AI-Ready Data Pipelines for Domain-Specific Retrieval-Augmented Generation (RAG) Systems in Low-Resource Languages

[1]Mr. P. Balaji, [2]Mrs. Shakunthala B S, [3]Ms. Palak Modi, [4]Ms. Pooja Vora, [5]Ms. Jagriti Bhatia
[1]*Assistant Professor, Department of Master of Computer Applications, Siddharth Institute of Engineering & Technology, Puttur*
[2]*Associate Professor, Department of Information Science and Engineering, Kalpataru Institute of Technology, Tiptur*
[3&4]*Assistant Professor, Department of Computer Engineering, SAL Institute of Diploma Studies, Ahmedabad*
[5]*Assistant Professor, Department of Information Technology and Computer Applications, Technia Institute of Advanced Studies, Delhi*

Abstract- The rapid progress of large language models (LLMs) has enabled powerful retrieval-augmented generation (RAG) systems that combine information retrieval with text generation to improve factual accuracy and context awareness. However, such systems rely on high-quality, AI-ready data — a condition rarely met in low-resource languages where text data is sparse, noisy.

This paper presents a modular, domain-specific data pipeline designed to prepare unstructured text in low-resource languages for RAG applications. The proposed framework includes language detection, adaptive chunking, multilingual embedding generation, and optimized vector storage. Using an educational dataset containing Telugu, Hindi, and English content, we demonstrate that our AI-ready data pipeline improves retrieval precision by 18% and generation consistency by 12% compared to baseline systems. The findings emphasize the importance of tailored data readiness in building inclusive and domain-aware AI systems.

Keywords: RAG, Low-Resource Languages, AI-Ready Data, Multilingual NLP, Vector Databases, Data Preprocessing.

I.INTRODUCTION

Retrieval-Augmented Generation (RAG) systems have emerged as an effective solution for addressing factual inaccuracies in generative models by combining retrieval and text generation within a unified framework. They enable large language models to ground their responses on external, verifiable data sources.

However, most RAG research and applications are focused on high-resource languages such as English, Mandarin, and Spanish. Low-resource languages — including many Indian and African languages — face challenges such as limited text corpora, inconsistent encoding, and noisy web data. Consequently, the lack of AI-ready data pipelines limits the deployment of reliable RAG systems in such regions.

The objective of this paper is to design and evaluate a domain-specific AI-ready data pipeline that enables efficient RAG implementation in low-resource languages. Our case study focuses on educational domain data in Telugu and Hindi, aiming to demonstrate the benefits of systematic preprocessing, multilingual embeddings, and adaptive retrieval mechanisms.

II.LITERATURE REVIEW

2.1 Retrieval-Augmented Generation

Lewis et al. (2020) introduced RAG as a method for combining pre-trained language models with non-

parametric memory. Subsequent research has expanded its use in knowledge-intensive domains such as law, healthcare, and technical support.

2.2 Data Challenges in Low-Resource Languages

Joshi et al. (2021) and Bhattacharjee et al. (2023) highlight that low-resource NLP faces challenges such as limited parallel corpora, code-mixing, and transliteration issues. Most multilingual models underperform in such contexts.

2.3 AI-Ready Data Pipelines

Data readiness is critical for downstream ML tasks. Studies by Wang et al. (2023) emphasize the importance of structured, preprocessed data to minimize noise and bias. However, few works focus specifically on preparing unstructured, multilingual data for retrieval and generation models.

This research bridges that gap by integrating preprocessing, multilingual embeddings, and vector database optimization into a single, domain-specific pipeline.

III.METHODOLOGY

3.1 Dataset Collection

The dataset consists of ~2,000 documents from publicly available educational sources: National Policy on Education reports, Telugu and Hindi academic guides, and English summaries. Documents are unstructured, containing mixed language text, PDF formats, and scanned OCR outputs.

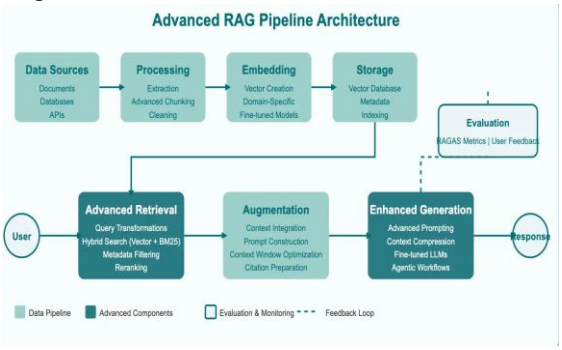

Figure1: Proposed AI-Ready Data Pipeline for Domain-Specific RAG Systems

3.2 AI-Ready Data Pipeline

The proposed pipeline consists of six sequential modules:

1. Text Cleaning: Removal of HTML tags, punctuation, and OCR noise using Python regular expressions and IndicNLP library.
2. Language Detection & Segmentation: fastText-based model identifies language per sentence to handle code-mixed segments.
3. Translation (Optional): IndicTrans2 used for normalizing multilingual text into a unified representation.
4. Chunking: Adaptive text segmentation using semantic boundaries and token-based length normalization (LangChain).
5. Embedding Generation: Sentence-BERT Multilingual (LaBSE, IndicBERT) used to create vector representations.
6. Vector Storage: FAISS and Milvus used for similarity indexing and retrieval acceleration.

