IIT ROORKEE

Misinformation Identification: Knowledge Graph based RAG Technique

Submitted by,
Jimmy Aghera
Department of Computer Science
and Engineering

Submitted to,
Prof. Durga Toshniwal
Department of Computer Science
and Engineering

Contents

- Introduction
- Problem Statement
- Literature Review
- Experiments
- Research Gap
- Proposed Methodology
- Gantt Chart
- References

Introduction

- The rapid spread of false information on social platforms leads to social unrest, public confusion, and critical health-related consequences.

- A Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs)-based solution that retrieves data from trusted sources to fact-check and verify content can help to solve this problem.

- This methods delivers a scalable and efficient system, which is capable of handling large volumes of data and preventing misinformation.

Problem Statement

- The rapid spread of misinformation on social media, combined with dynamic and evolving content, challenges traditional detection methods. So, need of automated misinformation identification is in need to counter misinformation as fast as possible.

- Existing detection methods, reliant on manual efforts or basic keyword filtering, fail to address the complexity and scale of modern misinformation campaigns, which often leverage AI, bots, and data-driven targeting. Therefore there is need of an automated, scalable, and accurate system to identify and mitigate misinformation on social media, thereby preserving the integrity of online communication and safeguarding societal trust.

Literature Review 1

TrumorGPT

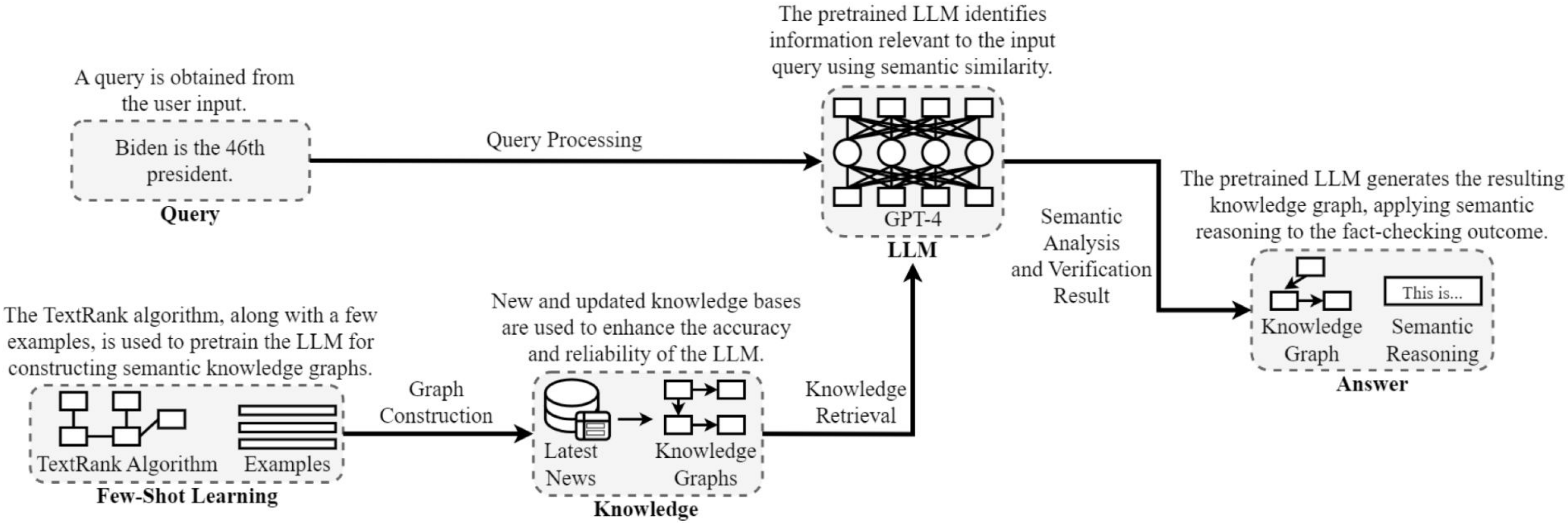

The pretrained LLM identifies information relevant to the input query using semantic similarity.
A query is obtained from the user input.
Biden is the 46th president.
Query
Query Processing
GPT-4
LLM
Semantic Analysis and Verification Result
The pretrained LLM generates the resulting knowledge graph, applying semantic reasoning to the fact-checking outcome.
This is...
Knowledge Graph
Semantic Reasoning
Answer
The TextRank algorithm, along with a few examples, is used to pretrain the LLM for constructing semantic knowledge graphs.
TextRank Algorithm
Examples
Few-Shot Learning
Graph Construction
New and updated knowledge bases are used to enhance the accuracy and reliability of the LLM.
Latest News
Knowledge Graphs
Knowledge
Knowledge Retrieval

Figure 1: The architecture of TrumorGPT, showcasing the workflow from user input to fact verification.

