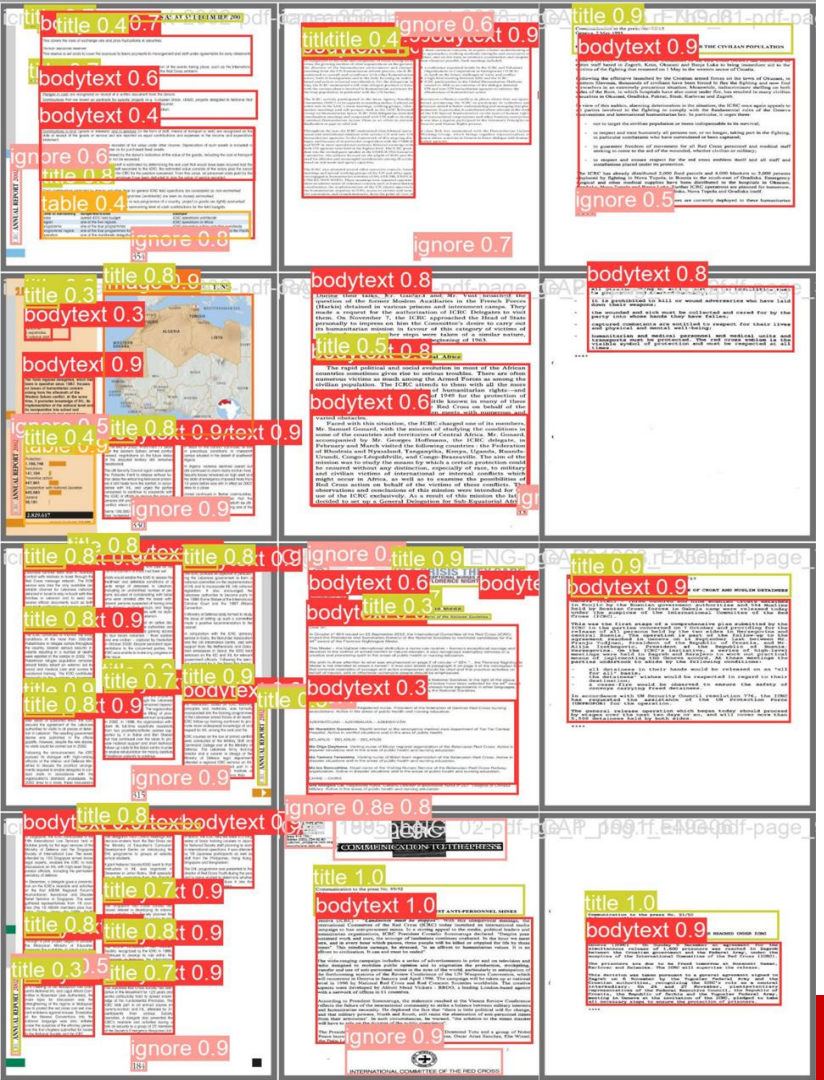


# Codex Project : RAG Model Development


Camille CHALLIER

**Supervised by:**

- Eduard Ďurech
- Mary-Anne Hartley
- Martin Jaggi

# CONTENTS

- CONTEXT
- AIM & OBJECTIVES
- METHODS & FINDINGS
  - Data processing → Retriever → Re-Ranker
- DISCUSSION
- FUTURE WORK
- QUESTIONS

# CONTEXT - *Automating Humanitarian Efforts with AI*

**International Committee of the Red Cross (ICRC)**

- World's largest humanitarian archives, amassed over 160 years.
- Need: Automate reading and parsing of vast archives.

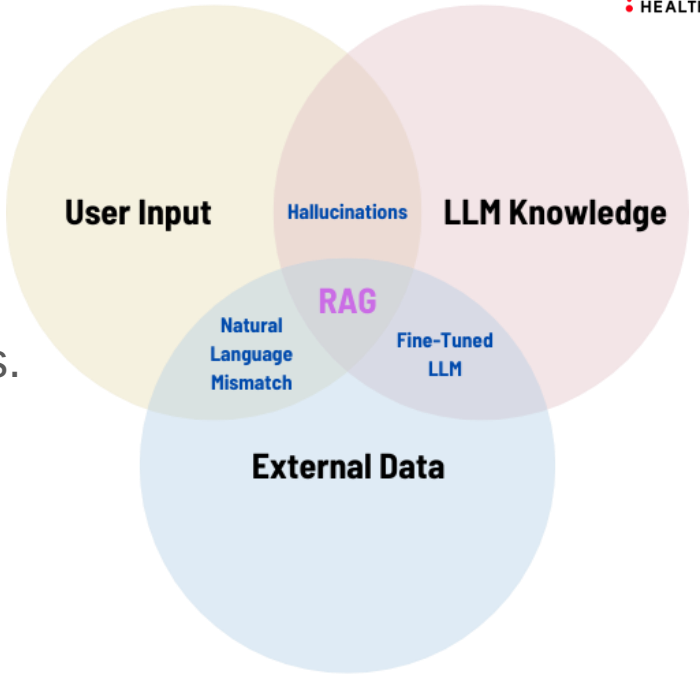
**World Health Organization (WHO)**

- Produces 20-30 major guidelines annually that require extensive review of academic papers and resources.
- Need: AI-assisted medical guidelines review

# CONTEXT - *Automating Humanitarian Efforts with AI*

**Problems :**

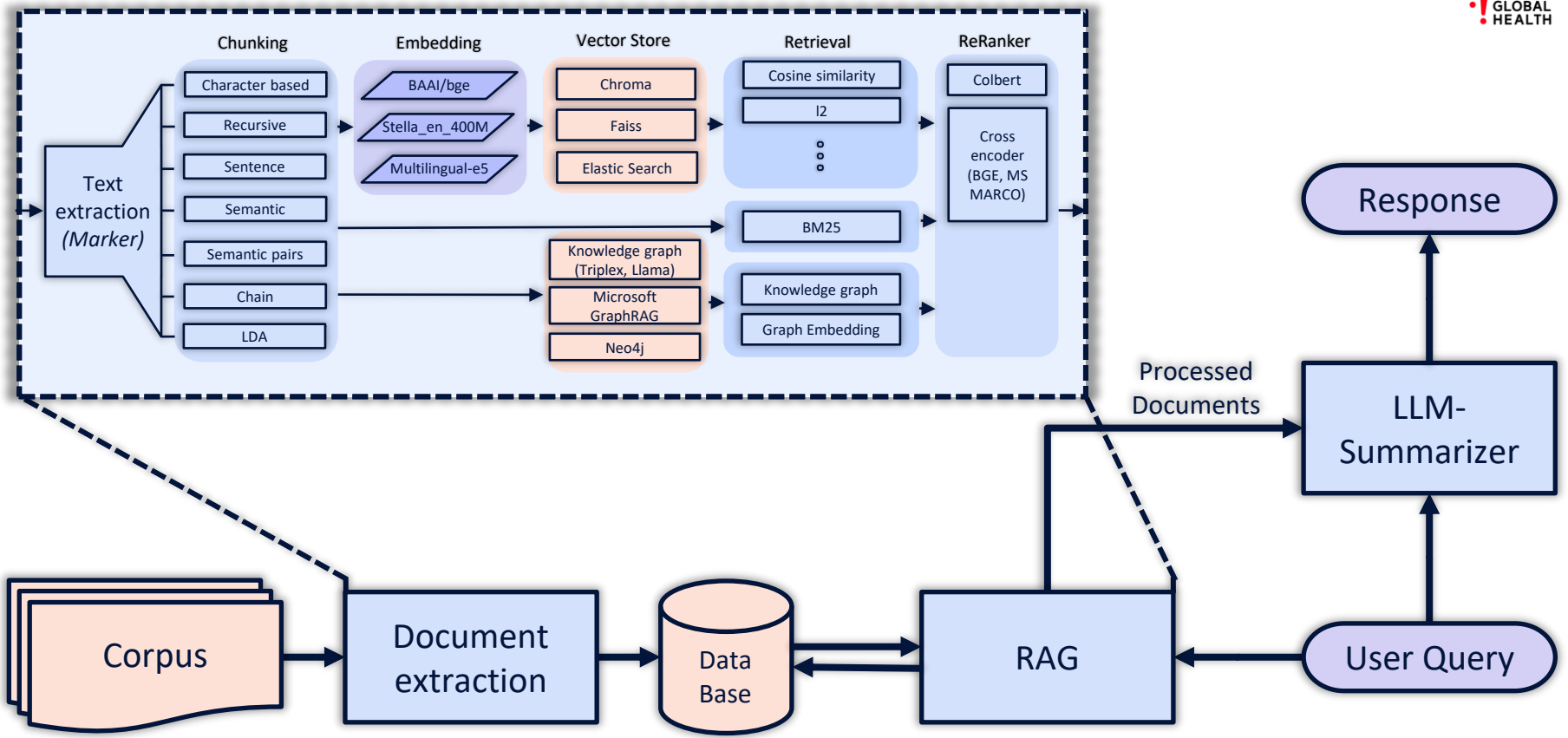
- Summarize TB-PB of data.
- Documents continuously updated.
- Citation of sources necessary (hallucinations).
- ICRC and WHO does not have super-computers.
- Multilingual data.
- Domain specific (NGO, medical).




**Solutions :**

- Retrieval-augmented generation (RAG)
- Use a small LLM with RAG to preprocess documents and embed them into an offline database => efficient & manageable.

# CONTEXT - *What is RAG ?*



# AIM & OBJECTIVES

**AIM:** Develop a RAG-based model to efficiently manage and accurately process the extensive, multilingual archives of ICRC and WHO.

**OBJECTIVES**

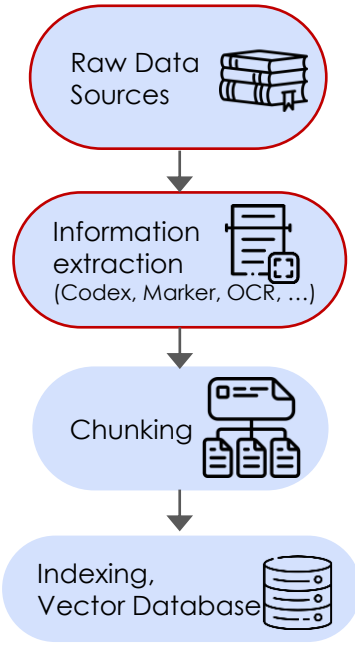
1. Construct a **vector database** by effectively processing the data.

