Conversational AI
Natural Language Processing (UCS664)

Project: RAG+LLM Legal QA Answering System in Hindi

Submitted By:
Shiven Khare – 102203852
Rehatman Kaur – 102203730

B.E. Third Year – COE

Group: 3CO20/3C45

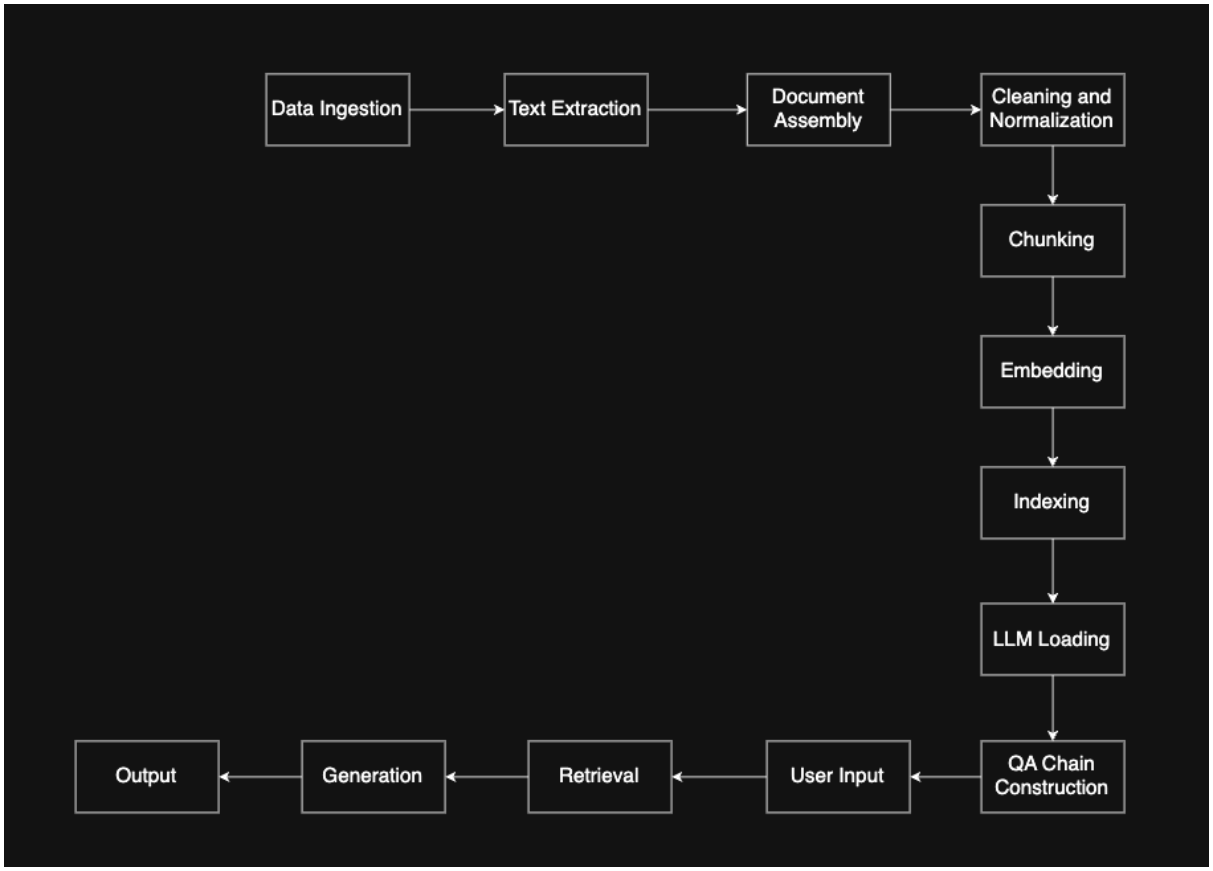
Submitted To:
Dr. Jasmeet Singh

Department of Computer Science and Engineering
Thapar Institute of Engineering and Technology
Jan – June 2025 (2425EVESEM 6th Semester)

Introduction to Project

In recent years, the sheer volume and complexity of legal documents—statutes, regulations, case law, and scholarly commentary—has made it increasingly difficult for non-experts and even seasoned practitioners to locate and interpret relevant provisions rapidly. In India, where legislation and constitutional text exist in both English and Hindi, there is a pressing need for tools to bridge the gap between dense legal language and everyday queries in the user's preferred language. This project presents a prototype **Legal QA Chatbot** designed to answer questions about the Indian Constitution and the Indian Penal Code (IPC) directly in Hindi, leveraging modern **Retrieval-Augmented Generation (RAG)** techniques and a lightweight multilingual language model.

