

FROM ENTERPRISE ORACLE APPLICATIONS TO PRODUCTION AI

HOW INFOSYS, NVIDIA, AND ORACLE ENABLE
ENTERPRISE-SCALE AI DEPLOYMENT



Table of Contents

The Enterprise AI Imperative	03
Why Oracle Applications Demand a Purpose-Built AI Approach	03
Infosys AI Workbench: Functional Layers Across the Lifecycle	03
Implementation Workbench	04
AMS Workbench (Application Managed Services)	04
Business Workbench	05
NVIDIA Acceleration & Safety Integration	05
Architecture: Client-First, Model-Agnostic, Enterprise-Ready	05
Five Core Architecture Design Principles	06
Layered Architecture in Practice	06
The Path from POC to Production: Navigating Enterprise Governance	07
Built to Pass Enterprise Reviews: Every Gate, Every Layer	08
Requirement-to-Design Mapping: How Every Enterprise Concern Is Addressed	09
Why This Combination Is Uniquely Powerful	10

The Enterprise AI Imperative

Deploying generative AI in a large enterprise requires more than simply connecting a model to business data. For organizations running mission-critical operations on Oracle Applications, AI must be demonstrably secure, governed, compliant, and deeply integrated with existing workflows. Infosys has built a

repeatable, production-grade approach that addresses this – Infosys AI Workbench: a modular, enterprise-ready solution powered by Oracle Cloud Infrastructure (OCI) and accelerated by NVIDIA, designed to move organizations decisively from AI experimentation to scalable production.

Why Oracle Applications Demand a Purpose-Built AI Approach

Oracle Fusion Applications and E-Business Suite are the backbone of enterprise operations — systems of record where accuracy, security, and performance are non-negotiable. Deploying AI in these environments requires deep Oracle domain expertise to work within role-based access controls and preserve system stability.

A key insight from Infosys’ field experience is worth stating directly: **AI deployment is not a conventional implementation — it must be treated as a controlled product release with formal**

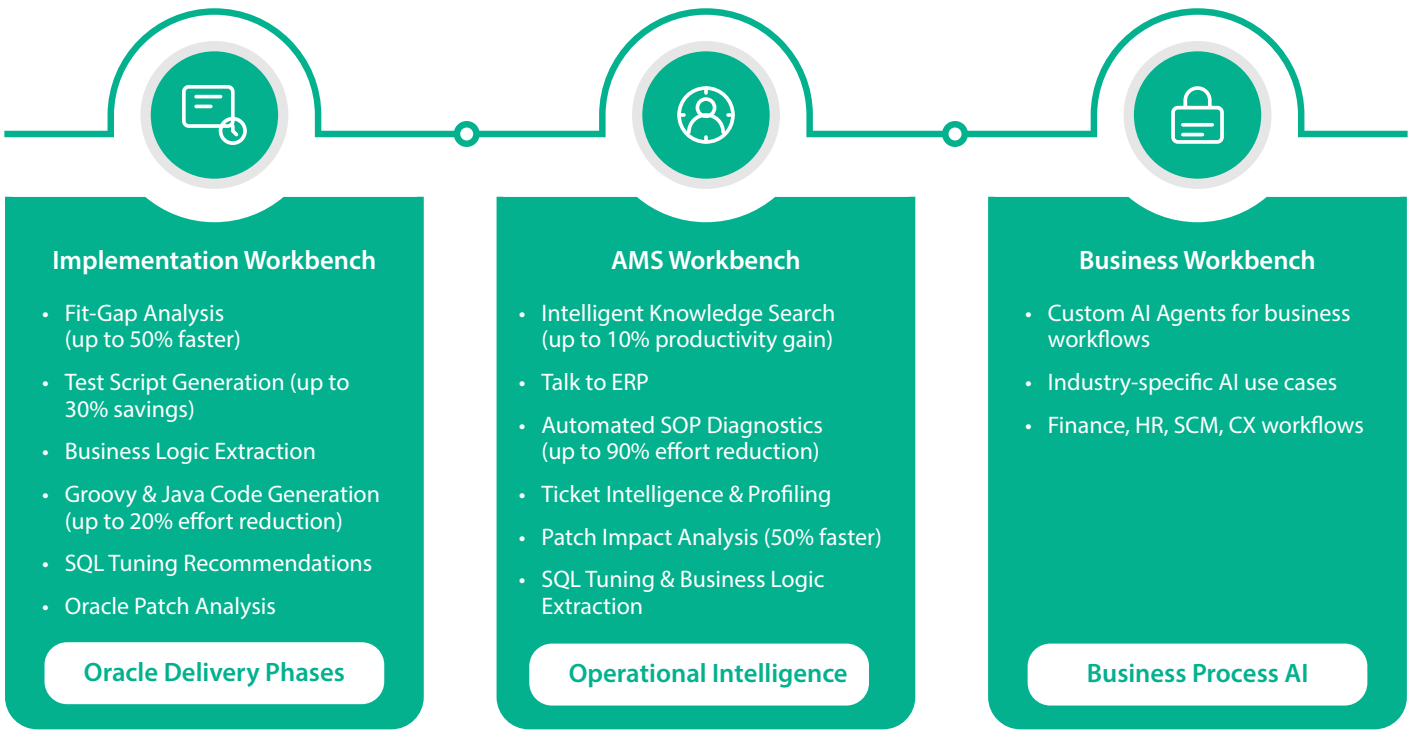
risk acceptance. Enterprise customers treat AI as a key enabler because AI can generate rich insights and provide predictive capabilities that traditional applications cannot. Introducing AI into an enterprise application landscape also introduces new threat classes — prompt injection, data exfiltration, and unsafe generated instructions. Vector indexes and embeddings may inadvertently expose sensitive data structures or content if not properly governed, and model interactions must be controlled in terms of what context is sent to LLMs, where data is retained, and which cloud regions are involved.

