, with the majority of usage occurring through personal, non-enterprise accounts including 73.8% of ChatGPT usage and over 94% of Gemini usage bypassing organizational security controls [23]. Readers should note that Cyberhaven is a data loss prevention vendor and that this data reflects its customer population rather than a randomized enterprise sample. The World Economic Forum (WEF) Responsible AI Innovation Playbook (2025) identifies uncontrolled AI proliferation as a systemic enterprise risk requiring proactive governance architecture not reactive policy enforcement after incidents occur [22].

Shadow AI, the unsanctioned use of AI tools by employees outside of IT-approved channels, represents one of the most acute governance challenges facing enterprises as of 2026. Reco.ai's State of Shadow AI Report (2025); produced by a vendor specializing in SaaS security and shadow IT detection found that enterprise organizations averaged 269 unsanctioned AI applications per 1,000 employees, with these unauthorized applications persisting undetected for over 400 days on average, creating persistent, long-horizon data exposure risks [24]. These figures are drawn from Reco.ai's own customer environments and should be treated as indicative rather than representative of all enterprises. When employees input confidential business data, customer information, or unreleased product details

into consumer AI tools, that data may be incorporated into vendor training datasets or transmitted to third-party servers in non-compliant jurisdictions.

5.2. Governance Structures to Manage Proliferation

Effective prevention of extreme AI proliferation requires structural governance interventions at three complementary levels. First, enterprise-level AI policy establishes the organizational guardrails within which AI adoption occurs. This includes AI acceptable use policies, data classification standards governing which data categories may be used with which classes of AI tools, and an approved AI tool registry. Industry data from Knostic (2025) indicates that formal AI governance programs remain the exception rather than the rule, with only roughly one in four organizations reporting fully operational AI governance despite widespread awareness of regulatory requirements [15].

Second, technical controls provide enforcement mechanisms that complement policy governance. API gateway management, data loss prevention (DLP) systems configured to detect and restrict sensitive data transmission to AI endpoints, and network monitoring can identify and block unauthorized AI tool usage before it creates material risk. Third, cultural and educational interventions address the human dimension of AI proliferation. The Stanford HAI AI Index 2025 found that AI-related incidents rose to a record 233 in 2024, a 56.4% increase over 2023, a count compiled from public incident reporting and therefore sensitive to reporting and detection rates, while a gap persists between organizations recognizing responsible AI risks and taking meaningful steps to address them [12]. The Wharton School's AI Adoption Report (2025) found that organizations prioritizing workforce training and capability-building alongside AI investment are better positioned to realize sustained returns, noting that human capital development, not technology alone, is the primary lever distinguishing organizations that achieve durable AI value [13].

5.3. Responsible AI Principles in Practice

Responsible AI is both a compliance obligation and a strategic differentiator. McKinsey's 2025 State of AI survey found that a majority of organizations report that AI has improved innovation, and nearly half report measurable improvement in customer satisfaction and competitive differentiation, qualitative enterprise benefits that accrue even among organizations that have not yet achieved material EBIT impact [2]. However, despite broad recognition of its importance, only approximately 6% of organizations qualify as AI high performers, achieving measurable enterprise-wide EBIT impact and significant value from AI, while the vast majority remain in the experimenting or piloting stages [2]. The NIST AI RMF identifies seven characteristics of trustworthy AI systems: "valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with their harmful biases managed" ([11], p: 12). Organizations should use these characteristics as evaluation criteria during AI selec-

tion and as operational standards during deployment and monitoring.

6. Research Design and Validation Approach

The preceding sections present a framework rather than report an experiment, and the appropriate methodological frame for this contribution is design science [25] [26]: the construction of a purposeful artifact to address a defined organizational problem, followed by evaluation of that artifact. This section makes the research design explicit. It describes how the framework was developed, defines the criteria by which its effectiveness can be measured, and specifies the retrospective, comparative case approach by which the framework is validated in Section 7.

