MEDICAL QUESTION ANSWERING SYSTEM

Presented by Binit KC

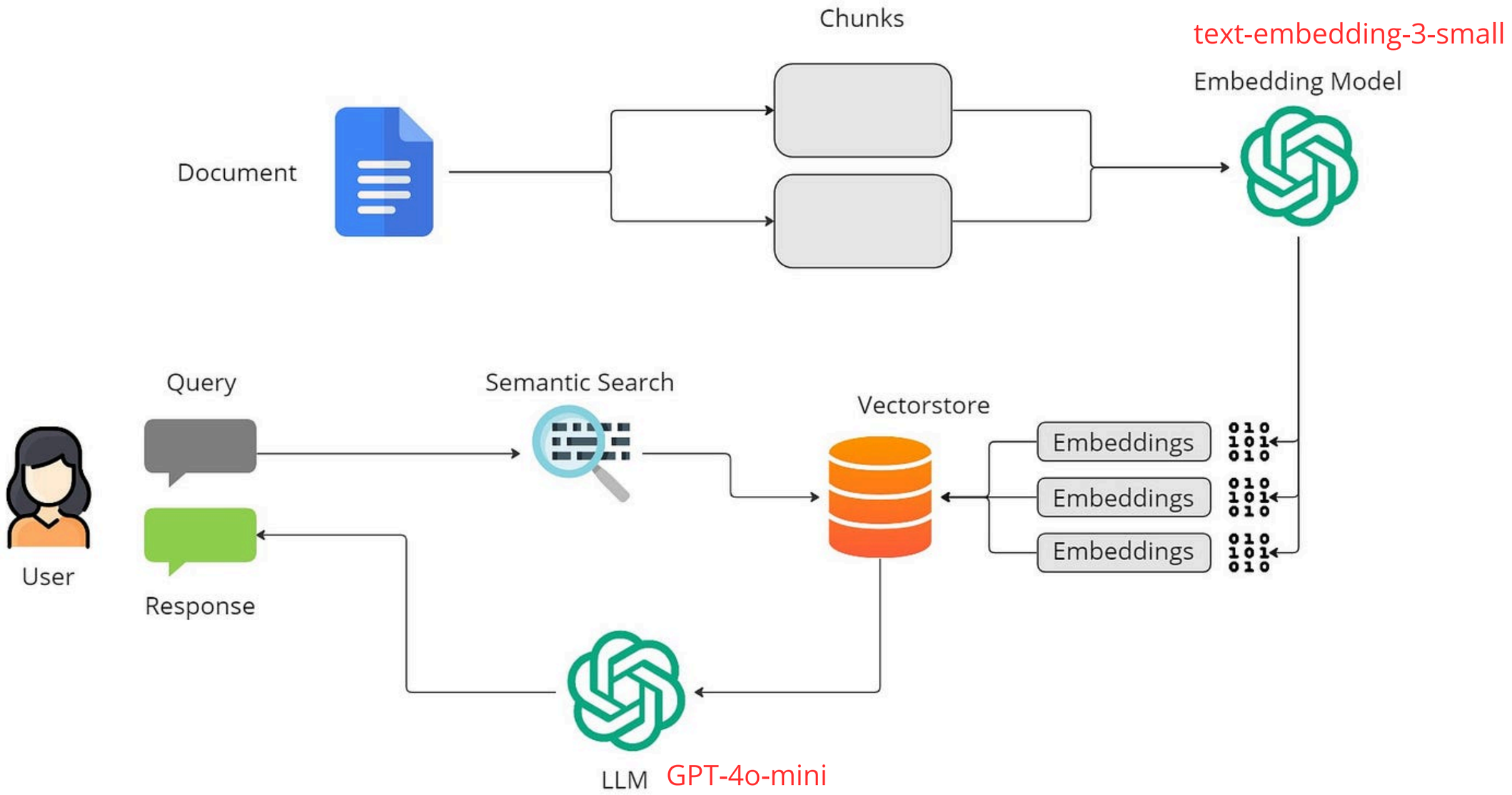
OVERVIEW

- Methodology
- Evaluation
- Problems Faced
- Future Enhancements
- Conclusion

Objectives

- Design effective retrieval strategies for medical content.
- Implement robust reasoning over retrieved contexts.
- Develop systematic evaluation methods.
- Build verifiable answer extraction.

Methodology



Methodology

1.Data collection

2.Document loading and text chunking

3. Embedding Creation and Vectore Store Creation

4. Retriever Selection

5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

7. Evaluation

https://www.nccih.nih.gov/health/providers/clinicalpractice

NIH National Center for Complementary and Integrative Health

Cardiology

- Vitamin, Mineral, and Multivitamin Supplements for the Primary Prevention of Cardiovascular Disease and Cancer *(USPSTF)*
- Soy Protein, Isoflavones, and Cardiovascular Health *(Circulation)*
- Management of Stable Ischemic Heart Disease *(Annals of Internal Medicine)*

Allergy and Immunology

- Guidelines for the Diagnosis and Management of Asthma *(NHLBI)*
- Diagnosis and Management of Food Allergy *(Journal of Allergy and Clinical Immunology)* [165KB PDF]
- Allergic Rhinitis and Its Impact on Asthma (ARIA) Guidelines: 2010 Revision *(Journal of Allergy and Clinical Immunology)*
- Clinical Practice Guideline: Allergic Rhinitis *(American Academy of Otolaryngology—Head and Neck Surgery)* [1KB PDF]

Gastroenterology

- ACG Clinical Guideline: Management of Irritable Bowel Syndrome *(American Journal of Gastroenterology)*
- Probiotics and Children *(Journal of Pediatric Gastroenterology and Nutrition)* [119KB PDF]
- Evidence-Based Recommendations on Management of Irritable Bowel Syndrome *(American Family Physician)*

Pediatrics

- Management of Children with Autism Spectrum Disorders *(Pediatrics)*
- Pediatric Integrative Medicine *(American Academy of Pediatrics)*
- Migraine Headaches in Children and Adolescents *(Journal of Pediatric Health Care)*
- A Practice Pathway for Identification, Evaluation, and Management of Insomnia in Children and Adolescents with Autism Spectrum Disorders *(Pediatrics)*

***selected 20 different guidelines from different medical fields**

Methodology

1.Data collection

2.Document loading and text chunking

3. Embedding Creation and Vectore Store Creation

4. Retriever Selection

5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

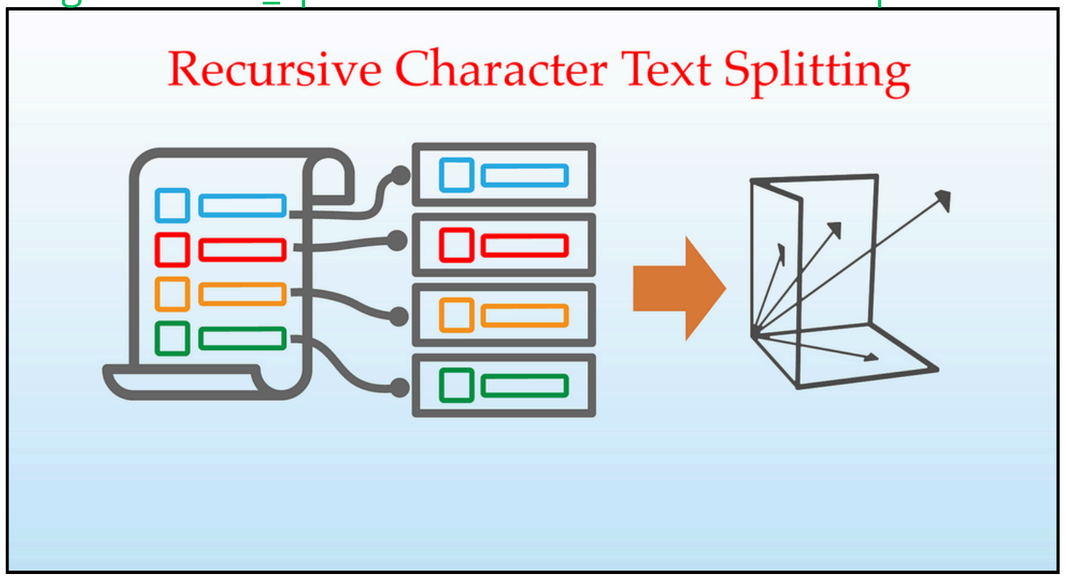
7. Evaluation

langchain.document_loaders.PyPDFLoader

```
1
2   loader = PyPDFLoader(f'/dbfs/mnt/{mount_point}/testfolder/sample.pdf')
3   pages = loader.load()
4   pages[0]
```

```
Out[7]: Document(page_content=' A Simple PDF File \n This is a small demonstration .pdf file - \n just for use in
anics tutorials. More text. And more \n text. And more text. And more text. And more text. \n And more text. And
re text. And more text. And more \n text. And more text. Boring, zzzzz. And more text. And more text. And \n more
ext. And more text. And more text. And more text. \n And more text. And more text. \n And more text. And more tex
And more text. And more \n text. And more text. And more text. Even more. Continued on page 2 ...', metadata={'so
t/pdfs/testfolder/sample.pdf', 'page': 0})
```

langchain.text_splitter.RecursiveCharacterTextSplitter



***Processed 696 pages into 4458 chunks**

Methodology

1.Data collection

2.Document loading and text chunking

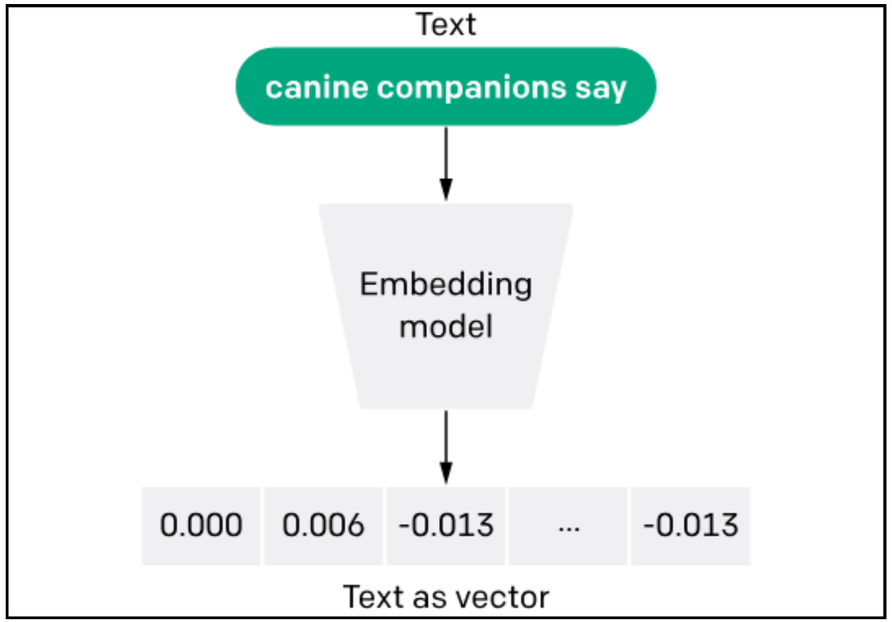
3. Embedding Creation and Vectore Store Creation

4. Retriever Selection

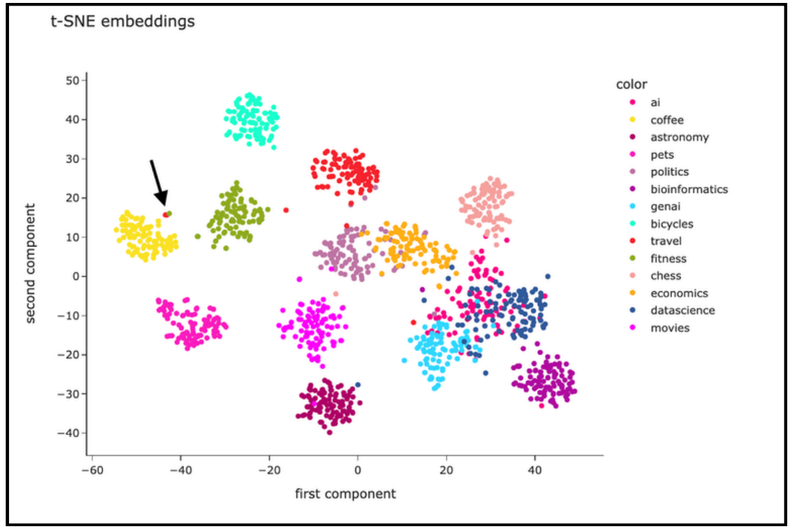
5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