Infosys AI Workbench: Functional Layers Across the Lifecycle

The Infosys AI Workbench, powered by Infosys Topaz Fabric, is a modular, enterprise-ready set of AI capabilities built for organizations running Oracle Fusion and related enterprise applications. Its applicability extends across the **full Oracle application portfolio** — and the broader enterprise application landscape — not just Oracle Fusion. It is delivered in three distinct workbench flavors.

Infosys AI Workbench- Three Flavors for the Full Oracle Lifecycle

Powered by Oracle OCI · Accelerated by NVIDIA



Implementation Workbench

The Implementation Workbench provides AI-assisted capabilities across core Oracle delivery phases:

01

Fit-Gap Analysis: AI-enabled analysis against Oracle product documentation and customer-specific knowledge that can reduce fit-gap assessment cycle times by up to 50%.

02

Test Script Generation: Automatically generates detailed test scenarios and scripts from business requirements, yielding up to 30% savings in preparation effort.

03

Business Logic Extraction: Analyzes PL/SQL application code to convert complex technical logic into human-readable formats, accelerating modernization.

04

Groovy & Java Code Generation: Automatically generates code aligned with Oracle Fusion Cloud standards, reducing manual coding effort by up to 20%.

05

SQL Tuning Recommendations: AI-driven query analysis that identifies performance bottlenecks and provides targeted optimization rewrites and recommendations.

06

Oracle Patch Analysis: Automated, AI-enabled impact analysis of quarterly patches across business processes, configurations, integrations, data models and Fast Formulas, delivering up to 50% faster analysis.

07

Talk to ERP: Enables rapid data exploration during implementation phases without requiring technical SQL expertise.

AMS Workbench (Application Managed Services)

The AMS Workbench focuses on AI-driven operational intelligence and support automation:

Intelligent Knowledge Search: Permission-aware natural language search across enterprise repositories that delivers ACL-enforced answers without exposing unauthorized content. Up to 10% improvement in team productivity and 30% reduction in onboarding time.

Ticket Intelligence & Profiling: AI-driven identification of recurring incident patterns, sentiment scoring, resolution hop analysis, and quality metrics to proactively improve operational support quality.

Talk to ERP: Natural language interaction with production Oracle Fusion data to query reports and business entities without SQL expertise, utilizing enforced guardrails to protect underlying systems.

Patch Impact Analysis: Proactive assessment of Oracle quarterly patch releases on running configurations, integrations, and business processes.

Automated SOP Diagnostics: Automatically analyzes incidents against registered SOPs to identify the most relevant resolution steps, delivering up to a 90% effort reduction in validation checks.

SQL Tuning & Business Logic Extraction: Provides both reactive support and proactive database stability management.

Business Workbench

The Business Workbench focuses on enterprise-specific AI augmentation opportunities through a library of pre-built use cases and custom agents to accelerate business workflows.

Custom AI Agents: Configurable, role-specific agents that automate custom business workflows where native, embedded AI is unavailable.