6.1. Framework Development

The framework was developed through a structured synthesis of three sources of evidence. The first is the established governance literature, principally the NIST AI Risk Management Framework (AI RMF 1.0) and ISO/IEC 42001:2023, which define the risk and management-system requirements that any responsible selection process must satisfy. The second is the peer-reviewed literature on large language model architecture, alignment, and limitations, which establishes the technical properties that determine enterprise suitability. The third is current industry evidence on adoption, abandonment, governance maturity, and proliferation, used to characterize the practical conditions under which selection decisions are actually made. The eight evaluation dimensions and the build, configure, or buy sequence were derived by mapping the requirements identified in these sources onto the decision points a technology executive must resolve, and by ordering those decision points so that non-negotiable constraints are resolved before discretionary comparison begins.

6.2. Success Criteria and Measurability

A selection methodology is only scientific if its effectiveness can be assessed against defined indicators rather than asserted. This framework defines a good selection decision as one that is defensible, durable, and economically accurate, and proposes four measurable indicators accordingly. The first is decision auditability: whether the selection produces a documented rationale, traceable to each of the eight dimensions, that can be presented to internal audit or an external regulator. The second is avoidance of premature abandonment: whether systems selected through the framework survive beyond proof of concept at a higher rate than the roughly thirty percent abandonment rate reported for generative AI projects [18], which serves as a baseline. The third is total-cost-of-ownership forecast accuracy: the variance between the three-year cost projected during selection and the cost realized in operation. The fourth is lock-in exposure: the presence or absence of contractual data-portability rights and a viable fallback option at the point of selection. These indicators are intended to be measured comparatively, against decisions made without the framework, in the field studies described below; they

are stated here so that the framework can be evaluated rather than merely adopted.

6.3. Validation through Standards Coverage and Retrospective Comparative Cases

Beyond demonstrating that the framework can be applied, its central design claim, that these are the right dimensions, evaluated in this order, requires validation that does not depend on primary data collection. Two complementary, panel-free approaches are used. The first is construct-coverage validation: the dimension-to-standard mapping in Table 3 establishes that the eight dimensions collectively cover the requirements of the NIST AI RMF and ISO/IEC 42001 without material gaps and without redundant criteria, which is evidence that the criteria set is complete and well-formed. The second is retrospective comparative case analysis: the framework is applied after the fact to documented enterprise AI deployments, both reported successes and the well-documented abandonments, to test whether its gates would have predicted or prevented the observed outcome. This approach uses real, citable evidence rather than recruited opinion, and it directly addresses whether the framework explains the failure patterns reported in the industry literature, such as abandonment after proof of concept for reasons of cost, unclear value, or inadequate controls. Section 7 applies this retrospective analysis to six documented enterprise deployments, four failures, a partial reversal, and a success, and Section 8.4 sets out the expansion and prospective testing that follow.

7. Retrospective Comparative Case Analysis

This section validates the framework by applying it retrospectively to six publicly documented enterprise AI deployments and testing whether its gates align with the outcomes that actually occurred. The cases were chosen to vary the binding constraint and the result: four are well-reported failures in which different gates were violated, one is a partial reversal of an over-automated rollout, and one is a disciplined success that followed the framework's logic. Sections 7.1 to 7.3 analyze three of these in detail and Section 7.4 treats the remainder more briefly, with Section 7.5 synthesizing the set. Evidence is drawn from public reporting, a civil tribunal decision, a university audit, and, for the success, vendor-published material that is read with appropriate caution. Because the outcomes were known before the analysis, this is a test of explanatory rather than predictive validity, and its limits are discussed in Section 8.2.

Case selection. The six cases were selected purposively rather than at random, against three criteria. First, each had to be publicly documented with a verifiable outcome, such as a tribunal decision, a university audit, named company statements, or sustained press reporting. Second, the set as a whole had to span different binding gates and both success and failure, so that the framework was tested against varied conditions rather than a single failure mode. Third, each had to be an enterprise-scale AI deployment from the period 2012 to 2025, most of them generative or conversational systems within the scope defined in Section 1.1; one

case predates the large language model era and is included because the gates it turned on are among the dimensions that, as Section 1.1 notes, are not specific to generative systems. This is a deliberately illustrative and non-exhaustive sample, not a random or systematic one. Because the cases were chosen with their outcomes already known, the analysis can show that the framework is consistent with the documented record but cannot establish predictive power, and it remains exposed to selection and confirmation bias; a systematically sampled and independently coded case set is identified in Section 8.4 as the next step. The cases are therefore offered as explanatory evidence, not as proof of effectiveness.

