

# Bilingual Retrieval-Augmented Generation (RAG) for Social and Political Inquiry

A Fact-Based, Multilingual Solution for Japanese-English News Analysis

Sotarriva Alvarez Isai Roberto

For: Collaboration across STEM and Liberal Arts: AI design concept and technology that supports digital humanities

**Institute of Sciences Tokyo**

# Table of contents



Motivation



Objectives



Workflow



Pre-processing



Query handling



Limitations and future improvements



Conclusion

# Motivation



## Language Barrier in Research

For non-native speakers, generating search queries in Japanese and navigating Japanese-language articles is often slow and imprecise.



## Need for a Seamless Bilingual Research Tool

Traditional search engines require separate queries in each language.



## Fact-Backed Responses

Fact-backed responses ensure that each answer includes **citations and direct links** to original sources, enhancing **transparency and credibility**.

# Objectives



Develop an AI solution for fact-backed responses to questions on social and political topics relevant to Japan.



Ensure AI system provides clickable citations linking to reputable news articles in English and Japanese.



Create a bilingual interface that serves both English and Japanese audiences effectively.



# Fact-aware LLMs

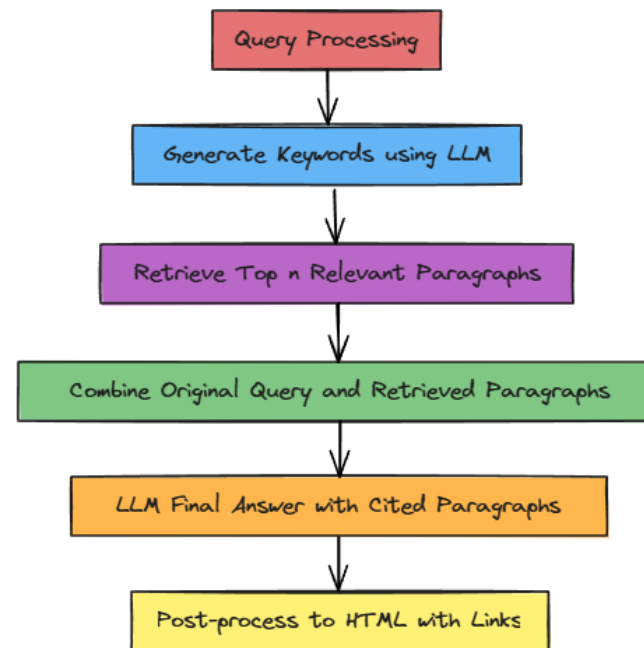
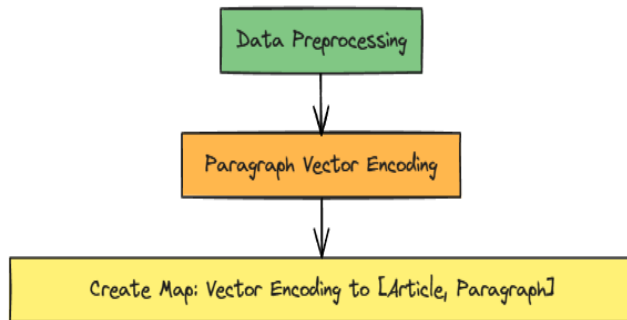
- Models with access to relevant facts are less likely to invent new “facts” to answer a query.
- Links to the original data sources are important for fact checking.

*Image author: DALLE 2024-11-15*

# Retrieval augmented Generation (RAG) implementation

During the query

Database encoding for  
fast article retrieval  
(One time only operation)





# Pre-Query Processing

## Original article

【ワシントン=飛田臨太郎】トランプ次期米大統領が自身の意向に沿わない米軍幹部を一斉に解任することを検討している。複数の米欧メディアが14日までに報じた。軍での女性の職務を制限する案もある。米軍における自身の権威を強める狙いがある。実現に向けて考えに同調するピート・ヘグセス氏を国防長官に起用した。

ロイター通信はトランプ氏の政権移行チームが米軍高官の「解任リスト」を作成したと伝えた。米軍制服組トッ...

## Split into sentences (minimum length 50 Japanese characters or 150 English characters)

【ワシントン=飛田臨太郎】トランプ次期米大統領が自身の意向に沿わない米軍幹部を一斉に解任することを検討している。

複数の米欧メディアが14日までに報じた。軍での女性の職務を制限する案もある。軍での女性の職務を制限する案もある。

米軍における自身の権威を強める狙いがある。実現に向けて考えに同調するピート・ヘグセス氏を国防長官に起用した。

...