Intelligent Document Extraction: Leverages NVIDIA NV Parse and Vision Language Models (VLMs) for structured data extraction from unstructured invoices, purchase orders, and contracts.

NVIDIA Acceleration & Safety Integration

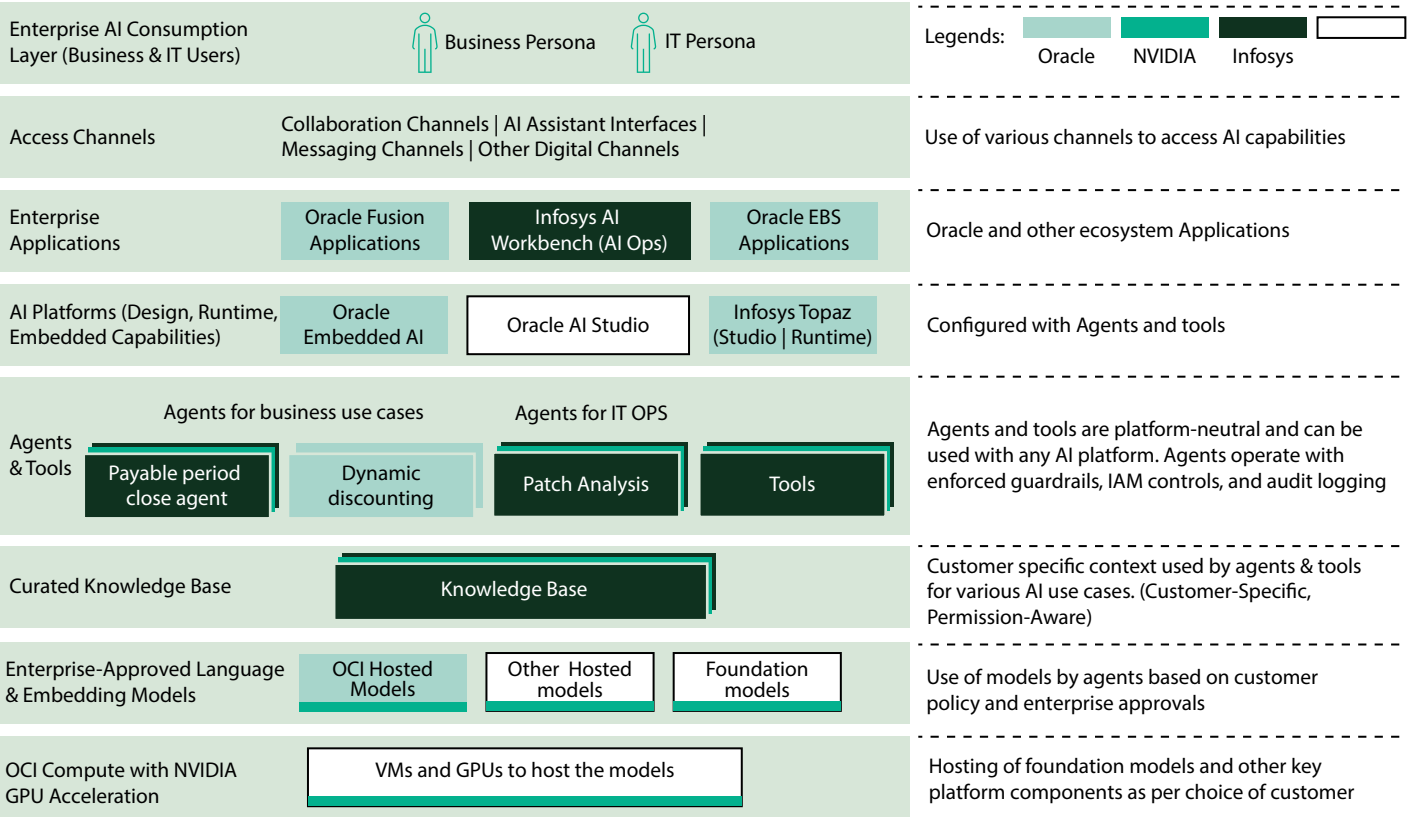
NVIDIA technologies are embedded directly into the Infosys AI Workbench pipeline to provide the performance and guardrails necessary for production. NVIDIA NeMo Guardrails act as the primary safety layer, enforcing real-time PII filtering, jailbreak detection, and topic controls. Performance is driven by NVIDIA cuVS using the CAGRA algorithm, delivering up to a 5x speedup

for GPU-accelerated vector index generation within Oracle Database 26ai—making real-time RAG feasible under live enterprise workloads. For strict compliance boundaries, NVIDIA NIM microservices enable fully self-hosted, containerized inference within the customer’s secure OCI environment.

