Agentic AI for Information Retrieval

The 'Agentic AI for Information Retrieval' solution is a cutting-edge agent-based AI platform and conversational interface. Users interact with a system that retrieves relevant information in real time, providing contextually-accurate and insightful responses, which is ideal for knowledge-driven environments.

This solution is particularly well-suited for organizations looking to streamline information access across large datasets; this can include RESTful APIs, databases and any size document repositories. The conversational AI enhances the user's experiences and accelerates their decision-making. The system is deployed on AWS and integrates key AWS services to deliver privacy, scalability, and security.

Key Features:

Agent-based data collectors: Agents for knowledge retrieval leverage intelligent entities to dynamically access, filter, and synthesize information from diverse data sources. These agents autonomously navigate databases, APIs, and document stores, applying context-aware reasoning to extract relevant insights.

Real-time Knowledge Retrieval: Leveraging Agentic AI and RAG (Retrieval-Augmented Generation), the system uses real-time conversation to retrieve information from knowledge bases.

Conversational Interactions: Users engage with the system through natural language queries, receiving precise responses augmented by the knowledge embedded within corporate information sources.

Advanced Metadata Filtering: Data and documents are tagged and filtered based on metadata, allowing the system to retrieve the most relevant information.

History-aware Retriever: The system considers the conversational history and can incorporate long-term memory. Together, these provide more accurate, context-aware responses, improving the overall user experience.

Cost Control: Built-in token, user, and role-based limitations help manage API usage and prevent excessive costs, ensuring that resource consumption is kept under control.

Correctness Testing: The solution includes built-in tools for evaluating the accuracy of the Agentic/RAG system. This feature enables admins and SMEs to test the system's performance against predefined benchmarks or real-world use cases, ensuring accuracy and reliability.

Customizable Information Retrieval: The solution offers a high degree of flexibility, allowing users to configure document ingestion, relational and non-relational data, API parameters, and role-based access controls.

Customizable UI: The chat client or other preferred user interface can be tailored to reflect the organization's brand in diverse business environments.

User Authentication and Access Control: The solution integrates with AWS Cognito and corporate Single Sign-On (SSO) systems to provide secure and seamless user authentication.

Data Protection: Sensitive data, such as conversation histories and document embeddings, are securely stored in DynamoDB and PostgreSQL. Data used by LLMs remains completely private and protected. Encryption at rest and in transit is employed to protect all stored information.

Fully Automated: Infrastructure is provisioned using an IaC platform, enabling easy replication, deployment, and management of the system. This reduces the risk of manual configuration errors and ensures consistent deployments across multiple environments.

Target Audience

The solution is ideal for enterprises seeking quick access to internal knowledge repositories in order to streamline decision-making processes. Typical use cases include:

Customer Support

Internal knowledge base for support teams to retrieve information while handling client inquiries.

Sales Teams

Sales professionals can quickly access product information, pricing details, competitive analysis, and case studies by querying the system.

Legal Teams

Quick access to legal documentation, contracts, and compliance guidelines.

Corporate Teams

Employees seeking information from HR documents, company policies, benefits information, or other internal manuals.

Contact Us

For further details about the technical architecture, high configurability, security and speed, please refer to the FAQ and other documents for this offering.

VIVID CLOUD