



AI / ML Reference Architecture Overview

Mike Rau

SVP, Enterprise Technical Strategy

Mark J Menger

Solution Architect, Business Development

Paul Pindell

Principal Solution Architect, Business Development

Ian Lauth

Senior Manager, Product Marketing for AI

Alysia Groves

Sr. Business Manager, Business Development

Eric Ji

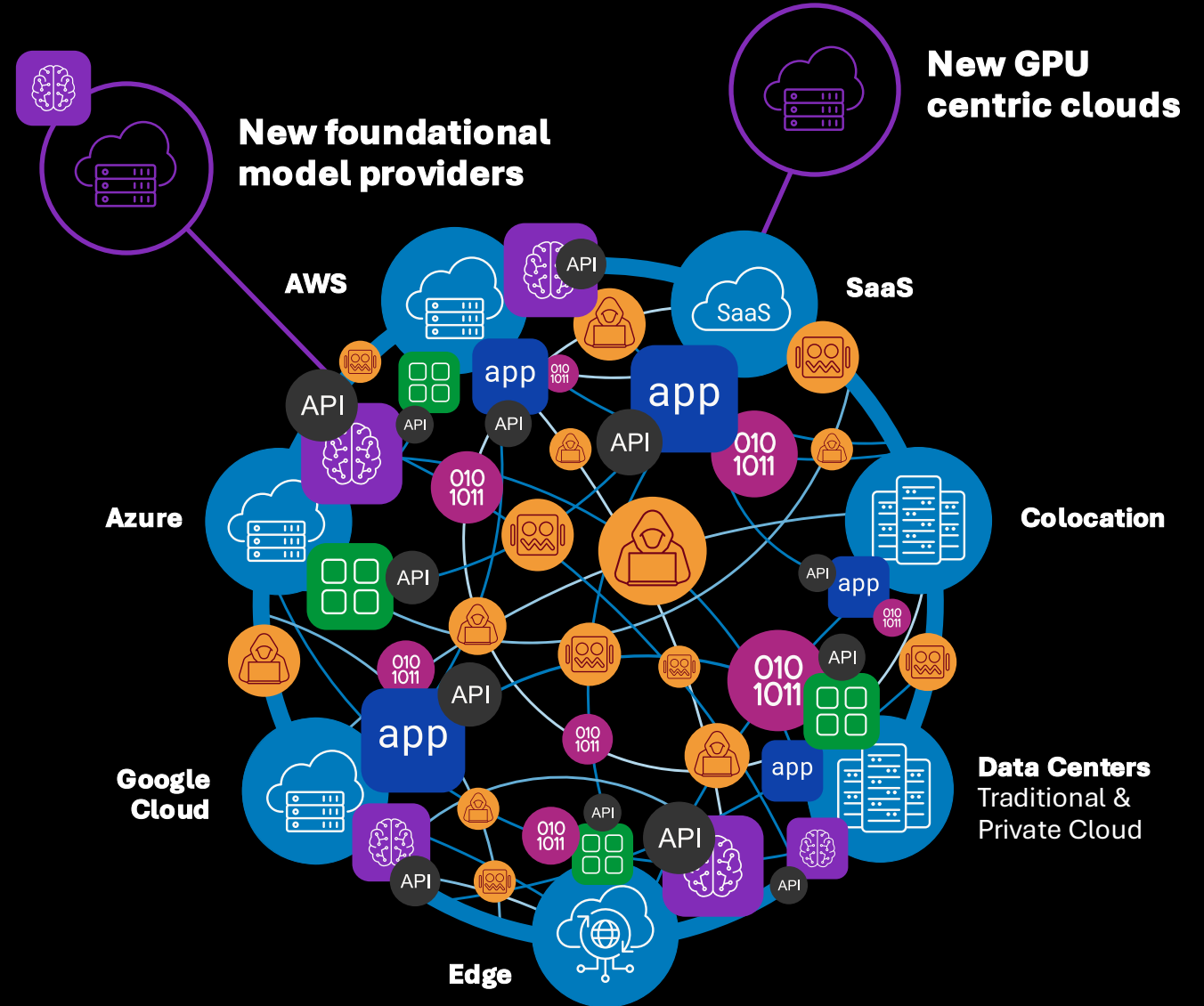
Senior Solution Architect, Business Development

Gregory Coward

Senior Solution Architect, Business Development

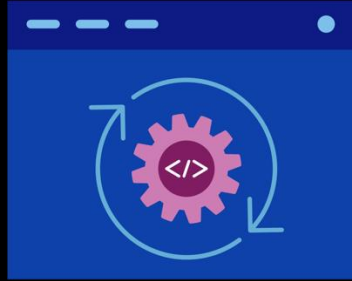
Generative AI threatens to make this scary complexity even more acute

- 1 Generative AI app experiences will be **multi-modal**
- 2 Generative AI apps will be highly **decomposed**
- 3 “**Data gravity**” will significantly influence placement of apps and models
- 4 Generative AI apps will be especially dependent on **APIs**





What are your objectives?



Are you building an **AI Product** or delivering **Operational Efficiency**?



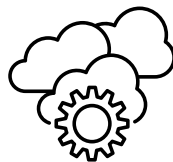
Do you want to **build, buy**, or **out-source** the solution?



How mature is your AI practice?
Are you **exploring, integrating**, or **transforming**?



Four Deployment Models



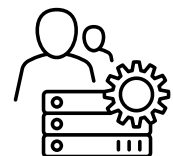
SaaS AI

The AI solution is provided as a **fully managed service** by a third-party provider. Customers can access and use the AI capabilities over the internet without worrying about the underlying infrastructure, maintenance, or updates, making it a **convenient and scalable option**.



Cloud-Hosted AI

The AI solution runs on cloud infrastructure provided by cloud service providers such as AWS, Google Cloud, or Azure. It offers **flexibility, scalability, and ease of integration** with other cloud services, while the **customer maintains control** over the configuration and management of their AI systems.



Self-Hosted AI

The AI solution is **deployed on the customer's own infrastructure**, such as on-premises servers or private data centers. This provides maximum control and customization options but **requires significant resources** for setup, maintenance, and management of the hardware and software components.



Edge-Hosted AI

The AI solution in an edge environment, **outside traditional cloud or data center infrastructure**. An example is a machine learning solution operating on a device like a kiosk in a retail storefront. This model **reduces latency, enhances privacy, and ensures real-time processing** by bringing the computation closer to the data source or end-user.

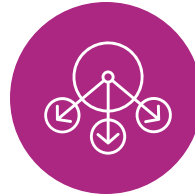


AI Ecosystem Considerations



OWASP LLM Top Ten

Educate developers, designers, architects, managers, and organizations about the potential security risks when deploying and managing LLM and Generative AI applications.



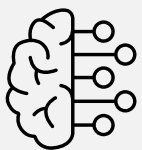
F5 Application Delivery Top Ten

The top unforeseen challenges that arise in today's hybrid multicloud application delivery model cause by too many point solutions, a lack of interoperability, multiple management consoles and manual complexity.



Design Requirements

Define the essential capabilities, technologies, and principles needed to address technical challenges and ensure effective solution implementation.



Seven AI Building Blocks

In this deck we will be showing two of the seven building blocks.

For access to the full deck, please reach out to your F5 account team or email businessdevelopment@f5.com

