

Agentic AI for Bharat's BFSI

A Practitioner's View

TECHNOLOGY
PEOPLE
TRUST
FOR A STRONGER BHARAT



FINANCIAL
INCLUSION
POWERED BY
INTELLIGENT
POSSIBILITIES

Foreword

India's financial services sector is entering its next phase of evolution. Digital adoption has expanded access, distribution has become more widespread, and financial awareness continues to steadily improve. Yet, for many customers, the challenge is no longer access to financial products and services; it is understanding what is right for them, making informed choices, and using financial solutions with confidence over time.

This makes the need for simplifying finance for Bharat an important priority which will ensure financial products, services and advisories are easier to understand, access, and use for people across the country. India's scale gives us a significant opportunity to address this. With over 950 million internet subscribers and more than 650 million smartphone users¹, digital platforms have the potential to take financial services deeper into the lives of millions of individuals and businesses.

Building on this digital foundation, artificial intelligence has a significant role to play in reshaping the financial services landscape. It enables institutions to better understand customer needs, deliver more relevant engagement, and provide personalised experiences at scale. For the BFSI industry, Agentic AI represents the next stage of this evolution. By connecting intelligence across customer journeys, it has the potential to transform how customers are educated, advised, acquired, onboarded, serviced, and retained. For financial institutions, Agentic AI brings greater context, consistency, and speed, while keeping the customer at the centre and trust at the foundation of every interaction.

At Aditya Birla Capital, we have embraced AI as a strategic enabler of our vision to simplify finance and deliver superior customer experiences. We have moved beyond experimentation to embed AI across our value chain, spanning origination, underwriting, sales, servicing, engagement, and operations. Through our Generative AI Centre of Excellence, established in 2023, we have scaled high-impact AI use cases that are improving efficiency, deepening customer engagement, and creating sustainable business value.

As AI adoption matures, the focus will increasingly shift from individual use cases to connected intelligence across the enterprise. We believe the future of financial services will be defined by how effectively institutions combine trust, data, technology, and process reimagination to deliver more personalised and seamless experiences at scale. Agentic AI can be a powerful enabler of that future, helping bridge the gap between access and meaningful participation, while creating new possibilities for customers, employees, and institutions alike.

This paper brings together our experiences and key learnings from scaling AI across the enterprise. We hope it offers practical perspectives for organisations navigating their own AI journeys and contributes to the broader dialogue on how AI can be applied responsibly to create sustainable value for customers, businesses, and all stakeholders for the long-term.

¹ International Trade Administration, India Country Commercial Guide, 2026





Contents

ES	Executive Summary	7
01	Advancing Financial Inclusion through Agentic AI	11
02	Trends Defining the Next Phase of Enterprise AI	15
03	Agentic AI in Action: Areas of Deployment at ABCL	23
04	Building the Human Operating Model for AI	39
05	Responsible AI at Scale: Governance, Cost and Observability	47
06	The Foundations Enabling AI at Scale	53
07	The AI Playbook: From Ambition to Execution	57
08	Conclusion	69
AP	Appendix	72

If you have ten minutes

The Executive Summary and the Conclusion

07 | 69

If you run a business line

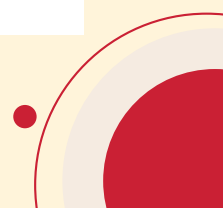
Eight areas and nine use cases, with what worked and what did not.

23

If you are building it

The 90-day, six-month and one-year decision path.

57



Executive Summary

Financial institutions are entering a new phase of AI adoption. The industry has largely moved beyond asking whether AI should be deployed and is increasingly focused on where it can create measurable value and how it can be scaled sustainably.

India has made significant progress in expanding financial access, but meaningful participation remains uneven. Reflecting this shift, the RBI's National Strategy for Financial Inclusion 2025-30² moves the inclusion agenda beyond access towards the effective use of equitable, responsible, suitable and affordable financial services, while emphasising financial resilience, financial education and customer protection. AI can help institutions advance this agenda by making personalised customer enablement, engagement and servicing more accessible while increasing economic viability.

Financial institutions are entering a new phase of AI adoption. The industry has largely moved beyond asking whether AI should be deployed and is increasingly focused on where it can create measurable value and how it can be scaled sustainably. While AI capability continues to advance rapidly, enterprise outcomes remain uneven. Many organisations have successfully launched pilots and individual use cases, but relatively few have translated those initiatives into enterprise-wide transformation. The challenge is no longer access to technology. It is the ability to integrate AI into workflows, connect it to trusted data, redesign ways of working and establish the operating disciplines & governance required for scale.

Aditya Birla Capital has moved beyond isolated pilots to apply agentic AI across the enterprise. Its deployments span customer acquisition and engagement, onboarding and underwriting, servicing and retention,

sales enablement, marketing personalisation, software development, and control functions such as audit, finance and compliance. These initiatives range from orchestrating end-to-end customer journeys and supporting frontline employees to accelerating software delivery and strengthening analytical coverage in control functions. Collectively, these deployments provide a practitioner's view of how agentic AI can create value across all functions when embedded within enterprise workflows.

ABCL's implementation experience shows that sustainable AI value depends less on model sophistication and more on how effectively AI is embedded into the enterprise. The strongest deployments began with a clearly defined business problem, redesigned the end-to-end workflow and established clear boundaries between AI execution and human judgement. They also required trusted data, connected systems and current business context to flow across each stage. Importantly, several solutions needed iteration to refine agent responsibilities, human hand-offs and operating controls as real-world usage revealed where the initial design fell short. The central learning is that agentic AI is not a one-time technology deployment, but a continuously evolving operating capability shaped by workflow design, workforce adoption, governance and measurable business outcomes.

Realising this potential requires workforce transformation to progress alongside technology adoption. Institutions need to redesign roles,

² RBI, National Strategy for Financial Inclusion 2025-30, 2025

performance measures and career pathways around human-AI collaboration, while providing employees with role-specific capabilities and clear decision rights. Managers play a critical role in translating AI into everyday working practices, and sustained adoption depends on employees understanding where AI can assist, where human judgement remains essential and who remains accountable for the outcome.

At the same time, scaling AI safely and sustainably requires responsible AI to operate as an embedded enterprise discipline, not a post-deployment control. As AI enters more consequential workflows, institutions require risk-based human oversight, decision-level traceability, continuous monitoring and clear accountability for outcomes. Security, observability and cost discipline must also be built into the operating model, enabling institutions to manage performance, risk and economics throughout the AI lifecycle.

Together, these lessons form the foundation of the implementation playbook presented in this report. The first 90 days focus on selecting the right business journey, establishing its baseline, assigning ownership and defining economic and risk boundaries. The first six months focus on controlled production deployment, supported by live monitoring, human oversight and clear measures of value. The first year focuses on converting early experience into reusable enterprise capabilities that support further adoption with consistent governance and operational control.

The next phase of AI adoption will be defined not by the number of use cases launched, but by the institutional capability created around them. Financial institutions that begin with clear business outcomes, build the required workforce and governance foundations, and expand with discipline will be better positioned to convert AI experimentation into sustained enterprise value and deeper financial participation.





01

The access gap

Depth over reach

A sustainable path

Advancing Financial Inclusion through Agentic AI

SECTION 1

Advancing Financial Inclusion through Agentic AI

Access has been won. The next financial inclusion frontier is depth: AI can help every customer understand, adopt and meaningfully use the right financial products.

India's financial inclusion agenda is entering its next and more consequential phase. With account ownership reaching close to 89% of adults³ and the RBI's Financial Inclusion Index improving from 64.2 to 67⁴, the country has made significant progress in expanding access to the formal financial system. But access alone is no longer an adequate measure of inclusion. For example, out of ~56 crore accounts opened under Pradhan Mantri Jan Dhan Yojana (PMJDY), ~13 crore (23%) accounts remained inoperative⁵ highlighting the gap between access and engagement.

The next financial inclusion challenge is not access to the system, but meaningful adoption of the products within it. In wealth management, retail mutual fund penetration remains at only around 3.6% of the population⁶, significantly below mature markets. The constraint is not a lack of products, it is the difficulty of helping first-time investors understand, evaluate and act on opportunities suited to their circumstances. For many consumers, particularly those outside major urban centres, personalised financial guidance remains too costly for traditional distribution models to provide economically.

Insurance reflects a similar gap between market growth and wider customer participation. Overall insurance penetration in India stood at 3.7% of GDP in FY25⁷, unchanged from the previous year and below the post-pandemic peak. Despite growth in premium collections, penetration has largely plateaued, indicating that growth is not translating into a corresponding expansion of protection among underserved segments. The challenge is no longer product availability, but the ability to engage, educate and serve millions of customers whose individual economic value may not support traditional advisor-led distribution.

AI can fundamentally change how institutions address this challenge. By making personalised guidance, timely engagement, onboarding and servicing available at unprecedented scale and at a fraction of the cost of conventional channels, AI can help institutions convert financial access into meaningful participation. It can enable them to reach and serve customer segments that have historically been too expensive or difficult to cover profitably, helping more individuals understand, adopt and meaningfully use financial products across their evolving needs. The underlying BFSI penetration metrics reveal both the scale of the gap and the opportunity ahead.

³ World Bank, The Global Findex Database, 2025

⁴ PIB, 67 and Rising: India's Financial Inclusion Gains Momentum, 2025

⁵ Sansad, Newly Opened Jan Dhan accounts, 2025

⁶ AMFI, Mutual funds route to Viksit Bharat @ 2047, 2025

⁷ PIB, Insurance for All: Expanding Coverage, Strengthening Social Security, 2026

The AI adoption frontier has already moved

The proportion of Indian banks with GenAI use cases under implementation increased from 10% in 2024 to 44% in 2025 and 86% in 2026⁸, making AI one of the sector's fastest-growing technology priorities. Institutions that have not yet started risk falling behind peers that are already accumulating operational experience and building the capabilities required to scale.

The question is no longer whether AI will become part of everyday financial-services operations, but how it can be deployed responsibly and at scale. The institutions that create sustained value will not necessarily be those that move fastest across every area. They will be those that start early, prioritise the right use cases and progressively build the governance, processes and controls required to scale with confidence.

A Sustainable Path to AI in BFSI

The value of agentic AI will be determined not by how widely it is deployed, but by where and how it is applied. As AI improves the productivity of advisors, contact centres, service teams and marketers, institutions will naturally seek to extend its use across customer interactions. The strongest near-term opportunity lies in deploying AI to expand reach and participation across the financial system. Across customer education, engagement, acquisition, onboarding and servicing, AI can remove friction, provide relevant information and support timely action.

Greater caution is required when AI moves from enabling decisions to determining outcomes, particularly in underwriting, claims adjudication and product recommendations with material customer impact. The defining boundary is therefore not between front-office and back-office activities, but between expanding financial participation and automating decisions that demand accountability, oversight and judgement. Agentic AI can become a powerful force for financial inclusion when it extends human capacity without displacing human accountability.

⁸ FIBAC, Winning in the AI Era, August 2026



Eight trends

The value chain

Where AI lands



Trends Defining the Next Phase of Enterprise AI

SECTION 2

Trends Defining the Next Phase of Enterprise AI

As AI capability becomes more accessible, enterprise advantage will come from connecting use cases into workflows, matching models to the task and governing every deployment for value, cost and accountability.

AI is emerging as a strategic lever for advancing financial inclusion and enhancing the productivity of financial institutions. Its potential extends beyond improving individual processes to helping institutions expand customer reach, deepen penetration, and reduce the cost of serving customers – 76% of business leaders in India are expecting GenAI to have a significant business impact⁹. The opportunity is therefore not limited to efficiency; it lies in enabling financial institutions to serve more customers and create more personalised and accessible financial experiences.

As AI capabilities become more accessible, competitive advantage is shifting away from access to models and towards the capabilities required to deploy them effectively. 63% of Indian organisations believe they are ready to leverage GenAI effectively¹⁰ but translating that readiness into value will depend on trusted data, integrated technology foundations, cost discipline, responsible governance and workforce readiness. Much like digital-first organisations reshaped customer experiences and business models over the previous decade, AI-first is emerging as the next enterprise paradigm, where intelligence is designed into workflows, decisions and operating models from the outset rather than added as a separate technology layer. AI-first

organisations will rethink how workflows are designed, experiences are shaped and tasks are executed and governed.

Trends shaping the next phase of the AI revolution

1. **AI focus is evolving from experimentation to scaled value:** GenAI adoption has gained significant momentum in India, with 47% of organisations operating multiple live use cases and 10% beginning to scale them across business functions¹¹. As portfolios of pilots and live applications grow, emerging pilot fatigue is encouraging organisations to become more selective. Leadership attention is shifting towards proven use cases with clear business ownership, measurable outcomes and credible pathways to scale.
2. **From standalone use cases to functional transformation:** AI leaders are moving beyond task-level assistance and embedding AI across complete functions and workflows. In India, 46% of organisations report deploying AI within multi-step workflows across systems, compared with

⁹ EY, The Aldea of India: Outlook 2026

¹⁰ EY, The Aldea of India: Outlook 2026

¹¹ EY, The Aldea of India: Outlook 2026

25% primarily using basic AI tools for individual tasks¹². Moving from individual applications to functional transformation requires organisations to redesign processes, integrate data and systems, establish clear business ownership, drive workforce adoption and measure outcomes across the end-to-end workflow.

3. **AI investment is shifting from efficiency-led automation to growth, risk and decision intelligence:** The first wave of AI in financial-services focused on high-volume activities such as servicing, collections and document processing, where efficiency gains could be demonstrated quickly. E.g., A digital-first life insurer in India has made AI-led underwriting its most mature AI implementation, using data-driven models to support risk assessment and policy issuance. Approximately 80% of non-medical, low- and medium-ticket applications are now issued on the same day¹³.
4. **AI agents are becoming the newest members of the workforce:** AI is evolving from a tool employees use into a digital teammate capable of researching, analysing information and executing tasks under human direction. 24% of Indian organisations have deployed autonomous AI agents¹⁴, moving the concept of a digital workforce closer to enterprise reality. As adoption matures, organisations are beginning to envisage employees working with one or more specialised agents aligned to their roles and workflows, making digital capacity an increasingly deliberate component of workforce design.
5. **AI will scale only as fast as the data, integration and operating foundations beneath it:** Access to capable AI models is becoming more widespread, but enterprise-grade deployment remains constrained by the foundations surrounding

them. 78% of Indian organisations identify integration and data-readiness as a major barrier to adoption¹⁵. Organisations that can establish trusted data, connect AI with core systems and build production-ready technology and operating capabilities will be better positioned to move from isolated deployments to reliable enterprise scale.

6. **Rising token consumption is driving focus towards AI unit economics:** As AI moves into production, enterprises are discovering that rapid adoption can generate significant and unpredictable consumption costs without a corresponding increase in business value. In one reported instance, uncapped employee usage resulted in a company spending US\$500 million on AI consumption in a single month¹⁶. The incident highlights the risks of "tokenmaxxing," where organisations encourage or celebrate rising AI usage without establishing whether the additional consumption is improving productivity, customer outcomes or financial performance.
7. **Small language models (SLMs) are emerging as the operational backbone for fit-for-purpose AI:** Enterprises are shifting from using the largest available models to selecting the right model for each task. Routine activities, such as retrieving product information for sales teams or assessing applicant eligibility in underwriting, can be handled by domain-specific SLMs with greater speed, control and cost predictability. The potential is already visible in India, where one voice-AI company reported that its purpose-built SLM reduced time to first byte to 53 milliseconds, compared with more than 200 milliseconds for global APIs, while delivering 20% to 30% better hallucination control¹⁷. Larger models can then be reserved for complex analysis and reasoning, creating a hybrid model portfolio aligned to the requirements of each workload.

