cognizant®

The business of agentic AI: Costs, commercial models and the path forward

Contents

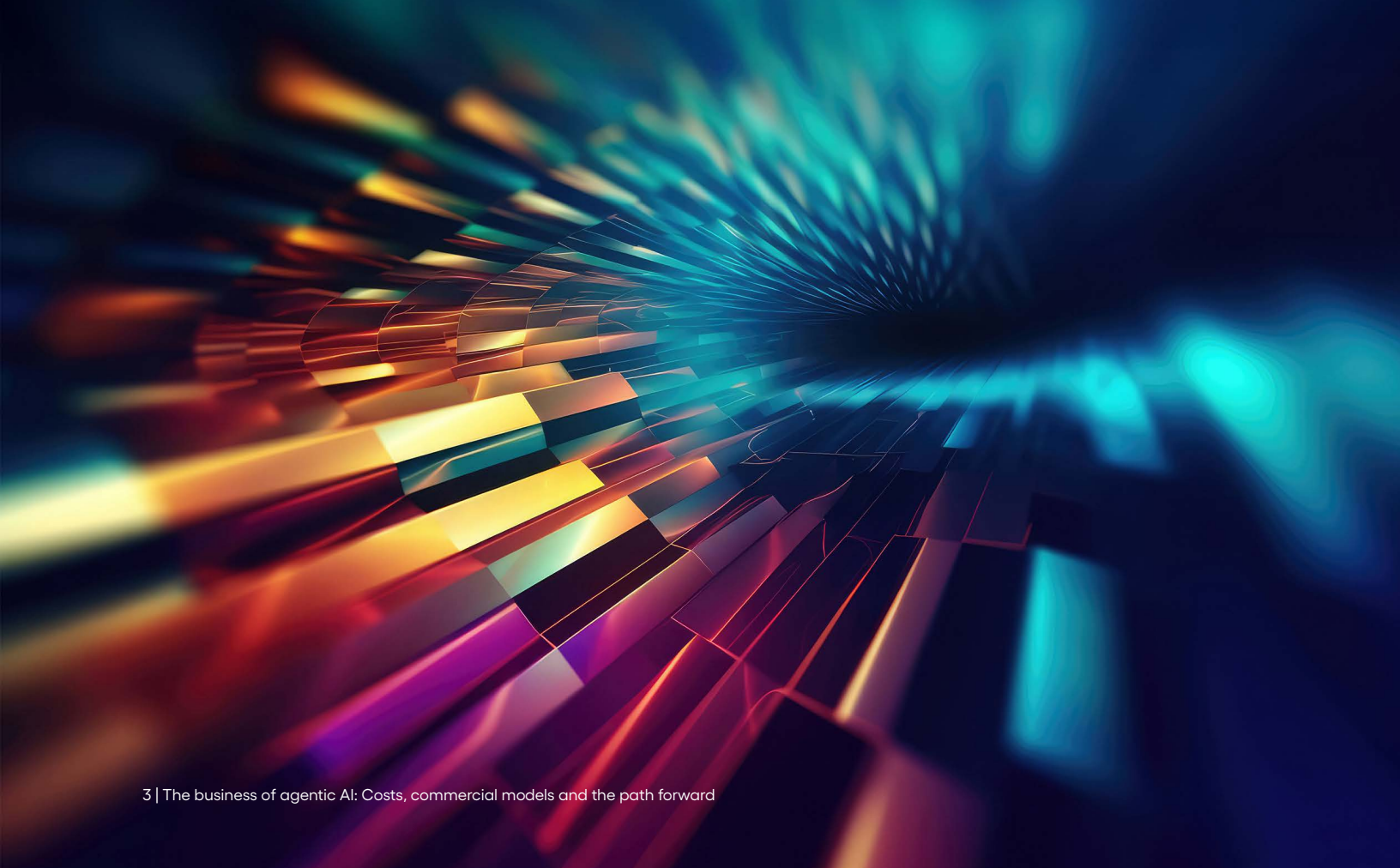
Executive summary

Enterprises today face a growing execution gap in agentic AI, driven by the disconnect between successful experimentation and the ability to operationalize intelligent agents reliably at scale. While pilots and proofs of concept continue to multiply, many organizations struggle with hidden costs, architectural fragility, governance exposure and uncertainty around workforce and commercial implications as agentic systems move closer to the core of operations.

