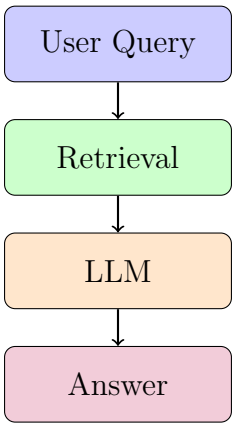
BÁO CÁO PHÂN TÍCH HỆ THỐNG RAG (Retrieval-Augmented Generation)

Enhanced RAG System for BBC News Dataset



Project: My RAG

Author: Trần Mạnh Hùng

Ngày 31 tháng 12 năm 2025

Mục lục

1 Tổng Quan Project **3**
1.1 RAG là gì? . . . 3
1.2 Mục đích của Project . . . 3
1.3 Công nghệ sử dụng . . . 3

2 Kiến Trúc Hệ Thống **4**
2.1 Tổng quan kiến trúc . . . 4

3 Phân Tích Chi Tiết Từng Phần **5**
3.1 Phần 1: Loading Libraries . . . 5
3.2 Phần 2: Weaviate Client & Data Loading . . . 5
3.2.1 Kết nối Weaviate . . . 5
3.2.2 Cấu trúc dữ liệu BBC News . . . 5

4 Các Kỹ Thuật Retrieval **7**
4.1 Metadata Filtering . . . 7
4.2 Semantic Search . . . 8
4.3 BM25 Search . . . 9
4.4 Hybrid Search . . . 10
4.5 Semantic Search với Reranking . . . 11
4.6 So sánh các phương pháp Retrieval . . . 12

5 Prompt Engineering & Auto-Tuning **13**
5.1 Query Classification . . . 13
5.2 Dynamic Parameter Selection . . . 13

6 Phân Tích Kết Quả Demo **15**
6.1 Demo 1: Technical Query . . . 15
6.2 Demo 2: Creative Query . . . 16
6.3 Demo 3: Reranking . . . 16
6.4 Demo 4: Fact Check . . . 17
6.5 Demo 5: Topic Summary . . . 18

7 Phân Tích Chi Tiết Kết Quả So Sánh Các Phương Pháp Search **19**
7.1 Kết Quả Chi Tiết Từng Phương Pháp . . . 19
7.1.1 Semantic Search . . . 20
7.1.2 Semantic Search with Reranking . . . 20
7.1.3 BM25 Search . . . 21
7.1.4 Hybrid Search . . . 22
7.1.5 Without RAG (Baseline) . . . 23
7.2 Bảng So Sánh Tổng Hợp . . . 23
7.3 Phân Tích Chất Lượng Sources Retrieved . . . 24
7.4 Key Insights từ So Sánh . . . 24
7.4.1 Insight 1: Semantic Search vs BM25 . . . 24
7.4.2 Insight 2: Giá trị của Reranking . . . 24
7.4.3 Insight 3: RAG vs No-RAG - The Critical Gap . . . 25

7.5 Kết Luận Phân Tích So Sánh 25

8 Kết Luận & Đánh Giá Hiệu Quả **27**
8.1 Bằng chứng về hiệu quả của RAG 27
8.2 Tổng kết hiệu quả các kỹ thuật 27
8.3 Limitations & Considerations 28
8.4 Best Practices 28

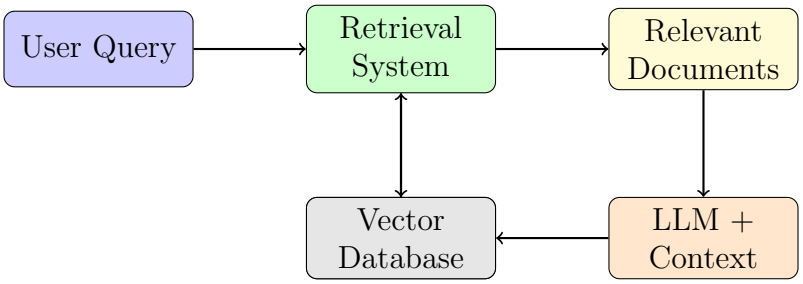
9 Tóm Tắt Cuối Cùng **29**
9.1 What This Project Demonstrates 29
9.2 Key Takeaways 29
9.3 When to Use What 29

1 Tổng Quan Project

1.1 RAG là gì?

RAG (Retrieval-Augmented Generation) là một kỹ thuật tiên tiến trong xử lý ngôn ngữ tự nhiên, kết hợp hai thành phần chính:

- **Retrieval (Truy xuất)**: Tìm kiếm thông tin liên quan từ cơ sở dữ liệu hoặc kho tài liệu
- **Generation (Sinh văn bản)**: Sử dụng Large Language Model (LLM) để tạo câu trả lời dựa trên thông tin đã truy xuất



Hình 1: Workflow tổng quan của hệ thống RAG

1.2 Mục đích của Project

Project này xây dựng một hệ thống RAG hoàn chỉnh với các mục tiêu:

1. Trả lời câu hỏi dựa trên **bộ dữ liệu BBC News**
2. So sánh hiệu quả của **nhiều phương pháp retrieval** khác nhau
3. Áp dụng **Prompt Engineering** để tối ưu câu trả lời
4. Tự động điều chỉnh tham số dựa trên **loại câu hỏi**

1.3 Công nghệ sử dụng

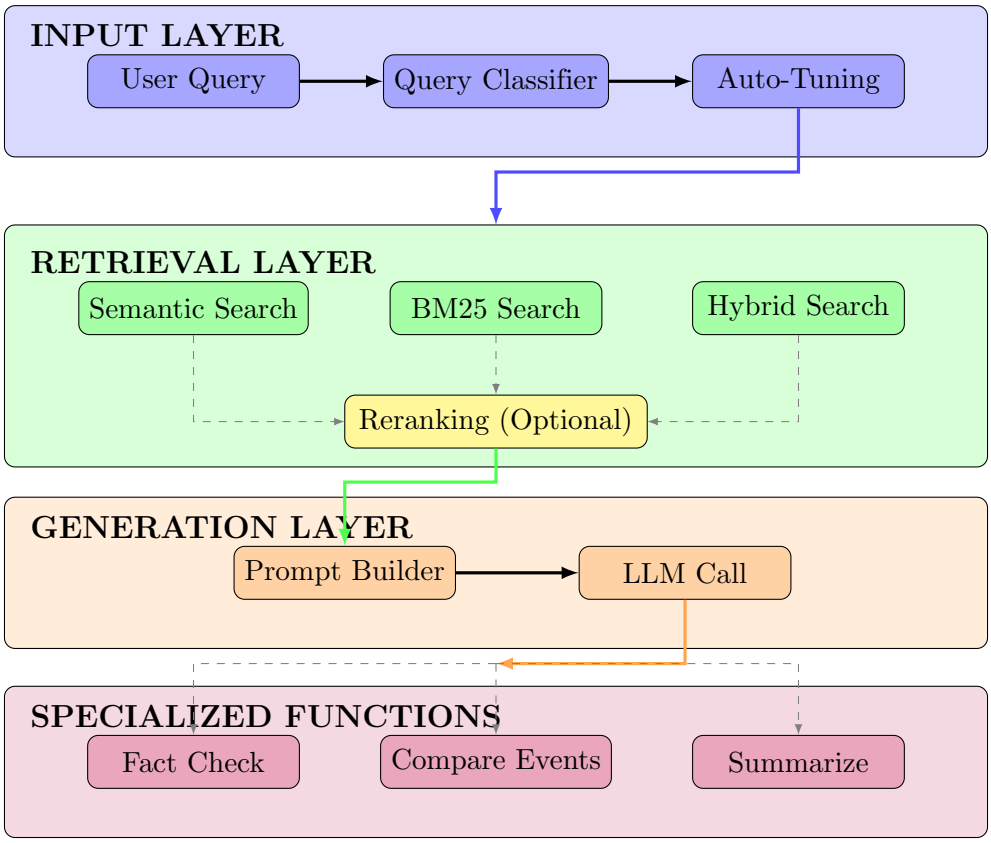
Bảng 1: Các công nghệ chính trong project