Architecture: Client-First, Model-Agnostic, Enterprise-Ready

The Infosys AI Workbench architecture is designed to be client-first and model-agnostic — decoupled from any single model provider so that OCI-hosted models, enterprise-approved third-party models, or future foundation models can be adopted based on customer policy, without vendor lock-in.

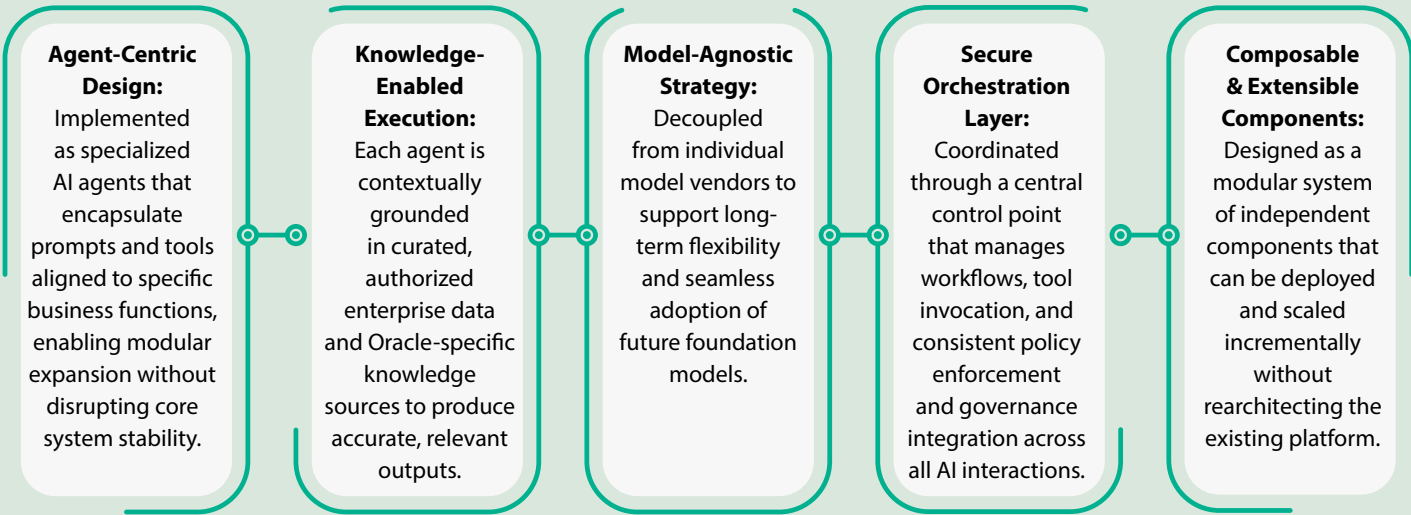
Client-First, Model-Agnostic AI Strategy



The following sections break down how these architectural layers translate into a structured, enterprise-ready deployment model.