¹² EY, The Aldea of India: Outlook 2026

¹³ Annual Reports, Stakeholder Discussions, EY Analysis

¹⁴ EY, The Aldea of India: Outlook 2026

¹⁵ EY, The Aldea of India: Outlook 2026

¹⁶ Fast Company, Jude Cramer, May 2026

¹⁷ EY, The Aldea of India: Outlook 2026

8. Governance, auditability and accountability are becoming essential components of AI operating model: As AI influences more decisions and executes activities with greater autonomy, governance can no longer remain a standalone compliance exercise. 64.5% of Indian organisations rate data governance and security concerns as very severe barriers to scaling GenAI¹⁸, reinforcing the need to embed transparency, monitoring and accountability throughout the AI lifecycle. Emerging frameworks, including the RBI's FREE-AI framework for financial services, further signal a shift from defining responsible AI principles to continuously demonstrating that AI systems and outcomes remain explainable, reliable and auditable in operation.

AI represents a fundamentally different wave of technological change. Unlike earlier technologies that largely digitised existing processes, AI is beginning to redesign how work is performed, decisions are made and value is created. For financial services, this shift has the potential to reshape not only individual institutions, but the structure and economics of the sector itself. As AI becomes embedded in workflows, decisions and

interactions, organisations will need to rethink not only how they deploy technology, but how they design the enterprise itself. The most significant impact of AI may ultimately be less about what organisations do, and more about how they operate.

AI in action across the financial-services value chain

Across financial services, AI adoption has typically gained early traction in high-volume operational areas such as customer servicing, contact centres, document processing and underwriting, where use cases can be integrated into defined workflows and outcomes demonstrated more readily. AI is now moving closer to the customer and the frontline, supporting prospect engagement, personalised interactions, advisor enablement and conversion. Realising value from these applications, however, depends more heavily on frontline adoption, high-quality customer and product context, and integration into existing ways of working.

The following section examines how AI is being applied across four core areas of the financial-services value chain:



1. Sales and Distribution

Financial-services players are embedding AI into customer acquisition and frontline sales enablement to make engagement more contextual, responsive and scalable. The shift is extending beyond isolated sales support towards connected capabilities that strengthen prospecting, conversion and distributor effectiveness.

¹⁸ EY, The Aldea of India: Outlook 2026

AI-led Conversational Lead Generation

Large Diversified NBFC

Uses voice and text interactions to engage customers, capture interest and generate new loan offers at scale

Vernacular Lead-to-Quote Acceleration

Leading Life Insurer

Uses ~4 multilingual WhatsApp journeys to engage prospects in their preferred language leading to 4X improvement in lead to quote conversion

AI Voice-led Prospect Nurturing

Large Retail-Focused NBFC

10 lakh+ monthly voice interactions to assist personal-loan prospects, answer queries and drive conversion. Enables scalable follow-up across large prospect pools

GenAI-led Humanised Prospect Engagement

Leading Life Insurer

Combines Customer 360 with a humanoid voice bot to engage prospects and generate leads.

AI-powered Relationship Manager Co-Pilot

Leading Private Bank

AI-assisted pitch practice to help prepare for customer meetings with relevant context and recommendations.

Sales and Advisory Copilot

Leading Health Insurer

Provides agents with explanations, plan comparisons, premium calculations, UW inputs and claims guidance during customer interactions.

2. Customer Onboarding & Underwriting

AI is reducing the friction created by document-heavy onboarding and underwriting processes, where information must be captured, verified and assessed across multiple systems. By improving the speed and consistency of processing and decision support, these capabilities are shortening the path from application to approval or issuance.

AI-led Digital Onboarding

Leading Life Insurer

Digitises the new business journey from customer onboarding to policy issuance, enabling faster, more consistent processing

AI-powered Identity and KYC Verification

Leading Diversified NBFC

Processes application documents and verifies customer identity through face matching during onboarding and KYC – 69.1 million documents processed

Zero-Ops Document Processing Engine

Leading Private Bank

Automates document checks, validations and reconciliations across operational workflows. Reduces manual effort via AI-led processing layer

AI-led Credit Document Validation

Large Retail-Focused NBFC

Reads, compares and validates documents across lending workflows.

GenAI-powered Document Intelligence

Leading General Insurer

Uses GenAI to extract, interpret and review information from documents at scale leading to 77% cost savings in OCR

AI-led Underwriting Decisioning

Leading Life insurer

Automates underwriting processing and decisioning for policy applications. Enables faster policy issuance and reduces errors

3. Servicing & Claims Management

AI is helping financial-services players deliver faster, always-available and increasingly conversational support across everyday servicing and claims. Customer-facing self-service and employee-facing assistance are improving accessibility, resolution and service productivity. In claims, where an insurer's promise is ultimately delivered, AI is reducing processing delays and improving consistency, enabling straightforward cases to progress more efficiently from intimation to settlement.

Omnichannel AI Servicing Concierge

Large Retail-Focused NBFC

Enables customers to access servicing across the app, website and WhatsApp through chatbots, voice systems and digital journeys.

Conversational Policy Servicing Engine

Leading Private Bank

AI bot and WhatsApp banking conversational interfaces support deposits, cards, loans, transactions and FAQs with 10 million+ customer conversations processed

AI-powered Inbound Voice Assistant

Leading General Insurer

Understands customer intent and resolves service requests across inbound calls. Manages 80,000+ calls per month, with 84% intent accuracy.

Vision-led Vehicle Damage Assessment

Leading General Insurer

75% of physical inspections are AI assisted – vehicle images analysed to detect damage and estimate the claim amount

AI-powered Health Claims Document Intelligence

Digital-first General Insurer

Classifies health-claim documents to accelerate claim processing.

GenAI-assisted Claims Investigation

Leading Life Insurer

Uses GenAI to assist claims teams in reviewing and investigating cases before settlement leading to average claim settlement time of ~2 days

4. Enterprise Workforce & Technology

AI is becoming embedded in the internal workings of financial institutions, augmenting employees and accelerating technology delivery. Applications range from enterprise knowledge assistants and frontline intelligence tools to software-development support and compliance workflow automation, enabling employees to access information faster, execute tasks more consistently and focus on higher-value activities.

AI-assisted Software Development

Large Diversified NBFC

AI generated 71% of business requirement documents and 39% of test cases, contributing to a 24% improvement in developer efficiency.

Domain Knowledge Assistants

Large Public-Sector Bank

Deployed 15 domain-specific GenAI assistants to provide employees with instant, contextual guidance on workflows, norms and procedures.

AI-powered Talent Marketplace

Large Private-Sector Bank

Connects employees with relevant internal roles, build skills and facilitated ~6,500 internal role movements in FY26.

Law-Enforcement Response Management

Large Public-Sector Bank

Uses AI/ML to automate law-enforcement requests, improving response accuracy, turnaround time and centralised monitoring.

Enterprise AI Workbench

Large Private Life Insurer

Provides reusable AI capabilities and enterprise knowledge tools to support employees across functions.

Enterprise Productivity Assistant

Leading General Insurer

Supports employees with contextual information retrieval, assisted responses and task automation across functions.



Eight areas

Nine use cases

Impact and learnings

Agentic AI in Action: Areas of Deployment at ABCL

SECTION 3

Agentic AI in Action: Areas of Deployment at ABCL

ABCL started its AI journey where value was visible and the blast radius was contained. It then built the foundations required to scale Agentic AI into increasingly complex and consequential journeys.

Aditya Birla Capital has moved beyond experimentation to deploy agentic AI across eight areas of the business, with several initiatives operating in production long enough to demonstrate measurable outcomes. These deployments span sales, customer onboarding, engagement and servicing, workforce productivity, software development, and core functions such as audit, finance and compliance.

ABCL was able to advance multiple deployments in parallel because it had already established several shared foundations: an enterprise orchestration layer, a common data platform and a dedicated AI Centre of Excellence. Institutions with more limited data, technology or AI engineering capacity may need to follow a more deliberate sequence. Lower-risk applications such as top-of-funnel engagement and sales coaching can provide practical starting points because they do not place AI in control of consequential customer decisions, have a relatively contained impact

if errors occur and can demonstrate measurable value within a shorter period.

The case studies that follow are intended to provide a practitioner's view of what it takes to move from deployment to business impact. For each area, the section sets out the business problem addressed, how the AI capability operates, the outcomes achieved and the lessons that emerged during implementation.

Note on reported outcomes: Results reported in this section are drawn from ABCL's internal measurement of the specific deployments named, not from a controlled trial or any industry benchmark. Each is stated against the pre-deployment baseline for that same journey and business line and covers use cases across organisations under ABCL rather than the group as a whole. They are offered as a practitioner's account of what moved, and an institution with a different starting baseline should expect a different magnitude.



AREA 1

Agentic Orchestration of Customer Journeys

Case Study 1

Problem Statement

Most institutions approach AI through point solutions: a chatbot on one page, a scoring model in a back-office queue or a voice bot for a specific call type. Each solution works in isolation and loses customer context as soon as the case moves to the next step. ABCL's first and broadest approach was to do the opposite.

AI Scope

The lending journey is organised around seven human roles, representing the loan officer, processor, underwriter, closer, servicing, compliance and legal

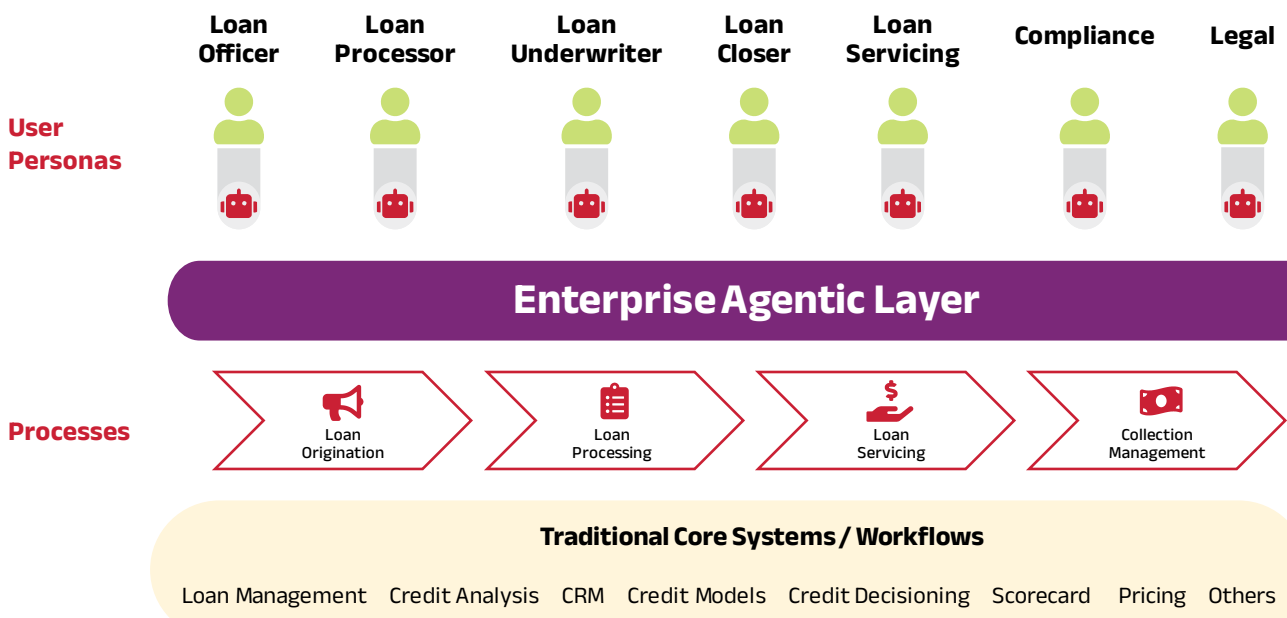
functions, rather than AI agents. Each is paired with an AI counterpart that performs the repetitive elements of the role, while the individual retains judgement and decision-making responsibility.

An orchestration capability coordinates the end-to-end journey across these agents. It determines which agent acts next, what context is carried forward from preceding steps, and when the process must pause for human intervention. First deployed across a lending journey, the layer supports every role from loan officer to legal and integrates with the core systems already used across credit analysis, CRM, credit models, decisioning, scorecards and pricing.

ABCL case study

Loan Underwriting and Back Office Automation

The use case has been proven end-to-end for loan underwriting and back-office automation, where a case moves through document collection, verification, risk assessment and issuance. Instead of automating each step in isolation, the orchestrator carries the case and its context across all four, turning what were separate queues into one journey. In practice, that meant GenAI extraction across 50+ document types, a dynamic checklist, cross-document triangulation and a case summary feeding a recommendation, each handed cleanly to the next agent rather than re-keyed at every desk.



Enterprise agentic layer spanning loan origination, processing, servicing and collection management, with an AI counterpart provisioned against every persona in the journey

Impact Areas

The most significant gains came from reducing the time cases spent waiting between steps. Credit memo preparation fell, representing close to a 90% reduction at a key bottleneck for business & SME Loans. Home finance & supply chain finance throughput increased 1.25 times, with 80% first-time-right processing.

The value came not only from accelerating individual activities, but also from reducing delays between them. Customer context captured upstream could flow through the journey, reducing the need for employees to reconstruct information at each stage. The resulting capacity could then be directed towards processing more cases and reaching more customers. These figures are based on ABCL's internal measurement against its pre-deployment baselines and do not represent controlled-trial results.

Key learnings

The most important lesson was organisational and architectural, rather than purely technical. Several early

initiatives were developed as standalone solutions, requiring additional effort to connect them into an end-to-end journey as deployment expanded.

The lesson was not that every institution must invest in a full orchestration layer before launching its first AI use case. Rather, institutions should establish early how customer context will move across workflow stages, where systems and solutions will integrate, and when human intervention will be required. Even when starting with simpler automation, designing for future interoperability can prevent today's point solutions from becoming tomorrow's integration challenge.

Case Study 2

Problem Statement

An underwriting decision may appear to be a single judgement, but it is built on dozens of individual verifications across identity, documents, disclosures, medical records and financial information. Each verification must be completed accurately before the final judgement can be made. Processing the entire case

through a single large model can make these checks expensive to execute, difficult to audit individually and vulnerable to the model drawing conclusions across steps for which it was not specifically instructed. AI can support the assessment, but the final underwriting decision must remain with a human.

AI Scope

Using Pints AI's Autothought, the underwriting workflow is decomposed into hundreds of small, single-purpose agents. Each agent performs a narrowly defined task that can typically be answered in near-binary terms:

Is the document what it claims to be? Does a disclosed condition appear in the medical report? Does the declared income support the proposed sum assured?

An orchestration layer brings these outputs together into a recommendation file for the human underwriter, who retains responsibility for the final decision. Each task is assigned to the smallest model capable of performing it effectively, with larger models reserved for the limited steps that require more complex analysis or reasoning.

