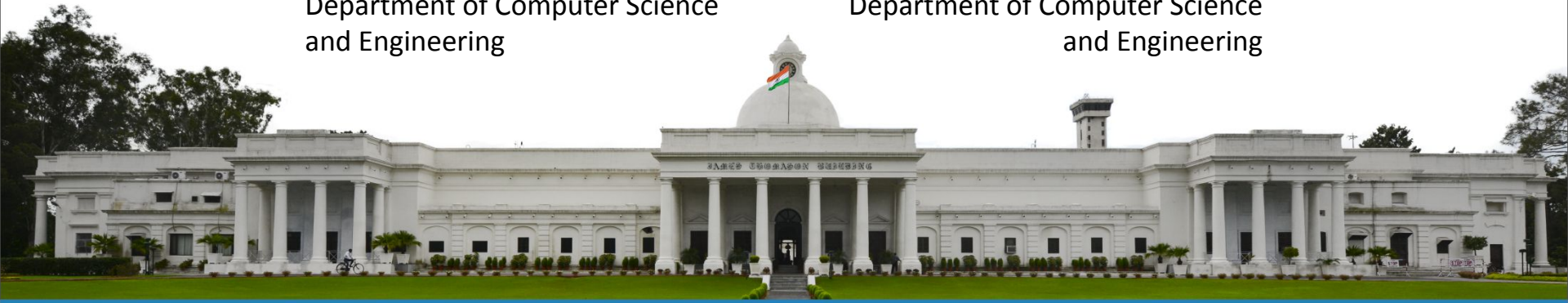
IIT ROORKEE

Misinformation Identification: An Agentic RAG based Technique

Submitted by,
Jimmy Aghera
Department of Computer Science and Engineering

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



Contents

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.
- Delivers a scalable and efficient system capable of handling large volumes of data, preventing misinformation and fostering trustworthy, accurate digital interactions

Motivation

- During the 2020 U.S. elections, 73% of Americans reported encountering claims on social media*
- Misinformation related to COVID-19 vaccines led to widespread hesitancy, exacerbating the pandemic's impact globally
- In 79 countries, governments used misinformation campaigns, with 57 countries deploying bots and 76 manipulating media to influence public opinion*

**Pew Research Centers and University of Oxford*

‘Digital platforms must take measures to curb fake news’

The Hindu Bureau
NEW DELHI

Information and Broadcasting Minister Ashwini Vaishnaw on Saturday said fair compensation by online platforms for the content created by conventional media was one of the four major challenges in the sector that needed to be addressed.

Speaking at an event organised by the Press Council of India on National Press Day, Mr. Vaishnaw said, “As we see that the consumption of news is rapidly shifting from the conventional modes to the digital media, the traditional media is losing out financially because of this change.”

“The investment made in creating a team of journalists, training them, having editorial processes, methods to check the veracity of the news, taking the responsibility for the content — all these investments which are huge both in terms of time and money are becoming irrelevant by the way these platforms are having a very unequal edge in terms of bargaining power they have vis-a-vis conventional media,” he said.

He said another key challenge was “fake news” and “disinformation”. Mr.

Ashwini Vaishnaw

Vaishnaw called for accountability in digital media to combat fake news.

The third challenge, he said, was algorithmic bias on platforms. “These algorithms are designed to maximise engagement. Because the engagement defines the revenue, so maximising the revenue becomes the objective of the platform. Unfortunately, these algorithms also tend to prioritise content that incites strong reaction regardless of the factual accuracy,” said the Minister.

He said the fourth challenge was the impact of artificial intelligence (AI) on intellectual property (IP) rights.

“The content produced by creators, musicians, filmmakers, writers, and authors, is being digested by the AI models. What happens to the IP rights of the creators,” he asked.

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

- Techniques like Neural Networks and LSTM-based models, improved accuracy by analyzing sequential data and linguistic patterns. However, they still had significant limitations, particularly when the data need to be verified multi-layered they fails due to vanishing gradient. These models required sequential data input, which made them slower and less efficient for misinformation detection.