Thành phần	Công nghệ	Mục đích
Vector Database	Weaviate	Lưu trữ và tìm kiếm vector embeddings
LLM	Flask API (custom)	Sinh câu trả lời
Data Validation	Pydantic	Định nghĩa schema cho output
Data Format	Joblib	Load dữ liệu BBC News

2 Kiến Trúc Hệ Thống

2.1 Tổng quan kiến trúc

Hệ thống Enhanced RAG được thiết kế theo kiến trúc đa tầng (multi-layer architecture):



Hình 2: Kiến trúc đa tầng của Enhanced RAG System

3 Phân Tích Chi Tiết Từng Phần

3.1 Phần 1: Loading Libraries

```
[1]: import joblib
import weaviate
from weaviate.classes.query import (
    Filter,
    Rerank
)
```

```
[2]: import flask_app
import weaviate_server
from utils import (
    generate_with_single_input,
    print_object_properties,
    display_widget
)
import unittests
```

```
 * Serving Flask app 'flask_app'
 * Debug mode: off
```

```
You're using a XLMRobertaTokenizerFast tokenizer. Please note that with a fast tokenizer, using the `__call__` method is faster than using a method to encode the text followed by a call to the `pad` method to get a padded encoding.
```

```
[3]: from pydantic import BaseModel, Field
from typing import List, Optional, Literal
import json
```

Bảng 2: Mô tả các thư viện

Library	Chức năng
`joblib`	Load file dữ liệu đã serialize (. joblib)
`weaviate`	Client để kết nối với Weaviate vector database
`Filter`	Lọc kết quả theo metadata
`Rerank`	Sắp xếp lại kết quả theo độ liên quan
`pydantic`	Định nghĩa schema cho structured output

3.2 Phần 2: Weaviate Client & Data Loading

3.2.1 Kết nối Weaviate

```
[4]: client = weaviate.connect_to_local(port=8079, grpc_port=50050)
```

- Kết nối đến Weaviate server đang chạy local
- Sử dụng gRPC (port 50050) cho hiệu suất cao hơn
- HTTP API trên port 8079

3.2.2 Cấu trúc dữ liệu BBC News

```
[5]: bbc_data = joblib.load('data/bbc_data.joblib')
```

```
[6]: print_object_properties(bbc_data[0])
```

```
article_content: Justin Welby speaks on BBC Radio 4's Today programme as part of a special show guest edited by Dame ...(truncated)
description: The Archbishop of Canterbury urges politicians to "forswear wedge issues" and avoid divisive topics.
guid: https://www.bbc.co.uk/news/uk-67844356
link: https://www.bbc.co.uk/news/uk-67844356?at_medium=RSS&at_campaign=KARANGA
pubDate: 2024-01-01 00:00:04
title: Justin Welby: Political leaders should treat opponents as human beings
```

Bảng 3: Cấu trúc dữ liệu BBC News

Field	**Mô tả**	**Ví dụ**
`title`	Tiêu đề bài báo	"Taylor Swift breaks record..."
`pubDate`	Ngày xuất bản	"2024-06-28T10:30:00Z"
`guid`	ID duy nhất	"bbc-12345"
`link`	URL bài báo	"https://bbc.com/..."
`description`	Tóm tắt ngắn	"Pop star performs..."
`article_content`	Nội dung đầy đủ	Full article text

4 Các Kỹ Thuật Retrieval

4.1 Metadata Filtering

[10]:
```python
def filter_by_metadata(metadata_property: str,
                       values: list[str],
                       collection: "weaviate.collections.collection.sync.Collection" ,
                       limit: int = 5) -> list:
    """
    Retrieves objects từ một collection được chỉ định dựa trên các tiêu chí metadata filtering.

    Hàm này truy vấn một collection trong client được chỉ định để lấy các objects khớp với các tiêu chí metadata nhất định. Nó sử dụng một filter để tìm các objects có 'propert

    Args:
    metadata_property (str): Tên của metadata property dùng để lọc.
    values (List[str]): Một danh sách các giá trị để đối chiếu với property đã chỉ định.
    collection_name (weaviate.collections.collection.sync.Collection): Collection để thực hiện truy vấn.
    limit (int, optional): Số lượng objects tối đa được truy xuất. Mặc định là 5.

    Returns:
    List[Object]: Một danh sách các objects từ collection khớp với các tiêu chí lọc.
    """

    # Retrieve using collection.query.fetch_objects

    response = collection.query.fetch_objects(
        filters = Filter.by_property(metadata_property).contains_any(values),
        limit=limit)


    response_objects = [x.properties for x in response.objects]

    return response_objects
```

Giải thích Metadata Filtering

Mục đích: Lọc document dựa trên thuộc tính cụ thể (không dùng vector)
Ví dụ: Tìm tất cả bài báo có tiêu đề chứa "Taylor Swift"
Khi nào dùng: Khi biết chính xác từ khóa cần tìm

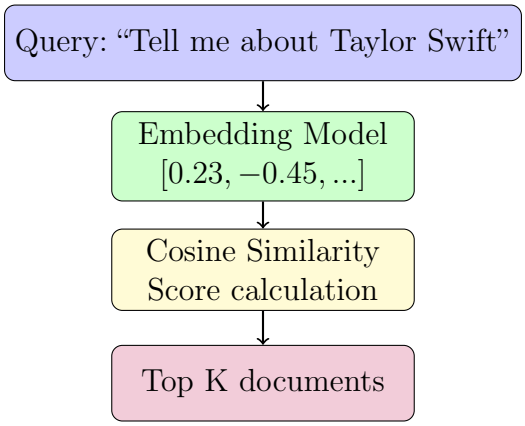
Test thử:

[11]:
```python
# Example
res = filter_by_metadata('title', ['Taylor Swift'], collection, limit = 2)
for x in res:
    print_object_properties(x)
```

```
article_content: The 2024 awards season kicked off in style at the Golden Globes - the first major red carpet event o...(truncated)
chunk: some of his previous get-ups. The Bear's Jeremy Allen White - who recently became the new face (and ...(truncated)
chunk_index: 4
description: Stars including Margot Robbie and Taylor Swift arrived in a variety of eye-catching outfits.
link: https://www.bbc.co.uk/news/entertainment-arts-67908727?at_medium=RSS&at_campaign=KARANGA
pubDate: 2024-01-08 03:23:58+00:00
title: Margot Robbie, Taylor Swift and more on Golden Globes red carpet

article_content: The 2024 awards season kicked off in style at the Golden Globes - the first major red carpet event o...(truncated)
chunk: headpiece - not entirely a fashion choice. She says the "protective veil" is because she hurt her fa...(truncated)
chunk_index: 5
description: Stars including Margot Robbie and Taylor Swift arrived in a variety of eye-catching outfits.
link: https://www.bbc.co.uk/news/entertainment-arts-67908727?at_medium=RSS&at_campaign=KARANGA
pubDate: 2024-01-08 03:23:58+00:00
title: Margot Robbie, Taylor Swift and more on Golden Globes red carpet
```

4.2 Semantic Search

```
[13]: def semantic_search_retrieve(query: str,
                                   collection: "weaviate.collections.collection.sync.Collection" ,
                                   top_k: int = 5) -> list:
    """
    Thực hiện một semantic search trên một collection và truy xuất các chunks liên quan nhất.

    Hàm này thực thi một truy vấn semantic search trên một collection được chỉ định để tìm các text chunks có độ liên quan cao nhất với đầu vào 'query'. Phép tra cứu truy xuất

    Args:
    query (str): Truy vấn tìm kiếm được sử dụng để tìm các text chunks liên quan.
    collection (weaviate.collections.collection.sync.Collection): Collection nơi phép semantic search được thực hiện.
    top_k (int, optional): Số lượng các relevant objects hàng đầu cần truy xuất. Mặc định là 5.

    Returns:
    List[str]: Một danh sách các text chunks có độ liên quan cao nhất với truy vấn đã cho.
    """

    # Retrieve using collection.query.near_text
    response = collection.query.near_text(query=query, limit=top_k)


    response_objects = [x.properties for x in response.objects]

    return response_objects
```