Web Apps & APIs



Inference



**Retrieval-Augmented
Generation**

Focus Area



**Agentic External
Services Integration**

Hybrid Multicloud & Data Ingest



**RAG Corpus
Management**

Focus Area



Fine-Tuning



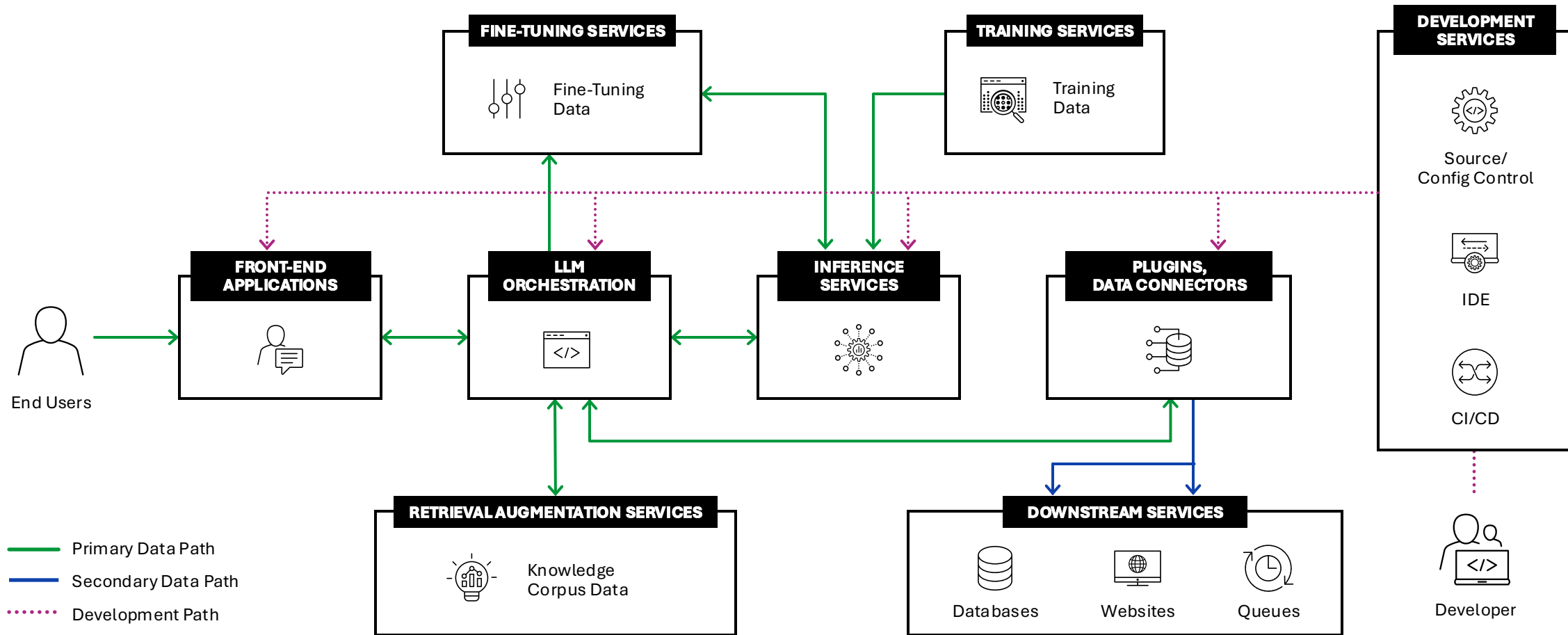
Training

App Development



Development

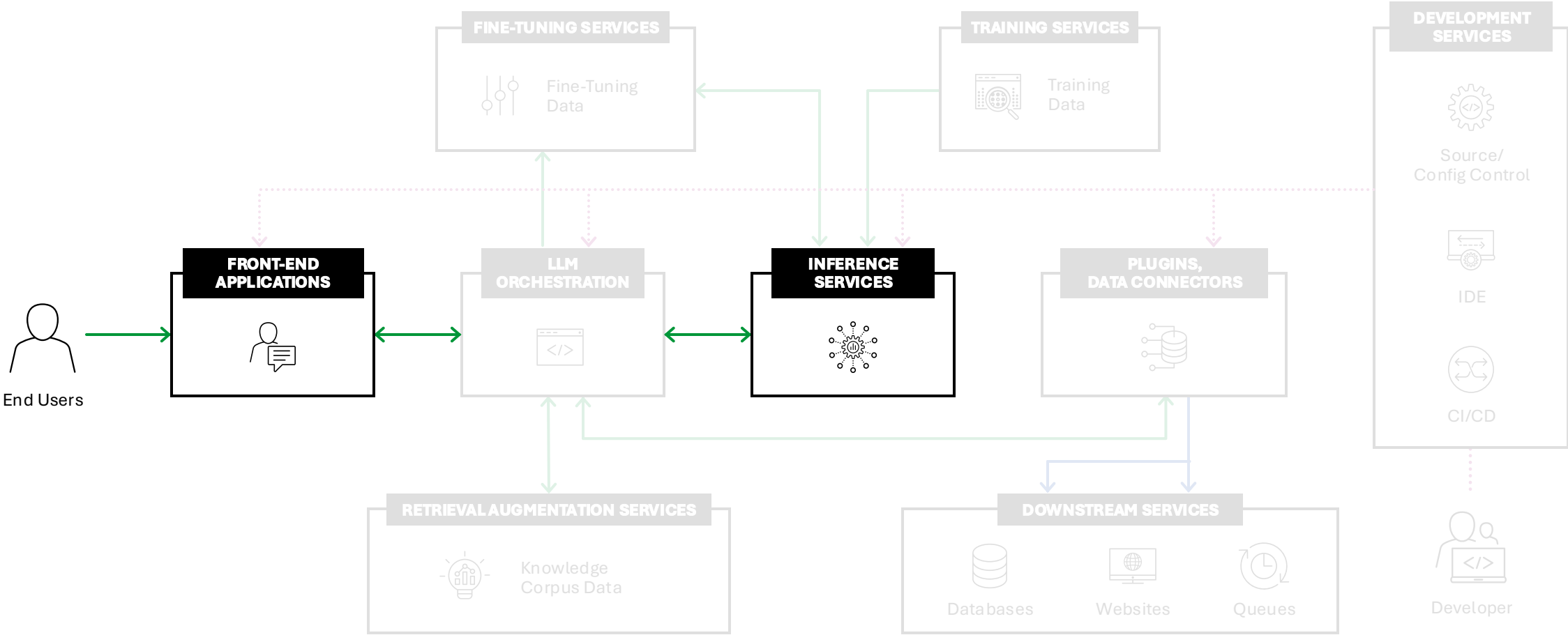
AI Component Architecture



Seven AI Building Blocks

Inference

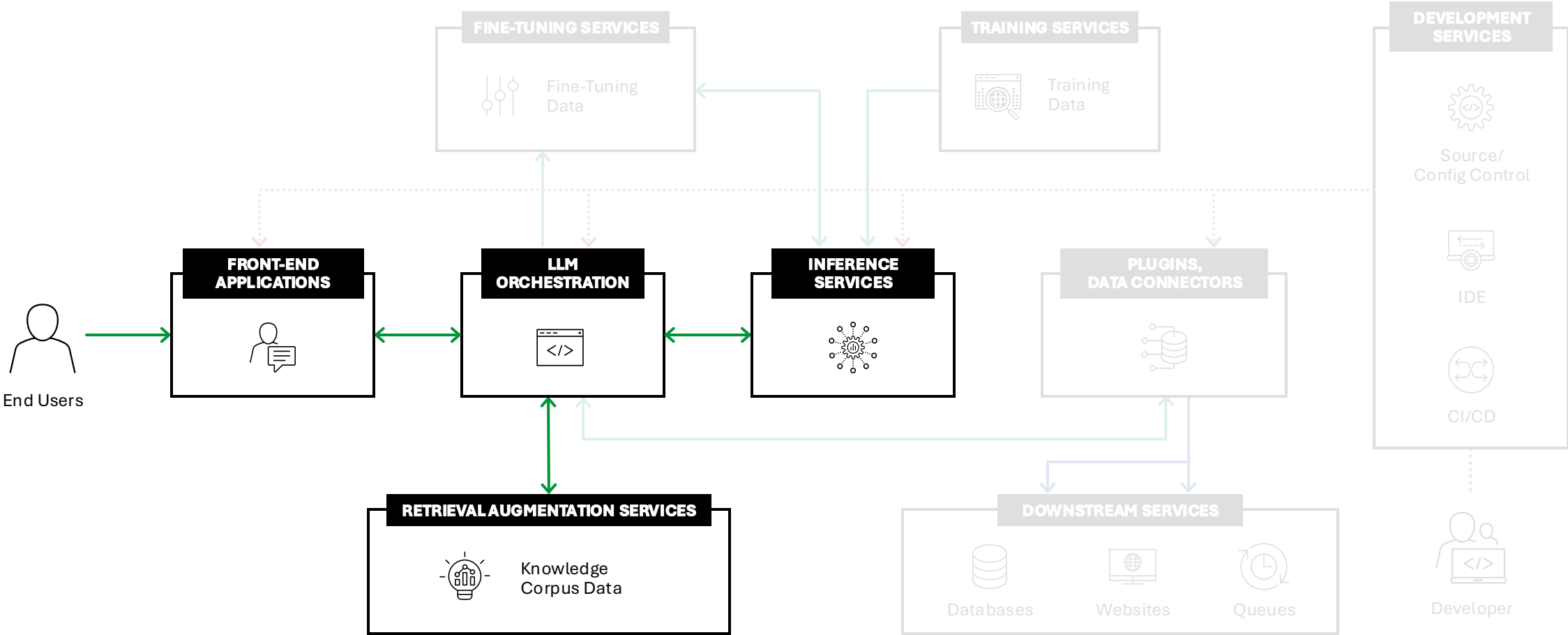
This building block involves the process of making predictions or generating outputs based on input data using pre-trained AI models. It's the core function where the AI system applies its learned knowledge to new, unseen data.



Inference with Retrieval Augmented Generation (RAG)



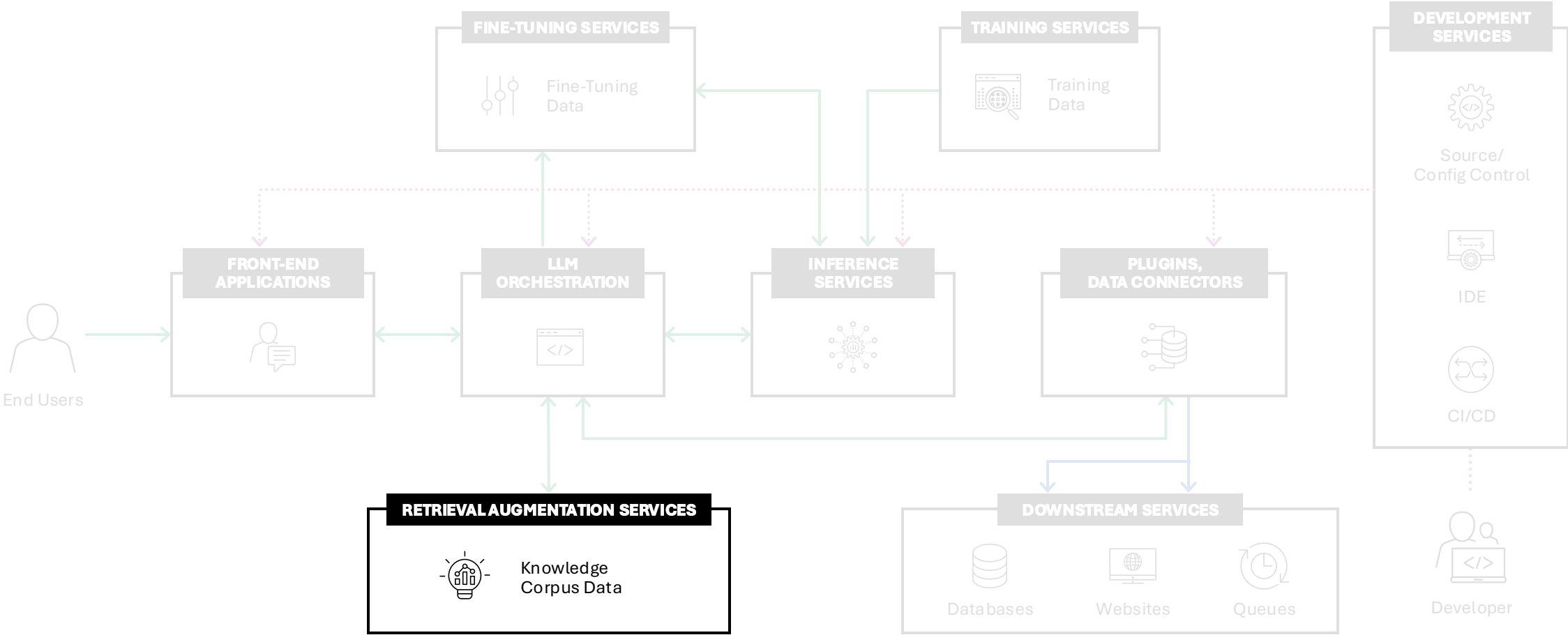
RAG combines the capabilities of retrieval and generation models to produce more informed and accurate responses. It retrieves relevant information from a predefined corpus and uses it to enhance the generation process, resulting in more contextually appropriate outputs.