Literature Review 1

- In TrumorGPT technique, graph generation is done with the help of data, having data source as DBpedia.

- Then the relevant graph is searched(community) with the help of similarity search technique and then feed into pre-trained LLM model along with the query.

- By checking the edges and vertex which will lead to sentence generation, it will give result for a given social media post is true or false.

- Accuracy for this approach was 90.5% for the dataset of 100, having 2 labels named as True and False.

Literature Review 2

KG-RAG

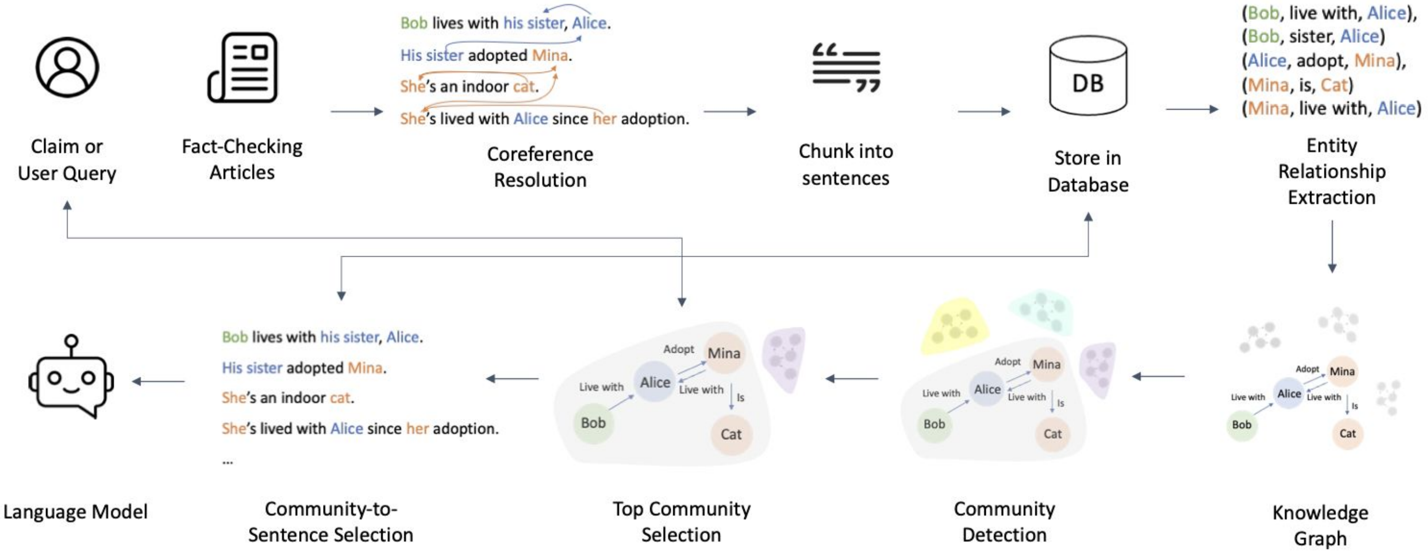


Figure 2: Workflow of Community KG-RAG

Literature Review 2

In KG-RAG technique also same as TumorGPT it tries to use knowledge graph technique, but the data is taken from debunking website like Factcheck, etc

Other then that they are using community based retrieval instead of direct searching with GQL, which first identify top community of graph and then find top relationship among that community based searching is done.

Accuracy for this approach was 56.24% for the dataset of 18,553, having 3 labels named as Supported, Refuted, NEI.