## Each sentence is encoded into a vector

[1.02,0.3,...]

[2.4,1.2,...]

[2.2,0.8,...]

# Step 1: Query Initiation and Keyword Generation

- Keyword Extraction: LLM generates a list of keywords from the user's question.
- Purpose: Optimizes article retrieval by narrowing down relevant content.

User query: 'What is the relationship between the USA and Japan?'

## Prompt

Your role is to be a search assistant. Provide a comma separated list of topics or keywords in Japanese related to the query. Example: For the query 'When was the opening ceremony of the Tokyo Olympics?' you could answer '東京、オリンピック、開会式、2020年、開催、日本', or for the query 'Is abortion legal in Japan?' you could answer '中絶、合法、日本'. Return at least one keyword.

Do not provide an explanation or details on the topics, just the keywords or topics since they will be automatically submitted to a search engine.

<start\_of\_turn>user

I want to know about: 'What is the relationship between the USA and Japan?'. What topics or keywords should I search for in the database?<end\_of\_turn><eos>

## Response

USA、日本、関係、同盟、友好、歴史、外交、経済、貿易



# Step 2: Article retrieval

**Keywords are encoded into vectors**

America,  
Security,  
Japan,  
Trump



[2.2,1.0,..],  
[3.2,1.5,...],  
[2.0,5.2,...],  
[5.0,1.2,...]



**Each article is assigned a score for each keyword (cosine similarity)**

[text1\_key1\_s, text2\_key1\_s, ...],  
[text1\_key2\_s, text2\_key2\_s, ...],  
[text1\_key3\_s, text2\_key3\_s, ...],  
[text1\_key4\_s, text2\_key4\_s, ...]



**An overall score for each article is calculated**

$$\text{Score}(\text{text}) = \sum_{k \in \text{Keywords}} \cos\_sim(k, \text{text})$$

[art1\_s, art2\_s, ...]



**The n articles with the highest score are retrieved.**

# Step 3: Prompt creation

Task it to read the article extracts



You are an assistant for a social studies research project. You have been asked to read the following news articles and provide a response to the user's question based on the information in the articles. (The articles are the result of an automatic search based on the user's query, some articles may not be relevant to the user's question)

=====

Provide extracts from the n most relevant articles



UID: @33289/0

Date: 2010-08-09

Source: nikkei.com

Content snippet: いま必要なことは日本の歴史、文化に深く根ざしている提案だと思いますサッカーのワールドカップ南アフリカ大会での日本代表の活躍でもわかるように、欧米が1人の圧倒的なスター(天才)を中心に戦うのに対して、日本は少し秀でたスターと平均的な力のある選手のチームワークによって勝ち抜いたそしてそれこそが日本のすばらしさでないでしょうか

.  
.  
.

=====

Additional instructions



Answer the question on the same language it is written (japanese or english).

Use [UID] to provide citations to the news articles, links in this format will automatically be transformed into hyperlinks to the original articles and transformed to look on the format desired by the user. Example, to cite the article with UID: @123/4, write [@123/4], not [UID:@123/4]. For multiple citations, place each citation in a separate pair of brackets. Example: [@123/4][@456/7] do not write them together like [@123/4, @456/7].

Use HTML tags to format the text. Example: <b>bold text</b>, <i>italic text</i>, <ul><li>unordered list item</li></ul>, <ol><li>ordered list item</li></ol>, <a href='link'>hyperlink</a>

Chat history



<start\_of\_turn>user

'What is the relationship between the USA and Japan?

<end\_of\_turn><eos>

Must follow guidelines

chat

# Step 4: Post-Processing and Output Formatting

```
* **Reconstruction and Economic Cooperation:** Following the economic collapse of the 1990s, both countries saw opportunities for cooperation in areas like technology and manufacturing. This economic relationship provided a platform for a more positive interaction. [UID: @241680/2]
* **Historical Issues:** However, unresolved historical issues like Japan's colonial past in Korea, particularly the forced relocation of Korean people and wartime labor practices, continued to cast a shadow on the relationship. [UID: @225205/2]
* **Diplomatic Tensions:** Political tensions persisted, with accusations and rhetoric often characterizing the relationship. Japan's stance on historical issues and its perceived lack of remorse created friction, particularly regarding issues like the comfort women system. [UID: @43946/0]
```



The relationship between South Korea and Japan has undergone significant changes since 2000, marked by both periods of rapprochement and renewed friction.