RAG Corpus Management



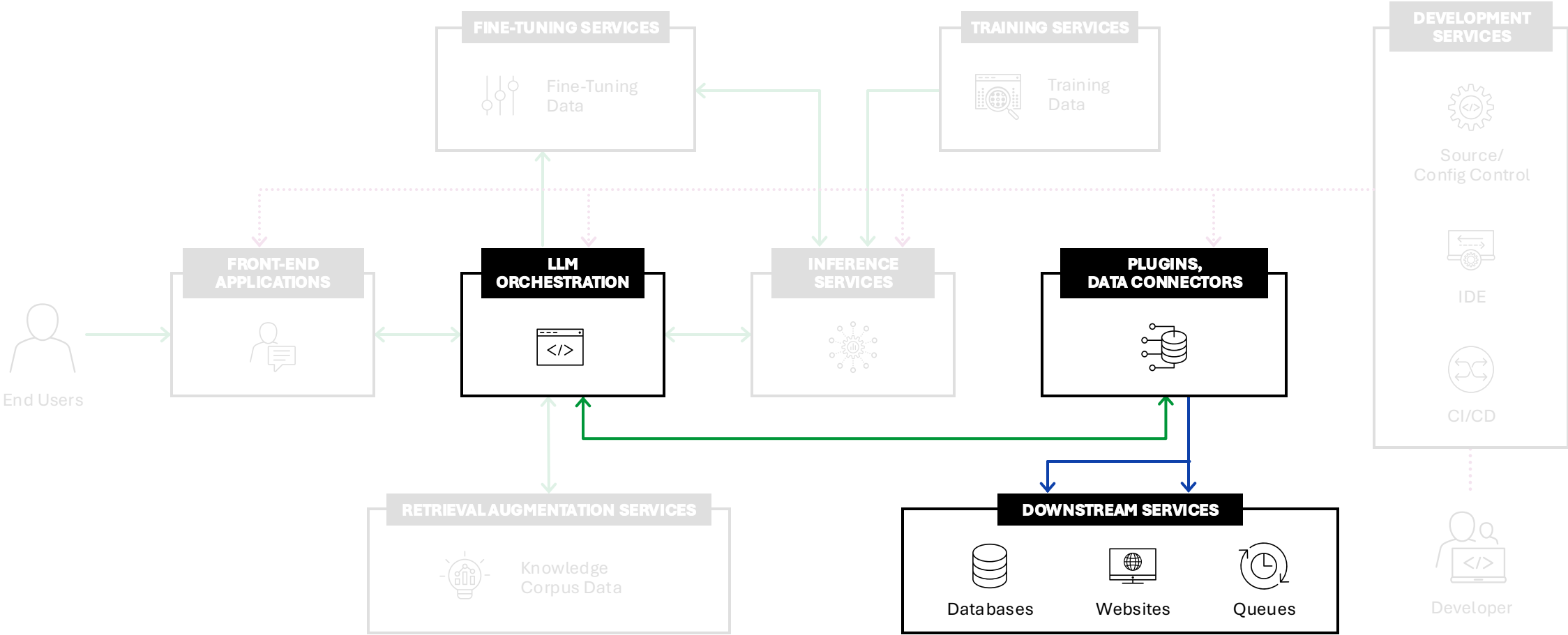
This focuses on maintaining and curating the database or corpus of information that the AI system uses for Retrieval-Augmented Generation. It includes updating, organizing, and ensuring the quality of the data to support accurate and relevant retrieval.



External Services Integration



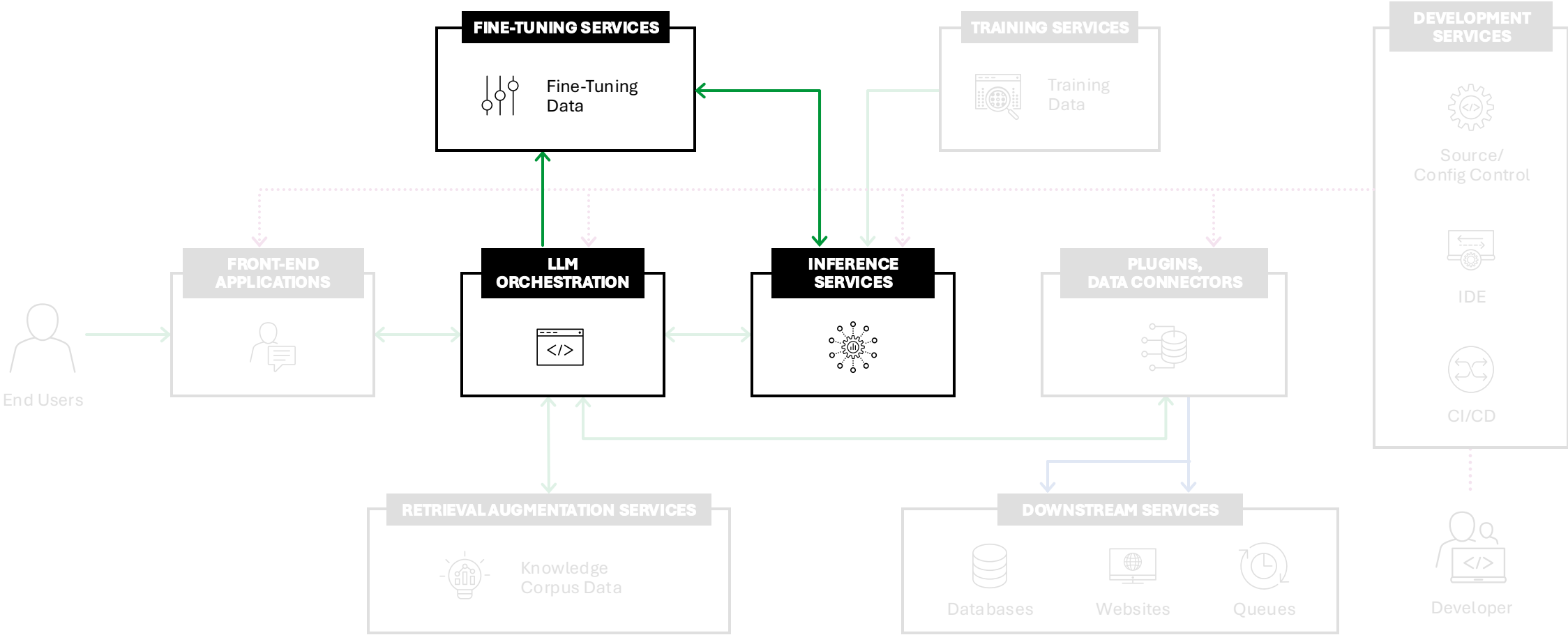
This involves connecting the AI system with external services and APIs, enabling it to interact, retrieve data, or perform actions based on user requests or model inference. It allows the AI to leverage external tools and databases to extend its functionality and autonomously make decisions or take actions as necessary.



Fine-Tuning

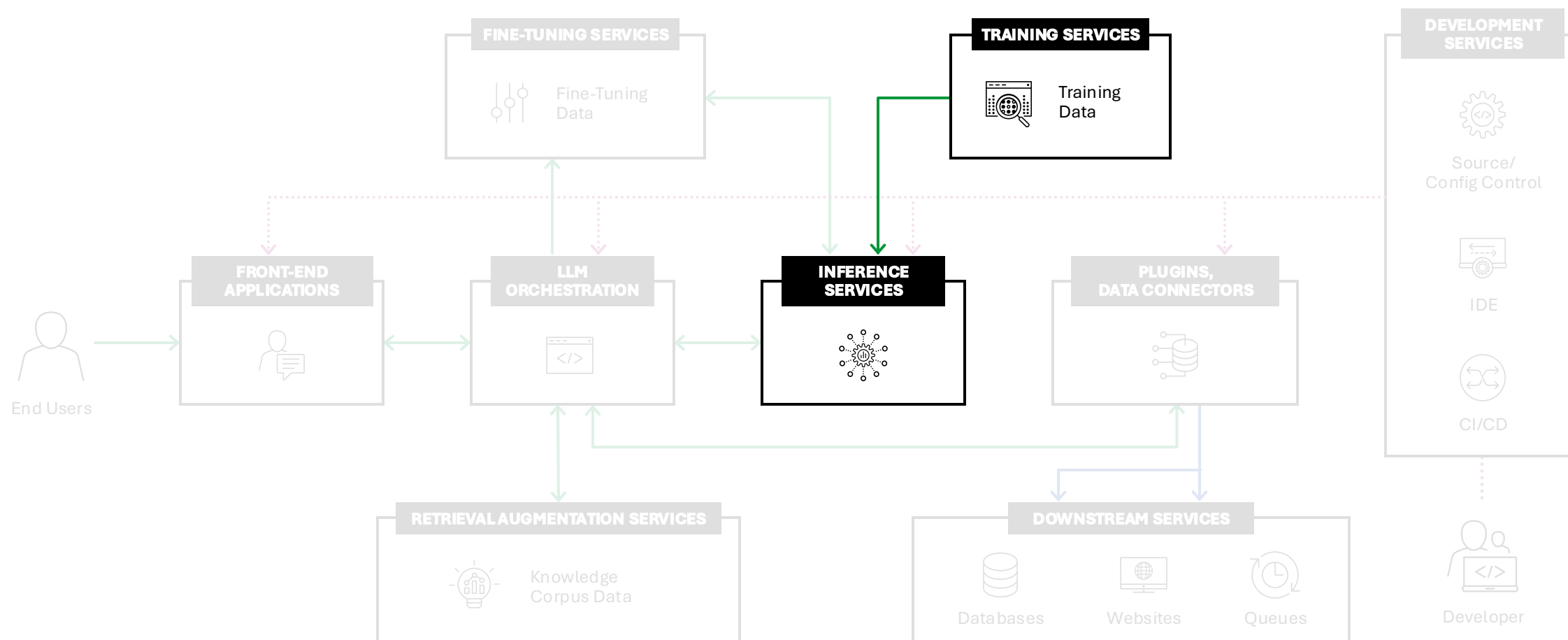


This process involves adjusting a pre-trained AI model on specific datasets to improve its performance for a particular task or domain. Fine-tuning helps tailor the model's capabilities to better meet the unique needs of specific applications or industries.



Training

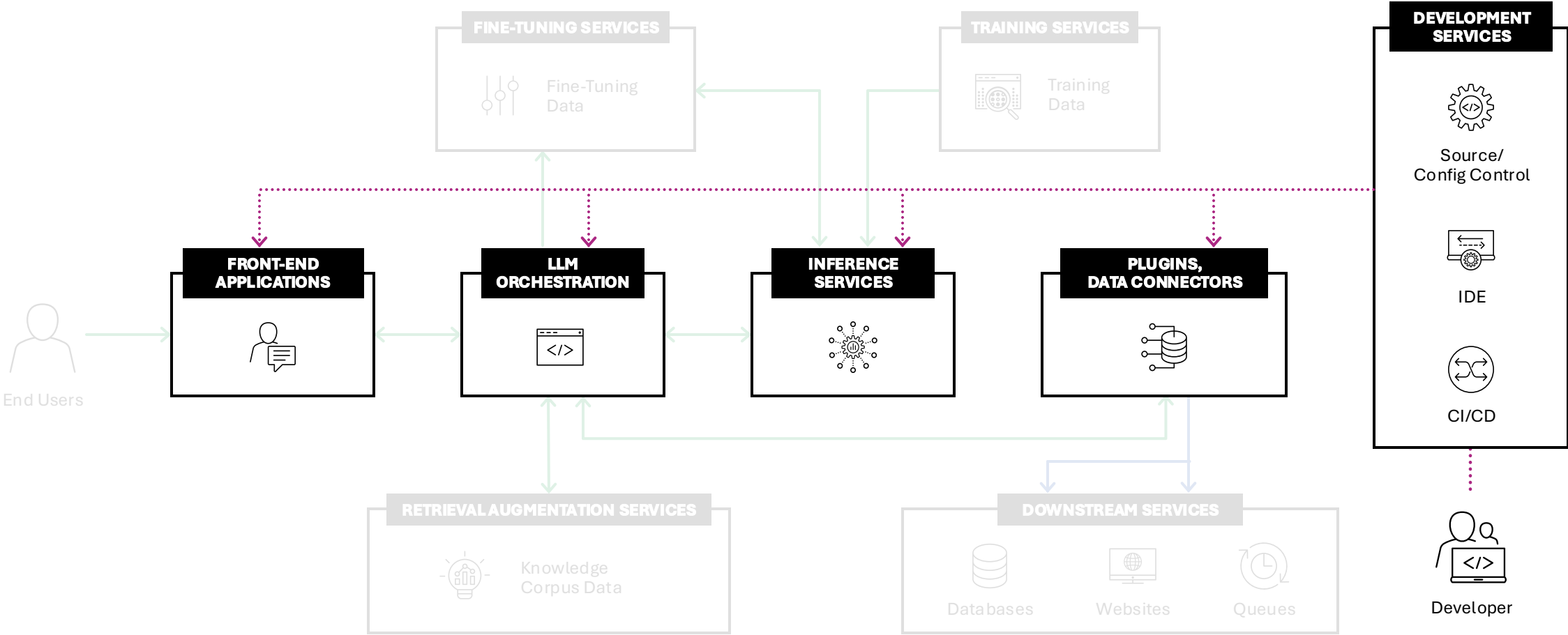
This is the process of teaching an AI model by exposing it to large amounts of data and allowing it to learn patterns and features. Training involves multiple iterations and optimizations to develop a model that can generalize well to new, unseen data.



