Enterprise AI Platform

Athena

AURDA

: From AI Chatbot Adoption to Enterprise-Wide Operations

uracle

AI Transformation : Three Structural Shifts

Generative AI is transforming not just model capability, but the very structure of software and services.

Transformation 01.

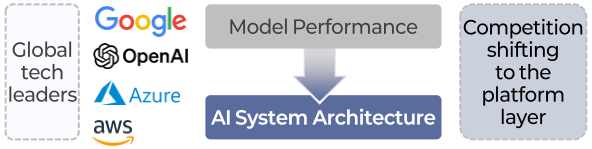
The End of the Model Race

As model performance converges toward parity, competitive advantage has shifted from "how smart is the model" to **"how effectively can it be applied to real work."**

POINT 01 **Performance Convergence**

- Leading models (GPT-4, Claude, Gemini) now **match human-level performance** on most benchmarks.
- Most models **score 80–90%+** on MMLU.

POINT 02 **A New Competitive Landscape**



What matters now is **AI system design**, not the model alone.

Transformation 02.

S/W Architecture Shift

Software is moving from tools users operate directly to systems where **AI performs complex tasks on their behalf**.

POINT 01 **AI Agent Market Growth**

- The AI agent market is projected to grow at a **45–50% CAGR** through 2030. (Markets and Markets/Gartner)

POINT 02 **Enterprise Software Restructuring**

- By 2028, **over 30%** of enterprise software will include agent-based capabilities. (Gartner)

GitHub Copilot | OpenAI Operator

Software is shifting toward **task execution**, not just tools.

Transformation 03.

SaaS → OaAS Emerges

Software is evolving from selling features to **delivering the outcomes users actually want, as a service**.

POINT 01 **Shifting Purpose of Enterprise AI Adoption**

- **Over 40%** of enterprises adopting generative AI **use it for task automation.** *McKinsey AI Report*

Report Writing | Code Generation | Data Analysis | Customer Response

POINT 02 **The Rise of Outcome-Based Software**

- Agent-based software is evolving from feature-driven SaaS to **outcome-driven OaAS.** *BCG/ITPro*

* OaAS : Outcome as Agentic Solution

Enterprises **now expect outcomes**, not just features.

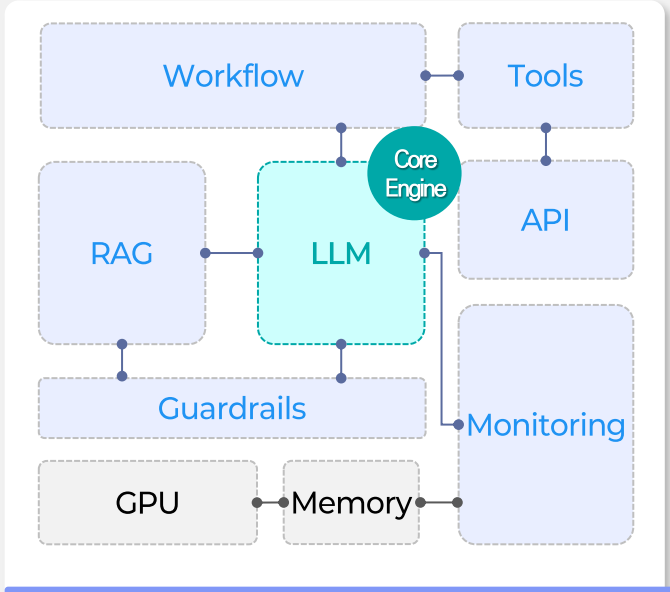
Our Approach : A Design Strategy for the AI Era

Responding to the shifts in generative AI requires an approach that designs AI system architecture, software methodology, and service models together.

Strategy 01.

Composite AI System Architecture

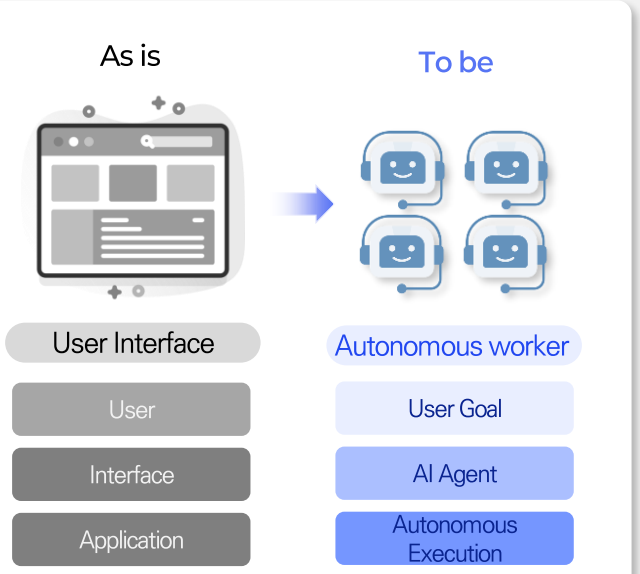
AI must be architected as a system of multiple components—**Workflow, Tools, Retrieval, and Memory**—not a single model.



Strategy 02.

Agentic Software

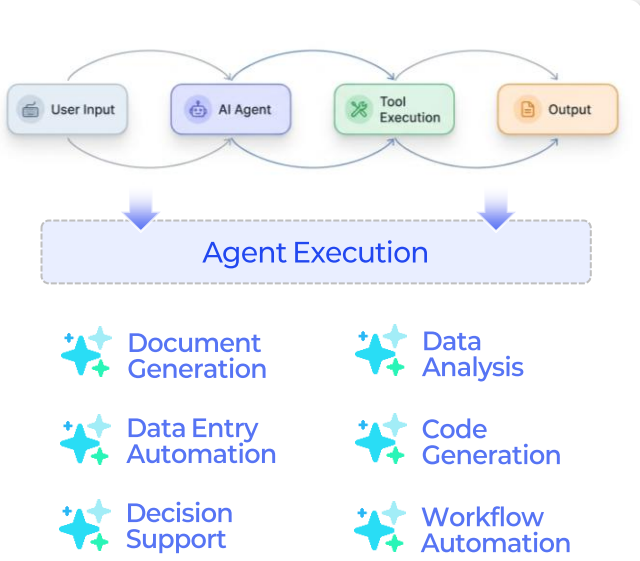
AI services must be built on **an agent-based architecture** that plans, executes, and uses tools—not just generates responses.



Strategy 03.

Outcome Automation

AI services must be designed around goal-driven agent workflows that **generate business outcomes directly**.



Realizing this architecture requires an **AI platform built to support agent execution and workflow orchestration.**

Our Solution : AI Product Suite (Athena/AURDA)

A unified platform for scaling AI company-wide—not scattered team-level pilots or short-term PoCs.