## Early 2000s: Reconstruction and Political Tensions:

- **Reconstruction and Economic Cooperation:** Following the economic collapse of the 1990s, both countries saw opportunities for cooperation in areas like technology and manufacturing. This economic relationship provided a platform for a more positive interaction. [\[@241680/2\]](#)
- **Historical Issues:** However, unresolved historical issues like Japan's colonial past in Korea, particularly the forced relocation of Korean people and wartime labor practices, continued to cast a shadow on the relationship. [\[@225205/2\]](#)
- **Diplomatic Tensions:** Political tensions persisted, with accusations and rhetoric often characterizing the relationship. Japan's stance on historical issues and its perceived lack of remorse created friction, particularly regarding issues like the comfort women system. [\[@43946/0\]](#)

- UID references are converted into clickable links.
- **\*\*?** are converted into bold
- \* ? Are converted into unordered lists.
- Line breaks \n are converted to HTML line breaks<br>

# Example Query

News App

How has the relationship between South Korea and Japan evolved since the year 2000?

The relationship between South Korea and Japan has undergone significant changes since 2000, marked by both periods of rapprochement and renewed friction.

**Early 2000s: Reconstruction and Political Tensions:**

- **Reconstruction and Economic Cooperation:** Following the economic collapse of the 1990s, both countries saw opportunities for cooperation in areas like technology and manufacturing. This economic relationship provided a platform for a more positive interaction. [\[ @241680/2 \]](#)
- **Historical Issues:** However, unresolved historical issues like Japan's colonial past in Korea, particularly the forced relocation of Korean people and wartime labor practices, continued to cast a shadow on the relationship. [\[ @225205/2 \]](#)
- **Diplomatic Tensions:** Political tensions persisted, with accusations and rhetoric often characterizing the relationship. Japan's stance on historical issues and its perceived lack of remorse created friction, particularly regarding issues like the comfort women system. [\[ @43946/0 \]](#)

**Late 2000s: Rising Cooperation and Complexity:**

- **Increased economic interdependence:** Continued economic cooperation and collaboration between Japan and South Korea found new ground, expanding into areas such as technology transfer and trade. [\[ @241680/2 \]](#)
- **ASEAN forum:** Participation in the ASEAN forum provided greater interaction and facilitated diplomatic discourse on issues of mutual interest, even as they continued to express discomfort about political decisions. [\[ @147864/0 \]](#)

Enter your query:

How has the relationship between South Korea and Japan evolved since the year 2000?

Submit

Clear Chat

# Checking the sources

The screenshot displays two overlapping windows from a web application. The top window, titled 'News App', contains a text input field with the query: 'How has the relationship between South Korea and Japan evolved since the year 2000?'. The bottom window, titled 'nan', shows the results of a search. It includes a '(No title)' header, a date of '2011-10-14', and a source 'sankei.jp.msn.com'. A snippet of a Japanese article is visible, mentioning the relationship between the two countries. Below the snippet, there is a list of sources with bullet points: 'Increased economic interdependence' and 'ASEAN forum'. At the bottom of the 'nan' window, there is a search bar with the same query, a 'Submit' button, and a 'Clear Chat' button.

News App

How has the relationship between South Korea and Japan evolved since the year 2000?

nan

(No title)

Date: 2011-10-14

Source: sankei.jp.msn.com

【ワシントン＝犬塚陽介】オバマ米大統領と韓国の李明博大統領は13日の米韓首脳会談で、軍事同盟の結束を経済分野にも広げ、米韓両国の「新時代」を築き出すことで一致した。李政権はアフガニスタンや核拡散防止など、オバマ政権の中核となる外交政策で積極的に米国と連携し、オバマ大統領の厚い信頼を得ている。鳩山政権下で冷えついた日米関係の修復に時間を割く日本をよそに、韓国が著しい台頭を遂げたことで、相対的な日本の存在感の低下も懸念されている。

- **Increased economic interdependence:** Continued economic cooperation and collaboration between Japan and South Korea found new ground, expanding into areas such as technology transfer and trade. [[@241680/2](#)]
- **ASEAN forum:** Participation in the ASEAN forum provided greater interaction and facilitated diplomatic discourse on issues of mutual interest, even as they continued to express discomfort about political decisions. [[@147864/0](#)]

Enter your query:

How has the relationship between South Korea and Japan evolved since the year 2000?