7. Evaluation

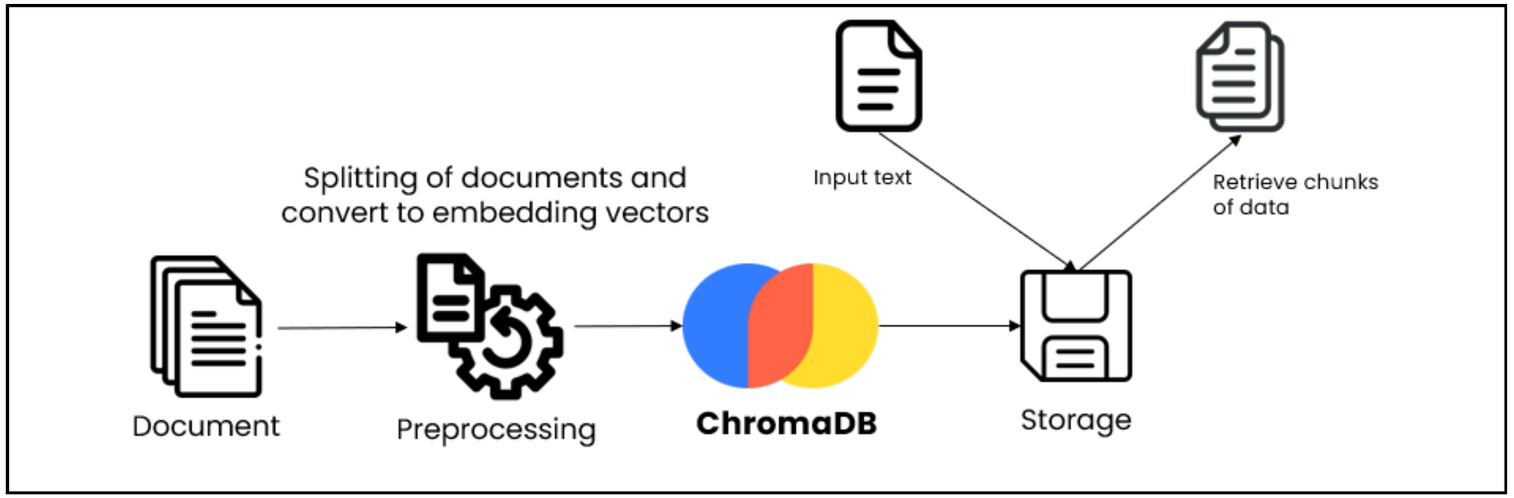


Embeddings capture the semantics of tokens

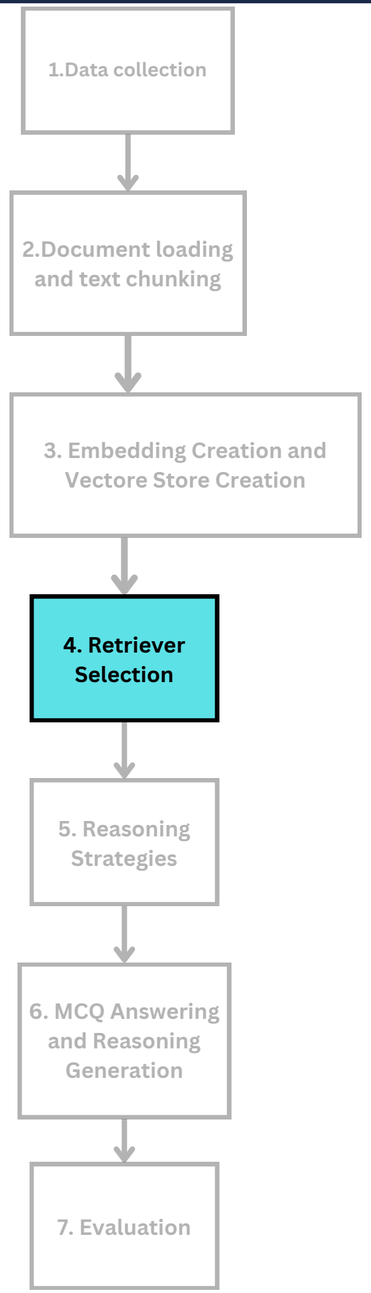


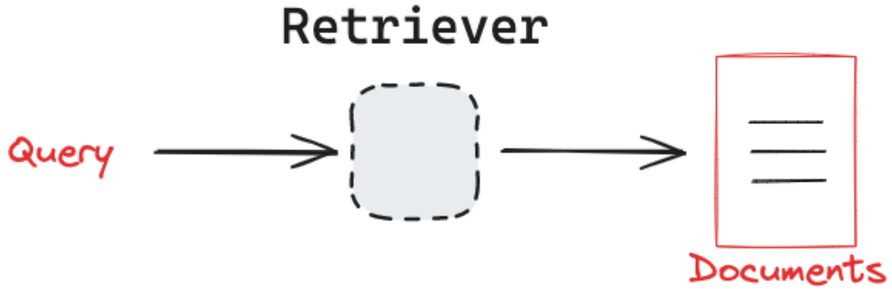
model="text-embedding-3-small" from OpenAIEmbeddings

- lightweight
- cost-effective



Methodology



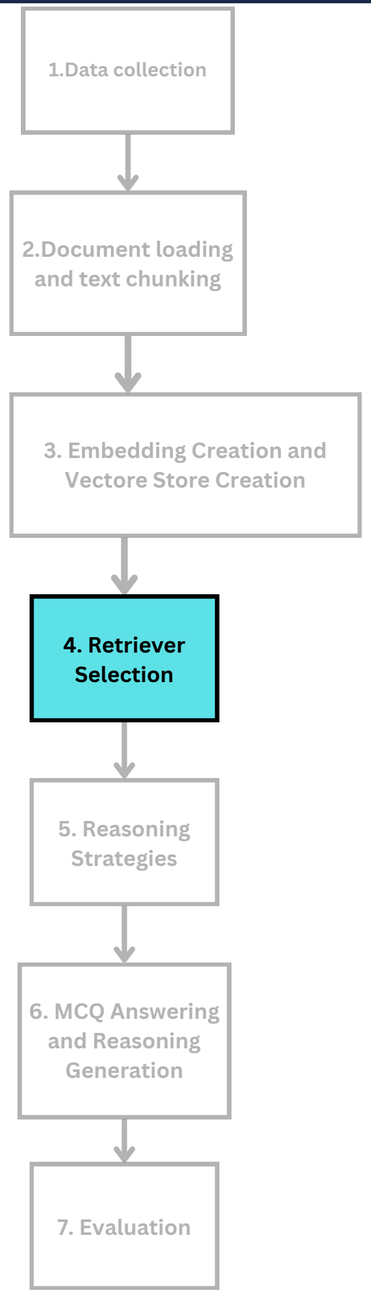


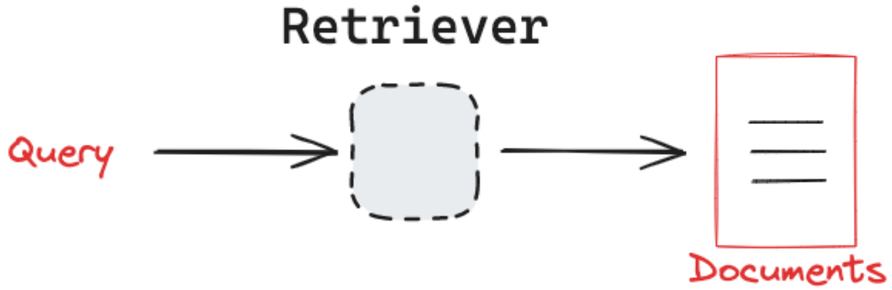
Single-Stage Retriever

A retrieval strategy that uses a single step to fetch relevant documents based on semantic similarity.

- **Working**:
  - Utilizes **Chroma** Vectorstore for storing document embeddings.
  - Performs similarity search by comparing the query's embedding with stored document embeddings.
  - Retrieves the top **k (=3)** documents ranked by similarity scores.
- **Advantages**:
  - **Fast and efficient**: Ideal for straightforward retrieval tasks.
  - **Semantic understanding**: Leverages **dense embeddings** to match the contextual meaning of the query and documents.
- **Use Case**:
  - Suitable for scenarios where **semantic relevance is sufficient without additional reranking or query expansion**.

Methodology





Two-Stage Retriever

- Working:
  - Stage 1: **Multi-Query Expansion**
    - Uses a Language Model (LLM) to generate multiple reformulated queries.
    - Retrieves documents for each query to ensure broader coverage of relevant information.
  - Stage 2: **Reranking**
    - Combines all retrieved documents and ranks them based on semantic similarity scores.
    - Returns the top k most relevant documents.
- Advantages:
  - **Improved Recall:** Expands query coverage to capture diverse aspects of the query.
  - **Enhanced Precision**: Reranking ensures only the most relevant documents are selected.
- Use Case:
  - Suitable for scenarios requiring deeper retrieval with a broader context and higher accuracy, such as medical or legal document retrieval.

Methodology

1.Data collection

2.Document loading and text chunking

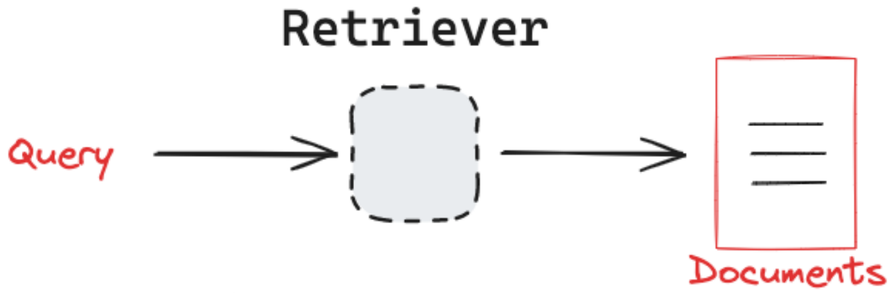
3. Embedding Creation and Vectore Store Creation