Athena - AI Agent Builder platform

Progressively expanding AI capabilities to fit each workflow

Expand: Extensions

Specialized feature modules for diverse work formats

- Agentic RAG
- Code Assistant
- OCR

Core: Agent Builder

Agents that go beyond simple answers to execute end-to-end workflows

- Workflow
- Multi-Turn
- Tool Integration
- Prompting

Base: RAG Pipeline

Turning enterprise knowledge into a searchable asset,
with high-quality retrieval

- Parsing/Embedding
- Vector Search
- Re-ranking

AURDA - Ops platform

Infrastructure and governance for reliably operating AI services at scale

Govern: Governance

Controlling access to data and features

- IAM / SSO
- RBAC
- Audit

Core: Operation

Stable operations for AI services at scale

- API Gateway
- CI/CD (GitOps)
- Auto-Scaling

Base: Infrastructure

The foundational infrastructure layer powering AI execution

- GPU Virtualization (MIG)
- Storage Management
- Networking

Key Capabilities : Unified Platform Architecture

From AI infrastructure to the Agent platform and real-world service deployment—delivered as one unified platform.

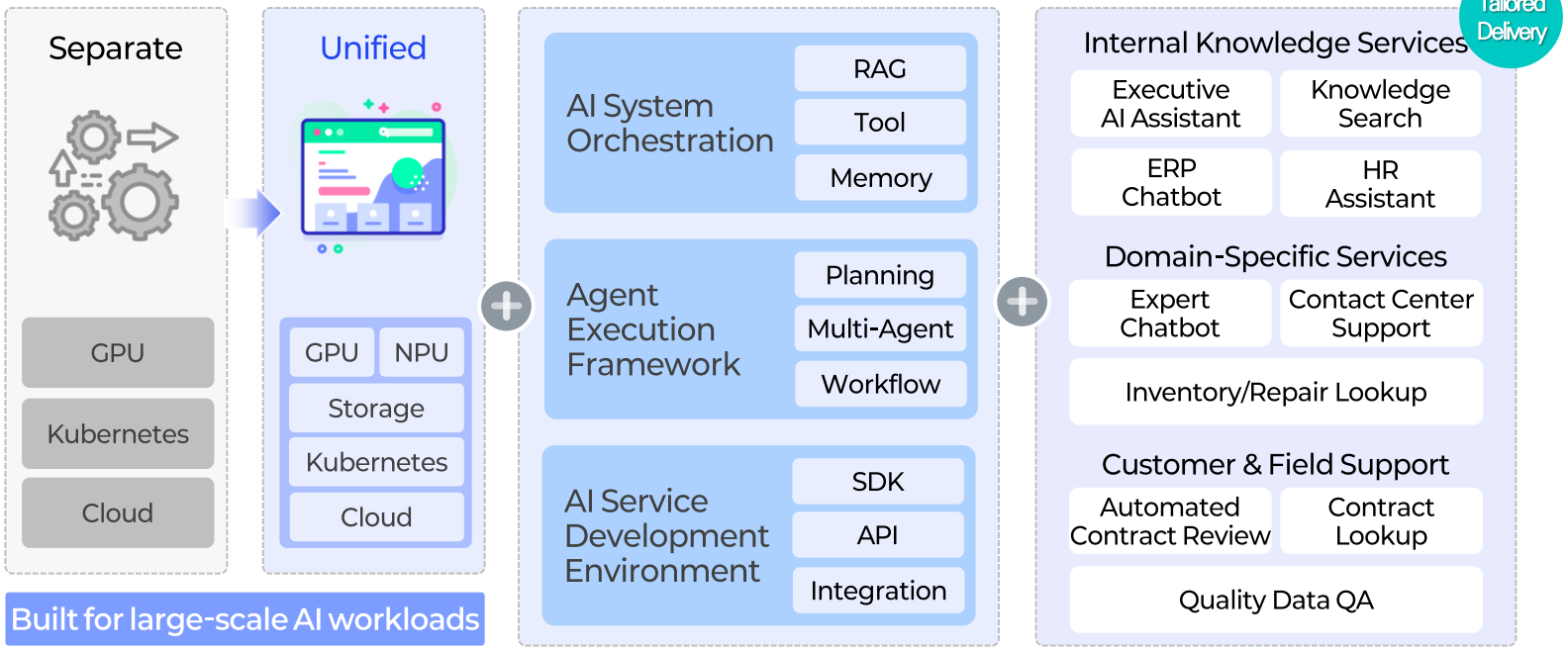
Integration | AI Technical Depth

A Unified Platform, from AI Infrastructure to Services

Not siloed builds—
an AI Operations Platform

Agent-based AI,
delivered as a platform

Powered by the platform,
Purpose-built AI services



Platform Impact

From infrastructure to Agent platform to services -one integrated AI operations framework.

1 Faster Delivery
- Less repetitive integration work
- More focus on service logic

2 Stable Operations
- Unified infra + Agent management
- Consistent deployment & monitoring

3 Sustainable Scaling
- Flexible for new AI services
- Built for many services, not one

Key Capabilities : Advanced RAG Processing

Agentic RAG enables multi-step reasoning for advanced RAG processing.

Integration | Intelligent Search

Next-Gen RAG Architecture

Technology Evolution

"Retrieval-based RAG"

Standard RAG

Single-Retrieval Response	Basic structure— retrieve relevant docs, answer directly
One-Way Execution	LLM answers immediately, even with inaccurate or incomplete results
Single-Query Focus	Can't combine, compare, or analyze multiple sources — limited on complex queries

Given a whole-life policy vs. a 10-year fixed annuity, which yields a higher one-year payout?

"Agent-Based RAG"

Agentic RAG Expansion

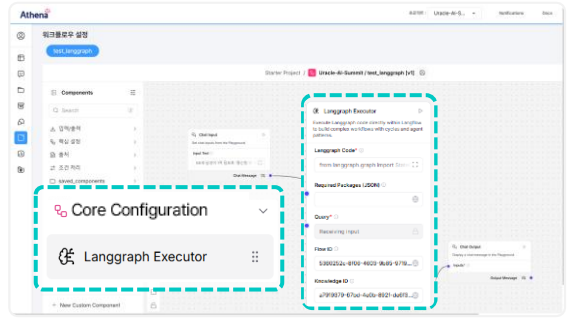
Extensible

Query Intent Analysis & Optimization	LLM interprets intent, rewrites query/logic — maximizes retrieval accuracy
Feedback Loop	Runs additional retrieval when results are insufficient — improves answer quality
Multi-Step Reasoning	Agent compares, analyzes, and computes across multiple documents — enables deep, complex answers

Agent Builder

Workflow	Tool Registry	MCP Repository
Multi-turn	Conversation History	Conversational Interface

\+ LangGraph



Simplifying Agentic RAG complexity into a platform

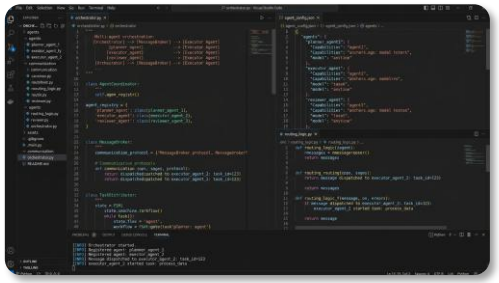
Key Capabilities : Agent Development Productivity

AI Code Assistant supports agent development—from coding to collaboration.