- **Limitations:** Short-Range Dependencies, sequential Processing, lack of Global Context

Literature Review

- Open-source LLMs face architectural constraints, making extensive customization challenging. While techniques like LoRA enable fine-tuning, studies[1] show a 63% accuracy in some cases, which remains insufficient for real-world misinformation detection.

- Real-time training for misinformation detection is impractical due to high computational demands, retrieval-augmented methods offer a viable alternative.

- **Limitations:** Fine-tuning requires labeled datasets and also high computational resources. Its reliance on static data limits adaptability to evolving misinformation patterns, and retraining with every new dataset is impractical for real-world applications.

Literature Review

- In TrumorGPT technique, graph generation is done with the help of data, having data source as DBpedia.
- Then the relevant graph is searched 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 it will give result for a given social media post is true or false.

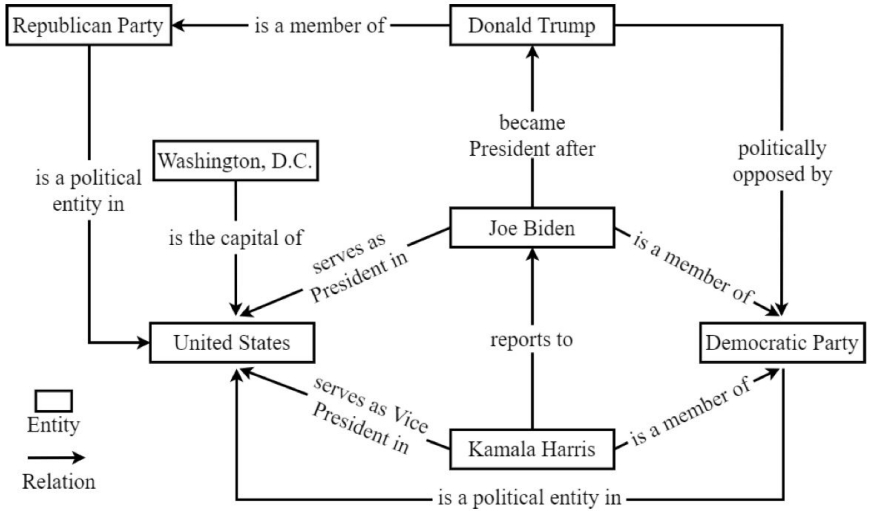


Figure 1: An illustrative semantic knowledge graph, highlighting key political figures and relationships in the United States[5]

