50.045 Information Retrieval Project

Team News Verifiers

Chirag Shivakumar (1004996)

Yap Zhan Hao, Sean (1005153)

Varshini (1005185)

Win Tun (1005265)

1 Introduction

1.1 Background

In today's world, where fake news is a growing concern, it's important to have tools that can tell real news from false. This report describes our work in creating a system that uses large language models (LLM) to find and verify news. We focused specifically on news from Singapore to keep our project manageable and relevant to a specific area.

Our main tool is the Retrieval-Augmented-Generation (RAG) model (refer to section 8 for the pipeline code) . This model is a mix of traditional information retrieval methods and new, advanced language-processing techniques. It looks at news stories and works out if they are true. By focusing on Singapore news, we made a system that understands local news better and is more effective in this area.

The goal of our project is to help in the fight against fake news and to show how such systems can work well for local news. In this report, we'll go over how we made the system, the challenges we faced, and how successful it was.

1.2 Problem Statement

How can we develop an LLM-based information retrieval system, specifically focusing on Singaporean news, that accurately verifies the authenticity of news content and combat the spread of fake news?

1.3 Definitions

Considering our problem statement, we outline several constraints. First, we assume that queries can take the form of either questions or statements relating to Singapore news stories. Secondly, when posed with a question, the IR system should provide a relevant and accurate response that addresses the question. Last but not least, when the query is a statement, the system should provide evidence to support or refute the

claim. For cases where no relevant information can be found, the system should refuse to respond to the query, or indicate that its response may be unreliable.

2 Dataset

2.1 Data Formatting

For our dataset, we used articles from The Straits Times to serve as a proxy for reliable news stories from Singapore. This will ensure that when the data retrieved is sent to the LLM, the LLM output will be based on accurate information and will be less-prone to hallucinations.

Since there are no existing datasets containing only Singapore related news articles, we obtained our articles by web scraping. We managed to find an existing code that scraped Straits Times articles from its sitemap (Ari's Scrapeyard), and adapted it to suit our needs.

We focused on collating recent Singapore related news dating from 2022 to present. Hence, any article url with the word 'Singapore' inside of it will be scraped. For each article, we recorded 4 data fields: url, datetime, headline and article (Fig.1)

	url	datetime	headline	article
0	https://www.straitstimes.com/singapore/citizen...	2022-04-06T15:54:56+08:00	Citizen archivist: Making historical records m...	"The newly launched Citizen Archivist Project ...
1	https://www.straitstimes.com/business/companie...	2022-04-05T12:58:38+08:00	China fillip put Singapore share investors in ...	"SINGAPORE - The local market started the week...
2	https://www.straitstimes.com/singapore/red-lio...	2022-08-10T12:02:25+08:00	Red Lions controversy: 5 things to know about ...	"In an embarassing U-turn, it has since droppe...
3	https://www.straitstimes.com/singapore/remembe...	2023-10-03T16:38:53+08:00	Remembering Lee Kuan Yew: 'The greatest C...	"On his regular visits to Hong Kong, Mr Lee Ku...
4	https://www.straitstimes.com/singapore/my-five...	2022-01-16T11:30:16+08:00	My five hours in the Padang queue which is Mr ...	"On Sunday night, I found some old YouTube vid...

Fig 1. Dataset format

2.2 Index Creation

We indexed these documents using Haystack's Elasticsearch DocumentStore class. Elasticsearch is a fast and scalable distributed search engine with full-text search

capabilities. It utilizes the inverted index as its underlying data structure. At the same time, it is straightforward to implement with Haystack.

