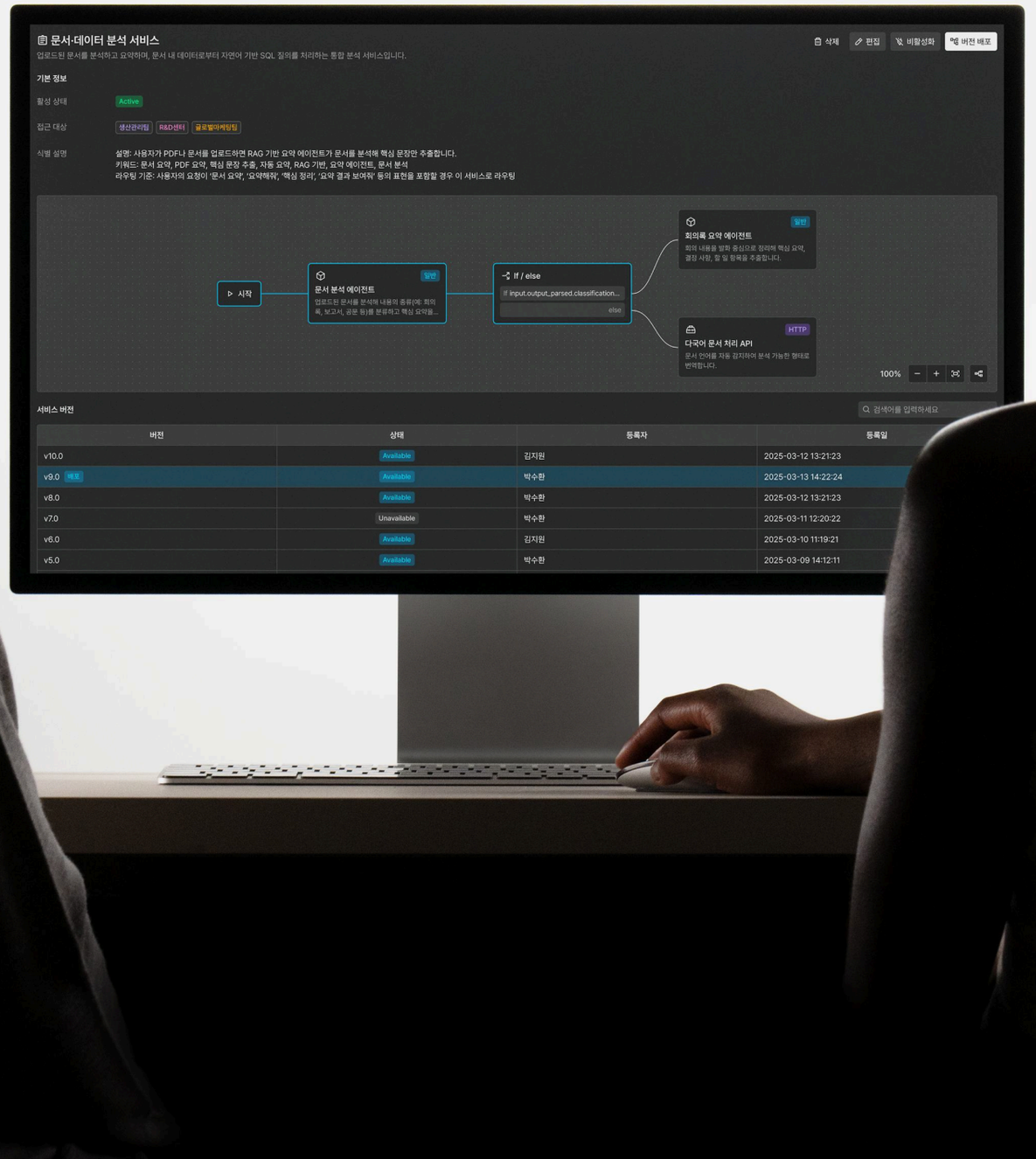
eXemble

Generative AI for Enterprise Data, Optimized for On-Premises

Enterprise AI Operations, Fully Realized On-Premises

The All-in-One LLMOps Platform. From deploying LLMs in air-gapped environments to document vectorization, agent creation, and security auditing — manage every aspect of your AI operations with eXemble. Securely leverage your data's potential to build and refine RAG-based AI services.



eXemble

Why eXemble

From Build to Operations, Unifying the Entire AI Service Lifecycle

As Generative AI adoption accelerates, enterprises confront new challenges in data security, model operations, and cost control. Public and financial institutions, in particular, must protect sensitive data while maintaining stable AI operations within air-gapped environments.

