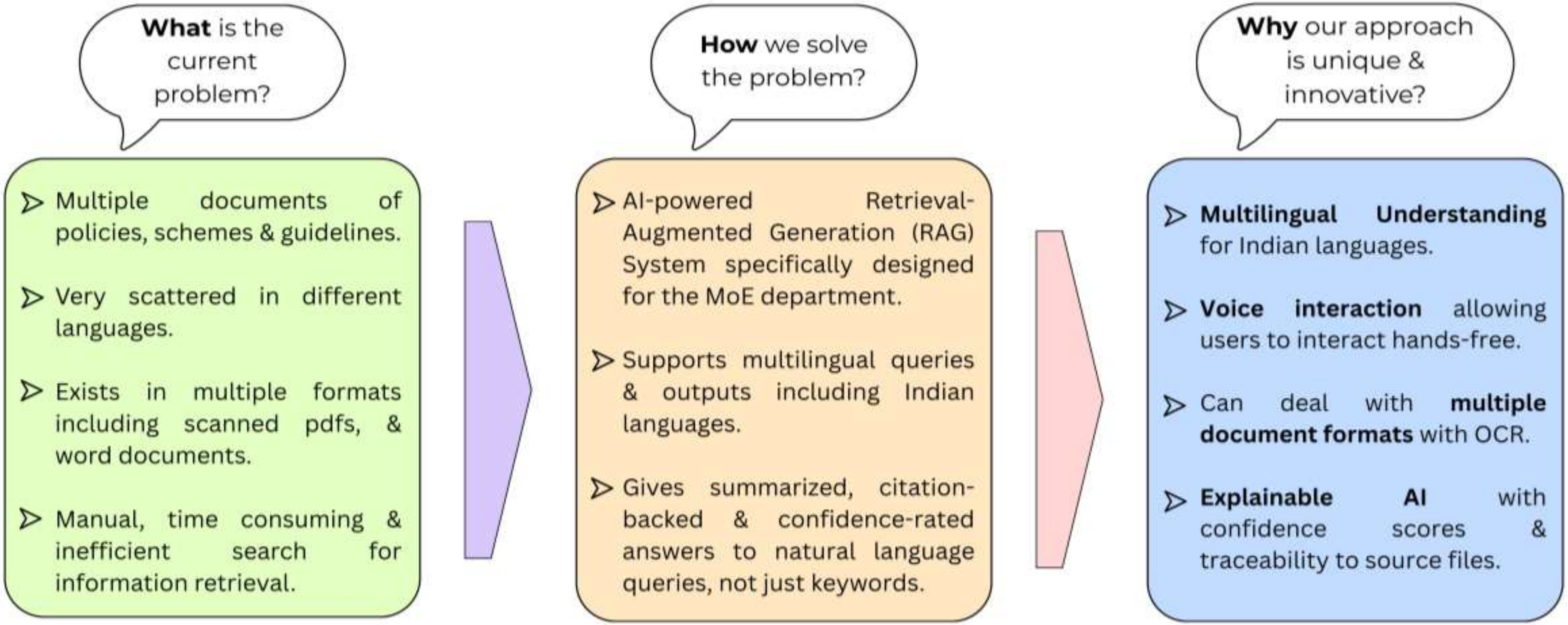
RAG CHATBOT FOR EFFICIENT RETRIEVAL OF INFORMATION

VidyAI

PROBLEM STATEMENT TITLE - RETRIEVAL OF DATA FROM KNOWLEDGE BASE OF COLLEGE WEBSITE AND STUDENT RESOURCE BOOK.

VidyAI

IDEA : AI –DRIVEN RAG CHATBOT FOR EFFICIENT & QUICK RETRIEVAL OF INFORMATION



What is the current problem?

- ⊳ Multiple documents of policies, schemes & guidelines.
- ⊳ Very scattered in different languages.
- ⊳ Exists in multiple formats including scanned pdfs, & word documents.
- ⊳ Manual, time consuming & inefficient search for information retrieval.