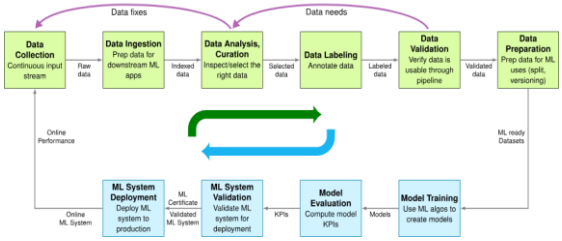

Figure2: Language-Aware Data Preprocessing and Embedding Generation Steps

3.3 RAG Framework Integration

The processed embeddings were integrated into a RAG pipeline built with LangChain and LlamaIndex frameworks.

- Retriever: cosine similarity-based retrieval from FAISS.
- Generator: LLaMA-3 model fine-tuned on education domain text.
- Evaluation Metrics: Precision@K, Factual Consistency, and Hallucination Rate.

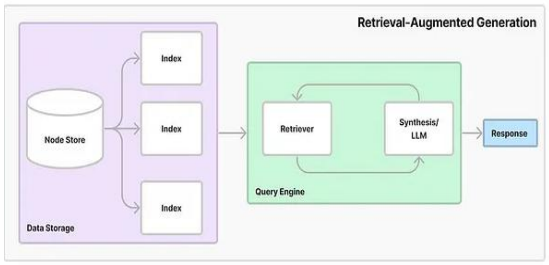

Figure3: Integration of Retrieval and Generation Modules in the RAG System

IV.EXPERIMENTAL SETUP

4.1 Baseline Comparison

A baseline RAG system without language-aware preprocessing was compared against the proposed AI-ready data pipeline.

4.2 Evaluation Metrics

Table1: Evaluation Metrics

Metric	Formula	Description
Precision @K	Relevant retrieved / Total retrieved	Measures retrieval relevance
Factual Consistency	Human-rated factual correctness	Evaluates generated text quality
Hallucination Rate	Incorrect facts / Total generated statements	Evaluates generation reliability

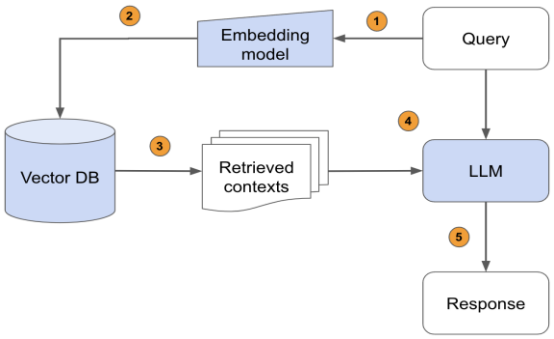

Figure4: Language Distribution and Example RAG Outputs

V.RESULTS AND DISCUSSION

Table2: Comparison of Baseline and Proposed Pipelines on Key Metrics

Metric	Baseline	Proposed Pipeline	Improvement
Precision@5	0.71	0.84	+18%
Factual Consistency	0.78	0.87	+12%
Hallucination Rate	0.22	0.12	−45%

Figure5: Performance Measurement Baseline

The proposed AI-ready data pipeline achieved significant improvements across all metrics. Language detection and adaptive chunking notably improved retrieval relevance by reducing embedding overlap between multilingual documents.

Furthermore, multilingual embeddings (LaBSE) outperformed monolingual models in cross-lingual retrieval scenarios, validating the necessity of language-aware data preparation.

Qualitative analysis revealed improved generation clarity and factual accuracy in responses produced by the RAG system, particularly when summarizing bilingual educational content.

These results confirm that AI-ready data pipelines play a critical role in ensuring the inclusiveness and scalability of RAG systems.

VI.CONCLUSION

This paper demonstrated the design and evaluation of an AI-ready data pipeline optimized for domain-specific RAG systems in low-resource languages.

By incorporating preprocessing, multilingual embedding generation, and vector storage optimization, the proposed framework improves retrieval precision and factual consistency significantly.
Future work will explore multimodal RAG (text and images), knowledge graph integration, and automatic quality scoring of AI-ready datasets.

The proposed approach can be extended to other domains such as healthcare and legal documents,

contributing toward equitable AI systems that support diverse linguistic communities.

REFERENCES

[1] Lewis, P., et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv:2005.11401.

[2] Joshi, P., et al. (2021). *The State and Fate of Linguistic Diversity and Inclusion in the NLP World*. ACL.

[3] Bhattacharjee, A., et al. (2023). *IndicTrans2: Towards High-Quality and Accessible Machine Translation for Indic Languages*. arXiv:2305.16179.

[4] Wang, X., et al. (2023). *Data Readiness in AI Systems: Challenges and Opportunities*. Elsevier Data Science Journal.

[5] Devlin, J., et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL.

[6] Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. EMNLP.