Development

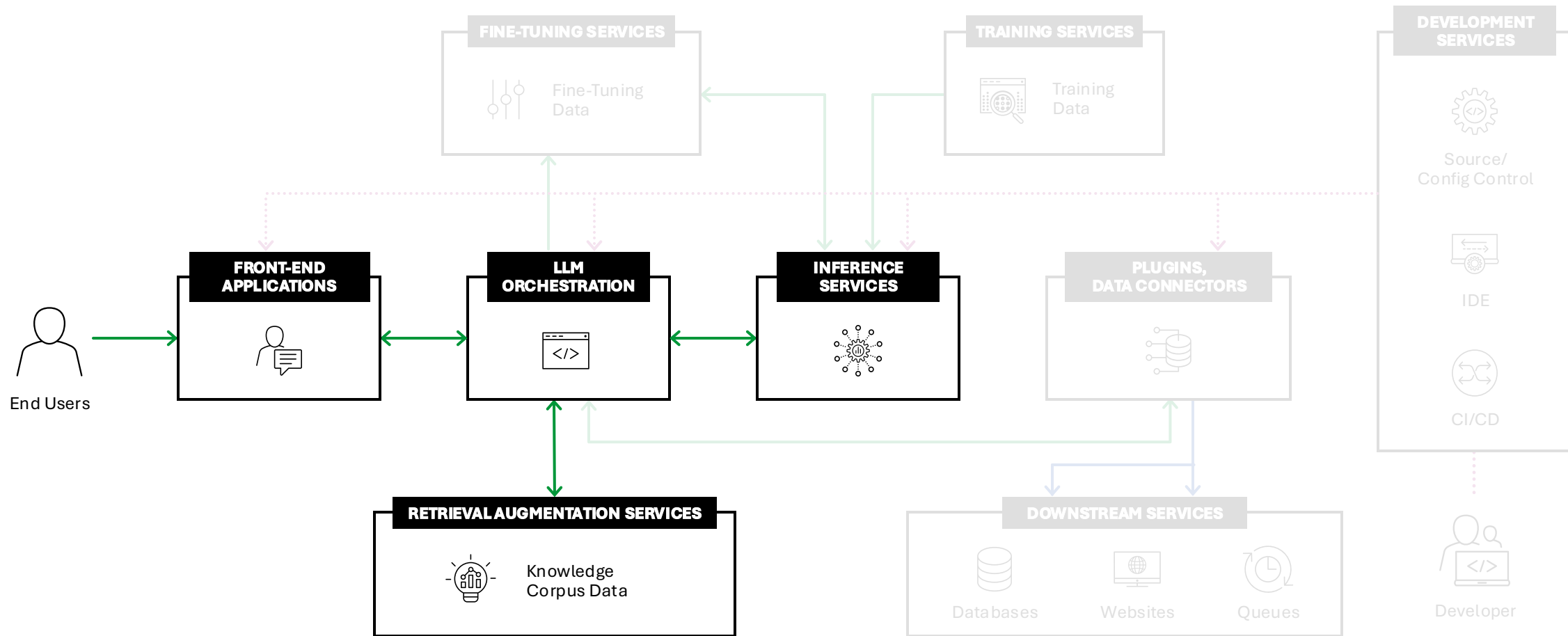


This encompasses the overall creation, testing, and deployment of AI solutions. It involves coding, integrating various AI components, and ensuring that the system is robust, scalable, and ready for production use.

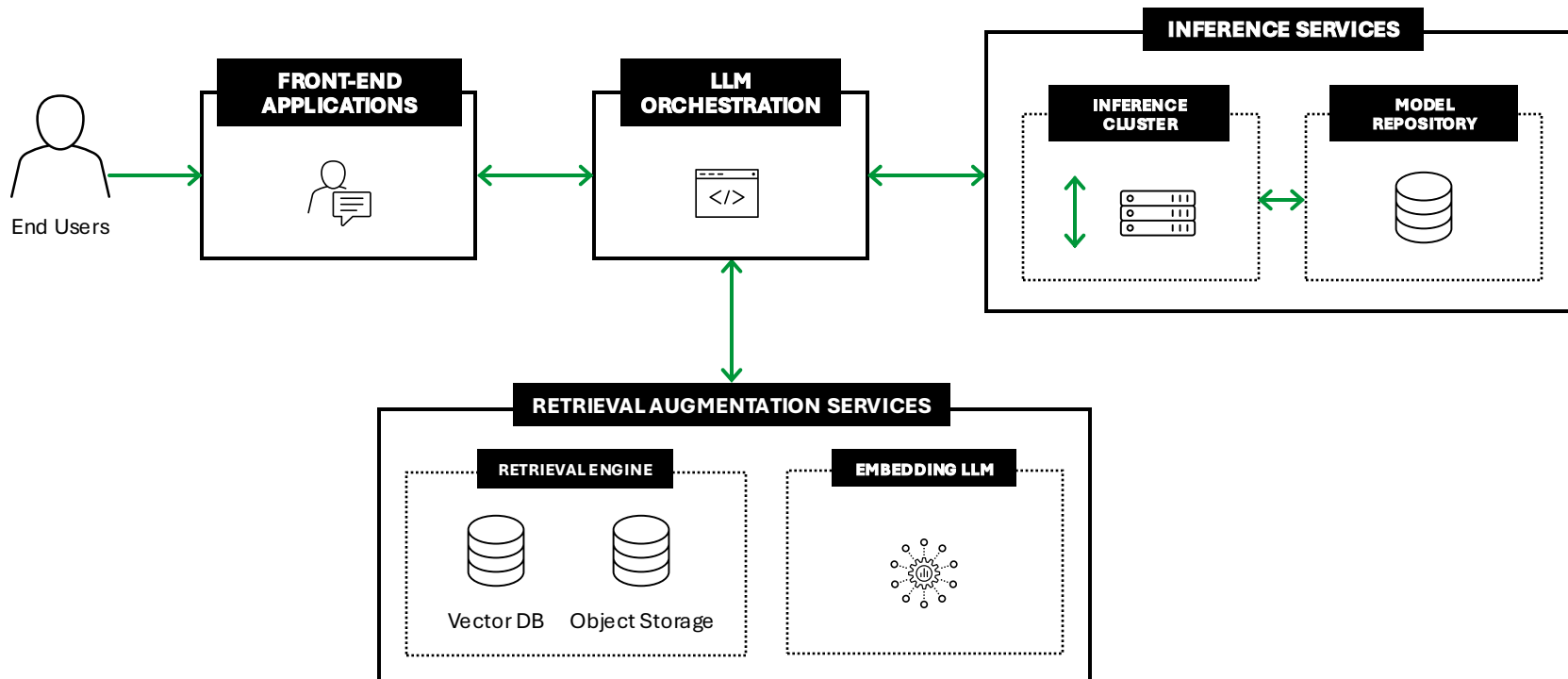


Inference with Retrieval Augmented-Generation (RAG)