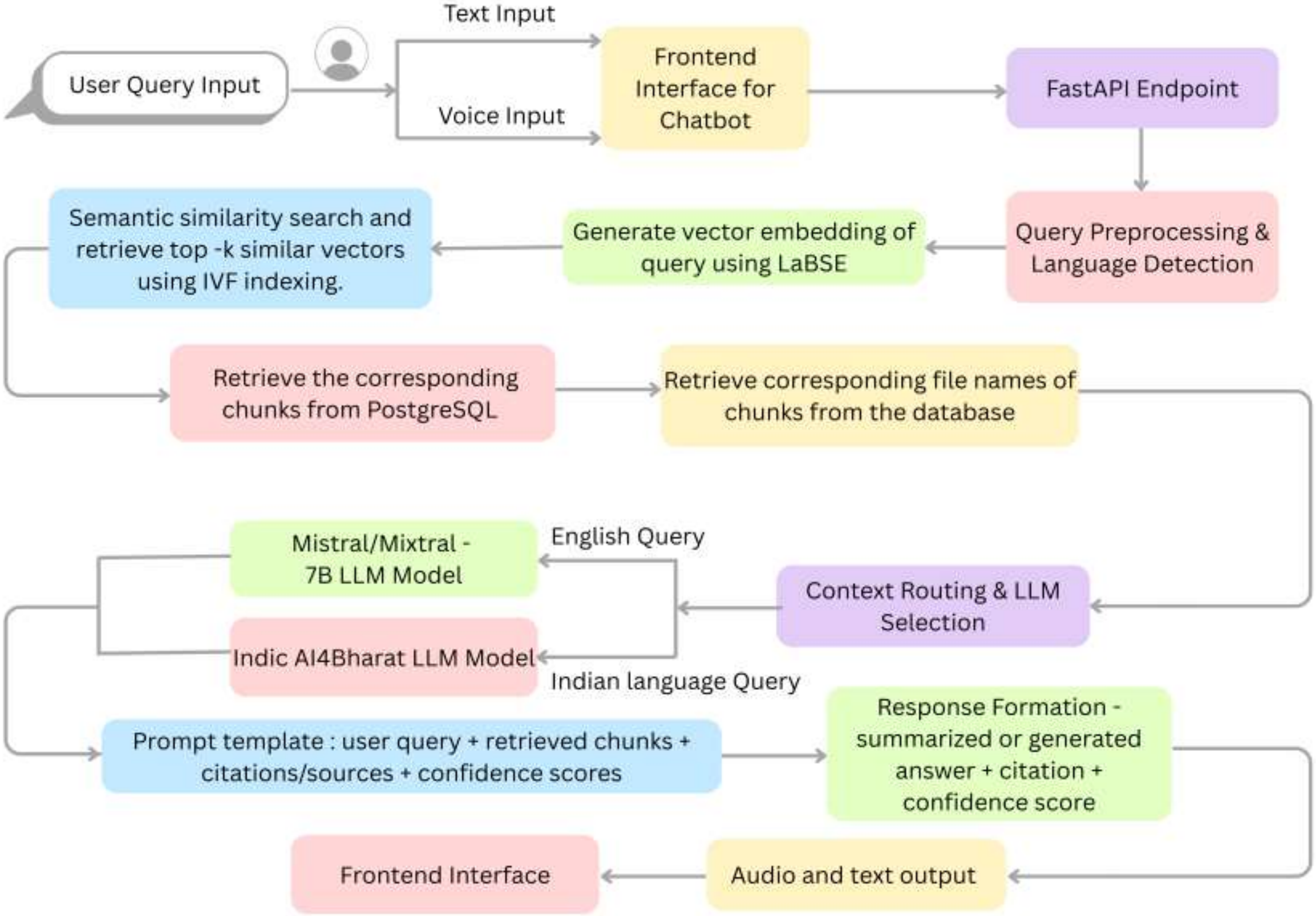
How we solve the problem?

- ⊳ AI-powered Retrieval-Augmented Generation (RAG) System specifically designed for the MoE department.
- ⊳ Supports multilingual queries & outputs including Indian languages.
- ⊳ Gives summarized, citation-backed & confidence-rated answers to natural language queries, not just keywords.

Why our approach is unique & innovative?

- ⊳ **Multilingual Understanding** for Indian languages.
- ⊳ **Voice interaction** allowing users to interact hands-free.
- ⊳ Can deal with **multiple document formats** with OCR.
- ⊳ **Explainable AI** with confidence scores & traceability to source files.

TECHNICAL APPROACH



User Query Input

Text Input

Voice Input

Frontend Interface for Chatbot

FastAPI Endpoint

Query Preprocessing & Language Detection

Generate vector embedding of query using LaBSE

Semantic similarity search and retrieve top -k similar vectors using IVF indexing.

