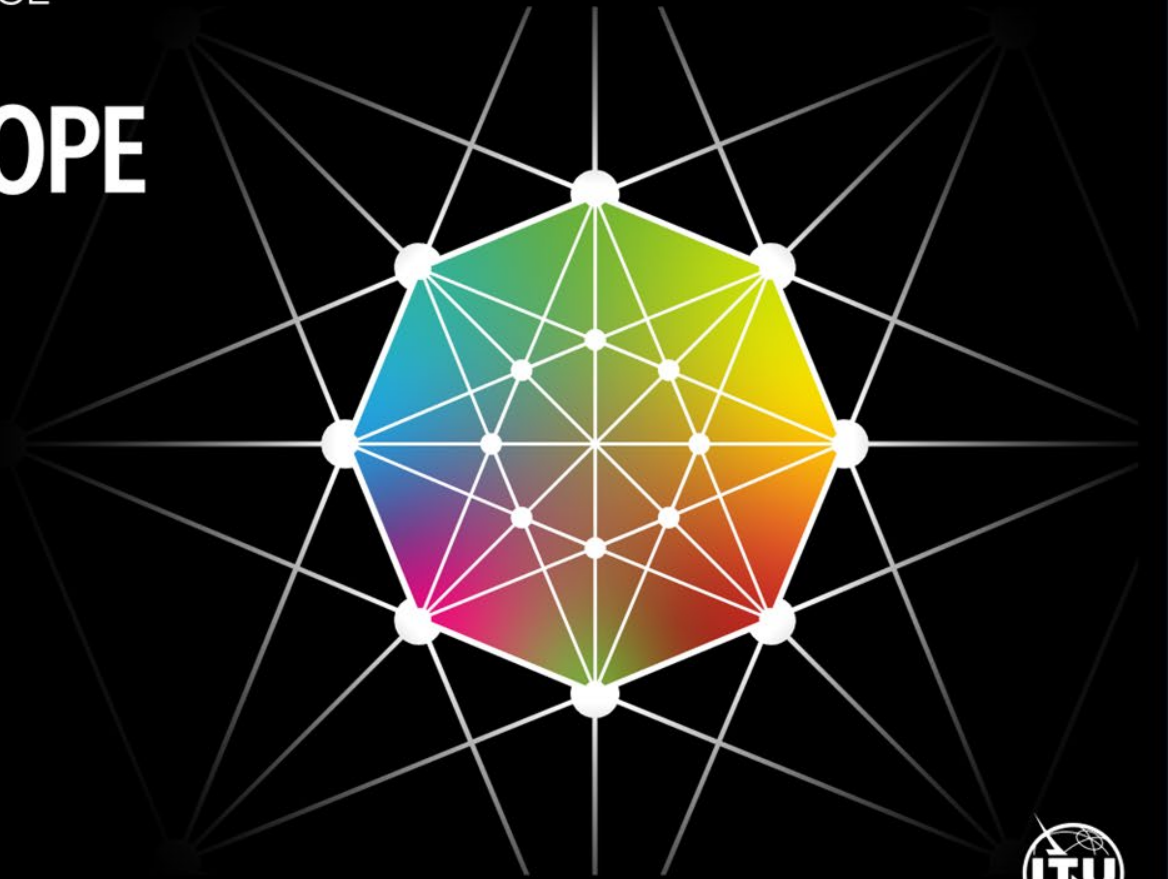
16TH ITU ACADEMIC CONFERENCE

ITU KALEIDOSCOPE
GENEVA2026

AI and frontier technologies for good

7-9 July 2026
Geneva, Switzerland

Hosted by AI for Good Global Summit



BRIDGING LANGUAGE DIVERSITY: A RETRIEVAL AUGMENTED GENERATION-BASED FRAMEWORK FOR INFORMATION EXTRACTION FOR SOUTH INDIAN LANGUAGE DOCUMENTS

Dr Gobi Ramasamy
CHRIST University

16TH ITU ACADEMIC CONFERENCE

ITU KALEIDOSCOPE GENEVA2026

Bridging Language Diversity :

A Retrieval Augmented Generation (RAG)-Based Framework for Information Extraction from South Indian Language Documents

01 Introduction & Problem Statement
Overview of information extraction challenges in South Indian languages and the limitations of existing approaches.