Literature Review

TrumorGPT

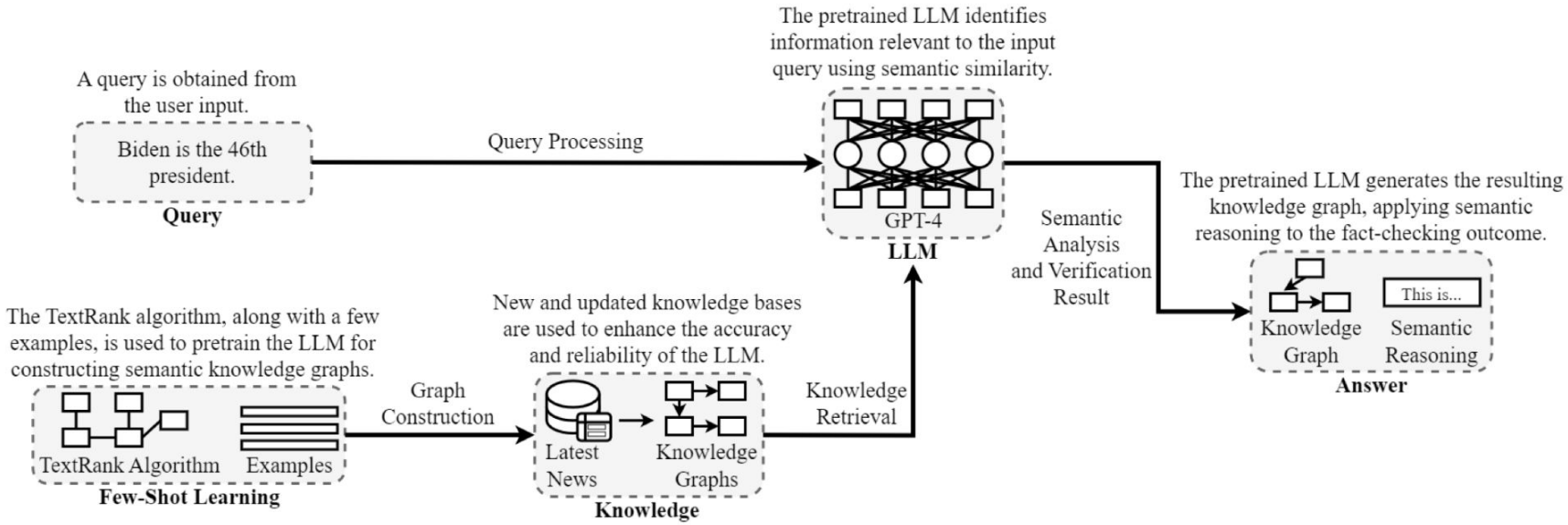


Figure 2: The architecture of TrumorGPT, showcasing the workflow from user input to fact verification[5].

Research Gap

- Traditional machine learning and RNN-based techniques are insufficient for this problem as they struggle with understanding the context of language and generating quick outputs for longer sentences.

- Fine-tuning approaches, though useful, rely on models trained on limited datasets that are not frequently updated, leading to reduced accuracy, as noted in the literature.

- While RAG-based techniques have shown superior performance, some, like TrumorGPT, lack evidence-based fact-checking, making them less trustworthy for users.

- Additionally, most RAG-based methods depend on top-K retrieval techniques, which can compromise the reliability of the generated answers.

Proposed Methodology

- The proposed approach leverages a vector database like Chroma to replace traditional knowledge graphs, enabling faster and more efficient retrieval using Locality Sensitive Hashing (LSH).

- Unlike static datasets like Wikipedia, this method fetches false claims from fact-checking websites such as Logically, Boom, Ptinews, and Altnews.

- For each user query (e.g., a social media post), a similarity search retrieves relevant information from the vector database. These retrieved chunks are injected into a pre-trained LLM like GPT-4, which, using prompt engineering and in-context learning (ICL), assesses the factual accuracy of the query. If no relevant or conclusive data is found in the vector database, an LLM agent is employed to address the query.

Proposed Methodology

- If relevant chunks are not found in the vector database, an agent will be triggered to search for new data across multiple sources using LLM tools. One of these tools will include the vector database itself, ensuring faster retrieval during the search process.

- The agent will also be activated in cases where the generated answer is ambiguous due to contradictory chunks or partial claim matching. Unlike a single pipeline, the agent's iterative functionality allows it to continuously seek conclusive evidence by re-querying until a definitive result is obtained, which is essential when the query does not align with any pre-existing vector database chunks.

- Additionally, all results fetched, whether from web searches or the vector database, include source links in their metadata, supporting evidence-based fact-checking.