Retrieve the corresponding chunks from PostgreSQL

Retrieve corresponding file names of chunks from the database

Context Routing & LLM Selection

English Query

Mistral/Mixtral - 7B LLM Model

Indian language Query

Indic AI4Bharat LLM Model

Prompt template : user query + retrieved chunks + citations/sources + confidence scores

Response Formation - summarized or generated answer + citation + confidence score

Audio and text output

Frontend Interface

TOOLS & FRAMEWORKS

Whisper

LangChain

python™

FAISS

Scalable Search With Facebook AI

MISTRAL AI_

8X7

Flask

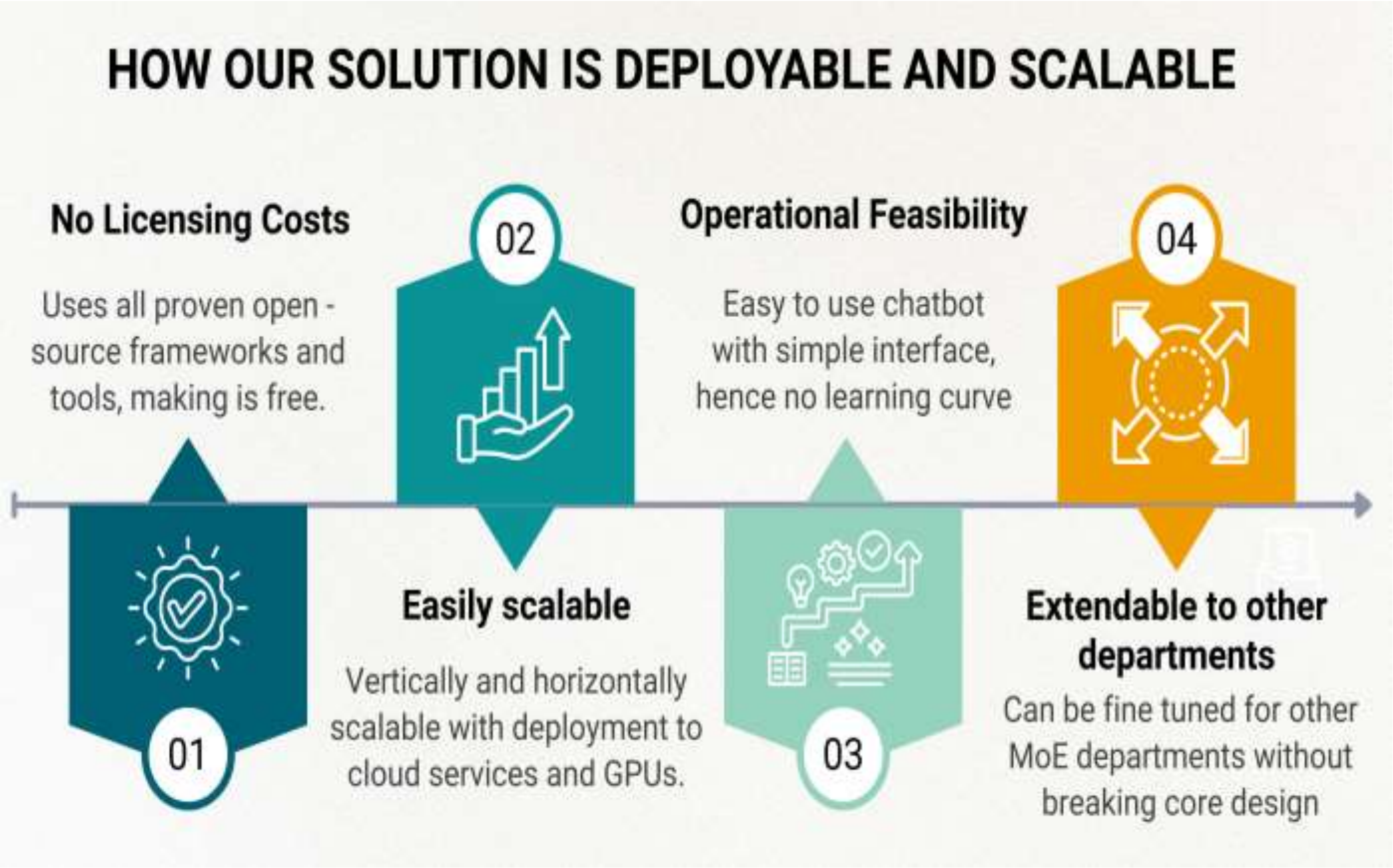
gtts

HTML

CSS

JS

FEASIBILITY AND VIABILITY



HOW OUR SOLUTION IS DEPLOYABLE AND SCALABLE

01 **No Licensing Costs**

Uses all proven open - source frameworks and tools, making is free.