However, public AI services transmit data to external servers, raising significant security concerns and often failing to address internal workflow specificities. eXemble unifies the entire AI service lifecycle—from model deployment and data vectorization to agent creation and security auditing—on a single platform optimized for on premises environments.

Customer Stories

Public Sector

Created AI Chatbot in Closed Environment

eXemble enabled us to build an AI chatbot using internal documents within a network-segregated environment. This led to a 70% reduction in manual processing time, while satisfying security audit requirements through granular access controls.

Financial Sector

Introduced NL2SQL-based Internal Data Query Service

Empowered non-developers to query internal databases using natural language via NL2SQL agents. Reduced dependency on data analysts by 50% and significantly accelerated data-driven decision-making.

Product Highlights

Enterprise AI for Air-Gapped Networks

eXemble provides secure AI that fully complies with security regulations by providing LLM serving and RAG pipelines in a closed environment.

Feedback-based Quality Enhancement

Automatically proceeds RLHF based on real-time user interactions and upgrades models for continuous improvement.

Automated Data Ingestion & Intelligent Vectorization

Vectorize formats (ex. PDFs, HWP, Image) and Maintain accuracy with auto re-vectorization when source data changes.

Fast Inference & GPU Optimization

Process high-volume requests with efficient serving engine and reduces costs through GPU anomaly detection.

No-Code AI Agent Workflow Design

Visually design workflows by connecting AI agents and tools via drag-and-drop. Build, test, and deploy with ease.

Unified Dashboard for Complete AI Observability

Visualize from AI performance metrics including monitor usage and latency to underlying infrastructure, all on a single panel.

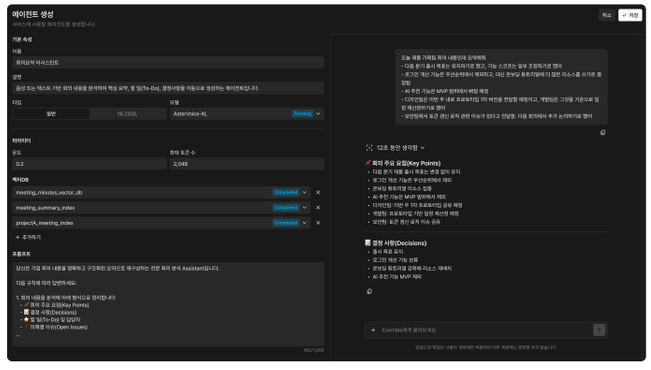
Agent Configuration

Direct creation of task-specific AI agents without coding knowledge. Visual design of complex AI services via a drag-and-drop workflow canvas combining agents and external tools.

1 Agent Builder

Effortless creation of purpose-built AI agents

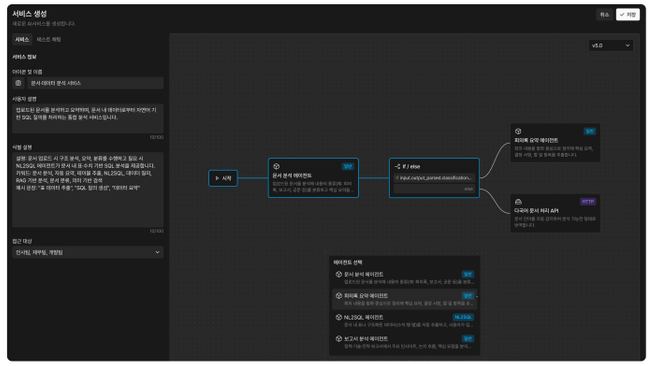
- Create agent with models, vector tables, and prompts
- Immediate validation of changes via real-time test chat
- Continuously refined by user feedback