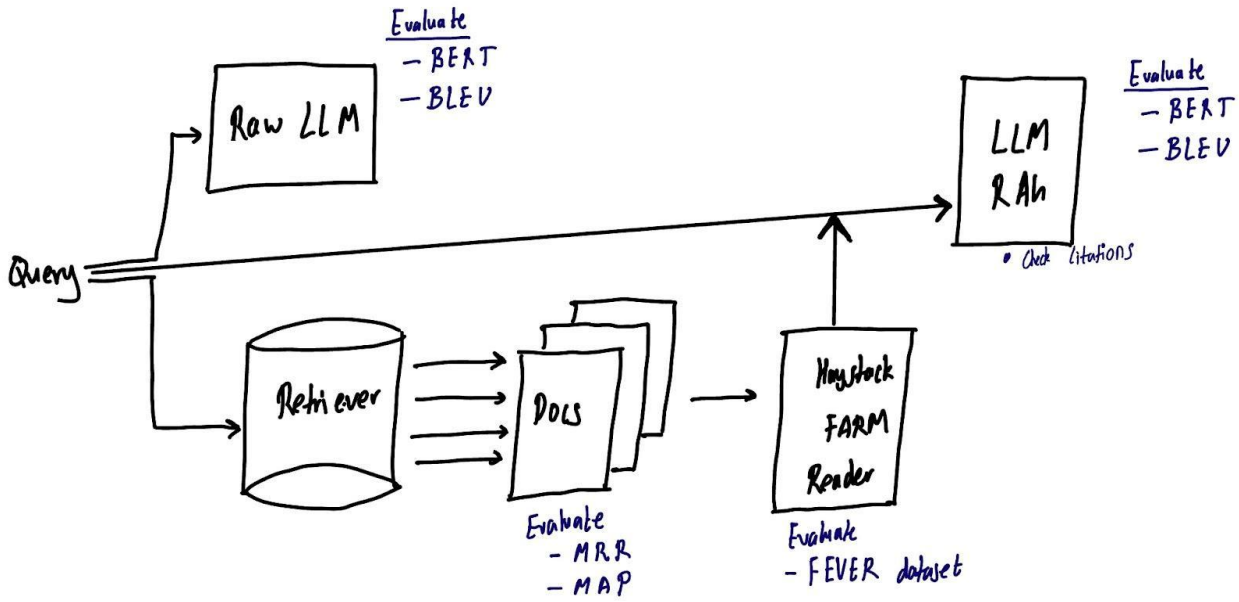
2.3 Test Questions

Using a subset of our dataset of news articles, we constructed a set of question-answer pairs to serve as the queries and gold answers for evaluating our LLM-supported information retrieval system. The table in Appendix A displays these test questions.

3 Methodology

3.1 Overview

The project leverages a Retrieval-Augmented Generation (RAG) model, integrating information retrieval and language model generation. It has been implemented using the Haystack framework. The focus is on verifying the truthfulness of news content, with a specific emphasis on Singapore.



3.2 Retrieval Techniques

We considered three different retriever systems to retrieve documents from the Elasticsearch DocumentStore, namely BM25, dense embeddings, and TF-IDF.

3.2.1 Overview

BM25

BM25 is a computationally efficient ranking function commonly used to score the relevance of documents to a given query. It utilizes the following formula.

$$\sum_{i}^{n} IDF(q_i) \frac{f(q_i, D) * (k1+1)}{f(q_i, D) + k1 * (1 - b + b * \frac{fieldLen}{avgFieldLen})}$$

With reference to the above,

$qi = the\ ith\ query\ term$

$f(qi,\ D) = the\ frequency\ of\ the\ term\ qi\ in\ document\ D$

$IDF(qi) = the\ inverse\ document\ frequency\ of\ qi$

$fieldLen = the\ length\ of\ document\ D$

$avgFieldLen = the\ average\ document\ length$

$k1 = 1.2\ (variable)$

$b = 0.75\ (variable)$

Where k1 is a variable which limits how much a single query term can affect the score of a given document, and b is another variable which controls the effect of the length of document D compared to the average document length.

Embedding Retriever

A transformer model is used to encode documents and the query. These document embeddings are then stored in a vector database and retrieved based on its similarity with the query embedding.

TF-IDF Retriever

$$Score = tf * idf$$

$$tf = no.\ of\ words\ in\ the\ query\ that\ occur\ in\ that\ document$$

$$idf = inverse\ of\ the\ fraction\ of\ documents\ containing\ the\ word$$

The tf-idf retriever works on the assumption that a document that has a greater overlap with the query words is more relevant and words that occur fewer times across the collection are more significant. A tf-idf score is computed for each document and the documents with a higher score are then retrieved.