02 Research Objectives
Goals of the study focusing on accurate, context-aware, and multilingual information extraction using RAG and LLMs.

03 Literature Review
Review of RAG, information retrieval, LLMs, and recent advancements for low-resource and Indian languages.

04 Proposed Methodology
Overview of the RAG pipeline, document processing, embedding, retrieval, reranking, and generation process.

05 System Architecture
Four-layer architecture: Frontend, Processing, RAG Pipeline Core, and LLM Layer with component interactions.

06 Key Features of the Proposed System
Multilingual interaction, multi-format support, OCR, reranking, translation, glossary extraction, and batch document processing.

07 Mathematical Model & RAG Framework
Mathematical formulation of dense retrieval, cross-encoder reranking, concept similarity, and final context selection.

08 Results & Discussion
Performance analysis of foundation models (Gemini, Llama, DeepSeek) in terms of latency, multilingual support, and context length.

09 Conclusion & Future Scope
Summary of findings and future enhancements including more Indian languages, domain expansion, and ontology-based RAG.

Motivation & Problem Statement

WHY THIS RESEARCH?

- India has **22** official languages, **121** additional languages, and over **1,600** dialects.
- South Indian languages (Malayalam, Kannada, Tamil, Telugu) are **morphologically rich** and **agglutinative**.
- Existing Information Extraction systems are primarily designed for **high-resource** languages.
- Large Language Models often generate **hallucinated responses** due to insufficient contextual grounding.

RESEARCH PROBLEM

How can Retrieval-Augmented Generation (RAG) improve accurate information extraction from multilingual South Indian languages while reducing hallucinations?

Research Objectives

OBJECTIVES >>>

- Develop a **multilingual RAG-based framework** for Information Extraction.
- Support **Malayalam, Kannada, Tamil**, and **Telugu** documents.
- **Reduce hallucinations** using document-grounded retrieval.
- Enable users to query documents in their **native language**.
- Improve semantic understanding using **multilingual embeddings**.
- Evaluate multiple **Foundation Models** for multilingual performance.

Literature Review

EXISTING APPROACHES

Traditional Information Extraction
Rule-based systems

Machine Learning

Knowledge Graphs

Large Language Models
- Strong reasoning capability
- Limited by hallucinations
- Depend on internal knowledge

RETRIEVAL-AUGMENTED GENERATION (RAG)

ADVANTAGES:

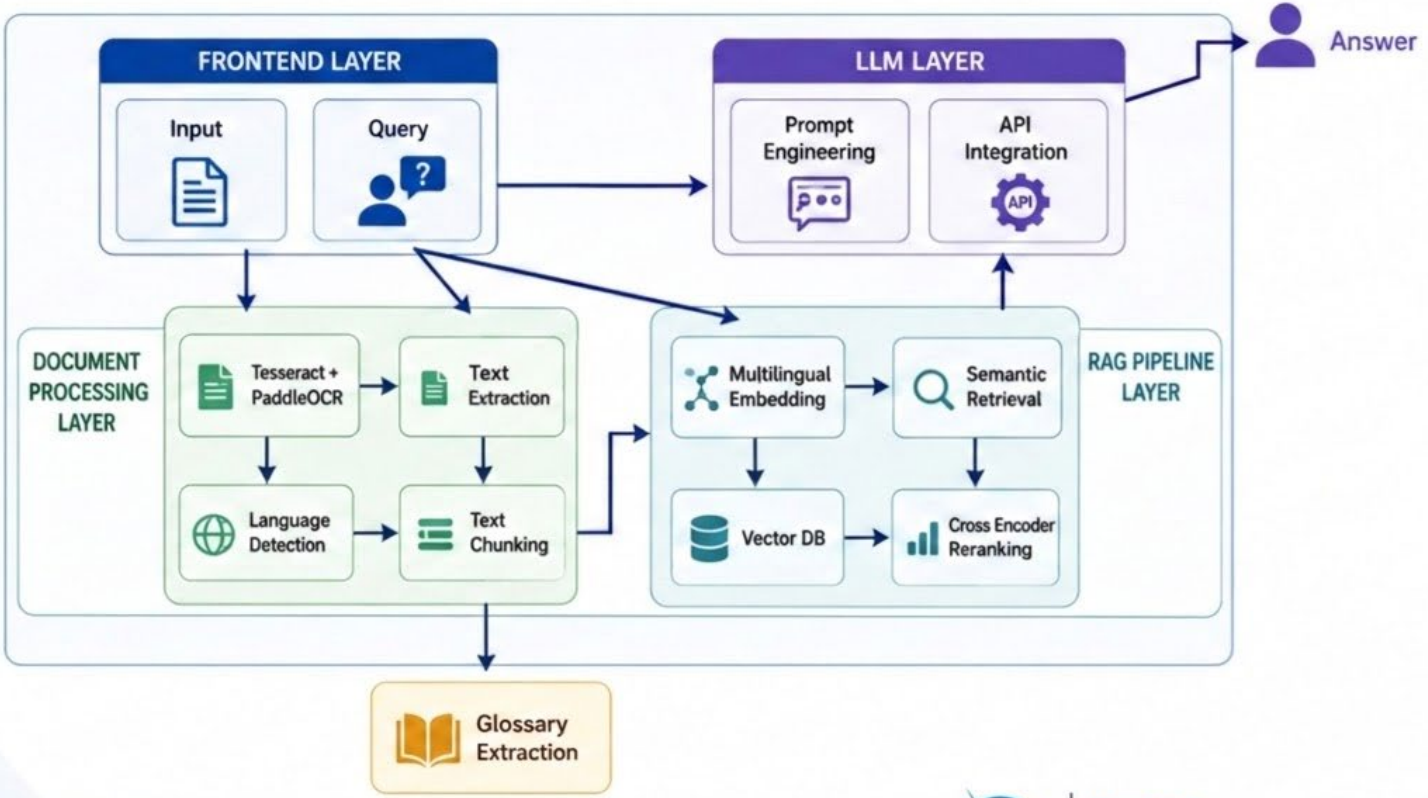
- Retrieves relevant document context
- Generates grounded responses
- Improves factual accuracy
- Better multilingual retrieval