4. Retriever Selection

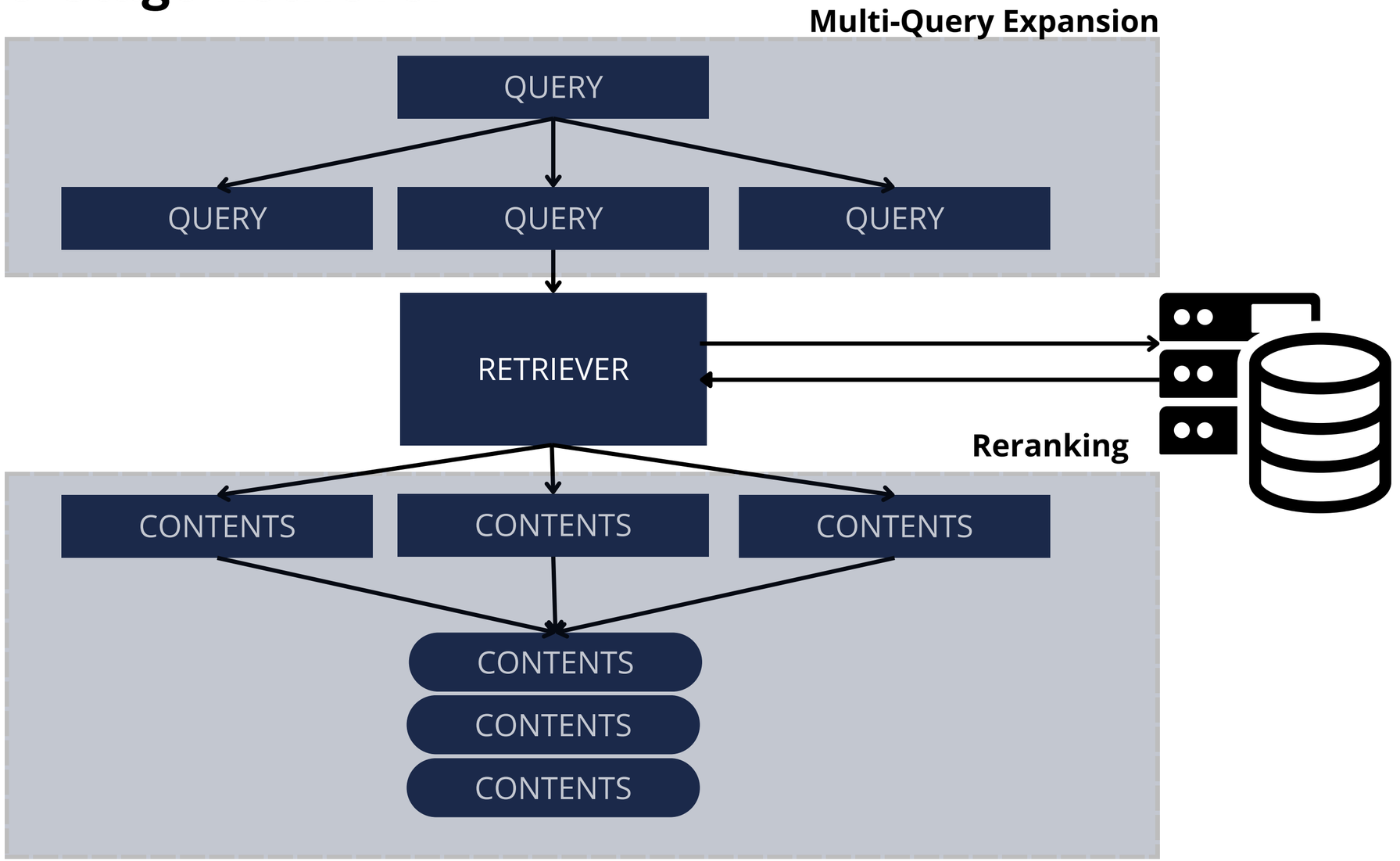
5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

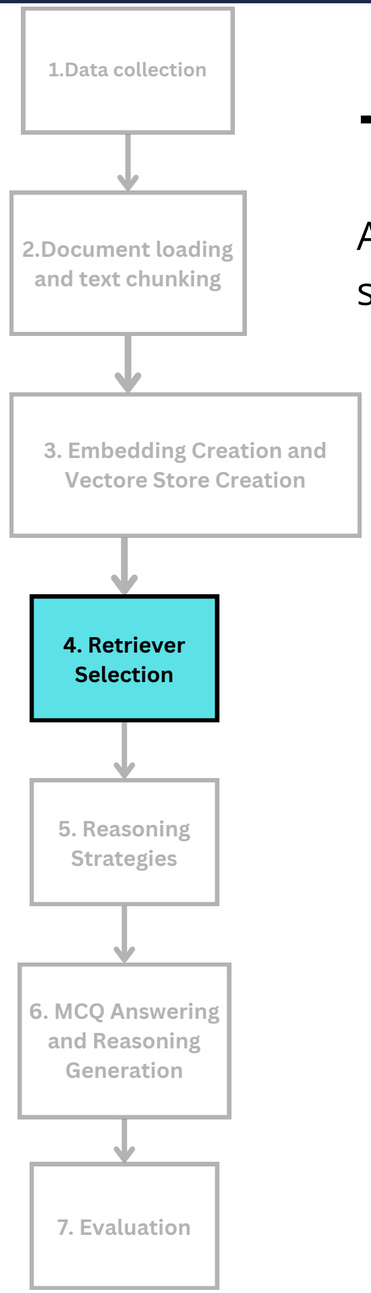
7. Evaluation

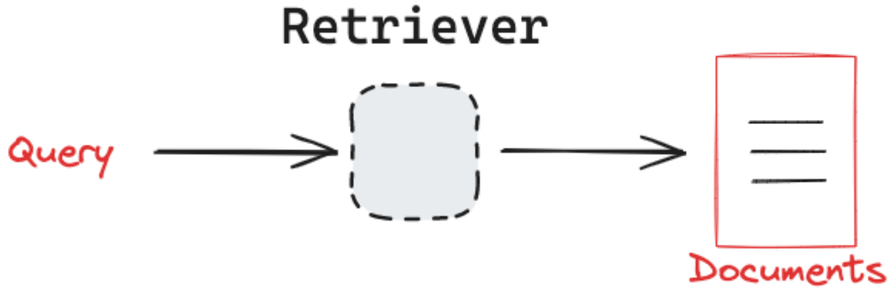


Two-Stage Retriever



Methodology





Three-Stage Retriever

A sophisticated retrieval strategy combining dense, sparse, and hybrid scoring techniques across three stages for maximum relevance.

- Working:
  - **Stage 1: Dense Retrieval**
    - Performs initial broad retrieval using semantic similarity with vector embeddings.
    - Captures contextually relevant documents.
  - **Stage 2: Multi-Query Expansion**
    - Utilizes an LLM to reformulate the query into multiple versions.
    - Retrieves additional relevant documents for a diverse query representation.
  - **Stage 3: Hybrid Reranking**
    - Combines scores from dense (semantic) and sparse (BM25) methods.
    - Uses a weighted formula to rank documents, ensuring both contextual and term-based relevance.

***Balances dense and sparse retrieval methods for nuanced understanding.**

Methodology

1.Data collection

2.Document loading and text chunking

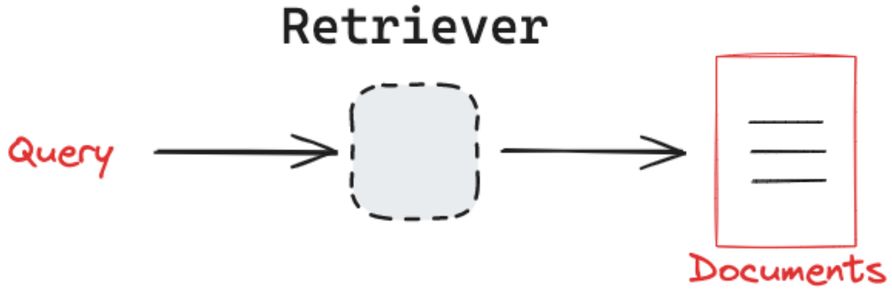
3. Embedding Creation and Vectore Store Creation

4. Retriever Selection

5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

7. Evaluation



What is BM25?

BM25 (Best Matching 25) is a probabilistic information retrieval model that ranks documents based on their relevance to a search query.

BM25 is based on the TF-IDF (Term Frequency - Inverse Document Frequency) model, but it introduces certain modifications to improve its performance

The BM25 score for a document with respect to a query is given by:

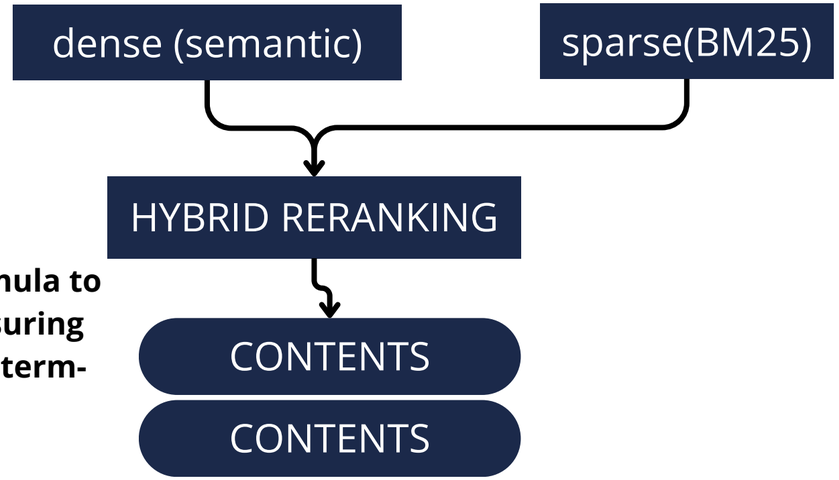
$$\text{BM25}(q,d) = \sum_{t \in q} \text{IDF}(t) \times \frac{\text{TF}(t,d) \times (k1+1)}{\text{TF}(t,d) + k1 \times (1 - b + b \times \frac{\text{len}(d)}{\text{avg\_len}})}$$

Where:

- `q`: The query (set of terms).
- `d`: The document.
- `t`: A term in the query.
- `IDF(t)`: Inverse Document Frequency for term `t`.
- `TF(t, d)`: Term frequency of term `t` in document `d`.
- `len(d)`: Length of the document `d`.
- `avg_len`: Average document length in the corpus.

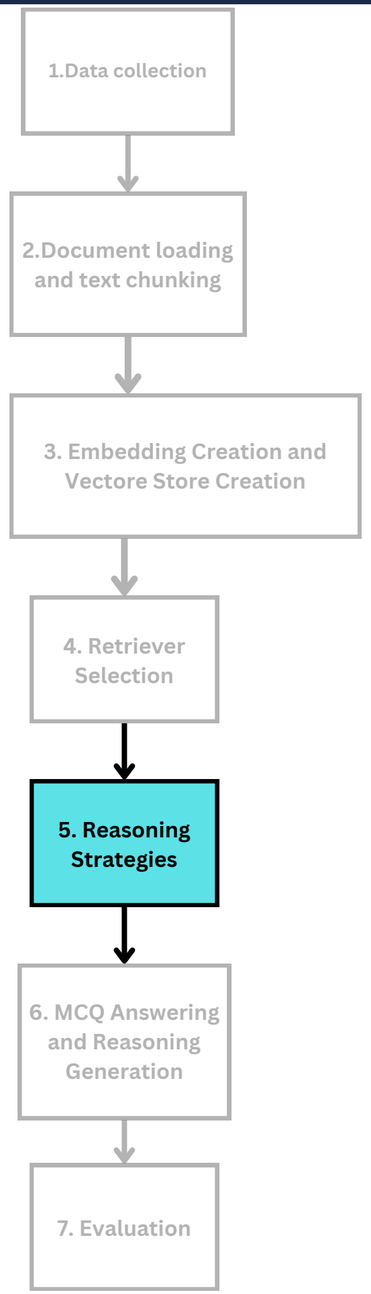
k1: Controls the saturation effect of term frequency. A higher value means more leniency for the term frequency, while a lower value means less.

b: Controls the length normalization, which compensates for document length variations (i.e., longer documents naturally have more occurrences of terms). A value of 1 means fully normalized, and 0 means no normalization.



Uses a weighted formula to rank documents, ensuring both contextual and term-based relevance.

Methodology



Prompt Template

template="""You are a medical expert analyzing a multiple choice question using chain-of-thought reasoning.
Context information:
{context}
Question: {question}
Options:
A) {optionA}
B) {optionB}
C) {optionC}
D) {optionD}

Let's solve this step by step:
1. First, let's understand what the question is asking for.
2. Then, analyze each option against the provided context.
3. Eliminate incorrect options with reasoning.
4. Confirm the correct answer with evidence from the context.