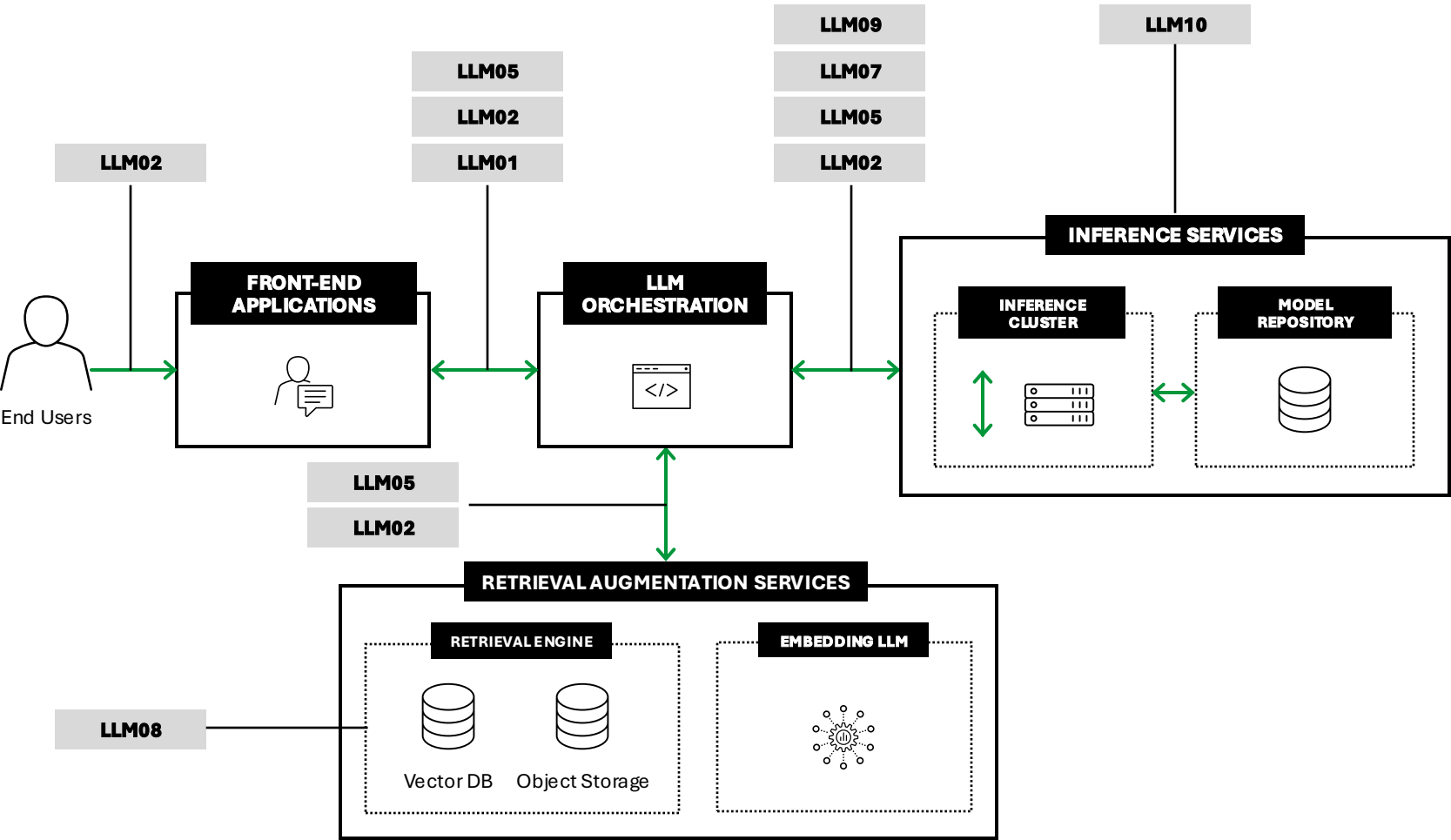
Featured AI Building Block



Detailed Component Architecture



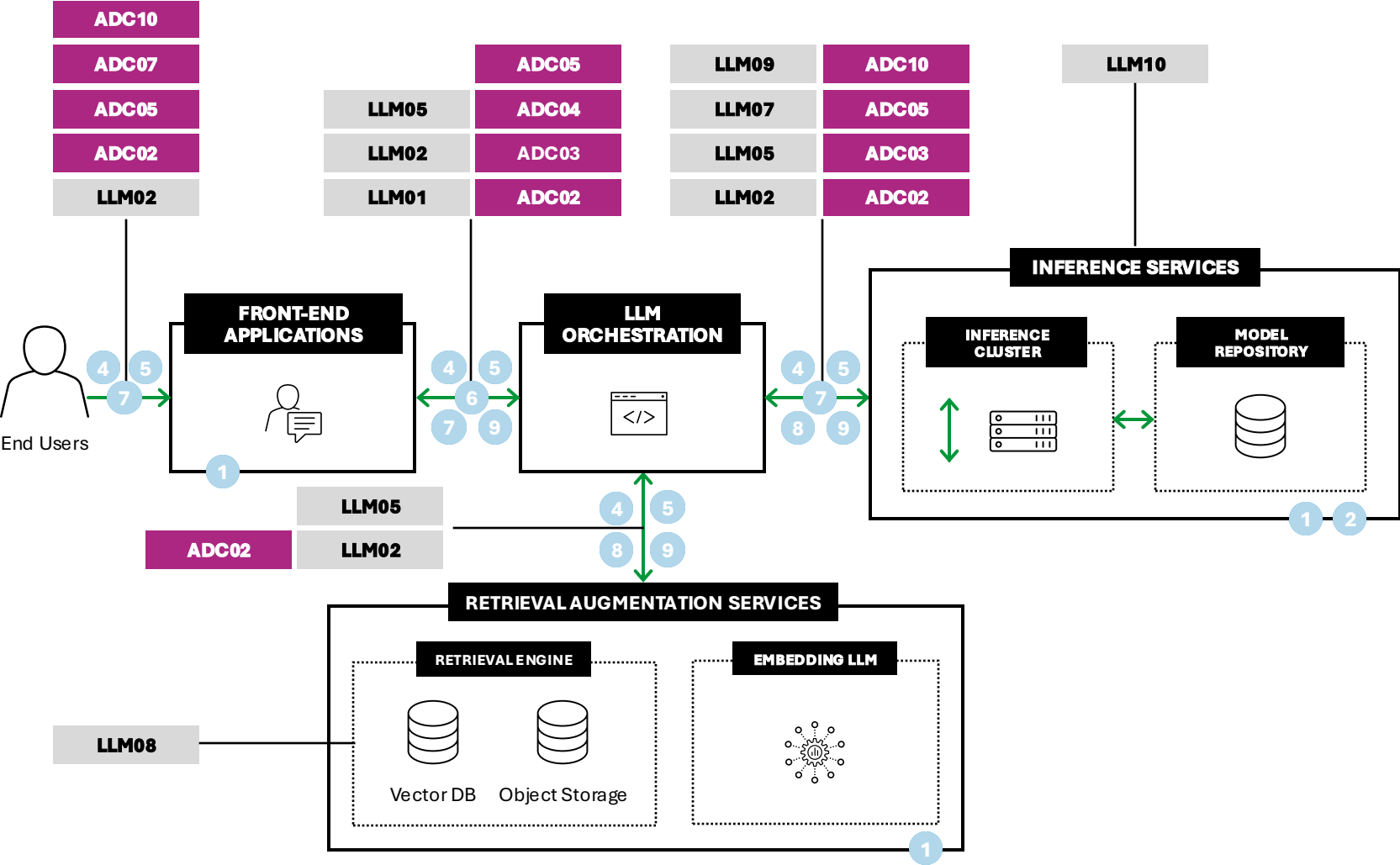
OWASP LLM Top Ten Insights



OWASP LLM Top Ten

LLM01	Prompt Injection
LLM02	Sensitive Information Disclosure
LLM03	Supply Chain
LLM04	Data and Model Poisoning
LLM05	Improper Output Handling
LLM06	Excessive Agency
LLM07	System Prompt Leakage
LLM08	Vector and Embedding Weakness
LLM09	Misinformation
LLM10	Unbounded consumption

F5 ADC Top Ten Insights



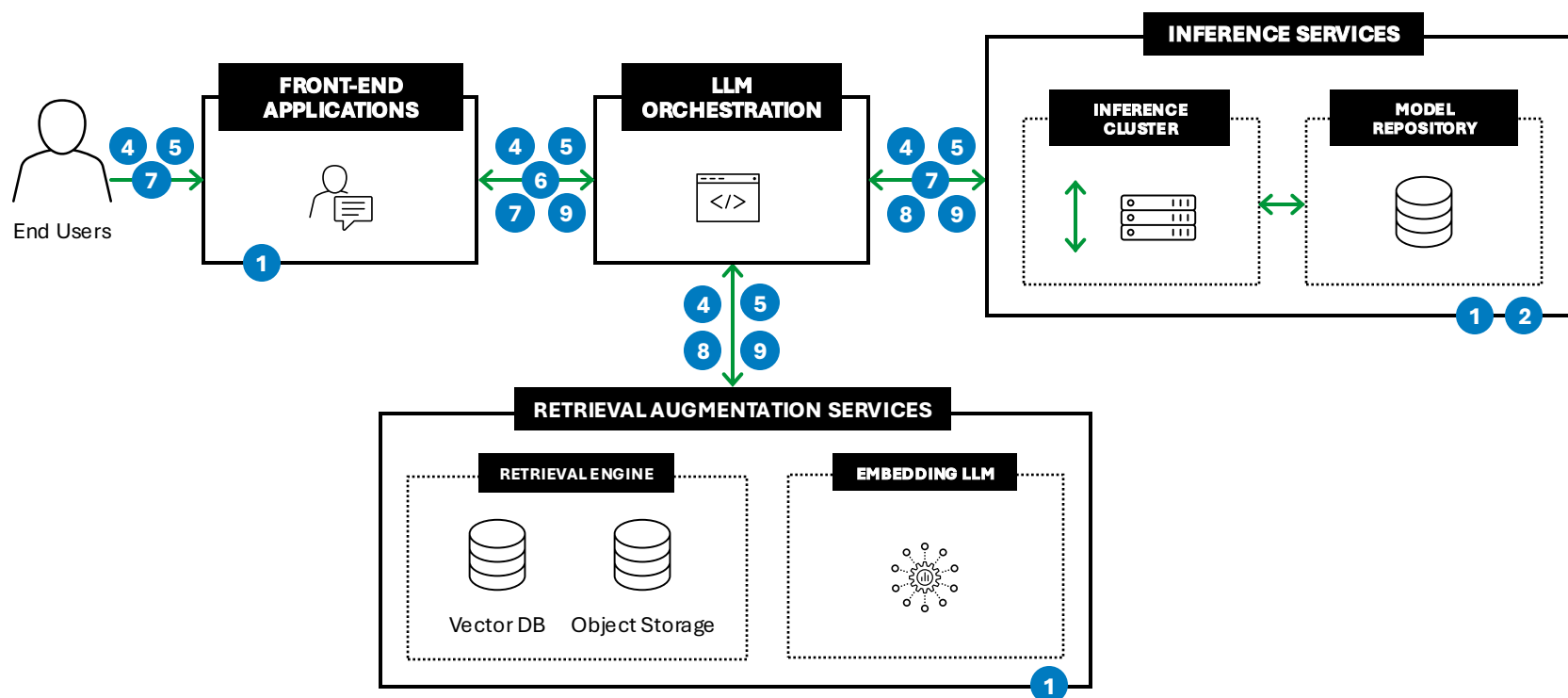
OWASP LLM Top Ten

LLM01	Prompt Injection
LLM02	Sensitive Information Disclosure
LLM03	Supply Chain
LLM04	Data and Model Poisoning
LLM05	Improper Output Handling
LLM06	Excessive Agency
LLM07	System Prompt Leakage
LLM08	Vector and Embedding Weakness
LLM09	Misinformation
LLM10	Unbounded consumption