The selection protocol can be stated explicitly. The inclusion criteria were three: a publicly verifiable outcome from an independent source (a tribunal decision, a published audit, a named company statement, or sustained press reporting); an enterprise-scale deployment rather than an individual or consumer use; and a proximate cause that could be attributed to a specific framework dimension. The exclusion criteria removed deployments known only through vendor marketing without independent corroboration, cases whose outcome was still unsettled at the time of writing, and cases in which no single binding gate could be identified. Outcomes were coded as failure when a deployment was abandoned, cancelled, or produced organizational liability; as a partial reversal when an initial rollout was later materially walked back; and as success when the system reached sustained production use without reversal. Each case was then gate-coded by mapping its documented proximate cause to the framework dimension it most directly implicated; where more than one dimension applied, the earliest gate in the sequence was recorded as binding. The six cases were chosen, against these criteria, to populate distinct binding gates (alignment and safety, capability, integration and data readiness, governance and compliance, and lifecycle management) and both outcome types, so that the analysis tested the framework across its range rather than at a single point.

7.1. Case 1: Air Canada, a Customer-Facing Deployment without Verification (Failure)

In 2022 Air Canada deployed a customer-facing website chatbot that, when asked about bereavement fares, gave a passenger inaccurate guidance; the airline declined the resulting refund, and in February 2024 the British Columbia Civil Resolution Tribunal held Air Canada liable for negligent misrepresentation by its chatbot [27]. Read through the framework, the deployment failed at the alignment-and-safety dimension and at the governance-and-compliance gate. The use was customer-facing and consequential, the category for which the framework requires verification layers, human-in-the-loop review, and documented accountability for outputs before deployment; none was evident, and the model produced confident but incorrect information of exactly the kind the alignment-and-safety dimension is meant to surface. The framework treats such a use as high-stakes and would have gated it on demonstrated output verification and a clear ownership-of-error position rather than on capability alone. The tribunal outcome, organi-

zational liability for the system's output, is the materialization of the risk the governance-and-compliance dimension exists to manage.

7.2. Case 2: McDonald's and IBM, Capability Unproven on Real Operating Data (Failure)

Between 2021 and 2024 McDonald's and IBM piloted an automated voice-ordering system at more than one hundred United States drive-thrus; after persistent ordering errors under real conditions (background noise, overlapping lanes, and varied accents), the companies ended the pilot in mid-2024 [28]. In framework terms the deployment failed at the capability-assessment gate as the framework specifies it: capability must be measured on the organization's own representative data and operating conditions, at the expected concurrency, rather than on benchmark or laboratory performance. A voice-ordering task in a noisy, multi-speaker drive-thru is precisely the edge-case-laden, latency-sensitive environment the framework directs evaluators to test before selection. The pilot reportedly performed acceptably in controlled settings but not in production, which is the failure mode the capability dimension targets; an evaluation against representative operating data would have surfaced the accuracy shortfall before scale-up rather than after.

7.3. Case 3: Morgan Stanley, Disciplined Selection of a Sanctioned System (Success)

From 2023 Morgan Stanley Wealth Management deployed an internal GPT-4 assistant that lets financial advisors query the firm's proprietary research; the firm reports adoption of approximately ninety-eight percent of advisor teams [29]. The deployment tracks the framework gates closely, which is why it serves as a positive control. Sourcing followed the build-configure-buy logic toward retrieval-augmented generation over proprietary documents rather than a full build, matching the data-readiness and organizational-capability analysis the framework prescribes. Capability was assessed through an evaluation framework run on the firm's own scenarios before deployment, satisfying the capability dimension as specified. Data governance was addressed through secure hosting and exclusion of data from vendor training, the data-governance gate for a regulated firm. Alignment and safety were managed by keeping advisors in the loop to review outputs, and integration fit was achieved by wiring outputs into existing advisor workflows. The adoption figure is firm- and vendor-reported and should be read as such, but it is the design choices, not the headline metric, that align with the framework: the deployment cleared in sequence the gates the two failures violated.