Proposed Methodology

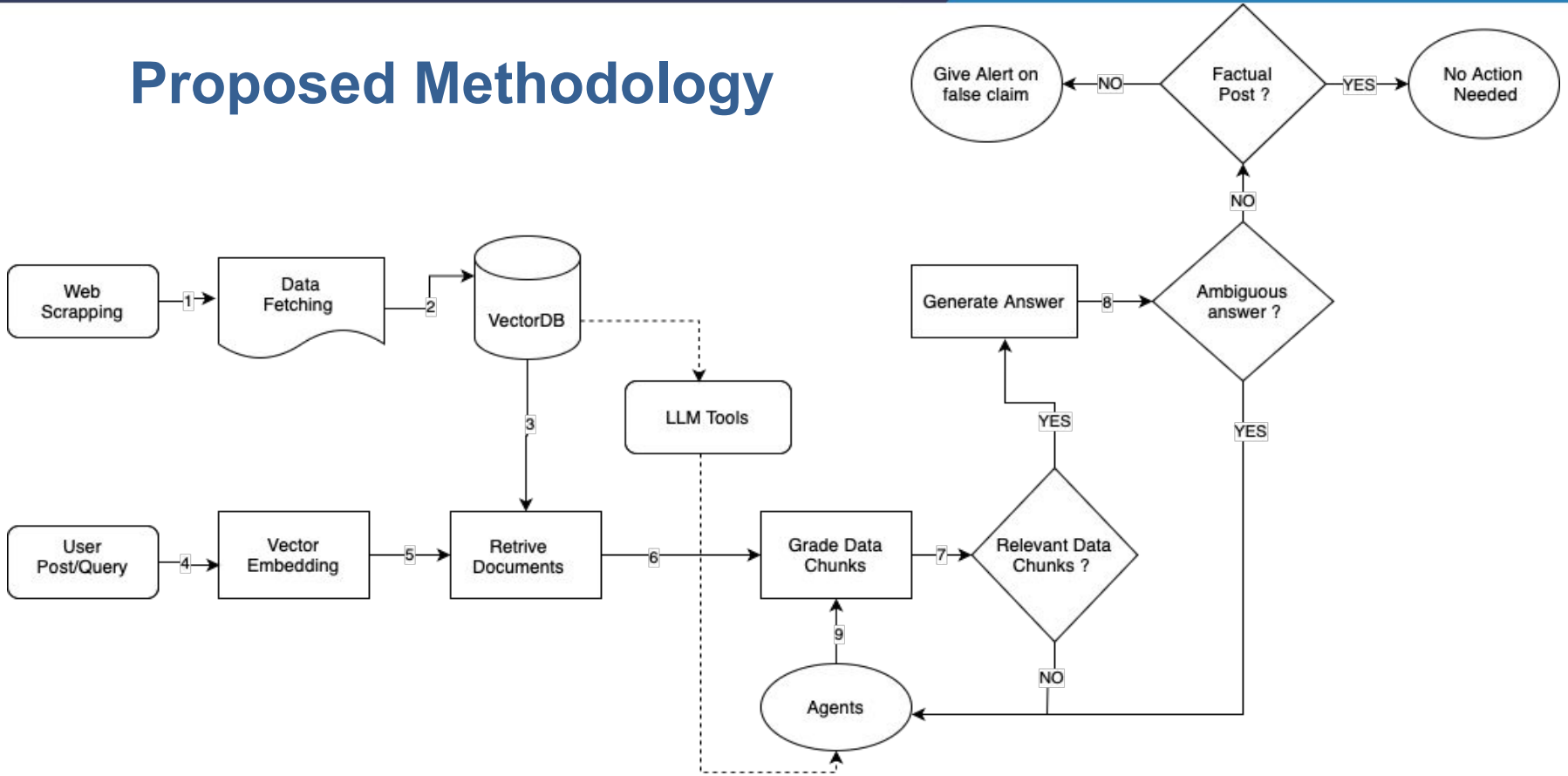


Figure 3: Workflow of Agentic RAG approach to detect misinformation

Proposed Methodology

- Example Workflow of a Claim by Agentic RAG:

Input Claim: "The average global temperature in 2023 was the highest ever recorded."

Decomposition:

Sub-claim 1: "What was the average temperature in 2023?"

Sub-claim 2: "Was this the highest ever recorded?"

Search Queries and Results:

Result: "1.2°C higher than pre-industrial levels in 2023."

Result: "Previous records show 2016 and 2019 were the hottest years before 2023."

Reasoning:

Sub-claim 1 resolved as True: "2023 temperature: 1.2°C higher."

Sub-claim 2 resolved as True: "Higher than 2016 and 2019."

Output:

Combined analysis: "True statement"

Dataset Used

- Data is scraped from (http://logicallyfacts.com).
- Current data have attributes as follows:
  - **Topic :** string
  - **Author :** string
  - **Date :** date
  - **Category :** enum('climate', 'conspiracies', 'economics', 'events', 'geopolitical', 'health', 'human-rights', 'media', 'politics', 'tech', 'conflict', 'sports')
  - **Context :** string
  - **Verdict_Status :** enum('Fact-Check', 'Unknown')
  - **Verdict :** enum('False', 'Misleading', 'Partly_True', 'True', 'Fake')
  - **Link :** url
- The Data size as of now is **8475** entries.

