SOLUTION BRIEF

RAG and Conversational AI for Internal Documents

The ‘RAG and Conversational AI for Internal Documents’ solution is a cutting-edge AI platform and conversational interface for internal users of any organization. Users interact with a system that retrieves relevant documents in real time, and provides contextually-accurate and insightful responses, which is ideal for knowledge-driven environments.

This solution is particularly well-suited for organizations looking to streamline document access across large datasets; this can include internal corporate manuals, policies, or any large document repository. The conversational AI enhances the user’s experiences and accelerates their decision-making. The system is fully deployed on AWS and integrates key AWS services to deliver scalability, security, and automation.

Key Features:

Real-time Document Retrieval: Leveraging RAG (Retrieval-Augmented Generation) technology, the system uses real-time conversation to retrieve documents based upon context and intent.

Integration with a SQL database: Extract information from databases using natural language.

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

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

VIVID CLOUD

History-aware Retriever: The system considers the conversational history to 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.

RAG Correctness Testing: The solution includes built-in tools for evaluating the accuracy of the Retrieval-Augmented Generation (RAG) system. This feature enables admins and SMEs to test the system's performance against predefined benchmarks or real-world use cases, ensuring that document retrievals and conversational responses are contextually accurate and reliable.

Highly Customizable: The solution offers a high degree of flexibility, allowing users to configure document ingestion, retrieval parameters, and role-based access controls to suit specific organizational needs. Custom workflows can be designed for different departments, ensuring that each user group accesses relevant information tailored to their roles. This adaptability ensures the solution fits seamlessly into 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. This ensures that only authorized users can access the system.

Data Protection: Sensitive data, such as conversation histories and document embeddings, are securely stored in DynamoDB and PostgreSQL, respectively. 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 small to medium enterprises that require quick access to internal knowledge repositories and need 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