At its core, our system ingests "born-digital" Hindi texts of the Constitution and IPC—alongside curated question-answer pairs from publicly available datasets—and converts them into a searchable knowledge base. We extract and normalize the raw Unicode text, split it into overlapping chunks to respect the model's context window and embed each chunk using a **multilingual sentence transformer**. These vectors are indexed in **FAISS**, enabling lightning-fast similarity search. When a user submits a Hindi question, the system retrieves the top-k relevant excerpts and "stuff" them into the prompt of a small, instruction-tuned multilingual model (bigscience/mt0-small). The model then generates a concise, accurate answer entirely in Hindi, accompanied by citations of the source chunks.



Dataset Used

Constitution of India and IPC PDFs in Hindi downloaded from, https://legislative.gov.in/constitution-of-india/ and https://vault.drishtijudiciary.com/hindi_file_uploads/1692628780_IPC.pdf.
Extracted text from the PDFs using OCR and created .txt files from them.

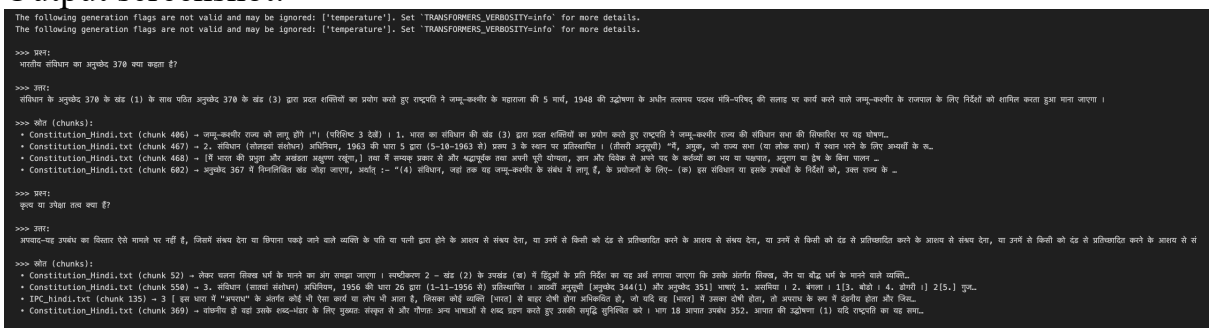
Model Used

For embedding Hindi chunks into vector space we used the multilingual MiniLM: sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

For generation we used instruction-tuned, multilingual MT0-small: bigscience/mt0-small

Result

Output screenshot:



Output Text:

The following generation flags are not valid and may be ignored: ['temperature']. Set `TRANSFORMERS_VERBOSITY=info` for more details.
The following generation flags are not valid and may be ignored: ['temperature']. Set `TRANSFORMERS_VERBOSITY=info` for more details.

>>> प्रश्न:
भारतीय संविधान का अनुच्छेद 370 क्या कहता है?

>>> उत्तर:
संविधान के अनुच्छेद 370 के खंड (1) के साथ पठित अनुच्छेद 370 के खंड (3) द्वारा प्रदत्त शक्तियों का प्रयोग करते हुए राष्ट्रपति ने जम्मू-कश्मीर के महाराजा की 5 मार्च, 1948 की उद्घोषणा के अधीन तत्समय पदस्थ मंत्रि-परिषद् की सलाह पर कार्य करने वाले जम्मू-कश्मीर के राजपाल के लिए निर्देशों को शामिल करता हुआ माना जाएगा ।

>>> स्रोत (chunks):
• Constitution_Hindi.txt (chunk 406) → जम्मू-कश्मीर राज्य को लागू होंगे ।"। (परिशिष्ट 3 देखें) । 1. भारत का संविधान की खंड (3) द्वारा प्रदत्त शक्तियों का प्रयोग करते हुए राष्ट्रपति ने जम्मू-कश्मीर राज्य की संविधान सभा की सिफारिश पर यह घोषण…
• Constitution_Hindi.txt (chunk 467) → 2. संविधान (सोलहवां संशोधन) अधिनियम, 1963 की धारा 5 द्वारा (5-10-1963 से) प्ररूप 3 के स्थान पर प्रतिस्थापित । (तीसरी अनुसूची) "मैं, अमुक, जो राज्य सभा (या लोक सभा) में स्थान भरने के लिए अभ्यर्थी के रू…
• Constitution_Hindi.txt (chunk 468) → [मैं भारत की प्रभुता और अखंडता अक्षुण्ण रखूंगा,] तथा मैं सम्यक् प्रकार से और श्रद्धापूर्वक तथा अपनी पूरी योग्यता, ज्ञान और विवेक से अपने पद के कर्तव्यों का भय या पक्षपात, अनुराग या द्वेष के बिना पालन …
• Constitution_Hindi.txt (chunk 602) → अनुच्छेद 367 में निम्नलिखित खंड जोड़ा जाएगा, अर्थात् :- "(4) संविधान, जहां तक यह जम्मू-कश्मीर के संबंध में लागू हैं, के प्रयोजनों के लिए- (क) इस संविधान या इसके उपबंधों के निर्देशों को, उक्त राज्य के …