Productivity | Dev Automation

AS-IS

IDE-Based Development



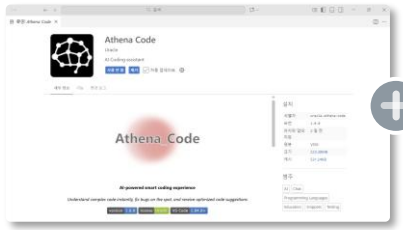
〈IDE development example〉

- RAG pipeline built from scratch in code
- Manual Tool/Memory/Retrieval integration
- Manual multi-agent orchestration

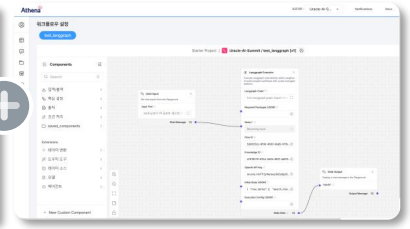
Agent complexity, compounding

TO-BE

IDE + Agent Platform Development



〈 Athena Code installed in IDE〉



〈Workflow in platform〉

- VSCode/Eclipse support
- API-based large-scale document processing
- Automated code generation & refinement
- Orchestration tool management
- Project-aware code analysis
- AI runtime and dev code separation

Lower complexity, higher productivity through collaborative development

Key Effects

A Shift in Development Paradigm

IDE-Based Agent Collaboration

- VSCode/Eclipse IDE support
- Code Assistant–driven generation & review
- Team–based agent development environment

Simpler RAG/Agent Builds

- Platform handles parsing, chunking, embedding
- Simplified Tool/Memory/Retrieval integration
- Less orchestration code overhead

Faster AI Service Development

- Agent workflow–based service delivery
- Platform–managed document processing & AI infra
- Consistent structure from dev to operations

Key Capabilities : Unified Agent Operations

Generative AI is reshaping not just models, but software and service architecture itself.

Reliability | Resource Optimization

AS-IS

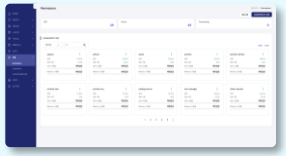
No Standardized Runtime or Ops Environment

Repeated Environment Setup	Complex Infra Config	Environment Mismatch	Lack of Ops Automation
Each agent/service built independently	Manual setup for GPU, Vector DB, Kubernetes	Dev/prod gaps reduce service stability	Manual monitoring, incident response, resource management

TO-BE

A Standardized AI Runtime & Ops Platform (PaaS-Based)

Standardized Environment

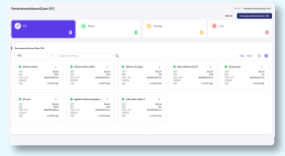


〈 Namespace Isolation〉

〈 Catalog Management〉

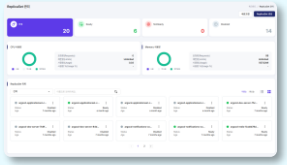
Unified Infrastructure

〈 GPU MIG 〉

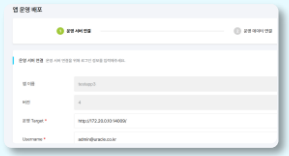


〈 Unified Storage Management〉

Scalable Services

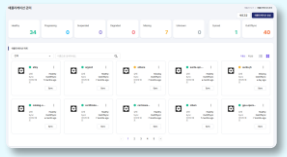


〈 Deployment Management〉

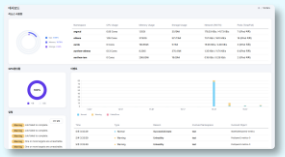


〈 CI/CD〉

Automated Deployment



〈 App Management〉



〈 Unified Monitoring 〉

Key Features : Multi-Model Management

Register and control private and public LLMs in a single catalog—enabling company-wide AI governance.

Private · Public LLM Management

Unify internal vLLM-served models and external API models in one catalog, under central control.

Zero-Downtime Model Switching, No Code Changes

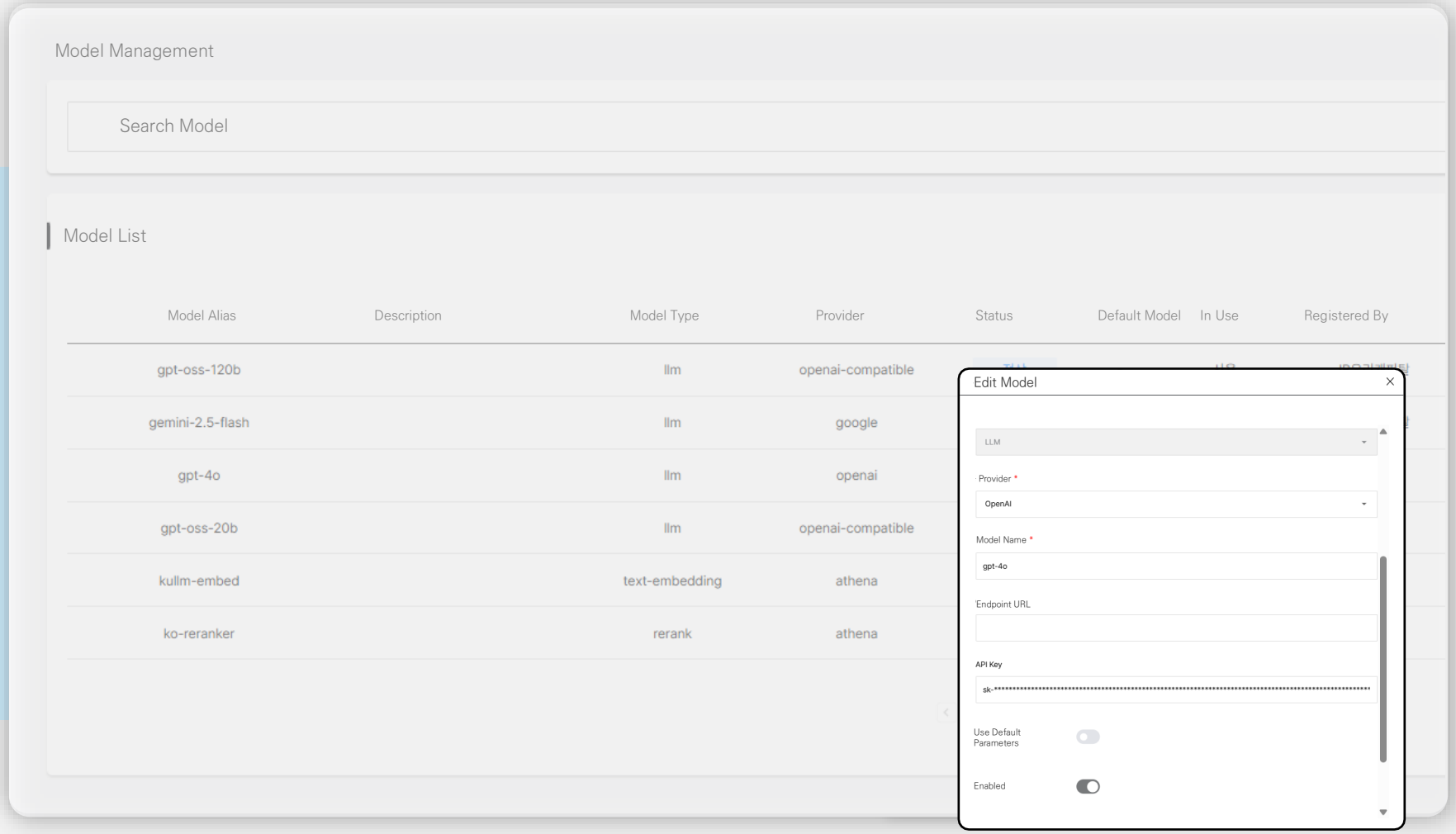
AI Gateway unifies heterogeneous models under a single API—swap the connected model without touching downstream code.

