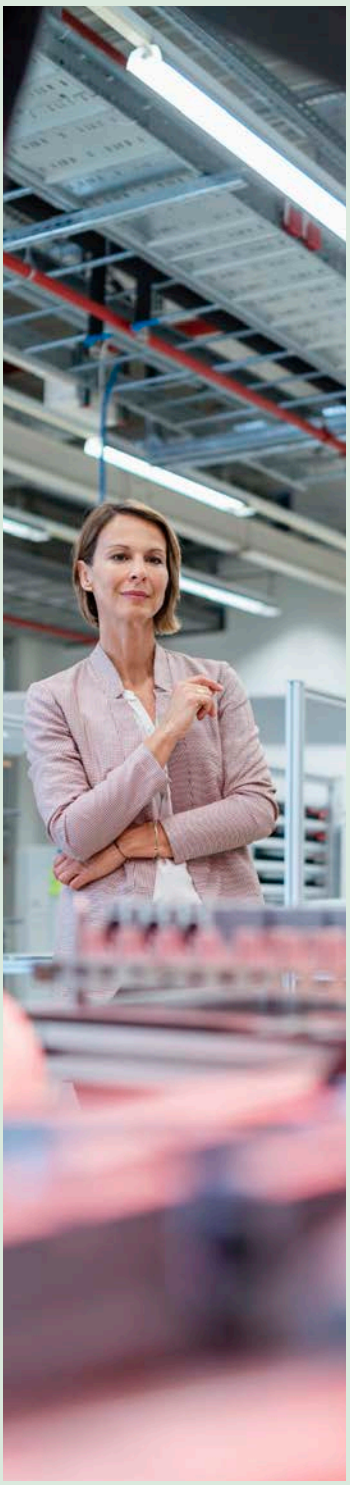
Infosys topaz

WHITE PAPER



SCALING AGENTIC AI: HOW INFOSYS BUILT A MULTI-AGENT PLATFORM FOR A GLOBAL QUICK SERVICE RESTAURANT ON AWS

aws

Infosys
Navigate your next

Executive Summary

Scaling enterprise AI requires a unified platform rather than isolated pilots. A global Quick Service Restaurant (QSR) enterprise faced fragmented AI efforts that duplicated work and limited scale. Infosys addressed this by building a centralized, multi-agent, multi-model platform on AWS using Infosys Topaz and Amazon Bedrock. By standardizing capabilities such as identity, orchestration, memory, and observability, and adopting a tiered model strategy that reduced inference costs by roughly 60 percent in internal benchmarks, the platform supports governed, scalable AI deployment. Backed by Amazon Bedrock AgentCore, a hybrid knowledge layer, and continuous data pipelines, it delivered faster insights, shorter legal review cycles, internally observed availability above 99.9 percent, and inference costs that tracked pre-launch projections. The engagement shows that enterprise AI success depends on platform maturity, not just model capability.

The Challenge: From Fragmented Pilots to Platform Thinking

Building a multi-agent AI platform at enterprise scale is not about whether large language models can solve a problem. The harder question is whether organizations can build the shared infrastructure to solve many problems at once, with the governance, identity, and cost controls production workloads demand. In this paper, we describe how Infosys partnered with a global Quick Service Restaurant (QSR) enterprise to build a unified, multi-agent AI platform on AWS, moving from fragmented pilots to a governed system that supports legal, operations, real estate, licensing, and communications use cases.

The QSR enterprise operates across multiple brands and a wide network of locations worldwide. Like many large organizations, it had reached an inflection point where individual teams were already exploring AI in meaningful ways.

- The legal function was experimenting with contract redlining and risk identification.
- Operations teams were building knowledge assistants to answer enterprise queries.
- Real estate teams were using AI for market and lease analytics.
- Licensing teams were navigating regulatory complexities, and communications teams were automating FAQ generation.

Each initiative showed value in isolation. But they were built independently, resulting in fragmented systems that could not scale cohesively across the organization.

A clear pattern emerged. Every team was rebuilding the same foundational layers: model access, data pipelines, identity controls, orchestration logic, and observability. This duplication was not just inefficient. It was the primary barrier to scaling AI adoption across the enterprise.

Infosys approached the problem with an AI-first, platform-driven strategy. The transformation was guided by Infosys Topaz, an AI-first set of services, solutions and platforms using generative and agentic AI technologies. Instead of extending isolated solutions, the focus shifted to building a unified, governed platform. The new platform supports multiple domain-specific agents while sharing a common backbone for identity, memory, model access, and observability.

The resulting architecture is a multi-agent, multi-model platform on AWS. Each use case is powered by its own agent. All agents operate within a shared, controlled infrastructure. At the core of this foundation is Amazon Bedrock, which provides serverless access to foundation models without the overhead of managing infrastructure.

External Document © 2026 Infosys Limited

Solution Architecture

Figure 1 illustrates the end-to-end architecture of the multi-agent AI platform, showing how the layers interconnect from the user interface through to the data and observability tiers:

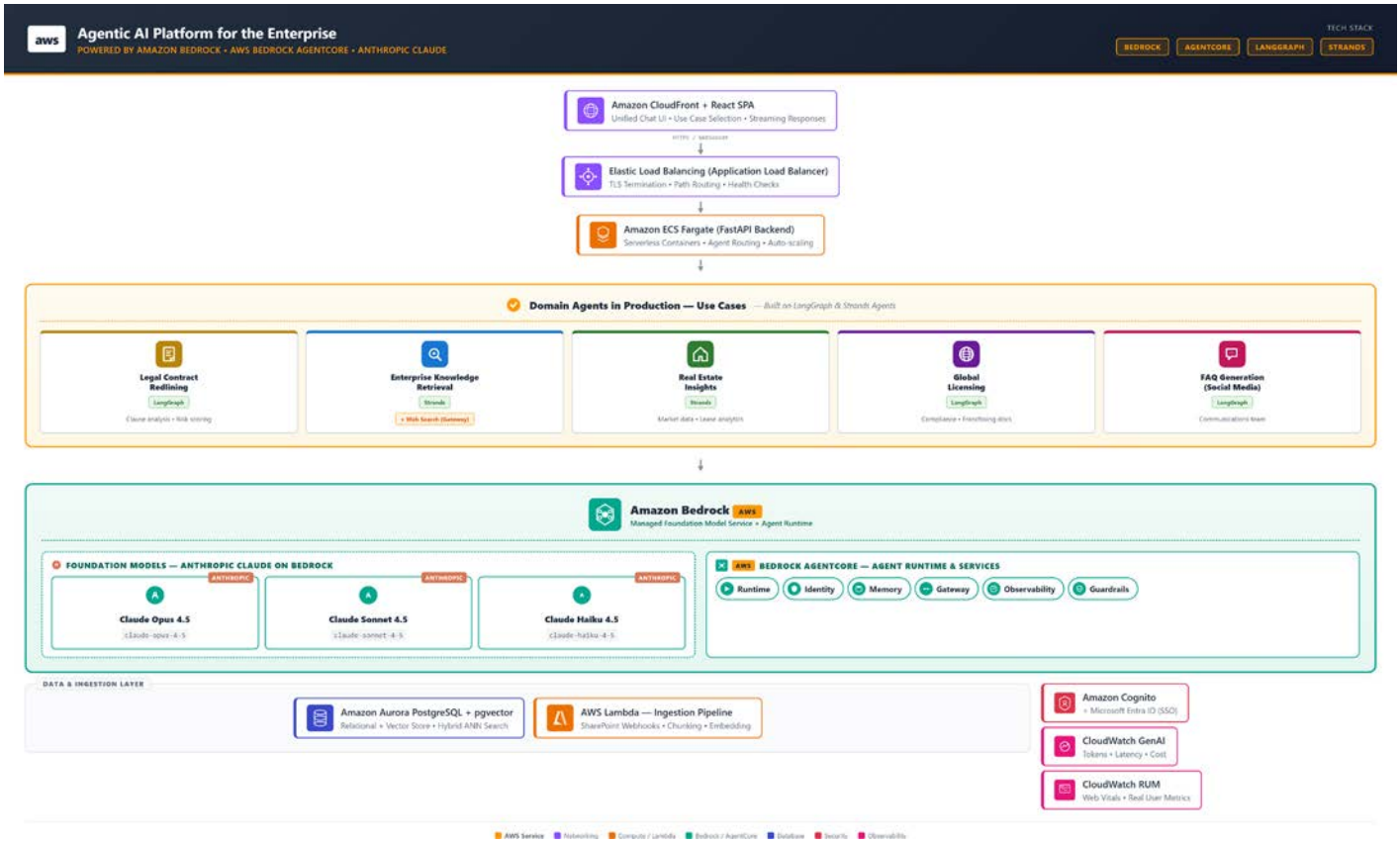


Figure 1: Agentic AI Platform Architecture - Powered by Amazon Bedrock, Amazon Bedrock AgentCore, and Anthropic Claude

The platform spans five layers: a React-based unified chat interface served via Amazon CloudFront; an application layer running on Amazon ECS with AWS Fargate for serverless container execution; Amazon Bedrock for managed foundation model inference; Amazon Bedrock AgentCore for agent runtime, memory, identity, gateway, and observability services; and a data and ingestion layer built on Amazon Aurora PostgreSQL with pgvector and AWS Lambda-powered ingestion pipelines.

External Document © 2026 Infosys Limited

Matching Models to Workloads

To balance performance, reasoning capability, and cost across diverse workloads, the platform adopts a tiered model strategy across Anthropic's Claude family. Claude Haiku 4.5 handles high-volume, low-latency work such as intent classification, FAQ generation, and document re-ranking, where sub-second response and low per-call cost matter most. Claude Sonnet 4.5 powers the medium-complexity middle: core reasoning, tool orchestration, and knowledge synthesis for most enterprise workflows. Claude Opus 4.5 is reserved for high-complexity tasks like deep legal analysis and multi-document reasoning, where its reasoning depth justifies the higher per-call cost.

Routing each task to the appropriate tier, rather than defaulting to the most capable model for every call, supports an internally observed reduction of roughly 60 percent in inference cost compared with a uniform single-model deployment. Prompt caching, cross-region inference, invocation logging, and guardrails further strengthen performance, scalability, and governance.

Operationalizing Intelligence with Amazon Bedrock AgentCore

Table 1: Amazon Bedrock AgentCore — Core Capabilities and Business Impact

While models define intelligence, consistent execution is supported by Amazon Bedrock AgentCore. Every agent runs within this managed runtime and inherits standardized capabilities for orchestration, execution, and governance.

Capability	Function	Business Impact
Managed Runtime	Handles execution, retries, workflows	Consistent and reliable agent behavior
Identity & Access Control	Restricts data and tools per agent	Strict domain isolation
Persistent Memory	Maintains session context	Supports multi-session workflows
Gateway (External Access)	Secure web/API integration	Augments internal knowledge with real-time data
Observability	Tracks model calls and execution	Full traceability and cost visibility

Amazon Bedrock AgentCore keeps every agent within defined boundaries. Legal workflows remain isolated from real estate or operations data through scoped identity and database controls. Persistent memory lets workflows like contract reviews extend across sessions, while gateway capabilities support controlled external augmentation. Every interaction is traceable through Amazon CloudWatch.



External Document © 2026 Infosys Limited

Designing the Right Agents for the Right Problems

Table 2: Agent Framework Selection — LangGraph vs. Strands Agents

As the platform evolved, it became clear that not all use cases require the same type of intelligence. A single approach to agent design would not scale effectively.

Framework	Primary Use Cases	Strength
LangGraph	Legal redlining, compliance workflows, FAQ generation	Multi-step reasoning and stateful execution
Strands Agents	Knowledge retrieval, real estate insights	Fast, tool-driven execution with AWS-native integrations

LangGraph supports structured reasoning workflows where decisions evolve step-by-step. Strands Agents support real-time, API-driven interactions with strong AWS integration. Together, the two frameworks provide both depth and speed across use cases.

Unifying Knowledge Across the Enterprise

The knowledge layer is the backbone of all agent interactions. It runs on Amazon Aurora PostgreSQL with the pgvector extension and supports hybrid retrieval that combines semantic search with structured filtering. Different content types are handled with different chunking strategies: legal documents at clause level, operational data at section level, and analytical content at paragraph level. This domain-aware processing reflects how readers actually use each document type and avoids the loss of precision that comes from one-size-fits-all chunking.

Queries are first classified by intent, then routed to the appropriate index and re-ranked before responses are generated. The trade-off is a small added latency in the query path in exchange for materially better precision, which the team judged worthwhile for legal and operations workloads where the cost of a wrong answer is high.

Keeping Knowledge Updated: Data Ingestion Pipeline

To keep agents working on current information, a serverless ingestion pipeline continuously processes enterprise data.

Table 3: Serverless Data Ingestion Pipeline — Steps and Outcomes

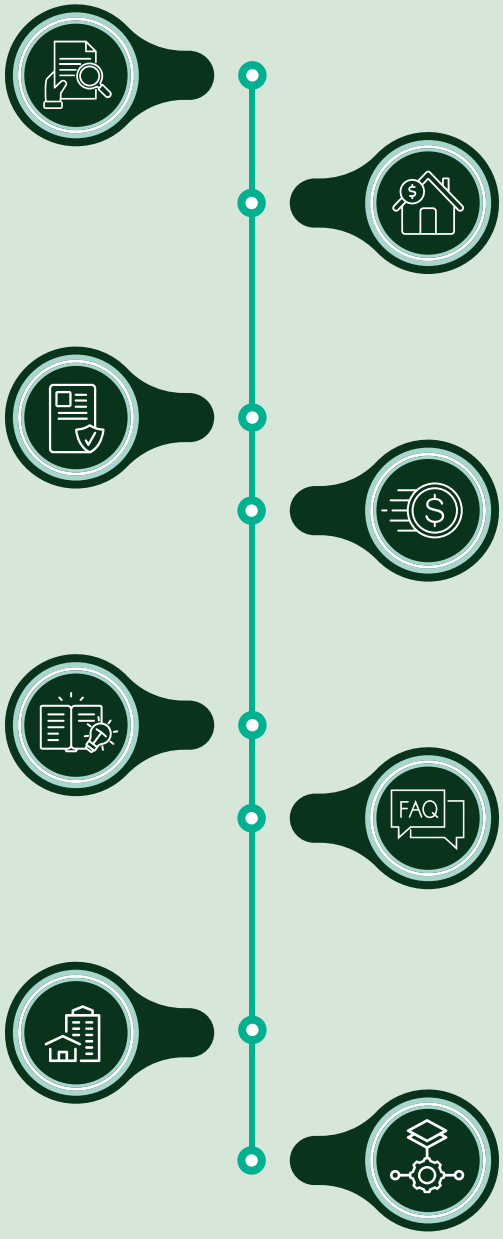
Step	Action	Outcome
Detection	SharePoint triggers ingestion	Continuous updates
Classification	Assign domain tags	Context-aware processing
Chunking	Domain-specific segmentation	Improved retrieval quality
Embedding	Amazon Bedrock models generate vectors	Semantic search enablement
Storage	Metadata-controlled storage	Secure and governed access

External Document © 2026 Infosys Limited

Customer Benefits

From a user perspective, the platform provides a unified interface across all use cases. It supports conversational interactions with clear source attribution and reasoning visibility, which strengthens transparency and trust. The implementation addresses three engineering challenges that determine whether multi-agent systems hold up under real workloads: strict isolation between agents and the data they can touch, precision retrieval through domain-aware processing and re-ranking, and cost governance through model-tier routing.

The platform is estimated to have delivered the following impact within weeks of being deployed*



Legal Contract Review
Initial analysis time reduced from 2–4 hours to 2-4 minutes per contract, with $345K+ in attorney time recovered annually and vendor onboarding reduced from 2–5 days to same-day turnaround.

Franchise Licensing Q&A
Query resolution time reduced from 30+ minutes of manual search to under 30 seconds with exact source citations and 100% audit trail coverage.

Operational Knowledge Access
Time to answer operational queries reduced from 15–30 minutes to under 10 seconds, with an estimated $86K in annual savings from staff time recovery.

Real Estate Site Insights
Historical site data access reduced from hours waiting for SQL queries to self-service in seconds, with $11M projected incremental revenue in Year 1 and $62M cumulative over 5 years (based on 5% AUV uplift on new locations).

Construction Build Costs
Manual data pulls replaced with instant natural language queries, with $3.75M estimated cost savings in Year 1 and $18.75M over 5 years (based on 5% build cost reduction).

Change Order Costs
No prior real-time visibility replaced with instant trend dashboards, with $3.75M estimated annual cost savings (based on 5% change order cost reduction).

LTO FAQ Generation
FAQ creation reduced from several hours of manual work to minutes of automated generation, with consistent brand voice enforcement and $80K–$100K estimated savings over 5 years.

Platform Performance
Internally observed availability above 99.9%, inference costs tracking pre-launch projections, and ~60% inference cost reduction via tiered model routing compared with a uniform single-model approach.

**Disclaimer: All financial values are projections from internal business case estimates at the time of use case chartering. Revenue projections are per the organization's Real Estate Strategy team. These are internal estimates and have not been externally audited.*

External Document © 2026 Infosys Limited

Key Learnings

The following lessons emerged from the engagement and are applicable to organizations considering a similar multi-agent AI platform approach:

- **Platform maturity outweighs model capability.** The primary barrier to scaling enterprise AI was not the quality of foundation models but the absence of shared infrastructure. Standardizing identity, memory, observability, and cost governance unlocked adoption across business functions that had previously stalled at the pilot stage.
- **Tiered model routing is essential for cost governance at scale.** Defaulting to the most capable model for every call is unsustainable. Routing high-volume, low-complexity tasks to Claude Haiku 4.5 and reserving Claude Opus 4.5 for deep reasoning tasks delivered a ~60% reduction in inference costs without sacrificing quality where it matters.
- **Domain-aware chunking directly impacts retrieval precision.** A single chunking strategy applied uniformly across legal contracts, operational SOPs, and analytical datasets produced unacceptable miss rates. Tailoring chunking to document type (clause-level for legal, section-level for operations, paragraph-level for analytics) improved answer accuracy materially.
- **Agent isolation prevents cross-domain data leakage.** In a multi-agent enterprise platform, strict RBAC and scoped database access per agent is non-negotiable. Amazon Bedrock AgentCore's identity and access controls enforced domain boundaries without requiring custom middleware.
- **Persistent memory transforms user experience over time.** Agents that retain context across sessions reduce the burden on users to repeat themselves and improve relevance progressively. This proved particularly valuable in legal and construction workflows where engagements span days or weeks.
- **Observability must be built in from day one, not retrofitted.** Per-agent, per-model cost tracking and full trace logging (via Amazon CloudWatch and AgentCore native tracing) allowed the team to identify cost anomalies, latency spikes, and quality regressions early, before they impacted users at scale.

Conclusion

The lesson from this engagement is straightforward. The true barrier to enterprise AI is rarely model capability. It is platform maturity. By combining Amazon Bedrock, Amazon Bedrock AgentCore, LangGraph, Strands Agents, and Infosys Topaz, organizations can move from disconnected experimentation to governed, scalable agentic systems, and remove the rebuild tax that slows every new initiative. What began as fragmented pilots is now a unified platform that extends to new use cases without rebuilding the foundation each time.

Teams considering a similar approach can start with the public AWS documentation for Amazon Bedrock and Amazon Bedrock AgentCore, review the open-source LangGraph and Strands Agents projects, and explore Infosys Topaz for end-to-end enterprise AI services. For organizations planning a comparable implementation, the Infosys AI and cloud engineering teams can discuss reference architectures, governance models, and deployment patterns tuned to specific industry needs.

External Document © 2026 Infosys Limited

About the Authors



Abhiram Gurijala is a **Digital Solution Specialist and AI Architect** with 19+ years of experience in designing and delivering enterprise-scale AI, Generative AI, and cloud solutions. He specializes in Agentic AI, Large Language Models (LLMs), and intelligent automation, with deep expertise across AWS, Microsoft Azure, and Google Cloud. Abhiram has led the successful delivery of AI-driven transformation initiatives across the QSR, Telecom, and BPM sectors, helping organizations accelerate innovation and realize business value through scalable, cloud-native solutions. He can be reached via LinkedIn.



Dhiraj Thakur is a **Solutions Architect** with Amazon Web Services, specializing in Generative AI and data analytics domains. He works with AWS customers and partners to architect and implement scalable analytics platforms and AI-driven solutions. With deep expertise in Generative AI services and implementation, end-to-end machine learning implementation, and cloud-native data architectures, he helps organizations harness the power of GenAI and analytics to drive business transformation. He can be reached via LinkedIn.

Infosys Topaz is an AI-first set of services, solutions and platforms using generative AI technologies. It amplifies the potential of humans, enterprises, and communities to create value. With 12,000+ AI assets, 150+ pre-trained AI models, 10+ AI platforms steered by AI-first specialists and data strategists, and a 'responsible by design' approach, Infosys Topaz helps enterprises accelerate growth, unlock efficiencies at scale and build connected ecosystems. Connect with us at infosystopaz@infosys.com.

For more information, contact askus@infosys.com

Infosys
Navigate your next

© 2026 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.