Conversational RAG System

Conversational RAG System with Dynamic Indexing

Advanced Question-Answering AI with Real-time Knowledge Updates

Presented by: BALAJI K

Date: AUGUST 2025

Problem Statement & Solution Overview

The Challenge

- Information Overload: Organizations struggle with scattered knowledge across multiple sources
- Static Q&A Systems: Traditional chatbots lack conversational context and real-time updates
- Manual Knowledge Management: Time-consuming process to keep documentation current

Our Solution: Conversational RAG System

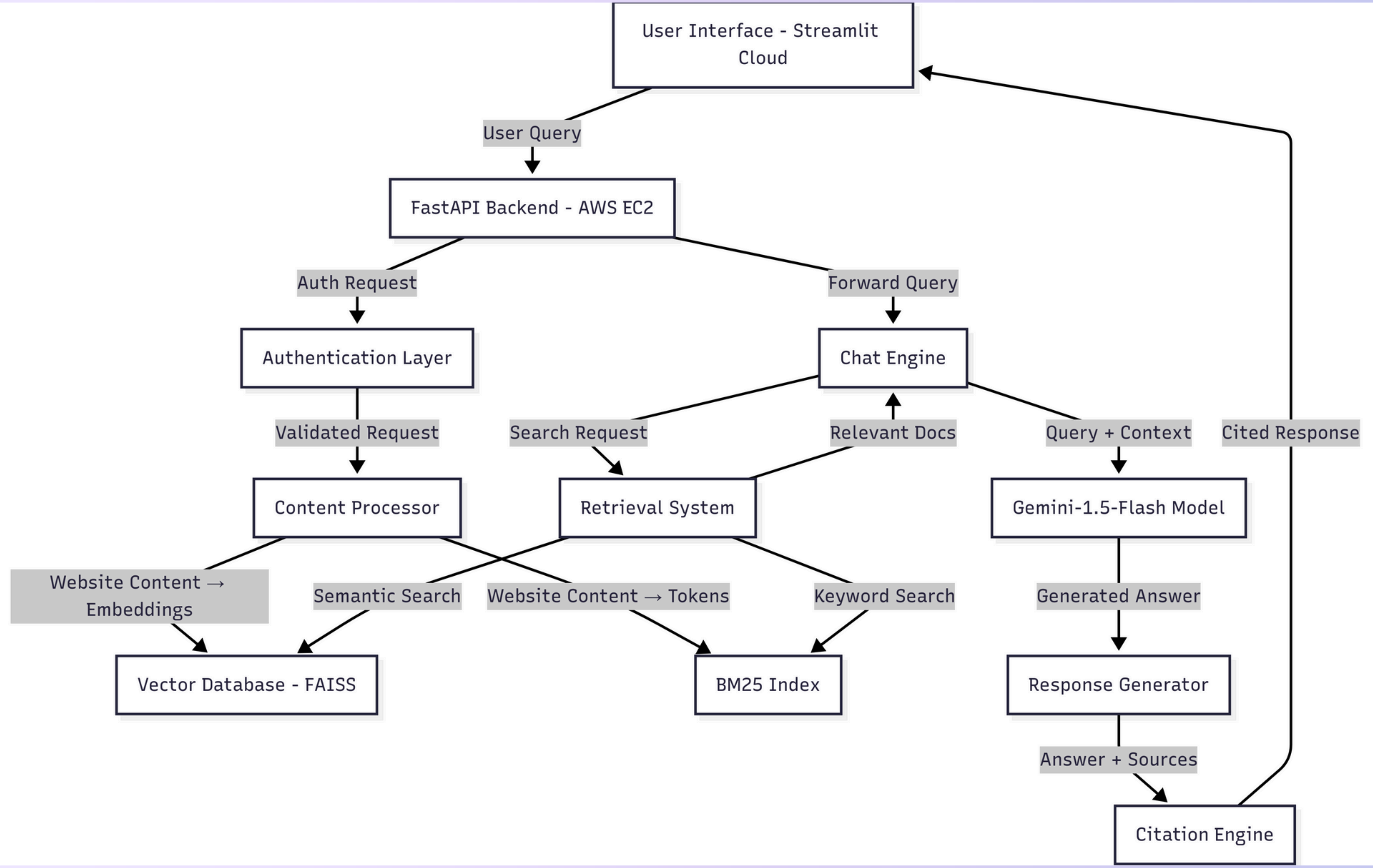
- Context-Aware AI: Maintains conversation flow like human dialogue
- Dynamic Knowledge: Real-time web content indexing and retrieval
- Multi-Source Intelligence: Unified access to diverse information sources
- Precise Citations: Transparent source attribution for every response

"Bridging the gap between human conversation and machine intelligence"

