Develop and deploy a Large Language Model (LLM) based tool for generating human-like responses to natural language inputs for networks not connected over the internet

MEMBERS:

KANAK MESHRAM

MIHIR SAHASRABUDHE

PRANAV JANJANI

SAMIUL SHEIKH

SHEENA D’MELLO

YASH DESHPANDE

- **Problem Statement ID: 1450**
- **Organization: National Technical Research Organisation,(NTRO)**
- **Category: Software**
- **Domain Bucket: Smart Automation**

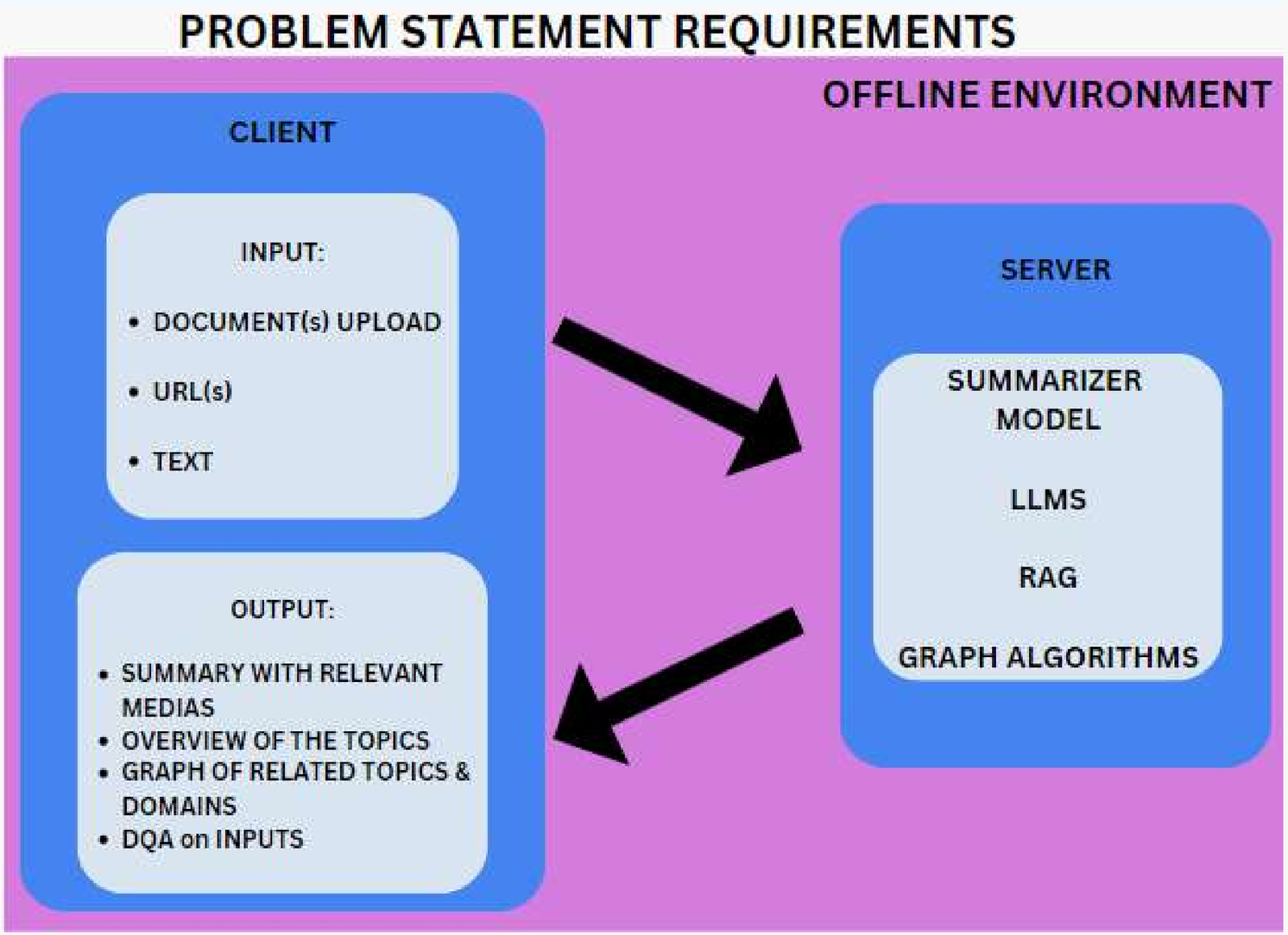
PROBLEM STATEMENT DETAILS

Document & Multimedia Information Retrieval:

- The input could be URLs, PDFs, or simple text. These may be multimedia and multilingual.
- For the extraction of images, text, and videos, we use Document segmentation.
- Through OCR and image captioning, we are able to generate a description of the image.
- Through transcript generations and frame-by-frame image segmentation, we are able to get textual representation. For the Multilingual approach, we employ Neural Machine Translation to convert to English for processing

Dataset

- We have a vast network of information using Microsoft Knowledge Graph and arxiv 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



TechStack:

- Web: HTML,CSS,JavaScript,TailWindCSS,ReactJS,NodeJS
- ML:NeblaGraph,LangChain, Pytorch,PytorchGeometric,Text2Cypher

CURRENT PROBLEMS IN SUMMARIZATION

REQUIREMENT

the automatic overview should show:

- no paper-unrelated "information noise" from the respective documents
- no dangling references to what is not mentioned or explained in the overview
- no text breaks across a sentence
- no semantic redundancy.

OPTIMAL SYSTEM

- shortens the source texts
- presents information organized around the key aspects to represent diverse views.
- produces an overview of a given topic.

QUALITY PARAMETERS

- clear structurean outline of the main content
- text within sections is divided into meaningful paragraphs
- gradual transition from more general to more specific thematic aspects
- good readability having diverse viewpoints clearly presented

SOLUTION STRUCTURE

MULTIMEDIA INPUT (single or multiple):
- TEXT
- PDFS/.DOCX
- URLS

DOCUMENT SEGMENTATION
- ABSTRACT
- INTRODUCTION
- METHODS
- RESULTS
- CONCLUSION
- CITATIONS/REFRENCES
- IMAGES
- EQUATIONS
- TABLES

INFORMATION RETREIVAL & CITATION SYSTEM

CONCEPT MAP GENERATION

SUMMARY GENERATION OVER SUBSECTIONS

GRAMMER CHECK & COMMON SENSE REASONING FOR QUALITY CHECK

OUTPUT:
- SUMMARY WITH RELEVANT IMAGES
- ONE-SHOT OVERVIEW OF THE TOPIC:
  - GRAPH OF RELATED TOPICS & DOMAINS
  - ALTERNATE VIEWPOINTS
  - RECOMMENDATIONS
- DQA on INPUTS

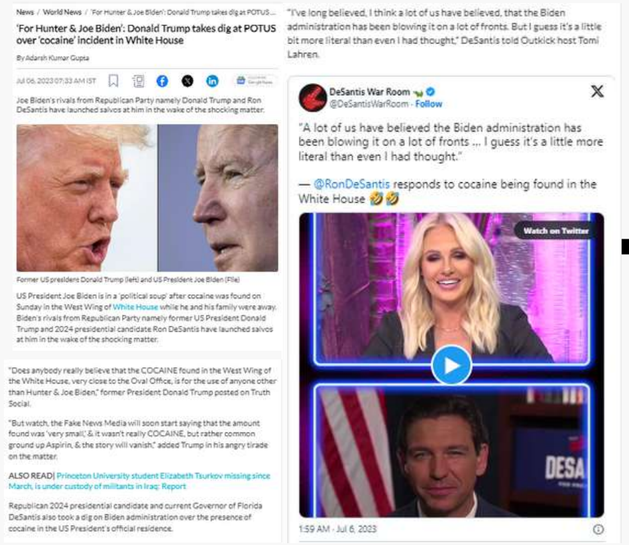
APPLICATION & ANALYSIS

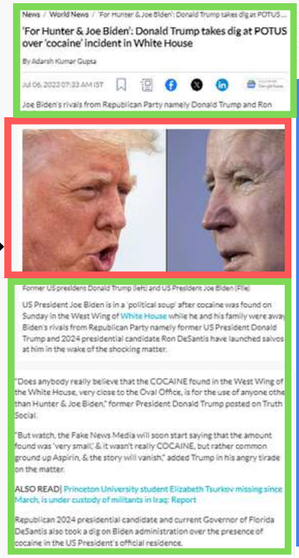
INFORMATION EXTRACTION AND CONVERSION TO TEXTUAL REPRESENTATION

NEWS ARTICLES

SPLITING DOCUMENT INTO MULTIMEDIA COMPONENTS

EXTRACTING INFORMATION FROM SEGMENTED DOCUMENT





TEXT

IMAGES

VIDEOS

OCR

IMAGE CAPTIONING

SCENCE GRAPH MAPPING

TRANSCRIPT GENERATION

FRAME-BY-FRAME IMAGES

TRANSLATION SYSTEM (IF MULTILINGUAL)

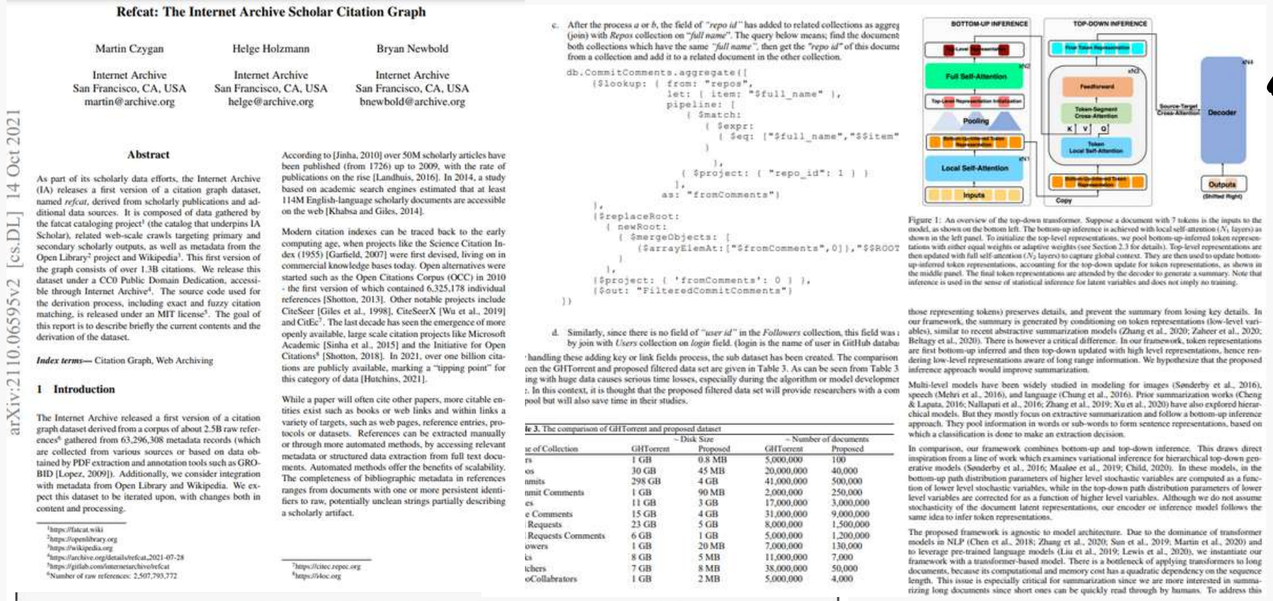
TEXTUAL REPRESENTATION OF NEWS ARTICLE

CLOSED CAPTIONS EXTRACTION/ TRANSCRIPT GENERATION SYSTEM FOR VIDEOS & AUDIO FILES

- **USAGE OF TRANSCRIPT GENERATION MODELS, USING whisper-large MODEL HAS HIGHER WORD ERROR RATE(WER) FOR A PARTICULAR LANGUAGE**
- **SO WE USED FINE-TUNED MODELS FOR GENERATING TRANSCRIPT OF THE VIDEOS AND AUDIO FILES FOR CONVERSION OF AUDIO TO TEXT FOR FURTHER PROCESSING AND DOWNSTREAM APPLICATIONS**

Language	Audio to Text (Transcript Generation)	**Word Error Rate (WER)**
Punjabi	DrishtiSharma/whisper-large-v2-punjabi-700-steps	24.476
Hindi	vasista22/whisper-hindi-large-v2	6.800
Gujarati	vasista22/whisper-gujarati-medium	12.330
Marathi	DrishtiSharma/whisper-large-v2-marathi	13.6440
Telugu	vasista22/whisper-telugu-large-v2	9.650
Kannada	vasista22/whisper-kannada-medium	7.650
Malayalam	DrishtiSharma/whisper-large-v2-malayalam	27.458
Tamil	vasista22/whisper-tamil-large-v2	7.500
Odia	Ranjit/Whisper_Small_Odia_10k_steps	19.772
Bengali	ucalyptus/whisper-large-v2-bengali-100steps	29.299
Assamese	DrishtiSharma/whisper-large-v2-assamese-5k-steps	20.917

SCIENCE & TECHNOLOGY DOCUMENTS INFORMATION EXTRACTION



Refcat: The Internet Archive Scholar Citation Graph

Martin Czygan — Internet Archive, San Francisco, CA, USA, martin@archive.org
Helge Holzmann — Internet Archive, San Francisco, CA, USA, helge@archive.org
Bryan Newbold — Internet Archive, San Francisco, CA, USA, bnewbold@archive.org

Abstract

As part of its scholarly data efforts, the Internet Archive (IA) releases a first version of a citation graph dataset, named *refcat*, derived from scholarly publications and additional data sources. It is composed of data gathered by the fatcat cataloging project[1] (the catalog that underpins IA Scholar), related web-scale crawls targeting primary and secondary scholarly outputs, as well as metadata from the Open Library[2] project and Wikipedia[3]. This first version of the graph consists of over 1.3B citations. We release this dataset under a CC0 Public Domain Dedication, accessible through Internet Archive[4]. The source code used for the derivation process, including exact and fuzzy citation matching, is released under an MIT license[5]. The goal of this report is to describe briefly the current contents and the derivation of the dataset.

***Index terms*—** Citation Graph, Web Archiving

1 Introduction

The Internet Archive released a first version of a citation graph dataset derived from a corpus of about 2.5B raw references[6] gathered from 63,296,308 metadata records (which are collected from various sources or based on data obtained by PDF extraction and annotation tools such as GROBID [Lopez, 2009]). Additionally, we consider integration with metadata from Open Library and Wikipedia. We expect this dataset to be iterated upon, with changes both in content and processing.

According to [Jinha, 2010] over 50M scholarly articles have been published (from 1726) up to 2009, with the rate of publications on the rise [Landhuis, 2016]. In 2014, a study based on academic search engines estimated that at least 114M English-language scholarly documents are accessible on the web [Khabsa and Giles, 2014].

Modern citation indexes can be traced back to the early computing age, when projects like the Science Citation Index (1955) [Garfield, 2007] were first devised, living on in commercial knowledge bases today. Open alternatives were started such as the Open Citations Corpus (OCC) in 2010 - the first version of which contained 6,325,178 individual references [Shotton, 2013]. Other notable projects include CiteSeer [Giles et al., 1998], CiteSeerX [Wu et al., 2019] and CitEc[7]. The last decade has seen the emergence of more openly available, large scale citation projects like Microsoft Academic [Sinha et al., 2015] and the Initiative for Open Citations[8] [Shotton, 2018]. In 2021, over one billion citations are publicly available, marking a "tipping point" for this category of data [Hutchins, 2021].

While a paper will often cite other papers, more citable entities exist such as books or web links and within links a variety of targets, such as web pages, reference entries, protocols or datasets. References can be extracted manually or through more automated methods, by accessing relevant metadata or structured data extraction from full text documents. Automated methods offer the benefits of scalability. The completeness of bibliographic metadata in references ranges from documents with one or more persistent identifiers to raw, potentially unclean strings partially describing a scholarly artifact.