02 **Easily scalable**

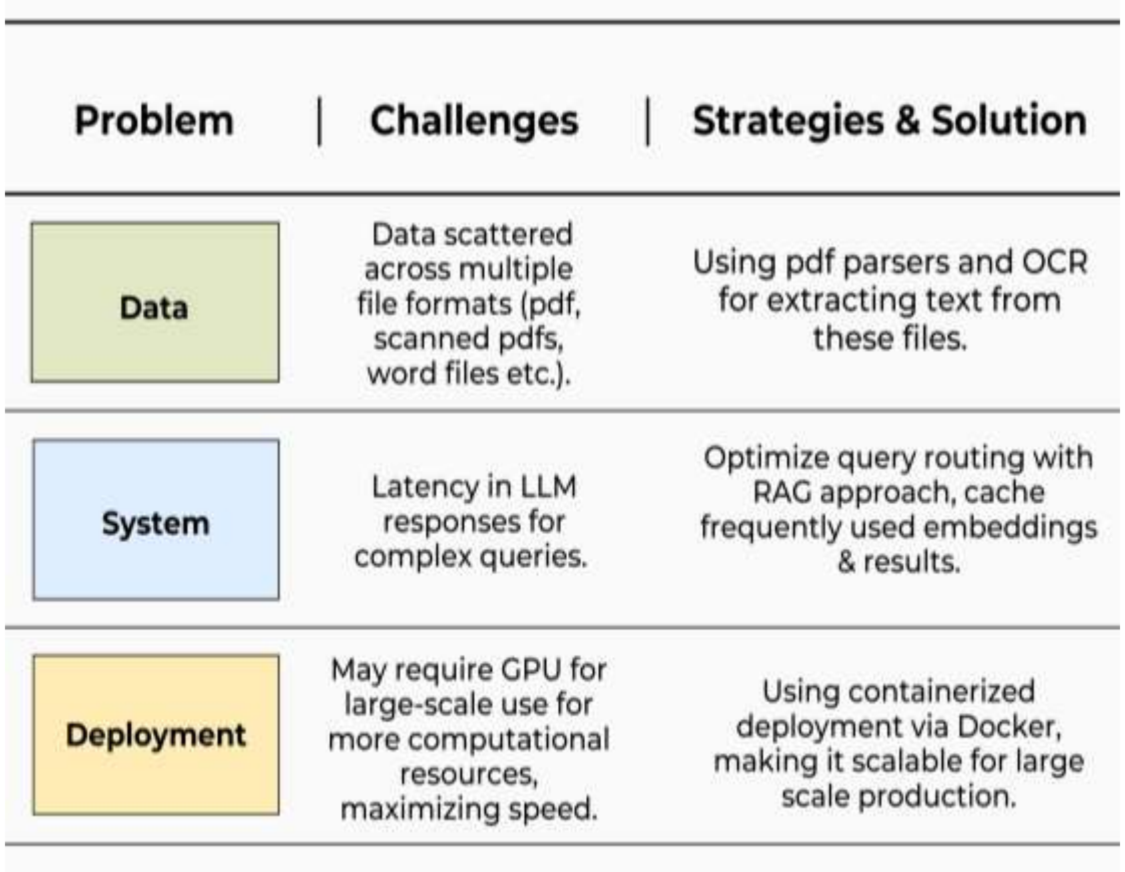
Vertically and horizontally scalable with deployment to cloud services and GPUs.