Team/Project-Level AI Resource Isolation

RBAC and namespace isolation apply usage permissions and parameter limits per team.

Model Quality & Ops Visibility

Track response time, token usage, and success rate per model in real time—catch anomalies early.



Key Features : RAG Pipeline Builder

From complex document parsing to Agentic RAG–based retrieval—design the entire pipeline on one platform.

Structuring Knowledge from Unstructured Documents

Parse PDF, HWP, and MS Office files, with configurable Parent–Child chunking strategies.

Precision Answers via Hybrid Retrieval

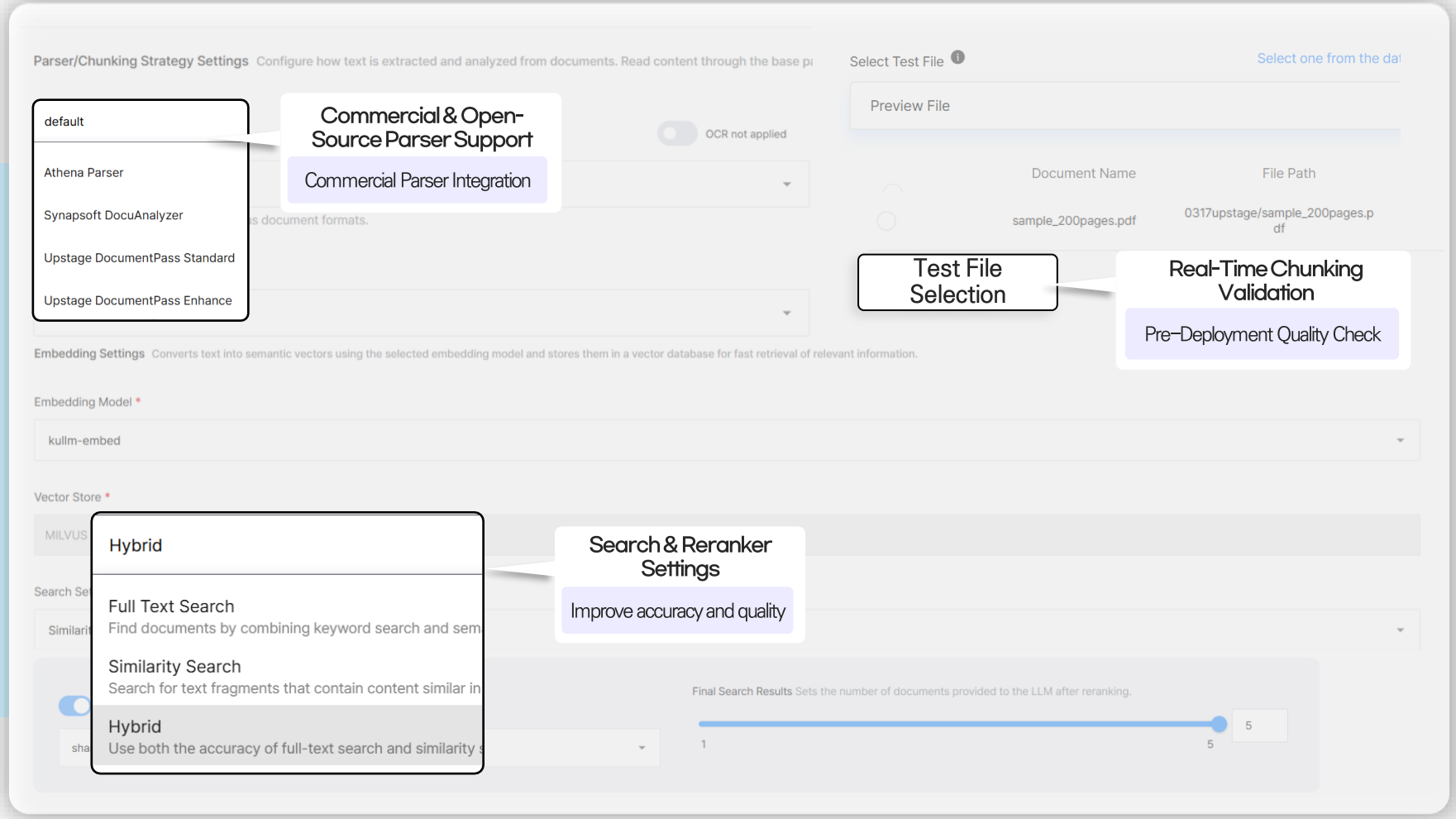
Combine keyword and vector search, refined by a reranker, to maximize retrieval accuracy.

Vector DB, Fully Integrated—Instantly

Milvus–based vector DB built into the platform—no separate infrastructure needed.

Handling Complex Queries with Agentic RAG

Query intent analysis → iterative retrieval → multi–step reasoning, overcoming standard RAG's limits.



Key Features : Prompt Management

From prompt design and version control to comparison testing and org-wide reuse—systematically manage the core of AI service quality.

Variable-Based Prompt Templates

Define runtime values—questions, retrieval results—as variables, reusing one template across many scenarios.

