Submitted By,

Muhammad Awaiz

Reg No: 221001 (BSAI-VI-A)

Assignment 2

Submitted to,

Dr. Humaira Waqas

Architecture / Approach for Retrieving Context

Objectives

- Efficiently break down documents into semantically meaningful chunks.
- Accommodate linguistic and structural variation across document types and languages.
- Enable accurate and relevant context retrieval from chunked data.

Chunking Functions

The system defines several specialized chunking methods based on document content types:

a. chunk_questions_answers(text)

- Designed for Spanish documents in Q&A format.
- Uses regular expressions to identify question marks **(¿... ?)** and extracts the subsequent answer until the next question.
- Combines each Q&A pair into a single chunk.

b. chunk_numbered_sections(text)

- Suitable for documents organized in numbered sections (e.g., medical guidelines or educational materials).
- Detects numeric headings using regex, extracts the section number, title, and the associated content.
- Useful for structured Spanish medical documents like infection-related topics.

c. chunk_by_number(text)

- Handles documents with prefixed section markers **(e.g., '.1', '1.')**.
- Common in Urdu documents with formal enumeration.
- Uses a multiline regex pattern to chunk text based on these markers.

d. chunk_urdu(text)

- Custom chunker for Urdu Q&A-like text, where questions end with the Urdu question mark (؟).
- Similar logic to chunk_questions_answers, adjusted for Urdu punctuation and structure.

e. character_text_splitter (text, chunk_size, overlap)

- Fallback or generic chunker for documents without a predictable structure.

- Utilizes LangChain's CharacterTextSplitter.
- Parameters like chunk_size and overlap are fine-tuned per document type to maintain semantic integrity across chunks.
- For best values of chunk_size and overlap I used ChunkViz.

Benefits

- **Language-aware**: Custom handling for Urdu and Spanish punctuation and semantics.
- **Format-adaptive**: Works with numbered formats, Q&A, and unstructured content.
- **Extensible**: Easy to add new strategies or adjust parameters for other documents.

Limitations

- Assumes consistent format per document; may fail if formats are mixed.
- Requires manual strategy assignment per document unless automated classification is added.

Results and its interpretation

Overview

This section outlines the experimental execution and outcomes of a multilingual document retrieval system, particularly focused on Urdu and Spanish medical texts. The pipeline includes data acquisition, parsing, chunking, embedding, vector storage in Pinecone, and ultimately, retrieval of a contextually relevant response from a language model based on a user query.

1. Dataset Acquisition

The system downloaded 10 multilingual medical documents in Urdu and Spanish from reputable sources such as:

- **FDA**
- **CDC**
- **CDA (California Dental Association)**
- **Immunize.org**
- **University of Michigan**

These documents covered diverse health topics including:

- **Urdu**: Anxiety, Rabies, RSV Infection, Asthma, Heart Health
- **Spanish**: X-rays, IUDs, Blood Clots, Bad Breath, Infections

Dataset Downloading

```
commands = [
    'curl -L -o "C:/Users/Lenovo/Downloads/Awaiz_Temp/Assignment_3/dataset/urdu/rabies.pdf"
https://www.immunize.org/wp-content/uploads/vis/urdu_rabies.pdf',
    'curl -L -o "C:/Users/Lenovo/Downloads/Awaiz_Temp/Assignment_3/dataset/urdu/rsvi.pdf"
https://www.immunize.org/wp-content/uploads/vis/urdu_iis_rsv.pdf',
    'curl -L -o "C:/Users/Lenovo/Downloads/Awaiz_Temp/Assignment_3/dataset/urdu/heart.pdf"
https://www.fda.gov/media/154604/download',
    'curl -L -o "C:/Users/Lenovo/Downloads/Awaiz_Temp/Assignment_3/dataset/urdu/asthma.pdf"
https://www.cdc.gov/asthma/pdfs/AsthmaFAQ-factsheet_ur-PK_508.pdf',
    'curl -L -o "C:/Users/Lenovo/Downloads/Awaiz_Temp/Assignment_3/dataset/urdu/anxiety.pdf"
https://www.fda.gov/media/173108/download?attachment',
    'curl -L -o "C:/Users/Lenovo/Downloads/Awaiz_Temp/Assignment_3/dataset/spanish/xray.pdf"
https://www.cda.org/wp-content/uploads/xrays_spanish.pdf',
    'curl -L -o "C:/Users/Lenovo/Downloads/Awaiz_Temp/Assignment_3/dataset/spanish/iud.pdf"
https://www.reproductiveaccess.org/wp-content/uploads/2024/05/2025-01-Copper-IUD-User-Guide-
Spanish_Final.pdf',
    'curl -L -o
"C:/Users/Lenovo/Downloads/Awaiz_Temp/Assignment_3/dataset/spanish/infection.pdf"
https://www.immunize.org/wp-content/uploads/vis/spanish_ppsv.pdf',
    'curl -L -o
"C:/Users/Lenovo/Downloads/Awaiz_Temp/Assignment_3/dataset/spanish/badbreath.pdf"
https://www.cda.org/wp-content/uploads/bad_breath_spanish.pdf',
    'curl -L -o
"C:/Users/Lenovo/Downloads/Awaiz_Temp/Assignment_3/dataset/spanish/bloodclot.pdf"
https://www.med.umich.edu/1libr/VTEprevention/PreventingClotsWhileHospitalized-SPN.pdf'

]

for cmd in commands:
    subprocess.run(cmd, shell=True, check=True)
```