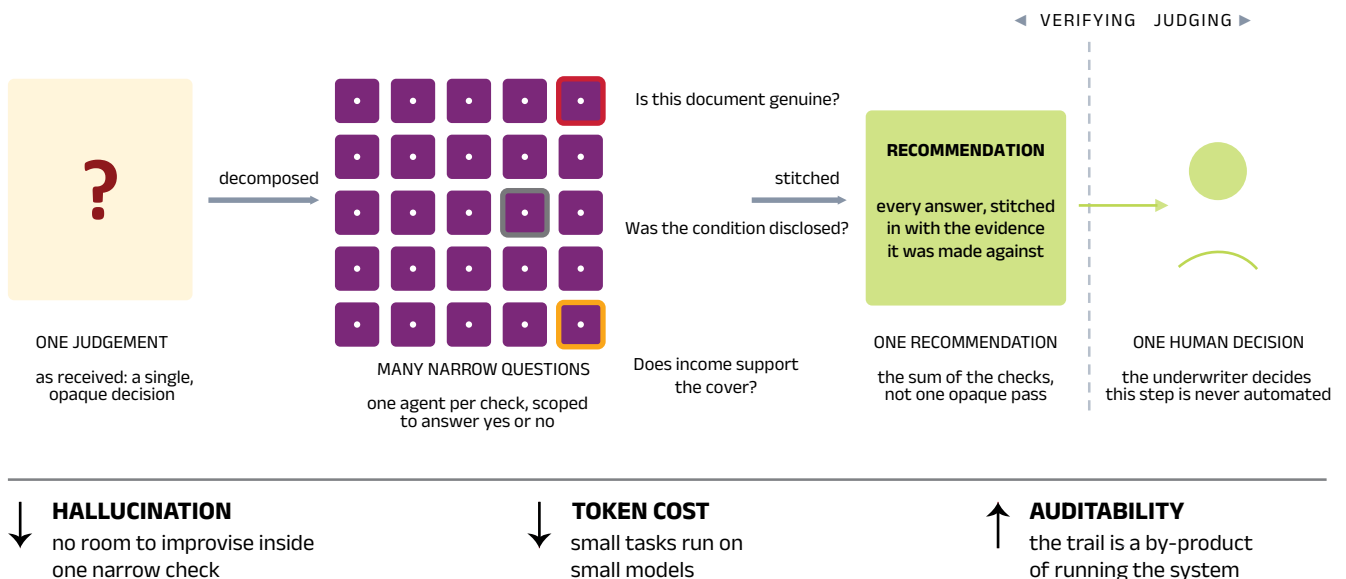
ABCL case study

Life Insurance Underwriting

In life insurance underwriting, a single decision is broken into hundreds of near-binary checks, each executed by its own agent and recorded with the evidence it was made against. The recommendation the underwriter receives is the sum of those checks rather than the output of one opaque pass, so a reviewer or a regulator can trace any element of it back to the record that produced it.

Decomposing an underwriting decision

The judgement is not automated. It is assembled.



Impact areas

The architecture improved reliability, cost efficiency and auditability by design. Hallucination risk is reduced because each agent performs a narrowly defined, near-binary check with limited scope for improvisation. Token costs are lowered by routing these smaller tasks to fit-for-purpose models, while larger models are reserved for activities requiring more complex reasoning. Auditability is strengthened because each check and its supporting evidence are captured during execution,

creating a traceable record rather than requiring the audit trail to be reconstructed afterwards.

Key learnings

The most significant effort lies in decomposing the underwriting workflow, not simply tuning the models. Converting an underwriting manual into hundreds of discrete, testable checks requires implicit judgement to be made explicit. This work must be undertaken with experienced underwriters, not designed around them.

AREA 2

Telesales and Cross-Sell

Problem Statement

A telecaller working from a generic script may increase the number of calls made, but not necessarily the quality or relevance of each conversation. The script lacks context on the customer's existing holdings and cannot determine which products may be appropriate for the customer's goals and risk profile.

AI Scope

An interactive audio-visual master agent coordinates the conversation and directs each activity to the

appropriate task agent. Six specialised agents sit beneath it: a purchase journey agent, drop-off journey agent, investment planning agent, calculation agent, product comparison agent and mutual fund awareness agent. Each accesses the relevant tools, APIs and information sources required to perform its specific task.

Customer interactions take place through voice calls and WhatsApp, with the master agent maintaining context across both channels and linking each interaction to the same underlying customer journey.

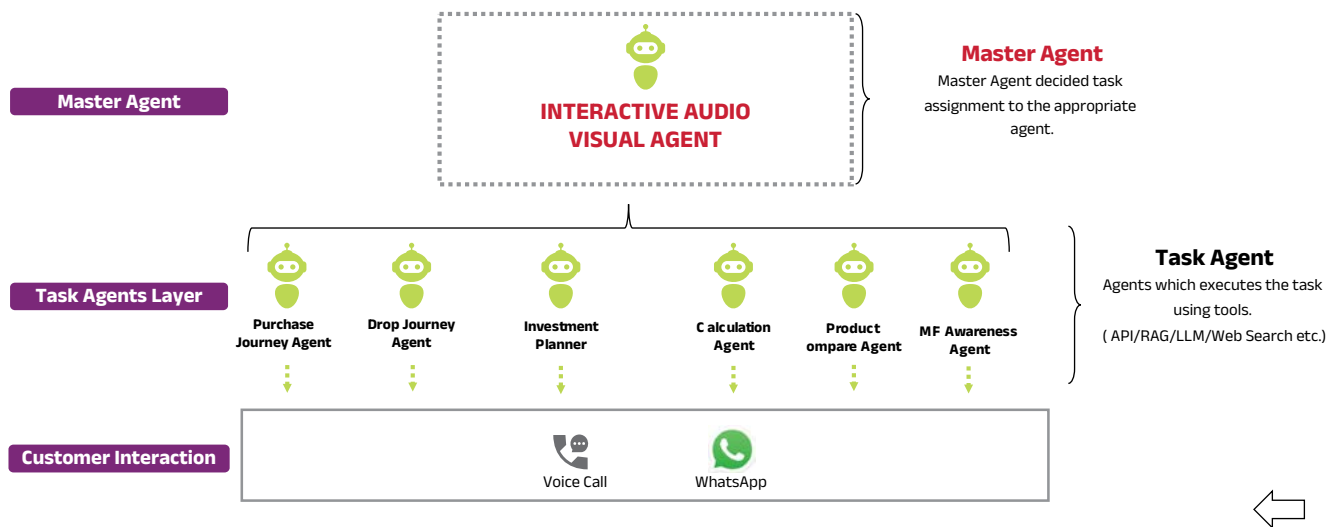
ABCL case study

Mutual Fund Telesales & Cross-Sell

Within the asset management business, the telesales and cross-sell system pulls a customer's existing folio and portfolio position, computes which scheme is actually suitable for that customer's stated goal and risk profile, and builds a pitch grounded in both, instead of a tele-caller working from a single generic script. The stated ambition for the system was to take outbound reach on mutual fund customers, both new and existing, toward the full base rather than a manually dialled sample.

Use Case

TeleSales Outbound Calls to Mutual Fund new/existing customers for increasing the reach to 100%



Use case: tele-sales outbound calling to mutual fund customers, built as a master agent over six task agents, with reach targeted at the full base rather than a sampled segment

Impact Areas

Customer conversations became more contextual and relevant, with the calculation and analysis agents assessing the customer's existing holdings before the pitching agent generated the conversation approach. Telecallers also spent less time retrieving information across systems and more time engaging with customers.

Key Learnings

Defining the right boundaries between task agents required multiple iterations. When agents were scoped too narrowly, the additional orchestration and hand-offs outweighed the benefits of decomposing the workflow. When they were scoped too broadly, it became difficult to identify whether an ineffective pitch originated from the calculation, product comparison or pitch-generation stage.

The key learning was to decompose the workflow into tasks that were distinct enough to measure and improve independently, but broad enough to avoid unnecessary orchestration complexity.

AREA 3

Top of Funnel Engagement

Problem Statement

Traditional broadcast marketing treats customers within a segment uniformly, reaching them at the same time and through the most cost-effective channel. While this approach can maximise reach, it does not distinguish between customers who are ready to act and those who have merely received the communication. As a result, engagement may increase without a corresponding improvement in customer intent or conversion.

AI Scope

The system brings together customer and product context to determine who to engage, when to engage them, through which channel and with what message. Signals are drawn from the ABCD app and website, performance marketing, the service platform and the cross-sell engine.

Three specialised agents work together:

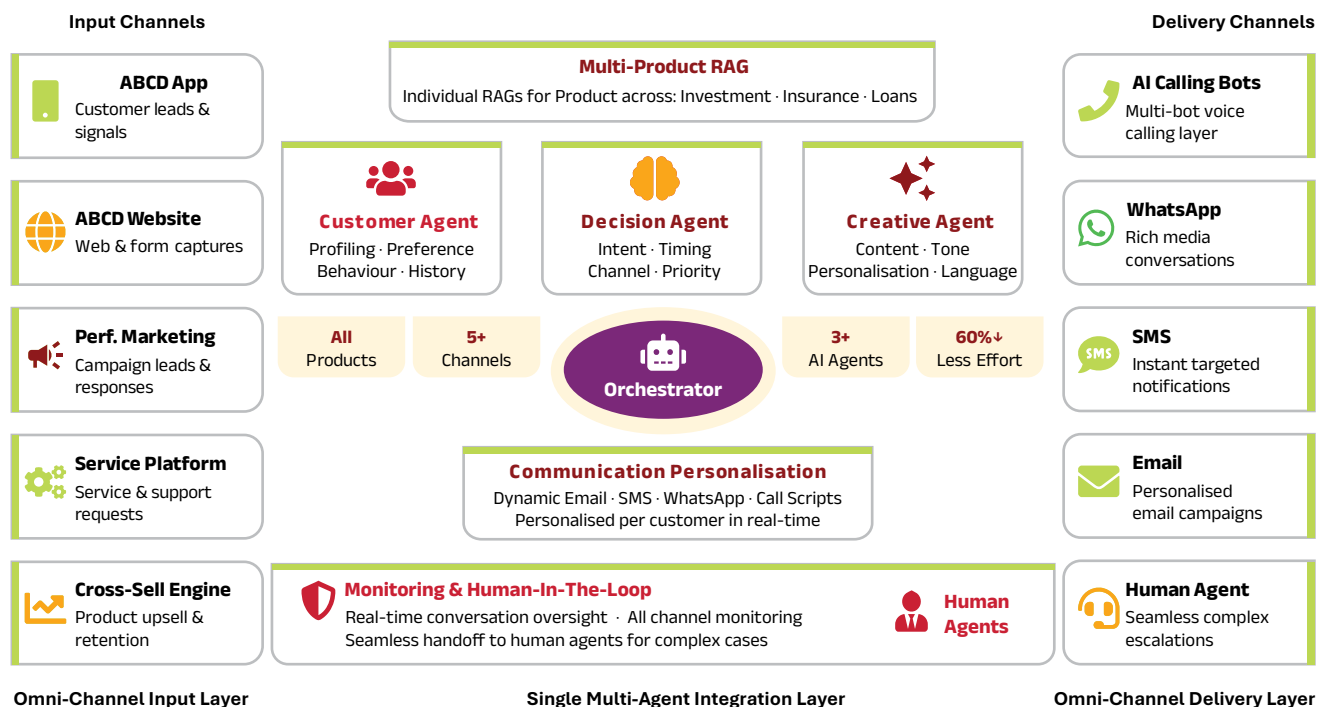
- An IQ agent builds a customer profile based on their preferences, behaviour and interaction history.
- A Decision agent assesses intent, timing, channel and priority.
- A Creative agent tailors the content, tone, language and level of personalisation.

An orchestration layer coordinates these agents across product lines and five or more channels. It personalises communications, including email, SMS, WhatsApp messages and call scripts, in real time before routing them through the appropriate delivery channel. Each interaction is monitored, with a defined hand-off to a human agent for complex or sensitive cases.

ABCL case study

NEXUS: Super App DigiGold Nudges

Within the ABCD super app, nudges prompt small, recurring gold investment, timed and personalised to each customer's own pattern of behaviour rather than sent on a fixed calendar to the entire base. The same architecture that powers the nudges reports a 60% reduction in the manual effort needed to run a campaign across products and channels, because the decision agent and creative agent do the segmentation and content work that a campaign team would otherwise do manually.



Omnichannel input layer, single multi-agent integration layer and Omnichannel delivery layer, with the orchestrator recording a 60% reduction in manual effort across the products and channels it covers

Impact Areas

Customer engagement became more targeted, with the focus shifting from the number of customers contacted to the relevance and effectiveness of each interaction. Thereby, increasing customer interactions & uplift across ToFu, MoFu & BoFu of product journeys.

Key Learnings

The effectiveness of personalised engagement depended as much on journey orchestration as on model accuracy. Customer signals and predictive insights created value only when supported by coordinated frequency, channel and suppression rules across the journey. Without this coordination, the same customer could receive repeated messages through different channels, increasing the risk of nudge fatigue and weakening the impact of communication.

AREA 4

Customer Service and Retention

Problem Statement

Renewal, retention and collections outreach is often broad-based and reactive, with customers approached through standardised journeys that do not adequately account for their risk, intent or servicing needs. This limits institutions' ability to identify customers who require early intervention, tailor engagement to the underlying issue and direct human effort towards cases where it can have the greatest impact.

AI Scope

This area brings together four connected capabilities across customer service and retention.

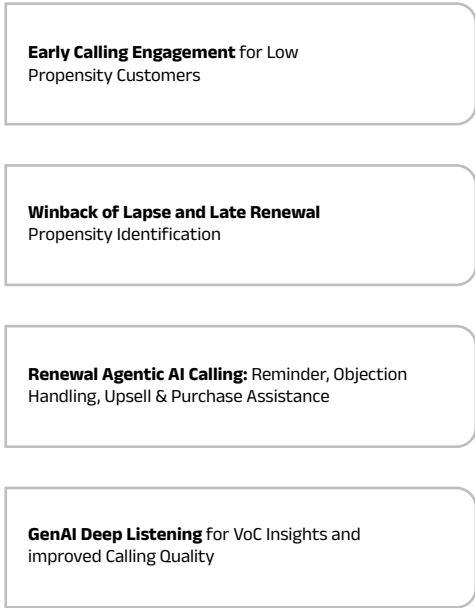
- Service Assist provides service representatives with relevant information during live customer interactions
- An email virtual assistant classifies incoming service and transaction emails by request or complaint type and drafts responses for review
- Speech analytics analyses inbound, outbound, NPS and telesales calls to identify interaction patterns and opportunities for improvement
- Voice bots manage renewal, SIP-retention and collection calls, with defined escalation to a human when a case becomes complex, sensitive or falls outside the bot's scope

ABCL case study

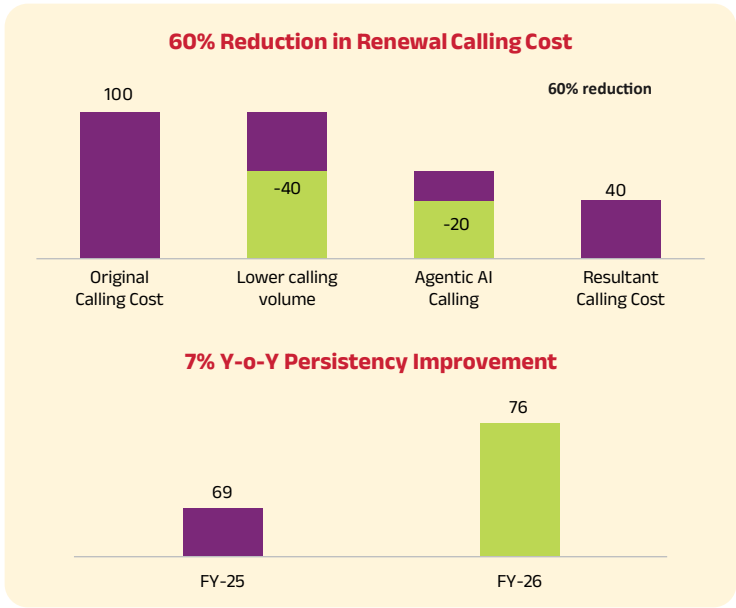
Health Insurance Renewal

Within health insurance renewals, the voice bot engages customers ahead of lapse, identifies low propensity and win-back candidates before they reach the renewal date, handles reminders, objection handling, upsell and purchase assistance directly, and applies GenAI deep listening on the resulting calls to improve calling quality over time.