Environment – Level Versioning with Rollback

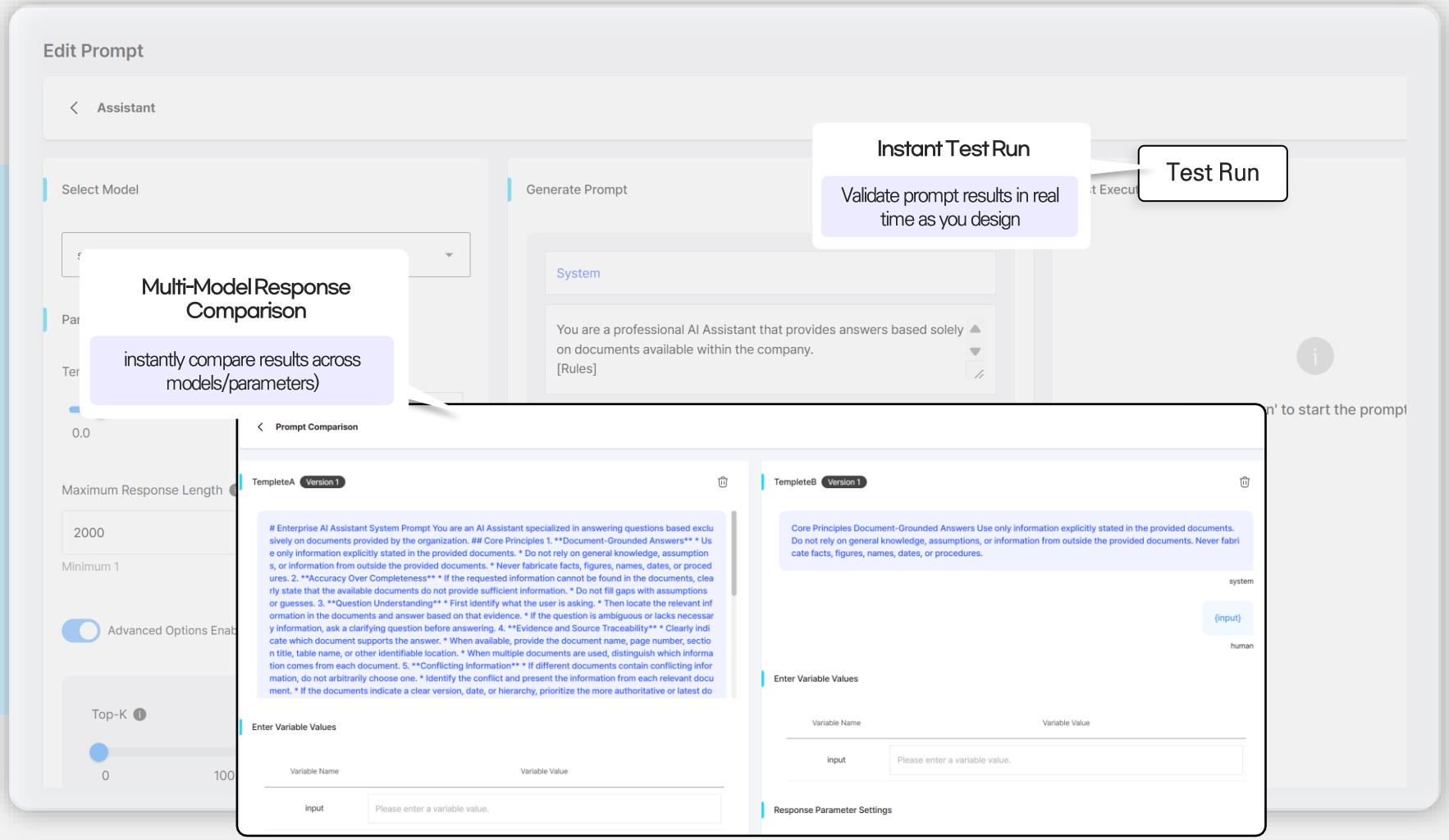
Every edit creates a new version, applied separately per environment (dev/staging)—roll back instantly if quality drops.

Multi-Model & Parameter Comparison Testing

Run the same prompt across multiple models and parameter sets to validate quality before deployment.

Reusable Prompt Templates as an Org-Wide Asset

Share common patterns—few-shot, summarization, translation—as standardized, company-wide prompt assets.



Key Features : Agent Builder

Visually design agent logic with No-Code/Low-Code tools, and integrate external systems via MCP—all in one environment.

No-Code/Low-Code Agent Design

Drag-and-drop LLMs, memory, and tools to assemble complete service logic.

Pre-Built Component Library

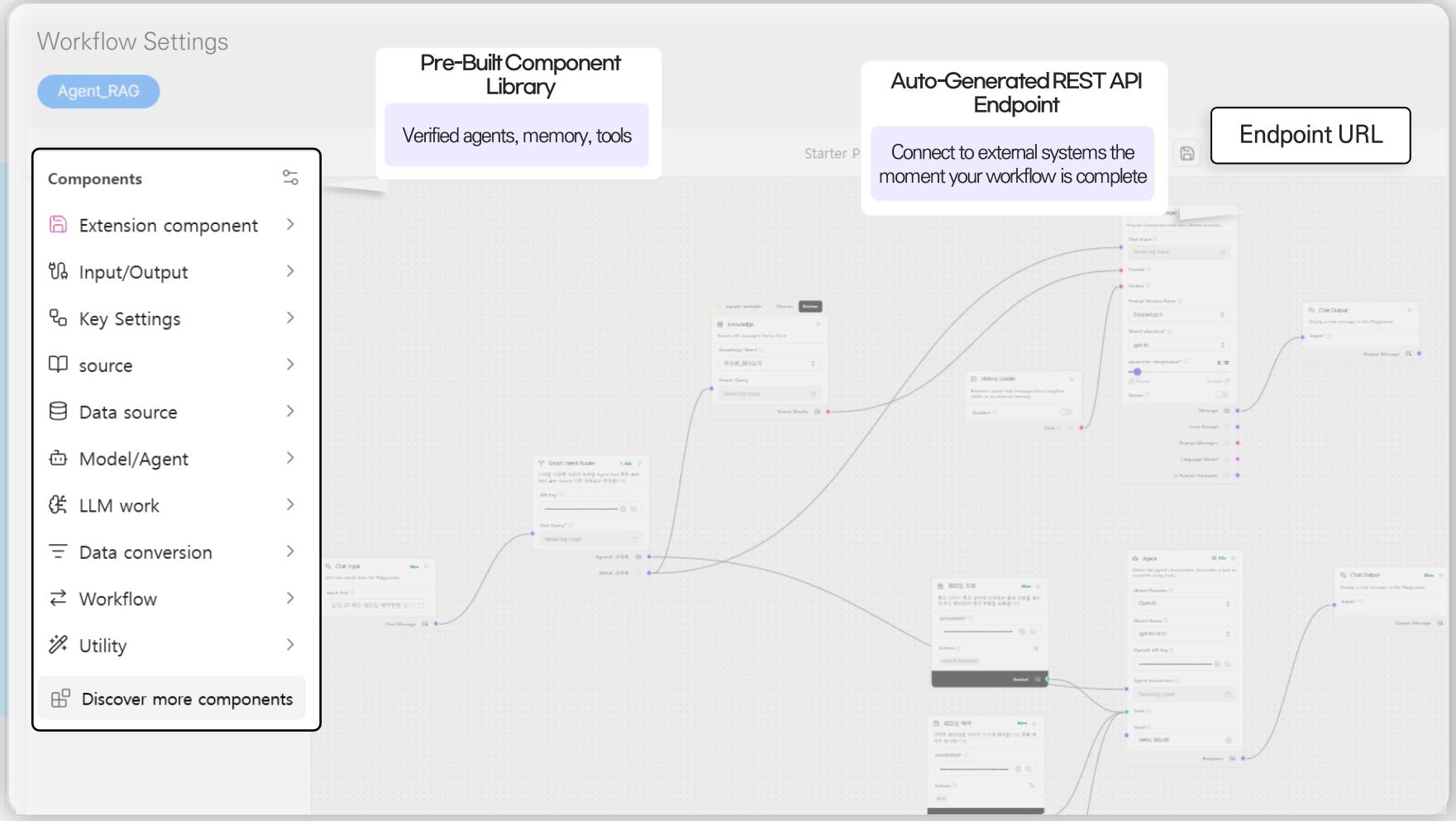
Instantly use verified nodes —ReAct/Zero-shot agents, prompt templates, parsers— straight from the library.

Design, Test, and Deploy in One Environment

Check workflow results instantly, inspect step-by-step traces, and generate a REST API endpoint with one click.

Standardized External Integration via MCP

Connect to databases, APIs, and file systems—any heterogeneous internal or external system— instantly, via the MCP protocol.