System Architecture & Tech Stack

Tech Stack

- Frontend: Streamlit Cloud (Interactive UI)
- Backend: FastAPI + AWS EC2 (t3.small)
- Search Engine: Hybrid (FAISS + BM25 + Cross-Encoder Reranker)
- LLM: Google Gemini-1.5-Flash
- Storage: Local FAISS indices (Static + Dynamic)



Key Features & Capabilities

Core Capabilities

- Conversational Context: Remembers previous exchanges, handles pronouns and references
- Hybrid Search: Combines semantic (FAISS) + keyword (BM25) + reranking for optimal accuracy
- Dynamic Indexing: Real-time web content ingestion and processing
- Health Monitoring: System status tracking and performance metrics

Advanced Features

- Multi-Source Management: Static knowledge base + live web content
- Citation Engine: Transparent source attribution for every response
- Batch URL Processing: Group indexing of multiple websites
- Intelligent Boundaries: Gracefully handles out-of-scope queries

Smart Conversation Flow

- Context Retention: "What is attention mechanism?" → "How does it help with long sequences?"
- Source Integration: Seamlessly blends information from multiple indexed sources

Live Demo Walkthrough

Demo Scenario: AI Knowledge Assistant

Demo Flow:

- System Dashboard: Health check, active sources, conversation metrics
- Conversational Interaction:
- Ask: "What is attention mechanism in transformers?"
- Follow-up: "How does it help with long sequences?" (context retention)
- Live URL Indexing: Add new AI research article in real-time
- Source Management: View indexed content and citations
- Context Demonstration: Multi-turn conversation with pronoun resolution

What You'll See:

- Real-time responses with source citations
- Seamless context handling across conversation turns
- Live content indexing from web URLs
- Transparent source attribution for every answer

Technical Challenges & Solutions

Major Challenges Overcome

Challenge	Solution Implemented
Memory Management	Efficient conversation pruning (10 pairs max) + optimized FAISS indices
Search Accuracy	Hybrid retrieval: BM25 + FAISS + Cross-encoder reranking
Context Handling	Conversation manager with pronoun resolution and topic continuity
AWS Deployment	Streamlined deployment with environment variable management
Real-time Indexing	Asynchronous web scraping with error handling and retry logic

Technical Innovations

- Conversation Memory Architecture: Context-aware prompt engineering
- Hybrid Retrieval Pipeline: Multi-stage relevance scoring
- Dynamic Index Management: Live content updates without system restart
- Graceful Error Handling: Robust failure recovery and user feedback

Current Results & Performance

Performance Benchmarks

- Response Time: ~2-3s average (including retrieval + generation)
- Memory Footprint: Efficient within 2GB RAM (t3.small instance)
- Sources Indexed: 7 total (4 static + 3 dynamic websites)
- Conversation Context: Supports 10 message pairs with full context retention
- Retrieval Quality: Hybrid search (BM25 + FAISS + reranker) for higher accuracy

Current Capabilities

- Document Processing: HTML articles, blog posts, technical documentation
- Conversation Sessions: Multi-turn dialogues with context preservation
- Real-time Updates: Dynamic content indexing without system downtime
- Source Attribution: 100% citation coverage for factual responses

System Reliability

- Uptime: Stable AWS deployment with health monitoring
- Error Handling: Graceful degradation for unsupported content types

Future Roadmap & Improvements

Immediate Enhancements (Next 3 months)

- Multi-format Support: PDFs, Word docs, plain text files
- Scalable Storage: Migration to AWS S3 for production-grade storage
- Enhanced Memory: Extended conversation context (50+ message pairs)

Advanced Features (6-12 months)

- Superior Search: Replace BM25 with advanced retrieval models
- Premium LLM Integration: GPT-4, Claude for higher-quality responses
- Domain Specialization: Finance, Healthcare, Legal-specific knowledge bases

Game-Changing Applications

- Education: AI tutor for complex academic subjects
- Enterprise: Company-wide knowledge assistant
- Research: Academic paper discovery and synthesis

Business Impact & Thank You

Real-World Applications

- Enterprise Knowledge Management: Centralized company documentation access
- Intelligent Customer Support: Context-aware help desk automation
- Educational Platforms: AI-powered learning assistants
- Research & Development: Multi-source information synthesis

Scalability Potential

- Horizontal Scaling: Multi-instance deployment for high-traffic scenarios
- Content Versatility: Adaptable to any domain or industry vertical
- Integration Ready: API-first design for seamless system integration

Technical Innovation

- Beyond Basic RAG: Conversational context + real-time updates
- Production-Ready: Full-stack deployment with monitoring and management
- Open Architecture: Extensible design for future enhancements

Thank You.

Live Demo Ready | GitHub Repository Available | AWS Deployment Active