F5 Application Delivery Top Ten

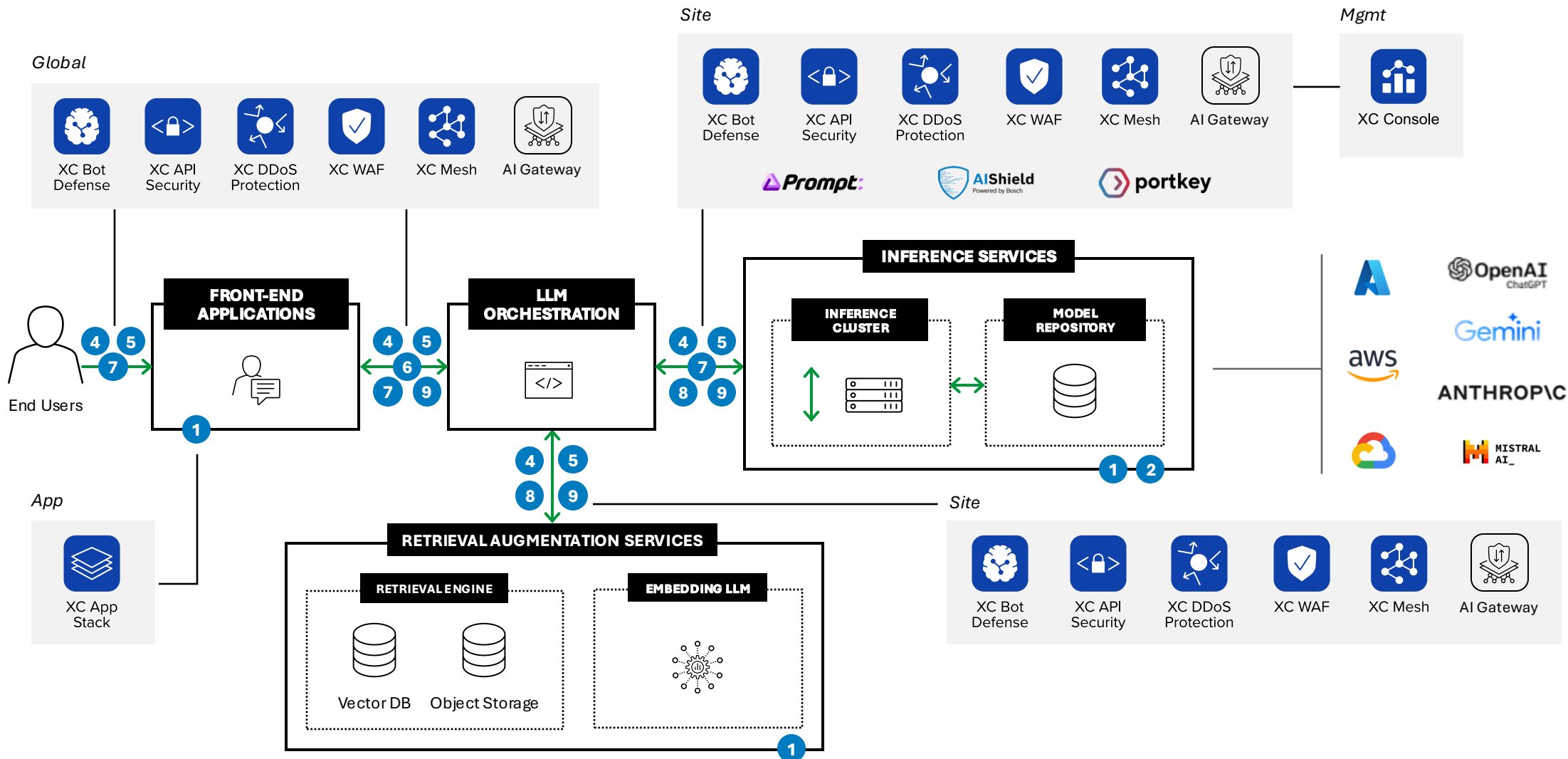
ADC01	Weak DNS Practices
ADC02	Lack of Fault Tolerance & Resilience
ADC03	Incomplete Observability
ADC04	Insufficient Traffic Controls
ADC05	Unoptimized Traffic Steering
ADC06	Inability to Handle Latency
ADC07	Incompatible Delivery Policies
ADC08	Lack of Security & Regulatory Compliance
ADC09	Bespoke Application Requirements
ADC10	Poor Resource Utilization

Design Requirements

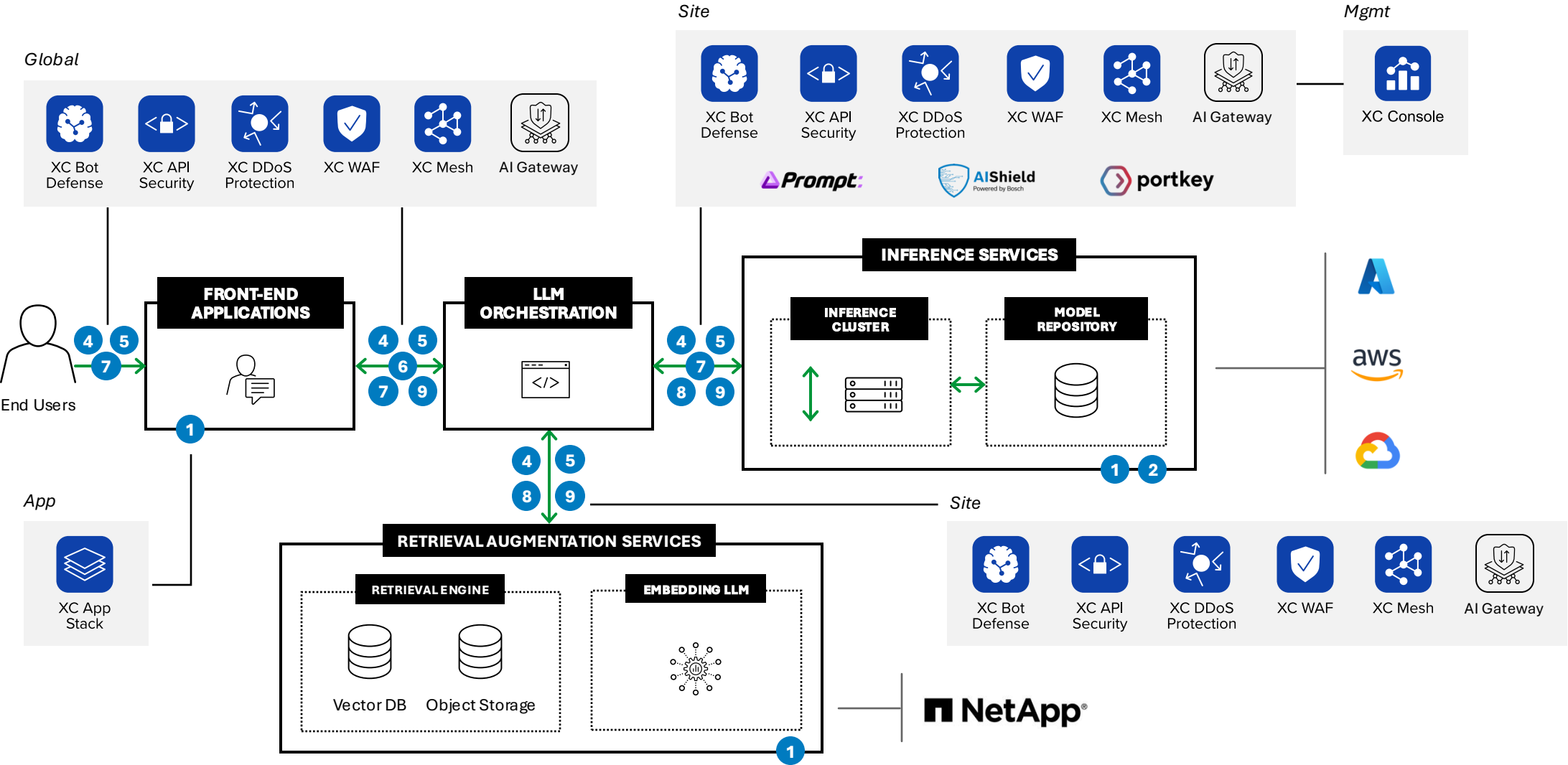


- 1 Distributed Compute Services
- 2 AI Compute Resources
- 3 Centralized Networking Management
- 4 Distributed App & API Security Services
- 5 Centralized Security Policy Management
- 6 AI/ML Data Loss Prevention
- 7 AI/ML Security
- 8 AI/ML Observability
- 9 Inter-Cluster Traffic Management