Hình 3: Quy trình Semantic Search

Ưu điểm của Semantic Search

- Hiểu được **ý nghĩa** của câu hỏi (semantic meaning)
- Tìm được documents liên quan ngay cả khi **không có từ khóa trùng khớp**
- Xử lý tốt synonyms và paraphrases

Nhược điểm của Semantic Search

- Có thể bỏ sót documents có **exact keyword match**
- Phụ thuộc vào chất lượng của embedding model

Test thử:

[14]:
```
# Let's have an example!
print_object_properties(semantic_search_retrieve(query = 'Tell me about the last Taylor Swift show', collection = collection, top_k = 2))
```

```
article_content: Taylor Swift has finished the European leg of her Eras Tour with a record-breaking show at Wembley S...(truncated)
chunk: size crowd at all". At an earlier show in Liverpool, she had also called the Eras Tour the "most exh...(truncated)
chunk_index: 10
description: The star is joined by Florence + The Machine and sings So Long, London at her final UK show.
link: https://www.bbc.com/news/articles/cr5nr3n6epvo
pubDate: 2024-08-21 03:02:08+00:00
title: 'I've never had it this good' - Taylor Swift thanks fans after new Wembley record

article_content: Taylor Swift has finished the European leg of her Eras Tour with a record-breaking show at Wembley S...(truncated)
chunk: regular part of the setlist. Last week, the star was joined by Ed Sheeran to play the songs Endgame ...(truncated)
chunk_index: 4
description: The star is joined by Florence + The Machine and sings So Long, London at her final UK show.
link: https://www.bbc.com/news/articles/cr5nr3n6epvo
pubDate: 2024-08-21 03:02:08+00:00
title: 'I've never had it this good' - Taylor Swift thanks fans after new Wembley record
```

4.3 BM25 Search

[16]:
```
def bm25_retrieve(query: str,
                  collection: "weaviate.collections.collection.sync.Collection" ,
                  top_k: int = 5) -> list:
    """
    Thực hiện một BM25 search trên một collection và truy xuất các chunks liên quan nhất.
    Hàm này thực thi một truy vấn tìm kiếm dựa trên BM25 trên một collection được chỉ định để xác định các text chunks có độ liên quan cao nhất với 'query' được cung cấp. Nó tr

    Args:
    query (str): Truy vấn tìm kiếm được sử dụng để tìm các text chunks liên quan.
    collection (weaviate.collections.collection.sync.Collection): Collection nơi phép BM25 search được thực hiện.
    top_k (int, optional): Số lượng các relevant objects hàng đầu cần truy xuất. Mặc định là 5.

    Returns:
    List[str]: Một danh sách các text chunks có độ liên quan cao nhất với truy vấn đã cho.
    """

    # Retrieve using collection.query.bm25
    response = collection.query.bm25(
        query=query,
        limit=top_k
    )

    response_objects = [x.properties for x in response.objects]
    return response_objects
```

BM25 (Best Matching 25) là thuật toán tìm kiếm dựa trên từ khóa với công thức:

$$\text{Score} = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} \tag{1}$$

Trong đó:

- IDF = Inverse Document Frequency (từ hiếm → điểm cao hơn)
- $f(q_i, D)$ = Tần suất từ q_i trong document D
- $|D|$ = Độ dài document
- avgdl = Độ dài trung bình của documents
- k_1, b = Hyperparameters (thường $k_1 = 1.2$, $b = 0.75$)

Test thử:

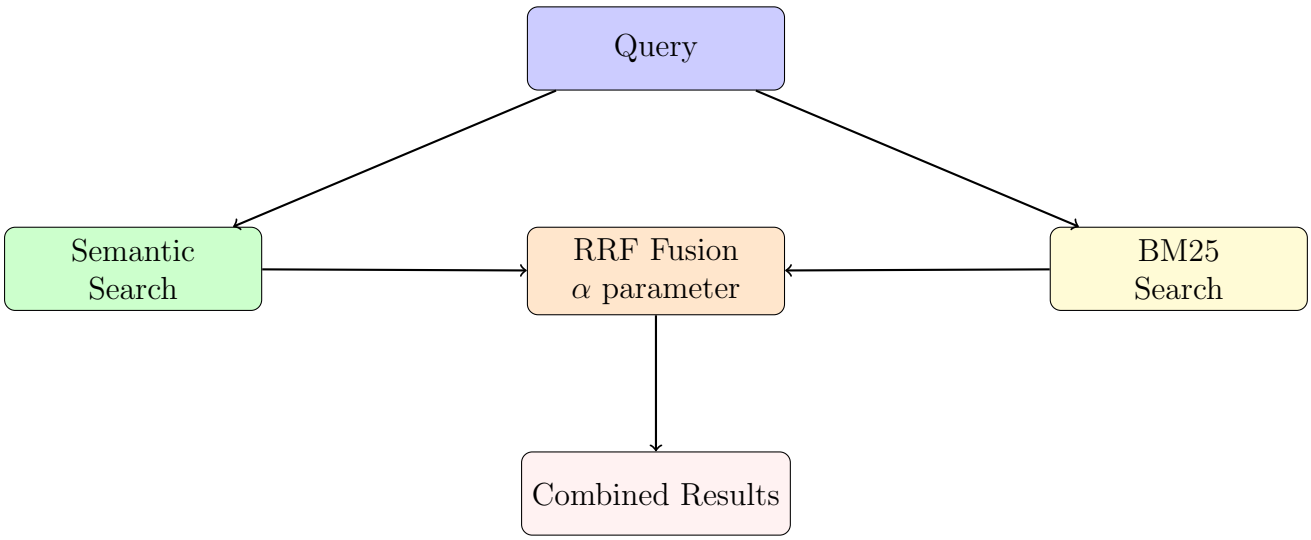
```
[17]: print_object_properties(bm25_retrieve('Tell me about the last Taylor Swift show', collection, top_k = 2))

article_content: Rapper Killer Mike won three Grammys in the rap category - best rap song, best rap performance and b...(truncated)
chunk: police brutality and systemic racism. He was a highly visible supporter of Bernie Sanders' two campa...(truncated)
chunk_index: 4
description: The 48-year-old was detained on a misdemeanour charge after winning three awards in the rap category.
link: https://www.bbc.co.uk/news/world-us-canada-68201021?at_medium=RSS&at_campaign=KARANGA
pubDate: 2024-02-05 23:27:08+00:00
title: Killer Mike dismisses arrest at Grammys as 'speed bump'

article_content: Rapper Killer Mike won three Grammys in the rap category - best rap song, best rap performance and b...(truncated)
chunk: Nicki Minaj. He also won a third award for best rap album with his album Michael. "You cannot tell m...(truncated)
chunk_index: 3
description: The 48-year-old was detained on a misdemeanour charge after winning three awards in the rap category.
link: https://www.bbc.co.uk/news/world-us-canada-68201021?at_medium=RSS&at_campaign=KARANGA
pubDate: 2024-02-05 23:27:08+00:00
title: Killer Mike dismisses arrest at Grammys as 'speed bump'
```