2. Precisely **retrieve** the most relevant documents.

3. Efficiently **re-rank** the retrieved documents for optimal relevance.

# METHODS - *Data*

| | Wikitext Dataset | ICRC Dataset |
|---|---|---|
| **QUANTITY** | 1293 articles / 100 millions tokens. | 193 documents |
| **QUALITY** | Extracted from the set of verified Good and Featured articles on Wikipedia. | Extracted using Marker and Surya for OCR. (https://github.com/VikParuchuri/marker) |
| **LANGUAGES** | English | French, English, German.. |
| **ADVANTAGES** | High quality data | Subset of the real data |



# METHODS - *Data*

**Test Queries :**

- A series of questions generated by LLaMA 3 8B based on 500 random data chunks.

**Evaluation :**

- Assess the RAG model's effectiveness in retrieving the specific document sections used to generate the query.

**Metrics :**

- Hit Rate : HR = $\frac{|\,U_{hit}^{L}\,|}{|\,U_{all}\,|}$,

Where $|\,U_{hit}^{L}\,|$ is the number of questions for which the correct answer is included in the top L retrieved chunks, $|\,U_{all}\,|$ is the total number of questions.

- Mean Reciprocal Rank : MRR = $\frac{1}{|\,U_{all}\,|}\sum_{u=1}^{|U_{hit}|}\frac{1}{rank_u}$

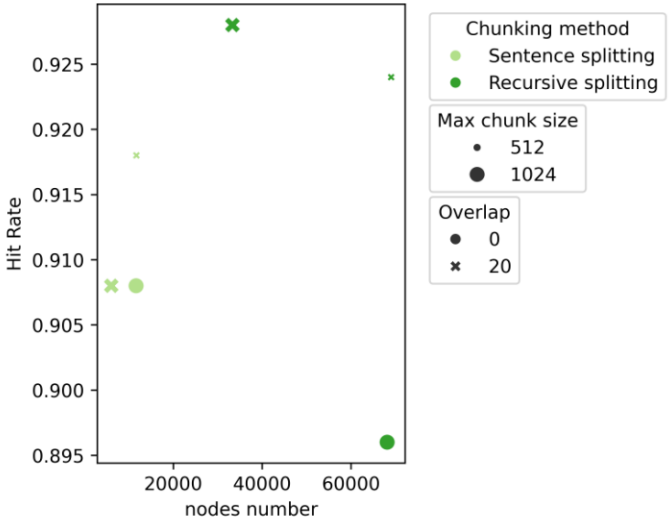
Where $rank_u$ is the rank of the correctly retrived chunks in the top L.

# OBJECTIVE 1 - *Database --Chunking Strategies*

## Method

- Character splitting
- Sentence splitting
- Recursive Character splitting
  ( ["\n\n", "\n", " ", ""] )

## Result



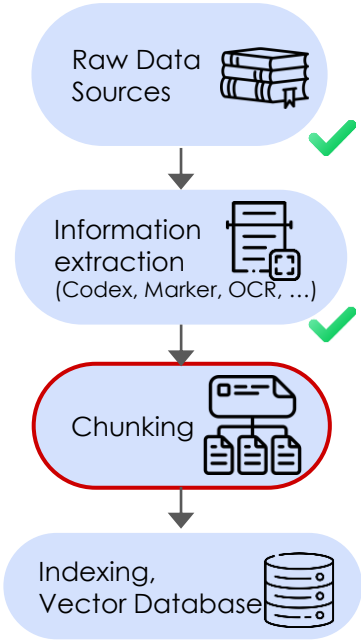
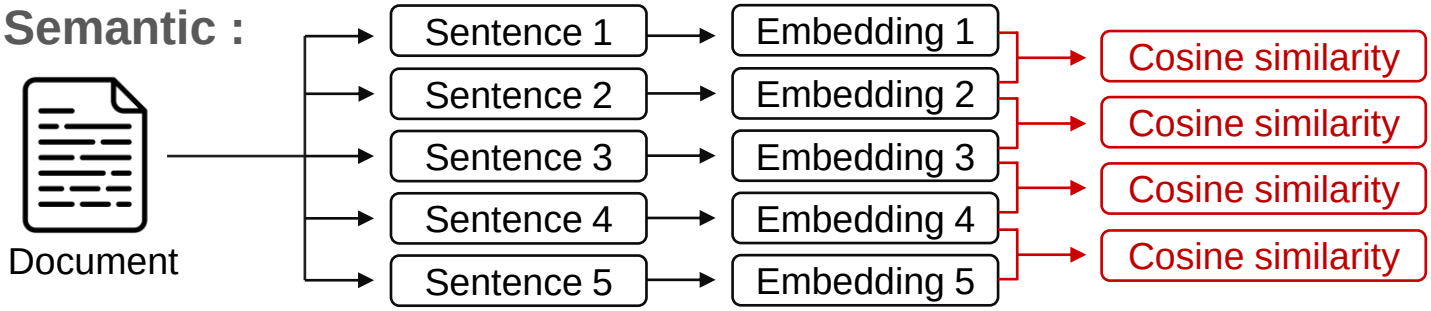



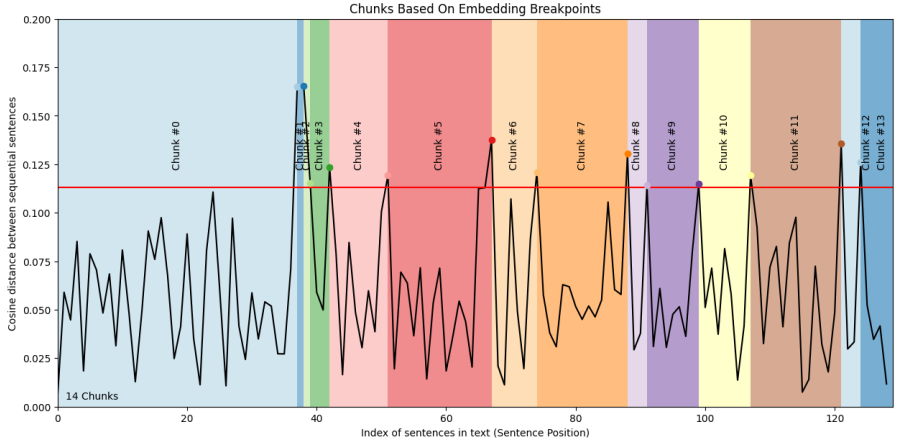

**Figure 2:** Hit Rate obtained using different chunking methods. The embedding model "bge-small-en" was used to retrieved the top 3 documents. It was tested on the WikiText datasets from Hugging Face "Salesforce/wikitext". The test queries consist of 500 precise questions.