3.3 Using a Reader for Document Digests

As the documents in our dataset are complete news articles, it is impractical to use the unprocessed documents for retrieval-augmented generation. Thus, we needed to select a method to efficiently obtain representative digests of documents. To do so, we used the FARMReader module from Haystack to extract keywords and contexts from documents using an encoder-based model. Assuming that the keywords were extracted with a high confidence score, the contexts for the extracted keywords serve as informative digests of the retrieved documents, as they contain the high-confidence keywords, as well as the words close to the extracted words.

For the FARMReader, we used the roberta-base-squad2 model, a version of the RoBERTa model that has been fine-tuned using the SQuAD2.0 dataset for the task of question-answering. It is a popular model that can interpret queries in the form of questions, and will enable us to extract relevant information from retrieved documents.

3.3 Generation Using LLM

Vicuna-7b-1.5 is an auto-regressive language model trained by fine-tuning Llama 2 on user-shared conversation data collected from ShareGPT. It has better conversational ability than its predecessor, which works well with our objective as we intend to pose questions to it and verify user shared news content. Vicuna can also easily be run on smaller GPUs due to its compact model size, making it a suitable option for our pipeline.

4 Evaluation

There are two goals for our evaluation. Firstly, the initial outputs of the LLM are verified against the retrieved documents from the retriever. Following which, our retriever is evaluated using selected metrics.

4.1 Verification

To verify the generated outputs of the LLM, we used the same LLM with a few-shot verification prompt to determine if a prediction was verifiable. We feed in the retrieved documents' context window, along with the initial query and LLM output to that query, to the LLM. Following which, we ask the LLM to verify if its initial output was correct.

```
def get_prompt_context(question, context,prediction):
  prompt = f"""
    Question: {prompt_queries[0]}
    Context: {prompt_context[0]}
    Answer: The President of Singapore was Halimah Yacob
    Given the above question and context, is the answer correct? Please return Yes or No.
    Yes

    Question:  {prompt_queries[0]}
    Context: {prompt_context[0]}
    Answer: 'Nicole Seah'
    For the above question, is the answer correct? Please return Yes or No.
    No

    Question: {question}
    Context: {context}
    Answer: {prediction}
    For the above question, is the answer correct? Please return Yes or No.
    """
  return prompt
```

4.2 Retriever Evaluation

To evaluate the retrievers, we assigned relevance data to all news articles for each test question using boolean retrieval. The effectiveness of the retriever is then assessed using metrics like Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP).

4.2.1 Assigning Relevance Data

The following figure illustrates how relevance data is assigned to documents for a particular test question.

```
# What's the new scheme for Covid-19 swabs in Singapore?
df.loc[contains(df, 'covid') & contains(df, 'swab'), 'rel_5'] = 1.0
```

Given a query, we identify the keywords and indicate that a document is relevant to the query if it has all the identified keywords. With these relevance indicators, we can evaluate the retrievers by their mean reciprocal rank and mean average precision.

$$MRR = \frac{1}{N}\sum_{q=1}^{N}\frac{1}{K^{(q)}}$$

$$K^{(q)} = rank\ position\ of\ first\ relevant\ document\ retrieved\ for\ query,\ q$$

$$N = no.\ of\ queries$$

$$MAP = \frac{1}{K}\sum_{k=1}^{K} Prec@k$$

$$Prec@k = \%\ of\ relevant\ documents\ in\ top\ k\ ranks$$

$$K = rank\ threshold$$

Doing so, we obtain the following results in the table below. From this, we concluded that BM25 retrieval performed the best for our use case. Thus, we selected this for our retrieval system in the final pipeline.

Retriever	Mean Reciprocal Rank	Mean Average Precision
TF-IDF	0.5174	0.4930
Embedding	0.6759	0.6158
BM25	0.7908	0.7181

4.3 LLM Evaluation

We considered three different metrics to evaluate the outputs of the LLM before and after RAG prompts were used.

1) BLEU score evaluates the similarity of machine-translated text to a set of high quality reference translations.
2) BERTScore leverages the pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity.
3) LLM Verification (section 4.1).

5 Retrieval Augmented Generation (RAG)

We perform RAG with 3 different attempts

5.1 RAG with Keywords

For the first attempt, we inserted only the extracted keywords as context in our augmented prompt. Such an approach is meaningful in our context since the main task of our IR system is to answer questions, or validate statements.