RESEARCH GAP

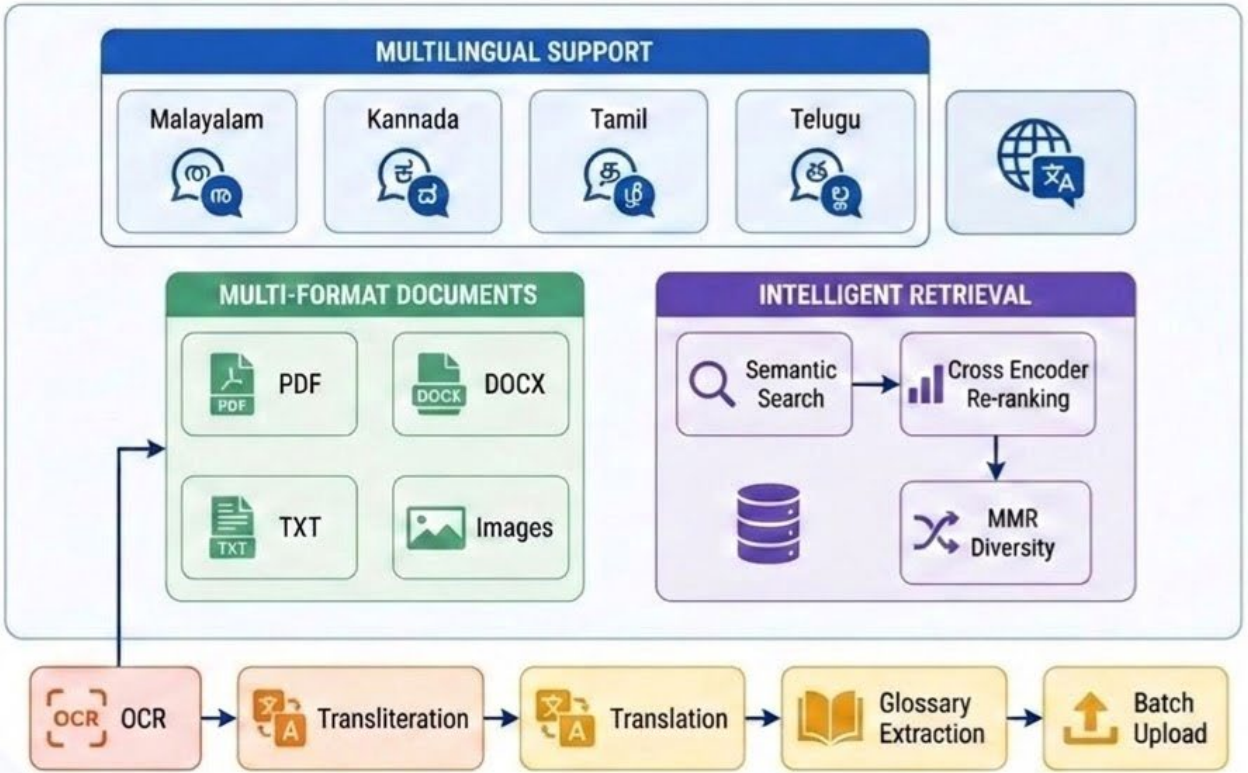
Limited work on **RAG** for South Indian languages.

Scarcity of **multilingual datasets** and **evaluation benchmarks.**

SYSTEM ARCHITECTURE



Key Features

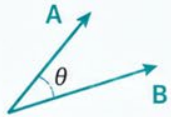


MATHEMATICAL MODEL

Information Retrieval Process using Retrieval-Augmented Generation (RAG)

1 Generate Embeddings

2 Compute Cosine Similarity



$$\text{Similarity} = \frac{A \cdot B}{\|A\| \, \|B\|}$$

3 Retrieve Top-k Chunks

- Chunk 1
- Chunk 2
- ...
- Chunk k

4 Cross Encoder Re-ranking

5 Concept Similarity

6 Prompt Construction

7 LLM Generates Final Answer

Accurate Answer

OUTCOMES

Better Context

Provides rich and relevant context for accurate information extraction.

Reduced Hallucination

Grounded responses using retrieved context minimize hallucinated outputs.

Higher Retrieval Accuracy

Improved semantic matching and reranking lead to more accurate results.

Comparison of Foundation Models

Model	Provider	Speed	Best For	Multilingual	Context	Cost
gemini-2.5-flash	Google Gemini	Very Fast	General QA, Quick responses	Excellent	1M tokens	Low
gemini-2.5-pro		Fast	Complex reasoning, Detailed answers	Excellent	1M tokens	Medium
llama-3.1-70b-versatile	Groq (Meta)	Very Fast	General purpose, Large context	Limited	128K tokens	Low
llama-3.1-8b-instant		Extremely Fast	Quick responses, Simple queries	Limited	128K tokens	Very Low
deepseek-chat	DeepSeek	Fast	General chat, Balanced performance	Limited	64K tokens	Low
deepseek-coder		Fast	Code generation, Programming	Limited	64K tokens	Low

Conclusion & Future Work

CONCLUSION

- Proposed a **RAG-based multilingual** information extraction framework.
- Successfully supports **South Indian language documents**.
- Combines **semantic retrieval** with **Large Language Models**.
- Provides **context-aware** and **accurate responses**.
- Significantly **reduces hallucinations** through retrieval and re-ranking.

FUTURE ENHANCEMENTS

- Extend support to **additional Indian languages**.
- Integrate **Ontology-based RAG (OWL-RAG)**.
- Improve **relation extraction** capabilities.
- Develop larger **multilingual benchmark datasets**.

Thank you!