>>> प्रश्न:
कृत्य या उपेक्षा तत्व क्या हैं?

>>> उत्तर:

अपवाद-यह उपबंध का विस्तार ऐसे मामले पर नहीं है, जिसमें संश्रय देना या छिपाना पकड़े जाने वाले व्यक्ति के पति या पत्नी द्वारा होने के आशय से संश्रय देना, या उनमें से किसी को दंड से प्रतिच्छादित करने के आशय से संश्रय देना, या उनमें से किसी को दंड से प्रतिच्छादित करने के आशय से संश्रय देना, या उनमें से किसी को दंड से प्रतिच्छादित करने के आशय से सं

>>> स्रोत (chunks):
• Constitution_Hindi.txt (chunk 52) → लेकर चलना सिक्ख धर्म के मानने का अंग समझा जाएगा । स्पष्टीकरण 2 - खंड (2) के उपखंड (ख) में हिंदुओं के प्रति निर्देश का यह अर्थ लगाया जाएगा कि उसके अंतर्गत सिक्ख, जैन या बौद्ध धर्म के मानने वाले व्यक्ति…
• Constitution_Hindi.txt (chunk 550) → 3. संविधान (सातवां संशोधन) अधिनियम, 1956 की धारा 26 द्वारा (1-11-1956 से) प्रतिस्थापित । आठवीं अनुसूची [अनुच्छेद 344(1) और अनुच्छेद 351] भाषाएं 1. असमिया । 2. बंगला । 1[3. बोडो । 4. डोगरी ।] 2[5.] गुज…
• IPC_hindi.txt (chunk 135) → 3 [इस धारा में "अपराध" के अंतर्गत कोई भी ऐसा कार्य या लोप भी आता है, जिसका कोई व्यक्ति [भारत] से बाहर दोषी होना अभिकथित हो, जो यदि वह [भारत] में उसका दोषी होता, तो अपराध के रूप में दंडनीय होता और जिस…
• Constitution_Hindi.txt (chunk 369) → वांछनीय हो वहां उसके शब्द-भंडार के लिए मुख्यतः संस्कृत से और गौणतः अन्य भाषाओं से शब्द ग्रहण करते हुए उसकी समृद्धि सुनिश्चित करे । भाग 18 आपात उपबंध 352. आपात की उद्घोषणा (1) यदि राष्ट्रपति का यह समा…