Prompt:

```
Given the context and scores, answer the question in no more than 50 words. Context: {<list of extracted phrases>]};
```

```
Question: {query};
Answer:
```

Sample Output: `The number of people arrested and charged in connection with the OCBC phishing scams is not specified in the given context`

5.2 RAG with Contexts

However, while justified, short keywords may not be sufficient for all possible query types. To account for this, we also considered RAG using the extended contexts for each extracted answer. This provides the LLM with additional semantic information that may be lost when only using keywords.

Prompt:

```
Given the context and scores, answer the question in 100 words or less. Context: {[<list of contexts>]};
Question: {query};
```

Sample Output:

```
Answer: Thirteen people aged between 19 and 22 have been arrested and charged in connection with the OCBC phishing scams.</s>
```

5.3 RAG for Question-Answering with References

As an attempt to provide citations in the RAG responses, we adapted a prompt template from Deepset's PromptHub (PromptHub). First, we attempted to prompt the LLM to generate the headlines and urls for the relevant documents in its output.

However, this provided poor results, as the references were often inexact. This is expected, since Vicuna is a decoder-based model and is thus prone to generating novel text rather than copying exact sequences of text.

To reduce the potential of hallucinations on the references, we instead decided to only prompt the LLM to provide index references within its output, using the document indices. Doing so, we can also validate the references by extracting the numbers from a generated response and comparing them to the indices of the provided documents.

Prompt

```
Create a concise and informative answer (no more than 50 words) for a given question based solely on the given documents. You must only use information from the given documents. Use an unbiased and journalistic tone. Do not repeat text. Cite the documents using Doc[num] notation. If multiple documents contain the answer, cite those documents like 'as stated in Doc[num], Doc[num], etc.'.
If the documents do not contain the answer to the question, say that 'answering is not possible given the available information.'
{[<list of contexts>]}
Question: {query}; Answer:
```

Sample Output: `"Thirteen people aged between 19 and 22 have been arrested for their suspected involvement in the recent spate of scams targeting OCBC Ban. " (Doc[582])</s>`

6 Results

The table below shows the 3 scores for the 20 queries both before and after query augmentation.

Method	BLEU Score	BERT Score	Verification
Raw LLM	0.0366	0.8509	69%

RAG 1	0.0421	0.8716	89.7%
RAG 2	0.0535	0.8656	86.2%
RAG 3	0.0470	0.8524	86.2%

There is a significant improvement across all 3 metrics once query augmentation is applied, demonstrating the effectiveness of the RAG method. The best performing of the 3 RAG methods was the RAG 1 (keyword and score augmentation) as it had higher BERT score and verification scores.

This shows that for this task a shorter, targeted augmentation of information in the query improves performance as compared to sending in larger chunks of context. RAG 3 had the lowest BERT score among the RAG methods. However, this could be attributed to the addition of Doc IDs as citation in its output, which could have led to a lower BERT score.

6.1 Final Output

Once we have the final predicted answer, we organize the final answer in a markdown format such that the document number, document heading and document URL is present. The Steps Include;

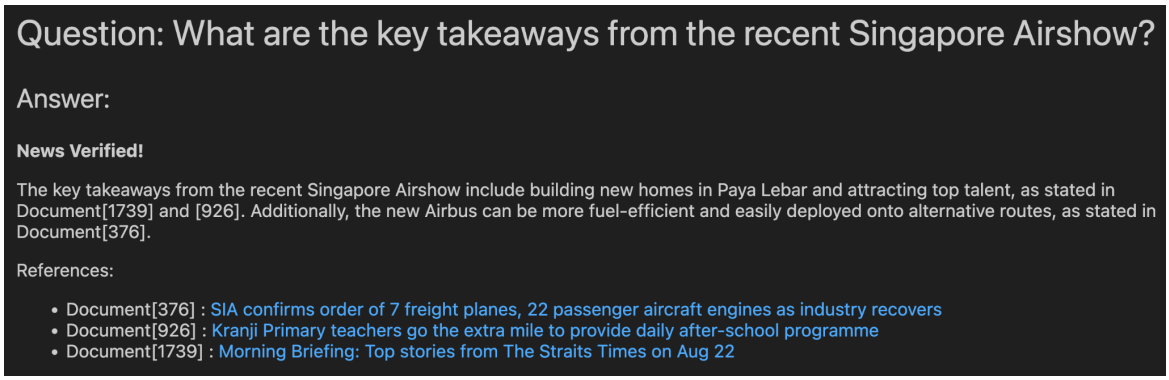
1. We first gather all the news information from our web-scraped dataset, consisting of the News Id, Heading and URL.
2. We check if there is a document number present in the query.
    a. If a document number is present, we output the news to be verified and we add the references below
    b. If there is no document number, we output the news to be unverified.