7.4. Additional Documented Cases

IBM and the MD Anderson Cancer Center spent roughly $62 million between 2012 and 2016 building the Watson-based Oncology Expert Advisor, which was cancelled before it was used on patients; a University of Texas audit attributed the

failure largely to the system never integrating with the hospital's electronic health records and to scope and management problems rather than to the underlying model [30]. In framework terms the deployment failed at the intersection of capability, integration fit, and data readiness: a build undertaken without a working data pipeline into the system of record cannot be sustained, which is exactly the organizational-capability and integration gates the framework places ahead of a build commitment. This case predates the generative AI era and is not itself a large language model deployment; it is included because capability, integration fit, and data readiness are among the six dimensions that Section 1.1 identifies as transferring beyond generative systems, and the build it represents failed on exactly those transferable gates.

New York City's MyCity chatbot, launched in 2023 on a commercial cloud model to advise small businesses, was found in 2024 to give confidently illegal guidance on labor, housing, and licensing rules, yet was kept online for an extended period [31]. The deployment failed at the governance-and-compliance gate for a public-facing advisory use: high-stakes regulatory advice demands the verification, audit, and accountability controls the framework treats as non-negotiable for that risk class, and their absence produced systemic misinformation rather than an isolated error.

Klarna's OpenAI-powered customer-service assistant, launched in 2024, handled a large share of contacts and was promoted as equivalent to hundreds of agents [32]; by 2025 the company publicly acknowledged that an aggressive AI-first posture had lowered service quality on complex cases and began rehiring human agents into a hybrid model [33]. Read through the framework, the early gates were cleared, capability and integration were real, but the lifecycle-management dimension was under-weighted: treating selection as a one-time replacement rather than a monitored, human-in-the-loop arrangement is the failure the lifecycle gate is designed to prevent, and the correction restored the human-oversight posture the framework prescribes.

7.5. Cross-Case Synthesis

The six cases vary the binding gate while holding the framework constant, and in each the gate the case turned on corresponds to the documented outcome (**Table 6**). The failures are explained not by weak technology in general but by a specific unaddressed gate (accountability and verification in the two public-facing chatbots, capability on real operating data in the drive-thru, and capability, integration, and data readiness in the healthcare build), while the disciplined success cleared those gates and the partial reversal corrected a lifecycle gate it had initially under-weighted. This is the pattern the framework predicts: outcomes turn on whether the early, non-negotiable gates were cleared and sustained, not on model quality in the abstract. The evidence is consistent with, though it does not prove, the central claim of the framework that sequencing these gates ahead of capability comparison is what separates durable selections from abandoned ones.

Table 6. Retrospective comparative case analysis: each documented outcome mapped to the framework gate the case turned on.

Case (sector)	Documented outcome	Gate the case turned on	Consistent with framework?
Air Canada chatbot (airline, customer-facing)	Inaccurate output; tribunal found the airline liable [27]	Alignment and safety; governance and compliance	Yes, gate unaddressed
NYC MyCity chatbot (public sector, advisory)	Public chatbot gave illegal regulatory advice [31]	Governance and compliance; verification	Yes, gate unaddressed
McDonald's and IBM drive-thru (food service, operations)	Persistent ordering errors; pilot ended in 2024 [28]	Capability on real operating data and concurrency	Yes, gate unaddressed
Watson at MD Anderson (healthcare, build)	About $62M; cancelled in 2016; never reached patients [30]	Capability; integration fit; data readiness	Yes, gates unaddressed
Klarna assistant (fintech, customer-facing)	Rapid automation, then partial rehire for quality [32]	Lifecycle management; human-in-the-loop	Partly, early gates met, lifecycle gate corrected later
Morgan Stanley assistant (financial services, internal)	High reported adoption of an internal assistant [29]	Data governance; capability evaluation; RAG sourcing; human-in-the-loop	Yes, gates cleared, success