2 Workflow Canvas

Complex AI made possible with node-based design

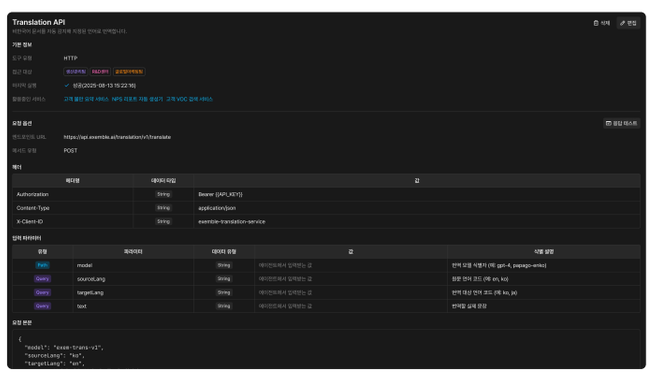
- Drag-and-drop integration between agents and tools
- Real-time workflow execution and tool invocation
- Easy root cause analysis through visual verification



3 External Tool Integration

Managing APIs and internal systems as reusable tools

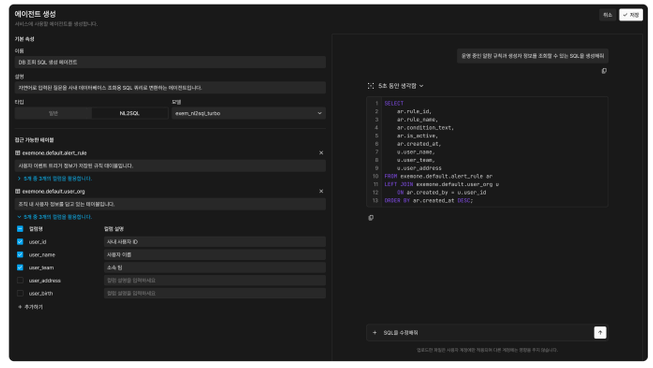
- Supports HTTP APIs and MCP Tools for reuse
- Instant functionality verification via test execution
- Dynamic parameters handling via AI or user input



4 Text-to-SQL Chart Generation

Automated natural language-to-SQL conversion

- Internal DB querying accessible to non-developers
- Expansion of data utilization scope
- Accelerated workflow via query result-based responses



Ops Monitoring

Key AI service metrics—usage, cost, and latency—all on a unified dashboard. Automatic version control for agents and services, enabling immediate deployment of optimized versions.

1 Unified Dashboard

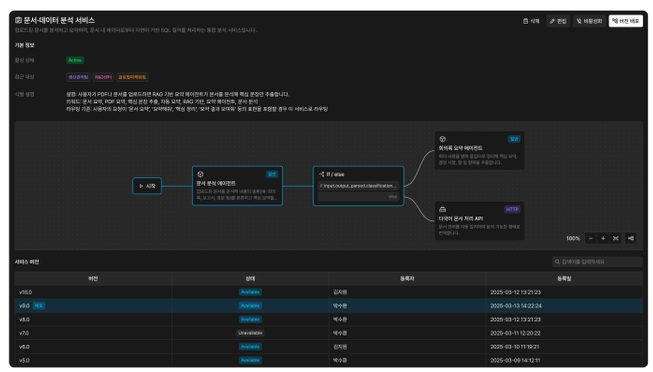
Monitor your AI service on a single interface

- Real-time tracking of service usage, costs, and latency
- Drill-down capability for specific service metrics
- Strategic capacity planning via performance trend

2 Auto Version Management

Tracking and managing all modification histories

- Automatic version generation triggered by changes
- Tracking modified values compared to previous versions
- Instant, seamless transition to desired version



3 Agent Evaluation

Optimal agent selection via comparative analysis

- Quantitative accuracy assessment based on datasets
- Automatic identification of top-performing agents
- Single-click deployment of highest-performing version