# OBJECTIVE 1 - *Vector database --Chunking Strategies*

**Method** Semantic and topic based chunking :

- **Semantic :**

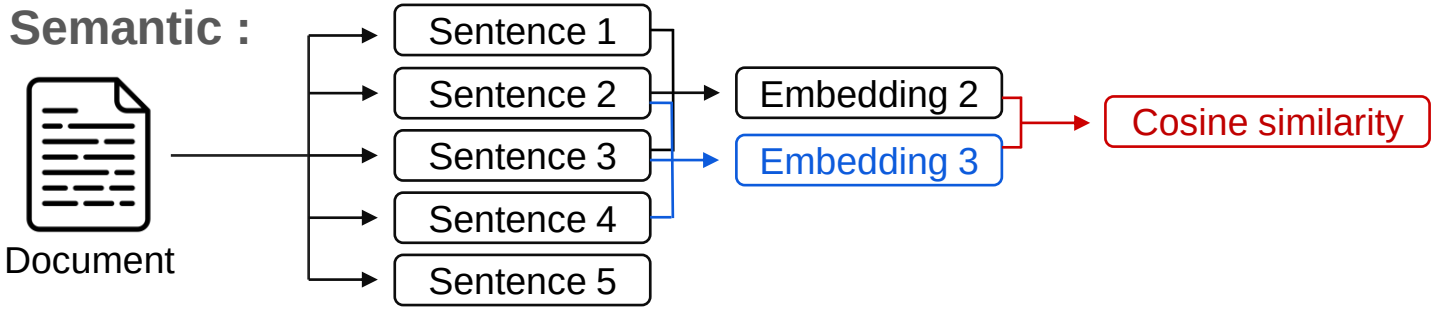


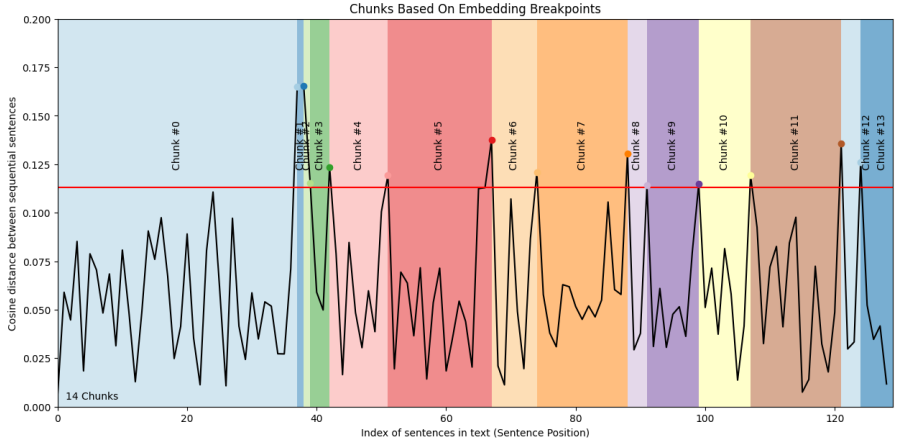



# OBJECTIVE 1 - *Vector database --Chunking Strategies*

**Method** Semantic and topic based chunking :

- **Semantic :**

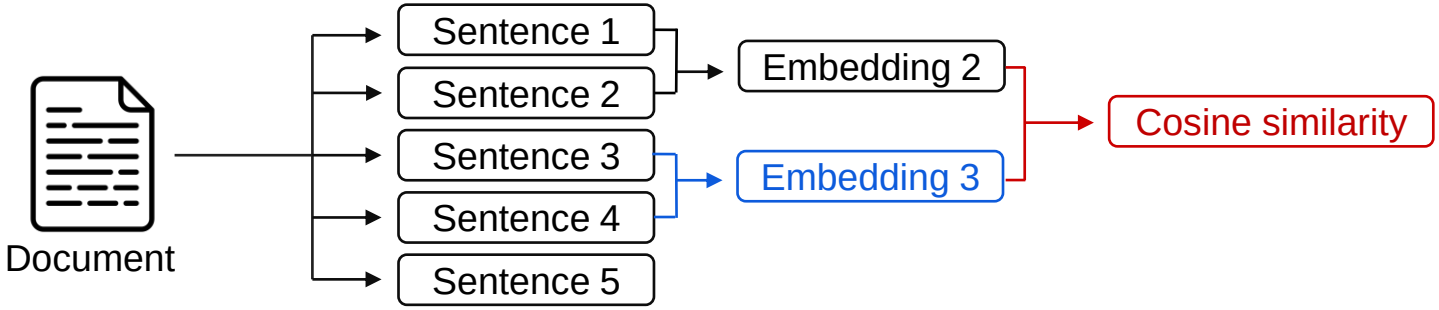


# OBJECTIVE 1 - *Vector database --Chunking Strategies*

**Method** Semantic and topic based chunking :

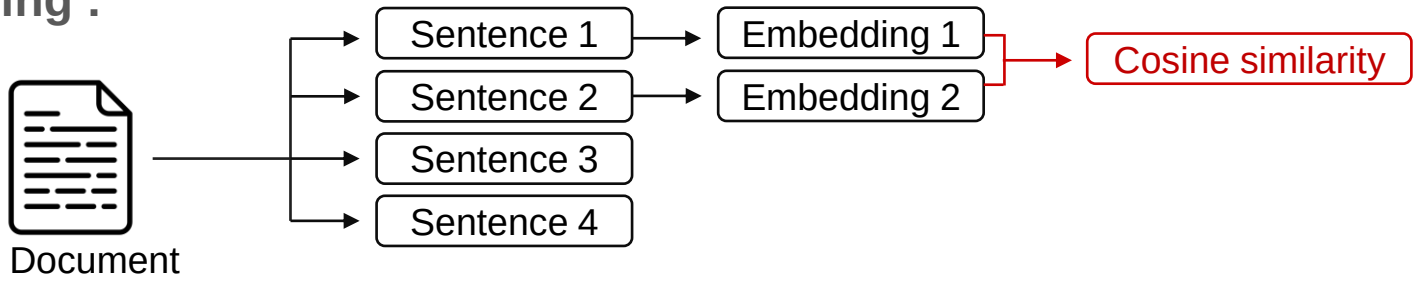
- **Semantic :**
- **Semantic pairs :**

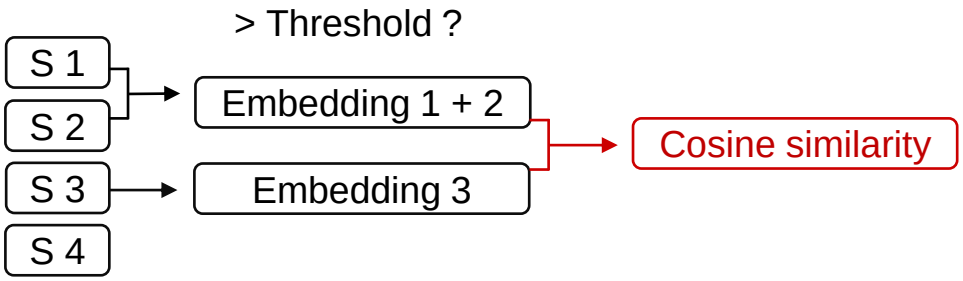


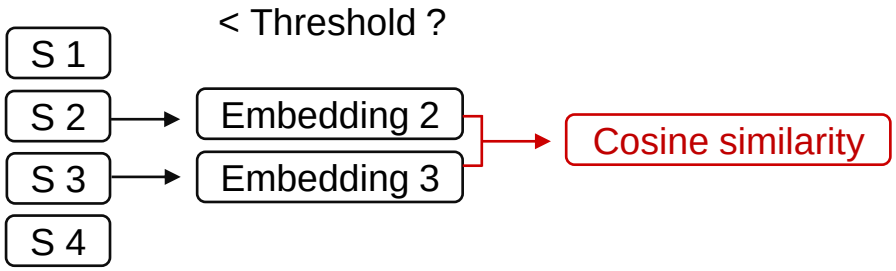
# OBJECTIVE 1 - *Vector database --Chunking Strategies*

**Method** Semantic and topic based chunking :

- **Semantic :**
- **Semantic pairs :**
- **Chaining :**



# OBJECTIVE 1 - *Vector database --Chunking Strategies*

**Method** Semantic and topic based chunking :

- **Semantic :**
- **Semantic pairs :**
- **Chaining :**
- **LDA**

Cosine similarity : $$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

With A and B two vectors embeddings

# OBJECTIVE 1 - *Vector database --Chunking Strategies*

## Results

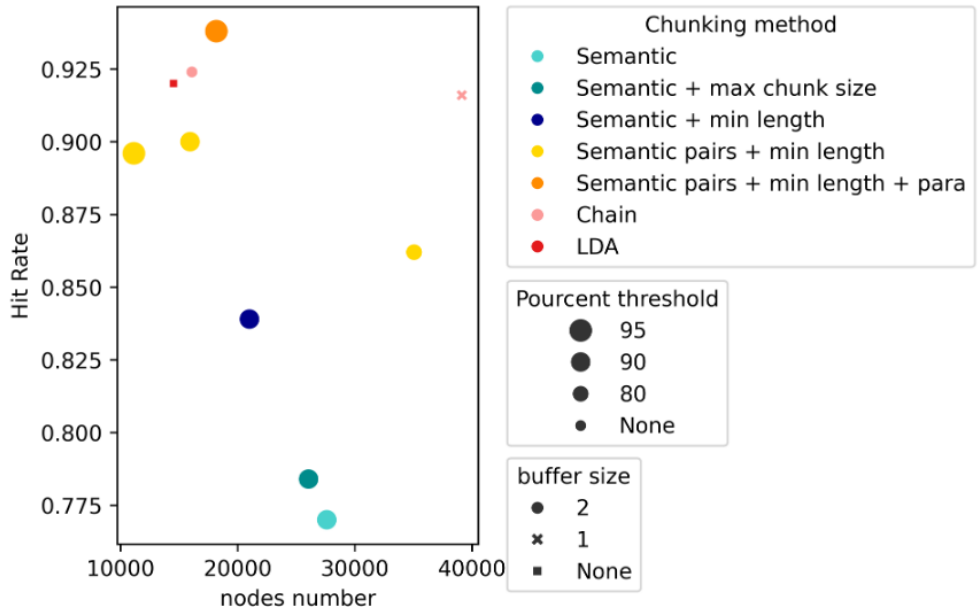
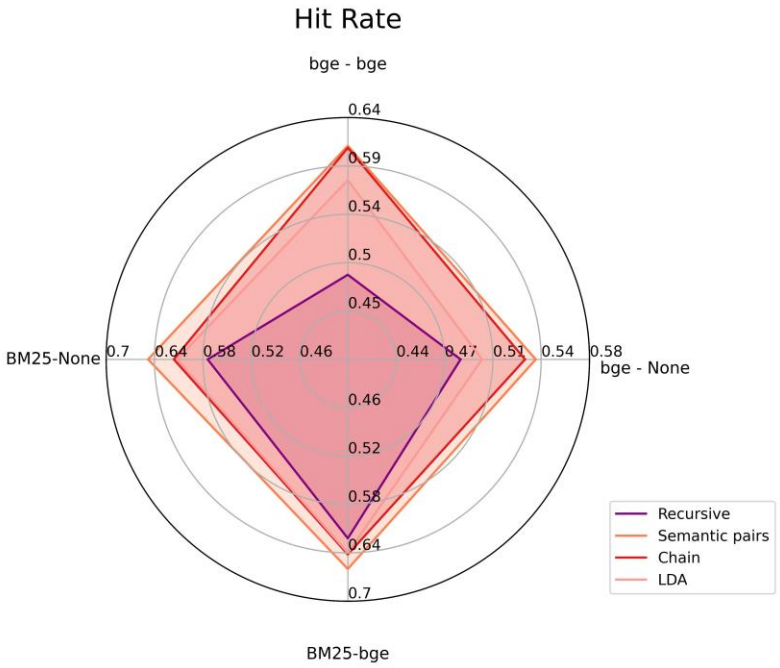
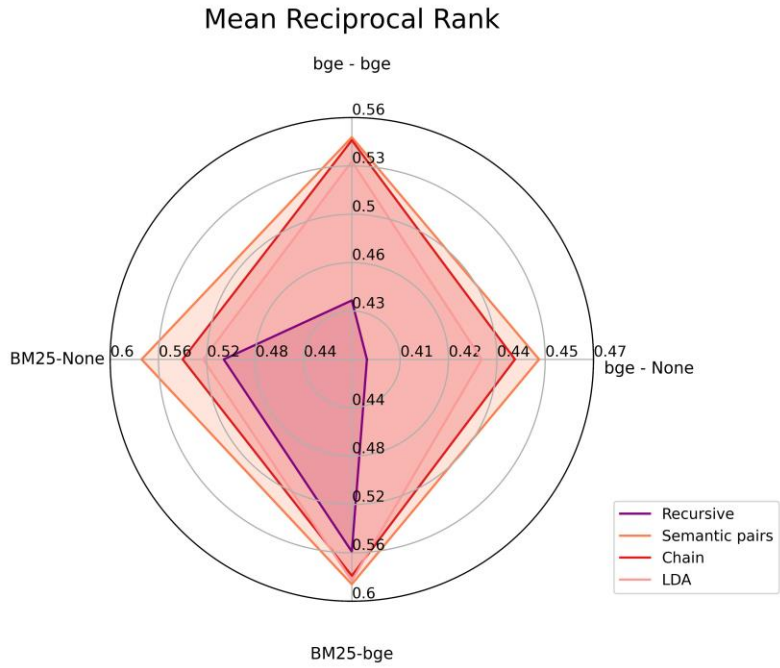



**Figure 3:** Hit Rate obtained using different semantic and topics-based chunking methods. The embedding model "bge-small-en" was used as RAG model and tested on the WikiText datasets.