Literature Review 3

Evidence-backed Fact Checking using RAG and Few-Shot In-Context Learning with LLMs

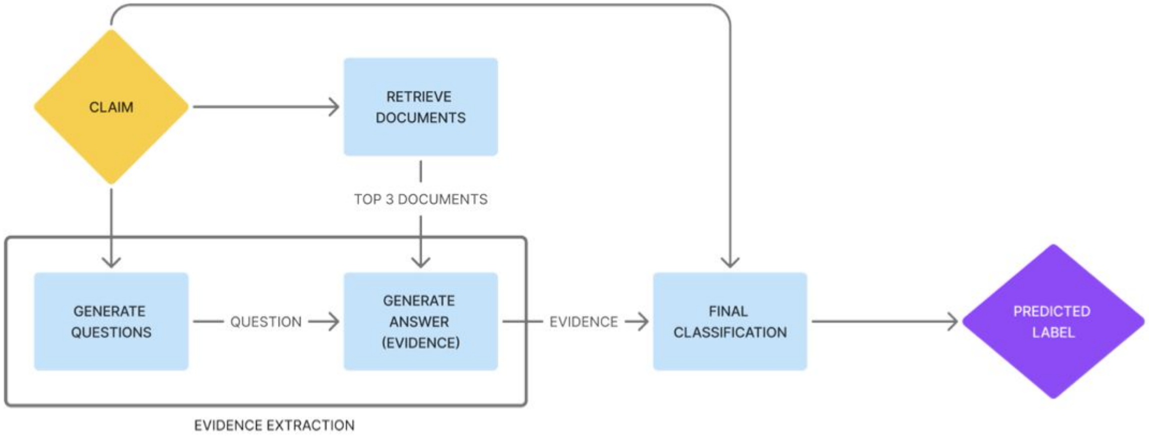


Figure 3: architecture

Literature Review 3

- This methodology take help of basic RAG technique and In-context learning for solving the fact checking problem.

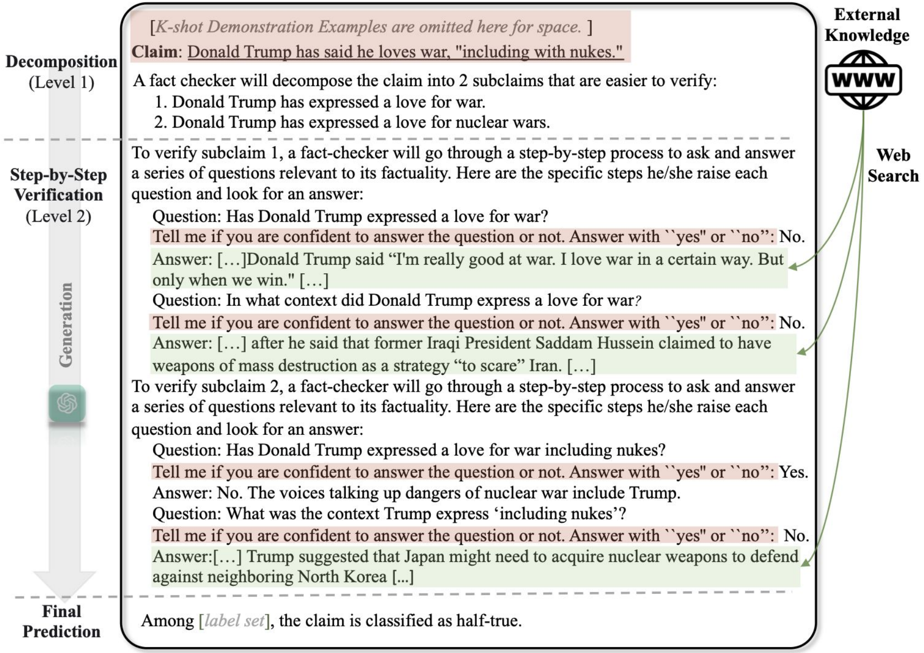
- Here they have used Averitec dataset having multiple labels like Support (S), Refute (R), Conflicting Evidence/Cherrypicking (C), or Not Enough Evidence (N).

- Accuracy for this approach was 63.6% for 500 labeled data.

Literature Review 4

HiSS

- This technique uses subclaim based approach which then checked individually with the help of external search.
- This methodology is completely based on ICL/prompting method, which can be challenging when user input counter statement along with the query for PT-LLM.
- F1 score for this techniques was shown to be 53.9 on RAWFC dataset



[*K-shot Demonstration Examples are omitted here for space.*]
Claim: Donald Trump has said he loves war, "including with nukes."

A fact checker will decompose the claim into 2 subclaims that are easier to verify:
1. Donald Trump has expressed a love for war.
2. Donald Trump has expressed a love for nuclear wars.