Dataset Analysis

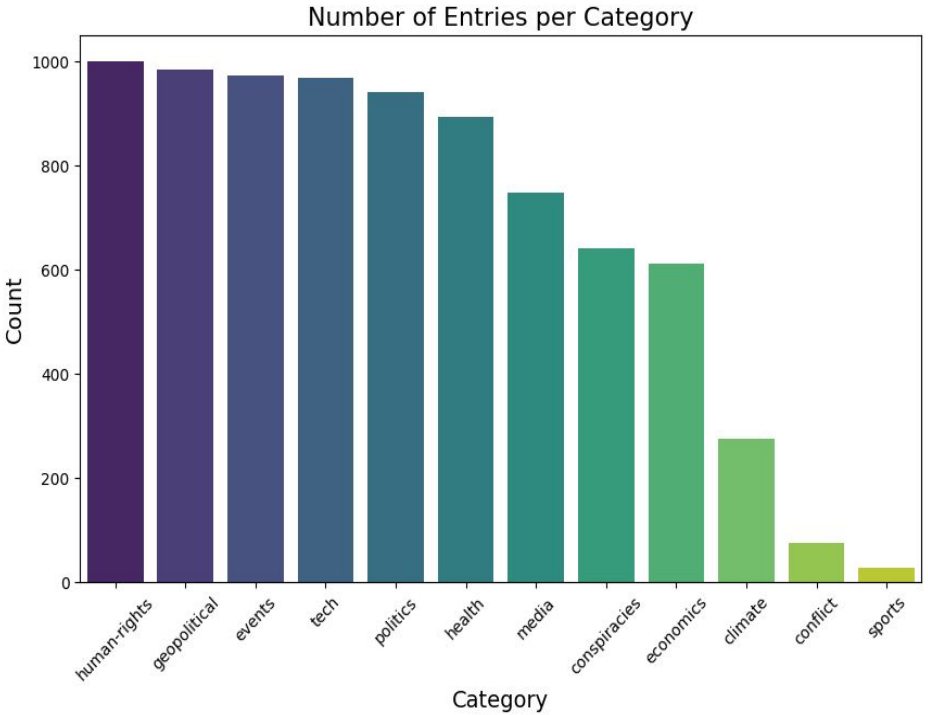


Figure 4: Categories

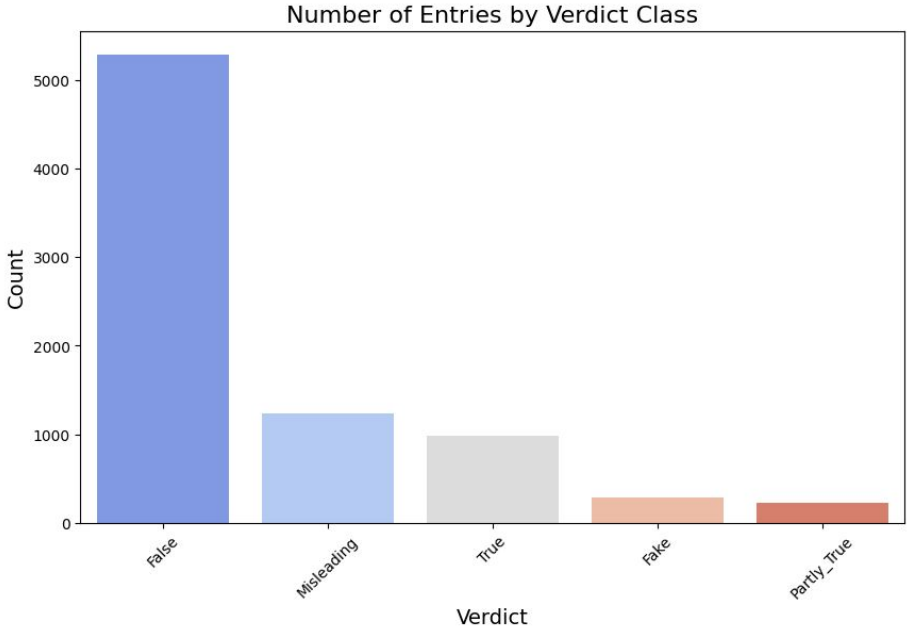


Figure 5: Verdict Classes

Dataset Analysis

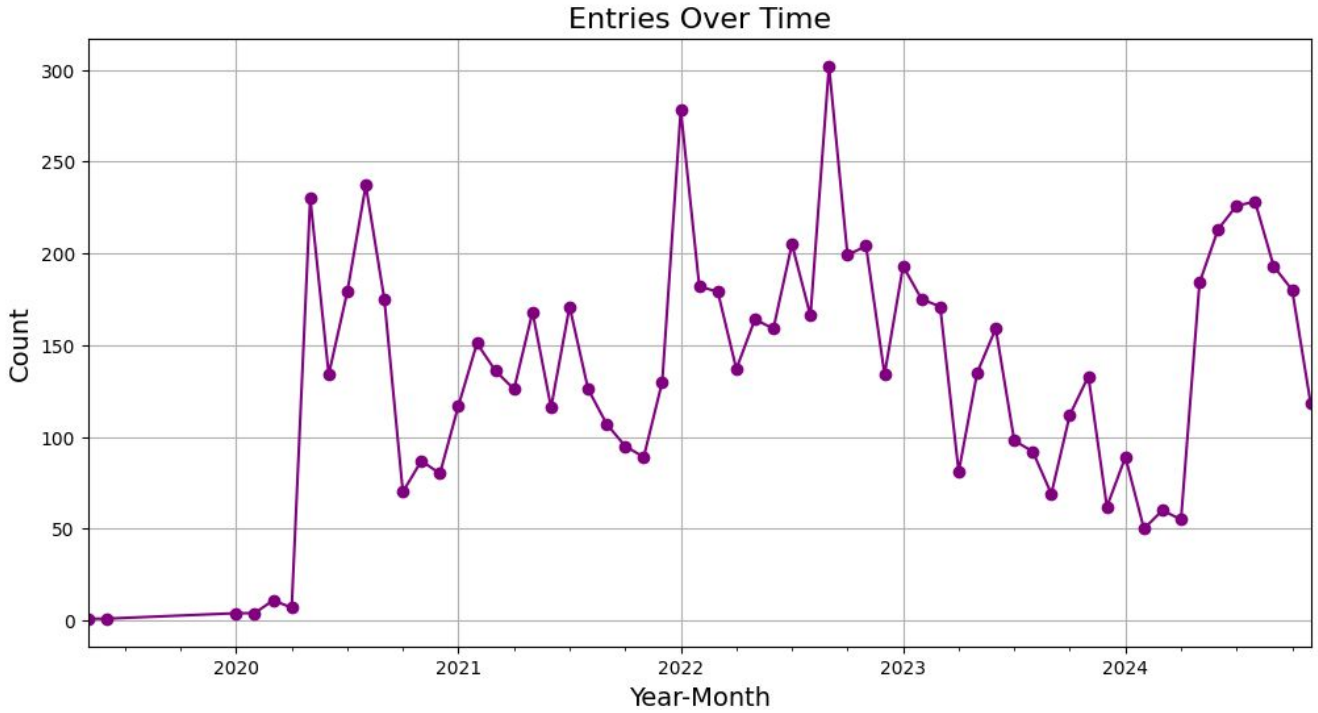


Figure 6: Factual Data timeline analysis

Experiments

- The testing of Politifact data was done to check the issues with RAG over Agentic 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%