[1] https://fatcat.wiki
[2] https://openlibrary.org
[3] https://wikipedia.org
[4] https://archive.org/details/refcat_2021-07-28
[5] https://gitlab.com/internetarchive/refcat
[6] Number of raw references: 2,507,793,772
[7] https://citec.repec.org
[8] https://i4oc.org

c. After the process *a* or *b*, the field of *"repo id"* has added to related collections as aggreg (join) with *Repos* collection on *"full name"*. The query below means; find the document both collections which have the same *"full name"*, then get the *"repo id"* of this docume from a collection and add it to a related document in the other collection.

```
db.CommitComments.aggregate([
    {$lookup: { from: "repos",
                let: { item: "$full_name" },
                pipeline: [
                    { $match:
                        { $expr:
                            { $eq: ["$full_name","$$item"
                        }
                    },
                ],
              as: "fromComments"}
    },
    {$replaceRoot:
      { newRoot:
        { $mergeObjects: [
            {$arrayElemAt:["$fromComments",0]},"$$ROOT
        ]
      },
    },
    {$project: { 'fromComments': 0 } },
    {$out: "FilteredCommitComments"}
])
```

d. Similarly, since there is no field of *"user id"* in the *Followers* collection, this field was ; by join with *Users* collection on *login* field. (login is the name of user in GitHub databa

handling these adding key or link fields process, the sub dataset has been created. The comparison een the GHTorrent and proposed filtered data set are given in Table 3. As can be seen from Table 3 ing with huge data causes serious time losses, especially during the algorithm or model developmer ; In this context, it is thought that the proposed filtered data set will provide researchers with a com pool but will also save time in their studies.

le 3. The comparison of GHTorrent and proposed dataset

	~ Disk Size		~ Number of documents	
ıe of Collection	GHTorrent	Proposed	GHTorrent	Proposed
rs	1 GB	0.8 MB	5,000,000	100
ers	30 GB	45 MB	20,000,000	40,000
mmits	298 GB	4 GB	41,000,000	500,000
mmit Comments	1 GB	90 MB	2,000,000	250,000
es	11 GB	3 GB	17,000,000	3,000,000
e Comments	15 GB	4 GB	31,000,000	9,000,000
Requests	23 GB	5 GB	8,000,000	1,500,000
Requests Comments	6 GB	1 GB	5,000,000	1,200,000
owers	1 GB	20 MB	7,000,000	130,000
ers	8 GB	5 MB	11,000,000	7,000
ichers	7 GB	8 MB	38,000,000	50,000
oCollabrators	1 GB	2 MB	5,000,000	4,000

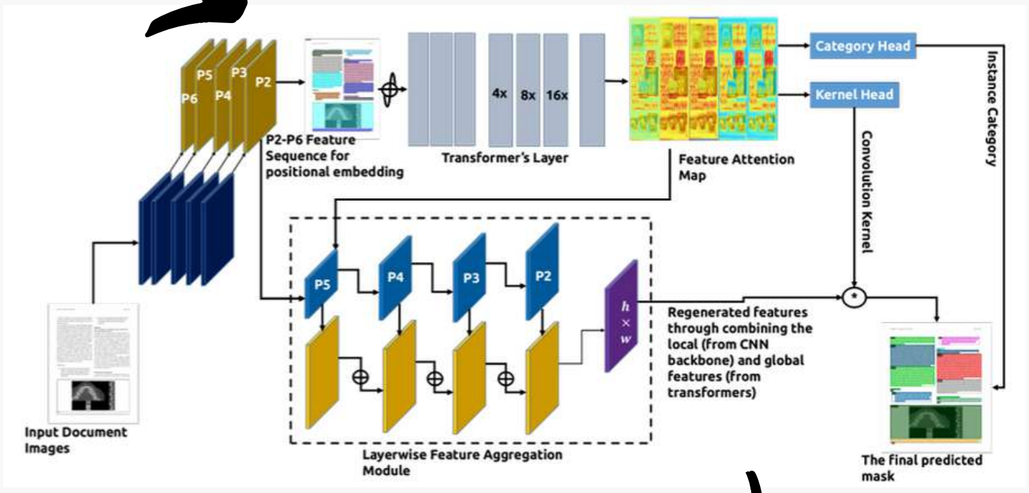
Figure 1: An overview of the top-down transformer. Suppose a document with 7 tokens is the inputs to the model, as shown on the bottom left. The bottom-up inference is achieved with local self-attention (N_1 layers) as shown in the left panel. To initialize the top-level representations, we pool bottom-up-inferred token representations with either equal weights or adaptive weights (see Section 2.3 for details). Top-level representations are then updated with full self-attention (N_2 layers) to capture global context. They are then used to update bottom-up-inferred token representations, accounting for the top-down update for token representations, as shown in the middle panel. The final token representations are attended by the decoder to generate a summary. Note that inference is used in the sense of statistical inference for latent variables and does not imply no training.

those representing tokens) preserves details, and prevent the summary from losing key details. In our framework, the summary is generated by conditioning on token representations (low-level variables), similar to recent abstractive summarization models (Zhang et al., 2020; Zaheer et al., 2020; Beltagy et al., 2020). There is however a critical difference. In our framework, token representations are first bottom-up inferred and then top-down updated with high level representations, hence rendering low-level representations aware of long range information. We hypothesize that the proposed inference approach would improve summarization.

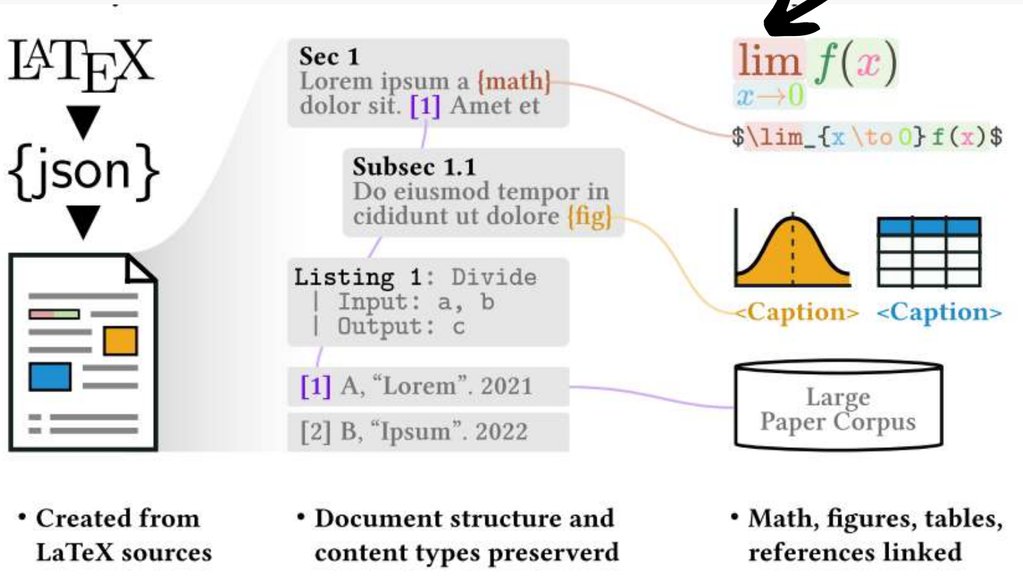
Multi-level models have been widely studied in modeling for images (Sønderby et al., 2016), speech (Mehri et al., 2016), and language (Chung et al., 2016). Prior summarization works (Cheng & Lapata, 2016; Nallapati et al., 2016; Zhang et al., 2019; Xu et al., 2020) have also explored hierarchical models. But they mostly focus on extractive summarization and follow a bottom-up inference approach. They pool information in words or sub-words to form sentence representations, based on which a classification is done to make an extraction decision.

In comparison, our framework combines bottom-up and top-down inference. This draws direct inspiration from a line of work which examines variational inference for hierarchical top-down generative models (Sønderby et al., 2016; Maaløe et al., 2019; Child, 2020). In these models, in the bottom-up path distribution parameters of higher level stochastic variables are computed as a function of lower level stochastic variables, while in the top-down path distribution parameters of lower level variables are corrected for as a function of higher level variables. Although we do not assume stochasticity of the document latent representations, our encoder or inference model follows the same idea to infer token representations.

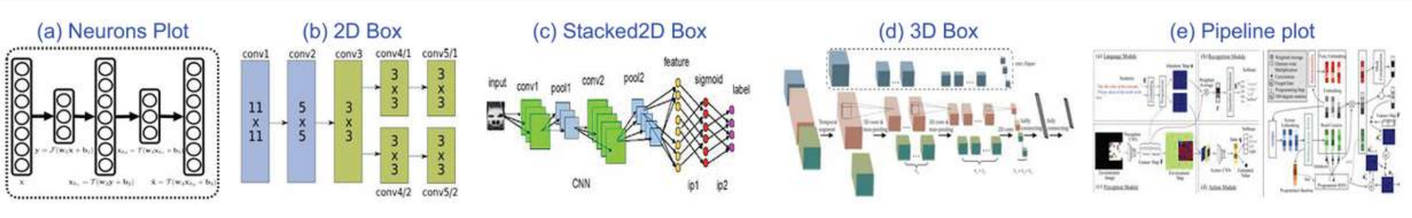
The proposed framework is agnostic to model architecture. Due to the dominance of transformer models in NLP (Chen et al., 2018; Zhang et al., 2020; Sun et al., 2019; Martin et al., 2020) and to leverage pre-trained language models (Liu et al., 2019; Lewis et al., 2020), we instantiate our framework with a transformer-based model. There is a bottleneck of applying transformers to long documents, because its computational and memory cost has a quadratic dependency on the sequence length. This issue is especially critical for summarization since we are more interested in summarizing long documents since short ones can be quickly read through by humans. To address this




```
/* - - - - - - - example paper (arXiv:2105.05862) - - - - - - - */
{ "paper_id": "2105.05862",
  "metadata": {...},
  "abstract": {...},
  "body_text": [...],
  "ref_entries": {...},
  "bib_entries": {...} }
/* - - - - - - - one of the sections in body_text - - - - - - - */
{ "section": "Memory wave form",
  "sec_number": "2.1",
  "sec_type": "subsection",
  "content_type": "paragraph",
  "text": "The gauge choice leading us to this solution does not fix
           completely all the gauge freedom and an additional constraint
           should be imposed to leave only the physical degrees of freedom.
           This is done by projecting the source tensor {{formula:7fd88bcd-
           9013-433d-9756-b874472530d9}} into its transverse-traceless (TT)
           components (see for example {{cite:80dbb6c8b9c12f561a8e585faceac5f
           4e104d60d}}). Doing this and without loss of generality, we will
           use the following very well known ansatz for the source term
           proposed in {{cite:bc9a8ca19785627a087ae0c01abe155c22388e16}}\n" }
/* - - - - - - - ref_entries entry for {{formula:7fd88...}} - - - - - - - */
{ "latex": "S_{\\mu \\nu }",
  "type": "formula" }
/* - - - - - - - bib_entries entry for {{cite:80dbb...}} - - - - - - - */
{ "bib_entry_raw": "R. Epstein, The Generation of Gravitational Radiation by Esc
                    aping Supernova Neutrinos, Astrophys. J. 223 (1978) 1037.",
  "contained_links": [
    { "url": "https://doi.org/10.1086/156337",
      "text": "Astrophys. J. 223 (1978) 1037.",
      "start": 87,
      "end": 117 }
  ],
  "ids": {...} }
```

TABLES & IMAGES



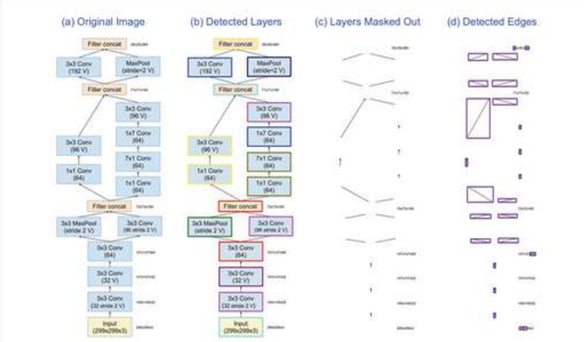
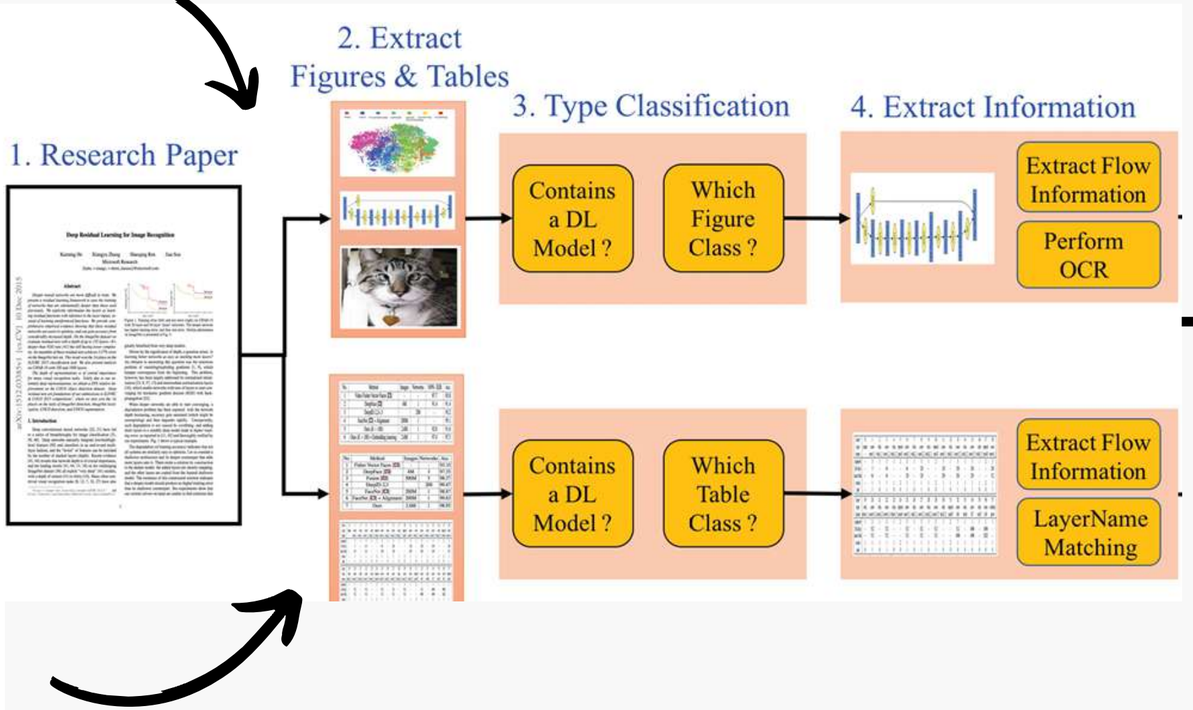
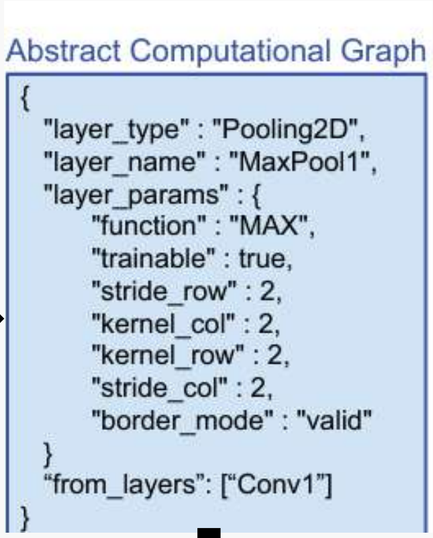


Figure 3: Illustration of the proposed flow detection approach from complex figures (Szegedy et al. 2017) (AAAI 2017) involving (b) node/ layer detection, and (d) edge detection.

layer type	0 input	1 conv	2 relu	3 conv	4 relu	5 mpool	6 conv	7 relu	8 conv	9 relu	10 mpool	11 conv	12 relu	13 conv	14 relu	15 conv	16 relu	17 mpool	18 conv
name	–	conv1_1	relu1_1	conv1_2	relu1_2	pool1	conv2_1	relu2_1	conv2_2	relu2_2	pool2	conv3_1	relu3_1	conv3_2	relu3_2	conv3_3	relu3_3	pool3	conv4_1
support	–	3	1	3	1	2	3	1	3	1	2	3	1	3	1	3	1	2	3
filt dim	–	3	–	64	–	–	64	–	128	–	–	128	–	256	–	256	–	–	256
num filts	–	64	–	64	–	–	128	–	128	–	–	256	–	256	–	256	–	–	512
stride	–	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
pad	–	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1

layer type	19 relu	20 conv	21 relu	22 conv	23 relu	24 mpool	25 conv	26 relu	27 conv	28 relu	29 conv	30 relu	31 mpool	32 conv	33 relu	34 conv	35 relu	36 conv	37 softmx
name	relu4_1	conv4_2	relu4_2	conv4_3	relu4_3	pool4	conv5_1	relu5_1	conv5_2	relu5_2	conv5_3	relu5_3	pool5	fc6	relu6	fc7	relu7	fc8	prob
support	1	3	1	3	1	2	3	1	3	1	3	1	2	7	1	1	1	1	1
filt dim	–	512	–	512	–	–	512	–	512	–	512	–	–	512	–	4096	–	4096	–
num filts	–	512	–	512	–	–	512	–	512	–	512	–	–	4096	–	4096	–	2622	–
stride	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1
pad	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0