The first one of this three part white paper series, "Confronting the AI velocity gap: A new architecture for enterprise operations", examined how enterprises can respond by rearchitecting operations around an AI led, human enabled operating model anchored in straight through processing (STP), progressive autonomy, modular design and robust governance. The second part, "From architecture to intelligence: Building the agentic technology stack", explored how intelligence itself must be structured as an architectural layer, introducing agentic systems, agent networks and ontology driven grounding as the foundation for scalable and trustworthy AI driven operations.

This third part of the white paper series, "The business of agentic AI: Costs, commercial models and the path forward", shifts the focus from architecture to execution. It addresses what it truly takes to design, deploy, govern and scale agentic systems as they transition from innovation environments into production, particularly within large scale, process intensive domains such as business process outsourcing (BPO). Rather than extending theory, it confronts the practical realities leaders face when autonomous systems begin to act within live enterprise workflows.

Through a pragmatic lens, this paper examines the often overlooked total cost of ownership of agentic AI and introduces a disciplined framework for building a credible business case. It explores how agentic systems reshape workforce models, redefine commercial constructs and force deliberate choices around ownership, risk, speed and partnership strategies. The paper concludes by positioning agentic AI not as an add on to existing operations, but as a catalyst for rearchitecting how work is executed, balancing autonomy with control to enable intelligent operations that scale with confidence and accountability.



From architectural intent to disciplined enterprise execution

The central premise of this white paper is that realizing value from agentic AI requires disciplined execution, not just architectural intent. As intelligent agents move from experimental deployments into live, mission critical operations, enterprises encounter challenges that extend well beyond model performance, spanning cost structures, governance, workforce impact and commercial viability. Without addressing these dimensions explicitly, the promise of agentic AI remains difficult to sustain at scale.

Building on the architectural foundations established in the first two parts of this white paper series, this third installment focuses on operationalizing agentic systems in real-world environments. It examines how enterprises can transition from pilot-driven adoption to production-grade deployment.

Rather than prescribing a single operating model, the paper explores the strategic choices leaders must make as autonomy increases across ownership models, risk allocation, workforce redesign and partner ecosystems. It also introduces a pragmatic lens on total cost of ownership, helping organizations move beyond surface-level ROI assumptions to construct business cases that withstand scale, time and operational complexity.

Finally, this paper provides a grounded perspective on what it takes to embed agentic AI responsibly into enterprise operations. It is intended to help leaders move from intent to impact, executing with clarity, governing with confidence and shaping operating models that balance autonomy with control as intelligent systems become integral to how work is delivered.

The true cost of agentic AI: What nobody is talking about

While the business case framework provides a high-level view, it is essential for leaders to have a granular, clear understanding of the cost components of an agentic solution. The sticker price of an LLM API call is a tiny fraction of the total cost of ownership. A truly comprehensive cost model must account for the full lifecycle of building and operating a production-grade agentic system.

A structural cost breakdown

Let us break down the cost structure into its core components, focusing on the relative magnitude and key drivers of each.

- Upfront build and implementation costs (large and one-time): This is the initial investment required to get the solution live. It includes professional services for process analysis, ontology development, agent decomposition, integration and front-end workflow creation. This is typically the largest single cost item and is often underestimated. It includes process discovery and redesign, ontology and knowledge graph construction, agent development and training, integration with legacy systems, front-end workflow development, testing and quality assurance, and change management and training.

Recurring technology costs (significant and ongoing): These are the ongoing costs of the technology stack:

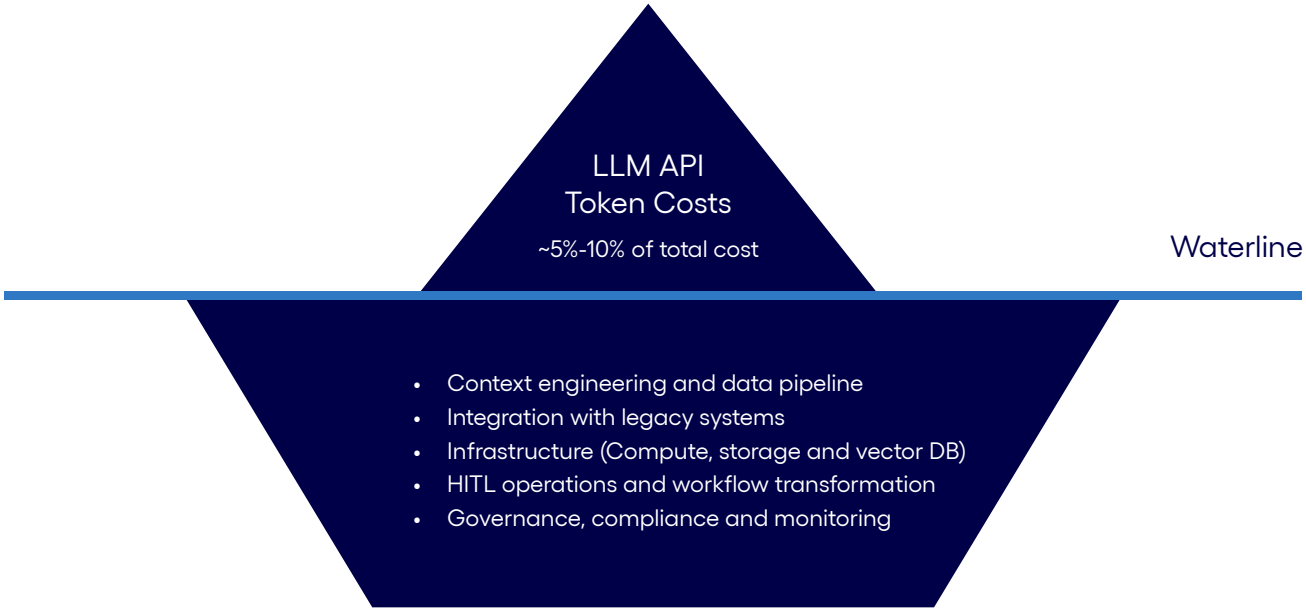
- LLM consumption: This includes costs starting from API calls to the underlying language models. This is driven by the number of transactions, the complexity of the reasoning (number of tokens per transaction), and the choice of model. Costs can be optimized by using smaller, faster models for routine tasks and reserving larger, more expensive models for complex reasoning.
- Infrastructure hosting: It accounts for the costs of cloud computing, storage and networking. This scales with transaction volume and the complexity of the agent ecosystem. Key cost drivers include the vector database, the knowledge graph, the orchestration platform and the real-time event bus.
- Software licensing: This includes recurring fees for all the commercial software components in the stack: databases, orchestration platforms, workflow tools, document intelligence services and more.

Recurring operational costs (variable and ongoing): These are the ongoing costs of running and maintaining the solution:

- Human-in-the-loop (HITL) operations: This covers the cost of human experts for handling low-confidence transactions. This cost is driven by two factors: the straight-through processing (STP) rate, which determines the volume of transactions requiring human review, and the average time required to resolve each HITL transaction. This cost should decrease over time as the system learns and the STP rate improves.
- Platform and model maintenance: This involves the cost of the engineering team responsible for monitoring system performance, managing model drift, retraining models and patching the technology stack.
- Human-over-the-loop (HOTL) governance and compliance: The cost of the team and tools required to continuously audit the system for compliance, fairness and risk. Unlike HITL operations, which involve active participation in individual transactions, HOTL governance is supervisory—encompassing policy setting, threshold calibration, exception pattern analysis and regulatory reporting. This cost remains relatively stable as the system scales, as it is driven by the complexity of the governance framework rather than transaction volume.

The visible cost of agentic AI is just the tip of the iceberg. The LLM token cost is only a small fraction of the total cost. A large part of the cost is below the waterline. They lie in the foundational build, supporting technology stack and the ongoing operational and governance overhead. Leaders who build their business cases based only on the visible cost run the risk of making misinformed decisions.

The true cost of Agentic AI: The iceberg model



Cost iceberg

Business case: A pragmatic framework for agentic AI

One of the most challenging aspects of agentic AI is building a compelling business case. The costs are real, immediate and quantifiable. The benefits are often strategic, long-term and harder to measure. This chapter provides a pragmatic framework for building a business case that is both honest and compelling.

The cost-value equation

A business case for agentic AI is not a simple cost-reduction calculation. It is a multidimensional equation that must account for both quantitative and qualitative factors:

Factor	Description
Direct cost savings	Reduction in labor costs from automating manual, repetitive tasks
Efficiency gains	Faster processing times, reduced error rates and lower rework costs
Quality improvement	More consistent, accurate and compliant outcomes
Scalability	Ability to handle volume increases without proportional cost increases
Reusability and platform value	Value of creating a reusable agentic platform that can be leveraged across the enterprise
Competitive differentiation	Ability to offer faster, more accurate and more personalized services than competitors
Risk reduction	Reduced operational risk from human error, improved compliance and better audit trails
Speed to market	Faster deployment of new capabilities and services