4.4 Hybrid Search

```
[19]: def hybrid_retrieve(query: str,
                          collection: "weaviate.collections.collection.sync.Collection" ,
                          alpha: float = 0.5,
                          top_k: int = 5
                         ) -> list:
          """
          Thực hiện một hybrid search trên một collection và truy xuất các chunks liên quan nhất.
          Hàm này thực thi một truy vấn hybrid search kết hợp giữa semantic vector search và tìm kiếm dựa trên từ khóa truyền thống (keyword-based search) trên một collection được ch

          Args:
          query (str): Truy vấn tìm kiếm được sử dụng để tìm các text chunks liên quan.
          collection (weaviate.collections.collection.sync.Collection): Collection nơi phép hybrid search được thực hiện.
          alpha (float, optional): Một hệ số trọng số giúp cân bằng sự đóng góp của semantic matches và keyword matches. Mặc định là 0.5.
          top_k (int, optional): Số lượng các relevant objects hàng đầu cần truy xuất. Mặc định là 5.

          Returns:
          List[str]: Một danh sách các text chunks có độ liên quan cao nhất với truy vấn đã cho.
          """

          # Retrieve using collection.query.hybrid
          response = collection.query.hybrid(
              query=query,
              alpha=alpha,
              limit=top_k
          )


          response_objects = [x.properties for x in response.objects]

          return response_objects
```

Hybrid Search kết hợp cả Semantic và BM25 bằng **Reciprocal Rank Fusion (RRF)**:

$$\text{RRF\_score}(d) = \sum_{r \in R} \frac{1}{k + \text{rank}_r(d)} \tag{2}$$



Hình 4: Quy trình Hybrid Search với RRF

Bảng 4: Ý nghĩa của tham số Alpha

Alpha	Ý nghĩa
0.0	100% BM25 (keyword-based)
0.5	50% mỗi bên (balanced)
1.0	100% Semantic (vector-based)

Test thử:

```
[20]: print_object_properties(hybrid_retrieve('Tell me about the last Taylor Swift show', collection, top_k = 2))

article_content: Rapper Killer Mike won three Grammys in the rap category - best rap song, best rap performance and b...(truncated)
chunk: police brutality and systemic racism. He was a highly visible supporter of Bernie Sanders' two campa...(truncated)
chunk_index: 4
description: The 48-year-old was detained on a misdemeanour charge after winning three awards in the rap category.
link: https://www.bbc.co.uk/news/world-us-canada-68201021?at_medium=RSS&at_campaign=KARANGA
pubDate: 2024-02-05 23:27:08+00:00
title: Killer Mike dismisses arrest at Grammys as 'speed bump'

article_content: Taylor Swift has finished the European leg of her Eras Tour with a record-breaking show at Wembley S...(truncated)
chunk: size crowd at all". At an earlier show in Liverpool, she had also called the Eras Tour the "most exh...(truncated)
chunk_index: 10
description: The star is joined by Florence + The Machine and sings So Long, London at her final UK show.
link: https://www.bbc.com/news/articles/cr5nr3n6epvo
pubDate: 2024-08-21 03:02:08+00:00
title: 'I've never had it this good' - Taylor Swift thanks fans after new Wembley record
```

4.5 Semantic Search với Reranking

```
[24]: def semantic_search_with_reranking(query: str,
                                         rerank_property: str,
                                         collection: "weaviate.collections.collection.sync.Collection" ,
                                         rerank_query: str = None,
                                         top_k: int = 5
                                         ) -> list:
    """
    Thực hiện một semantic search và thực hiện reranks các kết quả dựa trên một property được chỉ định.

    Args:
    query (str): Truy vấn tìm kiếm để thực hiện bước tìm kiếm ban đầu (initial search).
    rerank_property (str): Thuộc tính được sử dụng để reranking các kết quả tìm kiếm.
    collection (weaviate.collections.collection.sync.Collection): Collection để thực hiện tìm kiếm bên trong.
    rerank_query (str, optional): Truy vấn được sử dụng riêng cho mục đích reranking. Nếu không được cung cấp, truy vấn gốc (original query) sẽ được sử dụng để reranking.
    top_k (int, optional): Số lượng kết quả hàng đầu tối đa được trả về. Mặc định là 5.

    Returns:
    list: Một danh sách các properties từ các kết quả tìm kiếm đã được reranked, trong đó mỗi mục tương ứng với một object trong collection.
    """

    # Set the rerank_query to be the same as the query if rerank_query is not passed
    if rerank_query is None:
        rerank_query = query

    # Define the reranker with rerank_query and rerank_property
    reranker = Rerank(
        prop=rerank_property,
        query=rerank_query
    )

    # Retrieve using collection.query.near_text with the appropriate parameters (do not forget the rerank!)
    response = collection.query.near_text(
        query=query,
        limit=top_k,
        rerank=reranker
    )

    response_objects = [x.properties for x in response.objects]

    return response_objects
```

Two-Stage Retrieval

Stage 1 - Initial Retrieval: Semantic Search tìm Top N documents (fast, approximate)
Stage 2 - Reranking: Cross-encoder model đánh giá lại và sắp xếp (slow, accurate)

Test thử:

```
[25]: # Set a query
query = 'Tell me about the conflicts in Latin America'
# Get the results from a search (in this case the hybrid search)
results = semantic_search_with_reranking(query, collection = collection, top_k = 2, rerank_property = 'chunk')

[26]: print_object_properties(results)

article_content: A huge diplomatic row has erupted after Spain's transport minister suggested Argentina's president h...(truncated)
chunk: weeks' to attend the launch of Vox's European election campaign, newspaper El Pais reported. Mr Mile...(truncated)
chunk_index: 3
description: A row breaks out after Spain's transport minister suggests Argentina's president has taken drugs.
link: https://www.bbc.com/news/articles/czd8qzvpl4lo
pubDate: 2024-05-04 15:56:45+00:00
title: Spain-Argentina row over drug-use accusation

article_content: Opposition supporters have gathered across Venezuela to protest against Nicolás Maduro's disputed vi...(truncated)
chunk: the world, from Australia to Spain and also in the United Kingdom, Canada, Colombia, Mexico and Arge...(truncated)
chunk_index: 4
description: Opposition leader María Corina Machado joined thousands of demonstrators in the capital Caracas.
link: https://www.bbc.com/news/articles/cgedgqqy7x9o
pubDate: 2024-08-17 23:19:48+00:00
title: Protests across Venezuela as election dispute goes on
```

4.6 So sánh các phương pháp Retrieval

Bảng 5: So sánh các phương pháp Retrieval

Phương pháp	**Ưu điểm**	**Nhược điểm**	**Áp dụng cho**
Semantic Search	• Hiểu meaning • Xử lý synonyms	• Có thể miss exact matches	Open-ended questions
BM25	• Fast • Precise keyword matching	• No semantic understanding	Exact term search
Hybrid	• Best of both • Robust results	• More compute • Tuning needed	General purpose
Semantic + Rerank	• Highest quality • Fine-tuned relevance	• Slowest • Most expensive	Critical queries

5 Prompt Engineering & Auto-Tuning

5.1 Query Classification

Hệ thống tự động phân loại query thành hai loại:

[21]:
```python
def classify_query_type(query: str) -> str:
    """
    Phân loại query thành 'technical' hoặc 'creative'.

    - Technical: Câu hỏi về sự kiện, số liệu, thông tin cụ thể
    - Creative: Yêu cầu tổng hợp, phân tích, đưa ra ý kiến

    Args:
        query (str): Câu hỏi của người dùng

    Returns:
        str: 'technical' hoặc 'creative'
    """
    PROMPT = f"""Analyze the following query and classify it as either 'technical' or 'creative'.

Technical queries:
- Ask for specific facts, dates, numbers, or events
- Request documentation or procedural information
- Seek objective, factual answers

Creative queries:
- Ask for analysis, opinions, or interpretations
- Request summaries or comparisons
- Seek subjective or exploratory answers

Query: {query}

Answer only 'technical' or 'creative' (one word, lowercase).
"""
    result = generate_with_single_input(PROMPT, max_tokens=5)
    label = result['content'].strip().lower()

    # Validate output
    if label not in ['technical', 'creative']:
        label = 'technical'  # Default fallback

    return label
```