AI/Agentic AI Initiatives:



Business Outcomes:



Early engagement, win-back identification, agentic renewal calling and GenAI deep listening, mapped against a 60% reduction in renewal calling cost and a 7 point improvement in year on year persistency

Impact Areas

Renewal and retention engagement became more proactive and targeted. The system identifies customers at risk of lapsing before the renewal date, enabling earlier outreach and directing the human renewal team towards cases that require personal intervention.

The initiative reported a 60% reduction in renewal-calling costs and a 7% year-on-year improvement in persistency, from 69% to 76%. While the cost reduction is directly linked to lower calling volumes and the lower unit cost of AI-led calls, the persistency improvement

reflects the broader renewal programme enabled by the AI intervention.

Key Learnings

Clear escalation pathways are critical when AI is deployed in sensitive customer interactions. Renewal, retention and collection calls are subject to regulatory expectations around tone, disclosure and customer treatment. If a voice bot detects customer distress, a complaint or a request outside its defined scope, the interaction should be transferred immediately to a human rather than allowing the bot to continue.

AREA 5

AI Sales Coaching for Frontline Productivity

Problem Statement

Sales representatives often need to access product information, eligibility criteria, calculations and objection-handling guidance during customer

conversations. When this information is fragmented across documents and systems, representatives may lose valuable interaction time, affecting the quality and consistency of the conversation.

AI Scope

The initiative creates a one-stop, omnichannel sales-coaching layer that helps sales teams and advisors access ABCL's product and sales knowledge, practise customer conversations and receive contextual support throughout the sales journey.

An interactive audio-visual master agent directs each request to the appropriate specialised agent. These agents support:

- Product knowledge and eligibility queries
- AI-enabled role-play and practice
- Short sales-learning content and refreshers

- Product and financial calculations
- Sales engagement and activation
- Sales insights and contextual nudges
- Demand-funnel management
- Commission and rewards queries

Sales representatives can access these capabilities through the sales application, learning management system, voice calls or WhatsApp. The platform integrates with ABCL's existing data and CRM systems, allowing support to be delivered within established sales workflows rather than through a standalone tool.

ABCL case study

Finwise Sales Coach

Built on ABCL's Finwise platform, the sales coach gives a representative instant access to product policy, underwriting methods and eligibility criteria the moment they need it, coaches through role play scenarios, gives real time feedback during practice, and lets a representative simulate different customer scenarios before they are in front of a real one. Sales insights and nudges, demand funnel visibility and commission queries sit in the same layer, so a representative is not switching between four separate tools to get through a single working day.



One-stop omnichannel agentic coaching layer, from a single master sales coach agent down to eight task agents covering knowledge, role play, insights and rewards

Impact Areas

The sales coach improved frontline readiness by providing representatives with contextual support across product knowledge, customer conversations, calculations and objection handling. New joiners could practise a wider range of customer scenarios through AI-enabled simulations, while experienced representatives could access relevant information and guidance within the flow of work. This reduced dependence on document searches and supervisor support, helping representatives prepare for and navigate customer conversations more consistently.

Key Learnings

The value of an AI sales coach depends not only on the quality of its knowledge base, but also on whether representatives integrate it into their daily workflows. Keeping product, policy and eligibility information current is essential to maintaining user trust. However, sustained adoption also requires the solution to be easily accessible, relevant to real sales situations and embedded in the channels representatives already use.

AREA 6

Marketing and Personalisation

Problem Statement

Personalisation is often applied at the segment level, with broad customer cohorts receiving different versions of the same campaign. While this is more targeted than mass marketing, it does not fully use the individual-level customer signals already generated across top-of-funnel engagement and telesales.

AI Scope

The marketing and personalisation layer determines the most relevant creative, offer and channel for

each customer based on individual context rather than segment type alone. It then shares this next-best-offer signal with the top-of-funnel and telesales agents, enabling all three capabilities to operate from a common customer view.

This creates a connected decisioning layer across marketing and sales, helping ensure that customers receive more consistent and contextually relevant engagement across channels.



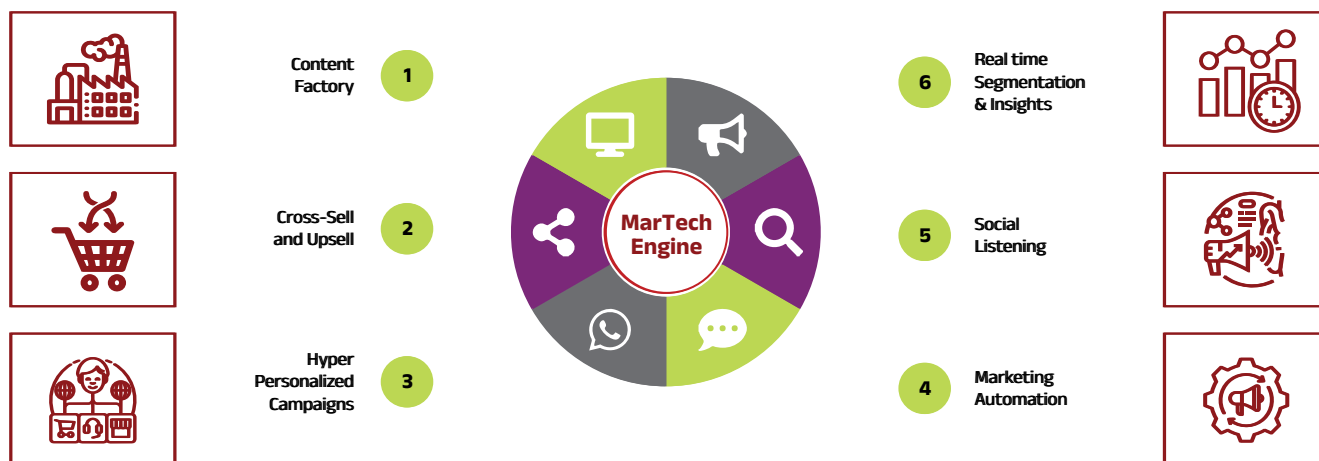
ABCL case study

Martech Platform

The platform is a 360-degree system for end-to-end marketing campaigns using generative AI, customised to each brand's guidelines, organised around six connected capabilities: a Content Factory; cross-sell and upsell; hyper-personalised campaigns; marketing automation; social listening; and real-time segmentation and insights. A design-led agentic workflow spans across campaign planning, calendar creation, content generation, validation and feedback, with marketers involved at key decision points.

MarTech Platform

A 360-degree platform for end-to-end marketing campaigns using GenAI & customized to brands guidelines



Built with security and privacy by design & hosted on our Secure Cloud, ensuring indemnified and plagiarism-free content

Impact Areas

The platform is at an early stage of deployment, and measurable business impact is yet to be established. Initial progress is reflected in the development of a shared personalisation layer that connects customer and product context across marketing, top-of-funnel engagement and telesales.

Key Learnings

Consent and customer communication preferences must be embedded into personalisation by design. Given the regulatory requirements governing marketing communication in India, consent cannot operate as a downstream compliance check. It must inform which customers can be engaged, through which channels and for what purpose before a campaign or interaction is generated.

AREA 7

Agentic Software Development

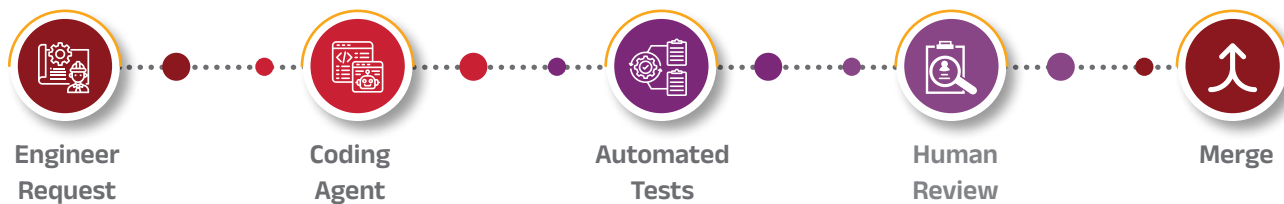
Problem Statement

Software delivery can be constrained by manual effort and delays across the development lifecycle, from business-requirement documentation and impact assessment to coding, testing and UAT. These bottlenecks limit how quickly technology teams can deliver and maintain the AI and digital capabilities required across the enterprise.

AI Scope

Within ABCL's technology teams, AI agents support multiple stages of the software-development lifecycle, including the creation of project documentation, code generation, pull-request reviews and test-case development. The agents work alongside technology teams, while engineers retain responsibility for reviewing and approving outputs.

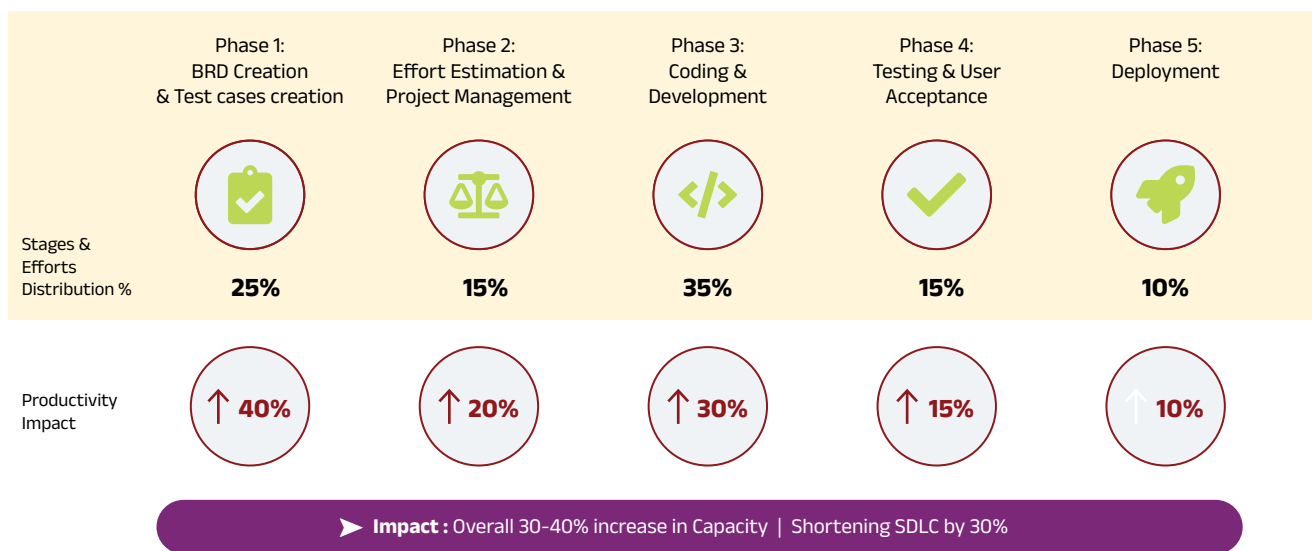
Access is managed in line with the controls applied to privileged internal users. Agents do not have standing write access to production systems, and human review remains mandatory before any changes are approved or deployed.



ABCL case study

SDLC & Testing across ABCL

Built on the IntelAct platform, the SDLC Assistant generates business requirement documents and project documentation from a user's own inputs, checked against standardised reference documentation so that requirements do not conflict and change impacts on existing systems are covered. It produces a structured BRD with linked test cases, a project plan and a dependency map, and writes into the code repository directly.



Impact Areas

The SDLC assistant accelerated delivery across multiple stages of the development lifecycle. It supported the creation of business requirement documents, project documentation and test cases, while helping technology teams identify dependencies and potential change impacts earlier in the process.

Key Learnings

Human review must remain mandatory for any AI-generated output that could affect production systems. ABCL treats the agent as a contributor whose output is reviewed and approved in the same way as work produced by an engineer. Agents do not have standing write access to production systems, ensuring that accountability for code changes and deployment decisions remains with the technology team.

AREA 8

Agents for Audit, Finance, Investment and Compliance

Problem Statement

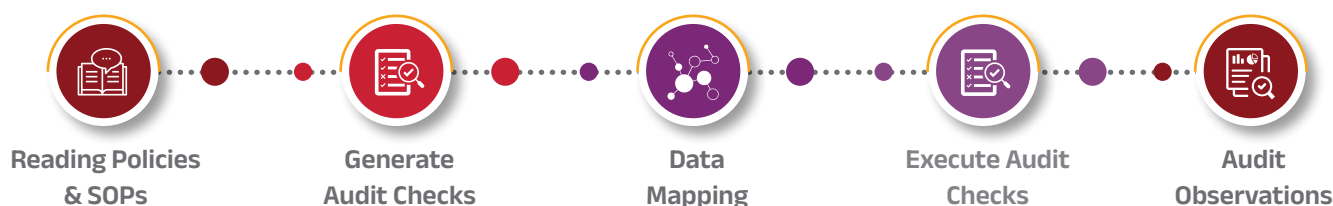
Decisions and findings across audit, finance, investment and compliance must be explainable and defensible to internal reviewers, auditors and regulators. This makes these functions particularly sensitive to AI systems whose analysis, evidence or reasoning cannot be reconstructed after an output is produced.

AI Scope

Across audit, finance, investment and compliance, AI agents perform financial checks, summarise

investment research for internal review and flag items in regulatory filings that require closer examination. The agents provide evidence-backed analysis and recommendations rather than making final decisions independently.

Human sign-off remains mandatory for consequential decisions, while a traceable record of the agent's analysis and supporting evidence is maintained for subsequent review. This enables AI to improve the speed and coverage of analysis without diluting human accountability



ABCL case study

AASTRA: AI-Enabled Automation of Audit Execution

AASTRA automates the execution stage of an internal audit. It reads the policies and standard operating procedures that govern a process, derives the audit checks those documents imply, maps each check to the underlying data structures in the warehouse, runs the checks and produces audit observations for the audit team to review.

Where an exception is found, the agent produces a deep-dive report on it, flags the anomaly and surfaces a recommended action, leaving the judgement on that action with the audit team.

Impact Areas

The deployment expanded the scope and consistency of control activities by automating defined stages of audit execution. Agents can interpret policies and standard operating procedures, generate audit checks, map them to underlying data and identify exceptions for review by the audit team.

Key Learnings

The pace of AI adoption should reflect the consequence and accountability associated with each use case. Audit, finance, investment and compliance require stronger traceability, human oversight and validation than lower-risk applications. Wider deployment should therefore follow only once outputs can be traced to their supporting evidence and consequential decisions remain subject to human review and sign-off.



Role clarity

Capability

Decision rights

Building the Human Operating Model for AI

SECTION 4

Building the Human Operating Model for AI

AI adoption is not a training problem. It is a management and work-design challenge: redesign the role, activate the manager and build a credible path for every employee.

The workforce challenge presented by AI is no longer primarily about technology adoption. It is about whether organisations can redesign workflows and tasks translate adoption into enterprise value. As AI becomes embedded in day-to-day activities, institutions must rethink how work is performed, decisions are made and accountability is shared between people and technology.

