TokenX Team

RAG Chatbot

A professional solution for intelligent Q&A over documents, powered by state-of-the-art retrieval-augmented generation.

V Varun S Sreelakshmi N Nandana J Jagan

LangChain Groq Gradio PDF & Text

Problem & Motivation

Why We Need Intelligent Document Q&A

Organizations often have vast amounts of knowledge trapped in static documents. Manually searching for answers is inefficient and unscalable.

Key Challenges:

- Time-consuming document searches
- Difficulty finding specific information
- Inconsistent answers across teams
- Scaling issues as document volume grows

How can we leverage modern AI to enable instant, reliable answers from enterprise text and PDF files?

Our Approach

Solution Overview

TokenX RAG Chatbot uses Retrieval-Augmented Generation (RAG) to synthesize answers directly from user-provided documents, utilizing advanced embeddings and scalable vector search.

How RAG Works:

- Documents are chunked and embedded into a vector space
- User queries retrieve the most relevant chunks
- LLM generates contextual answers using these chunks
- Responses are grounded in document content

Key Benefits:

Seamless Uploads

Support for PDF and TXT files with automatic processing

Contextual Q&A

Natural language interaction with document-aware responses

Accurate Results

Information drawn directly from source materials

Data Security

Documents remain within your control at all times

TokenX RAG Chatbot eliminates the need for manual document searches while providing reliable, source-grounded answers.

System Design

Technical Architecture & Components

Key Components:

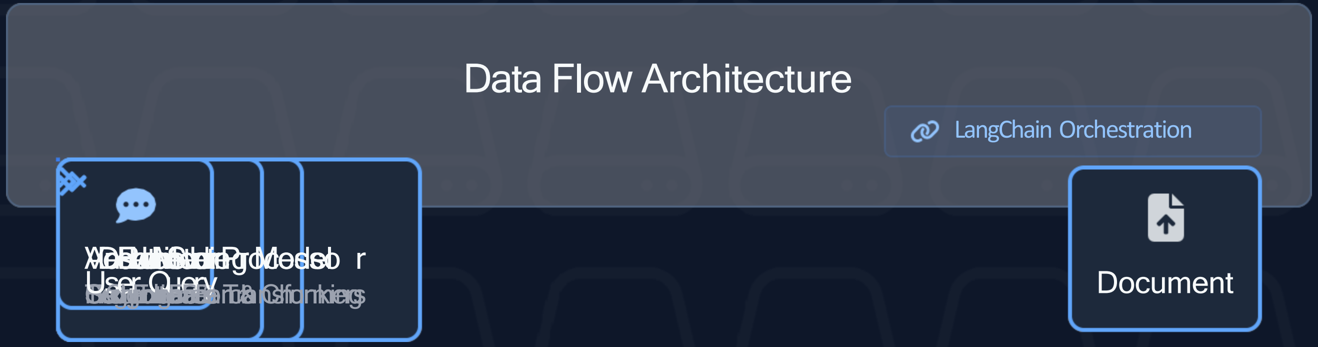
Document Ingestion
PDF & TXT extraction via PyMuPDF

Text Chunking
RecursiveCharacterTextSplitter for optimal segmentation

Embedding Generation
HuggingFace all-MiniLM-L6-v2 transformer model

Vector Store
ChromaDB for efficient similarity search

LLM Generation
Groq's Llama 3 (8B parameters) for answer synthesis



Data Flow Architecture

LangChain Orchestration

User Query

Document

LangChain orchestrates the entire retrieval-augmented

System Capabilities

Key Features & Advantages

Multi-Format Support

Seamlessly process PDF & Text files with intelligent extraction

Smart Chunking

Intelligent text segmentation for optimal context preservation

Vector Storage

Efficient semantic embedding storage with ChromaDB integration

Semantic Retrieval

Context-aware document search beyond keyword matching

Natural Language Q&A

Human-like responses grounded in document context

High Performance

Fast, responsive answers even with large document sets

Cloud Ready

Scalable architecture designed for cloud deployment

Customizable

Flexible configuration for different use cases & domains

Secure Workflow

Robust data handling with privacy-preserving features

TokenX's RAG architecture combines the precision of retrieval systems with the fluency of generative AI for superior document intelligence.

Technical Components

Technology Stack

Our RAG Chatbot leverages a modern stack of powerful AI and data processing technologies, all integrated in Python for rapid development and deployment.

LangChain

Orchestration framework that connects LLMs with external data sources and processing pipelines.

HuggingFace

Provides embedding models (all-MiniLM-L6-v2) for converting text into semantic vector representations.

ChromaDB

Open-source vector database for storing and retrieving document embeddings with semantic search capabilities.

Groq Llama 3

Large Language Model API service that generates coherent, contextually relevant responses based on retrieved content.

Gradio

Python library for quickly creating customizable web interfaces for machine learning models and workflows.

PyMuPDF

PDF parsing library that efficiently extracts and processes text content from document files.

All components are open-source or API-accessible, allowing for flexible deployment and customization.