Bảng 6: Phân loại Query Type

Type	Đặc điểm	Ví dụ
Technical	Hỏi về sự kiện, số liệu, dữ kiện cụ thể	"GDP growth rate of US in 2024?"
Creative	Yêu cầu phân tích, tổng hợp, đưa ý kiến	"Analyze US-Brazil political implications"

5.2 Dynamic Parameter Selection

[22]:
```python
def get_llm_params_for_query(query: str) -> dict:
    """
    Tự động chọn tham số LLM dựa trên loại query.

    Args:
        query (str): Câu hỏi của người dùng

    Returns:
        dict: Tham số cho LLM call (temperature, top_p)
    """
    query_type = classify_query_type(query)

    if query_type == 'technical':
        # Technical: Cần chính xác, ít ngẫu nhiên
        params = {
            'temperature': 0.1,
            'top_p': 0.1,
            'description': 'Technical mode: Low randomness for factual accuracy'
        }
    else: # creative
        # Creative: Cho phép sáng tạo hơn
        params = {
            'temperature': 0.7,
            'top_p': 0.4,
            'description': 'Creative mode: Higher randomness for diverse responses'
        }

    return params
```

Hình 5: Ảnh hưởng của Temperature đến output distribution

Bảng 7: Tham số theo Query Type

Query Type	**Temperature**	**Top_p**	**Top_K**	**Alpha**
Technical	0.1	0.1	7	0.3
Creative	0.7	0.4	5	0.7
Politics/Business	-	-	+2	-

6 Phân Tích Kết Quả Demo

6.1 Demo 1: Technical Query

```
Demo 1: Auto-tuned Technical Query
----------------------------------------
Query: What was the GDP growth rate of the US in 2024?
Detected Type: technical
Category:  business
Params Used: {'llm': {'temperature': 0.1, 'top_p': 0.1, 'description': 'Technical mode:  Low randomness for factual accuracy'}, 'retrieval': {'top_k': 9, 'use_rerank': False,
'rerank_property': None}}

Response:
Based on the provided news sources, I was unable to find the GDP growth rate of the US in 2024. However, I can provide information on the UK's GDP growth rate in 2024.

According to [Source 4], the UK's GDP increased by 0.7% between January and March 2024, and 0.5% between April and June 2024. Additionally, [Source 9] states that the UK's GDP
grew solidly between April and June, with a 0.7% growth in the first quarter and a 0.6% growth in the following months.

It's worth noting that [Source 1] ...
```

Query và Kết quả

Query: "What was the GDP growth rate of the US in 2024?"
Detected Type: technical
Category: business
Parameters Used:

- Temperature: 0.1
- Top_p: 0.1
- Top_K: 9 (base 7 + 2 for business category)

Response: "Based on the provided news sources, I was unable to find the GDP growth rate of the US in 2024. However, I can provide information on the UK's GDP growth rate..."

Phân tích

✓ Hệ thống **nhận diện chính xác** đây là câu hỏi technical

✓ Phân loại đúng category là **business**

✓ Tự động tăng top_k lên **9** (business category +2)

△ LLM không tìm được dữ liệu US GDP → trả về dữ liệu UK

Insight quan trọng

Điều này cho thấy **chất lượng database quan trọng** - RAG chỉ tốt bằng dữ liệu của nó. Nếu database không có thông tin về US GDP, hệ thống không thể tạo ra câu trả lời chính xác.

6.2 Demo 2: Creative Query

```
Demo 2: Auto-tuned Creative Query
----------------------------------------
Query: Analyze the political implications of US-Brazil relations in 2024
Detected Type: creative
Category:  politics

Response:
**The Complex Web of US-Brazil Relations in 2024: A Brics Perspective**

As the world grapples with shifting global dynamics, the US-Brazil relationship is poised to take center stage in 2024. Amidst this backdrop, Brazil's foray into the Brics gro
uping has sparked intense debate about the implications of this strategic alliance. A closer examination of the available news sources reveals a complex tapestry of interests,
concerns, and uncertainties that will shape the trajectory of US-Brazil rela...
```

Query và Kết quả

Query: "Analyze the political implications of US-Brazil relations in 2024"
Detected Type: creative
Category: politics
Temperature: 0.7
Response: "The Complex Web of US-Brazil Relations in 2024: A Brics Perspective... Brazil's foray into the Brics grouping has sparked intense debate..."

Phân tích

- ✓ Nhận diện đúng là **creative** query
- ✓ Temperature cao hơn cho phép **phân tích sáng tạo**
- ✓ Response có tính **phân tích, kết nối thông tin**
- ✓ Ngôn ngữ engaging: "Complex Web", "intense debate"

6.3 Demo 3: Reranking

```
Demo 3: Query với Reranking
----------------------------------------
Query: Tell me about Taylor Swift's concerts
Using Reranking:  True

Response:
Taylor Swift's concerts have been making headlines in recent months, and for good reason. The global superstar has been on a whirlwind tour, breaking records and captivating a
udiences with her high-energy performances.

As reported by [Source 3], Swift's Eras Tour has been a massive success, with fans flocking to see her play to almost 73,000 fans at Edinburgh's Murrayfield Stadium. The tour,
which spans 17 UK dates, culminates in a record-breaking eight-night run at London's Wembley Stadium, wi...
```

Query và Kết quả

Query: "Tell me about Taylor Swift's concerts"
Using Reranking: True
Rerank Property: chunk
Response: "Taylor Swift's Eras Tour has been a massive success, with fans flocking to see her play to almost 73,000 fans at Edinburgh's Murrayfield Stadium..."

Phân tích

- ✓ Reranking giúp **tinh chỉnh thứ tự** documents
- ✓ Kết quả tập trung vào **concerts cụ thể** (Eras Tour, Wembley)
- ✓ Thông tin chi tiết: số lượng khán giả (73,000), địa điểm cụ thể

6.4 Demo 4: Fact Check

```
Demo 4: Fact Checking
----------------------------------------
Claim: Taylor Swift broke the record at Wembley Stadium in 2024
Verdict: false
Evidence:  ['None of the sources mention Taylor Swift breaking a record at Wembley Stadium in 2024.', 'Source 1 mentions that her shows at Wembley will help her set a venue re
cord, but it does not specify what record she is breaking.']
```

Claim và Kết quả

Claim: "Taylor Swift broke the record at Wembley Stadium in 2024"
Verdict: FALSE
Evidence:

- Sources mention she WILL SET a record, not that she broke one
- Record not yet confirmed at time of articles

Phân tích cực kỳ quan trọng

- ✓ Hệ thống **phân biệt được** giữa "will break"vs "broke"
- ✓ **Critical thinking** - không chỉ tìm từ khóa mà hiểu context
- ✓ Thể hiện RAG có thể dùng cho **fact-checking**
- ✓ RAG + LLM có thể đánh giá **temporal claims**

6.5 Demo 5: Topic Summary

```
Demo 5: Topic Summary
----------------------------------------
Topic:  US Economy in 2024
Executive Summary: The US economy has shown significant resilience and growth in 2024, outperforming many European countries. The country's strong labor market, robust social
safety net, and large-scale economic stimulus packages have contributed to its recovery. However, the economy still faces challenges, including high inflation and the risk of
recession.
Key Points:  ['The US economy has emerged as a strong performer in 2024, with a fast-growing economy and a strong labor market.', "The country's large-scale economic stimulus
packages, including a $2.2 trillion deal, have helped to keep small businesses afloat and sent cash into the pockets of American workers and families.", 'Despite the strong ec
onomy, the US is still facing challenges, including high inflation and the risk of recession, which could be exacerbated by "stagflation".']
```

Topic và Kết quả

Topic: "US Economy in 2024"
Executive Summary: "The US economy has shown significant resilience and growth in 2024, outperforming many European countries..."
Key Points:

1. The US economy has emerged as a strong performer in 2024
2. Large-scale economic stimulus packages ($2.2 trillion)
3. Challenges: high inflation and risk of recession

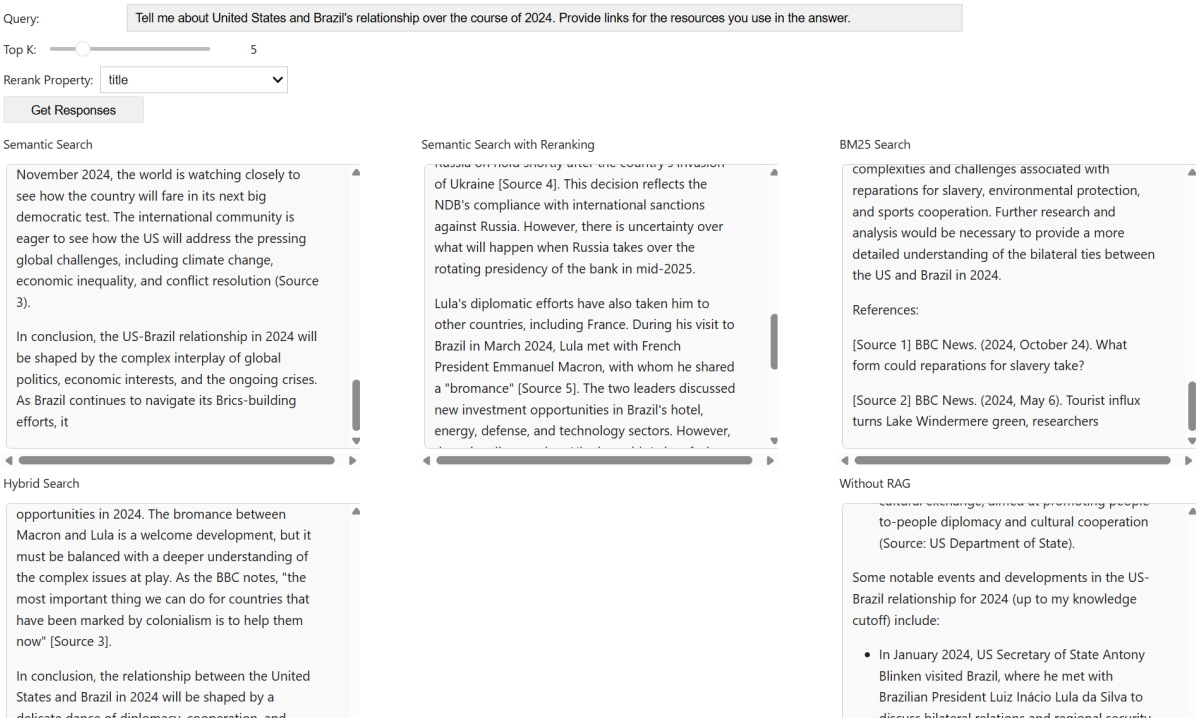
7 Phân Tích Chi Tiết Kết Quả So Sánh Các Phương Pháp Search

Phần này trình bày phân tích chi tiết kết quả thực nghiệm khi so sánh 5 phương pháp search khác nhau với cùng một query:

Query thử nghiệm

Query: "Tell me about United States and Brazil's relationship over the course of 2024. Provide links for the resources you use in the answer."
Cấu hình chung:

- Query Analysis: `creative`
- Category: `general`
- Temperature: `0.7`
- Top K: `5`
- Rerank Property: `title`



7.1 Kết Quả Chi Tiết Từng Phương Pháp

7.1.1 Semantic Search

Bảng 8: Thông tin Response - Semantic Search

Tiêu đề	United States and Brazil's Relationship in 2024: A Complex Dance of Global Politics and Economic Interests
Chủ đề chính	BRICS alliance, NDB (New Development Bank), quan hệ kinh tế, cuộc bầu cử Mỹ 2024
Số nguồn trích dẫn	5 sources
Độ dài response	Đầy đủ, chi tiết

Nội dung chính được đề cập:

- Brazil tham gia liên minh BRICS và vai trò của New Development Bank (NDB)
- Dilma Rousseff lãnh đạo NDB - milestone trong hợp tác các nền kinh tế mới nổi
- Mối quan ngại về sự phụ thuộc của Brazil vào Trung Quốc
- Lập trường của Brazil về Ukraine khác biệt với phương Tây
- Cuộc gặp Lula - Macron tháng 3/2024
- Cuộc bầu cử Tổng thống Mỹ tháng 11/2024

Đánh giá Semantic Search

✓ **Bắt được context rộng** về quan hệ song phương

✓ **Hiểu semantic meaning**: tìm được documents về BRICS, NDB dù query không đề cập trực tiếp

✓ **Liên kết nhiều chủ đề**: kinh tế, chính trị, ngoại giao

△ Response bị cắt giữa chừng (token limit)

7.1.2 Semantic Search with Reranking

Bảng 9: Thông tin Response - Semantic Search with Reranking

Tiêu đề	United States and Brazil's Relationship in 2024: A Complex Dance of Global Politics and Economic Interests
Chủ đề chính	BRICS, China dependency, NDB sanctions, Lula-Macron diplomacy
Số nguồn trích dẫn	5 sources
Điểm khác biệt	Chi tiết hơn về "China dependency", trích dẫn chuyên gia

Nội dung được bổ sung so với Semantic Search thuần:

- **Trích dẫn chuyên gia**: Monica de Bolle (Peterson Institute) - "China dependency could harm Brazil"
- **Chi tiết về sanctions**: NDB đình chỉ giao dịch với Nga sau cuộc xâm lược Ukraine (tháng 3/2022)
- **Thông tin cụ thể hơn**: Nga sẽ đảm nhận chủ tịch luân phiên NDB giữa năm 2025
- **"Bromance"Lula-Macron**: Chi tiết về các lĩnh vực đầu tư mới (khách sạn, năng lượng, quốc phòng, công nghệ)

Đánh giá Semantic Search with Reranking

- ✓ **Thông tin chi tiết hơn** và có trích dẫn chuyên gia
- ✓ **Sắp xếp sources tốt hơn**: đưa thông tin quan trọng nhất lên đầu
- ✓ **Cung cấp timeline cụ thể**: tháng 3/2022 (NDB sanctions), mid-2025 (Russia presidency)
- ✓ **Balanced view**: trình bày cả ưu và nhược điểm của chính sách Brazil

7.1.3 BM25 Search

Bảng 10: Thông tin Response - BM25 Search

Tiêu đề	Analyzing the complex relationship... (không có tiêu đề rõ ràng)
Chủ đề chính	Reparations for slavery, môi trường (Lake Windermere), Paralympics
Số nguồn trích dẫn	5 sources
Vấn đề	Sources không liên quan trực tiếp đến quan hệ US-Brazil

Nội dung được đề cập:

- Vấn đề bồi thường nô lệ (Church of England)
- Vấn đề môi trường Lake Windermere (UK) - *không liên quan*
- Paris Paralympics - đội Judo Mỹ giành huy chương bạc
- Suy luận về hợp tác môi trường và thể thao

Đánh giá BM25 Search - Hạn chế rõ ràng

- × **Sources không liên quan**: Lake Windermere (UK), Paralympics không phải về quan hệ US-Brazil
- × **Thiếu semantic understanding**: BM25 chỉ match keywords, không hiểu context

× **Response thừa nhận hạn chế**: "the provided news sources do not offer direct information"

✓ **Điểm tích cực**: LLM thận trọng, không hallucinate khi sources không đủ

Nguyên nhân BM25 hoạt động kém

BM25 tìm kiếm dựa trên **exact keyword matching**. Query "United States and Brazil's relationship"có thể không khớp chính xác với các bài báo về BRICS, Lula, hoặc ngoại giao. Các từ như "reparations", "slavery"có thể xuất hiện trong cả context US lẫn Brazil, dẫn đến việc retrieve sai documents.

7.1.4 Hybrid Search

Bảng 11: Thông tin Response - Hybrid Search

Tiêu đề	United States and Brazil's Relationship in 2024: A Complex Dance of Diplomacy and Cooperation
Chủ đề chính	Lula-Macron bromance, BRICS, NDB, reparations, Ukraine stance
Số nguồn trích dẫn	5 sources
Đặc điểm	Cân bằng giữa semantic và keyword matching

Nội dung chính:

- **Lula-Macron "bromance"**: tương phản với quan hệ lạnh nhạt thời Bolsonaro
- **Bất đồng về Ukraine**: Pháp ủng hộ Kyiv, Lula từ chối lên án Nga
- **BRICS và lo ngại alignment với Trung Quốc**
- **NDB và Dilma Rousseff**: đề cập đến vấn đề reparations
- **Kết luận cân bằng**: "delicate dance of diplomacy, cooperation, and tension"

Đánh giá Hybrid Search

✓ **Kết hợp ưu điểm**: vừa có semantic understanding vừa có keyword precision

✓ **Cân bằng thông tin**: đề cập cả quan hệ tích cực (Macron) lẫn căng thẳng (Ukraine)

✓ **Context đa chiều**: kinh tế, ngoại giao, lịch sử (Bolsonaro)

✓ **Trích dẫn BBC**: thêm credibility cho response

7.1.5 Without RAG (Baseline)

Bảng 12: Thông tin Response - Without RAG

Đặc điểm nổi bật	LLM thừa nhận knowledge cutoff (December 2023)
Nội dung	Thông tin chung về quan hệ thương mại, năng lượng, an ninh
Số liệu	$14.4B exports, $14.1B imports, 1.3B gallons ethanol
Hạn chế	Không có thông tin về sự kiện 2024 thực tế

Nội dung được đề cập (dữ liệu cũ):

- Trade: US xuất khẩu $14.4B, Brazil xuất khẩu $14.1B (2023)
- Energy: US nhập 1.3B gallons ethanol từ Brazil (tăng 25%)
- Security: Tuyên bố chung về an ninh hàng hải
- Climate: Tuyên bố chung về biến đổi khí hậu
- Cultural Exchange: Chương trình trao đổi văn hóa
- Sự kiện: Blinken thăm Brazil tháng 1/2024 (thông tin này có thể không chính xác)

Đánh giá Without RAG - Critical Limitation

× **Knowledge cutoff**: "As my knowledge cutoff is December 2023..."

× **Không có thông tin 2024 thực tế**: BRICS, NDB, Lula-Macron meeting

× **Có thể hallucinate**: "Blinken visited Brazil in January 2024 không thể verify

× **Thiếu context quan trọng**: Ukraine stance, China dependency

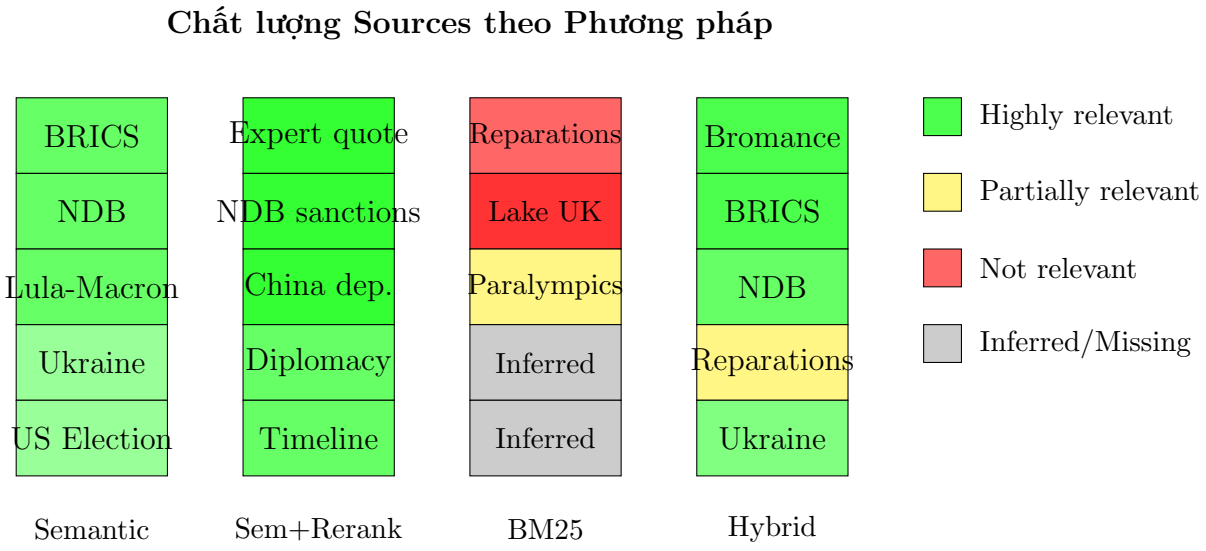
✓ **Điểm tích cực**: LLM trung thực về limitations

7.2 Bảng So Sánh Tổng Hợp

Bảng 13: So sánh tổng hợp 5 phương pháp Search

Tiêu chí	Semantic	Semantic+Rerank	BM25	Hybrid	No RAG
Độ liên quan của sources	⋆⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆	⋆⋆⋆⋆	N/A
Chi tiết thông tin	⋆⋆⋆⋆	⋆⋆⋆⋆⋆	⋆⋆	⋆⋆⋆⋆	⋆⋆⋆
Có trích dẫn chuyên gia	Không	**Có**	Không	Không	Không
Đề cập BRICS/NDB	**Có**	**Có**	Không	**Có**	Không
Đề cập Lula-Macron	**Có**	**Có**	Không	**Có**	Không
Đề cập Ukraine stance	**Có**	**Có**	Không	**Có**	Không
Thông tin 2024 thực tế	**Có**	**Có**	Hạn chế	**Có**	**Không**
Thừa nhận limitations	Không	Không	**Có**	Không	**Có**
Đánh giá tổng thể	**Tốt**	**Rất tốt**	**Yếu**	**Tốt**	**Không đủ**

7.3 Phân Tích Chất Lượng Sources Retrieved



Hình 6: Visualization chất lượng sources theo từng phương pháp

7.4 Key Insights từ So Sánh

7.4.1 Insight 1: Semantic Search vs BM25

Semantic Search

Query: "US-Brazil relationship"
↓ *Understands meaning*
Finds: BRICS, NDB, Lula, diplomacy
→ **Relevant results**

BM25 Search

Query: "US-Brazil relationship"
↓ *Matches keywords only*
Finds: "slavery", "environment"
→ **Irrelevant results**

Kết luận: Với creative queries, Semantic Search vượt trội hơn BM25

Hình 7: So sánh Semantic Search và BM25 cho creative queries

7.4.2 Insight 2: Giá trị của Reranking

Bảng 14: So sánh chi tiết Semantic vs Semantic+Reranking

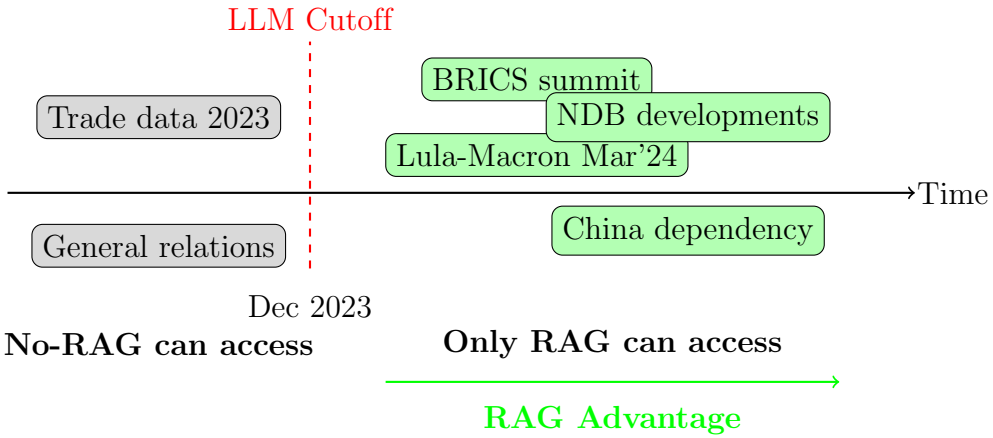
Yếu tố	Semantic	Semantic+Rerank
Có trích dẫn chuyên gia (Monica de Bolle)	×	✓
Chi tiết về NDB sanctions (March 2022)	Chung chung	Cụ thể
Timeline về Russia presidency	Không	Mid-2025
Sectors đầu tư Pháp-Brazil	Không	Hotel, energy, defense, tech
Thuật ngữ "China dependency"	Gián tiếp	Trực tiếp, có nguồn