Thought process:
1. Question Analysis:
[Analyze the key aspects of the question]

2. Context Analysis:
[Identify relevant information from the context]

3. Option Analysis:
A) {optionA}: [Reasoning]
B) {optionB}: [Reasoning]
C) {optionC}: [Reasoning]
D) {optionD}: [Reasoning]

4. Final Selection:
[Explain why the chosen option is correct]

Please provide your answer in the following format exactly:
Selected Option: [A/B/C/D]
Reasoning: [Concise explanation based on the above analysis]
Confidence: [Numerical value between 0.0 and 1.0 based on how well the context supports this answer]"""

*This template is an example of **Chain-of-Thought (**CoT) reasoning

Methodology

1.Data collection

2.Document loading and text chunking

3. Embedding Creation and Vectore Store Creation

4. Retriever Selection

5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

7. Evaluation

from langchain.chains import **LLMChain**

Input:

- Prompt
  - question
  - options
  - context
- Context (from retriever)

Output:

- Predicted option
- Reasoning
- Confidence-Score

Methodology

1.Data collection

2.Document loading and text chunking

3. Embedding Creation and Vectore Store Creation

4. Retriever Selection

5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

7. Evaluation

NotebookLM

mcqs.json


```
{
      "question": "Which enzyme is potentially more inhibited by organophosphate (OP) compounds?",
      "options": {
         "A": "AChE",
         "B": "BuChE",
         "C": "Both equally",
         "D": "Neither"
      },
      "correct_answer": "B",
      "explanation": "Butyrylcholinesterase (BuChE) is potentially more inhibited by organophosphate compounds than acetylcholinesterase (AChE) [1].",
      "confidence": 1
   },


{
 "question": "According to the tapering guide, when should a patient not lower their dose of prednisone again?",
 "options": {
 "A": "After one good day.",
 "B": "After one bad day.",
 "C": "Until a minimum of two good days in a row",
 "D": "After two bad days"
 },
 "correct_answer": "C",
 "explanation": "A patient should not lower their dose of prednisone again until they have had a minimum of two good days in a row, according to the tapering guide [conversation history].",
 "confidence": 1
 }
```


Using the same set of guidelines, I generated 100 MCQ questions along with their reasoning and correct answers using Google **NotebookLM**. I **hypothesize** that the answers produced by NotebookLM are accurate and serve as the ground truth.

Methodology

1.Data collection

2.Document loading and text chunking

3. Embedding Creation and Vectore Store Creation

4. Retriever Selection

5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

7. Evaluation

```
sample_mcq = MCQInput(
    question="Which treatment option is commonly used for actinic keratosis according to the guidelines referenced in the document?",
    options={
        "A": "Dexamethasone",
        "B": "Photodynamic therapy (PDT)",
        "C": "Intravenous fluids",
        "D": "Antibiotics"
    }
)
```

Question: Which treatment option is commonly used for actinic keratosis according to the guidelines referenced in the document?

Options:
A) Dexamethasone
B) Photodynamic therapy (PDT)
C) Intravenous fluids
D) Antibiotics

Selected Answer: B

Reasoning: The guidelines referenced in the document specifically mention that conditional recommendations are made for the use of photodynamic therapy (PDT) for the treatment of actinic keratosis (AK). This indicates that PDT is recognized as a treatment option for AK, whereas the other options (dexamethasone, intravenous fluids, and antibiotics) are not mentioned in the context of AK management.

Confidence Score: 0.9

Evaluation

1.Data collection

2.Document loading and text chunking

3. Embedding Creation and Vectore Store Creation

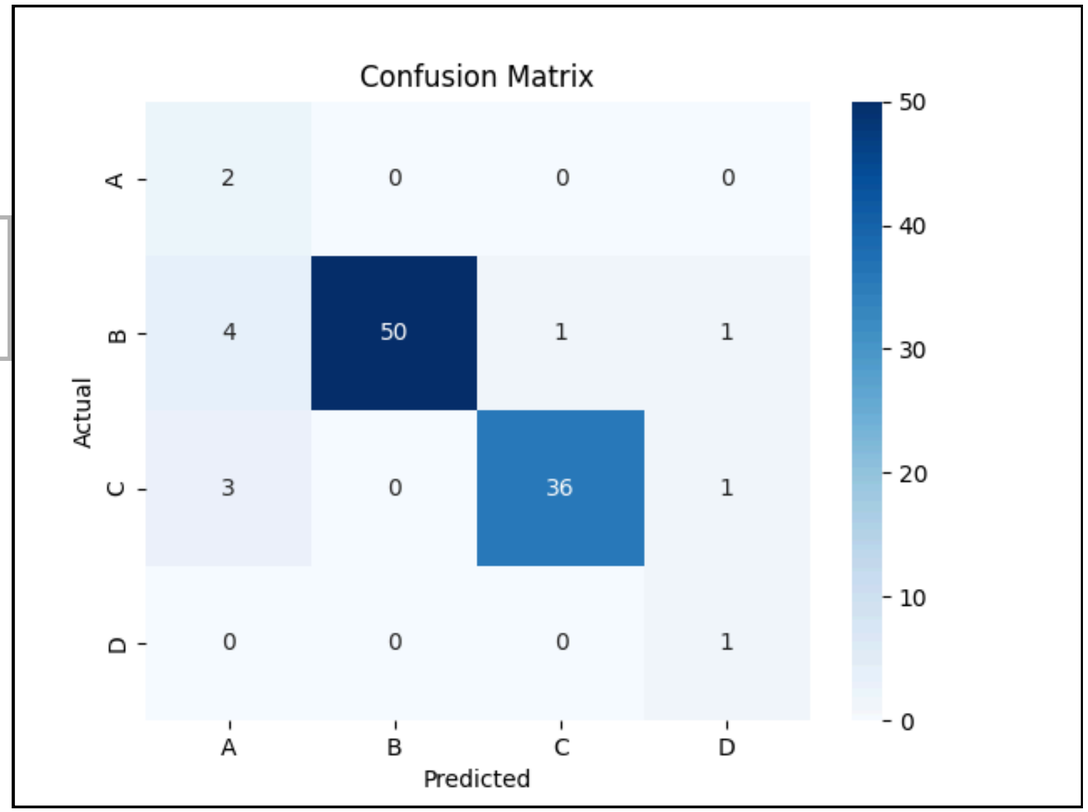
4. Retriever Selection

5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

7. Evaluation

Single-Stage Retriever



$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{All predictions}}$$

$$\text{Precision}_{\text{Class A}} = \frac{TP_{\text{Class A}}}{TP_{\text{Class A}} + FP_{\text{Class A}}}$$

$$\text{Recall}_{\text{Class A}} = \frac{TP_{\text{Class A}}}{TP_{\text{Class A}} + FN_{\text{Class A}}}$$

$$\text{Precision}_{\text{Macro-average}} = \frac{\text{Precision}_{\text{Class A}} + \text{Precision}_{\text{Class B}} + \ldots \text{Precision}_{\text{Class N}}}{N}$$

$$\text{Recall}_{\text{Macro-average}} = \frac{\text{Recall}_{\text{Class A}} + \text{Recall}_{\text{Class B}} + \ldots \text{Recall}_{\text{Class N}}}{N}$$

Accuracy: 0.89

- Micro Metrics:
  - Precision: 0.90
  - Recall: 0.89
  - F1-Score: 0.89
- Macro Metrics:
  - Precision: 0.63
  - Recall: 0.94
  - F1-Score: 0.68

Evaluation

1.Data collection

2.Document loading and text chunking

3. Embedding Creation and Vectore Store Creation

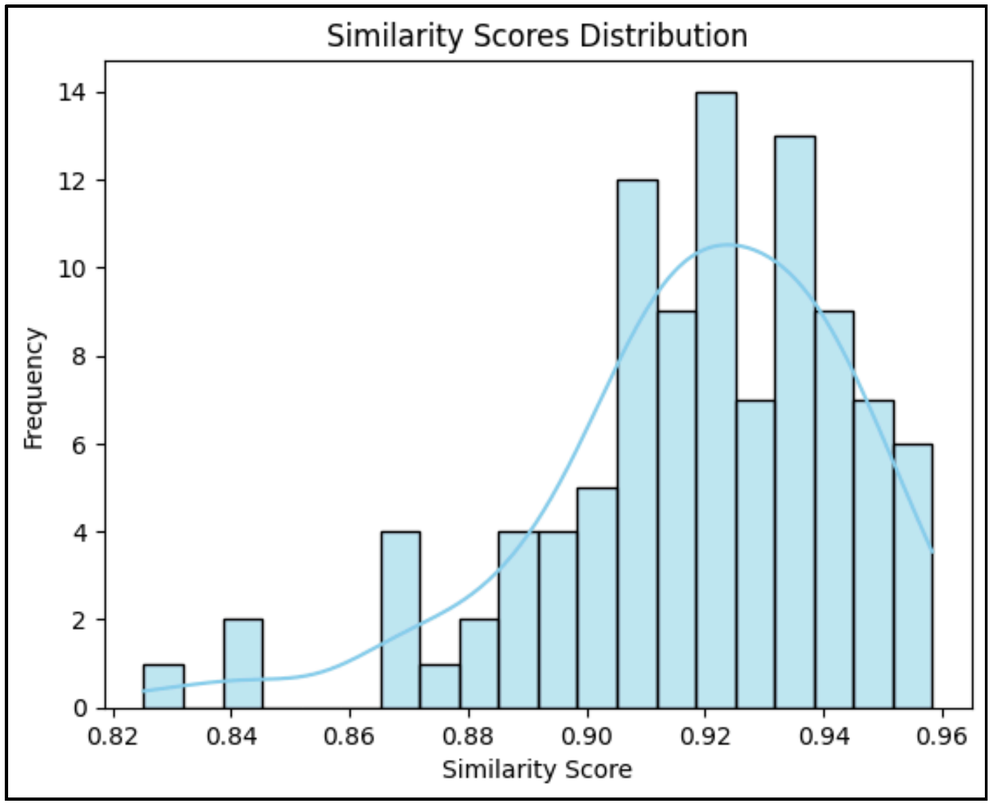
4. Retriever Selection

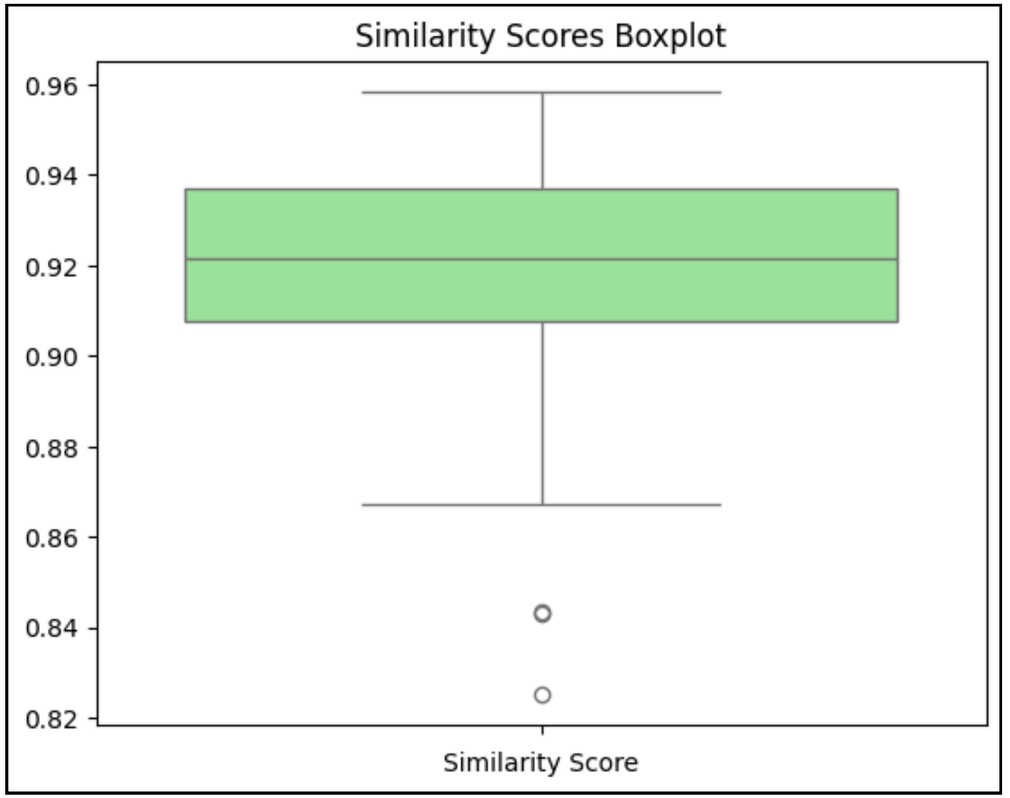
5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