2. Embedding Generation

The chunked documents were embedded using a multilingual embedding model. While the exact model is not specified in the code snippet, it likely uses **sentence transformers** or **LaBSE / multilingual-e5**, suitable for cross-lingual semantic understanding.

The result was:

- A list of vector representations of document chunks (**vector_store_input**)

- A reference to the embedding model for inference compatibility in later steps (model)

3. Retrieval and Response Generation

The system combines semantic search and multilingual LLMs to generate accurate, language-consistent answers for medical queries.

Language Detection

Upon receiving a query, the system detects its language using **langdetect**. This ensures the response is generated in the same language (e.g., Urdu or Spanish), preserving clarity and user understanding.

Challenge: Managing Large Retrieved Contexts (>10 Documents)

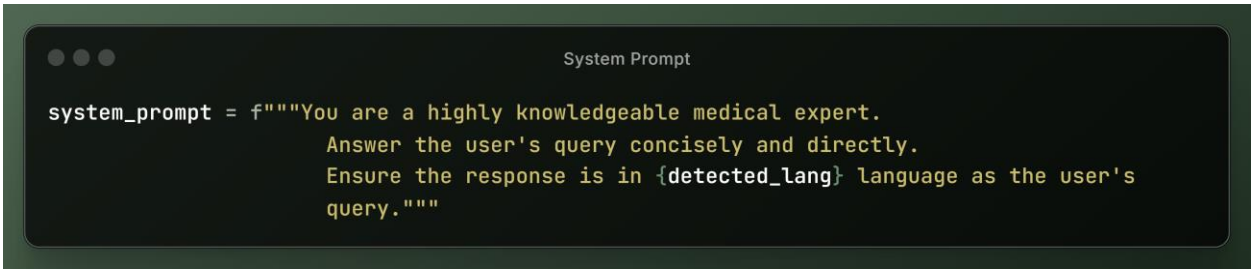
When the retrieval step yields a large number of documents (e.g., >10), selecting the most relevant and non-redundant subset becomes critical to address this, we experimented with three ranking strategies:

1. **Cosine Similarity** – Direct similarity scoring using embedding vectors.
2. **Maximal Marginal Relevance (MMR)** – Balances relevance with diversity to reduce redundancy.
3. **Cross-Encoder Re-Ranking** – Provides fine-grained re-ranking via a more expensive model.

