Explainable and Verifiable Retrieval-Augmented Conversational Agent with Hallucination Control

Nancy Jenifer A[1], Varssha B[2], Vishnupriya B[3]
[1,3]*Assistant Professor/ Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Sriperumbudur, Chennai*
[2] *Student/ Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Sriperumbudur, Chennai*

Abstract **— This paper proposes a hallucination-free Retrieval-Augmented Generation (RAG) enabled chatbot designed to improve the accuracy and reliability of responses generated by large language models. Traditional language models often produce hallucinated or incorrect information due to their reliance on pre-trained knowledge. To address this issue, the proposed system integrates document retrieval with transformer-based language models using LangChain and embedding models from Hugging Face Transformers. The system retrieves relevant information from external documents using vector similarity search implemented through FAISS and generates responses using a local LLM executed via Ollama. Experimental evaluation using BLEU score and qualitative analysis demonstrates that the RAG-based chatbot significantly reduces hallucination and improves response accuracy compared to standard LLMs. The system is suitable for applications requiring reliable and domain-specific information retrieval.**

Index Terms **—RAG, Hallucination, Chatbot, LangChain, Hugging Face, FAISS, Embeddings, LLM**

I. INTRODUCTION

Large Language Models (LLMs) have revolutionized natural language processing by enabling advanced conversational systems. However, one major limitation of LLMs is hallucination, where the model generates incorrect or fabricated information.

In critical domains such as education, healthcare, and research, inaccurate responses can lead to serious consequences. Traditional chatbots rely solely on pre-trained knowledge and lack access to real-time or domain-specific data

To overcome these challenges, Retrieval-Augmented Generation (RAG) **c**ombines information retrieval with text generation. This approach enhances the reliability of responses by grounding them in external knowledge sources

This paper presents a hallucination-free chatbot system that integrates retrieval mechanisms with transformer-based models to ensure accurate and context-aware responses

II. LITERATURE REVIEW

Several studies have explored hallucination and its mitigation in AI systems:

- Research on GAN-based hallucination focuses on image generation but lacks applicability to text-based systems.
- Studies on LLM hallucination in skill learning highlight the negative impact of incorrect AI-generated instructions.
- Multi-perspective consistency checking methods detect hallucinations but increase computational cost.
- Multimodal verification systems improve reliability but require large datasets and complex architectures.

These studies indicate the need for a scalable, efficient, and accurate hallucination reduction method, which is addressed by the RAG framework

III. PROPOSED METHODOLOGY

A. Data Collection

Documents such as PDFs, text files, and web data are collected as the knowledge base.

B. Data Preprocessing

Before processing, the collected data undergoes several preprocessing steps to improve quality and consistency:

- Text Cleaning: Removal of unwanted characters, special symbols, and formatting issues
- Tokenization: Splitting text into smaller units such as words or sentences
- Stopword Removal: Eliminating common words (e.g., "the", "is") that do not contribute to meaning
- Normalization**:** Converting text to lowercase and standard formats
- Chunk Preparation: Structuring text for efficient segmentation

These steps ensure that the data is clean, structured, and suitable for embedding generation, improving retrieval accuracy

C. Text Chunking

Large documents are divided into smaller segments using text splitters available in LangChain.

- Each document is split into fixed-size chunks (e.g., 500–1000 tokens)
- Overlapping between chunks is maintained to preserve context
- Chunking improves:
  - retrieval efficiency
  - semantic relevance
  - computational performance

This step is critical because smaller chunks allow the system to retrieve more precise and relevant information**.**

D. Embedding Generation

Each text chunk is converted into a numerical vector representation using models from Hugging Face Transformers.

Common models include:

- BERT
- Sentence-BERT

These models use the self-attention mechanism to capture relationships between words and generate context-aware embeddings**.**

Key advantages:

- Captures semantic similarity
- Enables meaningful comparison between text and queries
- Improves retrieval accuracy

E. Vector Database Storage

The generated embeddings are stored in a vector database such as FAISS.

Features of FAISS:

- Fast similarity search
- Efficient handling of large datasets
- Scalable indexing

Each embedding is indexed so that similar vectors can be retrieved using distance metrics such as cosine similarity or Euclidean distance**.**

F. Query Processing

When a user submits a query:

1. The query is preprocessed (cleaning and tokenization)
2. It is converted into an embedding using the same embedding model
3. The system performs similarity search in the vector database
4. Top-k most relevant chunks are retrieved

This ensures that the system retrieves contextually relevant information instead of relying solely on the LLM's internal knowledge.