7. Evaluation

Single-Stage Retriever





Evaluation

1.Data collection

2.Document loading and text chunking

3. Embedding Creation and Vectore Store Creation

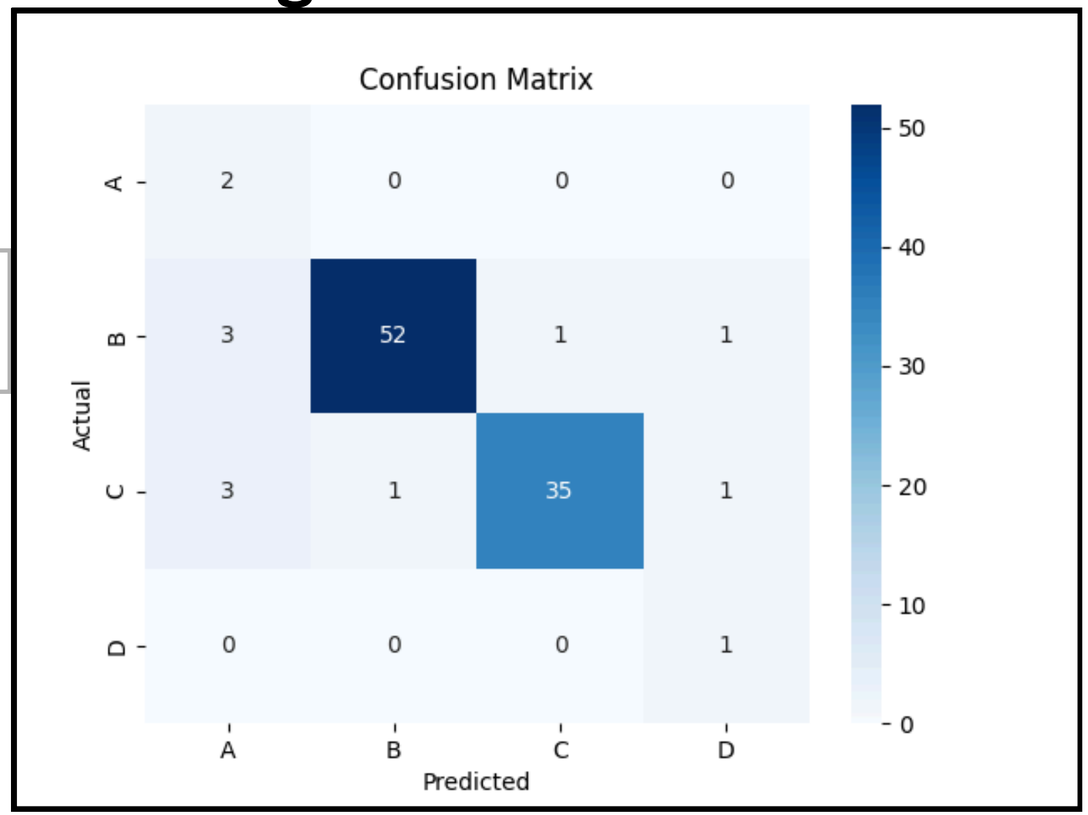
4. Retriever Selection

5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

7. Evaluation

Two-Stage Retriever



$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{All predictions}}$$

$$\text{Precision}_{\text{Class A}} = \frac{TP_{\text{Class A}}}{TP_{\text{Class A}} + FP_{\text{Class A}}}$$

$$\text{Recall}_{\text{Class A}} = \frac{TP_{\text{Class A}}}{TP_{\text{Class A}} + FN_{\text{Class A}}}$$

$$\text{Precision}_{\text{Macro-average}} = \frac{\text{Precision}_{\text{Class A}} + \text{Precision}_{\text{Class B}} + \ldots \text{Precision}_{\text{Class N}}}{N}$$

$$\text{Recall}_{\text{Macro-average}} = \frac{\text{Recall}_{\text{Class A}} + \text{Recall}_{\text{Class B}} + \ldots \text{Recall}_{\text{Class N}}}{N}$$

Accuracy: 0.90

- Micro Metrics:
  - Precision: 0.90
  - Recall: 0.90
  - F1-Score: 0.90
- Macro Metrics:
  - Precision: 0.63
  - Recall: 0.95
  - F1-Score: 0.69

Evaluation

1.Data collection

2.Document loading and text chunking

3. Embedding Creation and Vectore Store Creation

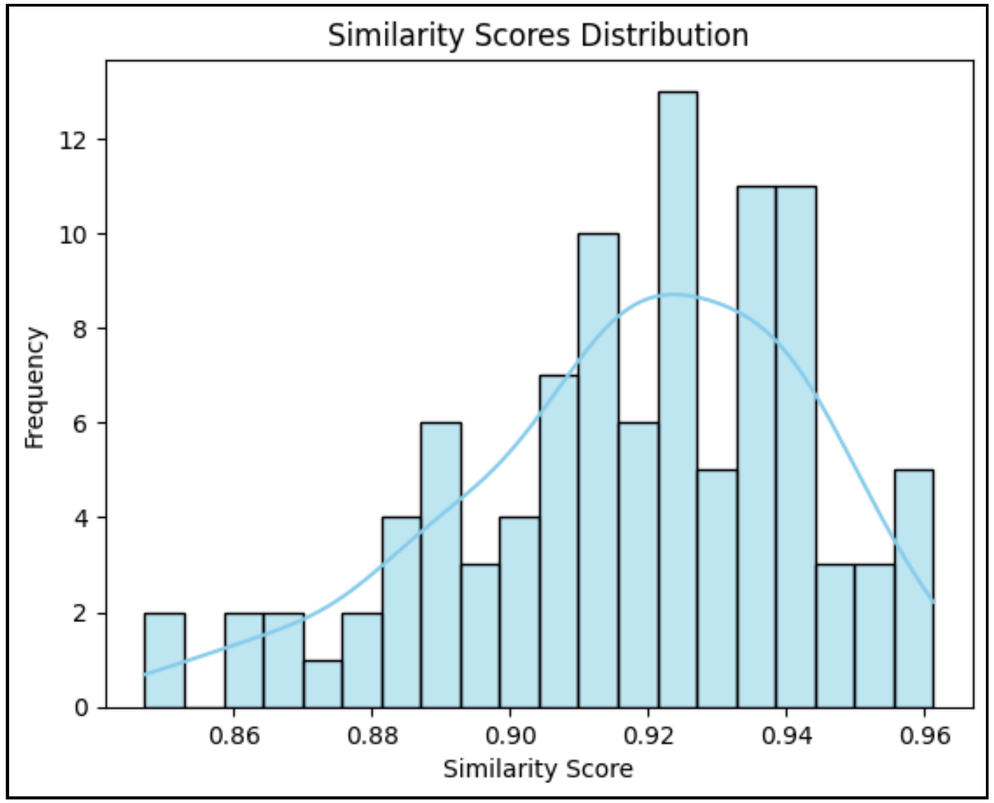
4. Retriever Selection

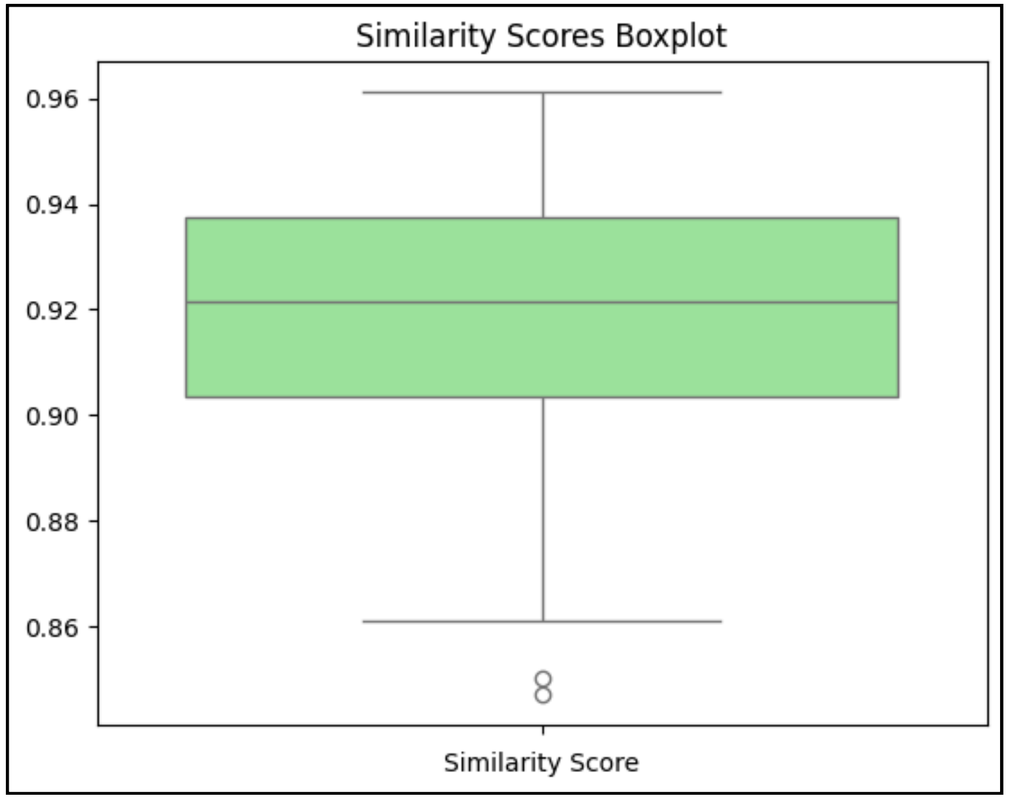
5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

7. Evaluation

Two-Stage Retriever





Evaluation

1.Data collection

2.Document loading and text chunking

3. Embedding Creation and Vectore Store Creation

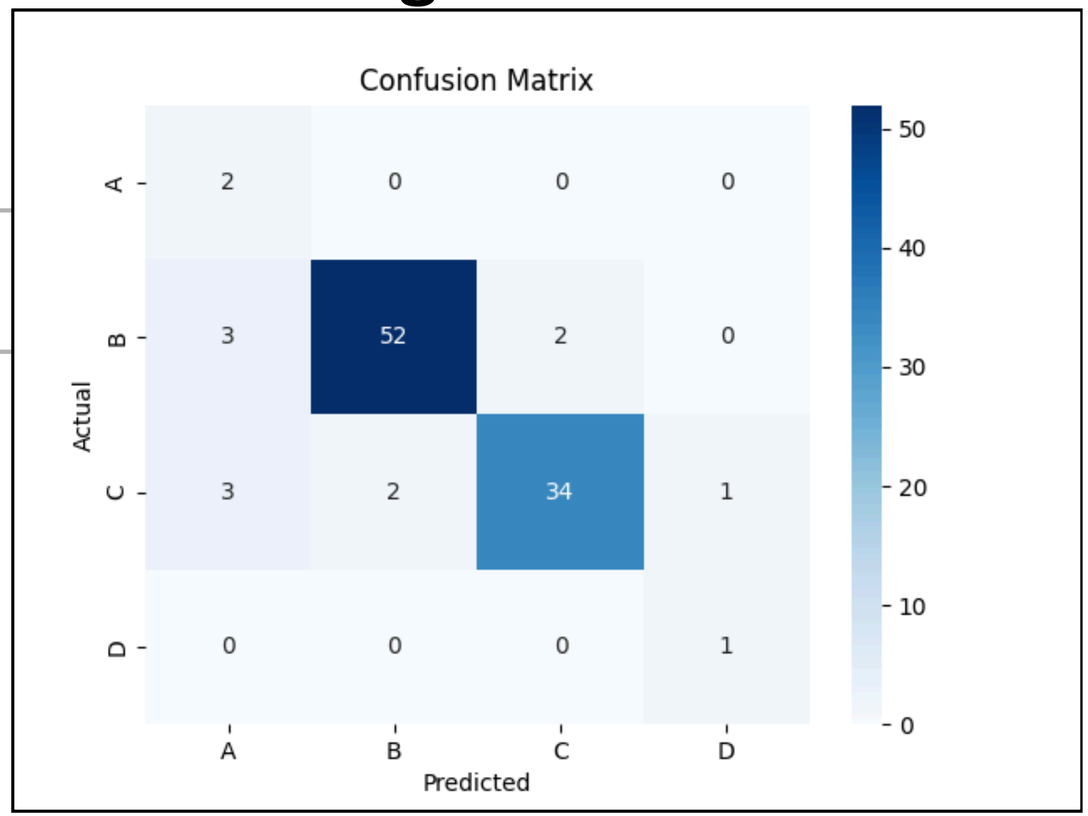
4. Retriever Selection

5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

7. Evaluation

Three-Stage Retriever



$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{All predictions}}$$

$$\text{Precision}_{\text{Class A}} = \frac{TP_{\text{Class A}}}{TP_{\text{Class A}} + FP_{\text{Class A}}}$$

$$\text{Recall}_{\text{Class A}} = \frac{TP_{\text{Class A}}}{TP_{\text{Class A}} + FN_{\text{Class A}}}$$

$$\text{Precision}_{\text{Macro-average}} = \frac{\text{Precision}_{\text{Class A}} + \text{Precision}_{\text{Class B}} + \dots \text{Precision}_{\text{Class N}}}{N}$$

$$\text{Recall}_{\text{Macro-average}} = \frac{\text{Recall}_{\text{Class A}} + \text{Recall}_{\text{Class B}} + \dots \text{Recall}_{\text{Class N}}}{N}$$

