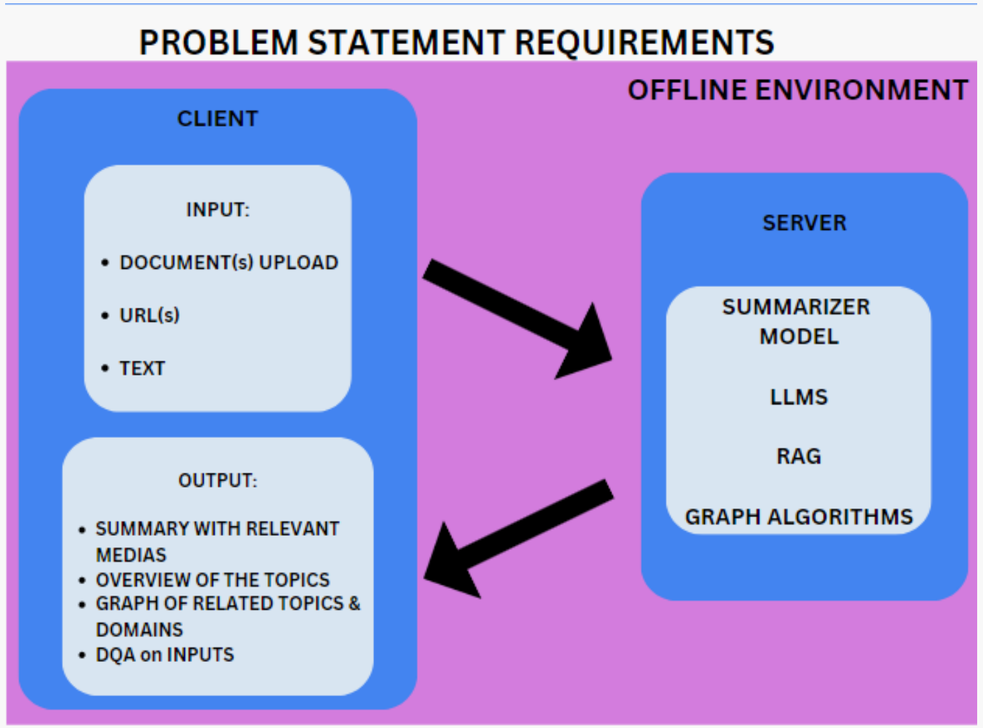
SMART INDIA HACKATHON 2023



Organization: NTRO

PS Code: 1450
Problem Statement Title: Develop and deploy a Large Language Model (LLM) based tool for generating human like responses to natural language inputs for network not connected over internet

