Retrieval-Augmented Generation System for Document Intelligence

An Explainable AI Framework for Knowledge-Intensive Domains

Arnav Bhardwaj

Problem & Motivation

Organizations across industries increasingly depend on large, complex document repositories spanning financial reports, legal contracts, healthcare records, and regulatory filings. Traditional large language models (LLMs) frequently generate plausible-sounding but factually incorrect responses, a phenomenon known as hallucination. These models also struggle to provide transparent sourcing for their answers.

In regulated, high-stakes domains where accuracy and accountability are paramount, this lack of explainability poses significant risks. Decision-makers need AI systems that can ground their responses in verifiable sources, provide clear citations, and confidently acknowledge when information is unavailable.

Factual Grounding

Answers anchored in actual documents

Clear Citations

Traceable source attribution

Honest Uncertainty

System acknowledges knowledge gaps

Objectives & Contributions

Objectives

- Design an end-to-end RAG pipeline capable of processing multi-document corpora efficiently
- Ensure explainability through page-level citations and complete answer provenance
- Implement intelligent guardrails that return *"I don't know"* for queries without supporting evidence
- Evaluate system performance across accuracy, relevance, faithfulness to sources, and response latency

Key Contributions

- A scalable, production-ready RAG framework for document intelligence applications
- An interactive Streamlit prototype enabling real-time question-answering with transparent citations
- A practical methodology for reducing LLM hallucinations through retrieval-grounded generation
- Comprehensive evaluation framework for assessing explainability and factual accuracy

Methodology & System Architecture

Our RAG system implements a six-stage pipeline that transforms raw documents into a trustworthy, explainable question answering system. Each stage is designed to maximize accuracy while maintaining transparency throughout the generation process.

Document Ingestion

PDF extraction, text cleaning, and quality preprocessing

Chunking & Tokenization

Intelligent segmentation into semantically coherent units

Embedding & Indexing

Vector embeddings stored in ChromaDB for efficient retrieval

Semantic Retrieval

Top-k most relevant chunks identified per query

LLM Answer Generation

GPT-based synthesis using only retrieved context

Explainability & Guardrails

Citations attached; insufficient context triggers "I don't know"

Evaluation & Results

Experimental Setup

Direct comparison between standalone LLM and RAG-enhanced responses using manually designed questions over authentic document sets

Performance Metrics

Evaluation across factual accuracy, hallucination rate, citation completeness, and end-to-end latency

Key Findings

Superior Factual Accuracy

RAG-enhanced system demonstrates significantly higher accuracy compared to LLM-only baseline

Reduced Hallucination

Grounding in retrieved context dramatically lowers false information generation

Complete Traceability

100% of generated answers include precise document and page-level citations

Honest Uncertainty

System correctly identifies and acknowledges unsupported queries with "I don't know" responses

Latency Performance

5-6s

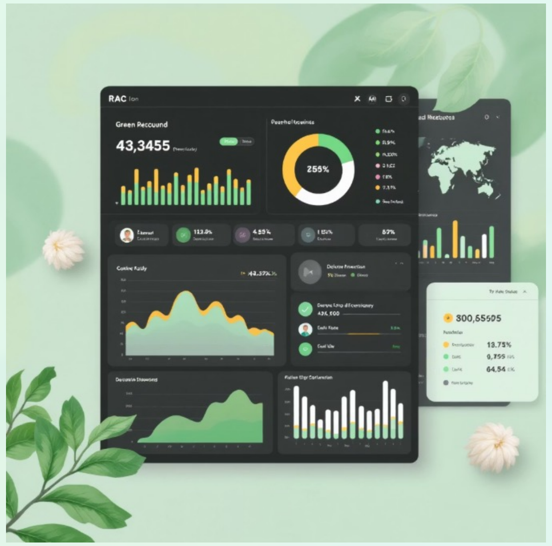
Retrieval Time

Semantic search and chunk extraction

8-10s

End-to-End

Complete query to answer generation



Conclusion & Future Work

Conclusion

This research demonstrates that Retrieval-Augmented Generation fundamentally transforms LLM reliability and transparency for knowledge-intensive applications. By grounding responses in verifiable document sources and implementing intelligent guardrails, our RAG system addresses critical limitations of traditional language models.

The integration of explicit citation mechanisms and uncertainty acknowledgment significantly enhances user trust, making this approach particularly valuable for regulated domains including healthcare, legal, and financial services where accuracy and accountability are non-negotiable.

Future Directions

1

Voice Integration

AI assistant with conversational interface (Real time Voice Assistant)

2

Multilingual & Multimodal

Extend to non-English documents and visual content processing

asb INTERACTION

Introducing Voice-Enabled RAG Assistants

Document Intelligence → Now with Real-Time Voice

Built for businesses that need accuracy & automation

Customers want:

⚡ Immediate answers

🎙 Verified information

Not guesses.

LinkedIn

The Solution: Voice + RAG

- ✗ Listens
- ✓ Retrieves answers from your documents
- ✓ Speaks back instantly
- ✓ Provides citations
- ✓ Says "I don't know" when unsure
- ✓ Works 24/7

Your website become a living, intelligent assistant.

WHAT IT CAN DO

- Explain your products/services
- Answer questions based on PDFs
- Handle finance/ legal/techinical content
- Speak in natural voice
- Learn from user questions

Built on:
Python -ChromaDB
GPT-40 mini
Sentence:Transformers
FastAPI - OpenAi
Realtime Voice

Coming Soon...

Want your own Voice-Enabled RAG Assistant?

Comment down "Interested"

Expected Launch: December 29th 2025

asb INTERACTION