Code

```
%pip install pdf2image pytesseract pdfplumber langchain-
community transformers torch faiss-cpu sentence-
transformers

import logging, re
from pathlib import Path

import torch
import pdfplumber
from pdf2image import convert_from_path
import pytesseract
from transformers import AutoTokenizer,
AutoModelForSeq2SeqLM, pipeline

from langchain.schema import Document
from langchain.text_splitter import
RecursiveCharacterTextSplitter
from langchain.embeddings import HuggingFaceEmbeddings
from langchain.vectorstores import FAISS
from langchain.chains.question_answering import
load_qa_chain
from langchain.chains import RetrievalQA
from langchain_community.llms import HuggingFacePipeline
```

```
logging.getLogger("pdfminer").setLevel(logging.ERROR)
logging.getLogger("pdfplumber").setLevel(logging.ERROR)
logging.getLogger("pdf2image").setLevel(logging.ERROR)

if torch.cuda.is_available():
    DEVICE, DEV_ID = "cuda", 0
elif getattr(torch.backends, "mps", None) and
torch.backends.mps.is_available():
    DEVICE, DEV_ID = "mps", 0
else:
    DEVICE, DEV_ID = "cpu", -1

print("Using device:", DEVICE)

def extract_text_ocr(pdf_path: Path, lang: str = "hin") -
> str:
    pages = convert_from_path(str(pdf_path), dpi=300)
    texts = []
    for i, img in enumerate(pages, 1):
        txt = pytesseract.image_to_string(img, lang=lang)
        texts.append(txt)
    return "\n".join(texts)

data_dir = Path("./data")
ocr_dir  = Path("ocr_texts")
ocr_dir.mkdir(exist_ok=True)

pdf_hi = data_dir/"Constitution_Hindi.pdf"
txt_hi = ocr_dir/"Constitution_Hindi.txt"
if pdf_hi.exists() and not txt_hi.exists():
    text = extract_text_ocr(pdf_hi, lang="hin")
    txt_hi.write_text(text, encoding="utf-8")

pdf_ipc = data_dir/"IPC_hindi.pdf"
txt_ipc = ocr_dir/"IPC_hindi.txt"
if pdf_ipc.exists() and not txt_ipc.exists():
    text = extract_text_ocr(pdf_ipc, lang="hin")
    txt_ipc.write_text(text, encoding="utf-8")

for name in ["Constitution_Hindi.txt","IPC_hindi.txt"]:
    src = data_dir/name
    dst = ocr_dir/name
    if src.exists() and not dst.exists():
        dst.write_text(src.read_text("utf-8"),
encoding="utf-8")
```

```
constitution_hindi =
(ocr_dir/"Constitution_Hindi.txt").read_text(encoding="ut
f-8")
ipc_hindi          =
(ocr_dir/"IPC_hindi.txt").read_text(encoding="utf-8")
docs = [
    Document(page_content=constitution_hindi,
metadata={"source":"Constitution_Hindi.txt"}),
    Document(page_content=ipc_hindi,
metadata={"source":"IPC_hindi.txt"}),
]
print("Total Hindi source docs:", len(docs))

docs = [
    Document(page_content=constitution_hindi,
metadata={"source":"Constitution_Hindi.txt"}),
    Document(page_content=ipc_hindi,
metadata={"source":"IPC_hindi.txt"}),
]
print("Total Hindi source docs:", len(docs))

embed_model = "sentence-transformers/paraphrase-
multilingual-MiniLM-L12-v2"
hf_embed   =
HuggingFaceEmbeddings(model_name=embed_model)

faiss_dir = Path("faiss_hindi_index")
if faiss_dir.exists():
    vectordb = FAISS.load_local(str(faiss_dir), hf_embed)
    print("Loaded existing FAISS index.")
else:
    print("Building FAISS index…")
    vectordb = FAISS.from_documents(hindi_chunks,
hf_embed)
    vectordb.save_local(str(faiss_dir))
    print("Built & saved FAISS index.")

model_id = "bigscience/mt0-small"
tokenizer = AutoTokenizer.from_pretrained(model_id)

dtype = torch.float16 if DEVICE in ("cuda","mps") else
torch.float32
model = AutoModelForSeq2SeqLM.from_pretrained(model_id,
torch_dtype=dtype).to(DEVICE)

hindi_pipe = pipeline(
    "text2text-generation",
```

```
    model=model,
    tokenizer=tokenizer,
    max_new_tokens=256,
    temperature=0.7,
    device=DEVICE_ID,
    do_sample=True,
    top_p=0.9,
)
llm = HuggingFacePipeline(pipeline=hindi_pipe)
print(f"Loaded {model_id} on {DEVICE}")

combine_chain = load_qa_chain(llm=llm,
chain_type="stuff")
qa_chain = RetrievalQA(
    combine_documents_chain=combine_chain,

retriever=vectordb.as_retriever(search_kwargs={"k":4}),
    return_source_documents=True,
)
print("Hindi RetrievalQA chain ready")

def run_hindi_qa(question: str, verbose: bool=True):
    prompt = f"Answer in Hindi: {question}"
    out = qa_chain({"query": prompt})
    ans = out["result"]
    srcs = out["source_documents"]
    if verbose:
        print("\n>>> प्रश्न:\n", question)
        print("\n>>> उत्तर:\n", ans)
        print("\n>>> स्रोत (chunks):")
        for d in srcs:
            m = d.metadata
            snip = d.page_content.replace("\n"," ")[:200]
            print(f" • {m['source']} (chunk {m['chunk']})
→ {snip}…")
    return ans, srcs

_ = run_hindi_qa("भारतीय संविधान का अनुच्छेद 370 क्या कहता है?")
_ = run_hindi_qa("कृत्य या उपेक्षा तत्व क्या हैं?")
```