8. Discussion

8.1. Relationship to Existing Frameworks

The framework is positioned against three bodies of prior work. Governance frameworks such as the NIST AI RMF and ISO/IEC 42001 specify what responsible AI management requires but are deliberately not selection methods: they govern systems already chosen and do not sequence the capability, sourcing, and lifecycle decisions that precede governance. Generic technology build-versus-buy and total-cost-of-ownership models address sourcing but predate the properties that make AI selection distinctive (probabilistic rather than deterministic behavior, alignment quality as a procurement variable, and the categorically different risk profile of agentic systems), and therefore omit the dimensions on which AI decisions most often fail. Capability benchmarks such as MMLU and HumanEval measure general model performance but, as the framework argues, frequently do not predict production accuracy on domain-specific tasks. The contribution of this paper is to integrate these otherwise separate concerns into a single sequenced process in which governance constraints, capability evidence drawn from the organization's own data, sourcing economics, and lifecycle readiness are evaluated as dependent gates rather than parallel checklists. Few existing frameworks integrate large language model capability assessment, AI governance standards, sourcing economics, build-configure-buy decisioning, lifecycle management, and proliferation control into a single sequenced enterprise selection method. Recent adjacent work, including buy-versus-build models for public-sector large language model adoption and operational and agentic-AI governance approaches aligned to the NIST AI RMF, addresses parts of this space, and Table 7 positions the present framework against those bodies of work.

Table 7. The proposed framework positioned against adjacent bodies of work.

Body of work	Primary focus	What it covers	What it does not provide
NIST AI RMF	AI risk governance	Govern, map, measure, and manage AI risk for systems in use	No selection or sourcing method; assumes the system is already chosen
ISO/IEC 42001	AI management system	Organizational processes for responsible AI governance	No capability evaluation, sourcing economics, or build-configure-buy decisioning
Operational and agentic AI governance frameworks	Governance of deployed and agentic AI	Control mapping and oversight, often aligned to the NIST AI RMF	Limited treatment of selection, sourcing economics, and lifecycle cost
Build-versus-buy and sourcing frameworks (including recent public-sector LLM models)	Sourcing decision	Build and buy trade-offs, sometimes sovereignty, cost, and sustainability	Generic or domain-specific; limited LLM capability evaluation, standards mapping, and proliferation control
TCO and IT procurement models	Lifecycle cost analysis	Multi-year cost modeling and vendor comparison	Not AI-specific; omit alignment, agentic risk, and governance gating
This framework	Sequenced enterprise AI selection	Integrates capability, governance standards, sourcing economics, build-configure-buy, lifecycle TCO, and proliferation control as ordered gates	Prospective validation still pending (Section 8.4)

8.2. Interpreting the Case Analysis

The case analysis provides convergent rather than conclusive support for the framework. In each of the six documented deployments the gates align with what actually happened: the failures violated a gate the framework evaluates early, and the success satisfied them. This is evidence of explanatory validity, the framework accounts for observed outcomes, but not of predictive power, because the cases were selected after their outcomes were known and were analyzed by the author. Retrospective fit of this kind is vulnerable to confirmation bias and to the small number of cases. The appropriate reading is that the framework is consistent with the documented record and has not been contradicted by it, which justifies the prospective testing described in Section 8.4 rather than a claim of demonstrated effectiveness. In the terms of Section 1.3, the six cases are consistent with propositions P1 and P2 and provide explanatory support for RQ2, while RQ1 is addressed by the dimension derivation in Sections 4 and 6.

8.3. Limitations

Several limitations bound the claims of this paper. First, the validation is retrospective and modest in scale: six publicly documented cases, selected after their outcomes were known and analyzed by a single author, can show that the framework is consistent with the record but cannot establish that it outperforms expert

judgment, and the analysis is therefore exposed to selection and confirmation bias. Second, the framework remains a conceptual artifact, partly informed by market and vendor evidence, that has not been tested prospectively and is likely to require sector-specific adaptation before use in any particular regulated domain. Third, the evidence base draws substantially on industry and vendor-produced sources whose limitations are noted at each point of use; although triangulated against peer-reviewed and standards literature, they reflect a fast-moving field in which figures date quickly. Fourth, the case evidence itself is uneven, the failures are documented through public reporting and, in one instance, a tribunal decision, whereas the success is reported partly through vendor-published material and should be read with that in mind. Fifth, the framework is calibrated to the senior technology executive and to generative and agentic systems, and its transferability to other decision-makers or to predictive, vision, and optimization AI has not been examined.