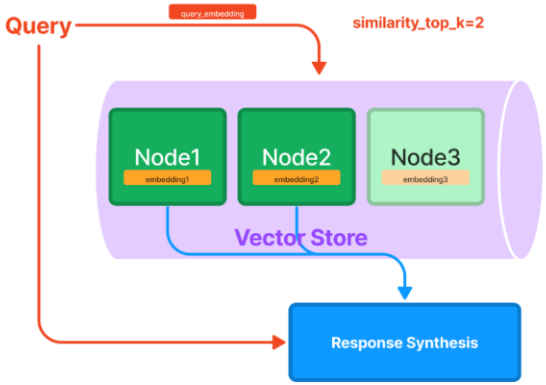
# OBJECTIVE 1 - *Vector database --Chunking Strategies*

## Results on ICRC dataset [real data : lower quality]



# OBJECTIVE 2 - *Retriever*




https://docs.llamaindex.ai/en

## Vector Store Index

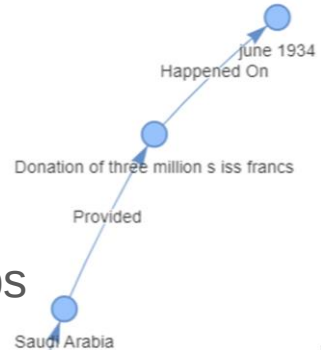
- Store each embedding.
- Fetch the top-k most similar chunks.

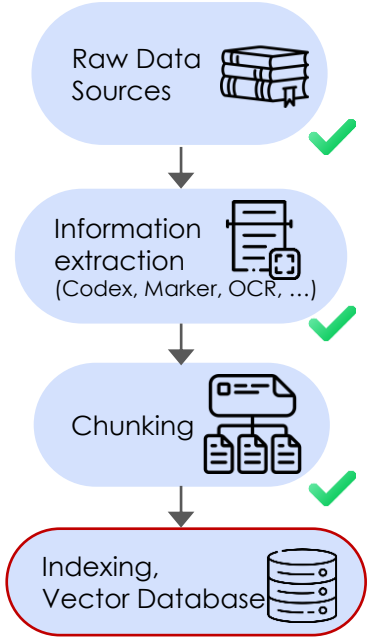
## Best Matching 25 (BM25)

- Ranks documents based on query term occurrence and rarity across the corpus.

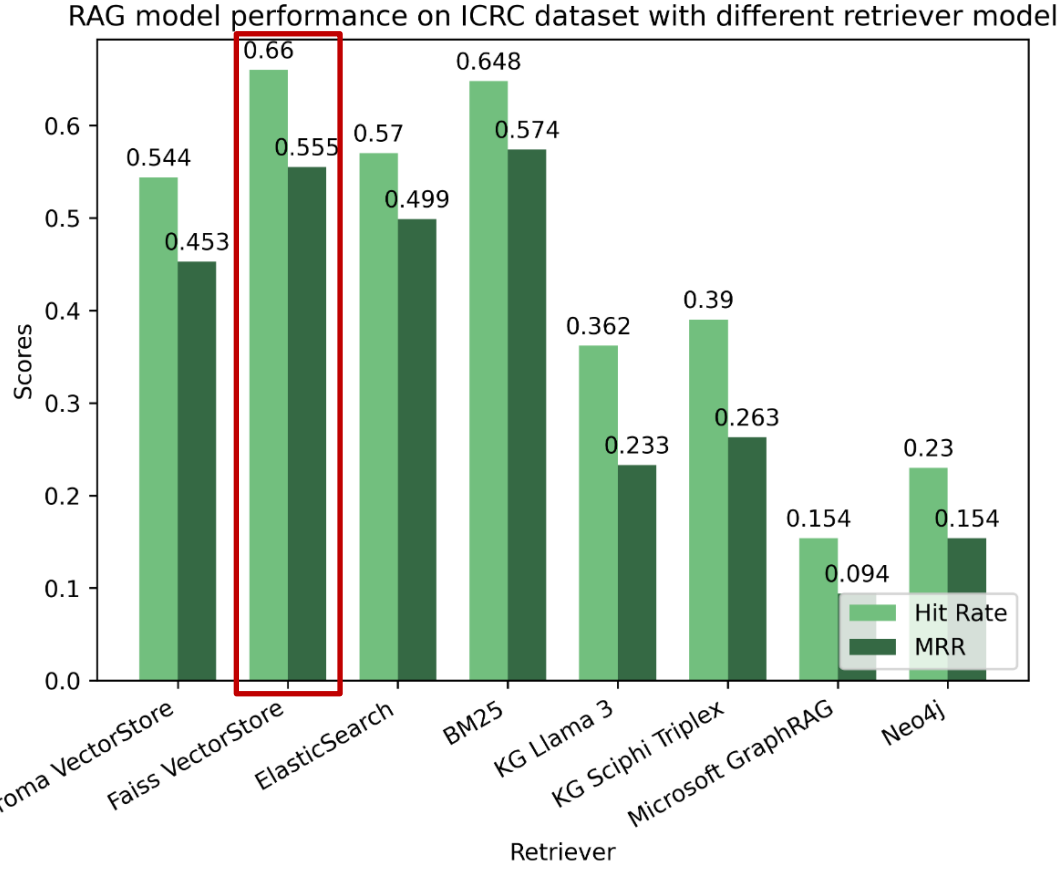
## Knowledge Graph

- Extract a knowledge graph using triplets.
- Query engine uses keywords from the query to retrieve relevant KG entities + explores relationships

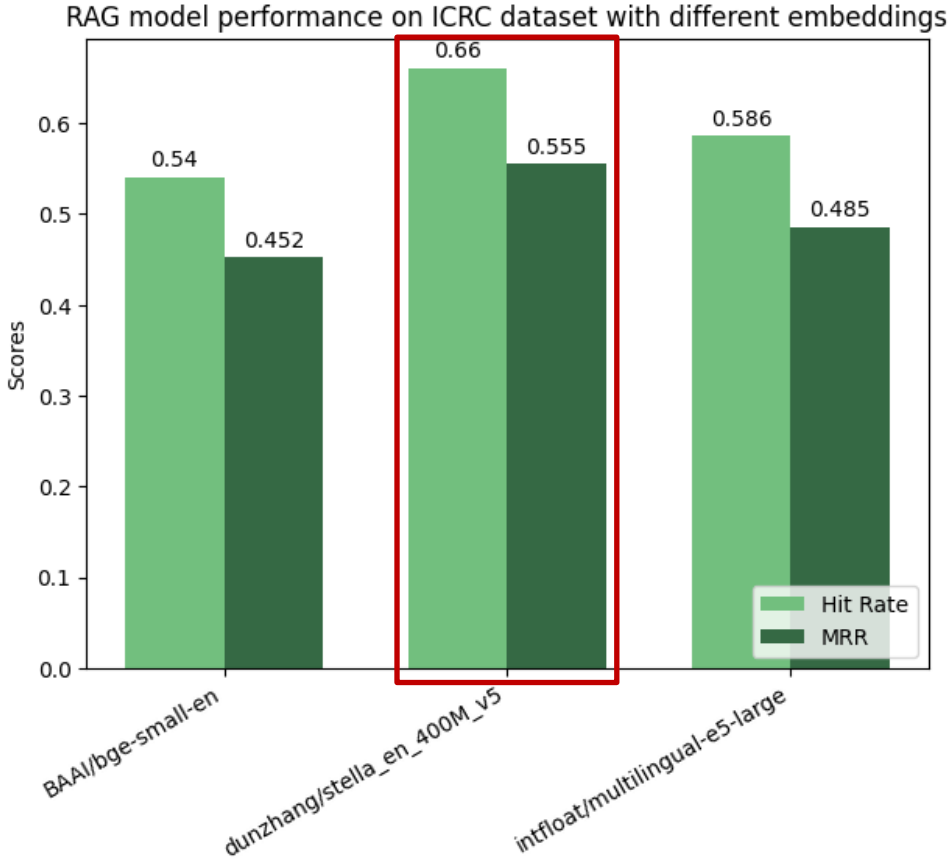


# OBJECTIVE 2 - *Retriever*



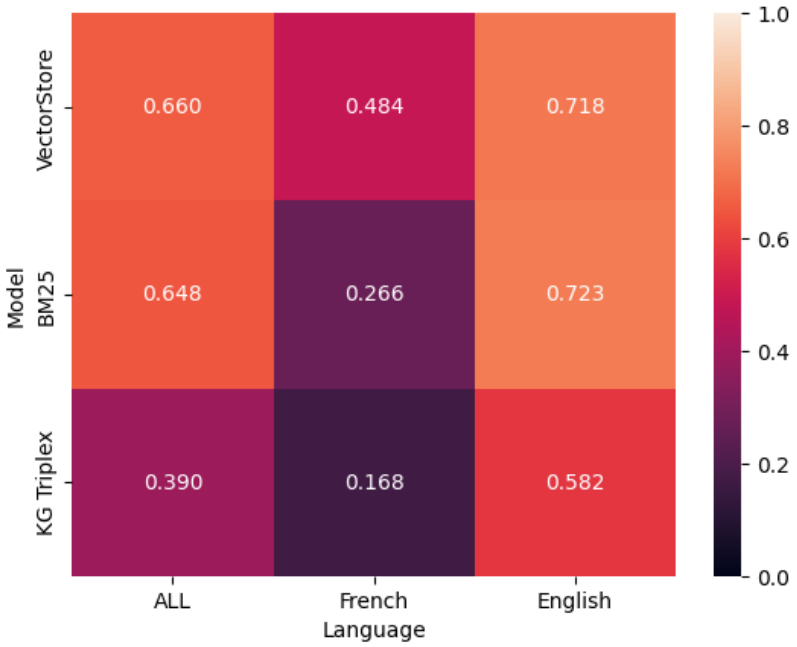
# OBJECTIVE 1 - *Vector database --Embeddings*



# OBJECTIVE 2 - *Retriever*

**Limitations :**

- Multiple languages (Non-English)
- Performance on broad questions?

# OBJECTIVE 2 - *Retriever*

## Precise questions :

data chunks → Llama 3 8b

Documents retrieved → Hit Rate + MRR

- 'How many Iranian prisoners were handed over to the International Committee of the Red Cross (ICRC) on January 29, 1984 at the Ankara airport in Turkey?'
- 'What was the amount donated by Saudi Arabia to the Red Cross in June 1934, and how does this donation contribute to the ICRC's activities worldwide?'