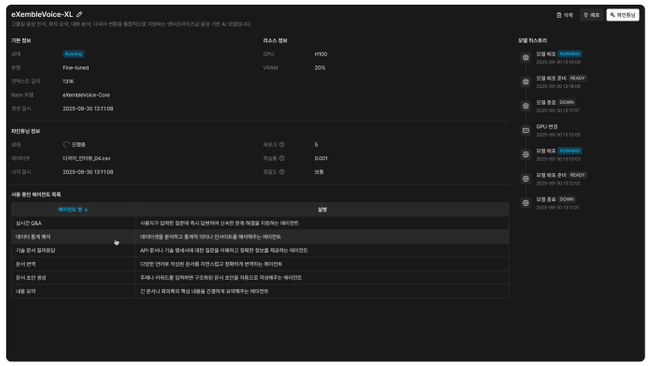
Model Operations

Stable LLM serving via high-efficiency engines. Simultaneous multi-model operation with automatic request routing for optimized operational efficiency and response quality.

1 Model & Version Control

Systematic management of all platform models

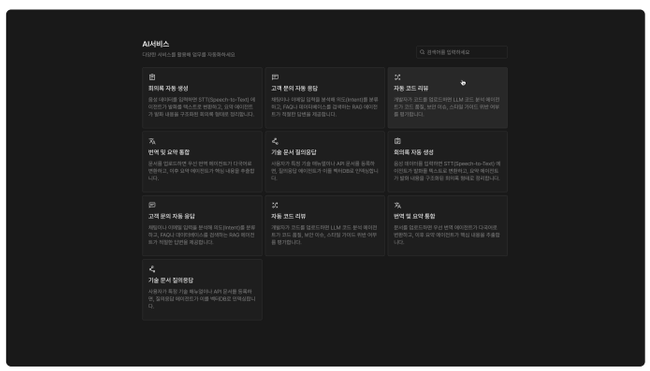
- Real-time access to model info and deployment status
- Complete change history preservation
- GPU selection and VRAM allocation for optimal serving



2 High-Efficiency Inference

Fast, stable processing of high-volume requests

- Enhanced processing speed and reduced operational costs
- Stable response speed even under heavy requests
- Availability assurance through automatic load balancing



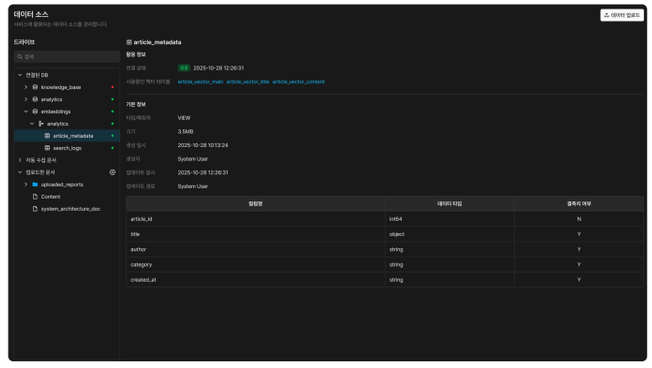
Data Unification

Centralize your scattered data assets—documents, databases, and ingested files—in one place. Automatically converts data into AI-ready vector indices and ensures up-to-date accuracy through real-time synchronization.

1 Data Source Integration

Explore and manage data assets on a single interface

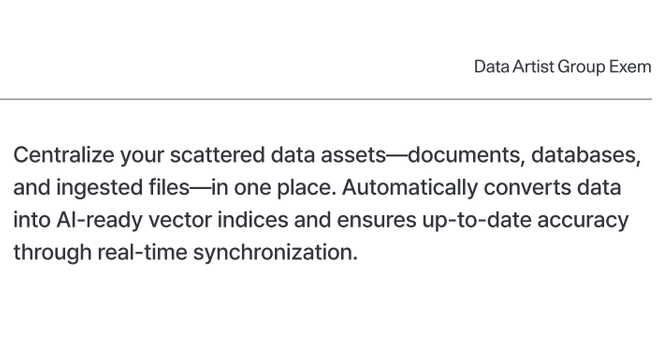
- Structure DBs and docs into a single hierarchy
- Real-time DB connection and vector usage monitoring
- Schema, table, folder, and file-level navigation