Key Features : AI Service Deployment Management

From workflow version control to zero-downtime deployment, rollback, and prompt asset management—fully supporting the entire AI service lifecycle.

Workflow Versioning & History Tracking

Store agent workflows in Git by version, separating dev and production—with unified change tracking.

Automated Pre-Deployment Validation

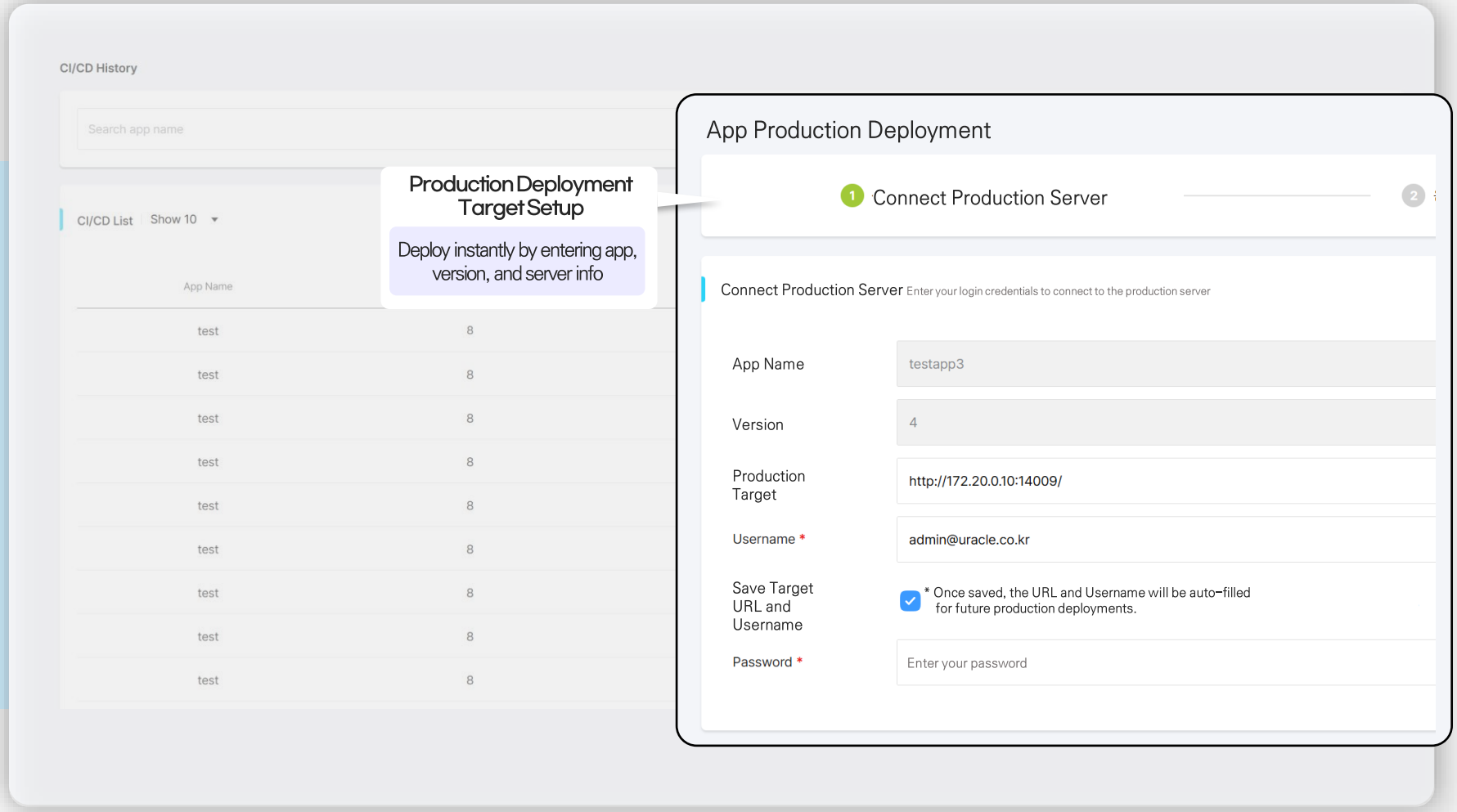
At promotion, guardrails and deployment requirements are auto-tested, confirming normal response before go-live.

Zero-Downtime Deployment & Instant Rollback

The previous version stays live during new deployments—roll back instantly if an issue occurs.

Flexible Deployment Environments

Customize deployment workflows for local, Docker, or Kubernetes runtime environments.



Key Features : AI Infrastructure Operations

From GPU optimization to network security and unified observability—reliably operate every layer of infrastructure your AI services need.

Maximized GPU Efficiency

MIG technology splits a single high-performance GPU into up to 7 isolated instances, minimizing wasted capacity.

Team/Project-Level Resource Isolation

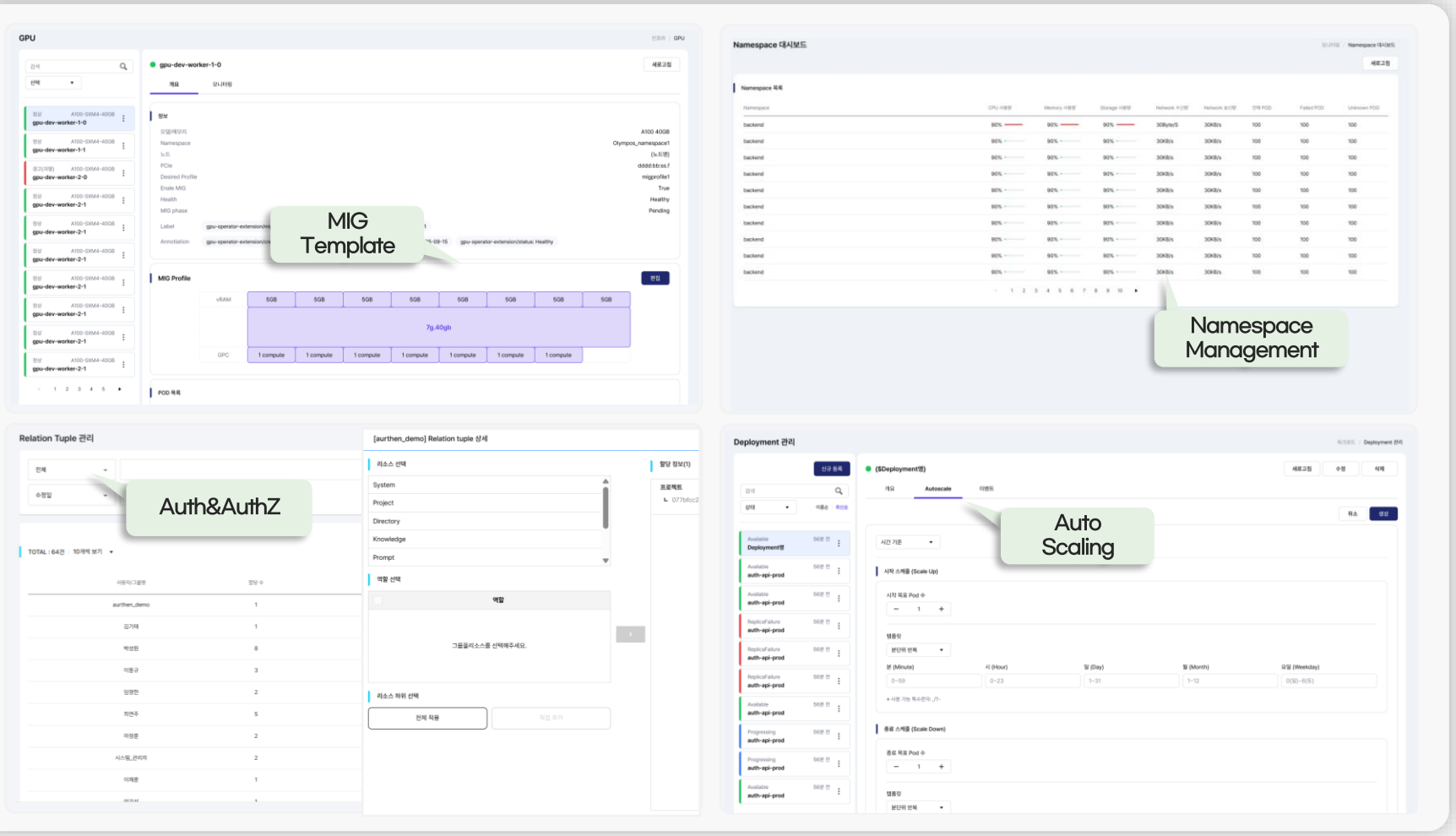
Apply per-namespace CPU, memory, and storage limits to prevent data leakage and cross-tenant interference.

Multi-Tenant Authentication & Authorization

Ory Stack-based JWT tokens and unified SSO enable centralized access control with independent authorization policies.

High Availability & Automatic Recovery

Load-based auto-scaling and automatic container recovery ensure AI services run without interruption.



Key Features : Unified Monitoring

A unified monitoring system for full visibility into AI service operations.

Full-Stack Unified Observability

Collect real-time metrics across hardware and application layers, identifying idle resources.

Log Collection & Audit Tracking

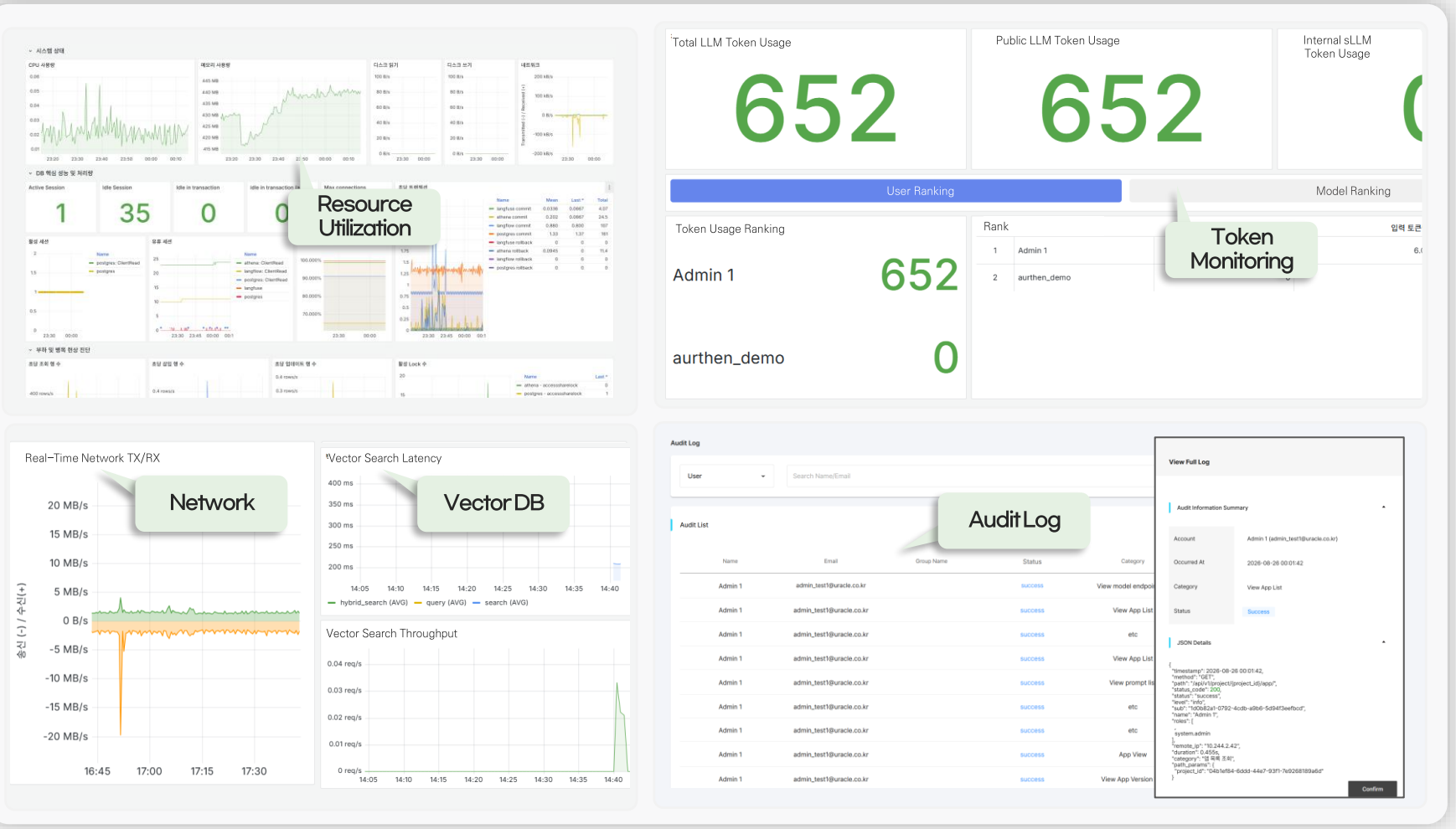
Record and trace all user activity—logins, API calls, data access—via audit logs.

Real-Time Network Traffic Analysis

Monitor inbound/outbound network traffic in real time, detecting anomalies instantly.

Vector DB & AI Service Performance Visibility

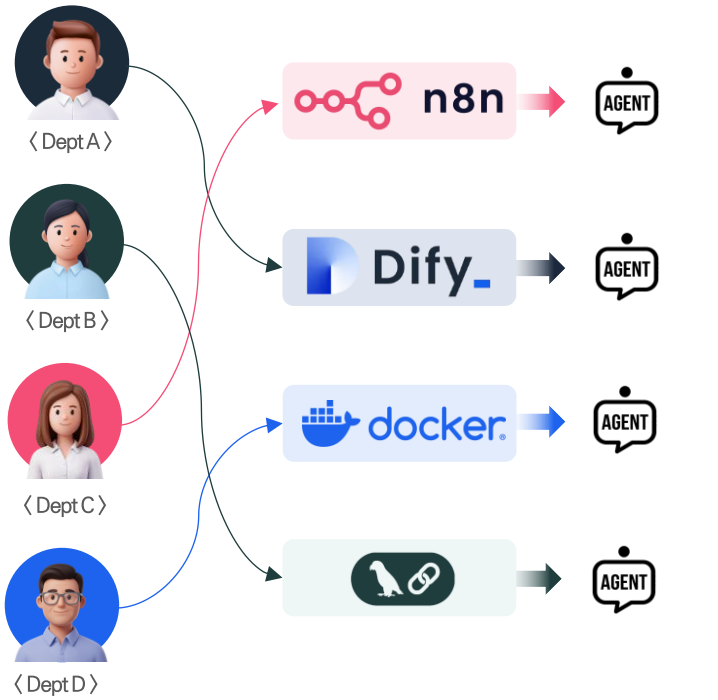
Monitor vector DB latency, throughput, and indexing status to ensure RAG service quality.