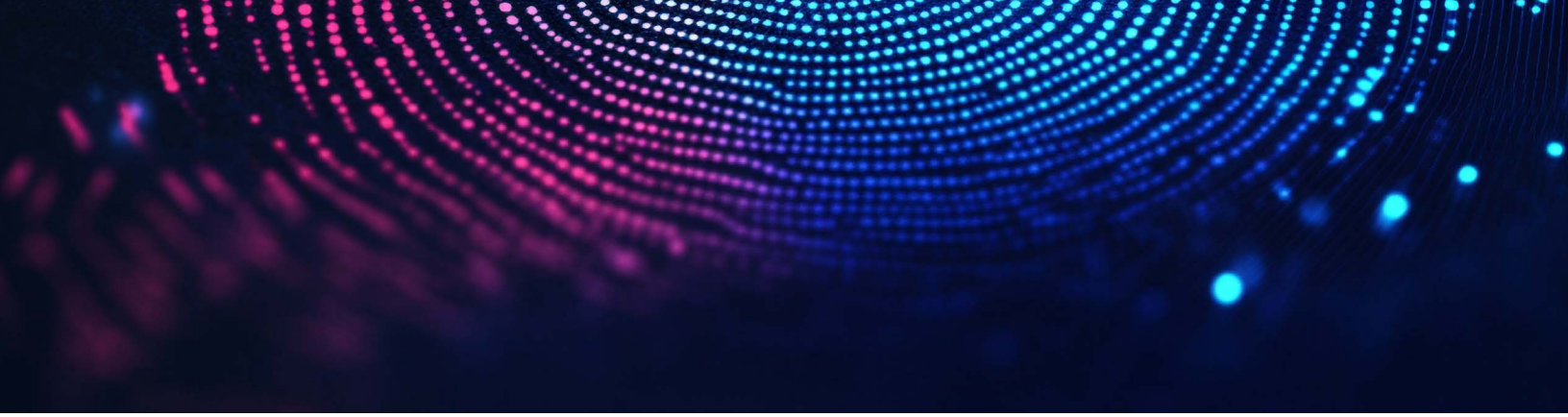
Building a business case

Instead of focusing on a single, speculative ROI number, a more effective approach is to build a scenario-based model that shows how the business case evolves as the solution matures. The model should clearly articulate the assumptions behind both the cost and value estimates and should include both quantitative and qualitative factors.

It is entirely possible that for some use cases, a purely financial ROI may not be compelling in the short term, especially when accounting for the full, upfront platform investment. However, the business case may still be overwhelmingly positive when strategic factors such as scalability, reusability and competitive positioning are considered. The goal is not to produce a magic number, but to provide a transparent, defensible framework for making a strategic investment decision.

The compounding effect

One of the most powerful aspects of the agentic business case is the compounding effect. Unlike traditional automation, which delivers a one-time efficiency gain, an agentic system is designed to learn and improve over time. The STP rate increases, the HITL cost decreases, and the quality of the output improves—all without additional capital investment. This means that the ROI of an agentic solution is not static. It accelerates over time. The business case in the third year will be dramatically better than the business case in the first year, because the system gets two additional years of learning and optimization.



The future BPO workforce: Beyond exception processing

The rise of agentic AI does not mean the end of human involvement in BPO. It means a fundamental transformation of the roles that humans play. The repetitive, rules-based work that has been the bedrock of the industry for decades will be increasingly automated. In its place, a new set of more sophisticated, more strategic roles will emerge. These are not the low-skilled exception handlers of the past; they are the highly skilled governors and trainers of the new agentic workforce.

The AI-led, human-enabled operating model

- The transition from a human-led, AI-enabled model to an AI-led, human-enabled model is not just a technology shift, it is a fundamental reimagining of the operating model. In the AI-led model:
- AI agents are the primary actors, handling the majority of transactions autonomously
- Humans provide strategic oversight, governance and continuous improvement
- Human intervention is a value-adding activity (training AI, handling genuinely novel situations), not a routine processing step
- The focus shifts from throughput ("how many transactions can a human process?") to outcomes ("how effectively is the system achieving its business objectives?")

Six sophisticated roles in the agentic era

The rise of agentic AI does not mean the end of human involvement in BPO. It means a fundamental transformation of the roles that humans play. The repetitive, rules-based work that has been the bedrock of the industry for decades will be increasingly automated. In its place, a new set of more sophisticated, more strategic roles will emerge. These are not the low-skilled exception handlers of the past; they are the highly skilled governors and trainers of the new agentic workforce.