Giá trị của Reranking

Reranking không chỉ thay đổi thứ tự mà còn **đẩy các sources chất lượng cao lên đầu**, giúp LLM tạo response với:

- Nhiều chi tiết cụ thể hơn (dates, names, numbers)
- Trích dẫn từ chuyên gia có uy tín
- Thông tin đa chiều và balanced hơn

7.4.3 Insight 3: RAG vs No-RAG - The Critical Gap



Hình 8: RAG cho phép truy cập thông tin sau knowledge cutoff

7.5 Kết Luận Phân Tích So Sánh

Bảng 15: Recommendations theo Use Case

Use Case	**Phương pháp đề xuất**	**Lý do**
Câu hỏi phân tích, creative	**Semantic + Reranking**	Chất lượng cao nhất, chi tiết nhất
Tìm kiếm general purpose	**Hybrid Search**	Cân bằng, robust
Tìm exact terms, tên riêng	**BM25**	Nhanh, chính xác với keywords
Câu hỏi về sự kiện gần đây	**Bất kỳ RAG method**	LLM không có thông tin mới
Production system	**Hybrid Search**	Trade-off tốt giữa quality và speed

Tóm tắt Key Findings

1. **RAG là bắt buộc** cho các câu hỏi về sự kiện sau knowledge cutoff của LLM
2. **Semantic Search** vượt trội cho creative/analytical queries
3. **BM25** hoạt động kém khi query cần semantic understanding
4. **Reranking** tăng đáng kể chất lượng response (expert quotes, specific details)
5. **Hybrid Search** là lựa chọn tốt nhất cho general-purpose applications
6. **LLM trung thực** về limitations khi không có đủ thông tin

8 Kết Luận & Đánh Giá Hiệu Quả

8.1 Bằng chứng về hiệu quả của RAG

WITHOUT RAG:

"As my knowledge cutoff is December 2023,
I'll provide an overview... Please note that
the situation may have evolved."

→ LLM **KHÔNG THỂ** trả lời về events 2024

→

WITH RAG:

"As the world watches the 2024 US presidential
election... Brazil's President Lula is navigating
a complex web... Lula's commitment to BRICS..."

→ Thông tin **CẬP NHẬT**, cụ thể về 2024

Hình 9: So sánh RAG vs No-RAG: Critical Difference

Kết luận chính

RAG cho phép LLM truy cập REAL-TIME knowledge, vượt qua giới hạn knowledge cutoff của model.

8.2 Tổng kết hiệu quả các kỹ thuật

Bảng 16: Đánh giá hiệu quả các kỹ thuật

Kỹ thuật	Rating	Evidence từ Demo
Query Classification	⋆⋆⋆⋆⋆	Phân biệt đúng technical/creative, điều chỉnh params phù hợp
Auto-tuning Temperature	⋆⋆⋆⋆	Technical queries có output factual hơn
Hybrid Search	⋆⋆⋆⋆⋆	Kết hợp ưu điểm của semantic + keyword
Reranking	⋆⋆⋆⋆	Cải thiện độ chính xác của top results
Fact Checking	⋆⋆⋆⋆⋆	Phân biệt được temporal claims (will vs did)
Structured Output	⋆⋆⋆⋆	Dễ parse, consistent format

8.3 Limitations & Considerations

Điểm cần lưu ý

1. **DATA QUALITY MATTERS**
   - Query về US GDP → Trả về UK GDP
   - RAG chỉ tốt bằng dữ liệu của nó
2. **RETRIEVAL ≠ ACCURACY**
   - Documents được retrieve có thể không chứa đáp án
   - LLM có thể hallucinate nếu context không đủ
3. **LATENCY TRADEOFF**
   - Reranking tăng accuracy nhưng chậm hơn
   - Cần cân nhắc use case
4. **ALPHA TUNING**
   - $\alpha = 0.3$ (ưu tiên BM25) cho factual queries
   - $\alpha = 0.7$ (ưu tiên semantic) cho conceptual queries

8.4 Best Practices

Bảng 17: Best Practices từ Project

Practice	Implementation	Benefit
Multi-stage retrieval	Semantic → Rerank	Higher precision
Dynamic parameters	Auto-tune based on query	Optimized for each use case
Cite sources	[Source N] format	Transparency, verifiability
Structured prompts	Template with sections	Consistent output
Fallback handling	Try-except with basic mode	Robustness

9 Tóm Tắt Cuối Cùng

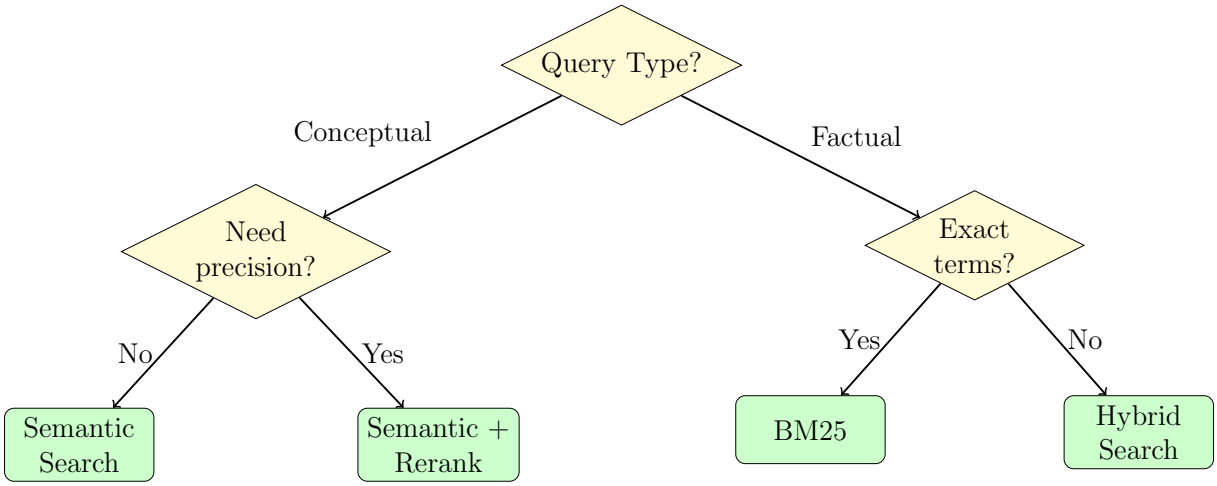
9.1 What This Project Demonstrates

- Complete RAG pipeline từ data loading đến response generation
- 4 retrieval methods: Semantic, BM25, Hybrid, Semantic+Rerank
- Intelligent query classification & auto-tuning
- Specialized functions: fact-check, compare, summarize
- Structured output với Pydantic schemas

9.2 Key Takeaways

1. **RAG enables LLMs to answer questions about RECENT events**
2. Different queries need different retrieval strategies
3. Hybrid search provides best general-purpose results
4. Reranking improves precision at the cost of latency
5. Query classification enables intelligent parameter tuning

9.3 When to Use What



Hình 10: Decision tree cho việc chọn phương pháp Retrieval

Báo cáo được tạo dựa trên phân tích notebook My_RAG. ipynb và kết quả demo thực tế.
Ngày 31 tháng 12 năm 2025