Team Name: **Aarohan915**

Team Leader Name**: Samiul Sheikh**

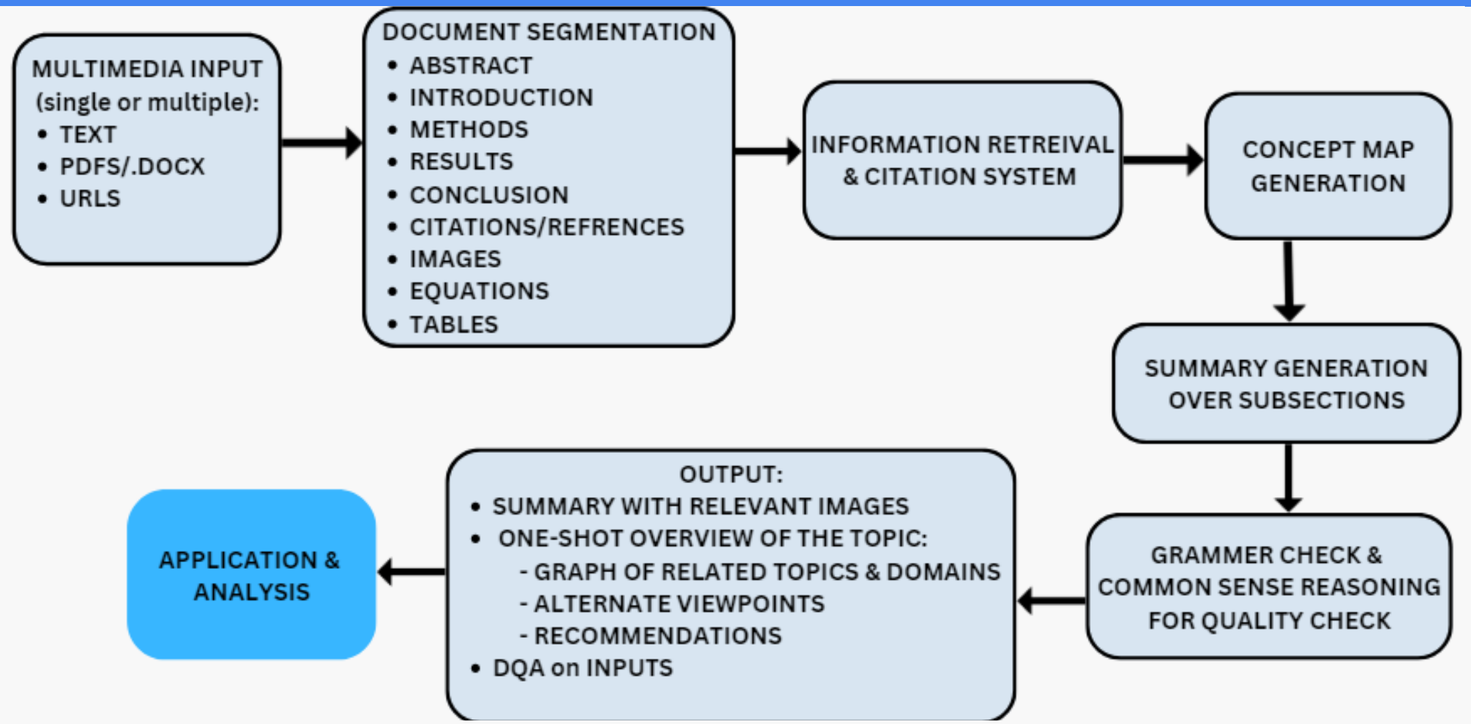
Institute Code (AISHE): **C-33641**

Institute Name: **V.J.T.I. Mumbai**

PROBLEM STATEMENT DETAILS

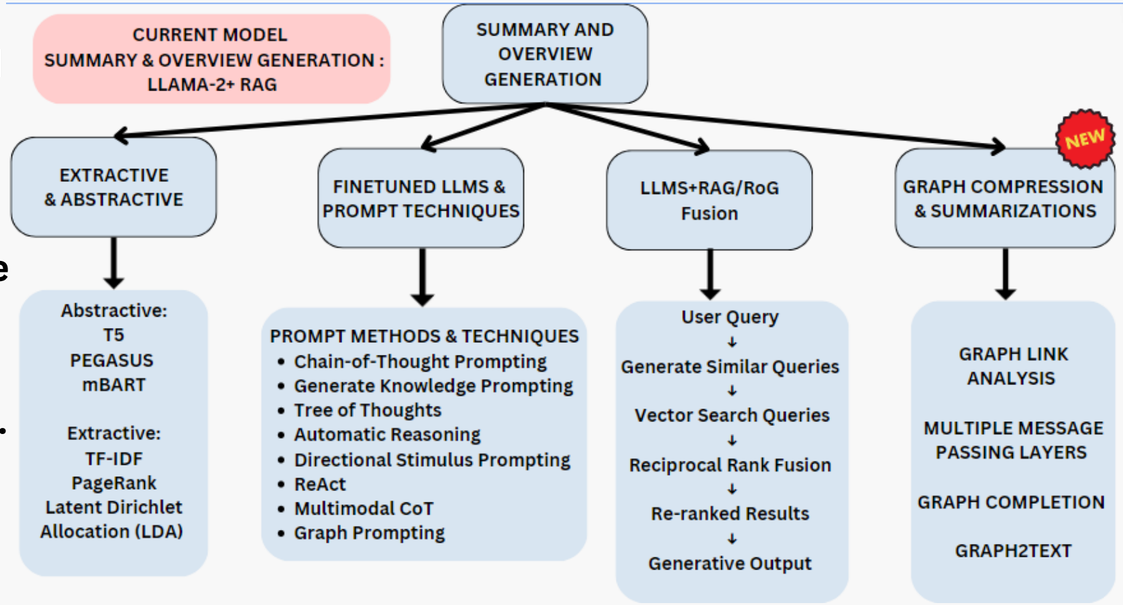
Dataset

- We have a vast network of information using Microsoft Knowledge Graph and Unarxiv Dataset.
- We have a corpus of 1.5M+ Research papers across 8 categories and 150+ subcategories. Each paper is divided into subsections.
- Each paper is present in .json format. We have over 20+GBs of data present that we are converting into a Knowledge Graph using NebulaGraph