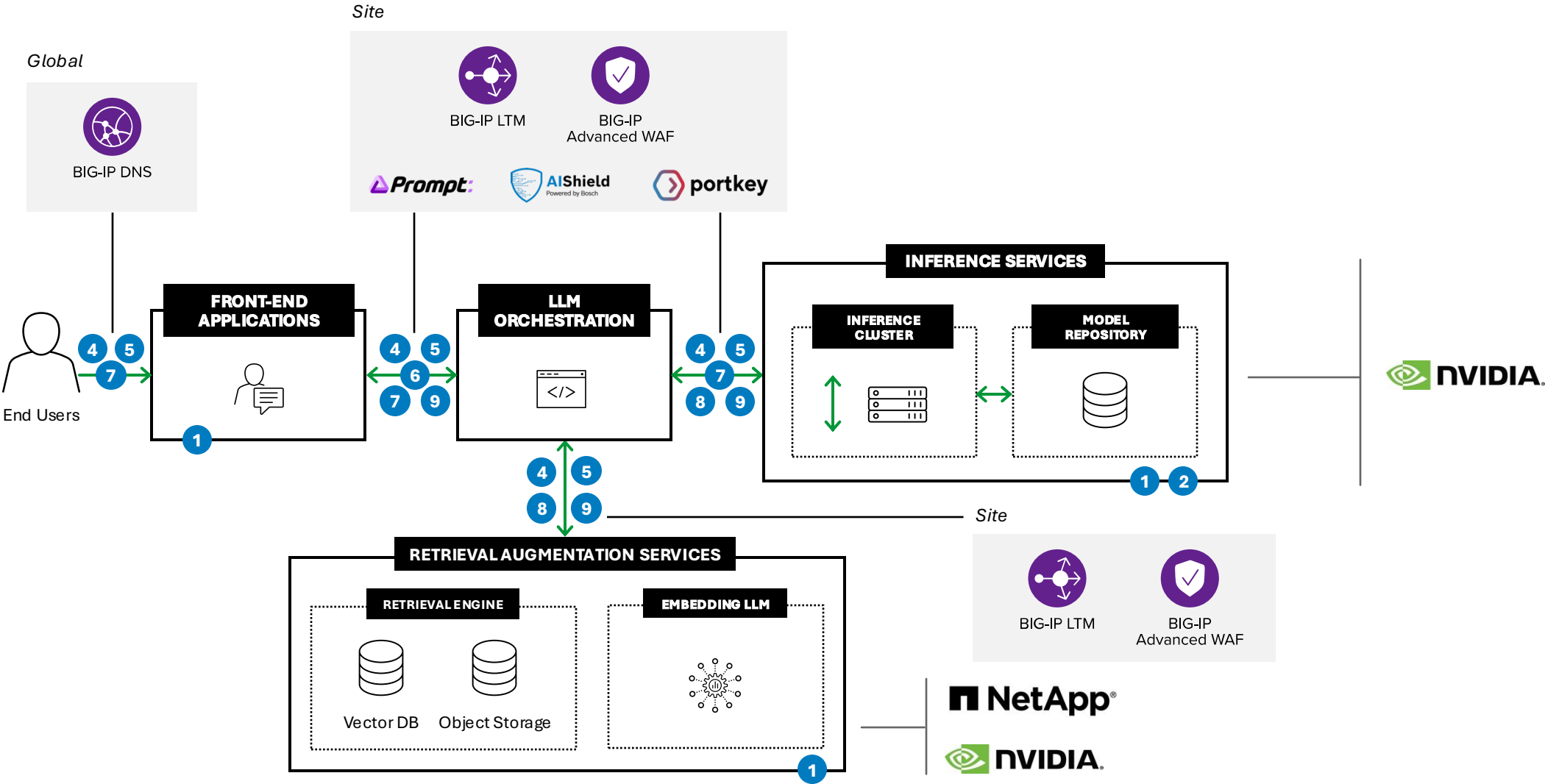
SaaS Deployment



Cloud-Hosted Deployment

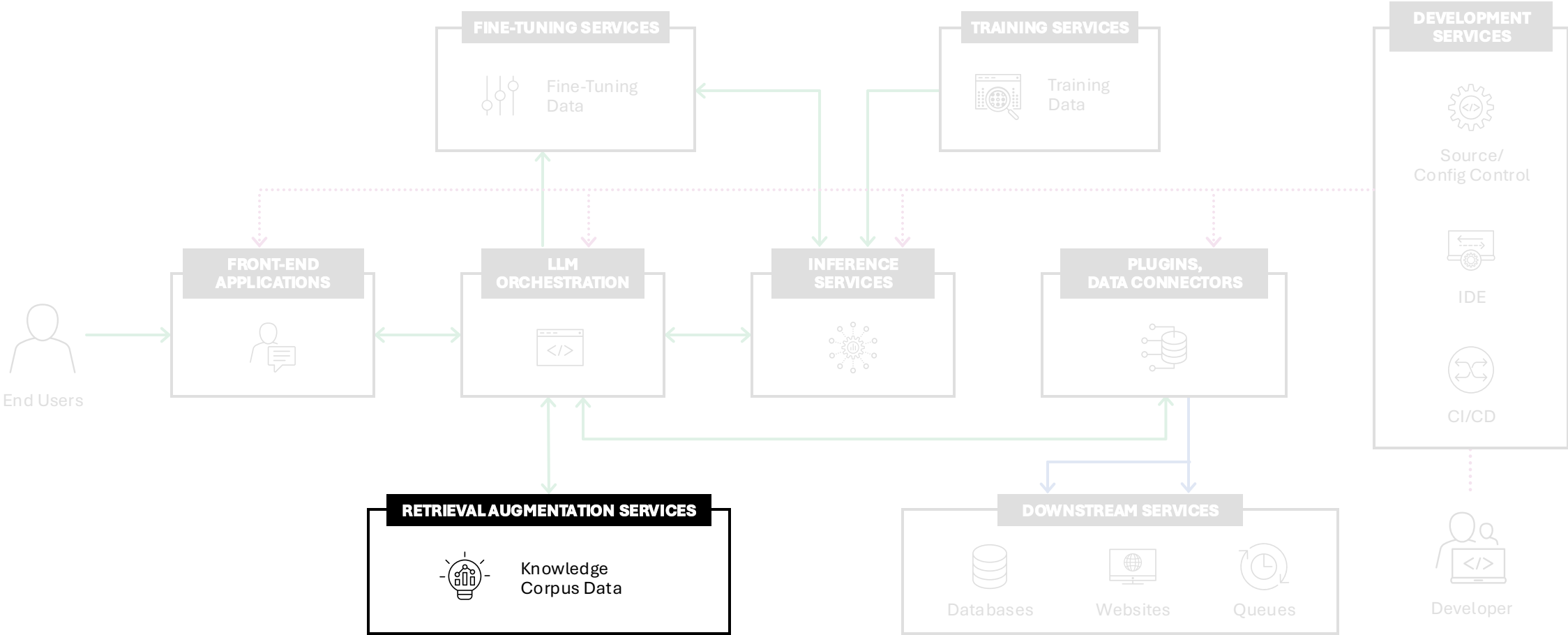


Self-hosted Deployment

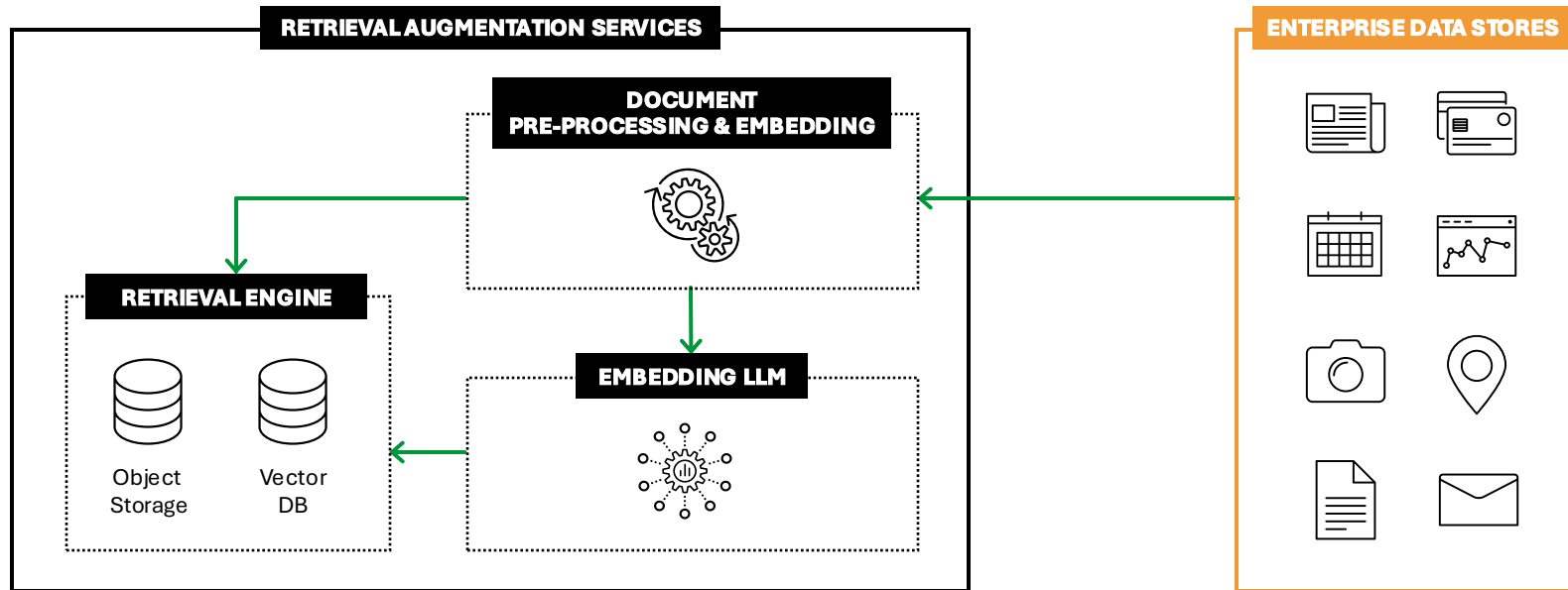


RAG Corpus Management

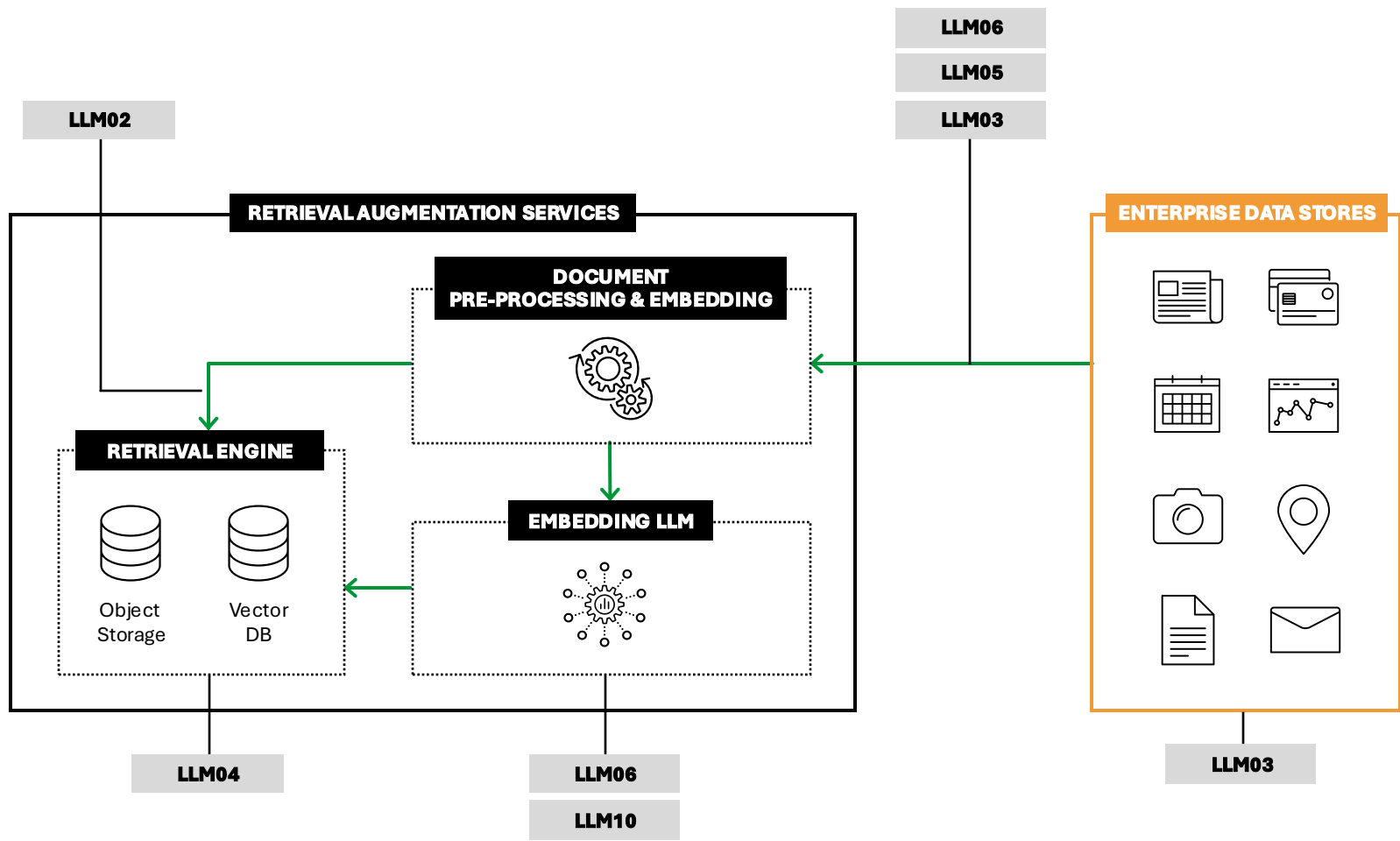
Featured AI Building Block



Detailed Component Architecture



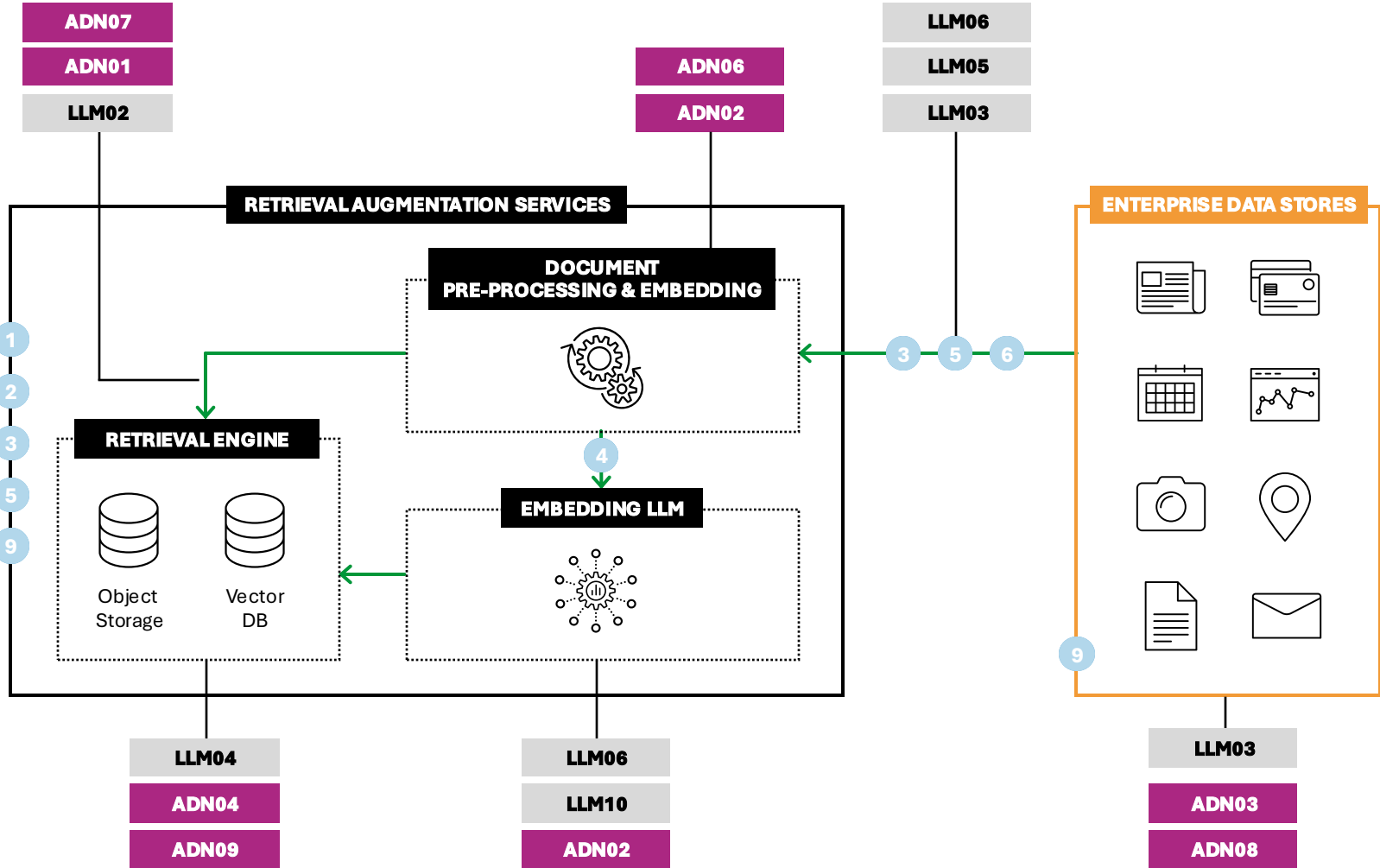
OWASP LLM Top Ten Insights



OWASP LLM Top Ten

LLM01	Prompt Injection
LLM02	Sensitive Information Disclosure
LLM03	Supply Chain
LLM04	Data and Model Poisoning
LLM05	Improper Output Handling
LLM06	Excessive Agency
LLM07	System Prompt Leakage
LLM08	Vector and Embedding Weakness
LLM09	Misinformation
LLM10	Unbounded consumption

F5 ADC Top Ten Insights



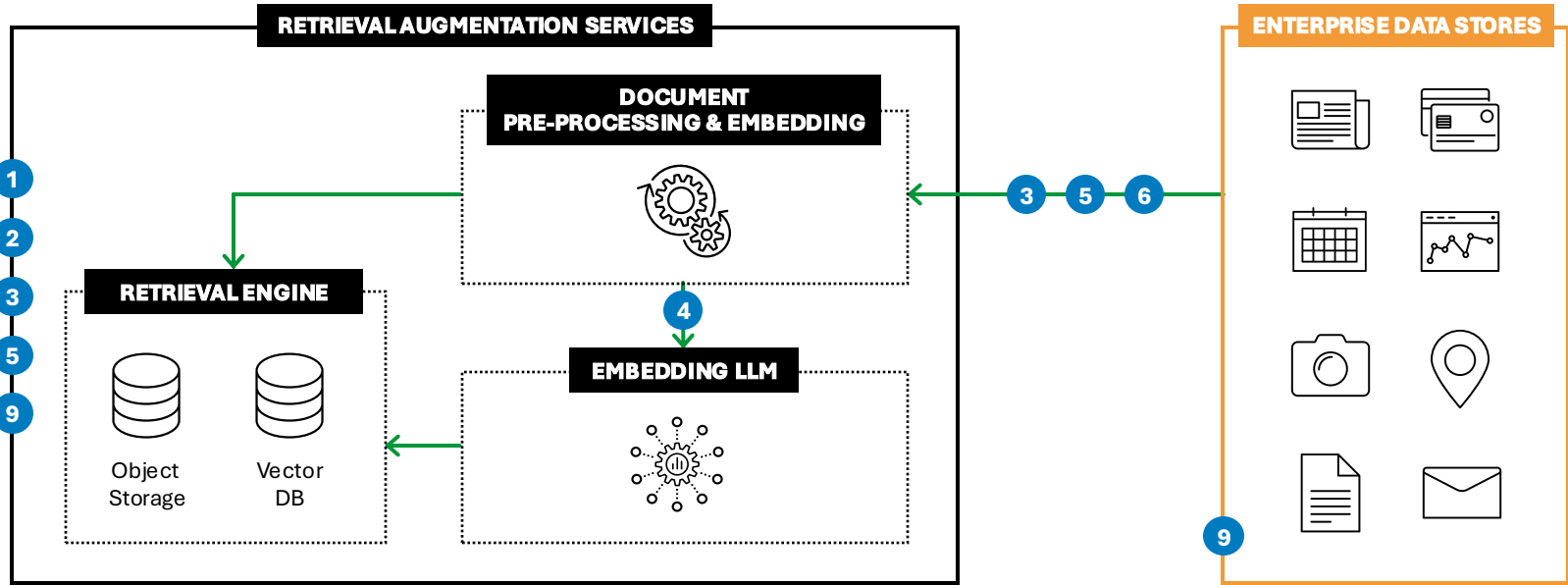
OWASP LLM Top Ten

LLM01	Prompt Injection
LLM02	Sensitive Information Disclosure
LLM03	Supply Chain
LLM04	Data and Model Poisoning
LLM05	Improper Output Handling