The Final output format of the News is;

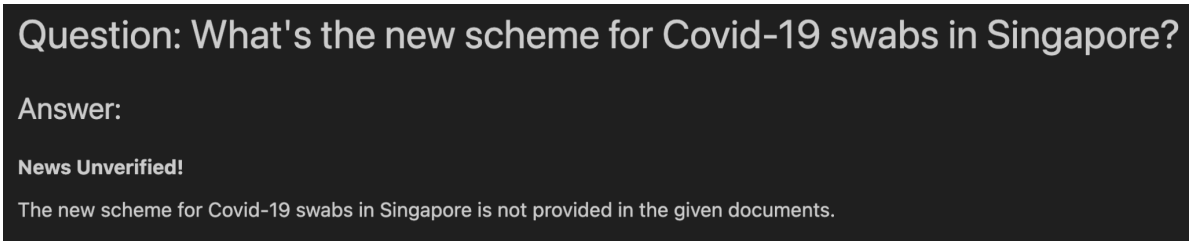
1. News Prediction: Verified or Unverified

2. The Predicted answer
3. References if the news is verified

6.1.1 Verified News



6.1.2 Unverified News



6.1.3 Markdown Format

```
**News Verified!**

The Prime Minister of Singapore is Lee Hsien Loong, as stated in Document[688], Document[1996], and Document[1774].</s>

References:

- Document[688] : [President Halimah to attend official mourning of UAE president in Abu Dhabi ](https://www.straitstimes.com/singapore/president-halimah-to-attend-official-mourning-of-uae-president-in-abu-dhabi)
- Document[1774] : [Brunei&#039;s Sultan Hassanal Bolkiah on two-day state visit to Singapore ](https://www.straitstimes.com/singapore/bruneis-sultan-hassanal-bolkiah-on-two-day-state-visit-singapore)
- Document[1996] : [PM Lee among top leaders to attend state funeral for ex-Japan PM Abe ](https://www.straitstimes.com/asia/east-asia/singapore-pm-lee-among-top-leaders-to-attend-state-funeral-for-ex-japan-pm-abe)
```

7 Conclusion

This News Verification project effectively integrated the Vicuna-7B model and used retrieval techniques like TF-IDF, Embedding, and BM25 Retrievers to enhance the system's capability to verify news content accurately.

A key achievement was the significant increase in news verification accuracy, from 69% to 89.7%, through the implementation of Retrieval-Augmented Generation (RAG) methods, comprising three approaches: extracted keywords/phrases, extracted context, and QA with references. This demonstrates the efficacy of the RAG approach in improving LLM output reliability. Lastly, the project's evaluation framework, using metrics such as BLEU and BERT Score, assessed the performance of its components.

In summary, the Looksy project significantly advances news verification by leveraging LLM's to develop a system that not only accurately identifies and verifies news content, but also shows notable improvement in verification accuracy with RAG-based outputs. This successfully combats fake news and helps setting a strong foundation for future advancements in information verification and retrieval.

7.1 Limitations

Data Bias and Scope: This project primarily focused on Straits Times news articles, which limits the applicability to other global news sources such as Channel News Asia.

Complexity of Fake News: Some sophisticated misinformation might still bypass the system’s verification process, especially if it mimics some factual news styles.

7.2 Future Considerations

Evaluate Retriever with Pseudo Relevance Data: Implementing an evaluation process using pseudo-relevance data can provide insights into the performance of the system, helping to fine-tune retriever accuracy and effectiveness.