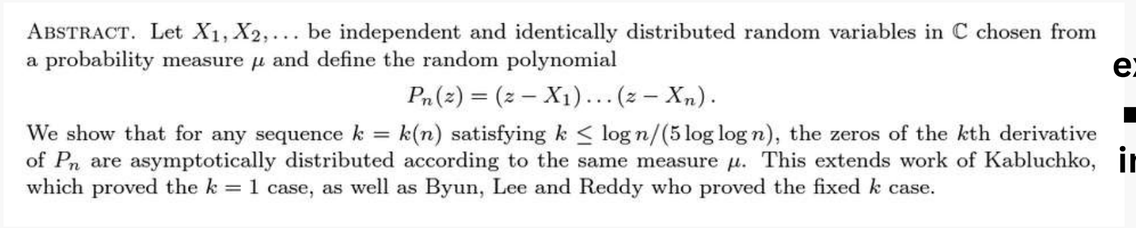




TEXTUAL DESCRIPTION

CONVERSION OF EQUATIONS INTO THEIR TEXTUAL REPRESENTATION

PERFORMING ITERATIVE CONTEXT-BASED DOCUMENT QUESTION ANSWERING



ABSTRACT. Let $X_1, X_2, \dots$ be independent and identically distributed random variables in $\mathbb{C}$ chosen from a probability measure μ and define the random polynomial

$$P_n(z) = (z - X_1) \dots (z - X_n).$$

We show that for any sequence $k = k(n)$ satisfying $k \leq \log n/(5 \log \log n)$, the zeros of the kth derivative of P_n are asymptotically distributed according to the same measure μ. This extends work of Kabluchko, which proved the $k = 1$ case, as well as Byun, Lee and Reddy who proved the fixed k case.

extracted data

in .json format


```
"abstract": {"section": "Abstract", "text": "  Let $X_1,X_2,\\ldots$ be
independent and identically distributed random\nvariables in
$\\mathbb{C}$ chosen from a probability measure $\\mu$ and define\nthe
random polynomial $$ P_n(z)=(z-X_1)\\ldots(z-X_n)\\,. $$ We show that for
any\nsequence $k = k(n)$ satisfying $k \\leq \\log n / (5 \\log\\log n)$,
the zeros of\nthe $k$th derivative of $P_n$ are asymptotically
distributed according to the\nsame measure $\\mu$. This extends work of
Kabluchko, which proved the $k = 1$\ncase, as well as Byun, Lee and Reddy
who proved the fixed $k$ case.\n"}
```


SUBSECTION CONTEXT

EQUATION VARIABLES & RELATIONS MAPPING

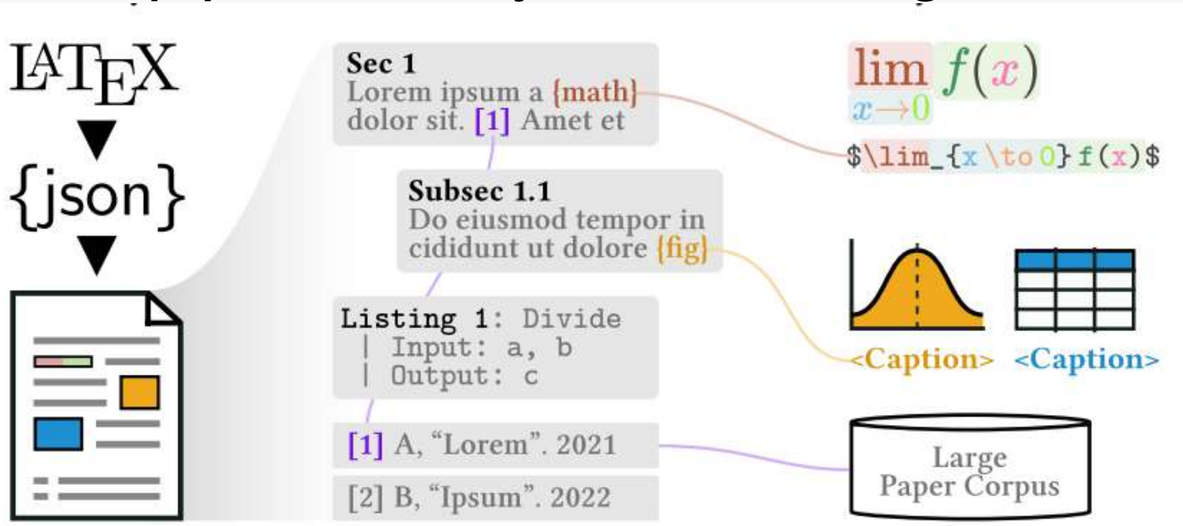
DOCUMENT QUESTION ANSWERING

TEXTUAL REPRESENTATION

CONVERSION OF TEXTUAL REPRESENTATION TO GRAPHICAL REPRESENTATION

DATASET:

- **Corpus of 1.5M+ Arxiv Papers over 8 categories(physics,math etc.) and 151 subcategories(astrophysics,convex optimization etc.)**
- **Each paper in form of .json format of a segmented document across subsections**



- **Created from LaTeX sources**
- **Document structure and content types preserverd**
- **Math, figures, tables, references linked**

	citing documents	references		cited documents
		outgoing	incoming	
full data set:	**1,043,126**	15,954,664	15,954,664	**2,746,288**
full text	1,043,126	15,954,664	7,181,576	736,597
linked to MAG	994,351	15,846,351	15,954,664	2,746,288
by discipline:				
physics	662,894	9,300,576	7,827,072	921,852
mathematics	237,422	3,426,117	5,062,033	906,301
computer science	111,694	2,526,656	1,876,401	425,860
other	31,116	701,315	1,189,158	492,275

```
/* - - - - - - - - example paper (arXiv:2105.05862) - - - - - - - - */
{ "paper_id": "2105.05862",
  "metadata": {...},
  "abstract": {...},
  "body_text": [...],
  "ref_entries": {...},
  "bib_entries": {...} }
/* - - - - - - - - one of the sections in body_text - - - - - - - - */
{ "section": "Memory wave form",
  "sec_number": "2.1",
  "sec_type": "subsection",
  "content_type": "paragraph",
  "text": "The gauge choice leading us to this solution does not fix
           completely all the gauge freedom and an additional constraint
           should be imposed to leave only the physical degrees of freedom.
           This is done by projecting the source tensor {{formula:7fd88bcd-
           9013-433d-9756-b874472530d9}} into its transverse-traceless (TT)
           components (see for example {{cite:80dbb6c8b9c12f561a8e585faceac5f
           4e104d60d}}). Doing this and without loss of generality, we will
           use the following very well known ansatz for the source term
           proposed in {{cite:bc9a8ca19785627a087ae0c01abe155c22388e16}}\n" }
/* - - - - - - - - ref_entries entry for {{formula:7fd88...}} - - - - - - - - */
{ "latex": "S_{\\mu \\nu }",
  "type": "formula" }
/* - - - - - - - - bib_entries entry for {{cite:80dbb...}} - - - - - - - - */
{ "bib_entry_raw": "R. Epstein, The Generation of Gravitational Radiation by Esc
                    aping Supernova Neutrinos, Astrophys. J. 223 (1978) 1037.",
  "contained_links": [
    { "url": "https://doi.org/10.1086/156337",
      "text": "Astrophys. J. 223 (1978) 1037.",
      "start": 87,
      "end": 117 }
  ],
  "ids" {...} }
```

KNOWLEDGE GRAPH CREATION

- For each category, we have a set of papers.
- For each paper, we have the textual representation of each subsection to convert it to a concept map and add it to the knowledge representation of that paper
- Thus for each paper, we have a knowledge graph containing concept maps for each subsection for each subcategory and thus for a category

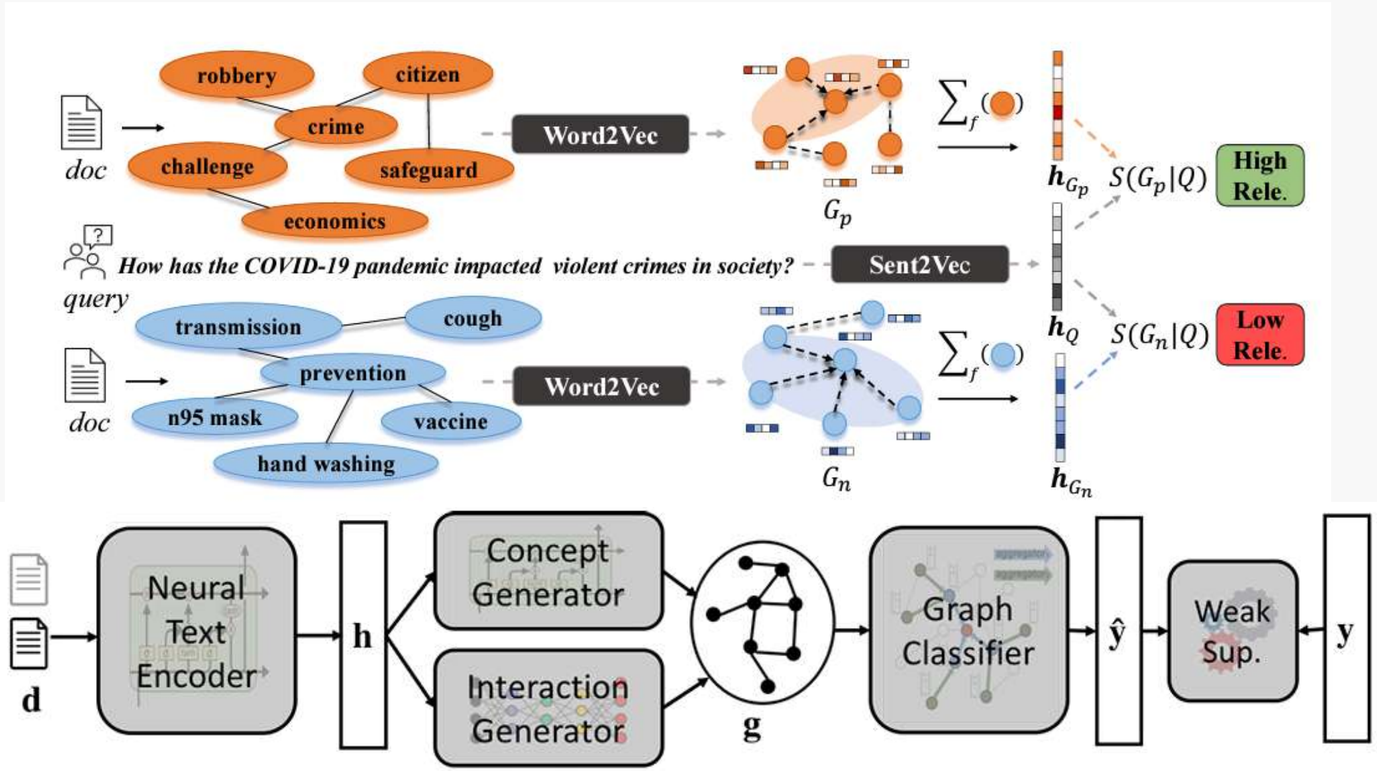
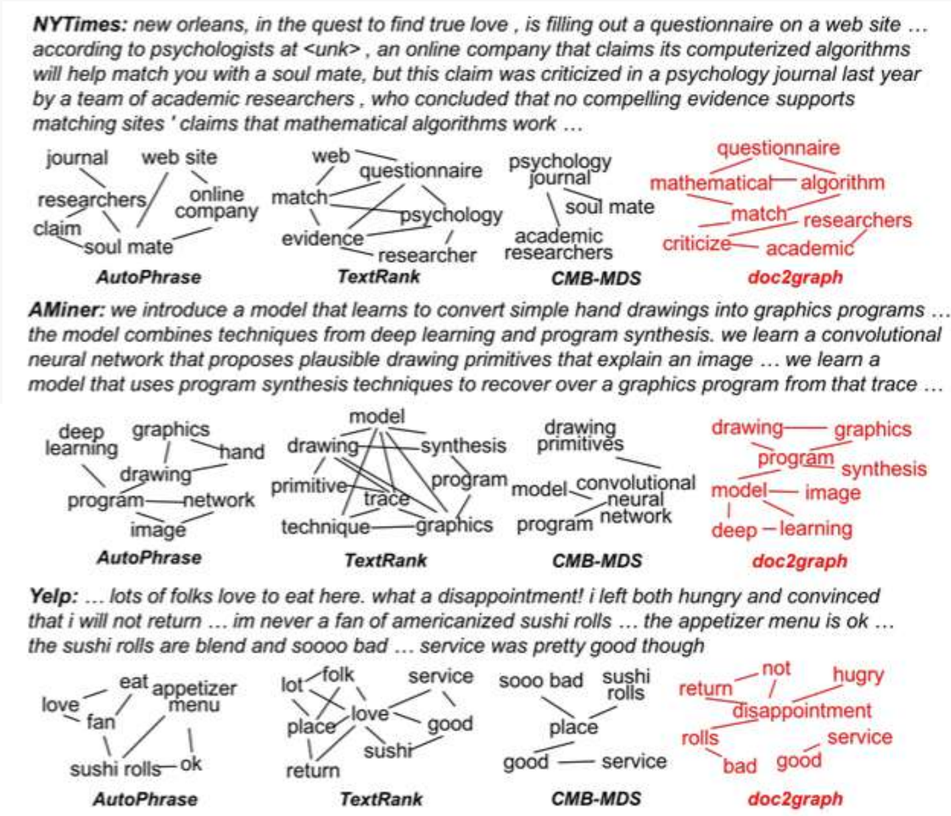
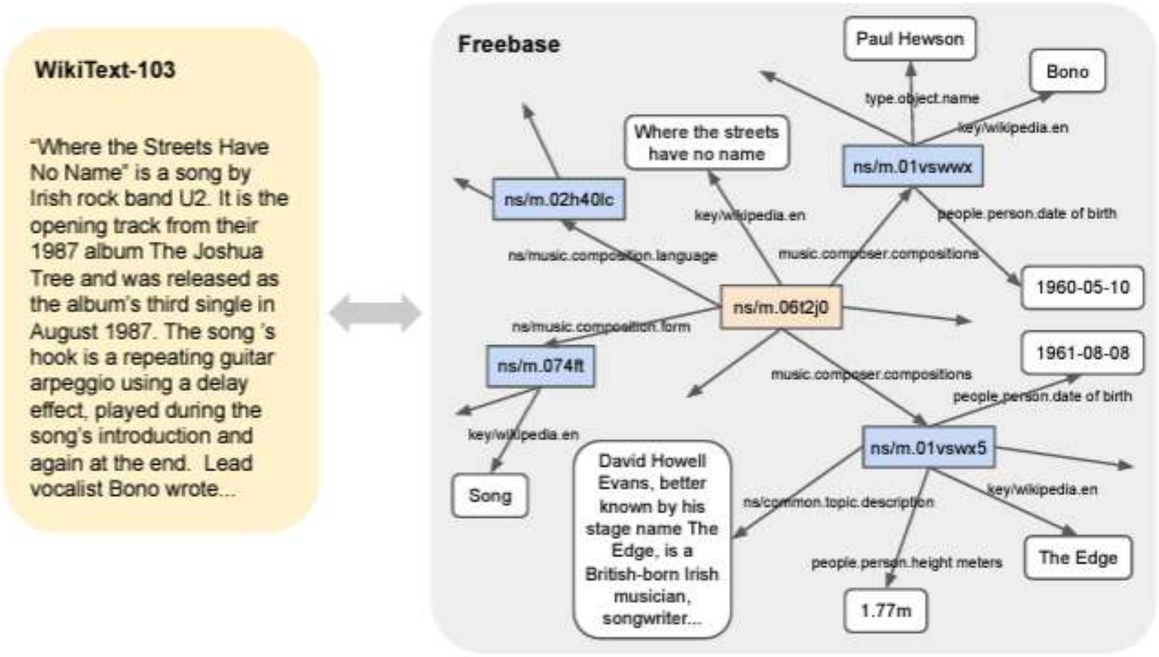
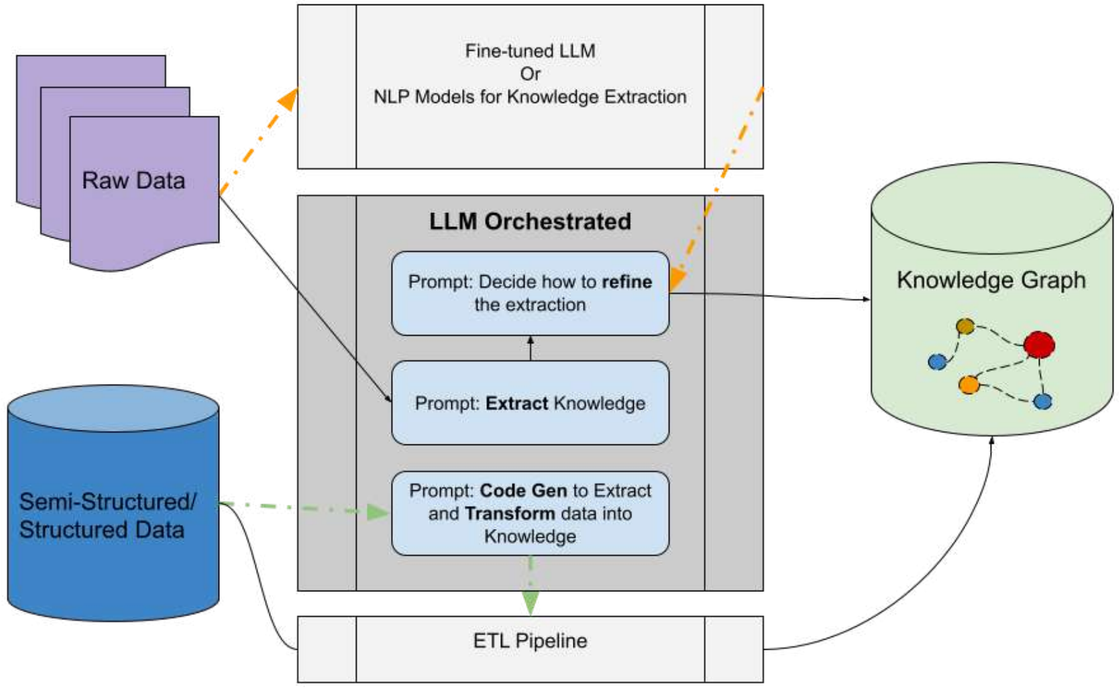


Figure 2: Overview of our proposed *doc2graph* neural framework: A neural text model encodes the content of each document **d** into an embedding vector **h**. A concept generator and an interaction generator use **h** to generate concepts C and weighted links **M** in parallel, which are directly combined as a concept map **g**. A graph classifier predicts a document label **ŷ** based on **g**, which can be directly trained towards the ground-truth document label **y**, while providing weak supervision to the intermediate generation of **g**.