To verify subclaim 1, a fact-checker will go through a step-by-step process to ask and answer a series of questions relevant to its factuality. Here are the specific steps he/she raise each question and look for an answer:
Question: Has Donald Trump expressed a love for war?
Tell me if you are confident to answer the question or not. Answer with ``yes" or ``no": No.
Answer: […]Donald Trump said "I'm really good at war. I love war in a certain way. But only when we win." […]
Question: In what context did Donald Trump express a love for war?
Tell me if you are confident to answer the question or not. Answer with ``yes" or ``no": No.
Answer: […] after he said that former Iraqi President Saddam Hussein claimed to have weapons of mass destruction as a strategy "to scare" Iran. […]
To verify subclaim 2, a fact-checker will go through a step-by-step process to ask and answer a series of questions relevant to its factuality. Here are the specific steps he/she raise each question and look for an answer:
Question: Has Donald Trump expressed a love for war including nukes?
Tell me if you are confident to answer the question or not. Answer with ``yes" or ``no": Yes.
Answer: No. The voices talking up dangers of nuclear war include Trump.
Question: What was the context Trump express 'including nukes'?
Tell me if you are confident to answer the question or not. Answer with ``yes" or ``no": No.
Answer:[…] Trump suggested that Japan might need to acquire nuclear weapons to defend against neighboring North Korea [...]

Among [*label set*], the claim is classified as half-true.

Figure 4: HiSS architecture

Experiments

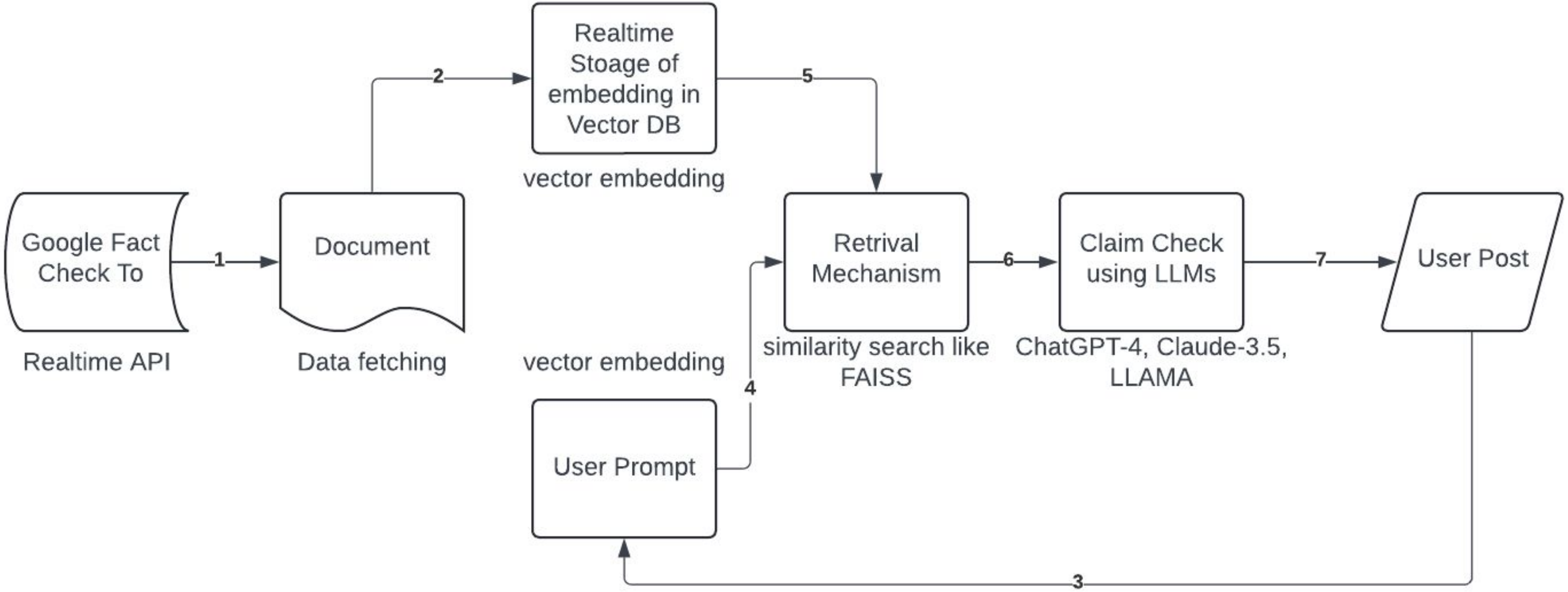


Figure 5: Basic RAG technique

Experiments

- The testing of Politifact data was done to check the issues with RAG and this experiments shows that claims which were marked false are more likely to be predicted as true.
  - Accuracy: 78.00%
  - Precision: 71.88%
  - Recall: 92.00%
  - F1 Score: 80.70%

- The major reason behind this inaccuracy was of astroturfing lie in between true claims.