Despite testing all three, the final pipeline used simple Cosine Similarity, as it provided a good balance of performance and speed.

```R
def get_relevant_info(query: str, index, model, namespace: str = NAMESPACE,
                      top_k: int = TOP_K):
    query_vector = model.encode(query)
    response = index.query(
        vector=query_vector.tolist(),
        top_k=top_k,
        namespace=namespace,
        include_metadata=True
    )

    retrieved_chunks = []
    for match in response['matches']:
        retrieved_chunks.append({
            "text": match['metadata']['text'],
            "source_link": match['metadata'].get('source_link', 'N/A')
        })

    doc_texts = [chunk["text"] for chunk in retrieved_chunks]
    if not doc_texts:
        return []

    doc_vectors = model.encode(doc_texts)

    # Cosine similarity and top 3 selection
    sims = cosine_similarity([query_vector], doc_vectors).flatten()
    top_indices = np.argsort(sims)[-3:][::-1]

    top_chunks = [retrieved_chunks[i] for i in top_indices]
    return top_chunks
```

System Prompt



Response Generation with GROQ

The retrieved chunks are combined with the original query and sent to a Groq-hosted LLM (meta-llama/llama-4-scout-17b-16e-instruct). The system instructs the model to act as a medical expert and respond in the detected language, ensuring contextual accuracy and appropriate tone.

Caching Mechanism

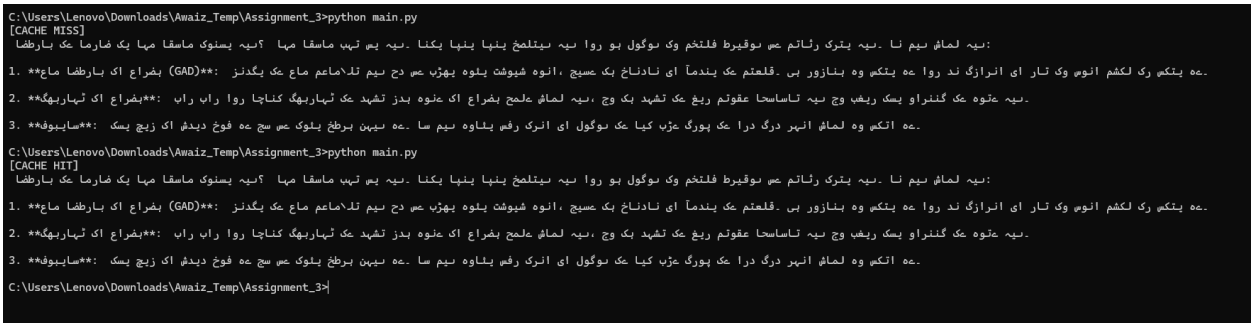
Responses are cached locally in a JSON file to avoid redundant computation. If a query has been previously processed, the system serves the cached response instantly.

Example Workflow

For the query:
"اضطراب کے امراض کی اہم اقسام کونسی ہیں؟"
the system:

- Detects it as Urdu,
- Retrieves anxiety-related content from the Urdu documents,

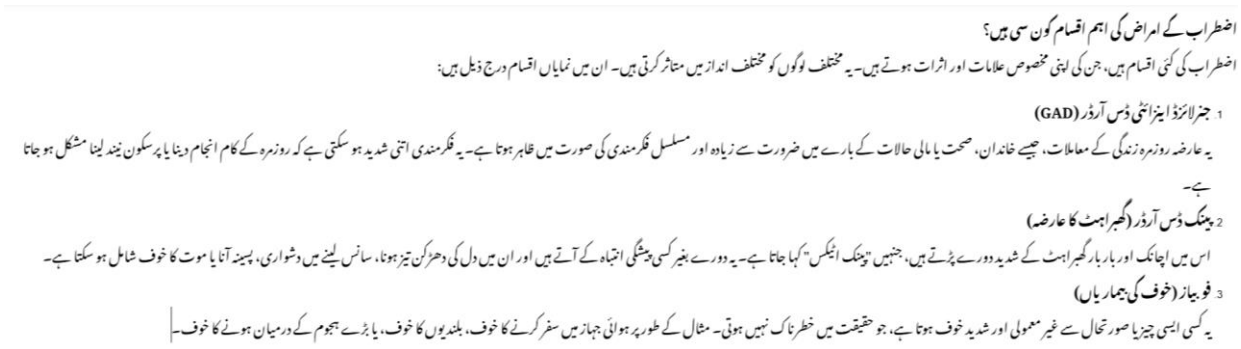
- Generates an Urdu response via the LLM,
- Caches the result for reuse.