In India, employee adoption is already well underway. EY's 2025 Work Reimagined research found that 88% of employees use AI at work and 37% use it daily¹⁹, signalling that AI is rapidly becoming part of how work gets done. Yet adoption does not automatically translate into organisational readiness. Many institutions continue to operate with roles, workflows, performance measures and governance structures designed for a pre-AI environment, creating a widening gap between individual usage and enterprise-wide value.

The next phase of workforce transformation will be defined by how effectively institutions close this gap. The organisations that benefit most will not necessarily be those with the highest AI usage, but those that integrate AI most effectively into roles and workflows. This requires clarity on how responsibilities are allocated, how roles are redesigned, how employees build proficiency and where human judgement and accountability must remain central.

Capability building must therefore extend beyond training employees to use AI tools. It requires redesigning the human side of the operating model so that AI augments workforce capacity while preserving trust, judgement and accountability. The six considerations that follow outline the capabilities required to make this transition successfully.

1) Build role clarity before adoption

Consideration: Apparent resistance to AI often reflects uncertainty about how roles, responsibilities and accountability will change, rather than an unwillingness to adopt the technology.

Indian employees are already embracing AI. The challenge is therefore less about creating willingness and more about clarifying how AI fits into day-to-day work. Employees need to understand which activities AI will support, where human judgement remains essential, what constitutes responsible use and who remains accountable for the outcome.

Role clarity is therefore a prerequisite for sustained adoption. When employees understand how their work will change, how success will be measured and where their expertise continues to matter, they are more likely to use AI confidently and responsibly.

¹⁹ EY, Can AI advance toward value if workforce tensions linger?, 2025

Practice: Define a Role Impact Charter

For every material AI deployment, create a concise Role Impact Charter before rollout, covering:

1. **Role of AI:** The activities AI will execute, recommend or materially assist.
2. **Impact on the employee:** The activities that will reduce, expand, change or emerge.
3. **Human responsibilities:** The judgements, decisions, approvals, exceptions and relationships that remain human-led.
4. **Responsible-use boundaries:** The approved tools, permitted data, prohibited activities and mandatory reviews.
5. **Accountability and escalation:** Who owns the outcome and what happens when an AI output is incorrect or uncertain.
6. **Performance and progression:** How KPIs, capability expectations and career opportunities will evolve.

2) Redesign Work Before Reskilling the Workforce

Consideration: AI capability programmes will not realise their full value if employees are trained to use AI while roles, workflows and performance measures remain unchanged.

In BFSI, the unit of transformation is the role and the workflow, not the AI tool. Providing sales managers, operations executives or service representatives with AI capabilities creates value only when routine activities are reduced, responsibilities are rebalanced and released capacity is directed towards higher-value work. This requires activities to be classified into four modes:

- Human-led
- AI-assisted
- AI-executed with human review
- AI-executed within defined policies

This helps institutions place AI where it adds value while preserving human judgement and accountability where these matter most.

Practice: Redesign the role before designing the training

For each AI-enabled workflow:

1. **Decompose the role:** Map activities at the task level and classify them across the four human-AI work modes.
2. **Design the future workflow:** Create a before-and-after view that removes redundant steps, reports, hand-offs and controls instead of layering AI onto the existing process.
3. **Define the capacity dividend:** Determine whether released capacity will support wider customer coverage, deeper advisory, faster exception handling, better quality or stronger controls.
4. **Rebase performance measures:** Align KPIs with the new workflow, balancing productivity with quality, customer outcomes and judgement.
5. **Pilot the redesigned role:** Evaluate the full operating model, including cycle time, rework, quality, workload, capacity released and business outcomes.
6. **Build training around the new role:** Finalise the curriculum only after defining how employees will work and the capabilities they will need.

Training employees to use AI without redesigning the underlying work risks making an inefficient process move faster. The objective should be to transform the work, not simply accelerate it.

3) Turn Managers into the Frontline of AI Adoption

Consideration: AI adoption accelerates when managers become credible users, role models and translators of how the technology applies to everyday work.

Gallup and NHRDN's study of Indian workplace culture identifies performance management (45%) and responsiveness to disruption (44%) among the key challenges facing senior HR leaders.²⁰ The implication for AI is clear: employees are influenced more by what managers practise and reinforce than by enterprise-wide communication alone. When managers continue to rely on legacy processes, employees have little incentive to adopt new ways of working. When managers actively use AI, demonstrate responsible application and reinforce the right behaviours, adoption becomes part of how the team operates.

Practice: Activate managers before their teams.

1. **Put managers one learning cycle ahead:** before a role is trained, its managers use the capability on real work and pass the same assessment.
2. **Give every manager a role-specific adoption brief:** what changes, which behaviours to encourage, what good AI-assisted work looks like, when to escalate.
3. **Run short workflow reviews for the first six to eight weeks,** covering one good use, one poor output, one lesson, turning tacit experiments into shared capability.
4. **Change what managers inspect.** If supervisors keep inspecting manual activity, employees optimise for the supervisor. Manager scorecards should include safe AI workflow penetration, quality, certification and outcomes.
5. **Create business-led practice leads** at branch, regional or functional level: credible practitioners in the actual role, not generic champions, backed by the central AI, risk and learning teams.
6. **Measure manager adoption before interpreting employee adoption.** A low-usage team under a non-using manager is a leadership signal, not an attitude problem.

4) Build Role-Based Proficiency, Not Generic AI Literacy

Consideration: India's AI skills challenge is shifting from building awareness to enabling employees to apply AI effectively and responsibly within their roles.

EY research finds that 87% of employees and 90% of employers view skill development as important, yet structured AI learning remains limited²¹. It also links deeper AI capability with greater time savings and lower attrition intent. AI capability therefore needs to extend beyond tool familiarity or prompt writing. In a regulated environment, employees must know how to apply AI within their workflows, validate outputs against policies and trusted sources, exercise human judgement and escalate issues appropriately.

Role-specific proficiency should therefore be the measure of capability, not the completion of a standard training programme. The required level of proficiency should reflect the employee's role, the way AI is used and the consequences of an incorrect output or action.

Practice: Build learning paths around four capability levels:

1. **Level 1, AI-aware:** understands the approved tools, data rules, limitations, security and the institution's human-AI principles.
2. **Level 2, AI-enabled practitioner:** uses AI effectively in the role's workflows, verifies outputs, spots common failures.
3. **Level 3, AI reviewer/supervisor:** evaluates AI-assisted work, handles exceptions, interprets uncertainty, coaches others.
4. **Level 4, AI workflow owner:** redesigns processes around human-AI collaboration, defines controls and quality measures, improves the system with technology teams.

²⁰ Gallup, The Culture Imperative: Insights for India's Next Chapter, 2025

²¹ EY, Work Reimagined Survey, 2025

For each material role, capability building should combine four elements:

- **Foundation:** Approved use, data, risk, security and accountability
- **Role-based practice:** Application to real or safely simulated work scenarios
- **Failure-based learning:** Identification, correction and escalation of deliberately flawed AI outputs
- **Observed application:** Demonstration of proficiency in the flow of work before independent use

Capability should be demonstrated through application, not attendance. Higher-consequence roles require deeper scenario-based learning, formal assessment and appropriate certification before production access is granted. Programmes should be jointly designed by business, AI, risk and compliance teams, with L&D orchestrating the learning journey. For multilingual frontline roles, training and job aids should also be available in the languages in which employees perform their work.

5) Redesign Career Pathways for an AI-enabled Workforce

Consideration: As AI takes on routine and entry-level activities, organisations must preserve the experiences through which employees develop domain expertise, professional judgement and readiness for more complex roles.

The long-term workforce challenge is not simply the automation of junior tasks, but the potential erosion of the traditional talent pipeline. First drafts, routine analysis and repetitive processing have historically provided employees with the foundational experience required to build expertise. As AI increasingly performs this work, organisations will need to create new pathways through which employees can learn, practise and progress.

This is particularly important in BFSI, where judgement is developed through repeated exposure to customers, cases, exceptions and consequential decisions. AI-enabled work must therefore be designed not only for immediate productivity, but also for long-term capability and career progression. For India's young workforce, the opportunity is to create credible pathways into emerging roles that combine domain expertise, human judgement and the ability to work effectively with AI.

Practice: Redesign the career architecture alongside the role.

1. **Create skill-adjacency maps for affected role families:** operations executive to exception specialist; service representative to complex-resolution specialist; analyst to AI-assisted reviewer; relationship manager to AI-augmented advisor; control analyst to AI assurance specialist.
2. **Establish a redeployment-first mechanism:** identify affected people early, map adjacent roles and certify them before defaulting to external hiring.
3. **Replace lost apprenticeship work deliberately** through curated exception libraries, supervised review of AI outputs, simulation, customer shadowing and progressively harder cases.
4. **Change the promotion criteria** to reward domain expertise, judgement, the ability to supervise AI and workflow improvement, not manual throughput.
5. **Link learning to mobility:** employees should be able to see which capabilities make them eligible for which next role. A certificate tied to a real transition is capability building; one that is not is just training.
6. **Track internal mobility from affected roles.** The board question is not how many were trained, but how many in materially changing roles gained an adjacent capability and moved into sustainable work.

6) Give Humans Explicit Decision Rights and Measure Value, Not Usage

Consideration: Employees can use AI confidently in consequential workflows only when the boundaries between AI authority and human accountability are clearly defined.

More than 55% of organisations cite incomplete or low-quality AI outputs as a challenge, according to NASSCOM's 2026 research²². In BFSI, employees therefore need clarity on when AI outputs can be acted upon, when they must be reviewed or challenged, and who owns the final decision. The ability to override AI is both a governance safeguard and an enabler of responsible adoption.

Oversight should reflect the consequence of the decision. Low-risk use cases may require limited human intervention, while decisions involving underwriting, claims, credit and compliance should remain accountable to a named human decision-maker. AI may inform or support these decisions, but it should not dilute human ownership of the outcome.

Practice: Establish a Human Decision Rights Protocol

For every material AI-enabled workflow, define:

1. **AI authority:** What the system may retrieve, generate, recommend or execute.
2. **Human authority:** What a named individual must review, approve or decide.
3. **Mandatory stop conditions:** When an output must not be acted upon without escalation.
4. **Challenge and override rights:** When employees may correct, reject or set aside an AI output.
5. **Escalation ownership:** Who resolves uncertainty or disagreement between human judgement and AI recommendations.
6. **Decision record:** What must be documented when an output is overridden or escalated.

Employees should be able to challenge AI outputs when they identify factual errors, missing context, policy concerns, privacy risks or insufficient confidence to act. Patterns in overrides and escalations should be treated as inputs for improvement, helping identify issues in models, data or processes.

AI programmes should be measured on workforce capability, adoption, quality, productivity and business impact. Usage metrics may indicate activity, but value is ultimately reflected in better outcomes, stronger controls and improved performance.

²² NASSCOM, WORK REIMAGINED: The Rise of Human-AI Collaboration, 2026





Human oversight

Cost operations

Observability

Responsible AI at Scale: Governance, Cost and Observability

SECTION 5

Responsible AI at Scale: Governance, Cost and Observability

Responsible AI is built from day one: accountable by design, observable in operation, secure by default and economically viable at scale.

As financial institutions move from AI experimentation to enterprise-scale deployment, governance can no longer be treated as a compliance activity performed after implementation. AI systems must be designed from the outset with appropriate controls for risk, accountability, transparency, security and monitoring. Nearly 58% of organizations acknowledge that they do not have a well-defined AI and data governance foundation, despite planning widespread AI adoption, highlighting a significant gap between AI ambition and governance readiness²³. Organizations that embed governance, accountability and trust mechanisms into AI design and deployment are better positioned to scale AI confidently, meet regulatory expectations and avoid costly remediation efforts. Building these trust mechanisms requires explainable decisions, human oversight, strong security controls, transparent monitoring and continuous quality management. AI initiatives that cannot demonstrate reliability, traceability and governance are unlikely to progress beyond pilots, regardless of their technical sophistication.

As organizations deploy AI across the value chain, traditional governance models become insufficient. AI introduces new challenges like cost management, model drift, autonomous decision-making, data access. To realize sustainable value, organizations must establish an enterprise operating model setup around 4

key pillars - regulatory compliance & risk management, cost & ROI management, observability & traceability, and information security. The below section deep dives into these 4 pillars and outlines key considerations and practices for each one of them.

1. Regulatory Compliance & Risk Management: Building to converging principles

Compliant AI with risk management frameworks must be embedded before deployment, not retrofitted after regulatory, customer or operational issues emerge.

As AI adoption accelerates across financial services, establishing robust risk controls, compliance and ethical guardrails is becoming as essential as the business case itself. Regulators are converging around four core principles: human oversight, decision-level traceability, board-level accountability and data protection by design. This is reflected in India's evolving regulatory landscape, including RBI's FREE-AI framework, which emphasizes board-approved AI policies, AI inventories, risk-based audits, human oversight and incident reporting across the AI lifecycle²⁴. SEBI's AI regulations also place accountability on financial institutions for AI-enabled decisions, irrespective of whether the solution is built internally or procured from a third party²⁵.

²³ IBM, Go further, faster with AI, 2025

²⁴ RBI's Free AI Framework

²⁵ SEBI, Regulation 16C, 10th Feb 2025

The business case for stronger governance is equally clear. 77% of organizations report that AI adoption is outpacing their governance capabilities²⁶, increasing exposure to compliance, operational and reputational risks. For example, in an insurance claims triage use case, an AI agent may summarize claim documents, flag missing evidence and recommend next actions, but final repudiation or approval should remain with a human claims assessor to ensure legal defensibility, customer fairness and regulatory compliance.

Consideration

Establish a risk-based AI governance framework that scales with the materiality of the use case, rather than applying uniform controls across all deployments. AI use cases should be classified based on customer impact, decision criticality, regulatory sensitivity, data sensitivity and degree of autonomy, with governance requirements increasing accordingly. Establishing these governance mechanisms early enables financial institutions to scale AI confidently while maintaining regulatory compliance, customer fairness, operational resilience and defensible decision-making.

Key practices

1. **Low-risk use case prioritisation:** Begin with low-risk, non-customer decisioning use cases such as sales coaching, employee copilots or customer engagement support, allowing governance controls to mature before applying AI to high-impact customer decisions.
2. **Human-in-the-loop governance:** Place human review on the decision rather than the task - a person approves the consequential output, while the agent performs the work leading to it.
3. **Executive accountability & ownership:** As organizations scale AI adoption, establishing clear executive and board-level accountability for material AI use cases may become increasingly important. Evolving governance models are expected to define ownership for outcomes, compliance, risk management and ongoing oversight throughout the AI lifecycle.

4. **Data protection by design:** Classify, label and control the data an AI system can access before deployment, ensuring privacy and data protection are embedded into solution design rather than added later.
5. **AI inventory & lifecycle governance:** Maintain a centralized AI inventory from the first pilot, capturing purpose, scope, data used, risks, controls, owners, third-party model details and lifecycle status for every AI use case.
6. **Third-party AI risk management:** Extend procurement and vendor governance to AI embedded within third-party products, including AI-specific disclosures, model change notifications and ongoing vendor compliance monitoring.

2. Cost & ROI Management: Establishing the ceiling before the pilot

AI cost must be governed before the first pilot starts. Waiting for the first invoice is too late.

