Multilingual PDF RAG System

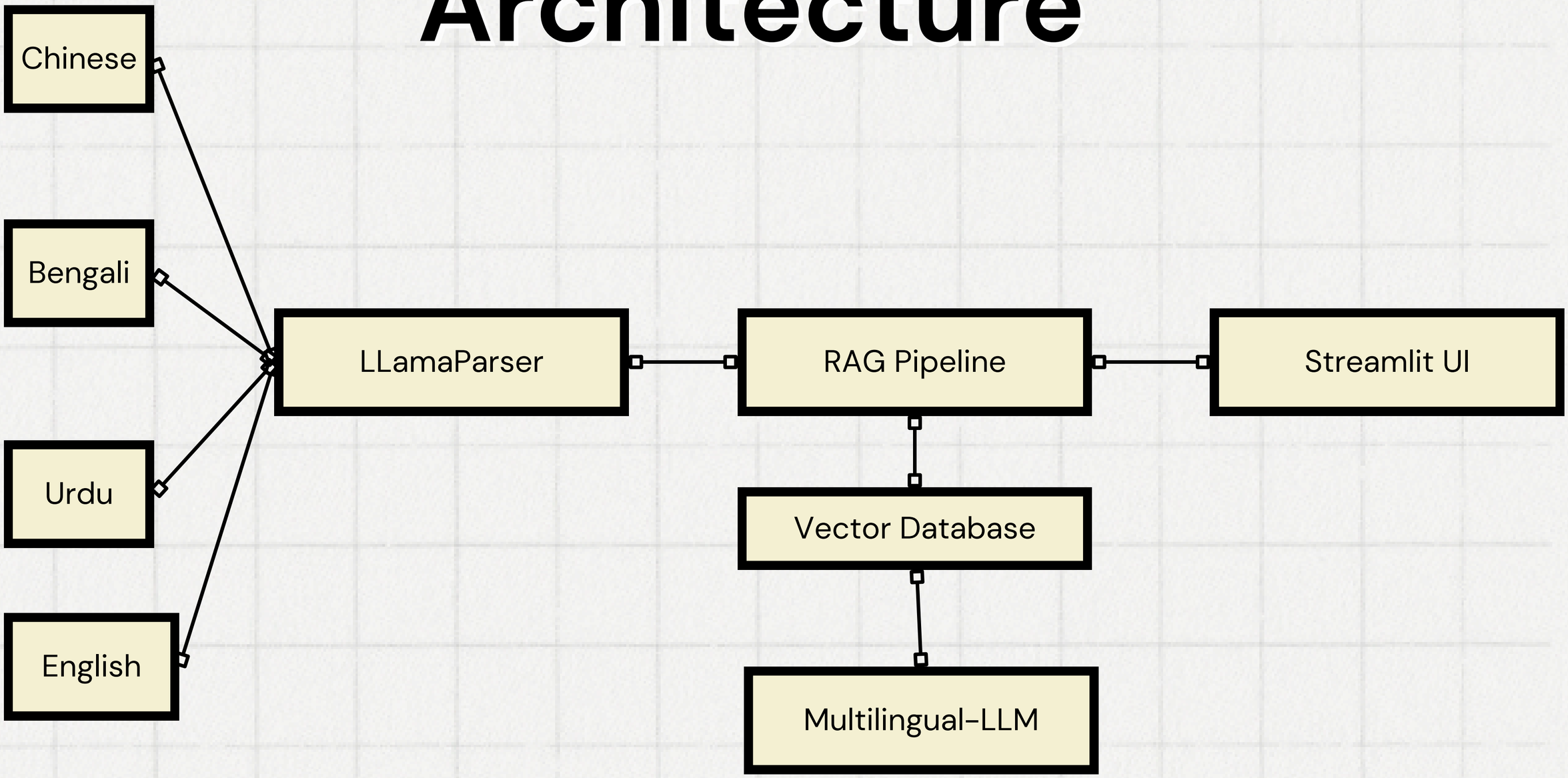
Presented by Saurabh Naik

Problem Statment

Develop a RAG pipeline for summarizing content and answering questions based on the input PDFs. The system should be scalable to handle large amounts of data (up to 1TB) and provide accurate, relevant responses.

PPT Contents

Architecture



Challenges Faced

1. Document parsing for other languages than English
2. Unable to use Llamaindex with huggingface models
3. Unable to use gemini due to limit request
4. Hybrid Search/ Chunks with metadata was not done due to change in framework
5. Llamaindex provides function for agentic workiflow of Summary and QA

Future Enhancements

01

Check how to work with huggingface with LLamaindex

02

Graph QA can significantly improve results

03

Better Transformer models suitable multilingual text

04

Combination of GraphQA and Vector/Hybrid Search

Thank you very much!