User Experience

Demo & User Interface

Interactive Workflow

1 Upload Document

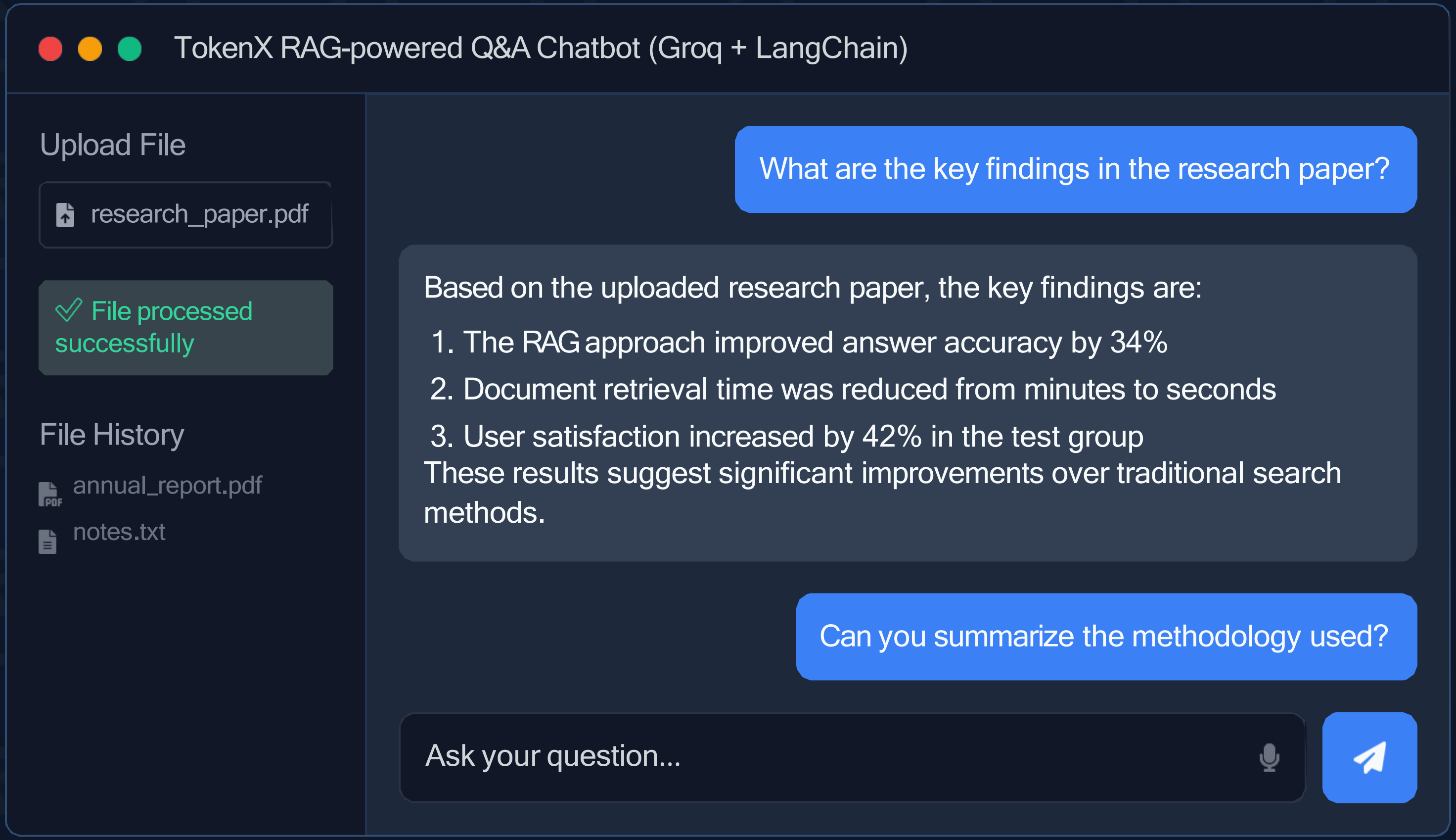
Upload your PDF or TXT file to the system

2 Ask Question

Type your question in natural language

3 Get Answer

Receive context-aware responses based on your document



TokenX RAG-powered Q&A Chatbot (Groq + LangChain)

Upload File

research_paper.pdf

✓ File processed successfully

File History

annual_report.pdf

notes.txt

What are the key findings in the research paper?

Based on the uploaded research paper, the key findings are:

1. The RAG approach improved answer accuracy by 34%
2. Document retrieval time was reduced from minutes to seconds
3. User satisfaction increased by 42% in the test group

These results suggest significant improvements over traditional search methods.

Can you summarize the methodology used?

The Gradio interface provides an intuitive experience with real-time feedback and document-grounded answers.

Looking Forward

Future Enhancements & Conclusion

Planned Improvements

- ### Multi-file & Multi-format Ingestion
  Support for multiple file uploads and additional formats (DOCX, CSV, HTML)
- ### Advanced Analytics & Source Citation
  Detailed analysis of document content with precise source references
- ### Fine-grained Document Access Control
  User-level permissions and role-based document access
- ### Enhanced Answer Explanations
  Deeper contextual reasoning and confidence scoring

Impact & Benefits

- 90% reduction in document search time
- Improved answer accuracy and relevance
- Enhanced knowledge sharing across teams
- Secure and compliant information access

Conclusion

TokenX's RAG Chatbot revolutionizes document-based Q&A with robust, production-ready AI solutions. Our implementation bridges the gap between static documents and dynamic knowledge retrieval, setting new standards for enterprise information access.

✉ contact@tokenx.ai