03 **Operational Feasibility**

Easy to use chatbot with simple interface, hence no learning curve

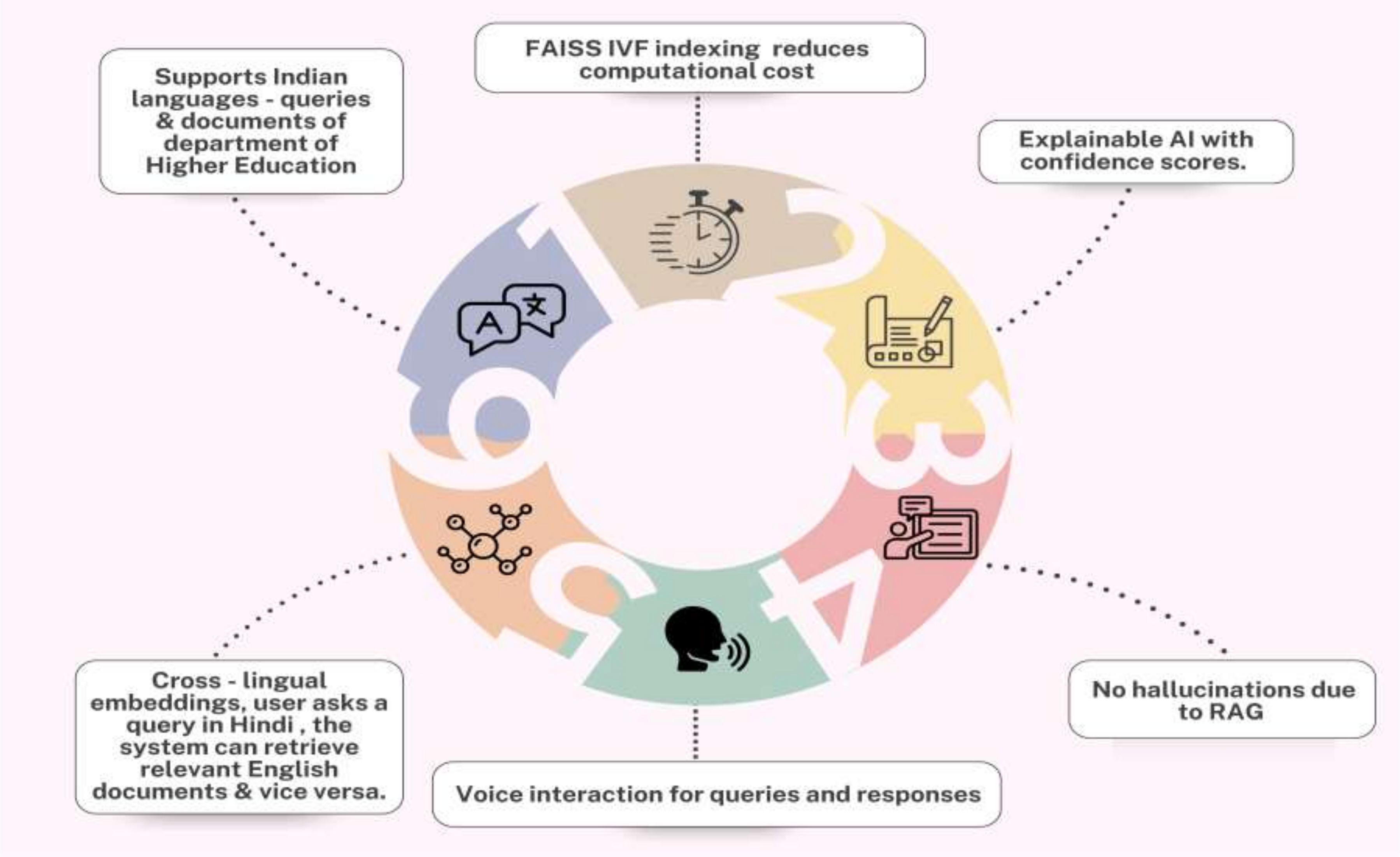
04 **Extendable to other departments**

Can be fine tuned for other MoE departments without breaking core design



Problem	Challenges	Strategies & Solution
Data	Data scattered across multiple file formats (pdf, scanned pdfs, word files etc.).	Using pdf parsers and OCR for extracting text from these files.
System	Latency in LLM responses for complex queries.	Optimize query routing with RAG approach, cache frequently used embeddings & results.
Deployment	May require GPU for large-scale use for more computational resources, maximizing speed.	Using containerized deployment via Docker, making it scalable for large scale production.

IMPACT AND BENEFITS



Supports Indian languages - queries & documents of department of Higher Education

FAISS IVF indexing reduces computational cost

Explainable AI with confidence scores.

No hallucinations due to RAG

Voice interaction for queries and responses

Cross - lingual embeddings, user asks a query in Hindi , the system can retrieve relevant English documents & vice versa.

RESEARCH AND REFERENCES

1) https://huggingface.co/sentence-transformers/LaBSE
2) https://huggingface.co/ai4bharat/Airavata?
3) https://python.langchain.com/docs/tutorials/rag/
4) https://www.education.gov.in/en/higher_education
5) https://towardsai.net/p/machine-learning/a-complete-guide-to-rag?
6) https://huggingface.co/docs/transformers/model_doc/rag?
7) https://github.com/stackitcloud/rag-template?
8) https://github.com/adithya-s-k/Indic-llm?
9) https://github.com/Zlash65/rag-bot-fastapi?