8.4. Future Work

Three lines of work follow directly. The first is to expand the retrospective analysis into a larger, systematically sampled set of documented deployments, including cases coded independently of the author, so that the framework's consistency with outcomes can be tested rather than illustrated. The second is to apply the framework prospectively in real organizations and to measure the four success indicators defined in Section 6.2 (decision auditability, post-proof-of-concept survival, total-cost-of-ownership forecast accuracy, and lock-in exposure) against a comparison set of decisions made without it. The third is to operationalize the ten-step sequence as a lightweight decision instrument and to develop sector-specific adaptations, beginning with the regulated environments in which the early compliance gates are most consequential. Together these would move the framework from one consistent with the documented record to one validated prospectively and instrumented for practice.

9. Conclusions

The arguments that follow are grounded in the framework developed above and in its illustrative application; they are advanced as a coherent and defensible decision approach whose empirical validation, through the retrospective comparative case analysis and field studies described in Sections 6 and 8, remains the subject of ongoing work.

The selection of AI systems in organizations is now a central strategic, operational, and governance decision. As the AI marketplace continues to expand and the capabilities of large language models become more accessible, organizations face increasing pressure to choose systems that align with business goals, risk tolerance, data requirements, and long-term operating models. In that environment, informed AI selection is not optional. It is essential to responsible adoption and sustainable value creation.

This paper argued that effective AI selection begins with a clearer understanding of how current AI systems work. Foundational technical literacy gives leaders and technology professionals the ability to evaluate claims more critically, ask better questions during procurement and design, set realistic expectations, and establish governance safeguards that reflect the actual behavior and limitations of these systems. Without that foundation, organizations are more likely to make decisions based on marketing narratives, incomplete comparisons, or short-term enthusiasm rather than disciplined evaluation.

The paper also presented a structured framework for AI selection that emphasizes capability assessment, alignment and safety, integration fit, data governance, total cost of ownership, vendor risk and lock-in, governance and compliance, and lifecycle management. Supported by the NIST AI Risk Management Framework and ISO/IEC 42001:2023, this approach helps organizations move beyond informal selection practices toward a more rigorous and repeatable decision process. The build, configure, or buy framework extends that discipline by giving decision makers a practical method for evaluating when AI should be internally developed, externally acquired, or strategically configured from existing platforms. Together, these frameworks support decisions that are more defensible, scalable, and aligned with enterprise reality.

Equally important, this paper has highlighted that organizations must address AI proliferation with deliberate governance. The expansion of shadow AI, uncontrolled data exposure, and fragmented tool adoption shows that AI selection cannot be separated from policy, architecture, technical enforcement, and workforce education. Organizations that act proactively in these areas are likely to be better positioned to realize the benefits of AI while limiting operational, ethical, and regulatory risk.

For leaders, the value of this paper is that it offers a clearer basis for making AI decisions that align innovation with governance, investment discipline, and organizational purpose. For technology professionals, it provides a structured way to assess AI systems in terms of architecture, integration, lifecycle implications, and operational fit. The broader contribution is a decision-oriented foundation for selecting AI in a way that is not only innovative, but also responsible, practical, and sustainable in the modern enterprise.

Conflicts of Interest

The author is employed as a Senior Managing Consultant for executive engineering and architecture at Crowe LLP and is the founder of Aletheon Labs, which develops enterprise AI and analytical-intelligence software in a domain related to the subject of this paper. This research was conducted independently of both organizations, and the framework presented is vendor-neutral and is not derived from, tied to, or intended to promote any specific product or service. The author declares no other conflicts of interest regarding the publication of this paper.