- AI agent orchestrators: They are the conductors of an agentic ecosystem. They design, configure and monitor end-to-end workflows, determining how different vertical and horizontal agents should collaborate to achieve a business outcome. They are responsible for the overall performance of the agent network and continuously optimize its configuration across the client's operational landscape.
- Context engineers: They work as librarians of an agentic enterprise. They are responsible for the entire context engineering lifecycle—finding, storing and searching information that fuels AI agents. They are experts in knowledge representation, data modeling and retrieval technologies. Recognizing this role as critical, Cognizant has already committed to deploying 1,000 context engineers.
- Ontology managers/knowledge curators: They play the role of domain experts who build and maintain enterprise ontology. They work with business stakeholders to define concepts, relationships and rules that govern a business domain and ensure that this knowledge is accurately represented in a machine-readable format. They are the bridge between business knowledge and machine understanding.
- AI performance analysts: Those in this role work as data scientists in the new BPO environment. They continuously monitor performances of AI agents, analyze root causes of errors and low-confidence predictions, and identify opportunities to improve systems through retraining, fine-tuning, or by making changes to their agent architecture.

- AI trainers: These are the human experts who handle low-confidence transactions routed by AI agents. Their role extends beyond processing exceptions: AI trainers provide high-quality, structured feedback that is used to improve AI performance through reinforcement learning from human feedback (RLHF), a technique in which human evaluations of agent outputs are used to iteratively refine the model's decision-making. In this sense, AI trainers serve as the teachers of the digital workforce. They are supported by cognitive assistants which are AI-powered tools that present the agent's proposed action, reasoning and confidence score, enabling AI trainers to review, correct and approve decisions efficiently.
- AI ethics and governance specialists: They are the guardians of responsible AI. They ensure that the agentic ecosystem is fair, transparent and compliant with all relevant regulations and ethical guidelines. They design and implement the governance frameworks that prevent bias, protect privacy and ensure that AI is aligned with human values and client obligations.

The interdisciplinary imperative

A common thread runs through all these new roles that are fundamentally interdisciplinary. They require a blend of deep domain expertise, strong technical literacy and critical thinking skills. The most valuable professionals in the agentic era will be those who can bridge the gap between business and technology and someone who can translate the nuances of a business process into the formal structures that an AI can understand.

This has profound implications for talent development. BPO providers can no longer compete simply by hiring and training for rote execution. They must become engines of talent transformation, reskilling their existing workforce and creating new career paths that lead to these more sophisticated, higher-value roles. As Ganesh Ayyar, President – Asia Pacific & Japan and Industry Solutions Group, Cognizant, observed in his Fortune interview on organizational transformation and AI adoption, this transformation requires a culture of experimentation and a willingness to embrace change at every level of the organization.

A new commercial landscape with evolving models for a new era

The transformation of BPO operations inevitably transforms the commercial models that govern the relationship between clients and service providers. The traditional pricing models: built around the cost of human labor are fundamentally misaligned with an operating model where AI agents perform a growing share of the work. A new commercial framework is emerging, one that aligns cost with value and creates a true partnership between clients and providers.

From inputs to outcomes

The fundamental shift is from pricing based on inputs (the number of people working on an account) to pricing based on outcomes (the results delivered by the process). This evolution will not happen overnight, but the direction of travel is clear.

Model type	Description	Best suited for
Hybrid pricing	A combination of a fixed fee for the AI platform and a per-transaction fee that varies based on the level of automation	Early-stage agentic deployments where the STP rate is still evolving
AI agent capacity	Analogous to the full-time equivalent (FTE) model, the client pays a recurring fee for a certain capacity of digital workers or AI agents	Steady-state operations with predictable volumes
Outcome-based pricing	The provider is compensated based on the business outcomes they deliver: accuracy rates, processing times, customer satisfaction scores, etc.	Mature agentic deployments with well-established performance baselines
Value/gain-sharing	A more advanced form of outcome-based pricing where the client and the provider agree to share the economic benefits of the transformation	Strategic partnerships with high trust and aligned incentives
Subscription/ Platform as a Service (PaaS)	As providers build out reusable, horizontal agentic capabilities, they can offer them to clients on a subscription basis	Horizontal capabilities (document processing, compliance checking) that are applicable across multiple clients

The path forward

The transition to these new commercial models will be gradual and will require a significant increase in trust and transparency between clients and providers. It requires sophisticated measurement and reporting capabilities to track outcomes accurately, and it requires a willingness on both sides to move away from the comfortable, but outdated logic of the full-time equivalent (FTE) model.

The most forward-thinking BPO service providers are already having these conversations with their clients, codesigning new commercial frameworks that reflect the shared journey of transformation. They understand that in the agentic era, the relationship is no longer about selling labor, it is about selling outcomes.

The strategic choice: How to execute your agentic transformation

Embarking on the journey to an agentic enterprise is not just a technical decision, it is a fundamental strategic choice about how to balance control, speed, risk and financial investment. There is no single right answer. The optimal path depends on an organization's culture, risk appetite, in-house capabilities and the nature of the processes being transformed. Based on our experience with large-scale enterprise transformations, we see five primary execution models, each with distinct pros, cons and strategic implications.

Strategic option	Description	Pros	Cons	Verdict
In-house	Build and manage the entire agentification journey internally	Full strategic control and IP retention; deep cultural integration	High upfront CAPEX; slower time-to-value; require significant in-house AI talent	Selectively recommended; best for core IP-sensitive functions
Advisory only	Leverage an external partner for strategy and architecture, but retain full control over execution	Access to best practices; flexible engagement model	Limited partner accountability for outcomes; client bears all execution risks	Viable, but not optimal
Outsource-led	The partner drives the agentification initiative on the outsourced scope, using savings to self-fund the transformation	Self-funding initiative; rapid fail-fast innovation; accelerated cycle time	Shared strategic control; require significant change management and trust	Highly recommended; an ideal engine for transforming high-volume, non-differentiating operations
Full hybrid	The partner manages the transformation for both outsourced and retained teams	Holistic transformation; maximized synergies across the enterprise	High governance complexity; significant coordination overhead	Future state; the end goal for a deeply integrated partnership
Strategic portfolio	A combined approach that leverages the best of the in-house and outsource-led models	Best of both worlds: speed plus control; self-funding innovation; diversified risk profile	Require robust dual governance; need careful resource balancing	The strategic choice; the optimal balance for most enterprises

The recommended path: A strategic hybrid portfolio

For most large enterprises, the most practical, highest-ROI and most accelerated path is the strategic portfolio approach. This hybrid model provides a pragmatic balance between the need for speed and the need for control. It works as follows:

- Outsource to innovate. A majority of high-volume, transactional operations are outsourced to a strategic partner. The partner takes on the responsibility for driving the agentification of these processes, using the cost savings from initial automation to fund the investment in the more advanced agentic capabilities, which are required to reach 80%–90% STP. This creates a self-funding engine for innovation and shifts a significant portion of the technology and talent investment risk to the service provider.
- Retain to differentiate. Simultaneously, the enterprise retains its most strategic, sensitive and differentiating functions in-house. It can then apply the learnings, best practices and even some of the reusable platform components from the outsourced transformation to its own internal projects, but without the pressure of having to build everything from scratch.

This model allows an enterprise to fail fast and innovate rapidly on scalable processes with a trusted partner, while maintaining full strategic control over its core differentiators. It is the fastest, most capital-efficient and most risk-mitigated path to becoming an agentic enterprise.

Making the decision: A framework for evaluation

When evaluating which execution model is right for your organization, consider the following dimensions:

Evaluation dimension	Key questions
Strategic sensitivity	Is this a core differentiating capability or a high-volume operational function?
In-house capability	Do you have AI engineering talent, context engineering expertise and platform infrastructure to build this in-house?
Speed imperative	How quickly do you need to realize value? Can you afford a multiyear internal build?
Risk appetite	Are you willing to share control with a partner in exchange for faster time-to-value and shared risk?
Financial model	Do you prefer high upfront CAPEX (in-house) or a self-funding OPEX model (outsource-led)?
Scale requirements	Is this a single process or a portfolio of processes that would benefit from a shared platform?