2 Automated Vectorization

Raw data into AI-understandable formats

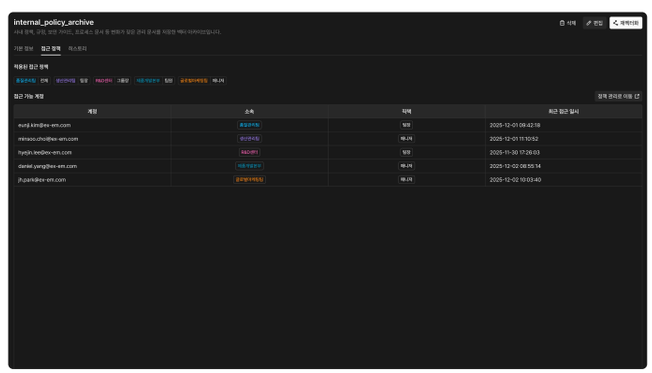
- Text extraction from PDFs, HWPs, and images
- Vectorizing from files to specific DB query results
- Automatically detect and update data changes



3 AI Search History Management

Vector table-based secure knowledge base

- Task status-based grouping and full history tracking
- Department and role-based access control
- Supports manual, operator-led re-vectorization



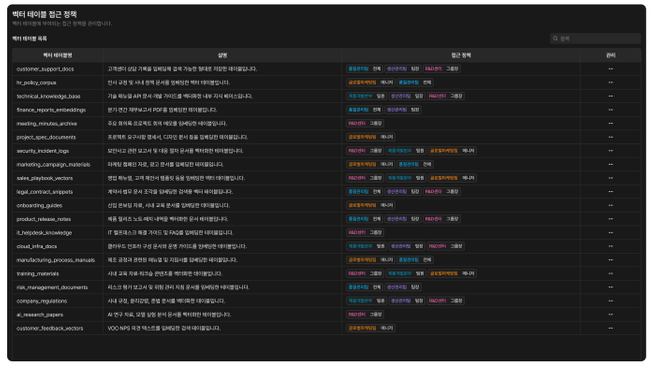
Security & Access

Protection of sensitive information via granular access controls at user and organizational levels. Clear limitation of data access scope by department and rank, with seamless alignment to internal security policies.

1 Access Control

Precise configuration based on affiliation and position

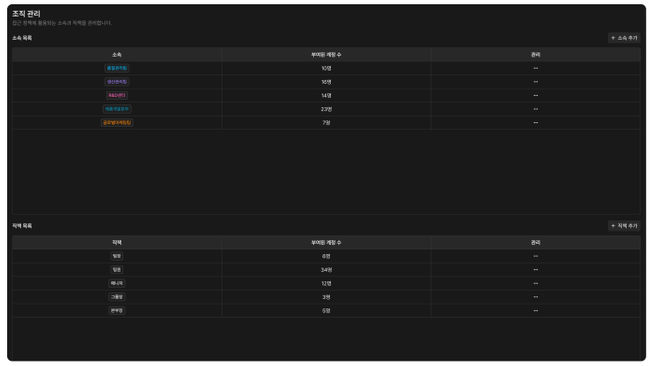
- Individual vector table access policy assignment
- Role-based management interface access restriction
- Account lockouts and permission history management



2 Organizational Management

Systematic user classification

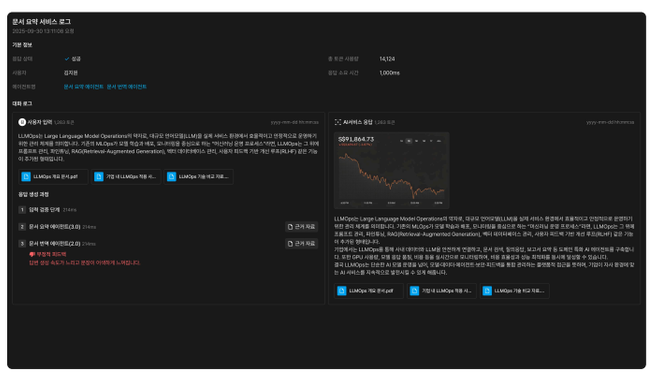
- Department and position-based hierarchy setup
- Batch permission assignment by organizational unit
- Preventing sensitive info exposure via data isolation



3 Secure AI Response Generation

Response scope configuration and history tracking

- Automatically blocking sensitive / inappropriate requests
- System prompt-based response boundary setting
- Comprehensive auditing tracking blocked responses



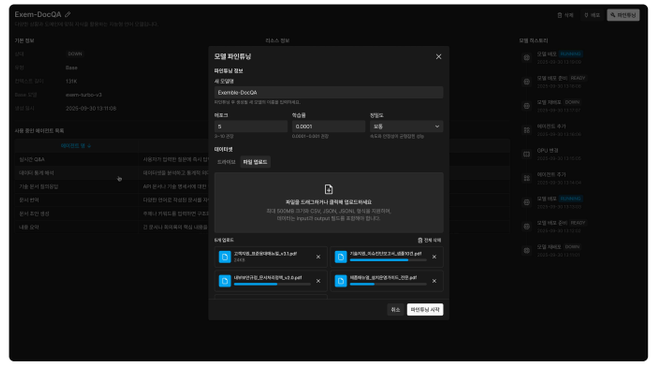
Quality Enhancement

Continuous elevation of AI quality via domain-specific fine-tuning and RLHF. Implementation of trustworthy AI services through transparent provision of reference sources for every response.

1 Fine-Tuning

Performance adjustment using domain-specific data

- Automatic format error inspection upon dataset upload
- Real-time monitoring of the fine-tuning status
- Automatically register training completed models



2 Feedback-Driven RLHF

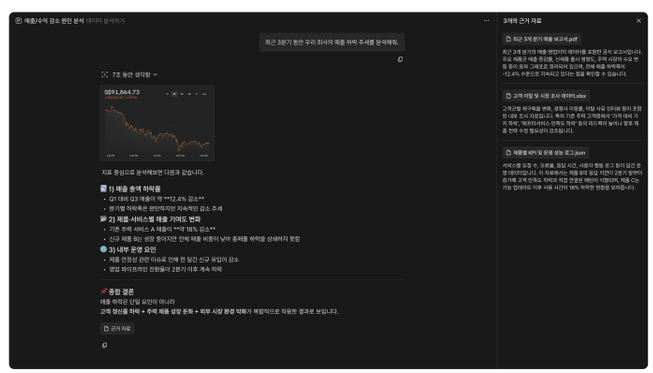
Ongoing enhancement reflecting user evaluations

- Utilize positive/negative feedback as training data
- Automatically integrate feedbacks and update models
- Add newest agent versions seamlessly

3 Source Traceability

Transparent display of sources behind AI responses

- Display references and sources used in responses
- RAG-based output reliability verification support
- Backtrack source to minimize hallucinations