LLM06	Excessive Agency
LLM07	System Prompt Leakage
LLM08	Vector and Embedding Weakness
LLM09	Misinformation
LLM10	Unbounded consumption

F5 Application Delivery Top Ten

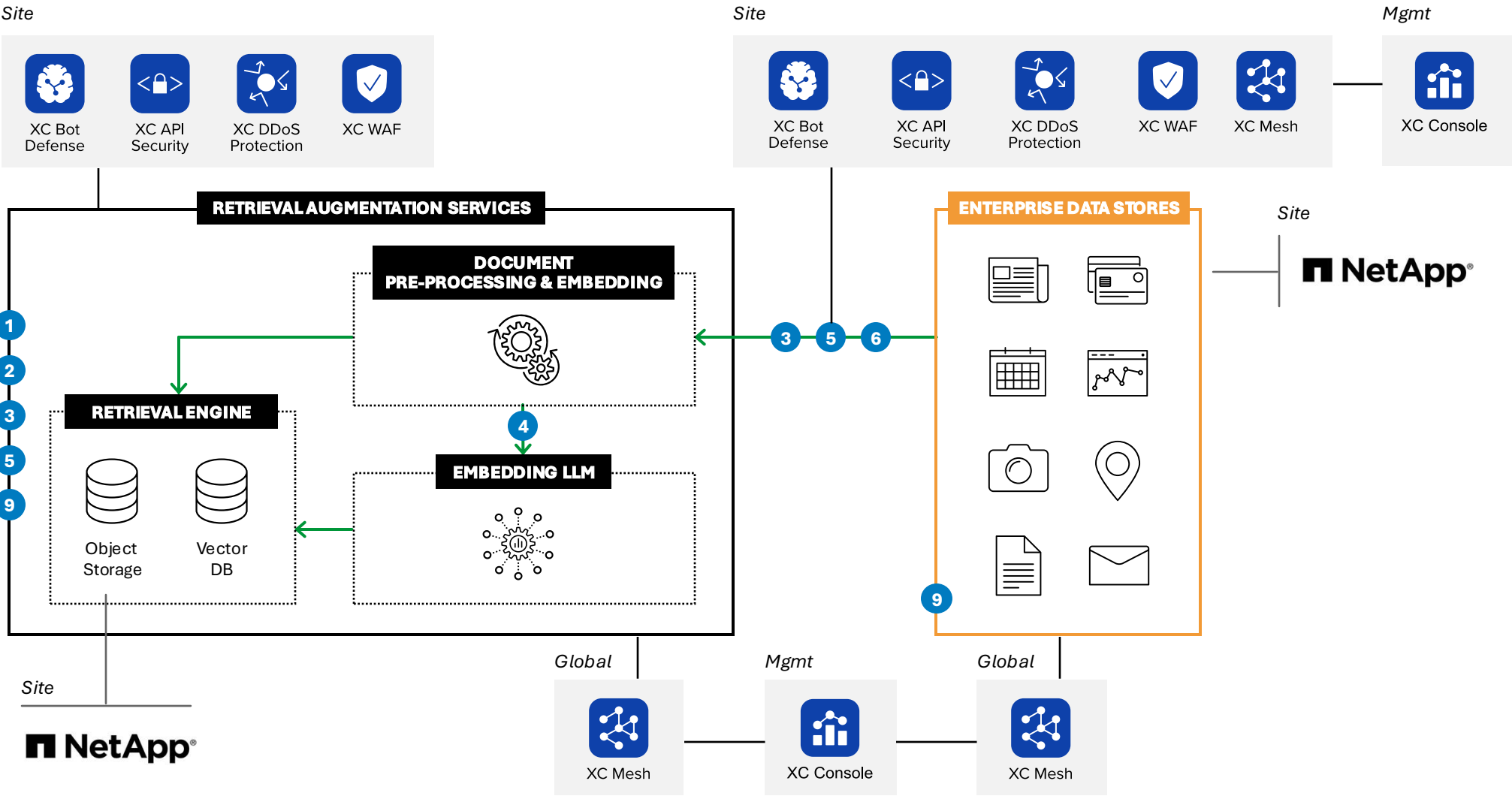
ADC01	Weak DNS Practices
ADC02	Lack of Fault Tolerance & Resilience
ADC03	Incomplete Observability
ADC04	Insufficient Traffic Controls
ADC05	Unoptimized Traffic Steering
ADC06	Inability to Handle Latency
ADC07	Incompatible Delivery Policies
ADC08	Lack of Security & Regulatory Compliance
ADC09	Bespoke Application Requirements
ADC10	Poor Resource Utilization

Design Requirements



- 1 Distributed Compute Services
- 2 AI Compute Resources
- 3 Centralized Networking Management
- 4 Distributed App & API Security Services
- 5 Centralized Security Policy Management
- 6 AI/ML Data Loss Prevention
- 7 AI/ML Security
- 8 AI/ML Observability
- 9 Inter-Cluster Traffic Management

Cloud Deployment



Self-Hosted Deployment