Five Core Architecture Design Principles

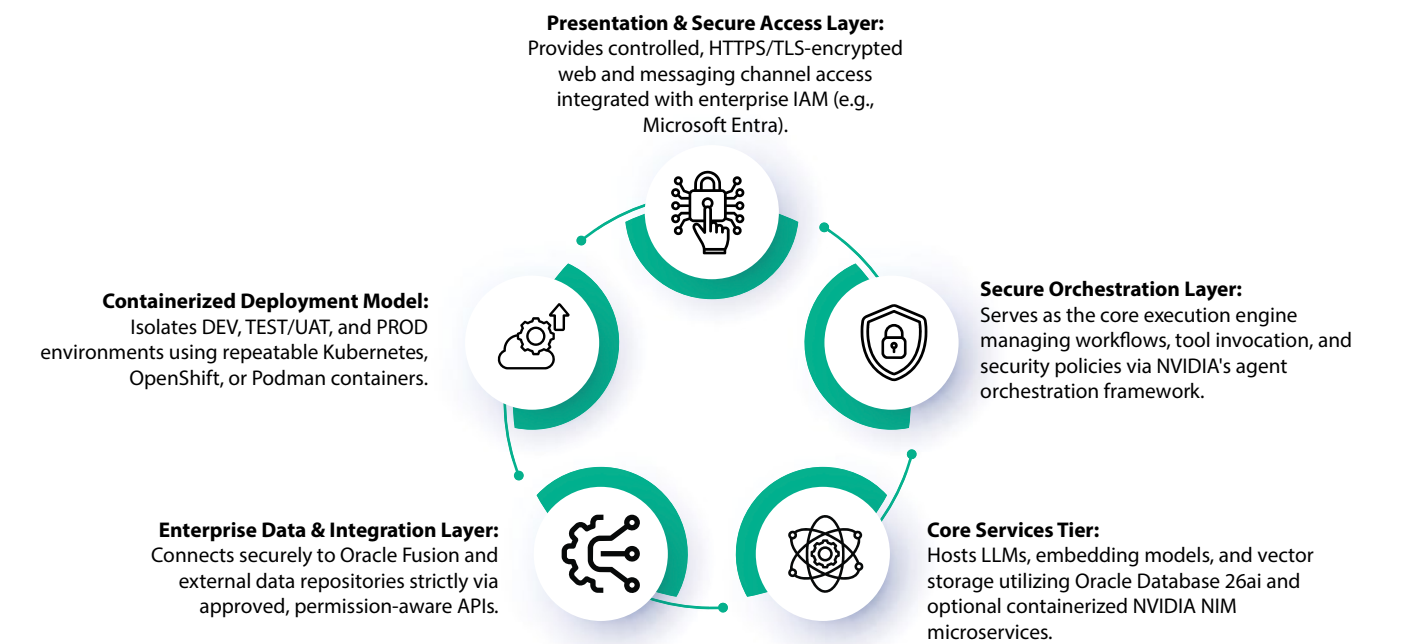
The Infosys AI Workbench architecture is built on a set of foundational design principles that ensure scalability, flexibility, and enterprise-grade governance. These principles provide the foundation for how the architecture is implemented and operated in practice.



Layered Architecture in Practice

The Infosys AI Workbench architecture is implemented as a **structured, layered design**. Each layer operates as a controlled boundary with clearly defined responsibilities and integration points, while remaining within enterprise security and governance controls.

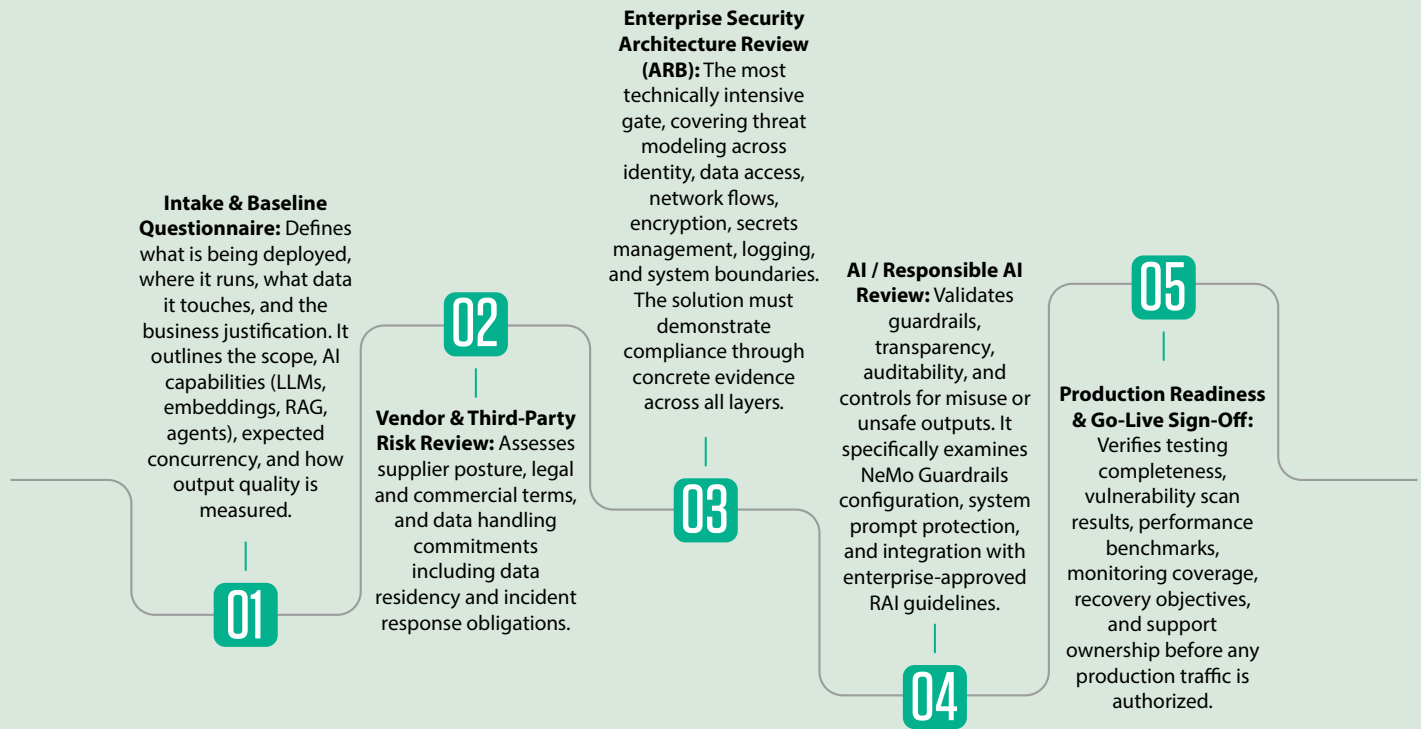
The following layers correspond to the architectural stack illustrated in the diagram above, grouping platform components into logical execution layers for enterprise deployment.