Accuracy: 0.89

- Micro Metrics:
  - Precision: 0.89
  - Recall: 0.89
  - F1-Score: 0.89
- Macro Metrics:
  - Precision: 0.66
  - Recall: 0.94
  - F1-Score: 0.72

Evaluation

1.Data collection

2.Document loading and text chunking

3. Embedding Creation and Vectore Store Creation

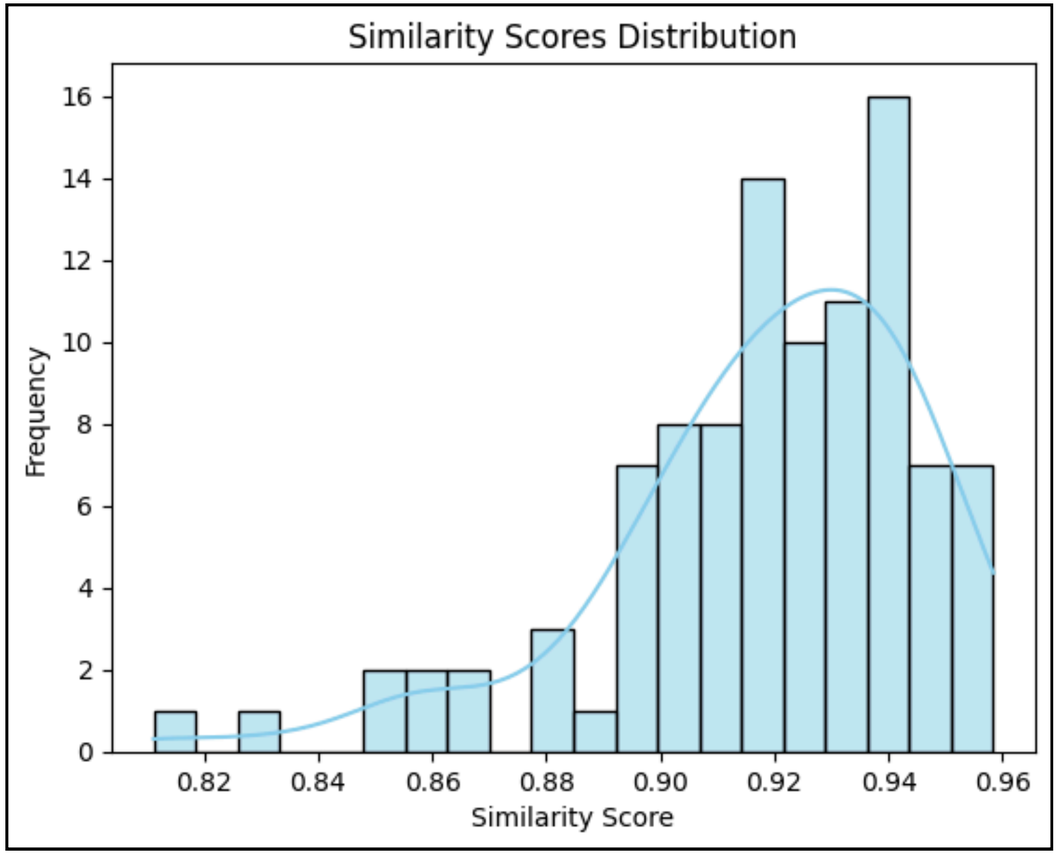
4. Retriever Selection

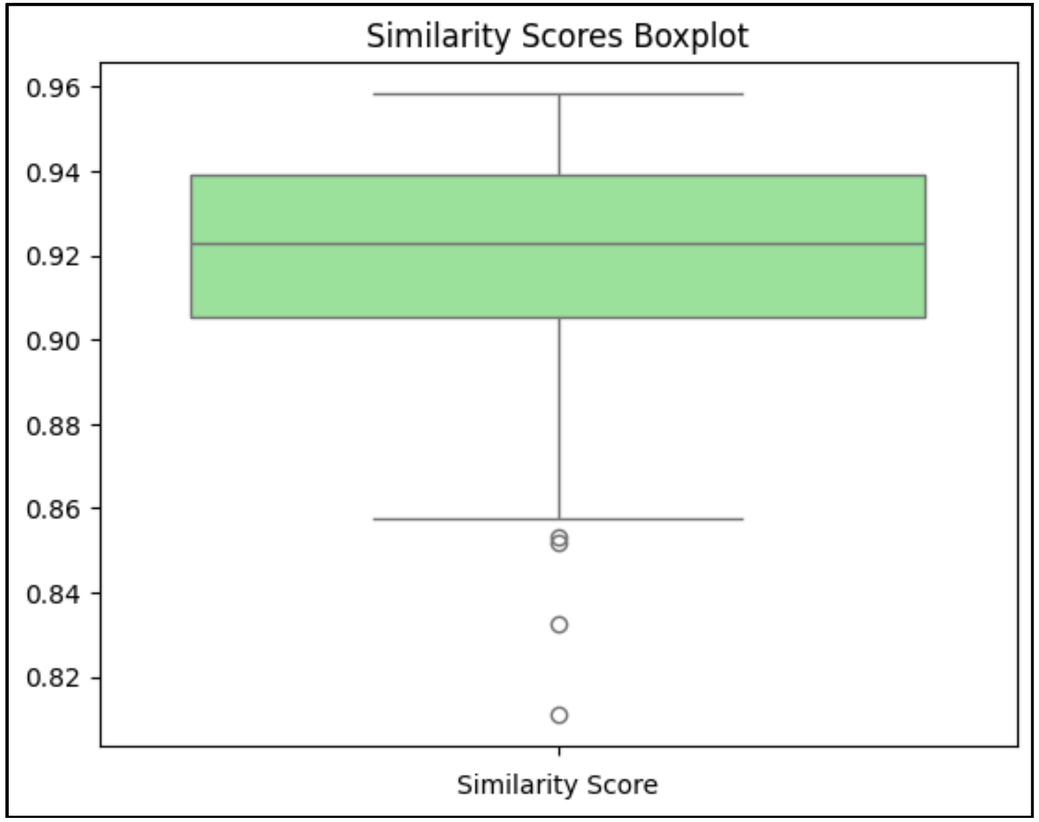
5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

7. Evaluation

Three-Stage Retriever





Comparison between single, two, and three-stage retriever

1.Data collection

2.Document loading and text chunking

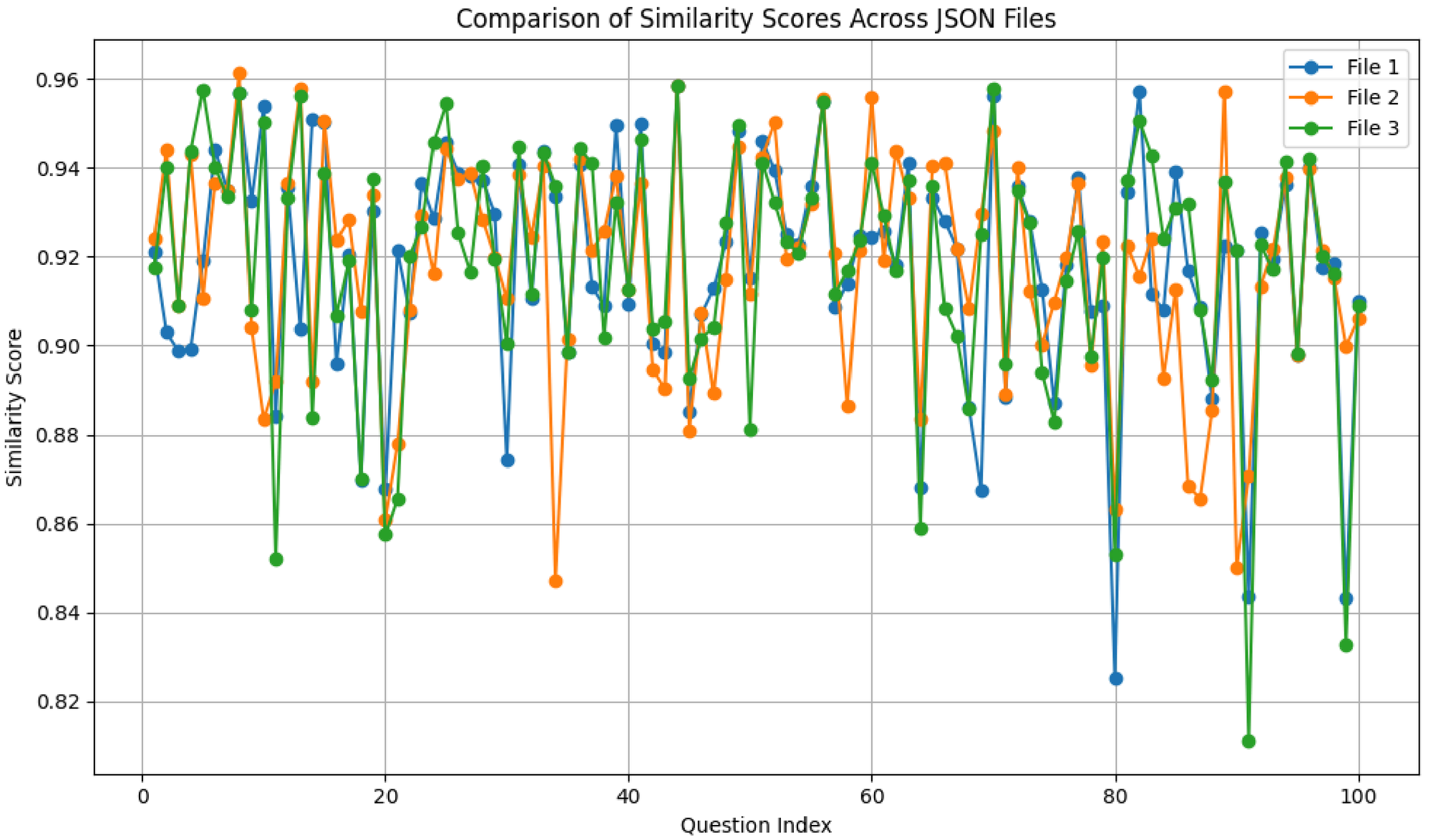
3. Embedding Creation and Vectore Store Creation