Answer Quality Check

To compare the quality of text generated from our RAG system compare the original and retrieved text

Retrieved text



اضطراب کے امراض کی اہم اقسام کون سی ہیں؟

اضطراب کی کئی اقسام ہیں، جن کی اپنی مخصوص علامات اور اثرات ہوتے ہیں۔ یہ مختلف لوگوں کو مختلف انداز میں متاثر کرتی ہیں۔ ان میں نمایاں اقسام درج ذیل ہیں:

1 جنرلائزڈ اینزائٹی ڈس آرڈر (GAD)

یہ عارضہ روزمرہ زندگی کے معاملات، جیسے خاندان، صحت یا مالی حالات کے بارے میں ضرورت سے زیادہ اور مسلسل فکرمندی کی صورت میں ظاہر ہوتا ہے۔ یہ فکرمندی اتنی شدید ہو سکتی ہے کہ روزمرہ کے کام انجام دینا یا پرسکون نیند لینا مشکل ہو جاتا ہے۔

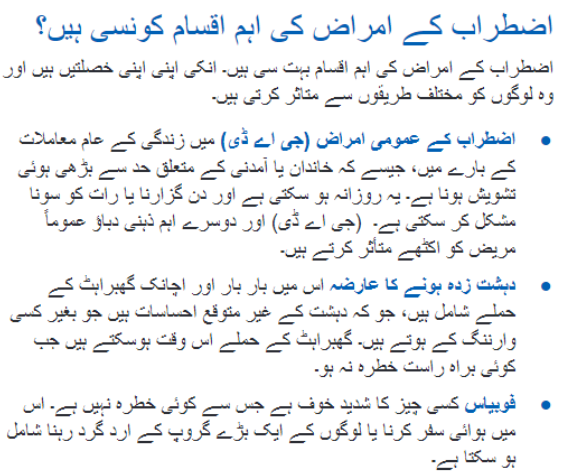
2 پینک ڈس آرڈر (گھبراہٹ کا عارضہ)

اس میں اچانک اور بار بار گھبراہٹ کے شدید دورے پڑتے ہیں، جنہیں "پینک اٹیکس" کہا جاتا ہے۔ یہ دورے بغیر کسی واضح خطرے کے آتے ہیں اور ان میں دل کی دھڑکن تیز ہونا، سانس لینے میں دشواری، پسینہ آنا یا موت کا خوف شامل ہو سکتا ہے۔

3 فوبیاز (خوف کی بیماریاں)

یہ کسی ایسی چیز یا صورتحال سے غیر معمولی اور شدید خوف ہوتا ہے، جو حقیقت میں خطرناک نہیں ہوتی۔ مثال کے طور پر ہوائی جہاز میں سفر کرنے کا خوف، بلندیوں کا خوف، یا بڑے ہجوم کے درمیان ہونے کا خوف۔

Original text from pdf



اضطراب کے امراض کی اہم اقسام کونسی ہیں؟

اضطراب کے امراض کی اہم اقسام بہت سی ہیں۔ انکی اپنی اپنی خصلتیں ہیں اور وہ لوگوں کو مختلف طریقوں سے متاثر کرتی ہیں۔

- **اضطراب کے عمومی امراض (جی اے ڈی)** میں زندگی کے عام معاملات کے بارے میں، جیسے کہ خاندان یا آمدنی کے متعلق حد سے بڑھی ہوئی تشویش ہونا ہے۔ یہ روزانہ ہو سکتی ہے اور دن گزارنا یا رات کو سونا مشکل کر سکتی ہے۔ (جی اے ڈی) اور دوسرے اہم ذہنی دباؤ عموماً مریض کو اکٹھے متأثر کرتے ہیں۔
- **دہشت زدہ ہونے کا عارضہ** اس میں بار بار اور اچانک گھبراہٹ کے حملے شامل ہیں، جو کہ دہشت کے غیر متوقع احساسات ہیں جو بغیر کسی وارننگ کے ہوتے ہیں۔ گھبراہٹ کے حملے اس وقت ہوسکتے ہیں جب کوئی براہ راست خطرہ نہ ہو۔
- **فوبیاس** کسی چیز کا شدید خوف ہے جس سے کوئی خطرہ نہیں ہے۔ اس میں ہوائی سفر کرنا یا لوگوں کے ایک بڑے گروپ کے ارد گرد رہنا شامل ہو سکتا ہے۔