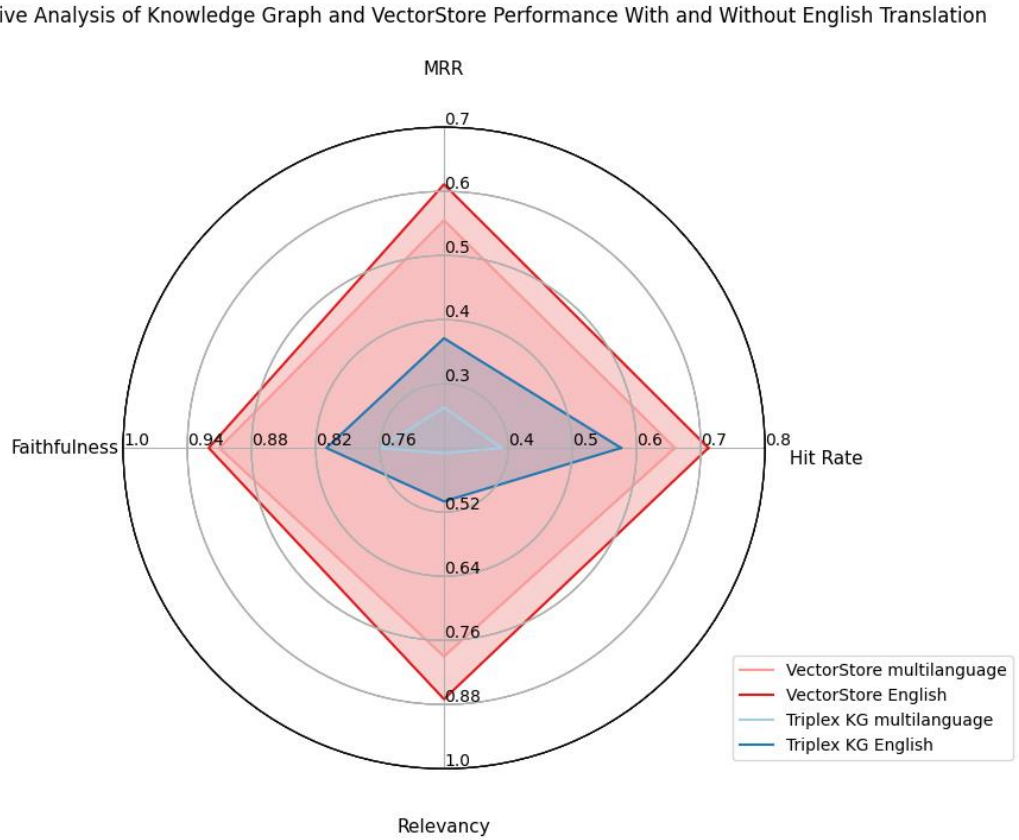
## Broad questions :

LDA (gensim) → List of word representing topics → Llama 3 8b

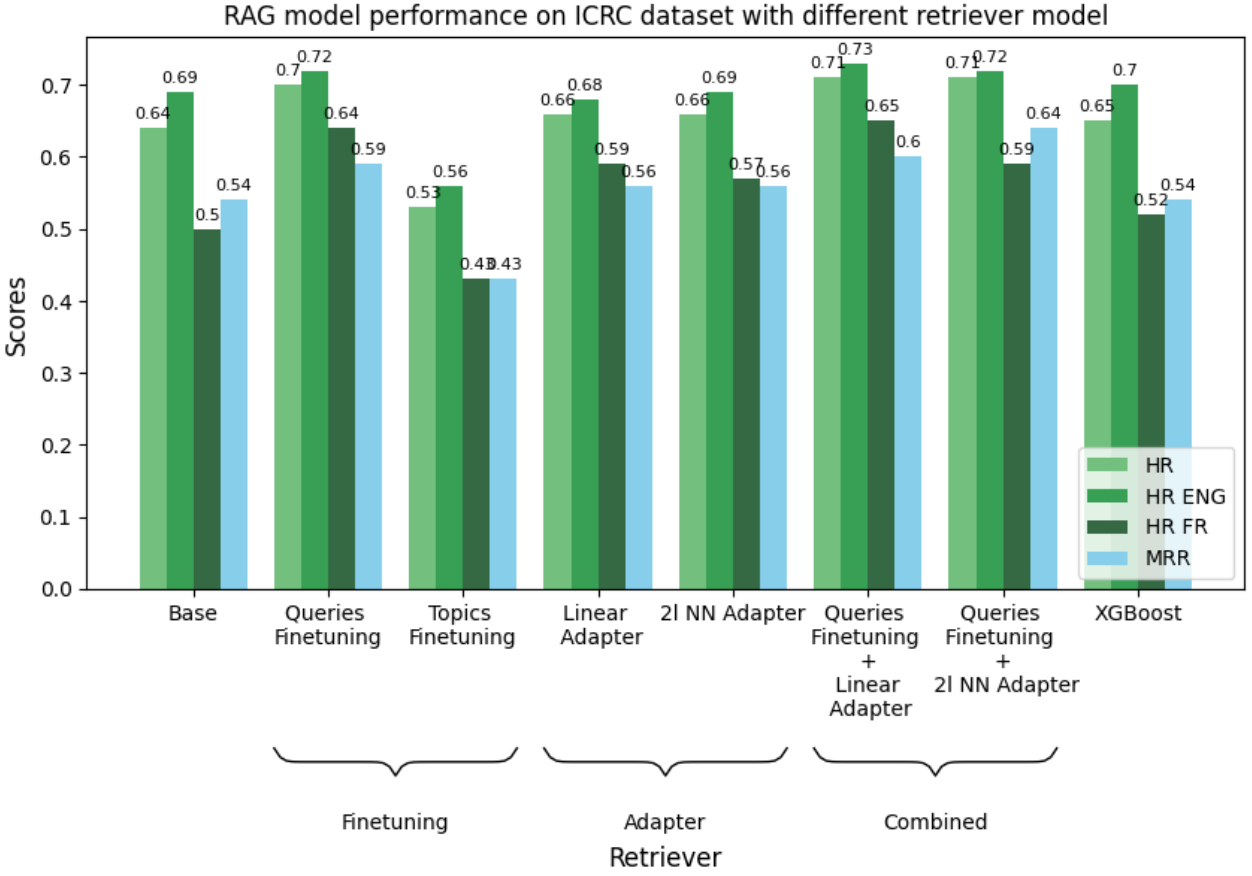
Documents retrieved → Summarizer head (generative llm) → Relevancy + Faithfulness

- 'How does the ICRC use the Restoring Family Links (RFL) program to reunite families who have been separated by conflicts or disasters?'
- 'How does the ICRC support victims in Colombia?'

# OBJECTIVE 2 - *Retriever*

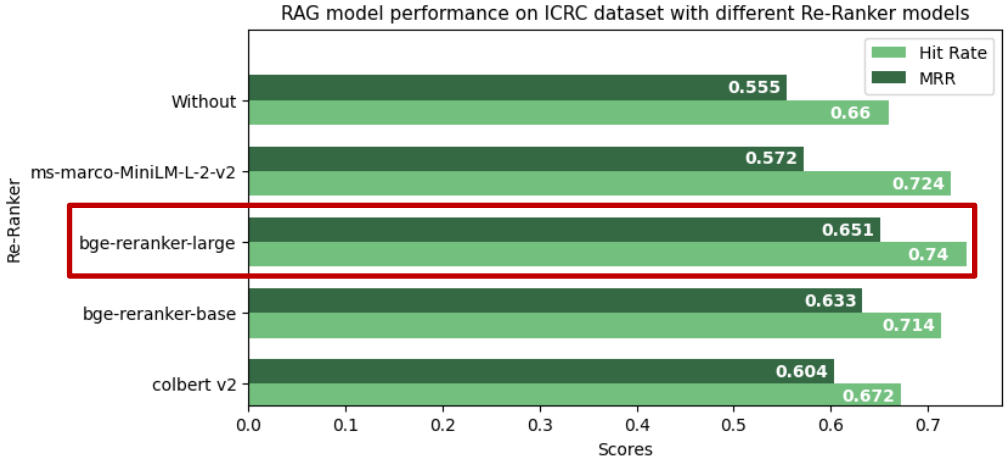


# OBJECTIVE 2 - *Retriever*



# OBJECTIVE 3 – *Re-Ranker*

- **ColBERT** v2 (Contextualized Late Interaction over BERT): A BERT-based search method utilizing late interaction for enhanced retrieval.
- **BGE Base**, **Large,** and **MS MARCO MiniLM** : Cross-encoders for multilingual/English language understanding.



# OBJECTIVE 3 – *Summarizer head*

| | HR | MRR | |
|---|---|---|---|
| Precise queries | 0.742 | 0.668 | |
| | Summarizer head | Faithfulness | Relevancy |
| Precise queries | Llama 3 8b | 0.93 | 0.98 |
| | Mistral 7b | 0.84 | 0.97 |
| | GPT2 | 0.32 | 0.24 |
| Broad queries | Llama 3 8b | 0.91 | 0.88 |
| | Mistral 7b | 0.89 | 0.90 |
| | GPT2 | 0.22 | 0.17 |

# DISCUSSION - *Take home messages*

- Character-based chunking methods work well when paragraphs separate subjects but often struggle to identify subjects within the text. Semantic pairing and chaining chunking methods produce better results. The LDA-based chunking also showed strong performance with lower computational demands.

- Fine-tuning the embedding model or using adapters significantly reduces the language performance gap, improving performance by capturing nuanced linguistic and domain-specific features.

- Using an appropriate reranker and embedding model significantly improves performance.

# DISCUSSION - *Take home messages*

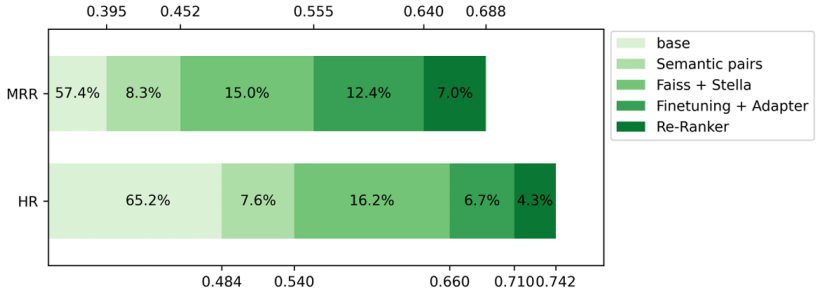



**Figure 12:** Improvements made at each stage of constructing the RAG model. The "base" model employs simple recursive chunking with retrieval using *bge-small-en*. The percentages reflect the level of improvement achieved at each step. The top axis represents MRR performance, while the bottom axis shows HR.