References

[1] Chui, M., Hall, B., Mayhew, H., Singla, A. and Sukharevsky, A. (2022) The State of AI in 2022, and a Half Decade in Review. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review

[2] Singla, A., Sukharevsky, A., Hall, B., Yee, L. and Chui, M. (2025) The State of AI in 2025: Agents, Innovation, and Transformation. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[3] Brynjolfsson, E. (2022) The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence. *Daedalus*, **151**, 272-287. https://doi.org/10.1162/daed_a_01915

[4] Soulami, M., Benchekroun, S. and Galiulina, A. (2024) Exploring How AI Adoption in the Workplace Affects Employees: A Bibliometric and Systematic Review. *Frontiers in Artificial Intelligence*, **7**, Article 1473872. https://doi.org/10.3389/frai.2024.1473872

[5] Watermann, L., Kubowitsch, S. and Lermer, E. (2025) AI and Work Design: A Positive Psychology Approach to Employee Well-Being. Gruppe. Interaktion. Organisation. *Zeitschrift für Angewandte Organisationspsychologie*, **56**, 311-320. https://doi.org/10.1007/s11612-025-00806-3

[6] Mienye, I.D., Jere, N., Obaido, G., Ogunruku, O.O., Esenogho, E. and Modisane, C. (2025) Large Language Models: An Overview of Foundational Architectures, Recent Trends, and a New Taxonomy. *Discover Applied Sciences*, **7**, Article No. 1027. https://doi.org/10.1007/s42452-025-07668-w

[7] IBM (2024) What Are Large Language Models (LLMs)? IBM Think. https://www.ibm.com/think/topics/large-language-models

[8] Zaza, Z. and Souissi, O. (2025) Architectural and Methodological Advancements in Large Language Models. *Engineering Proceedings*, **97**, Article 8. https://doi.org/10.3390/engproc2025097008

[9] Lu, W., Luu, R.K. and Buehler, M.J. (2025) Fine-Tuning Large Language Models for Domain Adaptation: Exploration of Training Strategies, Scaling, Model Merging and Synergistic Capabilities. *npj Computational Materials*, **11**, Article No. 84. https://doi.org/10.1038/s41524-025-01564-y

[10] Jeong, C. (2024) Fine-Tuning and Utilization Methods of Domain-Specific LLMs. https://arxiv.org/abs/2401.02981

[11] National Institute of Standards and Technology (2023) AI Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce. https://www.nist.gov/itl/ai-risk-management-framework

[12] Maslej, N., Fattorini, L., Perrault, R., Gil, Y., *et al.* (2025) Artificial Intelligence Index Report 2025. Stanford Institute for Human-Centered Artificial Intelligence (HAI). https://hai.stanford.edu/ai-index/2025-ai-index-report

[13] Wharton AI Program (2025) Gen AI Fast-Tracks into the Enterprise: Year Three Full Report. The Wharton School, University of Pennsylvania. https://ai.wharton.upenn.edu/wp-content/uploads/2025/10/2025-Wharton-GBK-AI-Adoption-Report_Full-Report.pdf

[14] International Association of Privacy Professionals (IAPP) (2024) AI Governance in Practice Report 2024. IAPP & FTI Technology. https://iapp.org/resources/article/ai-governance-in-practice-report

[15] Knostic (2025) The 20 Biggest AI Governance Statistics and Trends of 2025. Knostic AI. https://www.knostic.ai/blog/ai-governance-statistics

[16] Bisht, A., Yee, L. and Roberts, R. (2025) Open Source Technology in the Age of AI. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/open-source-technology-in-the-age-of-ai

[17] Gartner (2025) Hype Cycle for Artificial Intelligence, 2025. Gartner Inc. https://www.gartner.com/en/articles/hype-cycle-for-artificial-intelligence

[18] Gartner (2024) Gartner Predicts 30% of Generative AI Projects Will Be Abandoned after Proof of Concept by End of 2025. Gartner Newsroom. https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025

[19] European Parliament (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council, Artificial Intelligence Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

[20] International Organization for Standardization (2023) ISO/IEC 42001:2023, Artificial intelligence, Management System. https://www.iso.org/standard/81230.html

[21] IBM (2024) Guide for Implementing an AI Governance Framework. IBM Institute for Business Value. https://www.ibm.com/think/insights/ai-governance-implementation