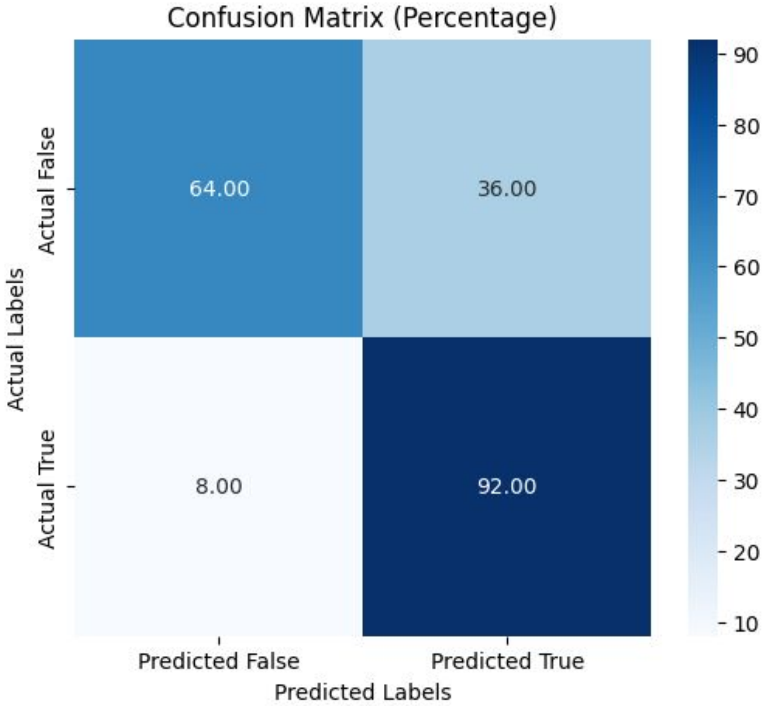


Figure 6: Confusion Matrix

Experiments

- When analysed the data which were having type-I error it was found that most of those claims were using techniques like "astroturfing" or "disinformation sandwiching". Example is as shown below

Nine different subsidies that the U.S. government gives to an industry that makes more money than any other industry, including refunds for drilling costs and refunds to cover the cost of searching for oil. Subsidies for oil and gas companies make up 88 percent of all federal subsidies. Just cutting the oil and gas subsidies out would save the U.S. government $45 billion every year.

- In the above claim all the statement are correct except last one, instead of “every year” its is “ten years”. This kind of techniques are widely used on social media, to counter this Agentic RAG can be helpful.

Research Gap

Overall it has been shown that graphRAG based techniques have shown higher accuracy over normal RAG based techniques.

In the Graph RAG techniques shown above there are two major problems

1. As time passes by the knowledge graph become much bigger and the relevant community detection in the graph become time consuming process.

2. Also, not use of multiple tools can lead to less accuracy on claims based on numerical calculations

3. Lack of early detection of claims, as rely on fact-checking website.

Proposed Methodology

In this method the claim is divided into multiple atomic claims, then the evidence gathering is done with the help of web search tools along with mathematical tools to check that retrieved document is sufficient the aviratic score is generated during training.

Then to Knowledge Graph is generated for the evidence gathered with the help of tool like Neo4j, after that alpha top-community(sub-graph) is identified and then after the relevant beta top-sentences will get identified.

In Last step with the help of ICL technique the top sentence and the user post will be imputed to identify claim as refuted/supported or partial supported.