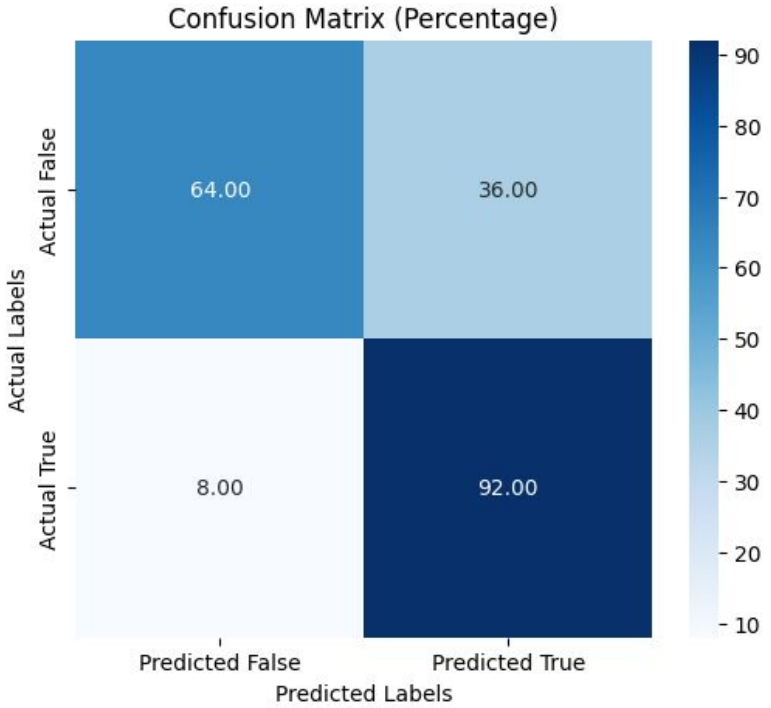


Figure 7: Confusion Matrix

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.

Gantt Chart

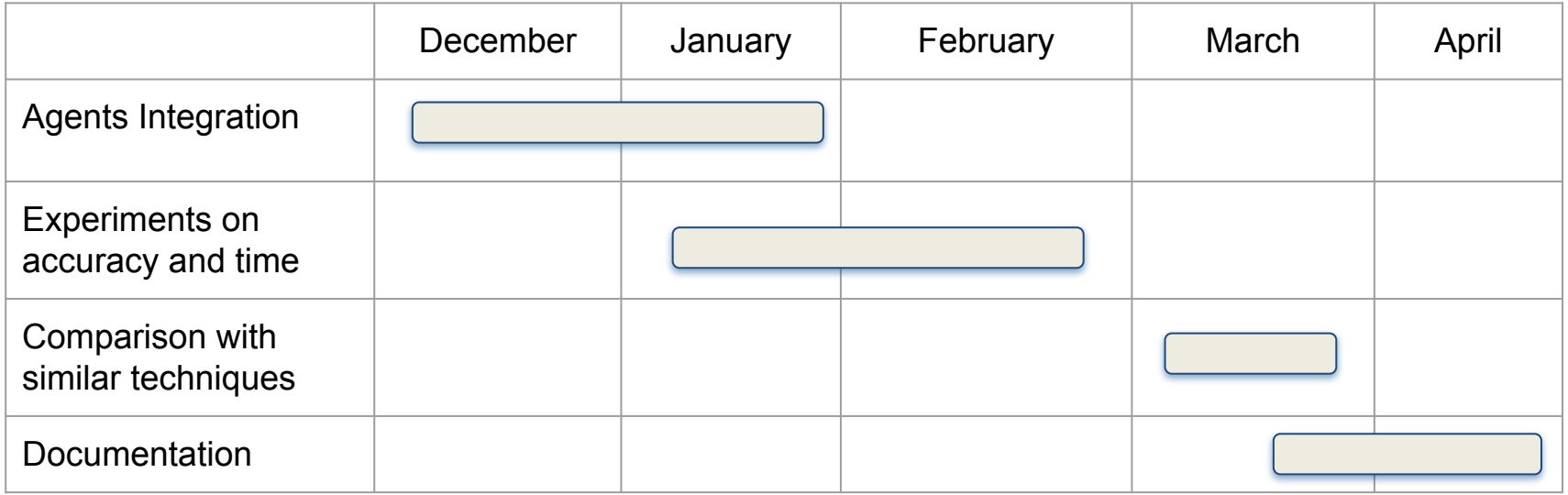


Figure 8: Gantt Chart

Conclusion and Future Work

- Conduct experiments to compare time and accuracy between the RAG framework and the proposed Agentic RAG approach. Also measure the improvement in accuracy over baseline algorithms to demonstrate the effectiveness of the proposed system.

- Integrate multilingual and image-processing agents to address misinformation in diverse languages and visual content.

- This project demonstrates a novel approach to misinformation detection by combining RAG with LLM-based agents. It successfully leverages fact-checking datasets and query mechanisms to debunk claims more effectively. The proposed methodology enhances adaptability to new data sources and ensures robust evidence-based fact-checking.

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]. Surendran, Pranav, et al. "Covid-19 fake news detector using hybrid convolutional and Bi-lstm model." 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, 2021.

[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.

Thank you