Submit

Clear Chat

# Handling Special Cases

KEYWORD GENERATION FAILURE->ORIGINAL QUERY USED AS FALLBACK FOR RETRIEVAL.

```
[2024-11-15 00:37:35,105] [DEBUG] [rag - find_similar_articles:81] Querying the chatbot for keywords...
[2024-11-15 00:37:35,105] [DEBUG] [chatbot - query:76] Prompt: Your role is to be a search assistant. Provide a comma separated list of topics or keywords in Japanese related to the query. Example: For the query 'When was the opening ceremony of the Tokyo Olympics?' you could answer '東京、オリンピック、開会式、2020年、開催、日本', or for the query 'Is abortion legal in Japan?' you could answer '中絶、合法、日本'. Return at least one keyword.
Do not provide an explanation or details on the topics, just the keywords or topics since they will be automatically submitted to a search engine.
<start_of_turn>user
I want to know about: 'How has the relationship between South Korea and Japan evolved since the year 2000?'. What topics or keywords should I search for in the database?<end_of_turn><eos>
<start_of_turn>model

[2024-11-15 00:37:41,557] [DEBUG] [rag - find_similar_articles:83] Keywords response:
[2024-11-15 00:37:41,557] [WARNING] [rag - find_similar_articles:86] No keywords found. Using the query as keywords. This might lead less accurate results.
[2024-11-15 00:37:41,557] [DEBUG] [sentence_vectors - find_similar:168] Finding similar articles...
[2024-11-15 00:37:42,994] [DEBUG] [sentence_vectors - find_similar:172] the database has 313225 articles
```

NO RELEVANT ARTICLES->NO CITATIONS ARE GENERATED IF NO PARAGRAPHS ARE RELEVANT.

## ## 竹島問題の複雑な背景と現在

竹島問題は、韓国と日本の歴史的、政治的争点のひとつで、両国の関係にとって大きな複雑さを生み出しています。

### 過去の歴史:

- **歴史的根源:** 日本と韓国の歴史は、明治維新、第二次世界大戦、そして朝鮮戦争など、複雑な関係を形成してきました。竹島は、戦後、韓国と日本の両国が領有権を主張する争点となりました。
- **日本の歴史:** 日本は、第二次世界大戦中に、韓国の侵略や強制移住の責任を認めていない場合があります。また、日本は、朝鮮戦争中に、韓国の人々を苦しめてきた歴史的誤解も抱えています。
- **韓国の歴史:** 韓国は、日本が韓国の人々を苦しめてきた歴史的誤解を明確に認識しています。

### 現在における状況:

- **領有権の主張:** 日本と韓国は、竹島の領有に関する法的根拠を示すことは難しいです。両国とも、竹島の歴史と文化的な関連性を含めた主張を行っています。
- **国際社会への影響:** 竹島問題は、両国の関係だけでなく、国際社会の関心を集めています。国際機関や地域諸国は、両国の意見を尊重しながら、対話を維持する必要があります。

# Limitations and Future Enhancements

- The LLM model used (Google Gemma) has a strong tendency to forget or ignore instructions on larger prompts (1000+ tokens)
- The news dataset is flawed. Many articles are incomplete, and some has garbage, access to a better-quality dataset is needed.
- Running a local LLM (like google Gemma) is a resources demanding task.

Inference on CPU takes several minutes, even hours.

Inference on CUDA takes seconds (in RTX3070 takes ~10-15 seconds per query).



# Conclusion

- The Bilingual RAG system is a powerful AI solution that provides fact-based, bilingual responses to complex social and political queries related to Japan.
- Key Strengths:** Accurate, citation-backed answers; seamless bilingual support for both Japanese and English users; enhances transparency and trust through direct source linking.
- Impact:** Facilitates informed decision-making and understanding by bridging language barriers and ensuring access to reputable information.

# Settings

- DATA source:

[https://www.kaggle.com/datasets/vyuholl/japanese-newspapers-20052021?select=japanese\\_news.csv](https://www.kaggle.com/datasets/vyuholl/japanese-newspapers-20052021?select=japanese_news.csv)

- Sentence transformer for encoding:

<https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1>

- LLM model:

- Model: **gemma-2-2b-jpn-it**

- Variant: **2b-v2**

- Weights from:

<https://www.kaggle.com/models/google/gemma-2-2b-jpn-it/>