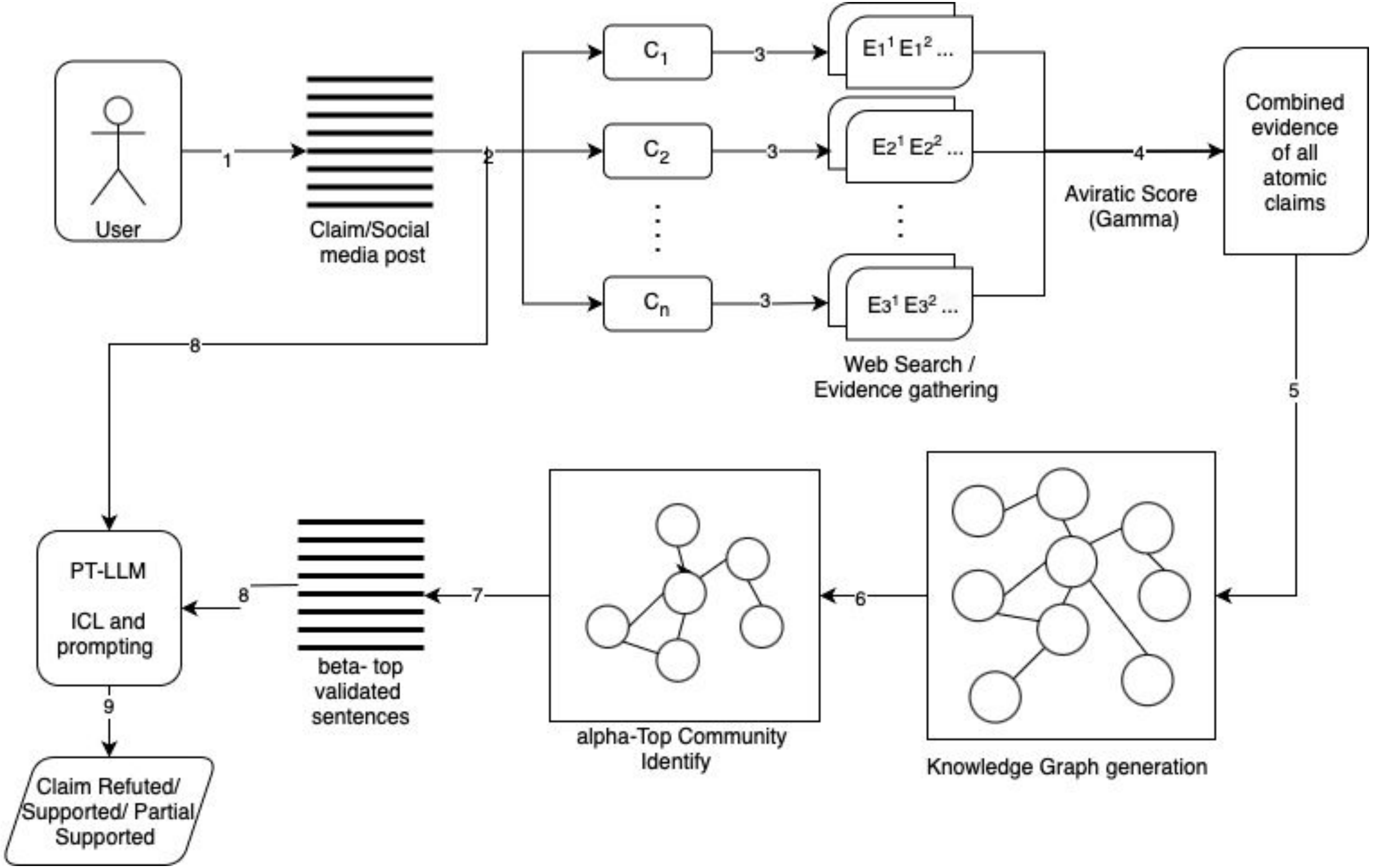

User
1
Claim/Social media post
2
C_1
C_2
C_n
3
3
3
E_1^1 E_1^2 ...
E_2^1 E_2^2 ...
E_3^1 E_3^2 ...
Web Search / Evidence gathering
4
Aviratic Score (Gamma)
Combined evidence of all atomic claims
5
Knowledge Graph generation
6
alpha-Top Community Identify
7
beta- top validated sentences
8
8
PT-LLM
ICL and prompting
9
Claim Refuted/ Supported/ Partial Supported

Figure 7: Work Flow of proposed Methodology

Proposed Methodology

- Example Workflow of a Claim:

Input Claim: "Donald Trump has said he loves war, including with nukes.”

Decomposition:

Sub-claim 1: "Donald Trump has expressed a love for war."

Sub-claim 2: "Donald Trump has expressed a love for nuclear wars."