4. Retriever Selection

5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

7. Evaluation



Evaluation

1.Data collection

2.Document loading and text chunking

3. Embedding Creation and Vectore Store Creation

4. Retriever Selection

5. Reasoning Strategies

6. MCQ Answering and Reasoning Generation

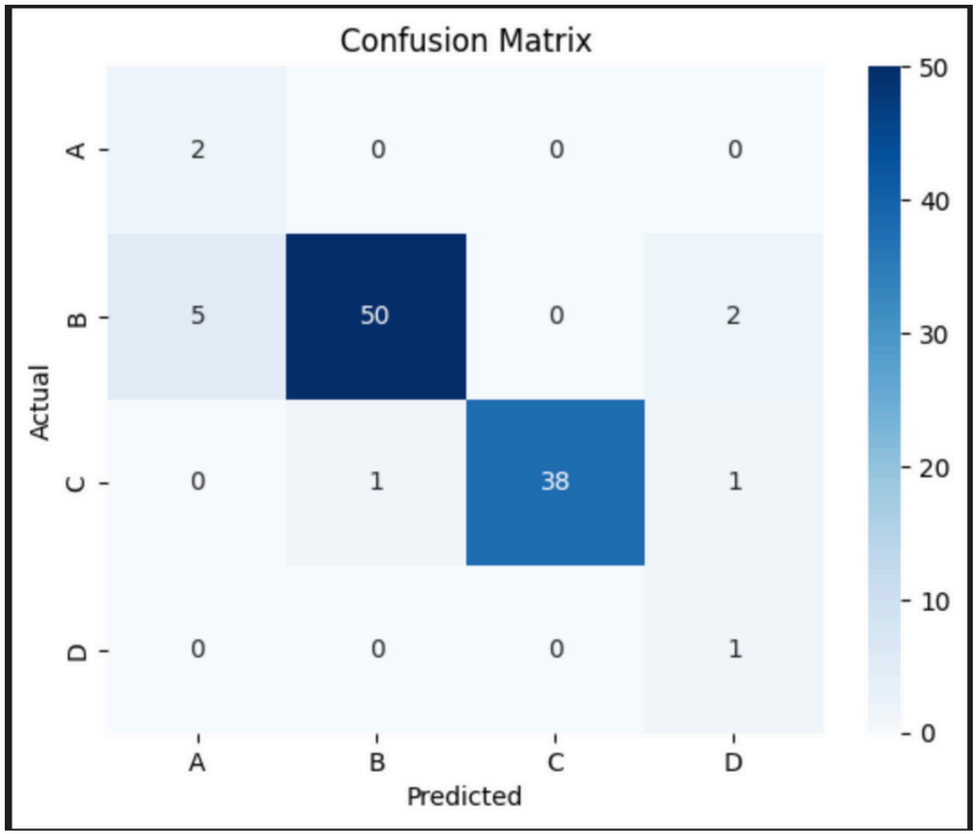
7. Evaluation

aleynahukmet/bge-medical-small-en-v1.5

fine-tuned on medical question-answer pairs

This model incorporates multiple datasets sourced from the Hugging Face platform, each contributing to the training and evaluation of the semantic search system:

- keivalya/MedQuad-MedicalQnADataset
- medalpaca/medical_meadow_wikidoc
- medalpaca/medical_meadow_medical_flashcards
- medalpaca/medical_meadow_wikidoc_patient_information



Accuracy: 0.91

- Macro Metrics:
  - Precision: 0.63
  - Recall: 0.96
  - F1-Score: 0.69
- Micro Metrics:
  - Precision: 0.91
  - Recall: 0.91
  - F1-Score: 0.91

Evaluation

1.Data collection

2.Document loading and text chunking

3. Embedding Creation and Vectore Store Creation

4. Retriever Selection

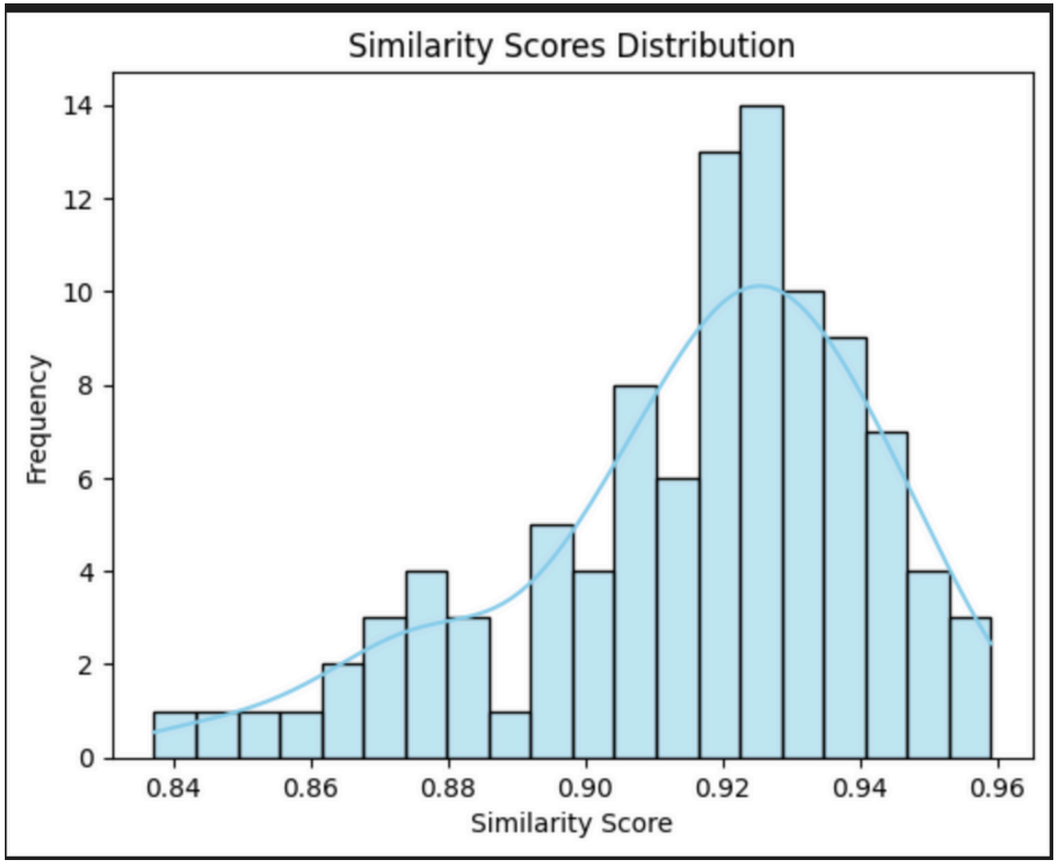
5. Reasoning Strategies

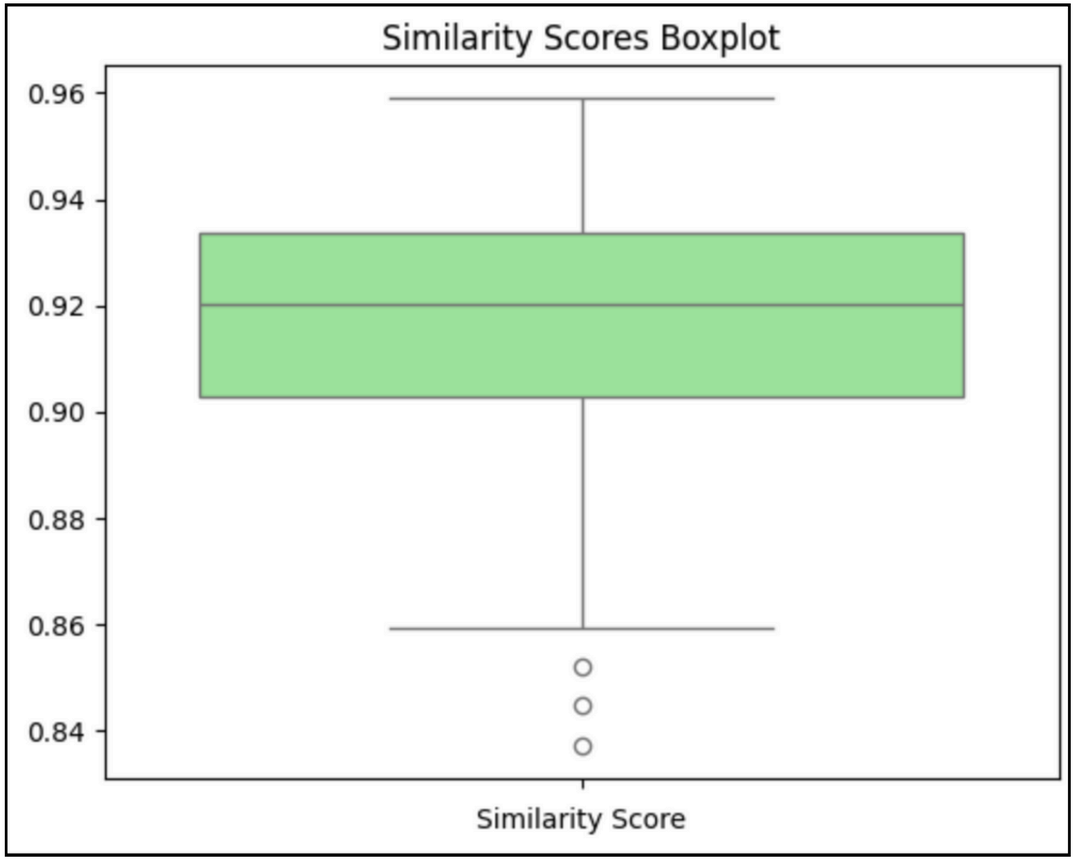
6. MCQ Answering and Reasoning Generation

7. Evaluation

aleynahukmet/bge-medical-small-en-v1.5

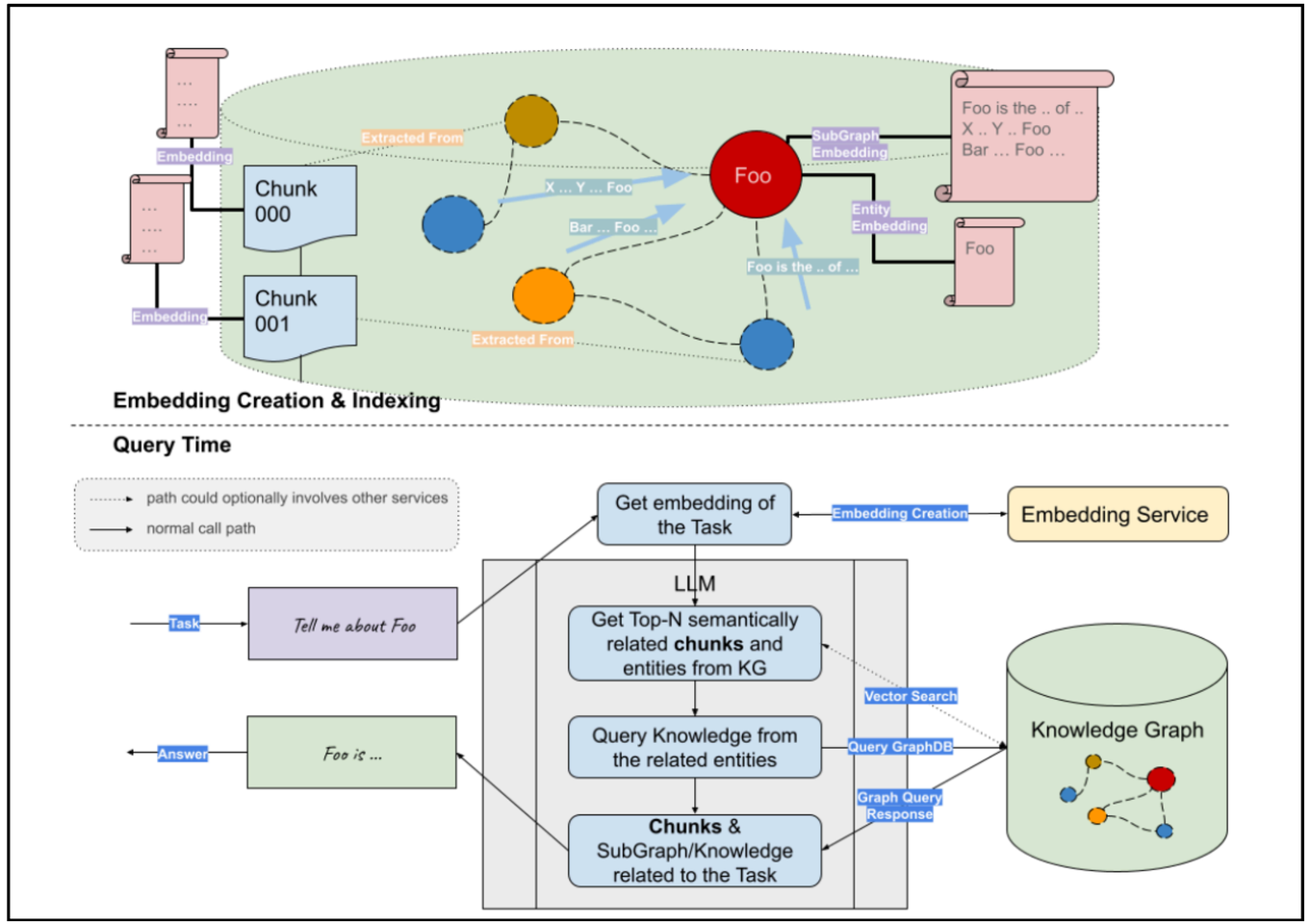
fine-tuned on medical question-answer pairs