AI value is determined not only by model performance, but by whether the organization can control cost, define unit economics and measure business outcomes. In agentic AI, the headline model price is rarely the only cost driver. Material costs can arise from unbounded token consumption, retry loops, repeated tool calls, agents running without sufficient workload and poor model routing. AI costs can increase significantly as solutions become more complex. Without upfront cost controls, a pilot can quickly become far more expensive than originally anticipated, highlighting the importance of disciplined AI FinOps, model selection and usage monitoring.

Consideration

Treat AI cost governance as a business discipline rather than a technology measure. Before launching a pilot, organizations should establish clear unit economics and define the expected business outcome for the use case. Cost controls should be embedded into the solution architecture through spend thresholds, model-routing

²⁶ IBM, New IBM Study Finds CIOs and CTOs Face Growing AI Control Gap as Enterprise Deployment Scales, 2026

policies, retry limits and kill-switch mechanisms, with ongoing monitoring. This ensures that AI investments remain economically viable as usage scales, while enabling management teams to continuously assess whether the value generated justifies the cost incurred.

Key practices

1. **Unit economics definition:** Define expected cost per conversation, lead, claim, service request or renewal before technology selection, creating a baseline for ROI measurement.
2. **Spend governance & control thresholds:** Set spend ceilings, usage thresholds and explicit failure conditions upfront, ideally enforced through an AI gateway rather than only through application logic.
3. **Agent loop & retry management:** Cap agent iterations, retries and recursive reasoning loops to prevent unexplained token consumption and uncontrolled expenditure.
4. **Intelligent caching strategy:** Apply prompt, context and response caching based on traffic type to reduce repeated processing and improve cost efficiency.
5. **Right-sized model routing:** Route low-complexity tasks to smaller, cost-efficient models and reserve premium models for tasks that genuinely require advanced reasoning.
6. **Enterprise kill switch controls:** Implement operator-accessible kill switches to suspend AI agents in the event of cost spikes, policy breaches or operational failures.
7. **Journey-level cost attribution:** Attribute AI costs by customer journey or use case rather than department, making unprofitable journeys visible for redesign or termination.
8. **Agent retirement & resource optimisation:** Retire agents that no longer carry meaningful workload and revoke associated identity, system access and tool permissions.

3. Observability & Traceability: Monitoring as the source of the audit trail

Effective observability ensures that audit trails are continuously created, rather than reconstructed after an issue occurs.

Observability and traceability are critical for understanding what an AI system is doing, why it produced a particular output and whether the system continues to operate within defined thresholds. Observability provides real-time visibility into AI behaviour, model calls, tool usage, latency, reliability and cost, while traceability enables reconstruction of decisions, source data, handoffs and internal system states for audit, compliance, incident management and dispute resolution. As AI adoption accelerates, the need for these capabilities is becoming increasingly important, with 70% of technology leaders reporting that AI and technology are being deployed faster than IT teams can effectively track and govern²⁷.

Consideration

Instrument observability and traceability from the first version of every AI use case, ensuring AI actions, decisions, tool calls and outputs are continuously monitored, re-playable and audit ready.

Key practices

1. **Baseline process measurement:** Measure the current manual process before automation, establishing baselines for quality, turnaround time, accuracy, cost and exception rates.
2. **End-to-end traceability:** Capture model invocations, tool calls, prompts, data sources, handoffs and outputs so each AI-driven decision can be replayed and audited.
3. **Open telemetry & portable logs:** Emit telemetry using open standards such as OpenTelemetry where possible, ensuring monitoring data remains portable across platforms and tools.

²⁷ IBM, New IBM Study Finds CIOs and CTOs Face Growing AI Control Gap as Enterprise Deployment Scales, 2026

4. **Continuous quality evaluation:** Evaluate a representative sample of live traffic using consistent evaluator definitions across development, testing and production environments.
5. **Threshold-based monitoring & response:** Define thresholds for quality, latency, hallucination, cost, policy breach and tool-call failure, with pre-agreed actions such as retrain, roll back, restrict or terminate.
6. **Automated detection & governance:** Automate anomaly detection, policy violation alerts and performance degradation monitoring, while keeping human judgement for approval, interpretation and accountability.

4. Information security: The agent as a privileged internal user

AI agents should be treated like privileged digital workers, with identity, access, monitoring and revocation controls from day one.

AI systems often operate with access to sensitive customer, employee, policy, claims, financial and enterprise data. This makes security, trust and data privacy foundational capabilities rather than technical add-ons. Every AI agent should operate under a managed identity, with least-privilege access, within a single registry and subject to the same identity, access and threat controls applied to privileged human users. With 97% of organizations experiencing AI-related breaches reporting inadequate AI access controls²⁸, these safeguards become even more important in multi-agent systems, where different agents may access different tools, databases, systems or external sources.

Consideration

Ensure every AI agent operates within a secure, privacy-first framework with managed identity, least-privilege access, centralized registry, content safeguards and clear delegation boundaries.

Key practices

1. **Managed agent identity:** Assign a unique managed identity to every AI agent before it runs, enabling traceability, access scoping, monitoring and revocation.
2. **Least-privilege access control:** Grant AI agents only the minimum data, system and tool access required for their designated tasks, reducing the risk of overexposure.
3. **AI security guardrails & content protection:** Apply prompt injection defences, content shields, input-output screening and secure communication controls across agent interactions.
4. **Adversarial testing & red teaming:** Run adversarial red-team scans during build and release cycles to identify vulnerabilities, misuse scenarios and model manipulation risks before deployment.
5. **Delegation boundaries & human authority:** Define activities that must never be delegated to AI, such as underwriting decisions, claims approvals, funds movement, adverse customer actions or external communication without review.
6. **Centralized agent registry & access governance:** Maintain a single registry of internal, third-party and shadow AI agents, with periodic permission reviews and access pruning.

²⁸ IBM Report, 13% Of Organizations Reported Breaches of AI Models or Applications, 97% of Which Reported Lacking Proper AI Access Controls, 2025





06

Data foundations

Architecture

Security

The Foundations Enabling AI at Scale

SECTION 6

The Foundations Enabling AI at Scale

The ability to scale Agentic AI will increasingly depend on the enterprise foundations beneath it: open standards that enable technology choice, and trusted data and connected workflows that enable business value.

This section examines two developments reshaping adoption in financial services: the emergence of interoperability standards that improve platform portability, and the growing importance of data readiness and workflow integration in determining business outcomes. Together, these developments have significant implications for how financial institutions select technology, design operating models and govern AI at scale. The section translates these shifts into practical considerations for building scalable, resilient and well-governed AI capabilities.

1) Interoperability standards and platform portability

Until recently, selecting an agent platform often meant committing to its broader technology ecosystem, including its tool integrations, agent communication mechanisms and telemetry standards. Changing platforms could therefore be costly and complex. As interoperability standards mature, financial institutions have greater flexibility to adopt AI without making an irreversible technology commitment.

Three developments are driving this shift:

1. Tool integration has largely converged around the Model Context Protocol (MCP). Anthropic transferred MCP to the Linux Foundation in December 2025 under the Agentic AI Foundation, supported by organizations including OpenAI, Google, Microsoft, AWS and Bloomberg. The protocol is therefore no longer controlled by a single vendor.

2. Agent2Agent (A2A) has emerged as the leading standard for agent-to-agent communication after being contributed by Google to the Linux Foundation, with broad industry backing and the subsequent merger of IBM's Agent Communication Protocol into A2A.
3. The growing availability of open-weight models and multi-model architectures has reduced dependence on any single model provider, making model selection an increasingly use-case-specific procurement decision rather than a long-term strategic commitment.

This has practical implications for financial institutions tools, models and collaborating agents are becoming increasingly portable across platforms, reducing vendor lock-in and concentration risk. As a result, the key barriers to adoption are shifting from technology availability to governance, security and implementation readiness. While some elements, such as guardrails, agent identity and observability standards, continue to evolve, the components that most directly influence switching costs have largely begun to standardize.

This flexibility also extends to deployment choices. Whether AI solutions are deployed in public cloud, data-resident environments or on-premise infrastructure, standardized interfaces make it possible to maintain a consistent architecture while aligning deployment decisions to data sensitivity, regulatory requirements and risk appetite. Adoption no longer requires a permanent technology commitment; it increasingly

becomes a business and risk decision governed by the needs of each use case.

Key Consideration and Practices

Ensure that agent architectures are designed for platform portability by standardising on open protocols at the tool, inter-agent, and telemetry layers.

1. **Standardise on Open Integration Protocols:** Require that tool integrations implement Model Context Protocol (MCP) and that inter-agent communication implements Agent2Agent Protocol (A2A), so that the orchestration platform remains a replaceable component.
2. **Implement Portable Telemetry Standards:** Emit observability telemetry on the OpenTelemetry semantic conventions for generative AI, accepting that these conventions remain in development, in preference to proprietary formats without a migration path.
3. **Plan for Vendor Exit and Concentration Risk:** Define exit arrangements and assess provider concentration risk before contracting for an agent platform, consistent with the expectations set out in RBI's draft Guidance on Regulatory Principles for Model Risk Management.
4. **Adopt Risk-Based Deployment Placement:** Determine the deployment environment for each AI use case - public cloud, data-resident region, or on-premise on the basis of the data classification of that use case, rather than as a single enterprise-wide architectural commitment.

2) AI will scale only as fast as its data and workflow foundations

Even a high-performing model can produce unreliable outcomes when customer, product or policy information is fragmented, inconsistently defined or disconnected across systems. As financial institutions embed AI across complex journeys, data readiness, semantic consistency and workflow integration become critical to ensuring that information and context move reliably across each stage.

This requires institutions to treat data readiness as a continuing operational capability, not a one-time technology prerequisite. Agentic AI demands ongoing oversight of how agents access, interpret and use information across the enterprise, alongside integration into the processes and systems where work is executed. The ability to scale AI will therefore depend as much on strengthening these foundations as on advancing the models themselves.

Key Consideration and Practices

Ensure that the data foundation for each agentic AI use case is governed as a first-class asset, including the retrieval layer, dynamic data access oversight, and the scoping of data sources to the requirements of the use case.

1. **Treat Knowledge Access as a Managed Capability:** Manage retrieval as a governed asset rather than a configuration setting. Version retrieval settings and citation requirements together with the index, and monitor retrieval health stale content, missing documents, decayed relevance as a live operational metric rather than a launch-time check.
2. **Enable Continuous Data Governance:** Uplift data governance from static, point-in-time review to real-time oversight of agent data access and use, proportionate to the risk materiality of the use case.
3. **Control Data Access by Design:** Scope the data before scoping the agent. Restrict each agent's access to the data sources its use case requires. This data-layer control provides much of the risk mitigation commonly attributed to runtime guardrails.
4. **Build a Shared Business Understanding Incrementally:** Build the business ontology incrementally along the path of deployed use cases, beginning with the single use case selected for the first deployment, rather than as a standalone multi-year enterprise data programme.



Three horizons

Cost discipline

Optimisation

The AI Playbook: From Ambition to Execution

SECTION 7

The AI Playbook: From Ambition to Execution

Start with the business outcome, not the technology: name the P&L line, set the boundaries and build the evidence before scaling the pattern.

This section focuses on the decisions required to move agentic AI from pilot to production. While many organisations concentrate on implementation activities, the success of AI programmes is often determined by the timing and quality of foundational decisions around ownership, governance, risk, controls and scale. The objective is therefore not to prescribe a project plan, but to provide a practical path for making the right decisions at the right stage of adoption. The approach is aligned to the principles of the NIST AI Risk Management Framework, helping financial institutions embed governance, risk assessment, measurement and ongoing management as AI adoption matures²⁹.

The roadmap is organised across three horizons: the first ninety days, the first six months and the first year. Together, these horizons describe how financial institutions can progress from establishing strong foundations to operationalising controls and ultimately scaling AI in a governed, resilient and sustainable manner.

Horizon 1 - The First Ninety Days: Decide, Baseline, Bound

The objective of this horizon is not a working agent. It is a defensible set of decisions and a measured baseline. A team that reaches day ninety with no agent but with both of these is ahead of one that reaches day ninety with a demonstration and neither.

Key consideration and practices

Ensure that the first ninety days produce a named journey, a measured baseline, bounded costs and failure conditions, and board-level accountability before any material build commences.

1. **Prioritize a High-Impact Journey:** Select a single end-to-end customer journey with a clear business owner, measurable outcome and low regulatory risk for initial deployment.
2. **Establish the Baseline:** Measure the current process, including effort, turnaround time, cost and error rates, to create a clear benchmark for evaluating impact.
3. **Define Economic and Risk Guardrails:** Set upfront limits on spend, iteration cycles and acceptable performance thresholds, along with predefined conditions for stopping the pilot.
4. **Assign Executive Accountability:** Establish named executive ownership and board-level oversight before deployment to ensure clear accountability for outcomes and risks
5. **Define Non-Negotiable Boundaries:** Create an explicit list of actions AI agents cannot perform, particularly those involving customer-impacting decisions, funds movement or regulatory obligations

²⁹ NIST, Artificial Intelligence Risk Management Framework, 2023

Completion criteria at day ninety

- One journey named, one P&L line named, one business owner named.
- A manual baseline: unit cost, cycle time, error rate, volume.
- A spend ceiling, an iteration cap and written failure conditions, agreed before the build starts.
- Every agent that will exist has a managed identity, an entry in a single registry, and least-privilege scoped access from day one, not from the security review.
- Tracing and baseline evaluations enabled in the first version, emitting on open standards.
- A named board-level owner and a written list of prohibited actions.

Horizon 2 - The First Six Months: Production Operation with Live Controls

The objective of this horizon is evidence that could be presented to a regulator and a CFO in the same meeting. The pilot must therefore run against real traffic, at small volume, with the controls live rather than planned

Key consideration and practices

Ensure that the pilot operates in production with human review placed on the decision, monitored thresholds with pre-agreed responses, line management carrying the change, and a data layer scoped to the journey.

1. **Keep Humans Accountable for Decisions:** Position human review at the decision point rather than every task, using risk-based sampling where full review is impractical
2. **Enable Controlled Overrides:** Provide reviewers with a defined ability to challenge or override AI outputs, with all interventions documented and tracked
3. **Monitor and Respond in Real Time:** Establish performance, risk and cost thresholds with predefined actions for breaches, supported by continuous monitoring and evaluation.

4. **Drive Adoption Through Line Managers:** Make branch and regional managers accountable for adoption, and measure leadership engagement alongside team usage.
5. **Build the Data Foundation Incrementally:** Develop the data definitions, knowledge assets and retrieval controls needed for the selected journey, and monitor them as operational assets.

Completion criteria at month six

- The agent has run against live traffic long enough to have a drift record, not merely a launch record.
- Unit economics per journey are reported alongside quality one dashboard, not two.
- At least one threshold has breached and the pre-agreed response has been executed. Thresholds that have never breached are decorative.
- A red/amber/green view exists that an executive can read and an auditor can trace back to individual decisions.
- The change was carried by line management, with adoption demonstrable by manager cohort.

Horizon 3 - The First Year: Scaling the Pattern, Not the Pilot

The objective of this horizon is a repeatable pattern and a governed fleet not a larger version of the first agent.

Key consideration and practices

Ensure that scaling reproduces the reusable pattern established by the first deployment its controls, architecture and gates rather than cloning the pilot, with central governance of how agents are governed and local freedom in how they are built.

1. **Scale the Pattern, Not the Use Case:** Document and standardize the reusable components of the first deployment, including governance, controls, monitoring and deployment practices, so they can be applied consistently across future journeys

2. **Balance Central Governance with Local Innovation:** Establish enterprise-wide standards for governance, security and oversight, while allowing business teams flexibility in how AI solutions are designed and deployed.
3. **Govern the Full AI Lifecycle:** Implement formal processes for reviewing, retiring and decommissioning AI agents, ensuring access, permissions and risks are removed when a use case is no longer required.
4. **Maintain Continuous Audit Readiness:** Maintain an up-to-date inventory, decision audit trail, customer disclosure records and validation evidence so regulatory and governance requirements can be met at any time

Completion criteria at month twelve

- A written internal pattern, and a second and third journey delivered against it materially faster than the first.
- A single registry covering every agent, including third-party and shadow agents.
- Cost attributed by customer journey, with at least one journey terminated or redesigned because its unit economics did not hold.
- A retirement schedule that has retired at least one agent.
- Governance artefacts produced by running the system, not assembled for an audit.

Across all three horizons, the emphasis is on bringing critical decisions forward and delaying technology choices until they are necessary. That sequencing is deliberate. Institutions that struggle are rarely those that selected the wrong platform, particularly in a market where growing interoperability has reduced the cost of switching. More often, challenges arise when organisations choose technology before defining the business journey, automate processes they have not measured, discover costs after deployment, or attempt to establish governance and auditability only

after questions are raised. These are fundamentally sequencing issues, and they are largely avoidable when the right foundations are established in the first ninety days.

The foundations of sustainable scaling

Deploying an agent is only the beginning; scaling it sustainably requires disciplined management of both cost and performance. Cost operations ensure AI investments remain economically viable as usage grows, while optimisation turns operational signals into measurable improvements in quality, reliability and efficiency. Together, they provide the mechanisms required to move from successful pilots to a scalable, governed AI operating model.

1. Cost Operations

Agentic AI programmes rarely struggle because of model capability alone; they struggle when adoption outpaces governance and cost visibility. While the core equation remains unchanged, $\text{Spend} = \text{Rate} \times \text{Usage}$, AI shifts both levers from procurement into architecture and design decisions such as model selection, prompting, orchestration and usage controls. As a result, many traditional FinOps principles still apply, but AI introduces new cost drivers that must be governed from the outset rather than addressed after deployment.

- **From infrastructure-centric cost management** → To token-centric cost management, where spend is driven by model usage, prompt design and token consumption.
- **From automatic cost allocation** → To engineered cost visibility, requiring deliberate tagging, monitoring and attribution mechanisms to track AI spend and value.
- **From periodic cost governance** → To continuous cost governance, as AI adoption can outpace traditional review cycles and requires ongoing oversight.
- **From cost optimization alone** → To cost-quality optimization, balancing model performance, accuracy and business outcomes alongside cost efficiency.

Key consideration and practices

Ensure cost is governed as an architectural property from the first horizon, with a single controlled entry point for model calls, clear cross-functional ownership, defined unit economics and a deliberate tagging strategy, rather than as a retrospective finance exercise once the AI estate has been established.

1. **Establish Cross-Functional Cost Ownership:** Form a cross-functional operating group covering business, engineering, data science, risk, finance and procurement, aligned on the trade-offs between performance, accuracy and cost before the first pilot begins.
2. **Start with Cost Transparency:** Begin with showback rather than chargeback. Making consumption visible to teams drives accountability, encourages responsible usage and helps identify unused or underutilised resources early.

3. **Define Cost Governance Rules:** Establish clear cost thresholds, performance expectations and compliance requirements, with named owners accountable for monitoring spend, forecasting budgets and optimizing deployments.
4. **Review Cost on a Fixed Cadence:** Conduct regular governance reviews covering budgets, model-selection trade-offs and optimization opportunities, rather than responding only when costs become a concern.
5. **Centralise Model Access Through a Gateway:** Route all model calls through a single gateway to enable consistent governance, cost visibility, access control and usage monitoring across use cases.

Without an AI gateway	With an AI gateway
A separate AI service and endpoint per use case, each holding its own master keys.	One mediated entry point, with granular access control and managed-identity or OAuth authorisation.
Uncontrolled token usage, with no ceiling per application or team.	Token rate limits and quotas per-minute ceilings and hourly-to-yearly quotas per app, team or department.
Opaque consumption that cannot be attributed to anyone.	Token metrics per consumer and usage reports per use case and team, enabling clean showback or chargeback.
No record of what was asked or what was answered.	Prompt and completion logging streamed out for billing, audit and compliance review.
Governance re-implemented for every new model or provider.	A unified model API many providers behind one endpoint, with the policy applied once.
Inconsistent performance and repeated spend on near-identical prompts.	Semantic caching that reuses completions for similar prompts, plus priority routing and circuit breakers.
6. Baseline Before Committing Spend: Assess both current costs and performance requirements before selecting a pricing approach, ensuring commercial	decisions align with expected usage patterns and business outcomes.

Pricing model	Best for	Watch-out
On-demand / pay-as-you-go	Variable, unpredictable workloads batch inferencing, ad hoc training	Needs vigilant monitoring of token and API-call volume
Reserved and committed use	Predictable, GPU-heavy training or inferencing	Upfront planning required to avoid underutilisation
Provisioned capacity	Low-latency real-time applications recommendations, customer- facing agents	Does not scale with spikes; carries low-utilisation risk
Spot / batch	Interruptible batch, retraining and burst workloads	Requires robust scheduling to handle interruptions
Tiered pricing	Scalable services with steadily growing usage	Forecast carefully to avoid unexpected step changes
Freemium / trial	Experimentation during trial or preview	Costs jump when the preview ends; usage limits apply

7. **Estimate Economics Early:** Size the customer journey and estimate token consumption before development begins to establish an initial view of operating costs and economic viability.

8. **Implement Cost Attribution from Day One:** Apply a consistent tagging strategy across projects, teams
- and workloads from the outset to support accurate allocation, reporting and optimization of AI spend

9. **Control Spend Through Usage Guardrails:** Set token quotas, spending limits and automated alerts to prevent cost overruns, while using consumption insights to drive ongoing optimization and efficiency improvements.

Instrument these four measures from Day 1

Measure	How it is calculated	What it tells the practitioner
Cost per inference	Total inference cost ÷ inference requests $\$5,000 \div 100,000 \text{ requests} = \0.05 per request	Running efficiency of a live agent; surfaces cost spikes caused by inefficient code, orchestration or retry loops
Cost per token	Total cost ÷ tokens consumed $\$2,500 \div 1,000,000 \text{ tokens} = \0.0025 per token	Where prompt trimming and semantic caching will pay back, without degrading output quality

Measure	How it is calculated	What it tells the practitioner
Training cost efficiency	Training cost ÷ performance metric A 95%-accurate model trained at \$10,000 = \$105 per point of accuracy	Whether the next increment of accuracy is worth what it costs to buy
Time to business value	Total value of the service ÷ daily cost of the alternative Forecast \$100k a month within one month; actual took five months at \$50k a month	Forecast against actual days to breakeven; the gap the sponsor is judged on

Cost operations are complete for a horizon when every model call in the estate passes through the gateway with no use case holding its own provider keys; when cost per inference and cost per token are reported for each live journey on the same dashboard as its quality metrics; when the forecast-to-actual gap on time to business value has been measured at least once and explained; and when the cross-functional owner group has met on cadence, with cost decisions documented rather than escalated.

Treating cost as a finance exercise that begins after go-live is the common failure. Architecture sets both variables in $\text{Rate} \times \text{Usage} = \text{Spend}$ through model choice, prompt design, caching and quotas so an FI that instruments before scaling can defend the spend and the decisions behind it separately. One that scales first has to defend both at once, to an audience that has already seen the invoice.

2. Optimisation

Two decisions largely determine an agent's cost and reliability. The first is the workflow design: whether a single agent performs retrieval, reasoning, tool execution and response generation, or whether these activities are distributed across specialised agents. The second is optimisation through prompt design, model selection and configuration. The first decision typically has the greatest impact, as it determines the overall architecture within which optimisation occurs.

Decomposing complex workflows into smaller, specialised agents can improve both cost and performance. Costs are reduced because each task can be matched to the most appropriate model, avoiding the use of expensive models where they are not required. Reliability improves because narrowly scoped agents operate within clearer boundaries and are less likely to generate unsupported outputs. Optimisation then enhances each component through ongoing tuning and evaluation. Together, workflow design and optimisation are complementary disciplines: one defines the structure, while the other improves performance within it. Where implemented at ABCL, the former is enabled through Pints AI's Autothought and the latter through Microsoft Foundry's Optimiser, although the sequencing is more important than the specific tools used.

Optimisation completes the observability loop

Optimisation is not a discipline bolted on after an agent ships. It is the fourth pillar of agent observability, and it works only because the three pillars before it produce the evidence it acts on. Reliable agent development needs observability to evaluate performance, quality and safety, to monitor and debug issues quickly, and to optimise agent performance.

Expressed as a loop, the four pillars carry distinct responsibilities.

Pillar	Role in the loop
Trace	End-to-end telemetry that gives execution flow visibility
Evaluate	Quality and safety scoring against defined evaluators
Monitor	Real-time issue detection across live traffic
Optimize	Continuous performance tuning to improve agent performance

Optimisation is continuous, not a one-time tuning pass

Optimisation runs continuously because agents are non-deterministic. Their behaviour changes over time, sometimes unexpectedly, even when nothing in the code has changed: model updates, prompt revisions, retrieval drift and long-tail traffic all shift live outputs. In production, monitoring and optimisation are not one-time tasks but ongoing processes.

The goal is not a single tuned configuration but a standing capability to detect regression, explain it and correct it, while balancing impact, quality, safety and cost.

Evidence-backed optimisation with the Agent optimiser

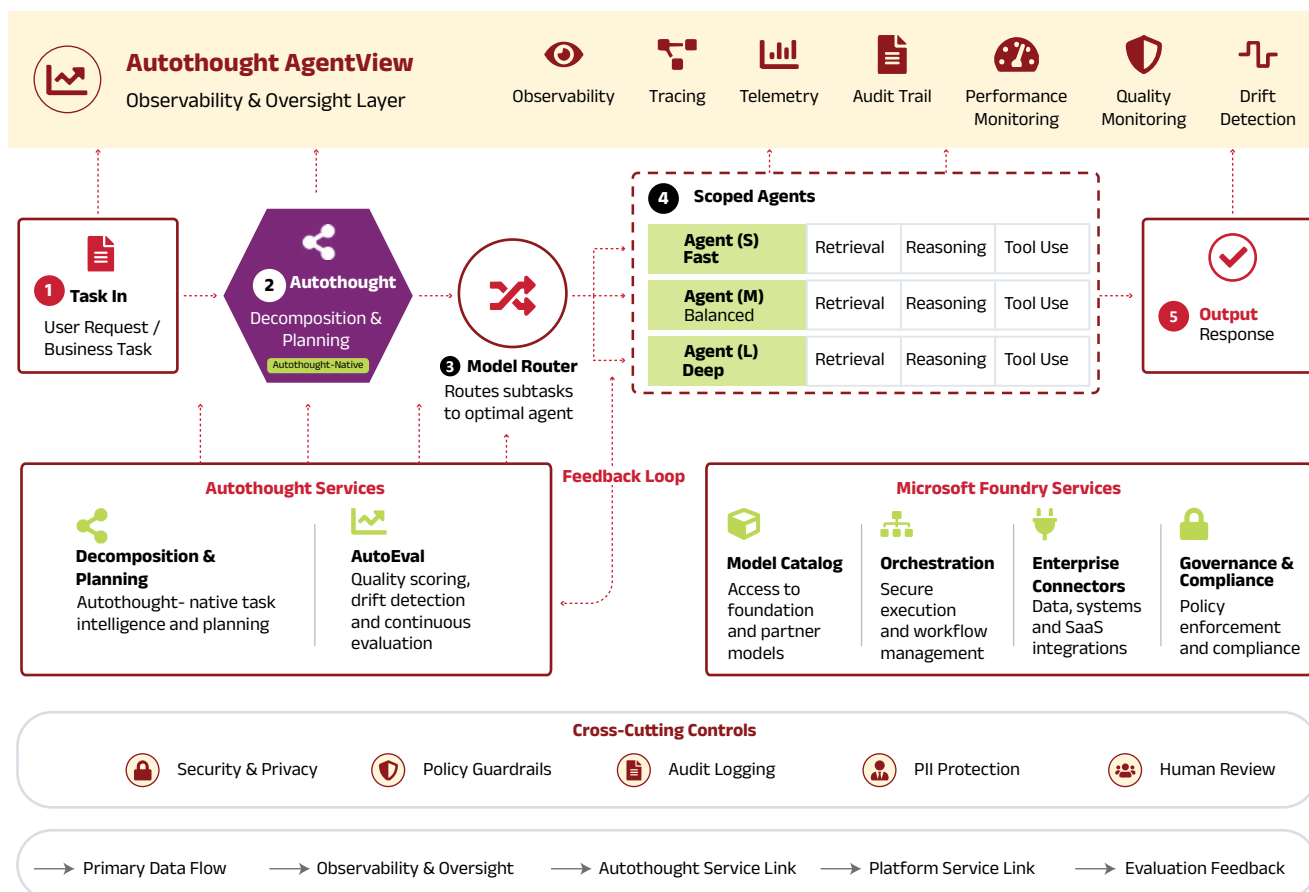
An agent optimiser is the clearest expression of the approach: it improves agents in production through governed, evidence-backed optimisation. Rather than asking an engineer to guess at a better prompt, it reads the agent's prompts and skills and searches for configurations that increase quality.

Each candidate carries measured quality and cost deltas, so the trade-off is visible before promotion.

Candidate	Configuration	Quality	Cost
Candidate 1	Prompt v4 · smaller model · 2 skills	+26%	-36%
Candidate 2	Prompt v5 · larger model · 4 skills	+16%	-10%

Illustrative candidate output, as presented in the Agent optimiser.

Both words matter. Evidence-backed means every proposed change arrives with evaluation results attached. Governed means the system proposes and a person disposes: candidates are surfaced for review, not applied automatically.



The optimisation toolkit

A control plane adds utilities for looking deeper into agent behaviour and driving the next change. The capabilities below are those an FI should expect of any platform it selects.

Capability	What it does
Evaluation comparison	Compares evaluation runs to confirm a change is a genuine improvement
Evals cluster analysis	Groups issues by type and nature to pinpoint failure patterns quickly
Chat with your data	Queries Foundry for insights and control over the project
Curated model upgrades	Moves to newer models through a curated, controlled upgrade path
Prompt optimiser	Tunes prompts while monitoring and remediating in production

Optimisation runs across the lifecycle, then across the fleet

The capability spans the workflow of plan, develop and operate. It begins in the agent playground with out-of-the-box evaluations and tracing for debugging, moves into CI/CD workflows as the agent transitions to code, and lands in production with fleet-wide visibility and control.

At production scale the signal should be unified across applications and agents, the AI platform, data and infrastructure, so that decisions are taken against the full picture rather than one layer of it.

Where the approach is heading

This direction extends the same evidence-then-decide pattern: version compared to set configurations side by side, rollouts and A/B testing to promote changes progressively, and actionable insights that recommend what to do next.

Optimisation stays inside the governance envelope

Because an optimisation changes what an agent says and does, it is treated as a release rather than a tweak. Sign-off on model and prompt changes stays in human hands, alongside the other judgement calls kept out of automation: triaging flagged interactions, interpreting fairness findings, and approving high-risk use cases.