8. Code Links

RAG Pipeline Code:
https://colab.research.google.com/drive/1gGm_F7NN6xwK-4aKmFc26ESL_dChx04I?usp=sharing#scrollTo=KmzvTyQzZkq7

Web scraping code:
https://colab.research.google.com/drive/1sfS0PhIDUJ_TtT9SKr1YKBD_H81K0mrH?usp=sharing

Github Repo: https://github.com/shaunnope/paperview/blob/master/extract.ipynb

Appendix A

Question	Answer
Who is the Prime Minister of Singapore?	The Prime Minister of Singapore is Lee Hsien Loong.

What is the top university in Singapore?	The National University of Singapore (NUS) is often ranked as the top university in Singapore.
How did Singapore fare against Japan in the recent badminton championship?	Singapore beat Japan 3-2 in the Asia Team Championships, moving closer to a medal.
What's the new scheme for Covid-19 swabs in Singapore?	A new scheme in Singapore offers 3,500 people free Covid-19 swabs at test centres.
What is the focus of Singapore's Budget 2022?	Budget 2022 aims to chart Singapore's 'new way forward together', focusing on recovery and growth post-pandemic.
What happened in the police confrontation at Clementi?	A knife-wielding man was shot by police during a confrontation at Clementi neighbourhood police centre.
What are the updates on the twin murder case in Upper Bukit Timah?	In the Upper Bukit Timah canal case, the father of the twins found dead received a second murder charge.
What is Singapore's approach to handling the climate crisis as per Budget 2022?	Budget 2022 outlines Singapore's decisive moves to tackle the climate crisis, focusing on sustainable development.

What is the goal of the Citizen Archivist initiative in Singapore?	The Citizen Archivist initiative aims to make historical records more accessible to the public in Singapore.
How did China's economic performance impact Singapore's stock market?	China's economic upturn positively influenced Singapore's share market, boosting investor confidence.
How is Lee Kuan Yew remembered in Singapore?	Lee Kuan Yew is remembered as a pivotal leader in Singapore, often hailed as 'The greatest Chinese outside mainland China'.
How has Singapore's stock market reacted to global economic trends recently?	Singapore's stock market has experienced fluctuations in response to global economic trends and key data from the US and China.
What measures are being taken in Singapore as the dengue peak season approaches?	Singapore is urging residents to be vigilant and reduce mosquito breeding spots as the dengue peak season approaches.
What is the new themed zone planned for Universal Studios Singapore in 2024, and which attraction will it replace?	In 2024, Universal Studios Singapore will introduce a new themed zone called Minion Land, inspired by the "Despicable Me" film franchise. This new attraction will replace the Madagascar themed area, specifically the "Madagascar: A Crate Adventure" ride, which is set to close on

	March 28 to make way for the construction of Minion Land.
What is the purpose of the recent property tax increase in Singapore as per the ST-UOB panel discussion?	The recent increase in property tax in Singapore, discussed at the ST-UOB panel, is intended as a wealth tax rather than a market cooling measure. This tax increase, announced in the Budget, primarily targets higher-end properties and is not expected to affect the majority of Singaporeans living in Housing Board flats. The adjustment is part of a broader effort to tax wealth more effectively, despite the challenges in implementing such a tax.
How many people have been arrested and charged in connection with the OCBC phishing scams?	13 people were arrested over OCBC phishing scams, and 7 of them have been charged.
How might the carbon tax hike affect the cost of living for Singaporean households, despite utility rebates?	Experts suggest that the carbon tax hike can still increase the cost of living for households, even after accounting for utility rebates.
What do fully vaccinated travelers need to know about Singapore's new border measures?	Fully vaccinated individuals planning to travel should look up the latest guidelines as Singapore's new border measures may impact their journey.

Which sectors could see potential wins and losses from Singapore's 2022 Budget?	The 2022 Budget will likely affect various sectors differently, and a detailed analysis is necessary to identify potential winning and losing stocks.
Has public transport ridership in Singapore returned to pre-pandemic levels?	Bus and train ridership in Singapore is on the rise but still hasn't reached the numbers seen before the Covid-19 pandemic.
What service allows Singaporeans to receive government payouts from ATMs?	Singaporeans receiving government payouts via cheques can now utilize the new GovCash service to get payments from ATMs.
How has the attendance at the Pink Dot event changed since it first started?	Since the first Pink Dot event in 2009, attendance has grown significantly, from 2,500 participants to 26,000 the previous year, showing a substantial increase in engagement and support for the LGBT community in Singapore.