CONVERSION OF TEXT DATA INTO GRAPHICAL REPRESENTATION





- WITH THE HELP OF REPRESENTATION LEARNING AND EMBEDDING METHODS WE CONVERT THE TEXTUAL REPRESENTATION OF DATA THAT HAS MULTIMEDIA INFORMATION IN IT TO A GRAPHICAL REPRESENTATION
- THIS IS DONE WITH THE HELP OF RETRIEVAL AUGMENTED GENERATION GRAPH METHODS THAT CAN EXTRACT RELATIONSHIPS AND KEY COMPONENTS FROM THE TEXT AND CONVERT IT INTO GRAPHS
- ESSENTIALLY EACH SUBSECTION IS CONVERTED INTO A CONCEPT GRAPH. THIS CONCEPT GRAPH IS EMBEDDED INTO A KNOWLEDGE GRAPH THAT IS THE CRUX OF OUR SYSTEM
- THE EMBEDDINGS GENERATION MODEL AND THE EMBEDDINGS GENERATED CAN BE STORED LOCALLY OR IN AN INTERMEDIATE STAGE FOR SECURITY REASONS NAD CAN BE RETRIEVED FOR SUMMARY GENERATION AND FACT CHECKING

CITATIONS AND DANGLING REFERENCES REMOVAL

PAPER(ie. Context)

DANGLING REFERENCES IN THE CONTEXT

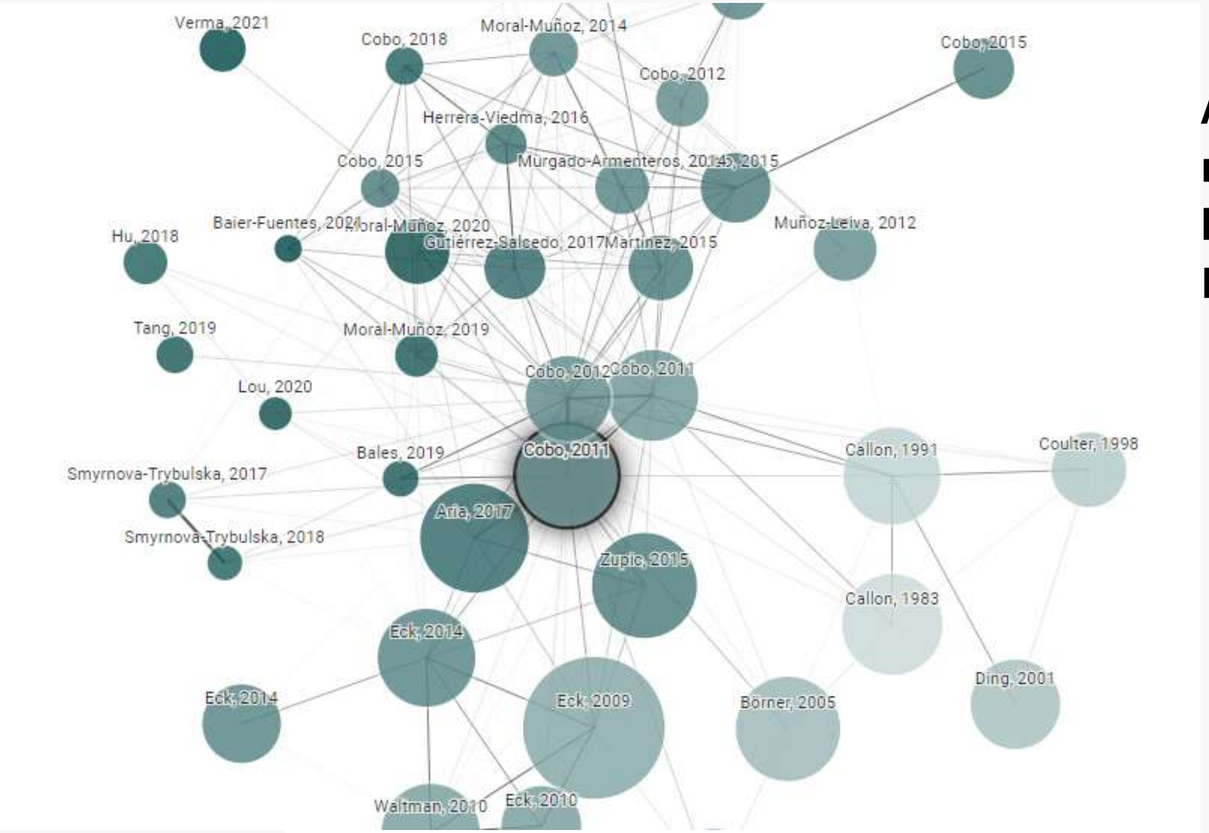
According to a previous paper [35], a typical RoBERTa-based classifier mislabels synthetic text to human by very basic differences such as changing 'a's to alpha or 'e's to epsilon. This vulnerability can be used to trick detectors of synthetic text either intentionally or accidentally.

json SUBSECTION

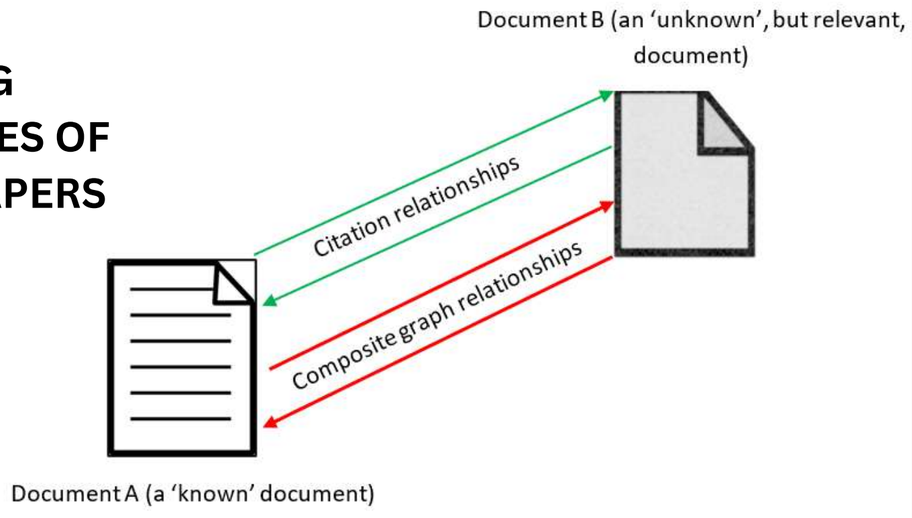

```
{"section": "Goals and Approaches", "sec_number": "3", "sec_type":
"section", "content_type": "paragraph", "text": "According to a previous
paper {{cite:1a20f851d9e2b957bccb4e91019402eb2cd497ff}}, a typical RoBERTa-
based classifier mislabels synthetic text to human by very basic
differences such as changing 'a's to alpha or 'e's to epsilon. This
vulnerability can be used to trick detectors of synthetic text either
intentionally or accidentally.\n", "cite_spans": [{"start": 30, "end": 79,
"text": "{{cite:1a20f851d9e2b957bccb4e91019402eb2cd497ff}}", "ref_id":
"1a20f851d9e2b957bccb4e91019402eb2cd497ff"}], "ref_spans": []}
```


REFERENCE SECTION


```
"1a20f851d9e2b957bccb4e91019402eb2cd497ff": {"bib_entry_raw": "Max Wolff
and Stuart Wolff. Attacking neural text detectors. 2 2020. URL
http://arxiv.org/abs/2002.11768.", "contained_arXiv_ids": [{"id":
"2002.11768", "text": "http://arxiv.org/abs/2002.11768.", "start": 73,
"end": 105}], "contained_links": [], "discipline": "Computer Science",
"ids": {"open_alex_id": "https://openalex.org/W3008544205",
"sem_open_alex_id": "https://semopenalex.org/work/W3008544205",
"pubmed_id": "", "pmc_id": "", "doi": "", "arxiv_id": ""}},
```



MAPPING DEPENDENCIES OF RESEARCH PAPERS



CITATION NETWORK USAGE

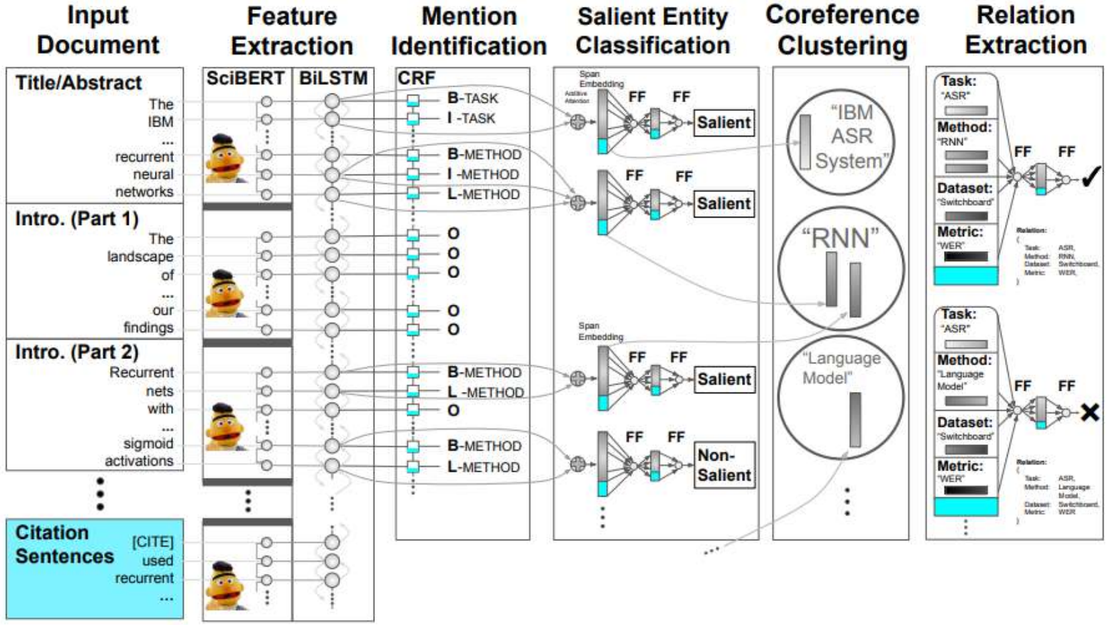
A citation graph, also known as a citation network, is a representation of the relationships among academic papers based on their citations.

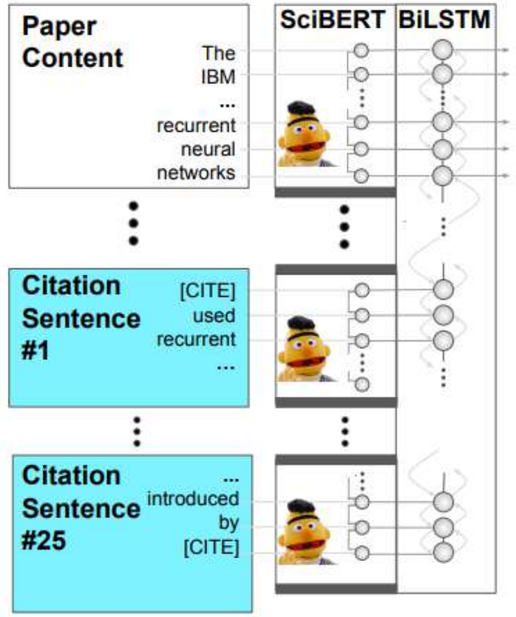
Key Components of a Citation Graph:

1. **Nodes (Papers/Publications): Each node in the citation graph represents a specific academic paper, article, or publication. These nodes contain metadata such as the title, authors, publication venue, abstract, and other relevant information.**
2. **Edges (Citations/References): The edges between nodes indicate the direction of citation from one paper to another. If paper A cites or references paper B, there will be a directed edge from node A to node B in the graph.**

- **We have created a huge citation network of Arxiv Research Paper spanning across 8 categories and 151 subcategories. Upon clicking on a paper in the interactive graph we can see the citations used by the paper and which papers have used it as a reference**
- **This can be exploited to extract dependencies and perform context infusion so that there are no dangling references in the summary generated and the hallucinations are decreased due to the presence of relevant information**

CONTEXT INFUSION OR CHASING CITATIONS





- Graph Context Infusion is a methodological approach that capitalizes on the structured information present in graphs to augment the understanding and performance of models that typically operate on non-graph or sequential data.
- It represents the synergy between structured knowledge representations and unstructured text, allowing machine learning models to leverage both sources of information for more sophisticated and context-aware processing.
- This approach is particularly useful for handling complex, multi-faceted documents where understanding relationships and entities is crucial to producing accurate and informative summaries.

We use these processes to perform context infusion for removal of dangling references:

- Graph-Based Features: Extract relevant features from the graph, such as centrality measures, node importance scores, or community structures. These features can help identify key elements that should be included in the summary.
- Graph-Based Ranking:Integrate graph-based features with traditional text-based features to rank the importance of nodes (sentences or content units) within the graph. This ranking can guide the selection of content for inclusion in the summary.
- Content Extraction:Use the rankings to extract the most important nodes from the graph. These nodes will form the basis of the summary.
- Summary Generation:Utilize the selected nodes to generate the summary. This could involve the extraction of key sentences or the generation of new sentences that connect the chosen nodes in a coherent and informative manner.
- Graph Connectivity:Ensure that the summary maintains the connectivity and flow of information as reflected in the graph. The graph structure can help in organizing the summary in a coherent way, where entities and concepts are connected logically.
- Relevance Assessment:Assess the relevance of the generated summary by comparing it to the graph structure. Ensure that important nodes and relationships are adequately represented in the summa
- Iteration and Refinement:Iteratively refine the summary generation process based on the feedback from the graph context. This may involve adding or removing content to ensure the summary aligns with the graph's structural and semantic information.