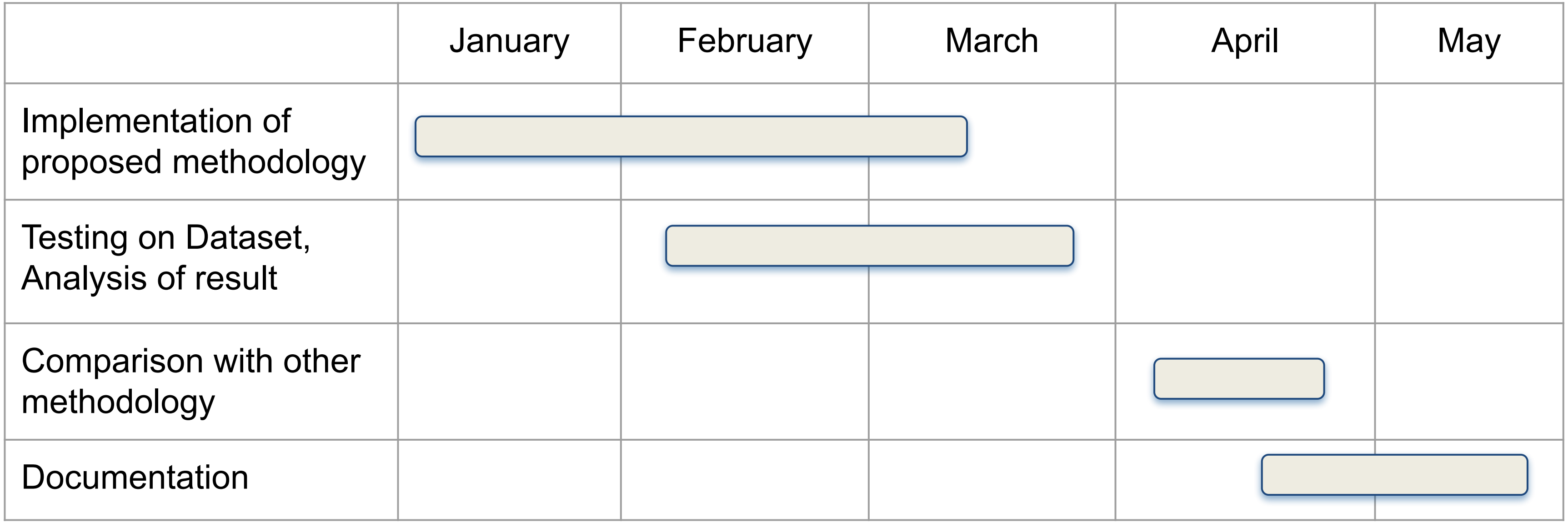
Reasoning by graph:

Donald Trump → President → USA , Donald Trump → loves → war , nuclear wars → Japan , etc

Output:

Combined analysis: "Partial Supported"

Gantt Chart



	January	February	March	April	May
Implementation of proposed methodology	■	■	■		
Testing on Dataset, Analysis of result		■	■		
Comparison with other methodology				■	
Documentation				■	■

Figure 8: Gantt Chart

Reference

[1]. Choi EC, Ferrara E. Fact-gpt: Fact-checking augmentation via claim matching with llms. InCompanion Proceedings of the ACM on Web Conference 2024 2024 May 13 (pp. 883-886).

[2]. Zhang Y, Sharma K, Du L, Liu Y. Toward Mitigating Misinformation and Social Media Manipulation in LLM Era. InCompanion Proceedings of the ACM on Web Conference 2024 2024 May 13 (pp. 1302-1305).

[3]. Singal R, Patwa P, Patwa P, Chadha A, Das A. Evidence-backed Fact Checking using RAG and Few-Shot In-Context Learning with LLMs. InProceedings of the Seventh Fact Extraction and VERification Workshop (FEVER) 2024 Nov (pp. 91-98).

[4]. Chang RC, Zhang J. CommunityKG-RAG: Leveraging Community Structures in Knowledge Graphs for Advanced Retrieval-Augmented Generation in Fact-Checking. arXiv preprint arXiv:2408.08535. 2024 Aug 16.

Reference

[5]. Hang, Ching Nam, Pei-Duo Yu, and Chee Wei Tan. "TrumorGPT: Query Optimization and Semantic Reasoning over Networks for Automated Fact-Checking." 2024 58th Annual Conference on Information Sciences and Systems (CISS). IEEE, 2024.

[6]. Zhang X, Gao W. Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method. InProceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) 2023 Nov (pp. 996-1011).

Thank you