The same model defines when a cycle begins. A reassessment trigger fires when an evaluation score drops below baseline, for example task adherence under 85%, or when a model, prompt or retrieval change is introduced. The response is to re-run evaluations and red team the agent, then re-approve before release.

The through-line is consistency. Optimisation here is not intuition applied to a live system; it is a measured loop in which traces supply the evidence, evaluations score it, monitoring detects when it moves, and every improvement is approved before it reaches users.

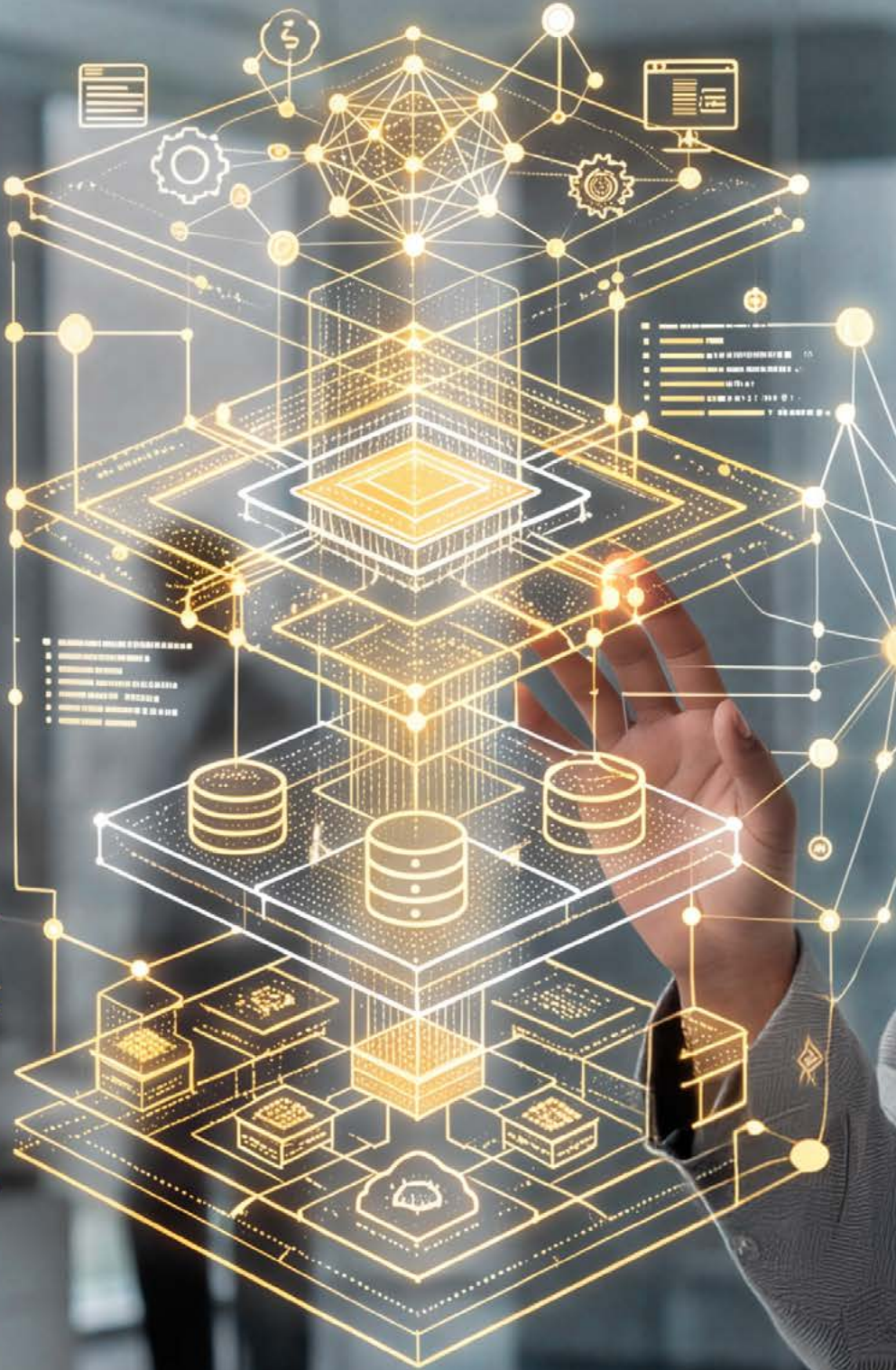


API/Services/
User Interface

Processing/
AI Models

Data Lake/
Storage

Infrastructure
(Cloud/On-Prem)



What holds

What is next

The measure

Conclusion

SECTION 8

Conclusion

The opportunity extends beyond large institutions

Aditya Birla Capital's experience demonstrates what becomes possible when AI is supported by enterprise data, technology, governance and workforce capabilities. Yet the lessons from this journey are not intended only for institutions with comparable resources. The larger opportunity lies in helping financial institutions of different sizes apply these lessons in ways that reflect their own readiness, risk profile and operating context.

This matters because India's financial system extends far beyond its largest institutions. The country has approximately 8,600 registered non-banking financial companies³⁰ and nearly 1,13,000 primary agricultural credit societies³¹, alongside a wide network of cooperative banks and other financial intermediaries. These institutions often hold relationships across the communities, occupations and income groups where the financial-participation gap remains most pronounced.

For many smaller institutions, however, the cost of meeting rising expectations across technology, compliance, customer service and risk management can constrain their ability to expand. Agentic AI has the potential to change what institutions can achieve with limited resources, extending operational capacity, improving customer engagement and strengthening employee productivity. The objective is not to replicate the capabilities of a large financial group, but to make sophisticated capabilities more accessible to institutions that play a critical role in financial inclusion.

The sequence matters more than the ambition

The central takeaway from this report is not that every institution should deploy AI across the enterprise immediately. The sequence of adoption matters more than the number of use cases launched.

Institutions can begin with use cases where outcomes are measurable and the consequences of error are contained, such as customer education, engagement, service support and employee productivity. As data, integration, governance and operating capabilities mature, AI can progressively extend into more complex and consequential workflows.

Five principles should guide this journey:

- **Start with a business outcome, not a technology capability:** Identify the customer journey, operational challenge or outcome the initiative is expected to improve before selecting the technology
- **Prioritise reach and participation alongside efficiency:** Productivity gains should not only lower the cost of serving existing customers. They should also expand the institution's capacity to reach and support customers who remain underserved
- **Keep humans accountable for consequential decisions:** AI can inform, analyse and recommend, but decisions that materially affect customers require clear human ownership
- **Treat data, integration and governance as part of the deployment:** These are not preparatory activities to be completed before the "real" AI work begins. They are the foundations that determine whether AI can operate reliably and responsibly

³⁰ RBI, List of NBFCs and ARCs registered with the RBI, 2026

³¹ PIB, National Cooperation Policy & Database, 2026

- **Redesign work, not just processes:** Sustainable value will depend on how effectively institutions redefine roles, workflows, decision rights and workforce capabilities around AI

From individual deployments to institutional capability

The institutions that create meaningful value from AI will not necessarily be those that deploy it most aggressively. They will be those that start with clear priorities, learn from contained deployments and build reusable capabilities that allow them to progress with confidence.

This means moving beyond isolated pilots towards an enterprise approach in which outcomes are measured,

costs are visible, controls are embedded and employees understand how AI changes their work. It also means recognising that different use cases should progress at different speeds. Customer engagement, employee productivity and control functions cannot follow the same implementation or governance pathway.

The deployments presented in this report reflect ABCL's experience, developed by its teams with technology partners including EY, Microsoft and Pints AI. The architectural choices and operating principles are specific to ABCL, but the broader lessons are technology-neutral and can be applied across different platforms and institutional contexts.

Today, a new employee is handed a laptop on the first day of work. Before long, the same employee may also be given a set of AI agents as standard tools for the role. The institutions that prepare for that future today will be best positioned to make AI a driver of productivity, responsible growth and deeper financial inclusion.

Appendix

Contributors

Aditya Birla Capital Limited

Anindya Karmakar | Sameer Jadhav | Udatta Ganguly
Ranabir Bose | Sumati Yadav
Paulami Roy | Amrita Siyan

Pints AI

Partha Rao | Calvin Tan

Ernst & Young LLP

Shruti Ladwa | Mayur Chawla | Pranab Ghosh
Sahil Kataria | Gandhar Parnandiwar | Devansh Gupta
Abhinav Suresh | Chetna Khanna

Every technical term used in this paper, in the sense this paper uses it, grouped by the part of the problem it belongs to. The definitions are practical rather than academic.

What an agent is

Agentic AI · AI agent: software that pursues a goal across several steps, planning, using tools and acting, rather than answering a single question. It holds credentials and takes actions that have consequences.

Large language model (LLM): the underlying system that reads and produces language. Agents are built on top of one.

Token: the unit in which an LLM's work is metered and billed, and the reason agent cost behaves as a usage function rather than a licence fee.

Copilot: an assistant that supports a person doing their work rather than acting on its own, so the person stays in the loop by construction.

Orchestration layer: the software that coordinates agents: sequencing their steps, routing work between them and enforcing their permissions.

Low-code · pro-code: build modes: assembling an agent from configurable blocks in weeks, versus engineering one fully in code over months. Matched to how consequential the task is.

What it knows

Retrieval-augmented generation (RAG): grounding an answer in the institution's own documents, retrieved at question time, instead of relying on what the model remembers.

Knowledge graph · ontology: a formal map of an institution's entities (customers, products, policies) and how they relate: the ontology is the schema, the graph the populated map.

Graph-based retrieval · GraphRAG: retrieval that runs through a knowledge graph, so an answer carries the path it was reached by, which is the property that makes it auditable.

Context engineering: deciding what an agent sees before it acts: the data, documents and constraints assembled around each task.

How it connects

A2A (Agent2Agent): the open standard for how agents communicate with one another. Donated by Google; governed under the Linux Foundation since June 2025.

MCP (Model Context Protocol): the open standard for how agents connect to tools and data sources. Donated by Anthropic to the Agentic AI Foundation, December 2025.

How it is governed

Observability: the ability to see, after the fact, what an agent did, on what basis and at what cost. The precondition for governing, auditing and improving it.

Kill switch: the authority, held by a named person, to stop an agent immediately and without escalation.

Per-journey unit economics: knowing what it costs to serve one customer through one journey, so profitable and unprofitable deployments can be told apart.

Blast radius: how far the consequences of an agent's worst plausible failure travel, and whether they can be contained and corrected.

How it fails

Prompt injection: an attack that hides instructions in content an agent reads, steering it to act against its owner's intent. Top of the OWASP list across editions.

Excessive agency: an agent holding more permissions than its current task needs: access granted for one purpose and retained for all others.

Data · model poisoning: corrupting what a system learns from, so it behaves wrongly later. A supply-chain risk for the data layer.

Frameworks and instruments named in this paper

FREE-AI: the RBI's Framework for Responsible and Ethical Enablement of Artificial Intelligence (August 2025); the regulatory anchor of Section 4.

NIST AI RMF: the US standards body's AI Risk Management Framework, published January 2023 and voluntary rather than certifiable. Its four functions govern, map, measure, manage are mapped onto the playbook's three horizons in 6.1.

OWASP LLM Top10 · MITRE ATLAS: the standard security references: OWASP ranks common LLM application risks, ATLAS catalogues real attack techniques against AI systems.

DPDP Act: India's Digital Personal Data Protection Act: the consent basis and fiduciary duties attaching to personal data an agent processes.

Persistency (13-month): the share of policies still in force a year after sale; the standard measure of whether customers were sold something they meant to keep.

Appendix to Section 5 Mapping the Practices to a Concrete Stack

The body of this section is deliberately technology-neutral: the controls matter more than the vendor. For an FI seeking a practical implementation example rather than a product evaluation, the table below maps each control to its implementation in one of ABCL's deployments, built on the Microsoft stack with Pints AI providing agent composition, routing and packet-level evaluation. While other deployments referenced in this paper have been delivered using different partners and technology stacks, the underlying controls remain consistent. Equivalent capabilities are available across other platforms; the key consideration is that the chosen stack is able to support each of the control requirements outlined below.

Control	Partner	Product
Task decomposition and agent creation	Pints AI	Autothought Agent Composer
Managed identity per agent	Microsoft	Microsoft Entra Agent ID
Single registry across the fleet, incl. third-party agents	Microsoft	Microsoft Agent 365
AI gateway: token quotas, caching, content safety at the edge	Microsoft	Azure API Management (AI Gateway)
Cost-optimised routing across packets, within the gateway ceiling	Pints AI	Autothought Model Routing
Data classification, labelling and loss prevention	Microsoft	Microsoft Purview
Conditional access and threat detection for agents	Microsoft	Microsoft Entra + Microsoft Defender
Tracing, continuous evaluation, cost per agent	Microsoft	Microsoft Foundry (Observability + Evaluation SDK)
Model evaluation and drift detection at the packet level	Pints AI	AutoEval
Retirement, revoking identity, access and registry entry	Microsoft	Microsoft Agent 365
Cost, drift and hallucination management at the individual use-case and agent level	Pints AI	Autothought AgentView (with inputs from Microsoft Foundry resources)

Appendix to Section 7 - The One-Page Readiness Checklist

Every item below is drawn from the completion criteria in Section 7. An FI can use it as a standing agenda for the governance forum that owns its agentic AI programme: each line is either true or it is not, and a line that cannot be answered is itself the finding.

Horizon	Check	Yes / No
Day 90	One customer journey named end to end, with a business owner who carries the number	☐
Day 90	The P&L line the journey affects is named, with its current value	☐
Day 90	A manual baseline exists: unit cost, cycle time, error rate, volume	☐
Day 90	A spend ceiling, an iteration cap and written failure conditions, agreed before build	☐
Day 90	Every agent has a managed identity, a registry entry and least-privilege access	☐
Day 90	Tracing and baseline evaluations enabled, emitting on open standards	☐
Day 90	A named board-level owner and a written list of prohibited actions	☐
Month 6	The agent has run on live traffic long enough to have a drift record, not a launch record	☐
Month 6	Unit economics and quality are reported on one dashboard, not two	☐
Month 6	At least one threshold has breached and the pre-agreed response was executed	☐
Month 6	An executive-readable RAG view exists, traceable back to individual decisions	☐
Month 6	The change was carried by line management, with adoption demonstrable by manager cohort	☐
Month 12	A written internal pattern exists, and journeys two and three were delivered faster than the first	☐
Month 12	A single registry covers every agent, including third-party and shadow agents	☐
Month 12	At least one journey has been terminated or redesigned on its unit economics	☐
Month 12	A retirement schedule exists and has retired at least one agent	☐
Month 12	Governance artefacts are produced by running the system, not assembled for an audit	☐
Ongoing	Every model call passes through the gateway; no use case holds its own provider keys	☐
Ongoing	Cost per inference and cost per token are reported for each live journey	☐
Ongoing	The forecast-to-actual gap on time to business value has been measured and explained	☐
Ongoing	The cross-functional owner group meets on cadence, with cost decisions minutes.	☐

Aditya Birla Capital Limited

CIN: L64920GJ2007PLC058890

Registered Office:

Indian Rayon Compound, Veraval, Gujarat - 362 266

Tel: +91 28762 43257

Corporate Office:

One World Center, Tower 1, 18th Floor,

Jupiter Mill Compound, 841,

Senapati Bapat Marg, Elphinstone Road, Mumbai,

Maharashtra - 400 013

Tel: +91 22 6723 9101

abc.secretarial@adityabirlacapital.com

www.adityabirlacapital.com

Toll free no.: 1800-270-7000



**ADITYA BIRLA
CAPITAL**