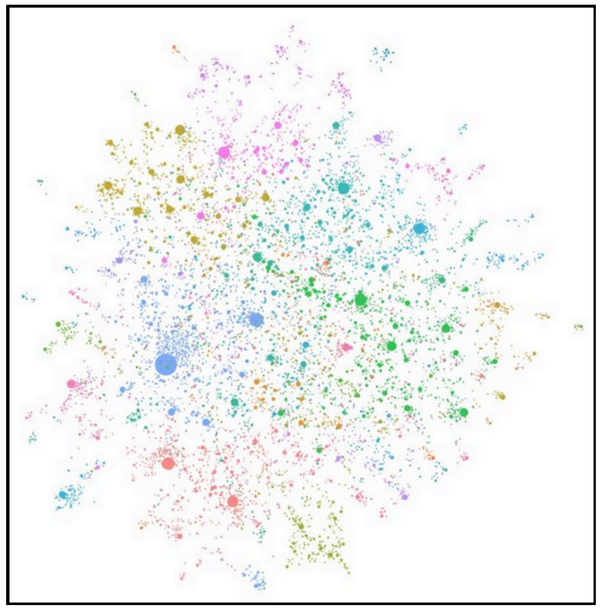
Research and Future Enhancement

Graph RAG: Retrieval-Augmented Generation with LLM Based on Knowledge Graphs



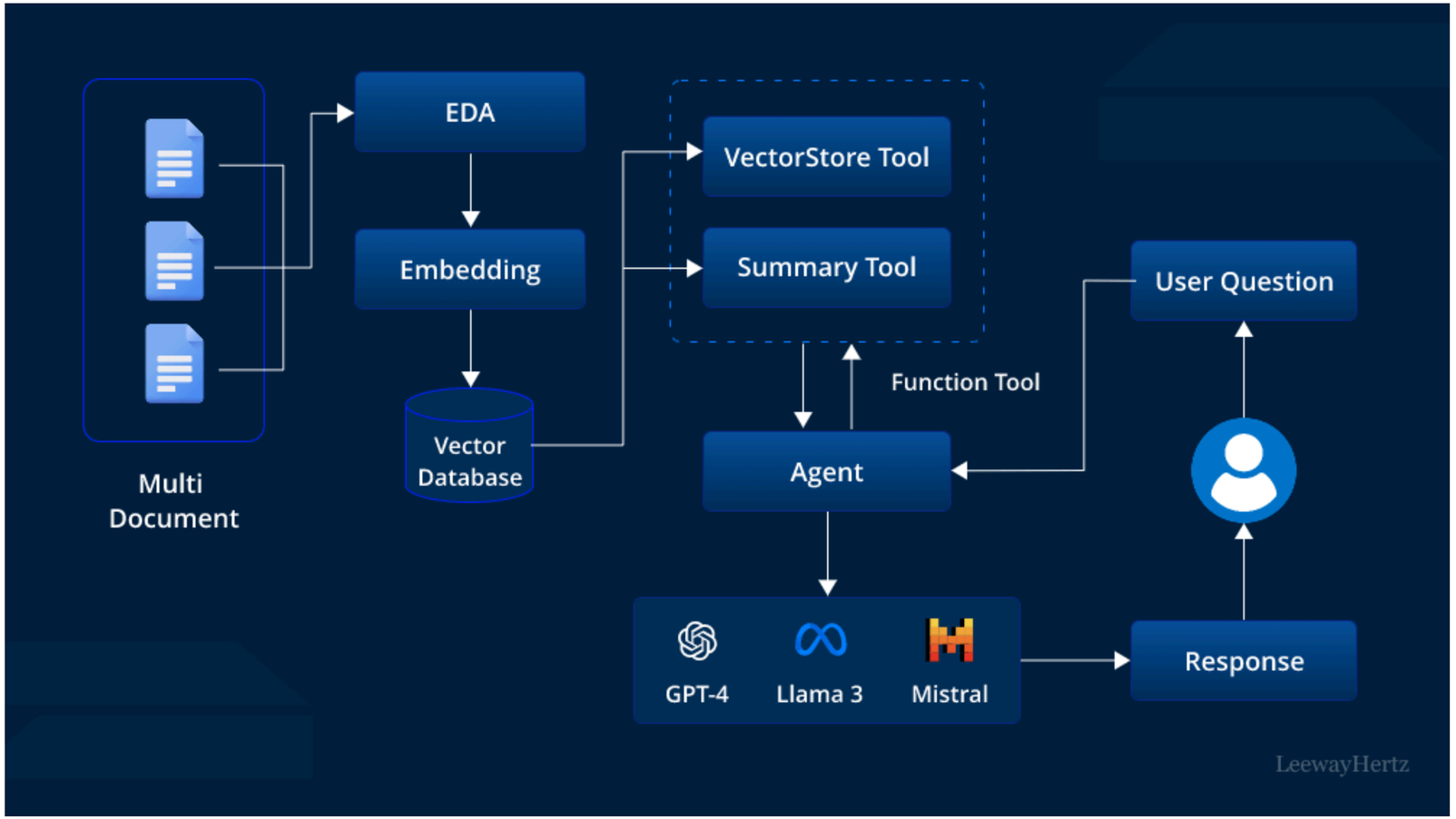
- GraphRAG identified entities and relationships from documents and developed a graph using LLM.
- Then, they used **clustering algorithm** to offer global information based on user query, offering better performance than naïve RAG on the traditional vector databases.

https://arxiv.org/pdf/2404.16130



Research and Future Enhancement

Multi Agent RAG System



- agents can assist in **planning** and **decision making** for complex tasks, rather than **simple retrieval.**

- A study developed an agent to answer questions related to rare diseases by expanding beyond RAG with additional tool functions, such as querying phenotypes and performing web searches. This approach improved the overall correctness from **0.48 to 0.75** compared to the GPT-4 baseline LLM

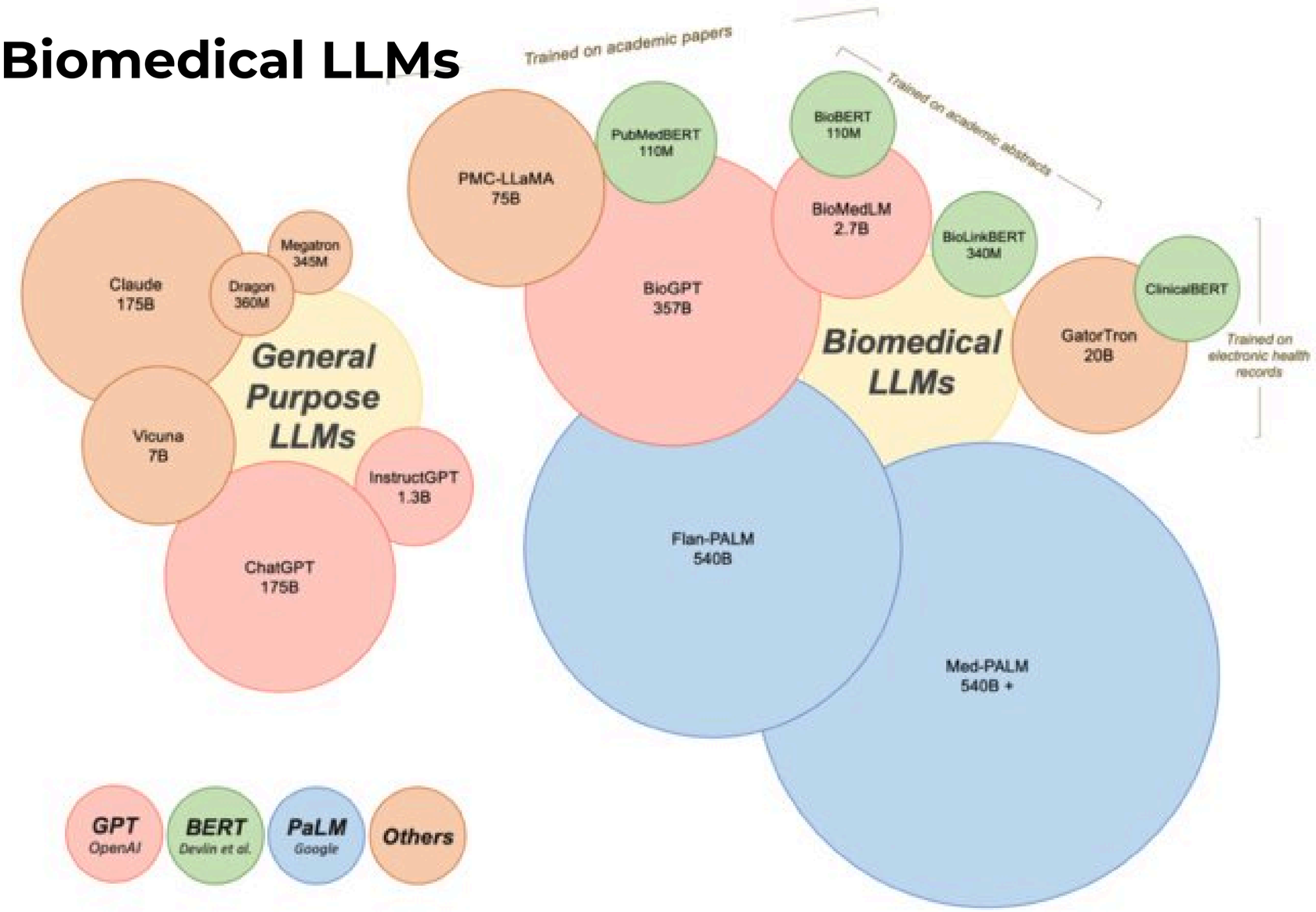
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10684379

source:leewayhertz.com

Yang J , Shu L , Duan H , et al. RDguru: a conversational intelligent agent for rare diseases . IEEE J Biomed Health Inform . Published online September 19, 2024.

Research and Future Enhancement

Biomedical LLMs



Current LLMs in medicine. Currently there are general purpose and biomedical LLMs used for medical tasks. While GPT by OpenAI, BERT by Devlin and colleagues, and PaLM by Google have led the development of LLMs with applications in medicine, other proprietary and open source LLMs also exist in this space. Circle sizes reflect the model size and the number of parameters used to build the models. LLMs with applications in medicine vary widely in how they were trained. BioMedLM 2.7B by GPT was trained on the corpus of PubMed articles and abstracts, for example, whereas ClinicalBERT was trained specifically on electronic health records. These differences in training and development can have important implications for how LLMs perform in certain medical scenarios.

https://www.researchgate.net/publication/373642018_Large_language_models_in_medicine_the_potentials_and_pitfalls/figures?lo=1

Problems Faced

- Lack of Domain Knowledge
- Time constraint for more Research and Implementation

THANK YOU!

For your attention