CURRENT MODEL
SUMMARY & OVERVIEW GENERATION :
LLAMA-2+ RAG

SUMMARY AND OVERVIEW GENERATION

EXTRACTIVE & ABSTRACTIVE

Abstractive:
T5
PEGASUS
mBART

Extractive:
TF-IDF
PageRank
Latent Dirichlet Allocation (LDA)

FINETUNED LLMS & PROMPT TECHNIQUES

PROMPT METHODS & TECHNIQUES

- Chain-of-Thought Prompting
- Generate Knowledge Prompting
- Tree of Thoughts
- Automatic Reasoning
- Directional Stimulus Prompting
- ReAct
- Multimodal CoT
- Graph Prompting

LLMS+RAG/RoG Fusion

User Query
↓
Generate Similar Queries
↓
Vector Search Queries
↓
Reciprocal Rank Fusion
↓
Re-ranked Results
↓
Generative Output

GRAPH COMPRESSION & SUMMARIZATIONS

NEW

GRAPH LINK ANALYSIS

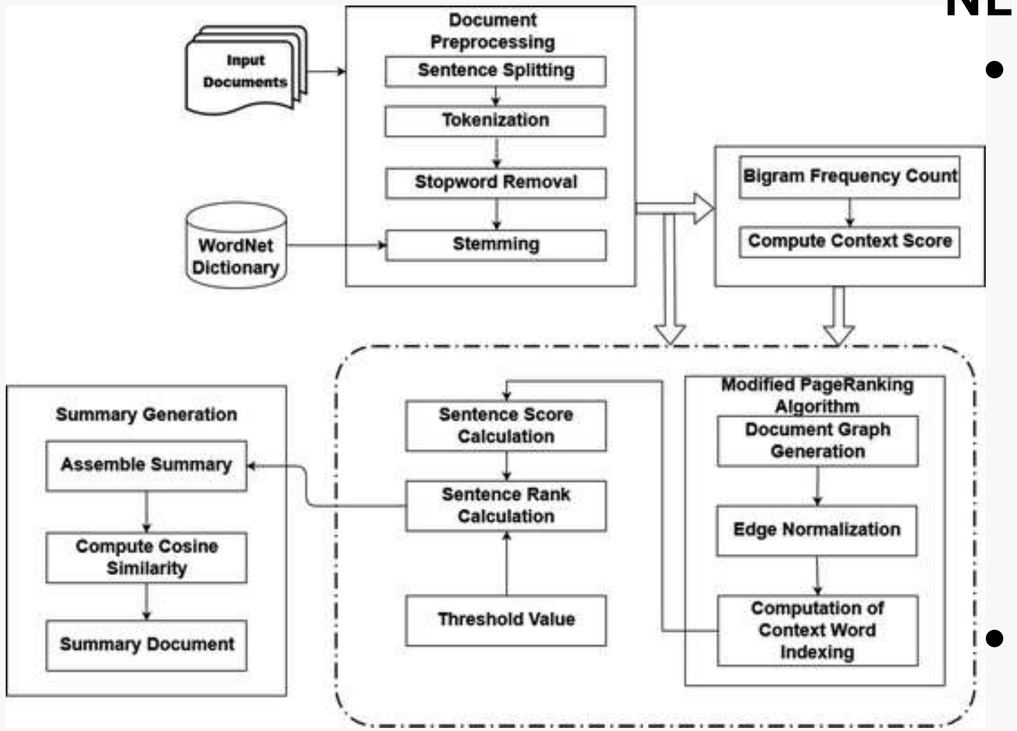
MULTIPLE MESSAGE PASSING LAYERS

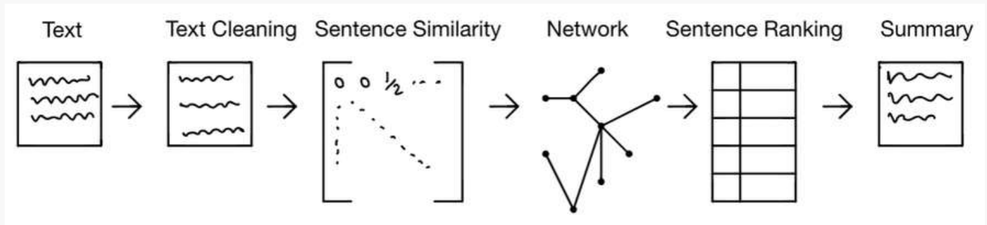
GRAPH COMPLETION

GRAPH2TEXT

EXTRACTIVE SUMMARIZATION

METHODS: TF-IDF & PageRank





DRAWBACKS:

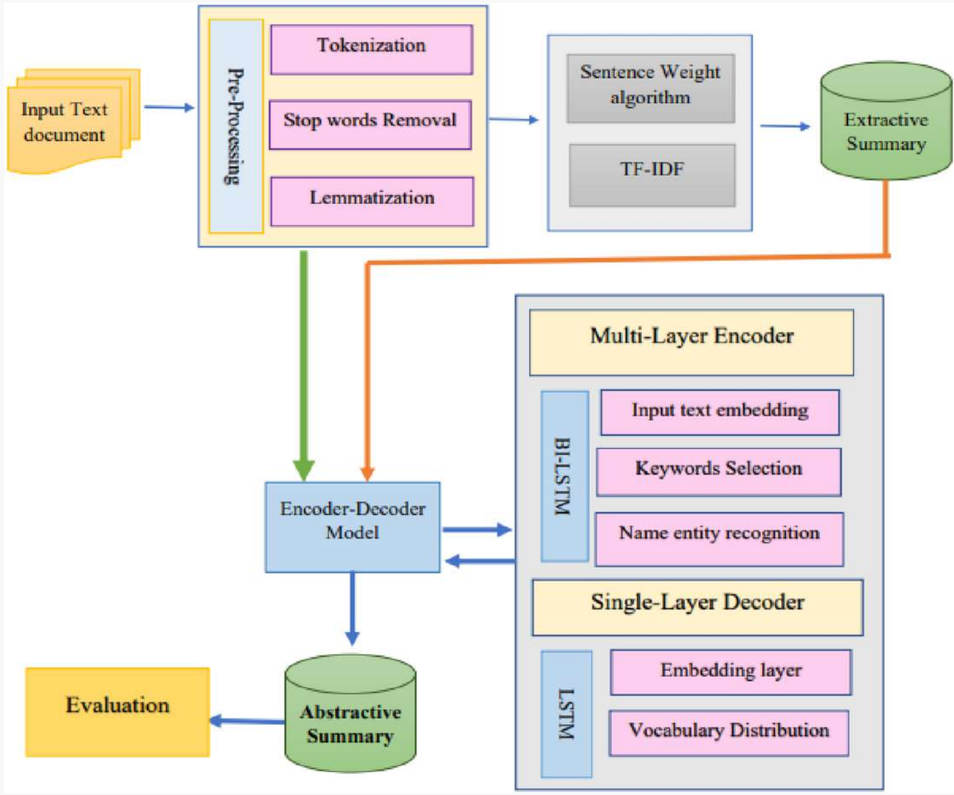
- DANGLING REFERENCES
- BAD QUALITY SUMMARY
- TEXT BREAKS

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

ABSTRACTIVE SUMMARIZATION

MODELS: T5,PEGASUS,Longformer



EMPLOYED METHODS

- Summarizer: allenai/led-large-16384-arxiv
- clustering or semantic search:
  - sentence-transformers/paraphrase-MiniLM-L6-v2
  -sentence-transformers/distilbert-base-nli-mean-token

LARGE LANGUAGE MODELS(LLMs)-PROMPT APPROACH

- TO INCREASE THE GENERATED SUMMARY'S QUALITY WE EMPLOY THE USAGE OF OFFLINE-CAPABLE LARGE LANGUAGE MODELS HAVING MORE PARAMETERS. EX.LLAMA-2, CLAUDE, BLOOM,privateGPT, MosicalML65b etc.

- IT ALSO GENERATES BETTER STRUCTURED OUTPUT PROVIDING LESSER AMOUNT OF HALLUCINATIONS AND BETTER SENTENCE COMPLETION CAPABILITIES

- IT STILL HAS SOME DRAWBACKS LIKE DANGLING REFERENCES, NON-INCLUSION OF EQUATIONS, HALLUCINATIONS

EMPLOYED MODELS

LLaMA by Meta

Claude

a BigScience initiative

BLOOM

METHODS

DIRECTED STIMULUS PROMPTING

Input text

Article: (CNN) For the first time in eight years, a TV legend returned to doing what he does best. Contestants told to "come on down!" on the April 1 edition of "The Price Is Right" encountered not host Drew Carey but another familiar face in charge of the proceedings. Instead, there was Bob Barker, who hosted the TV game show for 35 years before stepping down in 2007. Looking spry at 91, Barker handled the first price-guessing game of the show, the classic "Lucky Seven," before turning hosting duties over to Carey, who finished up. Despite being away from the show for most of the past eight years, Barker didn't seem to miss a beat.

Reference

Bob Barker returned to host "The Price Is Right" on Wednesday . Barker, 91, had retired as host in 2007 .

Directional Stimulus Prompting

Q: Summarize the above article briefly in 2-3 sentences **based on the hint**.
Hint: Bob Barker; TV; April 1; "The Price Is Right"; 2007; 91.

Model output

A: On April 1, Bob Barker returned to the TV show "The Price Is Right" after eight years . Looking spry at 91, Barker handled the first price-guessing game . Barker stepped down as host in 2007, but didn't seem to miss a beat .

ROUGE-1: **48.39** 👍

Standard Prompting

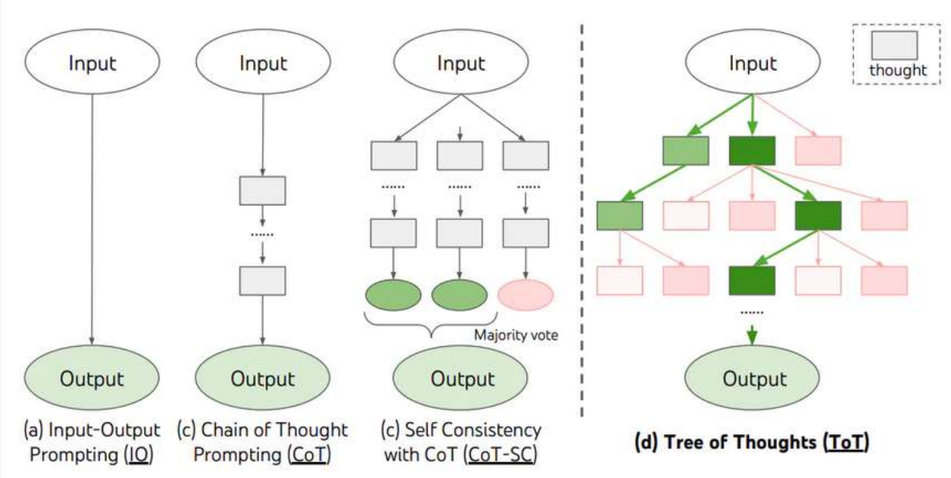
Q: Summarize the above article briefly in 2-3 sentences.

Model output

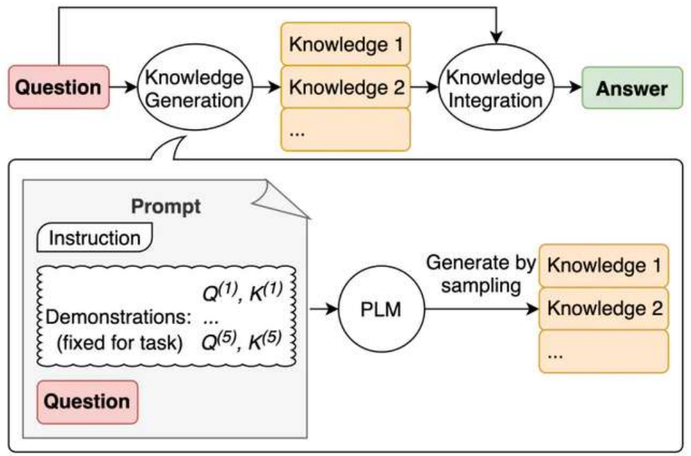
A: Bob Barker, the former host of \"The Price Is Right\", returned to the show for the first time in 8 years. He hosted the first game of the show, and then handed the hosting duties over to Drew Carey.

ROUGE-1: **34.48** 👎

CHAIN OF THOUGHT & TREE OF THOUGHT

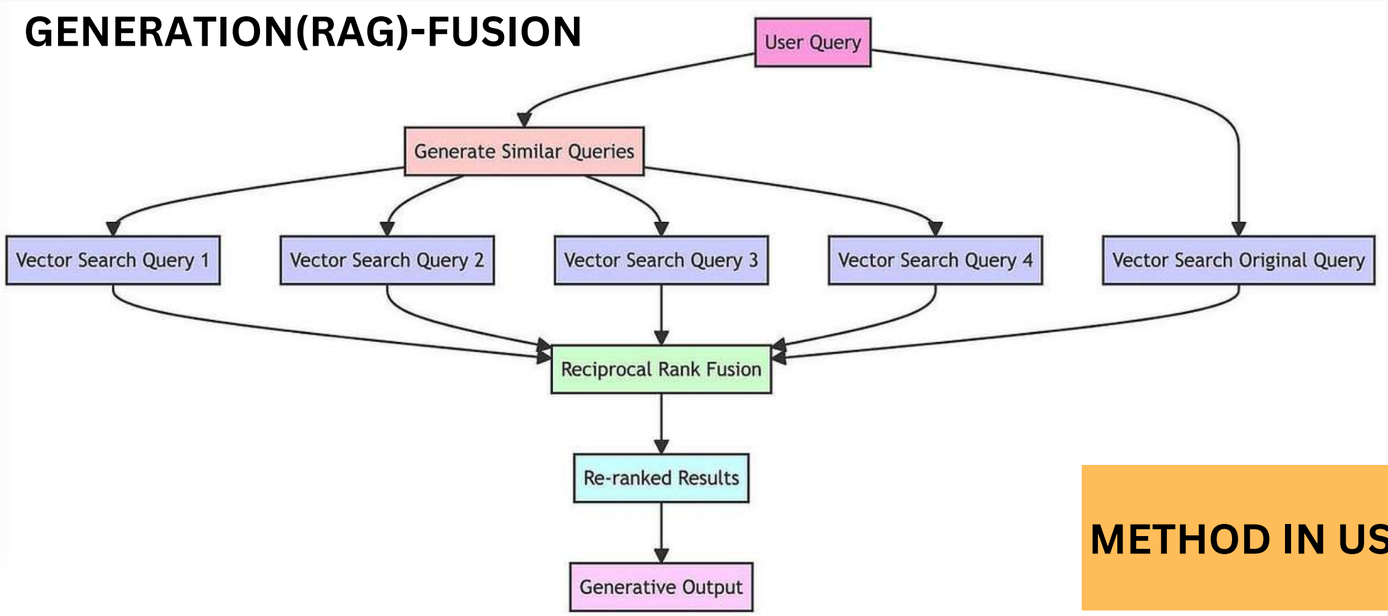


GENERATED KNOWLEDGE PROMPTING



LLMS+ RAG/RoG APPROACH

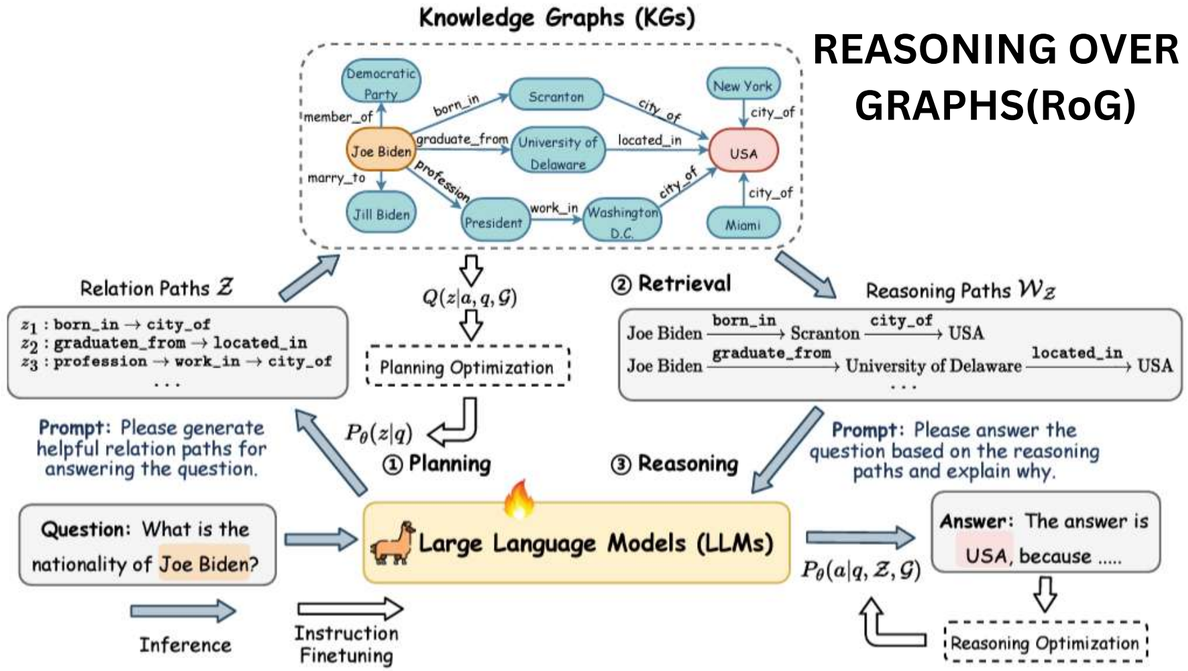
RETRIEVAL AUGMENTED GENERATION(RAG)-FUSION



- RAG combines an information retrieval component with a text generator model. RAG can be fine-tuned and its internal knowledge can be modified in an efficient manner and without needing retraining of the entire model.
- The documents are concatenated as context with the original input prompt and fed to the text generator which produces the final output. This makes RAG adaptive for situations where facts could evolve.
- This is very useful as LLMs's parametric knowledge is static. RAG allows language models to bypass retraining, enabling access to the latest information for generating reliable outputs via retrieval-based generation.