- **CONCLUSION :** The proposed RAG model offers an initial exploration of an efficient solution for handling large, multilingual, domain-specific archives of NGO documents.

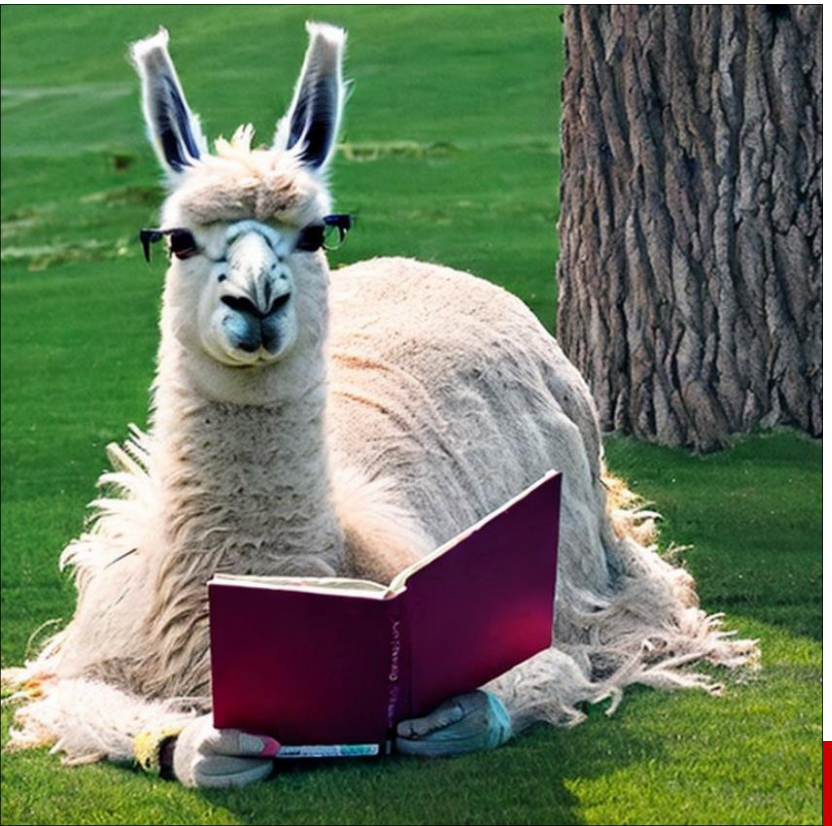
# DISCUSSION *- Limitations*

- **Limitation of this study :**
  - Data quality. ( Codex ☺ )
  - The test set may not always represent the questions ICRC workers need to ask.
  - Real-World Relevance; it would be valuable to use real questions and documents used from ICRC staff for RAG + LLM evaluation.
  - Resource Intensity : Knowledge graph, LLM inference with a lot of data.

# FUTURE WORK

- **Document Selection Mechanism**: Implementing a document selection process before RAG to improve retrieval relevance for specific queries.
- **Fine-Tuning Triplex:** Enhancing graph-based RAG performance by fine-tuning Triplex (linguistic and domain-specific features), and exploring hybrid retrieval strategies using both embeddings and knowledge graphs.
- **Summarization Head Improvements:** Fine-tuning different summarization models (e.g., LLaMA 3 8B, GPT) to improve summary accuracy and relevance for ICRC needs.
- **RLHF:** Exploring Reinforcement Learning from Human Feedback (RLHF) to refine RAG systems and capture the specific requirements of humanitarian queries.
- **Comparison with Commercial Models:** Evaluating commercially available models like GPT-4 and Ada embeddings to set a baseline for assessing future fine-tuned models.

# QUESTIONS



Thank you

Any question? !?

# APPENDIX

Prompting template used for generate the queries for both evaluation pipeline

| Precise queries prompt | Broad queries prompt |
|---|---|
| """<br>Context information is below.<br>---------------------<br>*context_str*<br>---------------------<br>Given the context information and not prior knowledge, generate only questions based on the below query. You are a Professor. Your task is to setup *num_questions_per_chunk* questions for an upcoming quiz/examination. The questions should be developped and diverse in nature across the document. The questions should not contain options, not start with Q1/ Q2.<br>Restrict the questions to the context information provided.<br>""" | """<br>Context information about the International Committee of the Red Cross (ICRC) and their interventions is below.<br>---------------------<br>*context_str*<br>---------------------<br>You are a Teacher/ Professor. Your task is to setup *num_question* questions for an upcoming quiz/examination. Given the context information and not prior knowledge, generate only questions based on the previous keyword themes.<br>""" |

# APPENDIX

Prompting template used for generate the topics used for training XGboost and finetuning based on topics.

| Topic extraction prompting template |
| --- |
| """<br>You are an AI assistant specialized in classifying PDF documents into broad categories. Your task is to analyze the given text and provide a single tag that best describes the content or purpose of the document. Use the most appropriate tag from the following list, or create a new tag if necessary: education, government, science, health, legal, environment, finance, social-sciences, history, human-rights, community-development, humanitarian-aid, advocacy, sustainability, migration-and-refugees, social-justice, crisis, war.<br><br>Respond with only the tag, nothing else. Here are some examples:<br><br>1. "February 7, 1984 lebanon : the icrc calls for immediate ceasefire. Since the deterioration of the situation in lebanon over the last few days, the civilian population has sustained hundreds of victims, dead and wounded, particularly in the south of the capital, beirut."<br>Humanitarian Aid<br><br>2. "M. comelio sommaruga, président du comité international de la croix-rouge (cicr) donnera une conférence de presse le lundi 8 février 1988 à 10 heures 30 m. sommaruga présentera le bilan des principales actions conduites en 1987 ainsi que les objectifs et les grandes options de l'institution pour l'année en cours."<br>Advocacy<br><br>Now, please classify the following text:<br>*context_str*<br>""" |

# APPENDIX

| Chunking | max chunk size | Pourcent thresl | buffer size | overlap | model | Hit Rate | MRR | model | Hit Rate | MRR | model | Hit Rate | MRR | nodes nb |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sentence splitter | 512 | None | None | 20 | bge - None | 0.918 | 0.814 | BM25-None | 0.882 | 0.778 | BM25-bge | 0.916 | 0.857 | 11717 |
| Sentence splitter | 1024 | None | None | 20 | bge - None | 0.908 | 0.782 | BM25-None | 0.866 | 0.776 | BM25-bge | 0.9 | 0.822 | 6057 |
| Sentence splitter | 1024 | None | None | 0 | bge - None | 0.908 | 0.783 | BM25-None | 0.838 | 0.740 | BM25-bge | 0.900 | 0.853 | 11640 |
| Recursive splitter | 512 | None | None | 20 | bge - None | 0.924 | 0.880 | BM25-None | 0.772 | 0.694 | BM25-bge | 0.832 | 0.812 | 69 025 |
| Recursive splitter | 1024 | None | None | 20 | bge - None | 0.928 | 0.853 | BM25-None | 0.784 | 0.715 | BM25-bge | 0.886 | 0.855 | 33319 |
| Recursive splitter | 1024 | None | None | 0 | bge - None | 0.896 | 0.857 | BM25-None | 0.716 | 0.662333 | BM25-bge | 0.784 | 0.768 | 68130 |
| Semantic | None | 90 | 2 | 0 | bge - None | 0.77 | 0.707 | BM25-None | 0.644 | 0.576333 | BM25-bge | 0.712 | 0.693 | 27611 |
| Semantic + max chunk size | 2048 | 90 | 2 | 0 | bge - None | 0.784 | 0.724 | BM25-None | 0.71 | 0.643 | BM25-bge | 0.774 | 0.743 | 26053 |
| Semantic + min length | None | 90 | 2 | 0 | bge - None | 0.839 | 0.780 | BM25-None | 0.70 | 0.618 | BM25-bge | 0.755 | 0.746 | 21008 |
| Semantic pairs + min length | None | 80 | 2 | 0 | bge - None | 0.862 | 0.805 | BM25-None | 0.698 | 0.621667 | BM25-bge | 0.776 | 0.751 | 35059 |
| Semantic pairs + min length + para | None | 95 | 2 | 0 | bge - None | 0.938 | 0.846 | BM25-None | 0.818 | 0.728 | BM25-bge | 0.872 | 0.831 | 18184 |
| Semantic pairs + min length | None | 90 | 2 | 0 | bge - None | 0.9 | 0.814 | BM25-None | 0.81 | 0.723 | BM25-bge | 0.896 | 0.87 | 15937 |
| Semantic pairs + min length | None | 95 | 2 | 0 | bge - None | 0.896 | 0.822 | BM25-None | 0.8 | 0.698 | BM25-bge | 0.858 | 0.818 | 11141 |
| Chain | None | None | 2 | 0 | bge - None | 0.924 | 0.89 | BM25-None | 0.776 | 0.719 | BM25-bge | 0.854 | 0.838 | 16095 |
| Chain | None | None | 1 | 0 | bge - None | 0.916 | 0.846 | BM25-None | 0.82 | 0.733 | BM25-bge | 0.869 | 0.823 | 39138 |
| Chain + para | 2048 | 0.4 | 2 | 0 | bge - None | 0.958 | 0.86 | BM25-None | 0.874 | 0.785 | BM25-bge | 0.93 | 0.878 | 12225 |
| LDA | None | None | None | 0 | bge - None | 0.92 | 0.826 | BM25-None | 0.83 | 0.751 | BM25-bge | 0.892 | 0.86 | 14525 |

**Figure 13:** Test results for HR and MRR using various chunking methods on the WikiText dataset from Hugging Face ("Salesforce/wikitext"). The evaluation involved 500 precise questions generated by the LLaMA 3 8B model based on these documents. The notation model refers to the embedding model name followed by the re-ranker name. bge represents the bge-small-en embedding model. The green rows represent the baseline, where paragraph separation is followed by method-based splitting. The purple row highlight the best-performing method. Nodes nb refers to the number of chunks created by the splitting method.