Document Segmentation & Multimedia Information Retrieval:

- The input could be URLs, PDFs, or simple text. These may be multimedia and multilingual.
- Through OCR and image captioning, we can generate a description of the image.
- Through transcript generations and frame-by-frame image segmentation, we can get textual representation. For the Multilingual approach, we employ Neural Machine Translation to convert to English for processing
- We include equations and table information with the help of Document Q&A.
- Using the references we will develop a Citation Network and develop a Knowledge Graph with a Concept Graph for each subsection of a research paper. This is done with the help of Retrieval Augmented Generation(RAG)



SUMMARIZATION

Knowledge Graph Summarization

- Each subsection is converted into a concept map using GNN methods like Retrieval Augmented Generation(RAG). This would make summarization and overview generation easier.
- The incoming S&T paper is put through a clustering method to find categories and subcategories for paper recommendation and citation network
- Now we employ Retrieval Augmented Generation over LLMs for summary generation over each subsection.
- The method of subsection summarization over multiple documents gives a structured overview from basics to advanced of all the novel methods and techniques in that particular research area based on category and subcategory
- An alternate method of using PageRank and Centrality measures on the graph embeddings after cluster generation of concepts.Overview can also be generated with the help of Latent Direchlet Allocation(LDA)

NLP Summarization

- Using open-sourced finetuned LLMs we were able to generate summaries over multiple documents and present them in a structured format of subcategories.
- This was done by employing embedding, clustering, and text-generation methods

MODELS COMPUTATIONAL REQUIREMENTS & INFERENCE TIME

TASKS	model used	memory capacity:	parameters	inference speed (on CPU):
DOCUMENT SEGMENTATION	docsegtr	1GB	-	20s
OCR	pytersseract	13.6 kB	-	3s
IMAGE CAPTIONING	Salesforce/blip-image-captioning-large	1.88GB	-	10s
TRANSCRIPT GENERATION	openai/whisper-large-v2	~6GBs	1550 M	54s-270s
SUMMARY GENERATION	meta-llama/Llama-2-7b-chat-hf	~4GB	70B	5 tokens/min
OVERVIEW GENERATION	meta-llama/Llama-2-7b-chat-hf	~4GB	70B	5 tokens/min
GRAMMER CHECK	meta-llama/Llama-2-7b-chat-hf	~4GB	70B	5 tokens/min
COMMON SENSE REASONING	meta-llama/Llama-2-7b-chat-hf	~4GB	70B	5 tokens/min
DOCUMENT QUESTION AND ANSWERING	meta-llama/Llama-2-7b-chat-hf	~4GB	70B	5 tokens/min

Concurrency Time: 1 min for upto 100 Requests

APPLICATIONS:

- DOWNLINK SATELLITE DATA ANALYSIS
- RESEARCH PAPER CODE GENERATION
- ADVANCED FAKE NEWS DETECTION SYSTEM
- IDENTIFYING & EXPLORING RESEARCH AREAS

MODELS:

- RAG: NebulaGraph, Text2Cypher
- Summarizer: allenai/led-large-16384-arxiv
- clustering or semantic search:

- sentence-transformers/paraphrase-MiniLM-L6-v2
-sentence-transformers/distilbert-base-nli-mean-token

Team Member Details

Team Leader: Samiul Sheikh
B.Tech, Electronics Engineering, Final Year, VJTI

Team Member 1: **Mihir Sahasrabudhe**
B.Tech, Electronics Engineering, Final Year, VJTI

Team Member 2: **Sheena Dmello**
B.Tech, Electronics Engineering, Final Year, VJTI

Team Member 3: **Pranav Janjani**
B.Tech, Computer Science, Pre-Final Year, VJTI

Team Member 4: **Kanak Meshram**
B.Tech, Electronics Engineering, Final Year, VJTI

Team Member 5: **Yash Deshpande**
B.Tech, Electronics Engineering, Final Year, VJTI

Team Mentor 1: **Dr. Faruk Kazi**
Category: Academic; Expertise: Cybersecurity; Domain Experience: 25 Years

Team Mentor 2: **Aum Patil**
Category: Industry; Expertise: AI; Domain Experience: 3 Years