METHOD IN USAGE

REASONING OVER GRAPHS(RoG)



NO DRAWBACKS

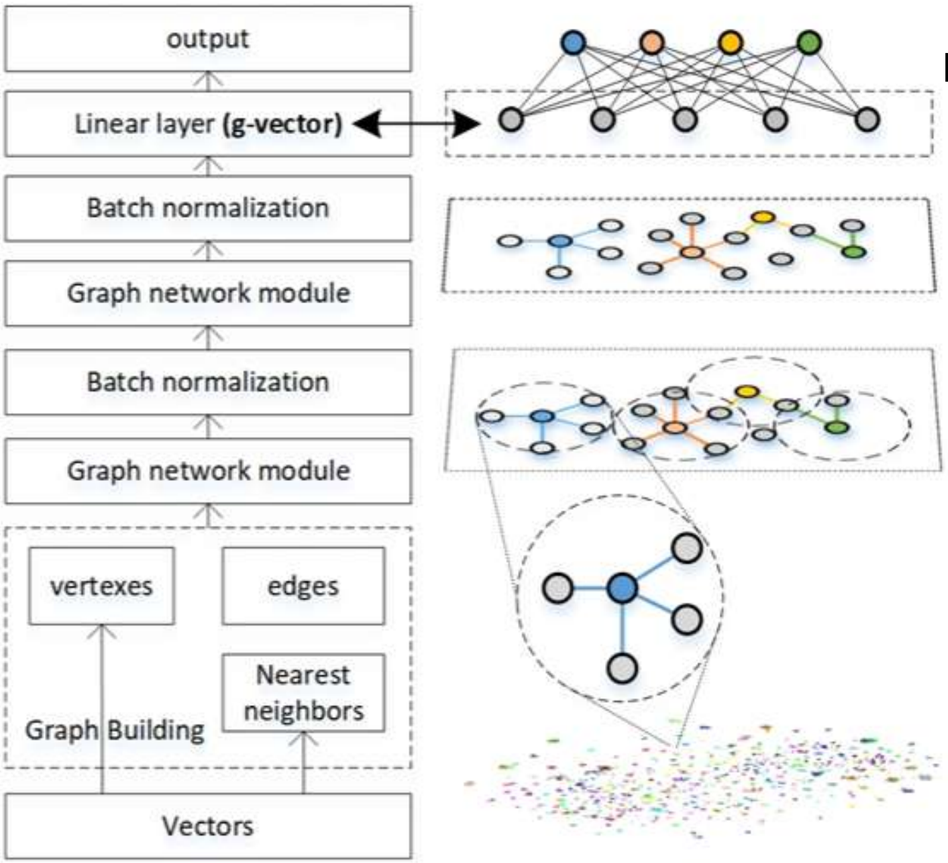
LESSER HALLUCINATIONS, NO DANGLING REFERENCES

STATE OF THE ART STRUCTURED SUMMARY GENERATION

USES EQUATIONS & TABLES INFORMATION

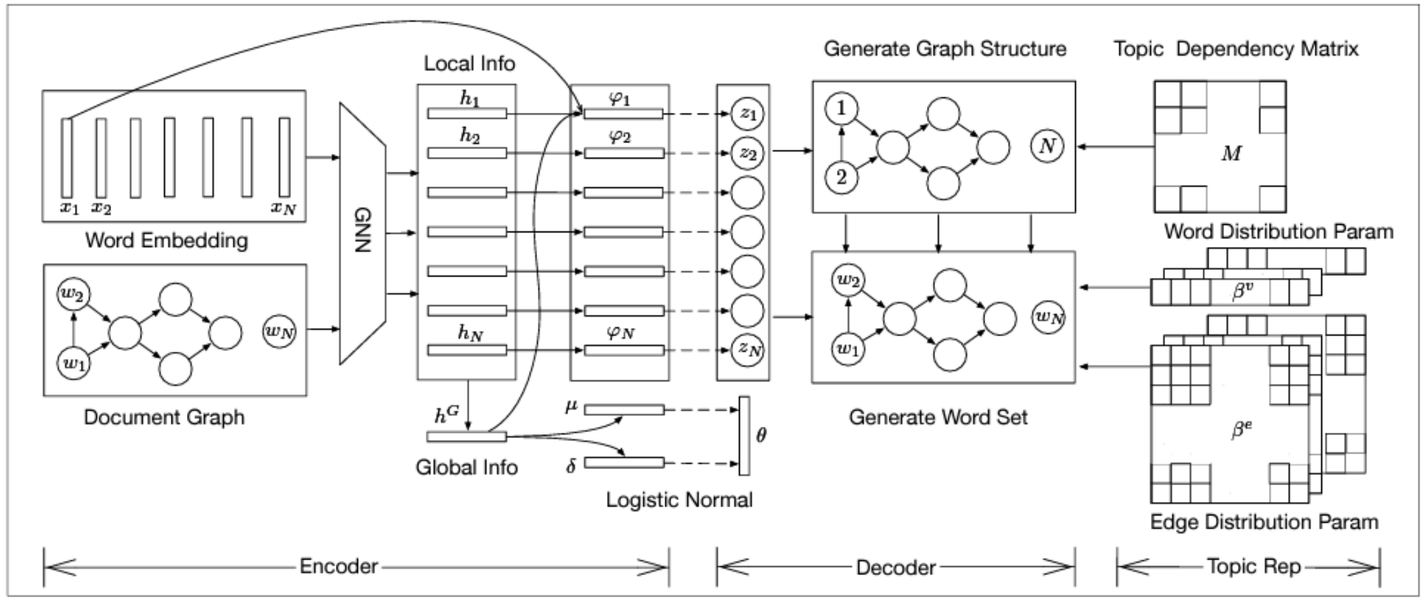
- Reasoning on graphs (RoG) synergizes LLMs with KGs to enable faithful and interpretable reasoning.
- We use a planning-retrieval-reasoning framework, where RoG first generates relation paths grounded by KGs as faithful plans.
- These plans are then used to retrieve valid reasoning paths from the KGs for LLMs to conduct faithful reasoning and generate interpretable results

GRAPH SUMMARIZATION & CONVERSION TO TEXT



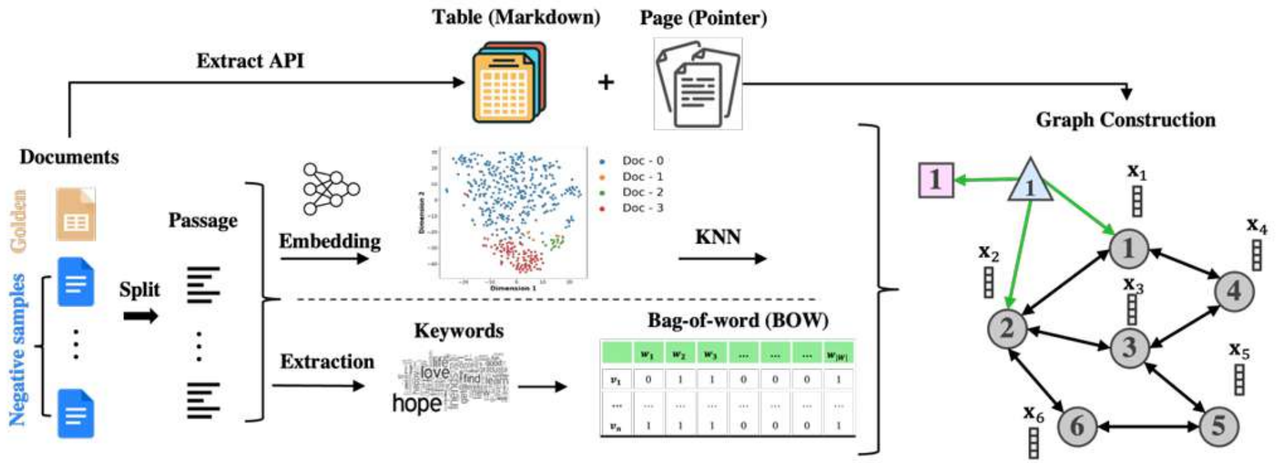
Knowledge Graph Summarization

- Each subsection is converted into a concept map using GNN methods.
- 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



This further enhances the summarization process, providing a comprehensive understanding of the research area.

GRAPH SUMMARIZATION & CONVERSION TO TEXT



- Graph2Text is a methodology that focuses on transforming graph structures into natural language text, aiding in the comprehension of the graph content. It involves generating human-readable summaries or descriptions from the graph data, enabling easier interpretation and analysis of the underlying relationships and entities within the graph.
- Approach:
  - Graph Encoding: Translate the graph structure into a format understandable by a language model.
  - Text Generation: Utilize natural language generation techniques to create coherent summaries or descriptions based on the graph structure.
- Components:
  - Graph Representation: Nodes, edges, properties, and relationships are encoded in a way understandable by the model.
  - Language Generation Model: Utilize techniques such as sequence-to-sequence models, often based on recurrent neural networks (RNNs) or transformer architectures, to generate text descriptions.

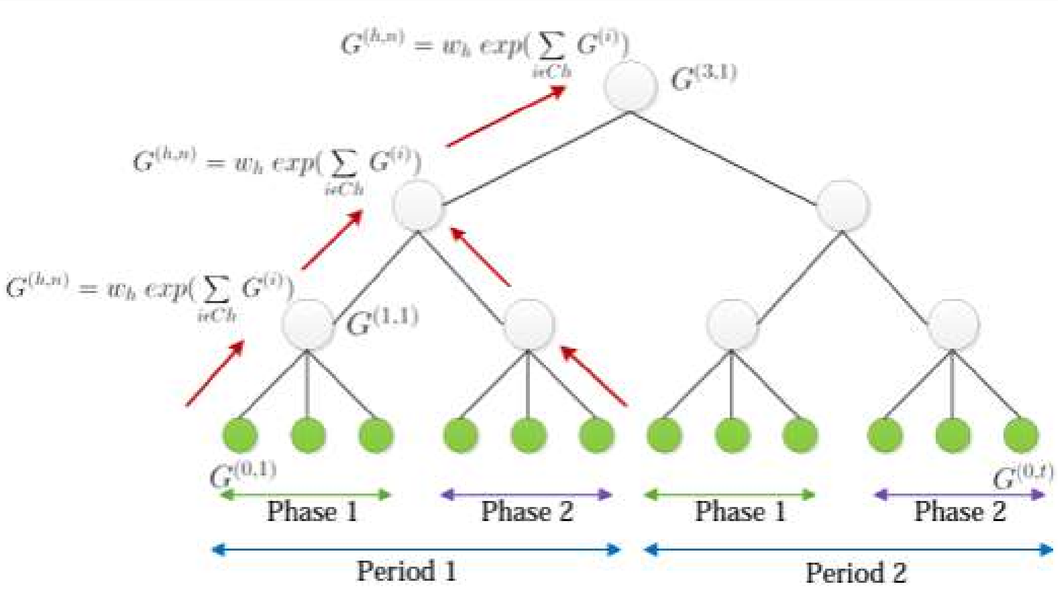
Applications

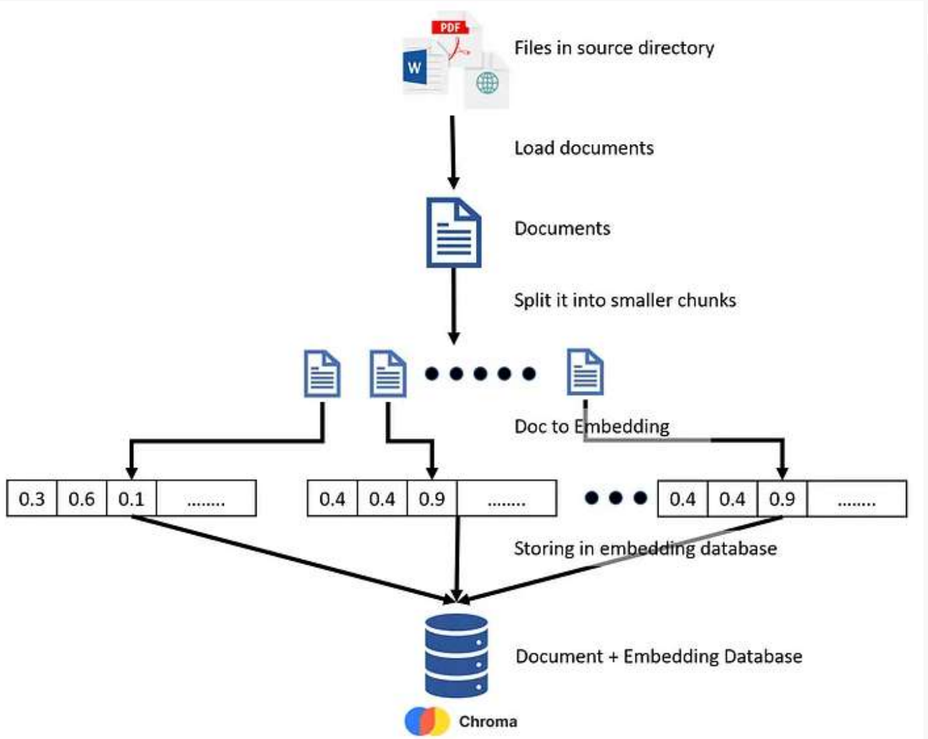
- Data Interpretation: Aid in understanding complex network structures by providing human-understandable summaries.
- Automated Reporting: Generating textual summaries for analytical reports based on large graphs or datasets.

Graph summarization is the process of condensing the information within a complex graph into a more compact and understandable representation without losing the essential structural and semantic information.

Techniques:

- Node/Graph Clustering: Identifying clusters or communities of nodes based on their connections, thus reducing the complexity.
- Subgraph Extraction: Selecting essential subgraphs representing important parts of the whole graph.
- Key Node Identification: Finding crucial nodes based on centrality, influence, or other metrics to represent the entire graph.



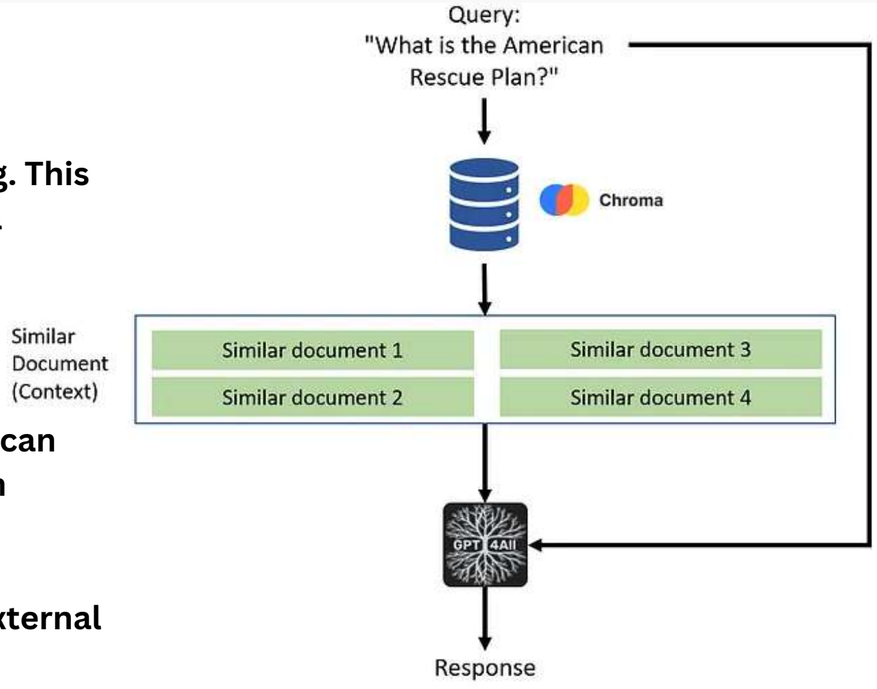


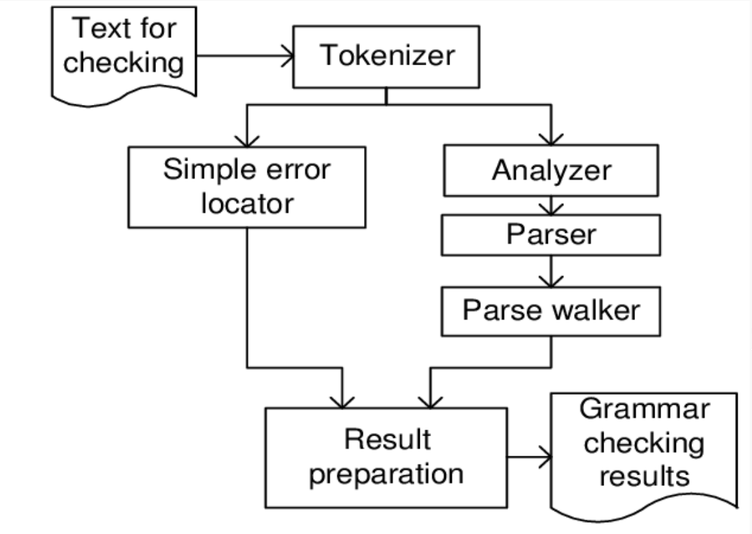
HOW WILL THE LLMs WORK OFFLINE

- THE LARGE LANGUAGE MODELS(LLMS) CAN BE DOWNLOADED AND UTILIZED
- THE GENERATION OF EMBEDDINGS AND STORAGE OF KNOWLEDGE GRAPH CAN BE DONE ON LOCAL SYSTEM ON CHROMADB
- INFERENCE CAN BE DONE BY SIMPLY RETRIEVING AND QUERYING OVER THE GRAPH AND PRESENTING THE OUTPUT
- WE HAVE VECTOR DATABASES(CHROMADB) EXPLICITLY DESIGNED FOR EFFICIENT STORAGE AND RETRIEVAL OF VECTOR EMBEDDINGS.