[22] World Economic Forum (2025) Advancing Responsible AI Innovation: A Playbook. WEF Insight Report. https://www.weforum.org/publications/advancing-responsible-ai-innovation-a-playbook/

[23] Cyberhaven (2024) Q2 2024 AI Adoption and Risk Report. Cyberhaven Labs. https://info.cyberhaven.com/hubfs/Content%20PDF/Cyberhaven%20Q2%202024%20AI%20Adoption%20and%20Risk%20Report%20052024.pdf

[24] Reco.ai (2025) 2025 State of Shadow AI Report. Reco. https://www.reco.ai/state-of-shadow-ai-report

[25] Hevner, A.R., March, S.T., Park, J. and Ram, S. (2004) Design Science in Information Systems Research. *MIS Quarterly*, **28**, 75-106. https://doi.org/10.2307/25148625

[26] Peffers, K., Tuunanen, T., Rothenberger, M.A. and Chatterjee, S. (2007) A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, **24**, 45-77. https://doi.org/10.2753/mis0742-1222240302

[27] Moffatt, V. (2024) Air Canada, 2024 BCCRT 149 (CanLII). (2024). British Columbia Civil Resolution Tribunal.

[28] Restaurant Dive (2024) McDonald's Ends IBM Drive-Thru Voice Order Test. https://www.restaurantdive.com/news/mcdonalds-ibm-drive-thru-automation-voice-ordering-ai/719085/

[29] OpenAI (2025) Morgan Stanley Uses AI Evals to Shape the Future of Financial Services. OpenAI. https://openai.com/index/morgan-stanley/

[30] The University of Texas System Audit Office (2016) Special Review of Procurement Procedures Related to the M.D. Anderson Cancer Center Oncology Expert Advisor project. The University of Texas System.

[31] Lecher, C. (2024) NYC's AI Chatbot Tells Businesses to Break the Law. The Markup.

https://themarkup.org/artificial-intelligence/2024/03/29/nycs-ai-chatbot-tells-businesses-to-break-the-law

[32] Klarna (2024) Klarna AI Assistant Handles Two-Thirds of Customer Service Chats in Its First Month. Klarna. https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/

[33] Shibu, S. (2025) Klarna Is Hiring Customer Service Agents after AI Couldn't Cut It on Calls, According to the Company's CEO. Entrepreneur. https://www.entrepreneur.com/business-news/klarna-ceo-reverses-course-by-hiring-more-humans-not-ai/491396

Appendix: AI Selection Scorecard

This appendix operationalizes the framework as a one-page scoring instrument so that a selection decision can be recorded, audited, and repeated. Items marked Gate are non-negotiable and are recorded as pass or fail; a selection should not proceed while any gate is failed. The remaining dimensions are scored from 1 (weak) to 5 (strong), and the sourcing recommendation is recorded at the foot of the scorecard. The instrument is intended for prospective use and refinement, consistent with the validation agenda in Section 8.4. Table A1 presents the scorecard.

Table A1. AI selection scorecard.

#	Dimension or gate	Assessment question	Score (1 to 5) or gate
1	Capability on own data	Does the model meet accuracy and latency targets on representative organizational data at the expected concurrency?	___/5
2	Alignment and safety	Are outputs verified, red-teamed, and human-reviewed in proportion to the risk class of the use?	___/5
3	Data governance	Are data residency, privacy, and training-exclusion requirements met?	Gate: pass/fail
4	Governance and compliance	Is there documented accountability, auditability, and regulatory fit for this use?	Gate: pass/fail
5	Integration fit	Does the system integrate with the systems of record and existing workflows?	___/5
6	Sourcing economics	Is the build, configure, or buy choice justified by differentiation, data, and capability?	___/5
7	Total cost of ownership	Is the three-year TCO base case defensible against buy or RAG alternatives (extendable to five years)?	___/5
8	Vendor risk and lock-in	Are data-portability rights and a viable fallback option contractually present?	___/5
9	Lifecycle management	Is there a monitoring, retraining, evaluation, and human-oversight plan?	___/5
10	Proliferation control	Is the system sanctioned and tracked, reducing shadow-AI exposure?	___/5
	Recommendation	Build, configure, or buy, with rationale traceable to the rows above	_____