Navigating the transformation

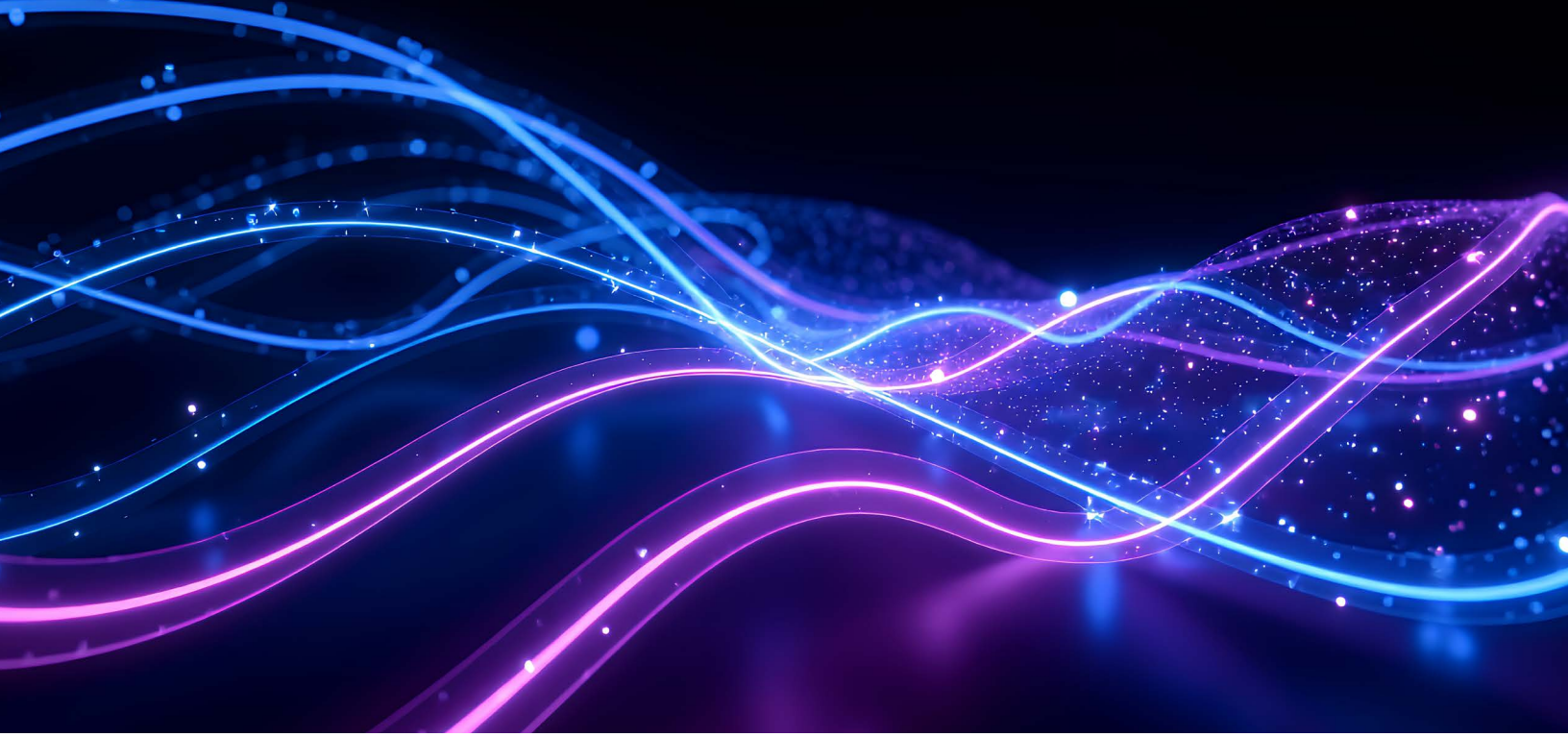
The BPO industry stands at a genuine inflection point. The convergence of LLMs, agentic AI frameworks and sophisticated engineering methodologies has created the conditions for a transformation that is more profound than anything the industry has experienced since the offshoring revolution of the early 2000s.

But transformation is not destiny. The organizations that will thrive in this new era are not the ones that adopt AI the fastest. They are the ones that adopt it the most thoughtfully with a clear-eyed understanding of both the opportunities and the challenges, a disciplined approach to engineering and governance and a genuine commitment to building the foundational capabilities that make reliable, scalable AI possible.

A successful roadmap is not about hitting arbitrary technology milestones; it is about progressively balancing autonomy and oversight, evolving from HITL operations in early stages to HOTL governance as confidence and system maturity increase. The three phases below map directly to the four-stage maturity model introduced in the first part of this series, translating those operating postures into a practical implementation timeline.