In this module, the developed Retrieval-Augmented Generation (RAG) system is integrated into an application environment for deployment and real-time interaction. The embedding models and vector database are initialized and stored to enable efficient retrieval operations, while the language model is executed locally using Ollama. Frameworks such as LangChain are used to manage the end-to-end pipeline, including document retrieval and response generation. The system can be deployed using a web framework such as streamlit to provide a user- friendly interface where users can input queries and receive accurate, context-aware responses

prediction. The deployed system allows users to input parameters and obtain predictive results efficiently, supporting practical implementation and accessibility of the predictive maintenance model

IV. PERFORMANCE EVALUATION

Metrics Used:

- BLEU Score
- Accuracy

- Response relevance

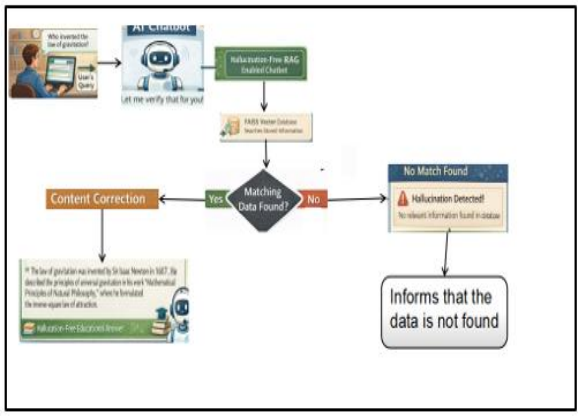
Results:

- RAG system shows higher BLEU scores compared to Standard LLM
- Reduced hallucination observed
- Improved domain-specific accuracy

The best performing model is selected for deployment.

TABLE I. TABLE TYPE STYLES

TableHead	TableColumnHead		
	Process	Tool/Model	Function
Data Collection	PDF Documents	Text Files	
Prepossessing	Cleaning & Tokenization	NLP Techniques	Prepares data
Embedding	Vector Creation	Transformer Models	
Storage	Indexing	Vector Database	Stores embeddings
Retrieval	Similarity Search	Search Algorithm	Finds relevant data



The figure illustrates the workflow of a hallucination-free Retrieval-Augmented Generation (RAG) chatbot system. Initially, the user submits a query, which is processed by the chatbot interface. The system then performs a similarity search in the vector database using FAISS to retrieve relevant information. If matching data is found, the retrieved content is passed to the response generation module, which produces an accurate and context-aware answer, thereby ensuring content correctness. If no relevant data is found, the system detects a potential hallucination and informs the user that the requested information is unavailable. This mechanism ensures that the chatbot avoids generating misleading or fabricated responses and maintains reliability in information delivery.

REFERENCES

[1] Evaluating the Effectiveness of Advanced Language Models in Detecting and Mitigating Hallucinations Using Structured Question-Answering, Novel Metrics, and Post-Hoc Retrieval — Rana Hassan Ajmal; Muhammad Umer Sarwar; Muhammad Kashif Hanif; Muhammad Irfan Khan, IEEE Access, vol. 13, 2025, E.

[2] Large Language Models in Human-Robot Collaboration with Cognitive Validation Against Context-Induced Hallucinations-Nadun Ranasinghe;Wael M. Mohammed;Kostas Stefanidis; Jose L. Martinez Lastra IEEE Access Year: 2025

[3] The Impact of LLM Hallucinations on Motor Skill Learning: A Case Study in Badminton — Yepeng Qiu, IEEE Access, vol. 12, 2025, IEEE.

[4] hallucinations in Medical AI: A Knowledge Graph-Augmented Retrieval System for Evidence-Based Age-Related Macular Degeneration Information: - Alexandru Lecu;Adrian Groza;Lezan Hawizy IEEE Access Year: 2025, Volume: 13, Publisher: IEEE

[5] T. Wei, H. Chen, W. Liu, L. Chen, P. Gu, and J. Wang, "Optical and SAR Cross-Modal Hallucination Collaborative Learning for

RemoteSensing Missing-Modality Building Footprint Extraction," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 19, 2026.

[6] C. Woesle, L. Fischer-Brandies, and R. Buettner, "A Systematic Literature Review of Hallucinations in Large Language Models," IEEE Access, vol. 13, 2025.

[7] N. Ranasinghe, W. M. Mohammed, K. Stefanidis, and J. L. Martinez Lastra, "Large Language Models in Human-Robot Collaboration with Cognitive Validation Against Context-Induced Hallucinations," IEEE Access, vol. 13, 2025

[8] A.Lecu, A.Groza and L.Hawizy Reducing Hallucinations in Medical AI: A Knowledge Graph- Augmented Retrieval System for Evidence –Based Agent"IEEE Access, vol. 13, 2025

[9] S. B Shah, S.Thap, A.Acharya, K. Rauniyar ,S.Poudel, S.Jain, A.Masood, And U.Naseem "Navigating the web of Disinformation and Misinformation: Large Language Models as Double-Edged Swords"IEEE Access, vol. 13, 2025

[10] B.Saha, U.Saha and M.Z. Malik,QuIM-RAG: Advancing Retrieval-Augmented Generation with Inverted Question Matching for Enhanced QA Performance"IEEE Access, vol. 12, 2024