Together, these layers translate the conceptual architecture shown in the diagram into a structured, enterprise-ready deployment model.

The Path from POC to Production: Navigating Enterprise Governance

Deploying generative AI in a large enterprise requires structured and accelerated reviews across customer architecture and governance processes. Most large organizations require five formal gates:



Built to Pass Enterprise Reviews: Every Gate, Every Layer

The Infosys AI Workbench for Oracle is natively designed for enterprise-grade deployment, with built-in controls aligned to common governance requirements:

- **Data Protection & Privacy:** Only the minimum necessary, authorized content is included in LLM prompts—no bulk data exposure. All data in transit is encrypted (HTTPS/TLS; TCPS with Oracle Wallet for database connections). Anonymization and masking controls are applied for sensitive content, and vector indexes are governed at the schema level with explicit grants to prevent exposing row-level transactional data.
- **Identity, Access Control & Oversharing Prevention:** Authentication is delegated entirely to the customer's existing IAM (Entra ID, SAML/OIDC) with role-based controls. Document permissions from SharePoint and other repositories are stored as metadata in the vector store and enforced at retrieval time (ACL-aware RAG). Oracle application integrations use scoped service accounts with entitlement enforcement via approved APIs.
- **Secrets Management & Token Handling:** All backend services authenticate using managed identities, eliminating hard-coded credentials or static tokens. Access tokens for systems like Oracle Fusion, ServiceNow, or databases are retrieved dynamically at runtime from enterprise-approved key vaults (e.g., Azure Key Vault) via cloud-native identity services. Secrets remain fully owned, stored, and rotated by the customer.
- **AI Use Case Evaluation & Continuous Quality Assurance:** Each AI use case onboarded defines objective, measurable evaluation criteria (accuracy, relevance, grounding, response latency) rather than relying on subjective user feedback. Evaluation mechanisms include curated test prompts, golden datasets, and rule-based scoring methods that can be re-executed periodically to detect behavioral drift or degradation.
- **AI Safety & Responsible AI:** NVIDIA NeMo Guardrails enforce input/output safety policies, topic restrictions, grounding requirements, and jailbreak detection at the orchestration layer. Alternatively, the platform can integrate with customer-provided RAI frameworks. System prompts are protected from end-user modifications, and all interactions are logged for audit.
- **Action Safety for Oracle Application Integrations:** SQL generation and automated execution use cases are protected by query allow/deny lists, auto-injected SQL LIMITs, concurrency caps, and rate limits that prevent unsafe patterns (DML/DDL) or entitlement bypass. Kill switches and rollback mechanisms halt AI-driven actions if anomalies are detected, using only Oracle-exposed APIs.
- **Monitoring, Observability & SRE Alignment:** Telemetry tracks agent performance, request latency, error rates, guardrail violations, and token consumption by feature and environment. Centralized, sanitized logging captures user interactions, agent execution flows, and tool invocations, integrating directly into the customer's existing observability platform.
- **Performance, Scale & Concurrency:** AI agents are implemented as stateless, horizontally scalable components. This enables multiple instances to be distributed across compute resources to handle increasing user volumes and data sizes without compromising reliability or system integrity.
- **Business Continuity, DR & Failure Handling:** Backup and recovery mechanisms are defined per customer policy for configuration data, knowledge artifacts, and vector indexes to meet enterprise RTO and RPO expectations. Separate DEV, TEST/UAT, and PROD environments follow controlled promotion mechanisms to keep production stable.
- **Operational Resilience & Change Management:** The solution is containerized (Podman/Kubernetes/OpenShift) for repeatable deployment and environment promotion. A versioning and change control framework governs prompt updates, model changes, and container patches. The Fusion Patch Analysis use case itself acts as a governance tool to proactively identify the impact of quarterly updates.
- **RAG Pipeline: Built for Enterprise Scale:** Hybrid retrieval patterns (dense + sparse search), Reciprocal Rank Fusion (RRF), and reranking improve relevance for large corpora. Scheduled sync mechanisms with delta detection re-embed only new or modified documents, while hashing and similarity checks reduce repeated content and prevent duplicate data indexing.
- **Token Economics & Cost Visibility:** Token consumption is tracked by use case and environment via telemetry to enable cost estimation, chargeback reporting, and ongoing optimization. Ballpark token consumption estimates by use case are provided upfront to support pay-as-you-go pricing discussions with cloud and model providers.