# APPENDIX

| Chunking | model | Hit Rate | MRR | model | Hit Rate | MRR | model | Hit Rate | MRR | nodes nb |
|---|---|---|---|---|---|---|---|---|---|---|
| agentic + semantic | bge - None | 0.969 | 0.896 | BM25-None | 0.908 | 0.821 | BM25-bge | 0.962 | 0.937 | 130 |
| LLM prompting | bge - None | 0.981 | 0.902 | BM25-None | 0.955 | 0.872 | BM25-bge | 0.970 | 0.948 | 130 |

**Figure 14:** Test results for HR and MRR using different chunking methods on a subset of 100 WikiText articles from Hugging Face ("Salesforce/wikitext"). This subset was selected due to computational constraints. The evaluation was conducted on 100 precise questions generated by the LLaMA 3 8B model based on these articles. In the results, the term model indicates the embedding model followed by the re-ranker name, with bge representing the bge-small-en embedding model.

| Chunking | model | Hit Rate | MRR | model | Hit Rate | MRR | model | Hit Rate | MRR | model | Hit Rate | MRR | nodes nb |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Recursive 2048 | bge - None | 0.484 | 0.395 | bge-bge | 0.484 | 0.439 | BM25-None | 0.574 | 0.506 | BM25-bge | 0.622 | 0.559 | 28998 |
| Semantic  b=1 2048 | bge - None | 0.518 | 0.443 | bge-bge | 0.61 | 0.554 | BM25-None | 0.602 | 0.526 | BM25-bge | 0.632 | 0.564 | 20835 |
| Semantic pairs 1024 | bge - None | 0.54 | 0.452 | bge-bge | 0.612 | 0.547 | BM25-None | 0.648 | 0.574 | BM25-bge | 0.66 | 0.586 | 23823 |
| Semantic pairs 2048 | bge - None | 0.522 | 0.439 | bge-bge | 0.614 | 0.555 | BM25-None | 0.608 | 0.526 | BM25-bge | 0.646 | 0.584 | 20420 |
| Chain 1024 | bge - None | 0.49 | 0.41 | bge-bge | 0.568 | 0.508 | BM25-None | 0.568 | 0.499 | BM25-bge | 0.61 | 0.537 | 21450 |
| Chain  2048 | bge - None | 0.532 | 0.444 | bge-bge | 0.61 | 0.545 | BM25-None | 0.616 | 0.54 | BM25-bge | 0.642 | 0.579 | 21333 |
| LDA 2048 | bge - None | 0.5 | 0.433 | bge-bge | 0.578 | 0.531 | BM25-None | 0.612 | 0.523 | BM25-bge | 0.64 | 0.583 | 23481 |

**Figure 15:** Test results for HR and MRR using various chunking methods on the ICRC dataset. The evaluation involved 500 precise questions generated by the LLaMA 3 8B model based on these documents. The notation model refers to the embedding model name followed by the re-ranker name. bge represents the bge-small-en embedding model. The purple row highlight the best-performing method.

| Embedding model for semantic splitting | model | Hit Rate | MRR | model | Hit Rate | MRR | model | Hit Rate | MRR | model | Hit Rate | MRR | nodes nb |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| w601sxs/b1ade-embed | bge - None | 0.54 | 0.452 | bge-bge | 0.612 | 0.547 | BM25-None | 0.648 | 0.574 | BM25-bge | 0.66 | 0.586 | 23823 |
| intfloat/multilingual-e5-base | bge - None | 0.536 | 0.459 | bge-bge | 0.586 | 0.532 | BM25-None | 0.620 | 0.556 | BM25-bge | 0.648 | 0.587 | 23844 |
| Lajavaness/bilingual-embedding-large | bge - None | 0.506 | 0.416 | bge-bge | 0.582 | 0.519 | BM25-None | 0.608 | 0.525 | BM25-bge | 0.62 | 0.557 | 23795 |
| dunzhang/stella_en_400M_v5 | bge - None | 0.536 | 0.44 | bge-bge | 0.596 | 0.527 | BM25-None | 0.61 | 0.532 | BM25-bge | 0.64 | 0.565 | 23806 |

**Figure 16:** Test results for HR and MRR using various embeddings model used for semantic splitting on the ICRC dataset. The evaluation involved 500 precise questions generated by the LLaMA 3 8B model based on these documents. The notation model refers to the embedding model name followed by the re-ranker name. bge represents the bge-small-en embedding model. The purple row highlight the best-performing method.

# APPENDIX

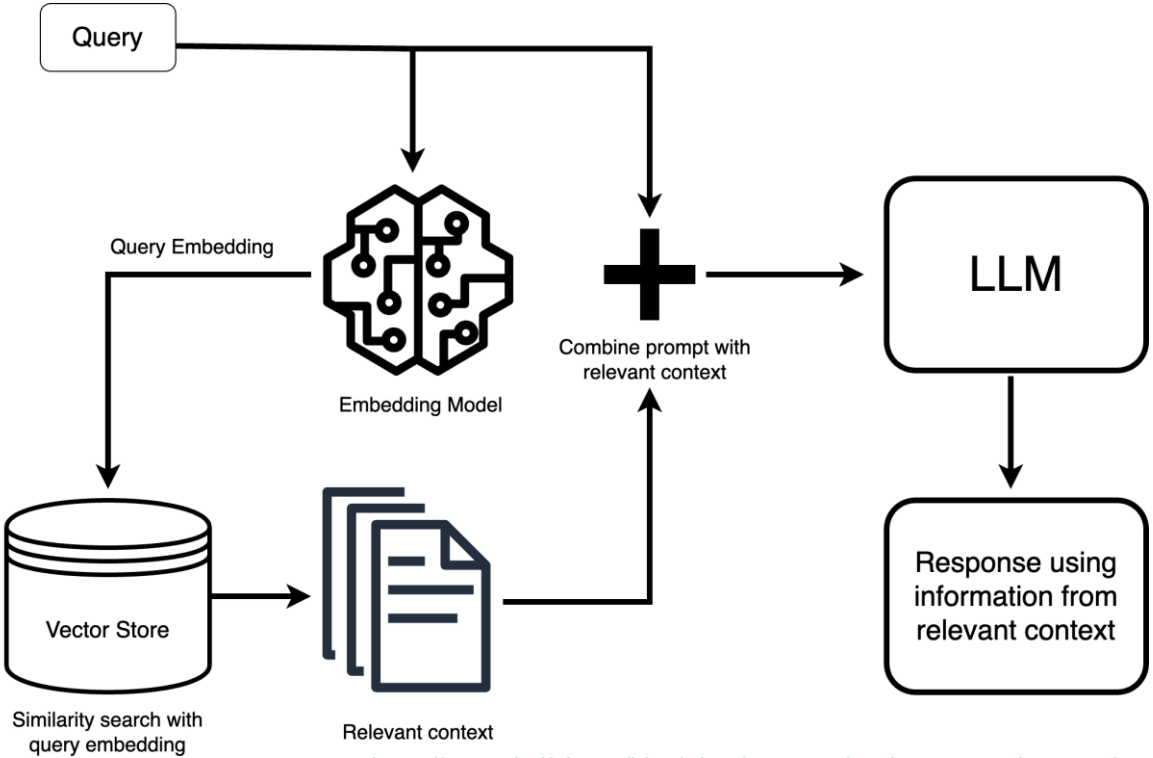
| Retriever Embedding model | Hit Rate | MRR | Reranker | Hit Rate | MRR |
|---|---|---|---|---|---|
| BAAI/bge-small-en | 0.54 | 0.452 | bge | 0.612 | 0.547 |
| dunzhang/stella_en_400M_v5 | 0.66 | 0.555 | bge | 0.714 | 0.633 |
| intfloat/multilingual-e5-large | 0.586 | 0.485 | bge | 0.64 | 0.575 |

**Figure 17:** Test results for HR and MRR using various embeddings model used for retrieval on the ICRC dataset. The evaluation involved 500 precise questions generated by the LLaMA 3 8B model based. bge-small-en was the embedding model used. The purple row highlight the best-performing method.

| Model | epochs | batch size | learning rate | HR | MRR |
|---|---|---|---|---|---|
| Base | / | / | / | 0.638 | 0.541 |
| Queries Finetuned | 4 | 10 | 2,00E-05 | 0.7 | 0.588 |
| Queries Finetuned | 2 | 10 | 2,00E-05 | 0.681 | 0.589 |
| Queries Finetuned | 4 | 4 | 2,00E-05 | 0.671 | 0.574 |
| Topics Finetuned | 4 | 10 | 2,00E-05 | 0.526 | 0.432 |
| Linear Adapter | 4 | 10 | 2,00E-05 | 0.656 | 0.558 |
| Linear Adapter | 4 | 10 | 0.01 | 0.336 | 0.256 |
| 2 layers NN Adapter | 25 | 10 | 2,00E-05 | 0.639 | 0.543 |
| 2 layers NN Adapter | 4 | 10 | 2,00E-05 | 0.662 | 0.563 |
| XGBoost | 4 | 10 | 2,00E-05 | 0.65 | 0.54 |