Using offline(not connected over the internet) large language models offers several advantages:

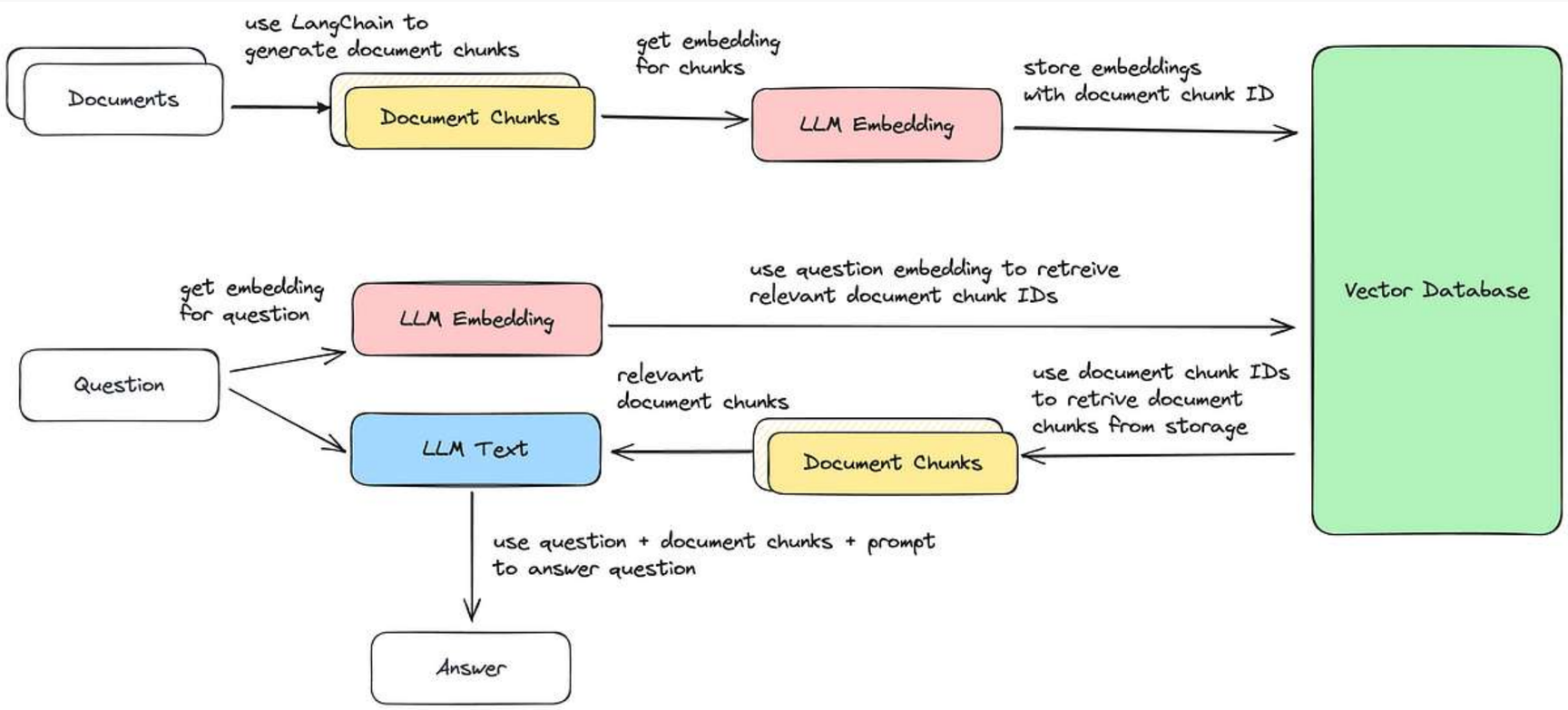
(1) Privacy and Data Security: Offline models don't require data to be sent to a remote server for processing. This mitigates potential privacy concerns as sensitive data remains on the local device, reducing the risk of data breaches or unauthorized access.This also prevents data leakages from occuring.
(2) Low Latency: With an offline model, there's no dependency on internet connectivity, resulting in faster response times. This is particularly beneficial in applications where real-time or low-latency responses are crucial, such as in certain customer service interactions or edge computing scenarios.
(3) Resource Efficiency: Using offline models reduces the need for continuous internet connectivity, which can save on bandwidth and potentially reduce costs associated with server usage and cloud services.Relying on offline models decreases reliance on cloud services or external infrastructure, which can be particularly beneficial in remote or underdeveloped areas with limited access to robust internet connections.
(4) Isolation from External Variabilities: An offline model is not affected by changes or fluctuations in the external environment, such as server downtime or variations in remote services, ensuring consistent performance.





GRAMMAR CHECK & FACT CHECKING SUMMARY GENERATED

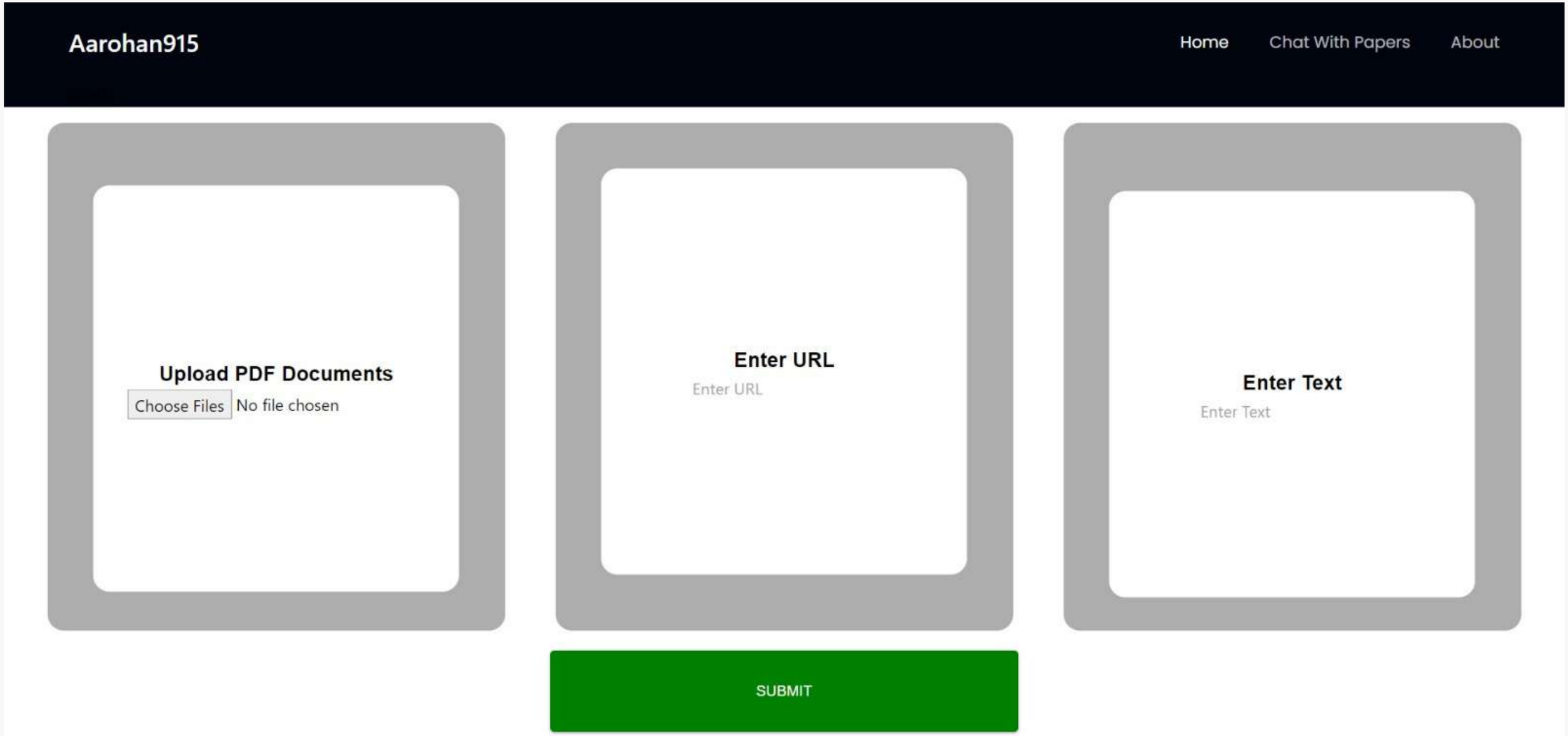
TO MAKE SURE THE SUMMARY GENERATED ABIDES BY GRAMMAR RULES. WE APPLY MULTILINGUAL LLMS-BASED MODELS(LLAMA-2) TO CHECK THE SENTENCE STRUCTURE AND MAINTAINING CONTEXTUAL INTEGRITY OF THE SUMMARY GENERATED.



USAGE OF RAG AND RoG GREATLY REDUCED HALLUCINATIONS BUT TO MAKE SURE THAT CONTEXT INFUSION HAS TAKEN PLACE PROPERLY WE PERFORM COMMON SENSE REASONING TO CHECK IF THERE IS ANY ANOMALY IN THE REASONING OF THE SUMMARY GENERATED

THIS CAN BE ACHIEVED WITH THE HELP OF DOCUMENT QUESTION ANSWERING AND MULTIMODAL CoT PROMPTING TO GO OVER THE INCLUDED IMAGES AND TABLES IN THE SUMMARY

LANDING PAGE



OUTPUT

DIFFERENT LEVELS OF SUMMARY WITH REVLEVANT MEDIAS (IMAGES,TABLES ETC)

KEYWORDS,CONCEPTS AND THIER DESCRIPTION

OVERVIEW OF THE TOPIC

- PRESENTING ALTERNATE VIEWPOINTS
- RELEVANT FEILDS TO EXPLORE
- RECOMMENDATIONS IN TERMS OF VIDEOS, ARTICLES AND PAPERS

GRAPH OF THE PAPER/ARTICLE AND ITS CITATIONS

GRAPH OF RELEVANT FEILDS & DOMAINS TO EXPLORE

CHAT WITH THE DOCUMENT (DOCUMENT QUESTION AND ANSWERING)

DOWNSTREAM APPLICATIONS

NTRO MAINLY DEALS WITH ISRO SATELLITE DATA & RESEARCH AND ANALYSIS WING(RAW) DATA

DOWNLINK SATELLITE DATA ANALYSIS & INFERENCE

- Using Physics Informed Neural Networks(PINNs) and information from the Knowledge Graph related to instrumentation, geography etc.

- We can perform real-time analysis on satellite data. We can also perform summarization to define the priority of data for downlink analysis and generation of overview and scientific reports.

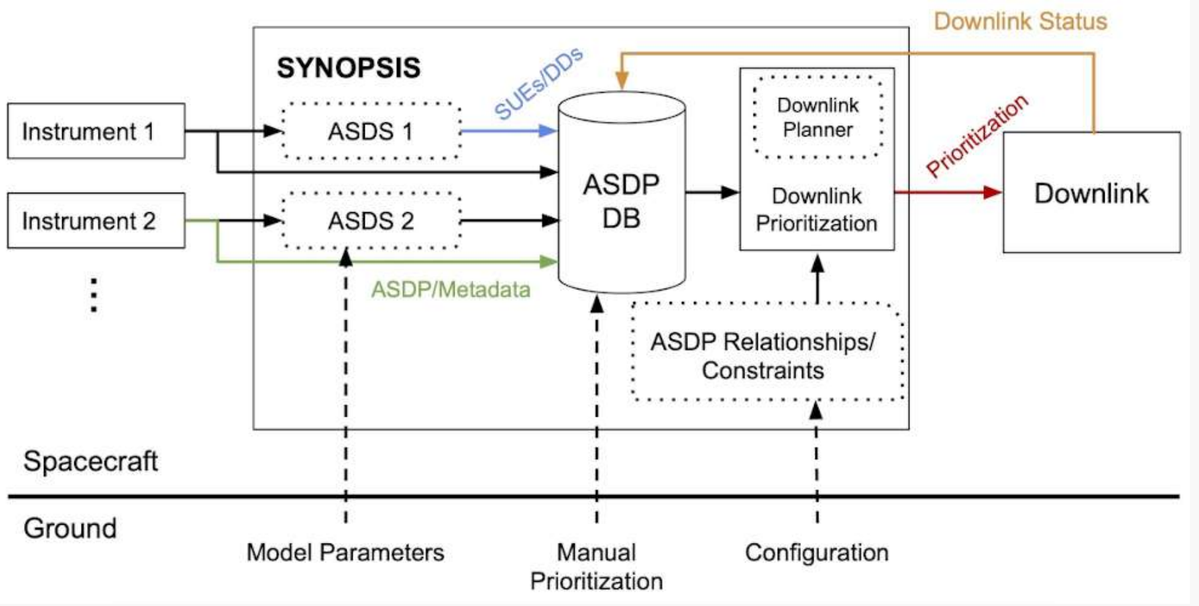
- Modern science datasets from missions in operational environments may have 500+ simultaneous measurements at each of millions of time samples.

- Scientists would often like to look through the record and discover not only expected trends but ones they did not initially guess, while Ops personnel perform the same task under serious time pressure should an anomaly occur.

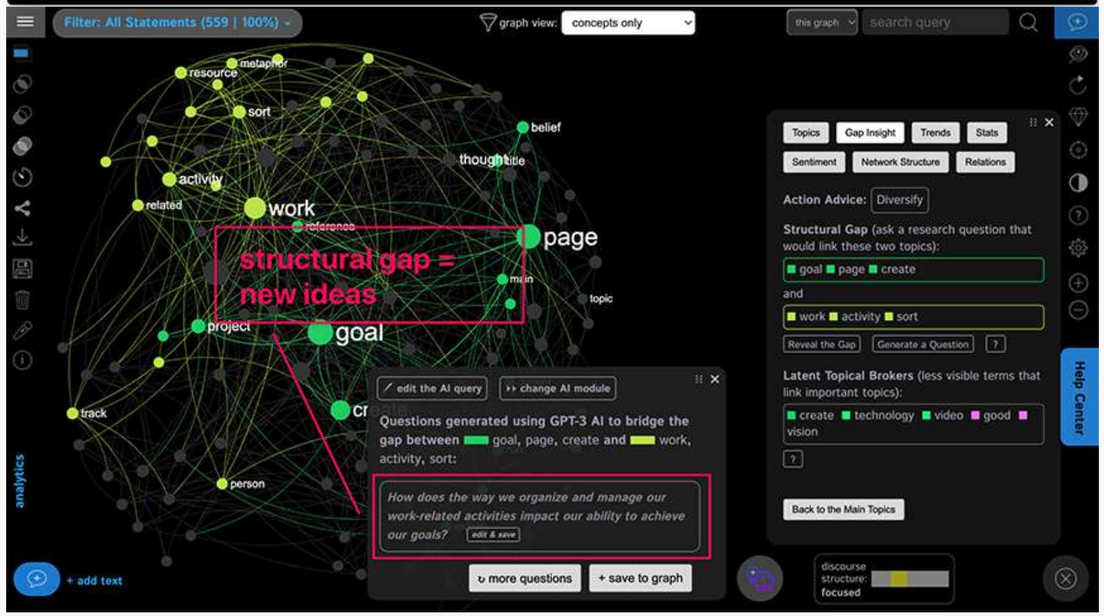
- In both cases, the optimal environment for this rapid exploration of large data would be one where visualizations were clear, interactive, and responsive, permitting the investigator to “play” with the data and gain rapid insight, falsify hypotheses, and make discoveries.

- Machine Learning (ML) has proven invaluable in providing some of these key data insights, but to do so in a statistically robust and reliable manner requires a data science professional and a lot of custom Python code, losing any sense of interaction and play.

- Usage of a desktop-like environment with standard scientific graph types that are robust to rapid, powerful exploration.



- For example, all graphs are linked such that selecting data on one graph reveals those same points in all others regardless of type including histograms and heat maps.
- Machine Learning comes into play as well, allowing users to request high-level directions like “Show me more like this region,” “How many groups best describe all of these observations,” and “Predict this value based on these 500 other values, and show me the small set of variables that were responsible for 90% of the prediction.”
- All of this power normally requires considerable tuning of complex parameters in the ML algorithms.
- We achieve this tuning visually, by showing demonstrations of what would happen for various hyperparameters and permitting the user to select visually the desired outcome.