Requirement-to-Design Mapping: How Every Enterprise Concern Is Addressed

Enterprise Requirement	Design Elements	Evidence Artifacts
Prevent repository oversharing; enforce per-document access	ACL-aware RAG; permissions stored as metadata; SSO integration	Access-control narrative; retrieval flow; ACL alignment proof; KB access model
Scale to large repositories; keep index current	Scheduled sync with delta detection; batch re-embedding; deduplication	Sync design; delta detection logic; ingestion logs; dedup rules
Secure data sent to LLMs	Data minimization; TLS encryption; anonymization/masking; tenant boundary	Data-flow diagram; LLM data-handling statement; masking approach
Prompt injection & unsafe output	NeMo Guardrails; protected system prompts; output filtering	Guardrails config; policy list; test results
Oracle Fusion access via supported APIs	Oracle-exposed services only; scoped service accounts; entitlement enforcement	Integration design; role model; API list; least-privilege evidence
Performance protection for Fusion access	Allow/deny lists; auto-injected LIMITs; rate limits; concurrency caps; kill switch	Performance strategy; NFRs; limits config; monitoring dashboards
Demonstrate measurable AI quality, fitness for purpose, and ongoing behaviour stability	Objective evaluation definition per use case; baseline and recurring evaluation execution; evidence retention	Evaluation criteria definition; curated test prompts and results; scoring outputs; trend/comparison reports
Change/patch management for AI + Oracle quarterly patches	Versioning; controlled rollout; rollback plans; Patch Analysis workflow	Release checklist; rollout/rollback plan; patch analysis docs
Secure secrets & token management	Managed identity-based auth; runtime token retrieval from key vault; short-lived tokens	Managed identity flow; key vault integration design; token acquisition sequence
Token economics & cost visibility	Telemetry-based tracking by feature; periodic cost review	Token assumptions; telemetry report; monthly usage output