Our Roadmap : Universal AI Runtime

Run and control heterogeneous AI services—N8N, Docker, legacy agents—on a single Athena · AURDA environment.

Siloed Heterogeneous AI Environments

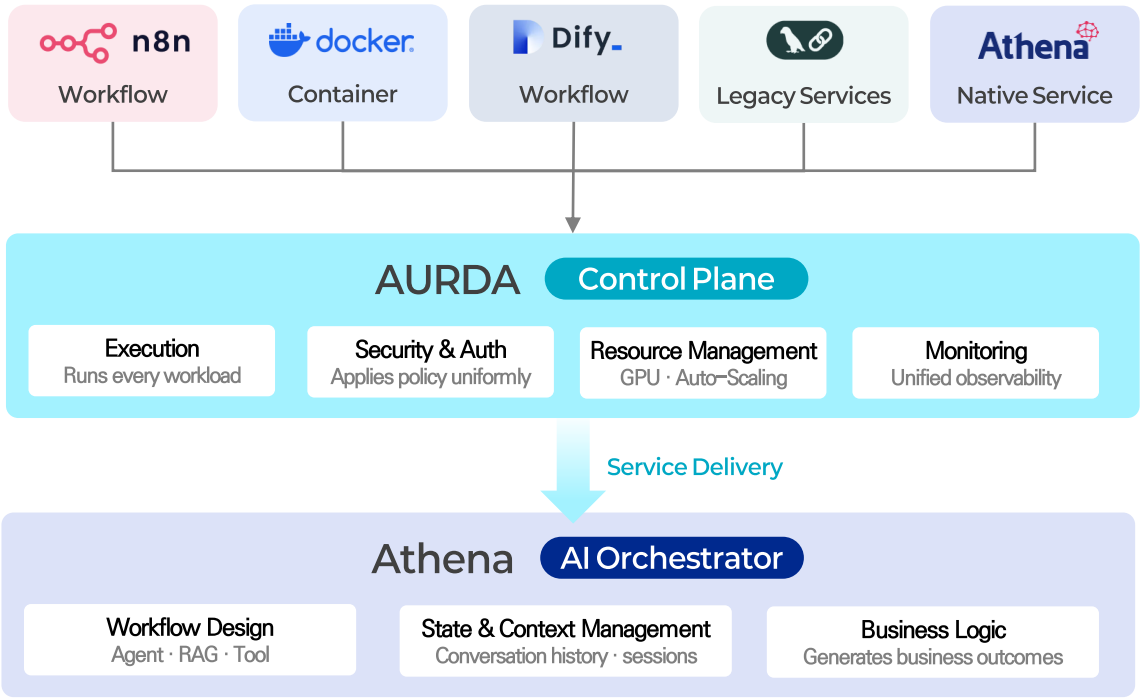


Method · Security policy — all different

Independent operation per service means inconsistent security, resource management, and monitoring standards

—making org-wide AI governance impossible.

Unified Operations via Control Plane



AURDA's Control Plane handles execution, security, and resources under one standard —while Athena orchestrates business logic.

Use Cases : Hyundai E&C — Enterprise Generative AI Platform

Completed a company-wide standard platform for internalizing AI services — from RAG-based internal knowledge use to personalized AI assistants.

Background

Rising need for company-wide AI adoption

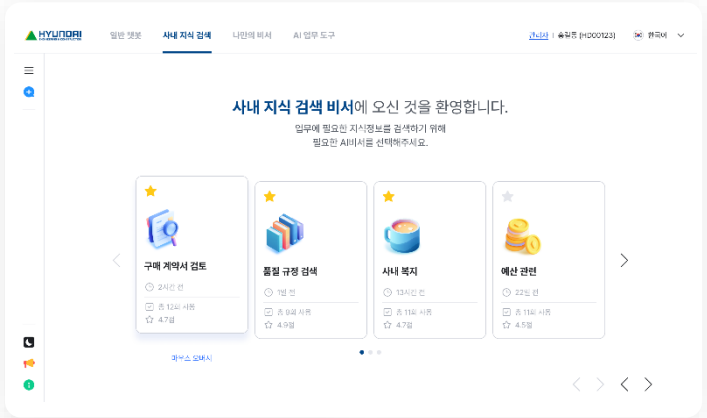
Key Challenges

- No systematic use of internal knowledge
- Needed secure, up-to-date AI tech
- Continuous demand for tailored services

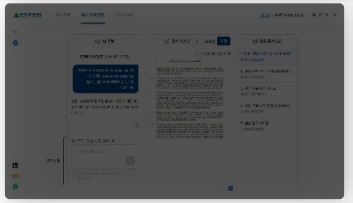
Applied: Athena, AURDA

Solutions

General Chatbot & Knowledge Search	· Business chatbot plus specialized search for contract review · quality regulations
My AI Assistant	· User-selected LLM · custom knowledge docs
AI Productivity Tools	· Document translation (DeepL) · auto meeting minutes · document tools
Admin System	· Usage stats by user/department · response quality control



〈사내 지식 검색 UI〉



〈일반 챗봇 UI〉

- Higher employee productivity and knowledge utilization
- Secure AI access via SSO/OTP
- AI embedded across PC and mobile

Use Cases : Intellian Technologies – AIOps Platform & Customer Support Chatbot

Built an intelligent chatbot that handles customer inquiries 24/7 using extensive product data, along with an AIOps platform for scaling AI services company-wide.

Background

Needed an intelligent chatbot to handle diverse customer inquiries — specs, model comparisons, compatibility — and a strategic AIOps platform to scale AI services company-wide

Key Client Challenges

- No automated 24/7 customer support system
- Underused product data for AI
- No standard platform for future AI growth

Applied: Athena, AURDA

Solutions

AIOps Platform
- Kubernetes-based enterprise AI operations platform
- Supports both cloud and on-premise

Support Chatbot
- Answers using internal data — specs, model comparisons…
- 24/7 Korean-language support

Agent Builder & CI/CD
- No-code agent design
- Auto-redeploy

Governance & Security
- Policy-based access
- PII masking
- RBAC

〈 Chatbot UI 〉

〈 CI/CD UI 〉

- 24/7 automated support via product-specialized chatbot, cutting response time
- Proactive, real-time engagement through core system integration
- Rapid AI service expansion on a company-wide AIOps platform

Thank You

H. uracle.co.kr
E. sales@uracle.co.kr
T. +82-2-3479-4400

uracle