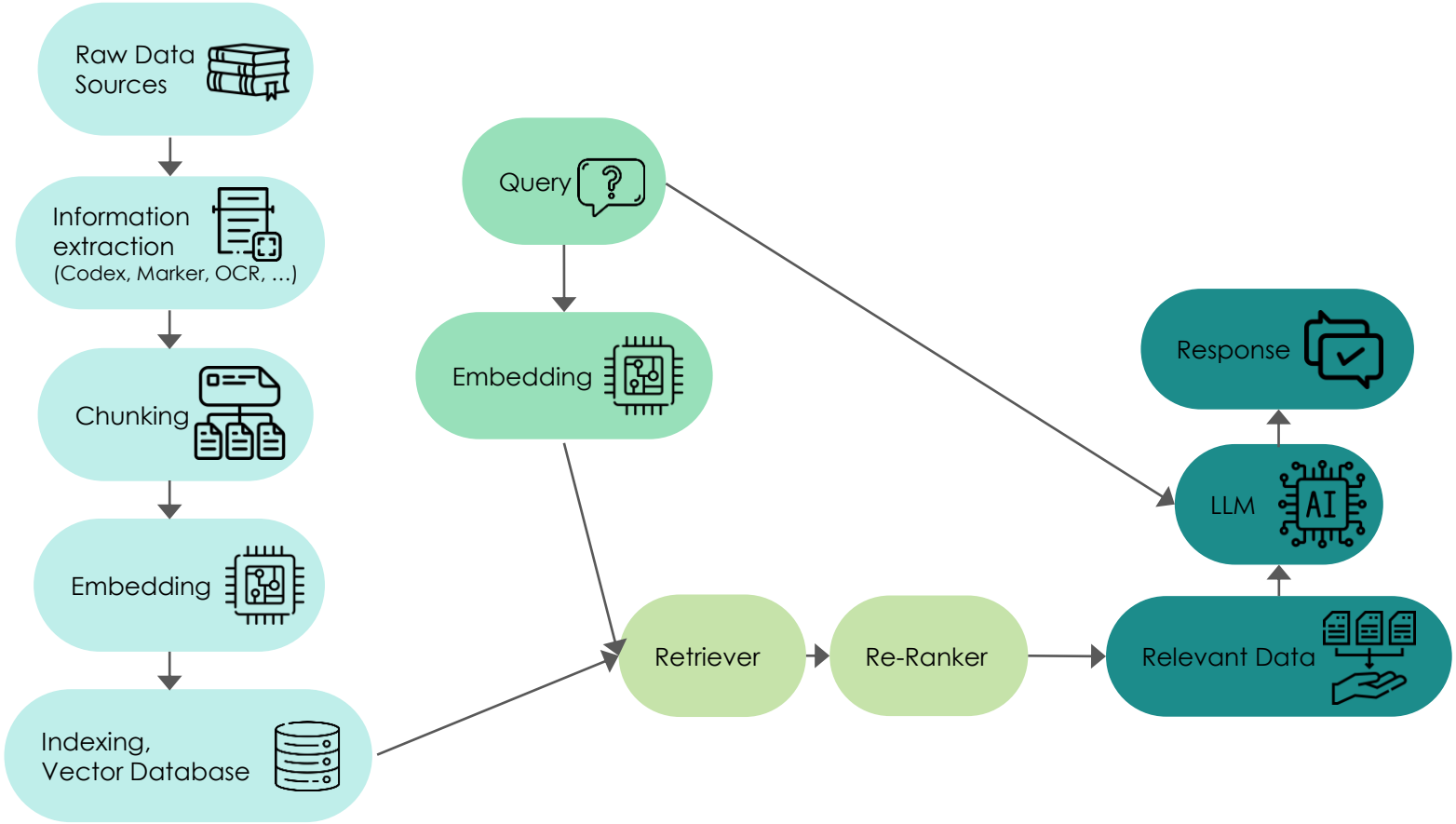
**Figure 18:** Test results for HR and MRR with different improvements to the embedding model used for retrieval on the ICRC dataset. The evaluation was conducted using 500 precise questions generated by the LLaMA 3 8B model, based on these documents. A 7:3 train test split was used. Stella en 400M v5 was the embedding model used. The purple row highlights the best-performing method.

# APPENDIX

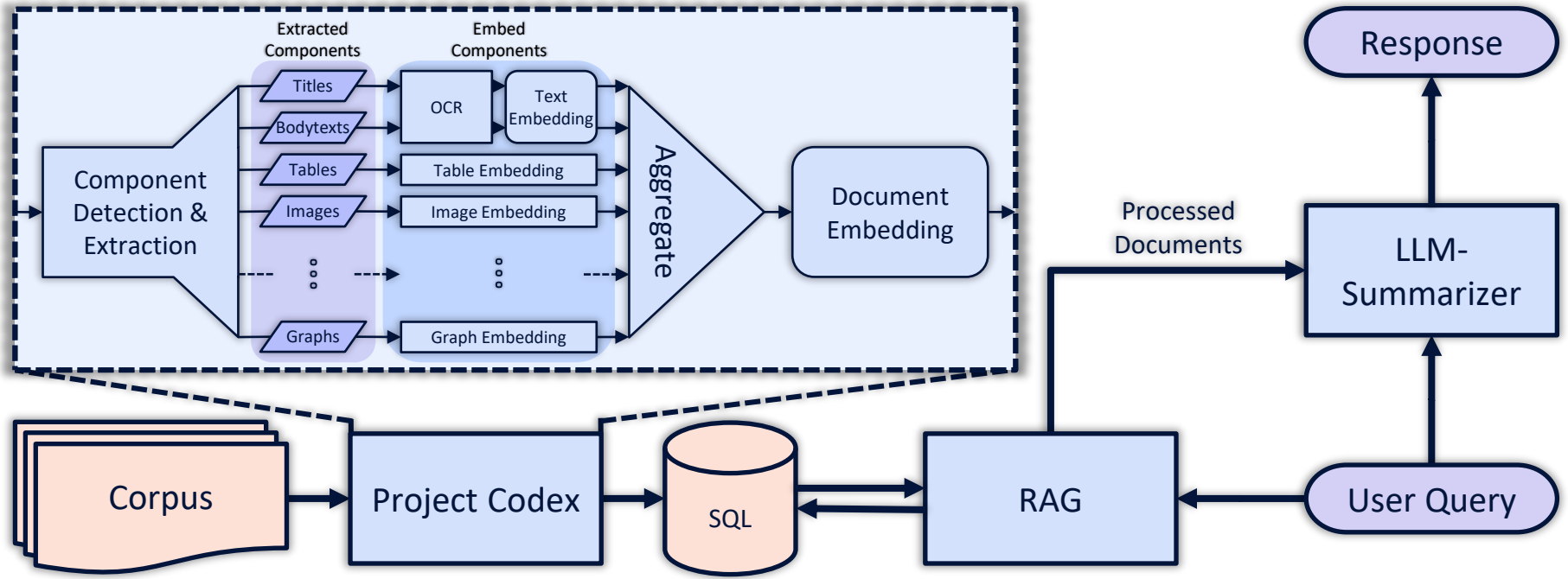



https://www.clarifai.com/blog/what-is-rag-retrieval-augmented-generation

# APPENDIX


Raw Data Sources
Information extraction (Codex, Marker, OCR, ...)
Chunking
Embedding
Indexing, Vector Database
Query
Embedding
Retriever
Re-Ranker
Relevant Data
LLM
AI
Response

# APPENDIX


Extracted Components
Embed Components
Titles
Bodytexts
Tables
Images
Graphs
OCR
Text Embedding
Table Embedding
Image Embedding
Graph Embedding
Component Detection & Extraction
Aggregate
Document Embedding
Response
Processed Documents
LLM-Summarizer
Corpus
Project Codex
SQL
RAG
User Query

# APPENDIX

BM 25 :

$$BM25 = \sum_{t \in q} \log \left[ \frac{N}{df(t)} \right] \cdot \frac{(k_1 + 1) \cdot tf(t, d)}{k_1 \cdot \left[ (1 - b) + b \cdot \frac{dl(d)}{dl_{avg}} \right] + tf(t, d)}$$

- $k_1$, $b$ – parameters
- $dl(d)$ – length of document $d$
- $dl_{avg}$ – average document length

- tf : Term Frequency
- df : Document Frequency

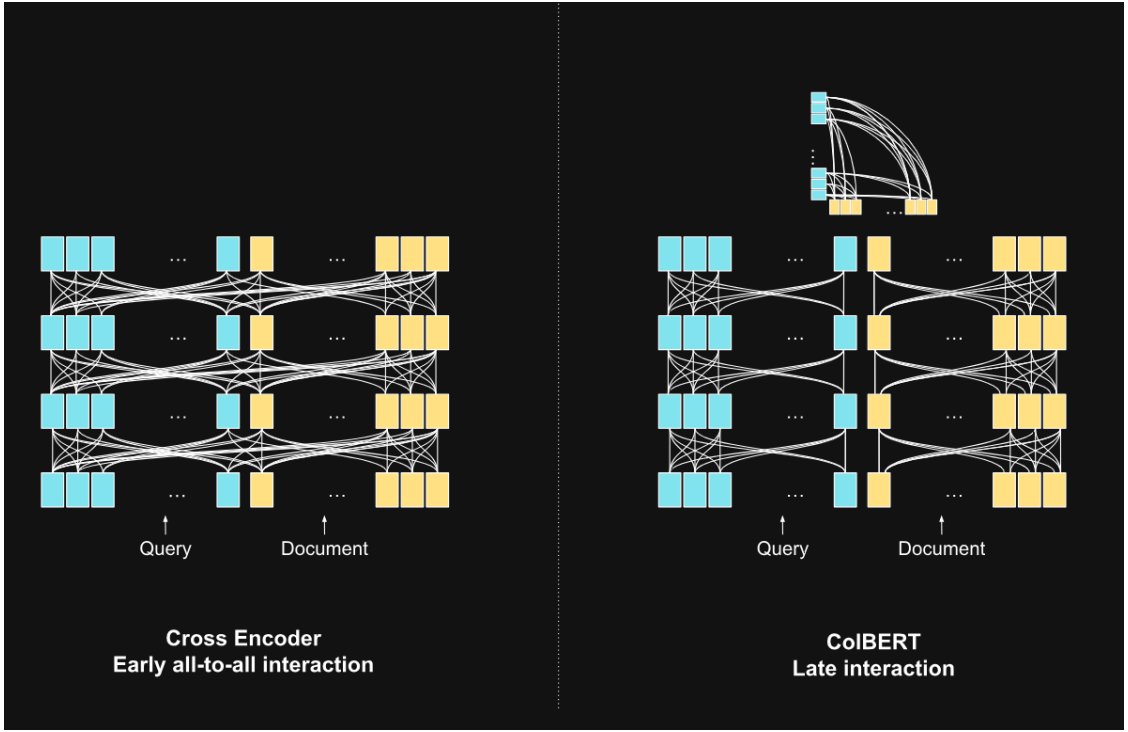
# APPENDIX

**Faiss :**

**1. K-means clustering**: This algorithm breaks the data into clusters, which helps narrow down the search space by focusing on the most relevant clusters during queries.
**2. Product quantization (PQ)**: PQ compresses vectors into shorter codes, reducing memory usage significantly and speeding up the search without a big drop in accuracy.
**3. Optimized product quantization (OPQ)**: An enhanced version of PQ, OPQ rotates the data to better fit the quantization grid, improving the accuracy of the compressed vectors.

# APPENDIX

## Cross encoder and ColBERT




https://jina.ai/news/what-is-colbert-and-late-interaction-and-why-they-matter-in-search/