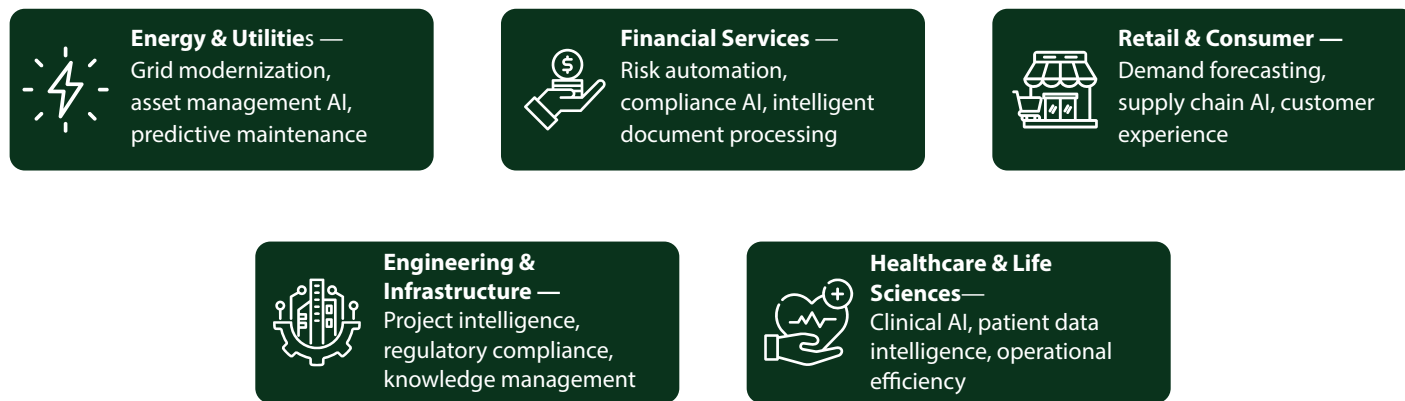


A Repeatable Pattern Built for Scale — Across Industries

What makes this approach particularly valuable is that it is not a one-off implementation — it is a **standardized, repeatable deployment pattern**. Once a customer approves the architecture through their governance gates, expanding to new use cases, business units, or AI models becomes a seamless process as the security and governance foundations are already accepted and documented.

The core solution — agents, tools, knowledge bases, and workflows — is intentionally vendor-agnostic. LLM providers, vector databases, and orchestration platforms can be adapted based on customer policy or technology preference without redesigning the underlying solution. This directly addresses one of the most common enterprise concerns during review: avoiding lock-in to a single AI vendor at a time when the technology landscape continues to evolve rapidly.

This pattern applies across industries and application landscapes:



Why This Combination Is Uniquely Powerful

The Infosys, NVIDIA, and Oracle collaboration represents a mature, production-tested approach to enterprise AI that brings together three essential strengths:

Infosys Oracle Services

Infosys has deep Oracle domain expertise, a battle-tested enterprise governance framework, and reusable deployment patterns built from real production experience across regulated industries. Winner of the Oracle AI Innovation Award 2025 — recognized for leading AI-driven innovation across enterprise clients globally, Infosys operates across 56 countries, with 12,000+ AI assets via Topaz and 150+ pre-trained models.

Oracle Cloud Infrastructure

Secure, enterprise-grade cloud hosting with native AI services, GPU infrastructure, Oracle AI database 26ai with NVIDIA cuVS-powered vector search (now GA as of GTC 2026), and the Oracle Fusion AI Agent Marketplace enabling enterprise-scale agent deployment.

NVIDIA AI Stack

Delivers the underlying acceleration, compute intelligence, and safety guardrails that make production enterprise AI feasible. The stack embeds NVIDIA NeMo Guardrails directly into the request pipeline for real-time compliance and PII protection, while utilizing NVIDIA cuVS to accelerate vector searches up to 5x within Oracle databases. Additionally, it provides multi-agent orchestration via the NVIDIA's agent orchestration framework, self-hosted inference options using NVIDIA NIM microservices, and specialized document extraction utilizing NV Parse and Vision Language Models (VLMs).

For organizations running Oracle Applications who are ready to move beyond experimentation, this is what production AI looks like: governed from the ground up, scalable by design, secure at every layer, and deeply integrated with the systems and workflows that matter most.

To learn more about Infosys AI Workbench:

[Click to explore more](#)



About the Authors:



Asad Ansari

Asad is the Go-to-Market (GTM) Lead at NVIDIA, where he drives strategic initiatives for global system integrators through the NVIDIA-Oracle partnership. Prior to joining NVIDIA, Asad spent years advising Fortune 500 clients, delivering high-impact strategic cloud architecture, digital transformation, and large-scale implementation engagements.



Deepak Munjal

Deepak has over 35 years of industry experience, having held engineering, sales, and leadership roles at leading technology companies, including Hewlett Packard Enterprise, Cisco, and Oracle. He joined Oracle nine years ago, initially serving as a Product Manager for Oracle Cloud Infrastructure (OCI). Today, he is a Technical Alliance Consultant within Oracle's Channels and Alliances organization, where he works closely with Global System Integrators to support and accelerate their cloud transformation and business growth initiatives.



Saurabh Muley

Saurabh is a Senior Principal Technology Architect and Head of AI and Automation initiatives within the Infosys Oracle practice. With nearly 30 years of experience driving large-scale digital transformations for Fortune 500 organizations, he specializes in architecting next-generation enterprise solutions using Generative AI, agentic frameworks, and vector databases. Saurabh is the lead architect behind multiple widely adopted AI-driven enterprise platforms and holds advanced NVIDIA professional certifications in Agentic AI and Multimodal Generative AI.

For more information, contact askus@infosys.com



© 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.