Architecture

1 Data Source

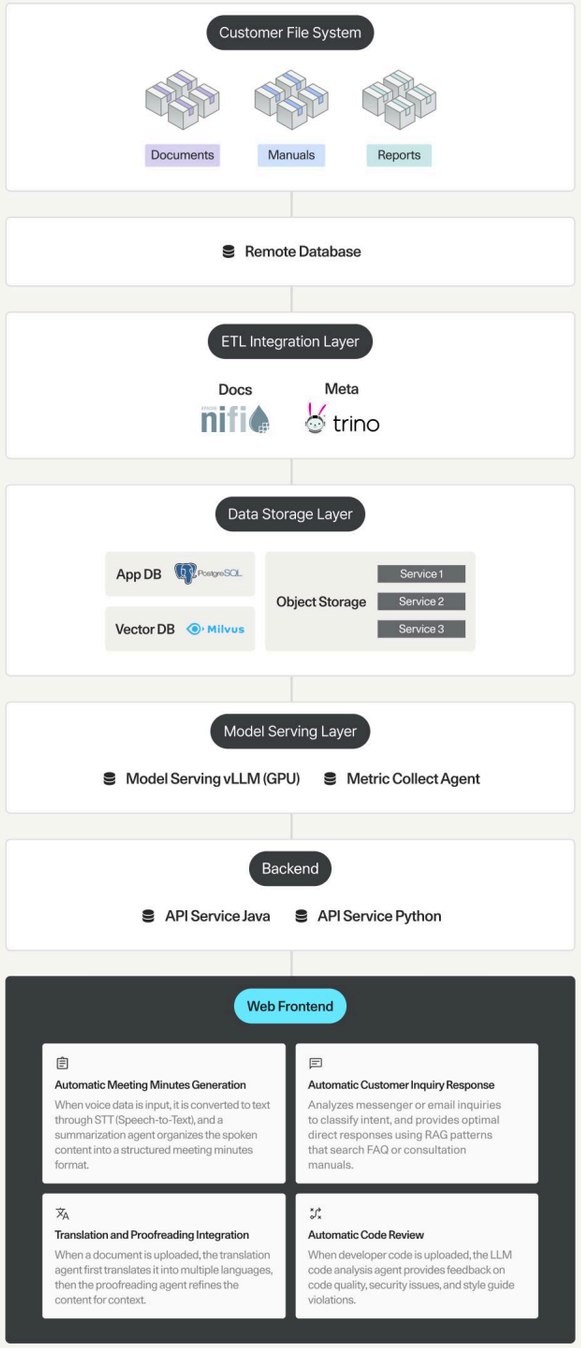
- Leverage diverse enterprise data assets (documents such as manuals and reports, as well as databases from business systems and operational DBs.)
- Customer data remains in existing systems, utilized without any direct modification to the original source.

2 Data Gathering and Processing

- Automated integration of diverse data sources through data ingestion pipelines.
- Distributed query engine for big data, enabling large-scale data retrieval and analytics.
- Automatic reprocessing triggered upon data change detection.

3 AI Service Embodiment

- High-performance LLM serving that maximizes GPU resource efficiency.
- Users can easily build task-specific AI services and chatbots through a canvas-based workflow.
- Automatic collection of logs, performance metrics, and usage history for every request.
- Consistent enforcement of security policies, access controls, and audit logs at the service level.



Platform Specs

Web Browser

Supported Browsers: Chrome, Edge
Recommended Resolution: 1920×1080 / Minimum Resolution: 1440×900

Platform Server Requirements

* May vary depending on client data size and retention period

Platform Operations Server

OS: Rocky Linux 8.10 or higher
Kernel: 4.18.0 or higher (Rocky Linux) / 5.15.0 or higher (Ubuntu)
CPU: Recommended 16 Core / Minimum 8 Core
Memory: Recommended 64GB / Minimum 32GB
Disk (DB): Recommended 500GB / Minimum 300GB

Model Serving Server
(Based on 1,000 concurrent users)

OS: Rocky Linux 8.10 or higher
Kernel: Recommended 5.15.0 / Minimum 4.18.0
CPU: Recommended 48 Core / Minimum 24 Core
Memory: Recommended 256GB / Minimum 128GB
GPU: Recommended H100 × 1 / Minimum A30 × 4
Disk: Recommended 4TB / Minimum 2TB

Data Collection Server

OS: Rocky Linux 8.10 or higher
CPU: Recommended 48 Core / Minimum 24 Core
Memory: Recommended 256GB / Minimum 128GB
OS Disk: Recommended 200GB / Minimum 100GB
Disk: Recommended 4TB / Minimum 2TB
*OS area and data area must be separated

Distributed File System
(3-Node cluster configuration)

OS: Rocky Linux 8.10 or higher
CPU: Recommended 24 Core / Minimum 4 Core (per node)
Memory: Recommended 256GB / Minimum 64GB (per node)
Disk: Recommended 1TB × 3 (total 3TB) / Minimum 100GB × 3 (total 300GB)

www.ex-em.com

Data Everywhere,
Make it Matter

40, Magokjungang 8-ro 5-gil, Gangseo-gu, Seoul, Korea
TEL 02-6203-6300 **FAX** 02-6203-6301

exem