- **1–6 months (foundational build and supervised deployment):** Aligning with cognitive assistance (0%–20% STP) of the maturity model in stage 1. The focus is on building the core platform (ontology, context engineering and agent architecture) and deploying the first set of agents in a highly supervised, cognitive assistance mode. The goal is to establish a baseline, prove the technology and begin capturing the training data needed for the next phase. Key deliverables include the enterprise ontology for the target domain, the context engineering pipeline, the first set of vertical and horizontal agents and the HITL workflow and cognitive assistant tools. A good place to start would be using AI agents to build the standard operating procedures and training content. Then move into building QC and QA agents and finally, using this enhanced learning to build the vertical and horizontal agents that would perform the business function with high degree of accuracy and help with progressive autonomy of the agents.
- **7–18 months (gradual autonomy and confidence building):** Aligning with supervised autonomy (20%–60% STP) of the maturity model in stage 2. Based on the performance in the first phase, the system is gradually given more autonomy. The STP rate is increased for low-risk, high-volume transaction types. The key milestone is not a specific STP percentage, but the demonstrated ability of the system to reliably identify and route low-confidence transactions to humans. Key deliverables include calibrated confidence thresholds, expanded agent coverage, the RLHF feedback loop and performance dashboards. During this phase, the governance model begins transitioning from predominantly HITL (where humans review individual transactions) toward selective HOTL oversight (where humans monitor aggregate performance patterns and intervene at the policy level rather than the transaction level).
- **19–24+ months (governed scaling and outcome optimization):** Aligning with governed autonomy through near-full autonomy (60%–90%+ STP) of the maturity model in stages 3 and 4. With confidence established, the focus shifts to scaling the solution across more complex processes and optimizing for business outcomes. The milestones are now business-centric, such as reduction in cost per transaction, improvement in customer satisfaction and the successful retirement of legacy technical debt. Key deliverables include scaled agent deployment, outcome-based performance metrics, new commercial model implementation and workforce transformation programs. At this maturity level, the operating model is predominantly HOTL, where humans provide supervisory governance, set policies and thresholds, monitor dashboards and handle escalations, while AI agents operate autonomously within defined parameters.

This risk-based, phased approach ensures that the organization can move at a speed that matches its comfort level and the demonstrated performance of the technology. It is the key to achieving a pragmatic balance between efficiency and security.



What next?

The future of BPO is neither purely human nor purely AI. It is a carefully orchestrated partnership between human intelligence and artificial intelligence, evolving progressively from human-led to AI-led, governed by robust frameworks and measured by real business outcomes. The organizations that master this partnership, those that become true agentic enterprises, will define the next era of the industry.

The key principles that should guide every organization on this journey are:

- Start with STP as the north star. Design every process for STP, not for exception handling alone (refer part 1 of this whitepaper series, "Confronting the AI velocity gap: A new architecture for enterprise operations", for details).
- Build the foundation first. Invest in ontology, the context engineering pipeline and the agent architecture before chasing use cases. (Refer part 2 of this whitepaper series, "From architecture to intelligence: Building the agentic technology stack", for details).
- Earn trust through transparency. Use confidence-based routing, full explainability and progressive autonomy to build trust with stakeholders. applying HITL governance in early stages and transitioning to HOTL oversight as system maturity and stakeholder confidence increase.
- Invest in people. The transformation of the workforce is as important as the transformation of technology. Build new roles, create new career paths and invest in training that will make the transition successful.
- Align incentives. Adopt commercial models that align with the interests of clients and providers around shared outcomes, not labor inputs.
- Adopt a portfolio strategy. Rather than applying a single execution model across all processes, use the strategic portfolio approach: outsource high-volume, transactional operations to a trusted partner to accelerate innovation, while retaining strategic, differentiating functions in-house to maintain competitive control.

The journey begins with a single step, an honest assessment of where you are today, a clear vision of where you need to be and the courage to start building.

Related reading

- "Confronting the AI velocity gap: A new architecture for enterprise operations"
- "From architecture to intelligence: Building the agentic technology stack"

References

Ganesh Ayyar, President – Asia Pacific & Japan and Industry Solutions Group, Cognizant. Perspectives on organizational transformation, experimentation culture, and the role of positive fear in driving AI adoption. Referenced in Fortune, "Cognizant Executive on AI, Company Culture Change, Hierarchy, Failure, and Fear," June 2025. Available at: https://fortune.com/2025/06/12/cognizant-executive-ai-company-culture-change-hierarchy-failure-fear/

Author

Anoop Nair
Senior Vice President, Global Head of FSI - IOA
anoop.nair@cognizant.com

Follow

cognizant

Cognizant (Nasdaq-100: CTSH) engineers modern businesses. We help our clients modernize technology, reimagine processes and transform experiences so they can stay ahead in our fast-changing world. Together, we're improving everyday life. See how at www.cognizant.com or follow us **@Cognizant.**

World Headquarters
300 Frank W Burr Blvd
Suite 36, 6th Floor
Teaneck, NJ 07666, USA
Tel: (201) 801-233

European Headquarters
280 Bishopsgate
London
EC2M 4AG
England
Tel: +44 (01) 020 7297 7600

India Corporate Office
Siruseri-Software Technology Park of India (STPI)
SDB Block – Ground floor north wing
Plot No H4, SIPCOT IT Park
Chengalpattu District
Chennai 603103, Tamil Nadu
Tel: 1800 208 6999

APAC Headquarters
1 Fusionopolis Link, Level 5
NEXUS@One-North, North Tower,
Singapore 138542
Tel: + 65 6812 4000

WF 4486200