IDENTIFYING & EXPLORING RESEARCH AREAS

- WITH THE HELP OF LINK ANALYSIS AND PREDICTION WE CAN DETECT UNTAPPED RESEARCH AREAS WITH THE HELP OF CITATION GRAPH

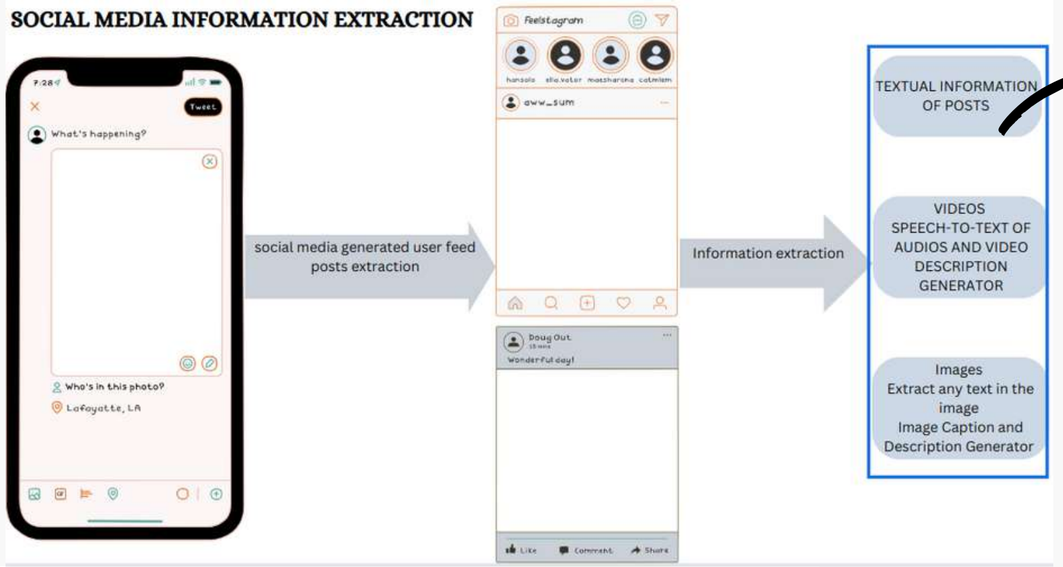
- THIS CAN BE DONE AT MICROLEVEL IE. A HIGHLY SPECIFIC RESEARCH IS LIKE ‘MULTIPLE DOCUMENT SUMMARY GENERATION USING KNOWLEDGE INFUSION VIA SUBGRAPH MAPPING'

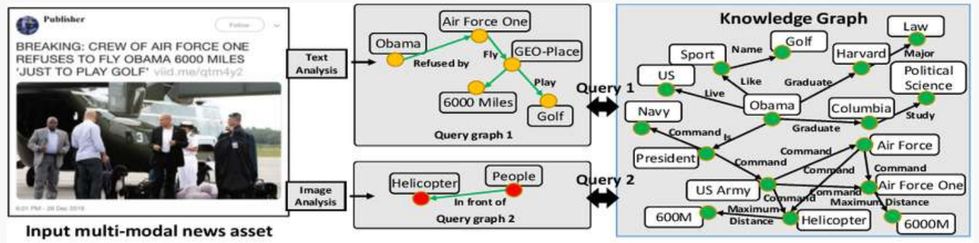
- OR AT A MACRO LEVEL IE. OVER DIFFERENT CATEGORIES OF RESEARCH AREAS LIKE THE INTERACTION OF QUANTUM CHEMISTRY AND COMPUTATIONAL METHODS LIKE GRAPH NEURAL NETWORKS

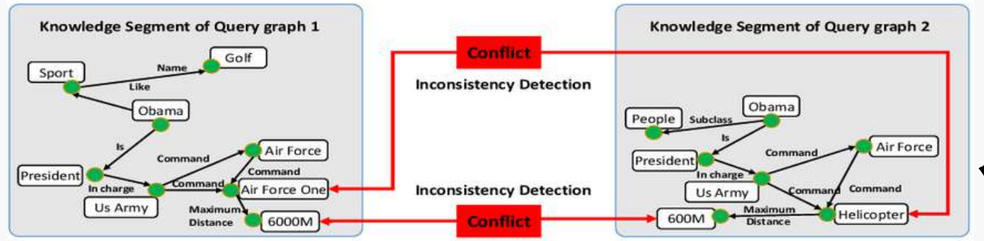
- THIS CAN HELP RESEARCHERS ANALYSE A FIELD AND WORK ON THE UNTAPPED AREAS. THIS CAN ALSO HELP INSTITUTIONS AND ORGANIZATIONS PREDICT FUTURE AREAS OF RESEARCH AND DEVELOPMENT AND KEEPING TRACK OF NEW DEVELOPMENTS IN THE FIELD

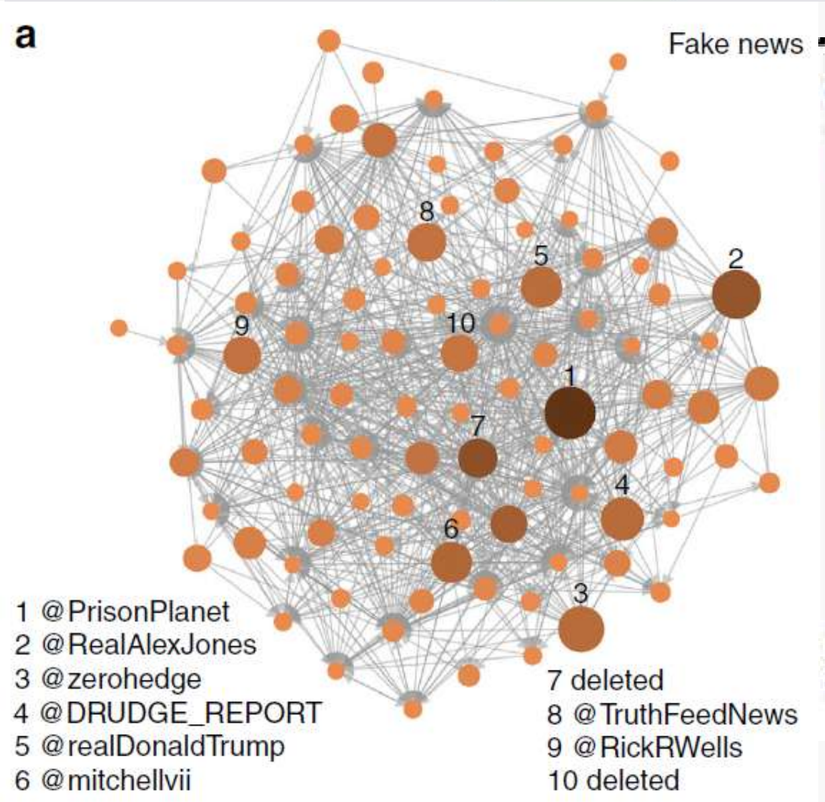
- THE GOVERNMENT CAN ALSO KEEP TABS ON RESEARCH AND INVEST IN UPCOMING LABS AND TECHNOLOGIES INCREASING THE TECHNOLOGICAL ADVANCEMENTS OF THE NATION

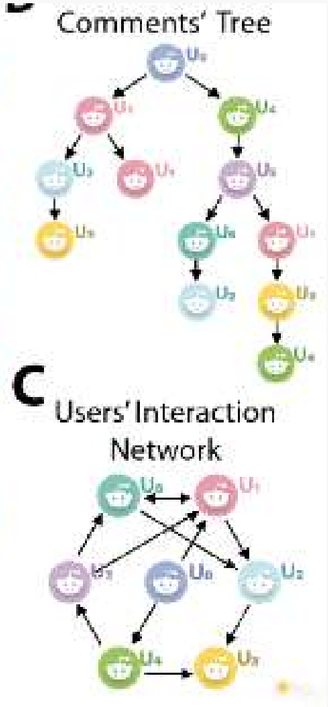
ADVANCED FAKE NEWS DETECTION SYSTEM .











ANALYSIS	
USERNAME & USER_ID	
URL ON POST	SENTIMENT ANALYSIS
TEXT ON THE POST	TYPE OF SPEECH (OFFENSIVE,HATESPEECH ETC.)
CATEGORY OF NEWS (SPORTS,EDUCTION,BUISNESS,SCIENCE & TECHNOLOY,MEDICINE & HEALTH)	CLICKBAIT PROBABILITY
POLITICAL VIEW EXPRESSED	EXTENT OF REPOST (NO OF ACTITVITY ON POST)
TYPE OF PORPAGANDA TECHNIQUE USED	COMMUNITY & ECHOCHAMBER CATEGORY PREDICTIONS
KEYWORDS	SECONDARY USER IDS

ADVANCED FAKE NEWS DETECTION SYSTEM CNTD.

1. Multiple Factchecking LLMS

Using GPT-4, Alpaca, GPT-3.5, LLaMA all fine tuned on our customized database and knowledge graph for fact checking

2. Hybrid Graph Neural Network and Transformer based Large Language Model

Graph Neural Networks:

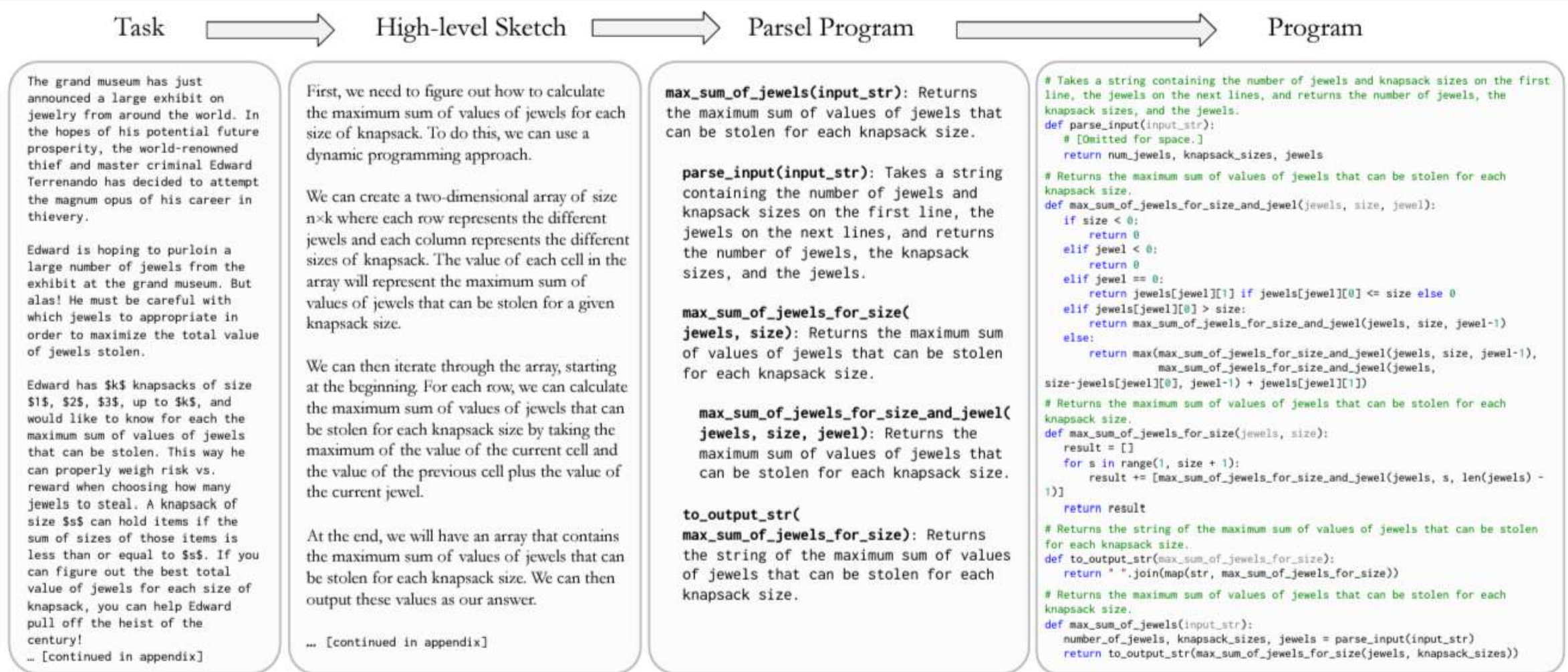
- Advanced graph neural networks like Bi-Directional Graph Convolutional Networks(BiDGCN), Graph Convolutional Networks for Fake News(GCNFN), User Preference-aware Fake News Detection (UPFD)
- all have proven very good results on rumor detection on social media and fake news probability detection based on user history and social graphs
- Fine-tuning using our centralized database the customized Large Language Model will be able to detect fake news and trace it very quickly

- **USING PRESS INFORMATION BUREAU PRESS RELEASES DATA AS A KNOWLEDGE GRAPH WE CAN PERFORM SUMMARIZATION AND QUERIES ON THE GRAPH FOR FACT-CHECKING**
- **CREATING TIME-SERIES BASED CLUSTERS OF INFORMATION FOR CRIME ANALYSIS AND PROVIDING FACTUAL INFORMATION**
- **THIS WILL ALSO HELP US IN ANALYSING THE EVOLUTION & SPREAD OF INFORMATION RELATED TO A PARTICULAR EVENT AND ANALYSE IMPACT ON SOCIAL MEDIA BY SOCIAL MEDIA ANALYSIS**
- **WITH THE HELP OF ECHO CHAMBER PREDICTION AND ANALYSIS OF NEWS AND PERSONS INTERACTION, WE CAN REDUCE THE POOL OF SUSPECTS THAT PROPOGATE FAKE NEWS AND OTHER CYBER CRIMES**

RESEARCH PAPER CODE GENERATION

USING THE PREPROCESSING STEPS WE CAN EXTRACT METHODS ,DATASETS & RESULTS FOR ABSTRACT SYNTAX TREE FOR CODE GENERATION.

WE CAN EXTRACT HIGH LEVEL TASKS AND USING TREE OF THOUGHTS PROMPT METHODOLOGY PERFORM CODE GENERATION FROM LLMS LIKE BLOOM AND LLAMA-2



IMPACT

OUT OF THE 17 U.N. SUSTAINABLE GOAL OUR SOLUTION FOCUSES ON

BY PROVIDING STATE-OF-THE-ART SUMMARIZATION METHODS WE CAN IMPROVE THE ACCESSIBILITY OF QUALITY EDUCATION AND INFORMATION TO EVERYONE

BY PROVIDING A CITATION NETWORK AND UNTAPPED RESEARCH AREAS WE CAN FOSTER COLLABORATION ACROSS INSTITUTIONS BY GATHERING THE BEST MINDS FOR RESEARCH AND INNOVATION

BY PROVIDING FACTUAL INFORMATION OF REAL-TIME NEWS WE CAN TACKLE MISINFORMATION & FAKE NEWS PROPAGATION BY PRESENTING ALL FACTS RELATED TO AN EVENT

BY PROVIDING STRUCTURED INFORMATION SPANNING ACROSS MULTIPLE AGENCIES AND COUNTRIES. WE CAN FURTHER THE COMMON CAUSE OF SUSTAINABLE DEVELOPMENT. AS WE KNOW INFORMATION IS KNOWLEDGE

CONCLUSION

PROBLEM	OUR SOLUTION
OFFLINE FUNCTIONALITY	DOWNLOADING STATE-OF-THE-ART LIGHTWEIGHT MODEL QUANTIZED FOR SUMMARIZATION & VECTOR EMBEDDING GENERATION AND STORAGE AT LOCAL STORAGE FOR RETRIEVAL AND QUERYING ABILITY TO LOAD AND GENERATE RESPONCES IN LOW RESOURCE COMPUTATIONAL ENVIRONMENT
ABLILITY TO INCLUDE IMAGES,EQUATIONS AND TABLES	USAGE OF OCR,SCENE GRAPH GENERATION,DOCUMENT QUESTION AND ANSWERING WE ARE ABLE TO INCLUDE IMAGES, EQUATIONS AND TABLES
MULTILINGUAL	WE HAVE SUPPORT FOR OVER 10+ MOST POPULAR LANGUAGES
DANGLING REFERENCES	USAGE OF CITATION NETWORK AND KNOWLEDGE INFUSION FOR CONTEXT MAINTAINENCE AND BETTER QUALITY OF OVERVIEW PRESENTING DIVERSE VIEW POINTS
CONTEXTUAL INTEGRETY & FACT CHECKING	USAGE OF RAG REDUCES HALLUCINATIONS AND USING FINETUNED LLMS FOR GRAMMER CHECK AND CONTEXTUAL INTEGRITY CHECK

WE HAVE CREATED A SOLUTION FULFILLS ALL THE REQUIREMENTS OF THE PROBLEM STATEMENT WITH STATE-OF-THE-ART RESULTS USING MINIMUM COMPUTE TIME, STORAGE WITH OFFLINE FUNCTIONALITY AND MORE

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

THANK YOU

Team Leader: **Samiul Sheikh**

B.Tech, Electronics Engineering, Final Year, VJTI

Team Member 1 Name: **Mihir Sahasrabudhe**

B.Tech, Electronics Engineering, Final Year, VJTI

Team Member 2 Name: **Sheena Dmello**

B.Tech, Electronics Engineering, Final Year, VJTI

Team Member 3 Name: **Pranav Janjani**

B.Tech, Computer Engineering, Pre-Final Year, VJTI

Team Member 4 Name: **Kanak Meshram**

B.Tech, Electronics Engineering, Final Year, VJTI

Team Member 5 Name: **Yash Deshpande**

B.Tech, Electronics Engineering, Final Year, VJTI

Team Mentor 1 Name: **Faruk Kazi**

Category: **Academic** Expertise: **Cybersecurity** Domain Experience: **25 Years**

Team Mentor 2 Name: **Aum Patil**

Category): **Industry** Expertise: **AI** Domain Experience: **3 Years**: