Optimizing Information Retrieval: A Hybrid Model Leveraging MAR and RAPTOR Frameworks

George Washington University

Prepared by: Timur Abdygulov
Supervised by Amir H. Jafari
December 10, 2024

Table of Contents

Introduction

Retrieval-Augmented Generation (RAG):

- Combines retrieval mechanisms with large language models (LLMs).
- Enhances contextual understanding and reasoning capabilities.

Key Challenges in Retrieval Systems:

- Limited adaptability for dynamic queries.
- Poor handling of multi-hop reasoning and long-context synthesis.

Existing Models:

- **MAR:** Excels at static memory recall but struggles with evolving tasks.
- **RAPTOR:** Effective for dynamic and hierarchical reasoning but lacks memory persistence.

Contribution:

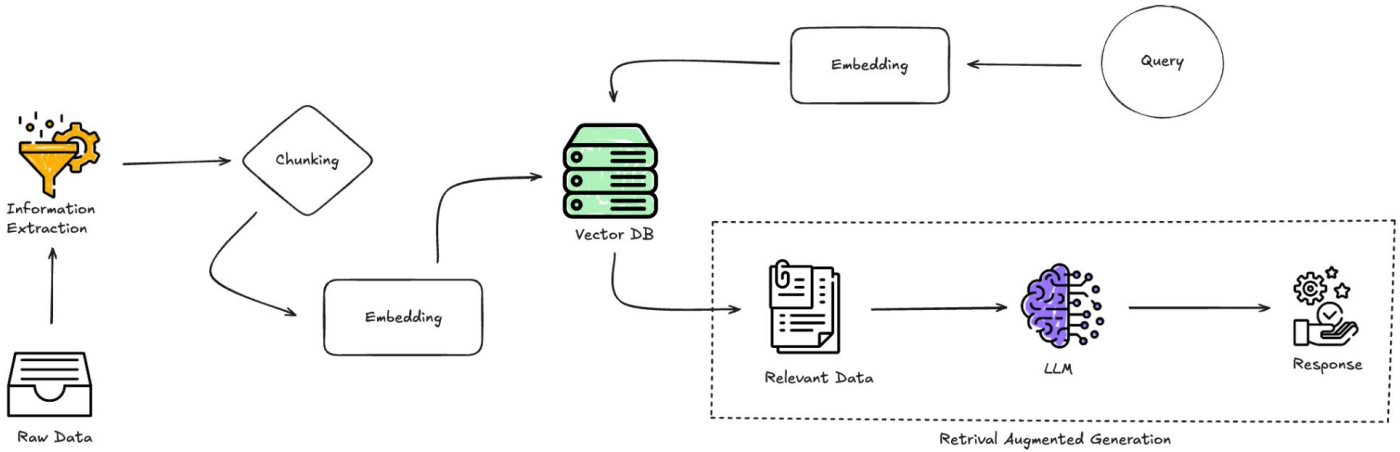
- Hybrid Model integrates MAR's memory with RAPTOR's hierarchical reasoning.
- Balances static recall and dynamic adaptability for diverse query challenges.

Methodology Overview

- Evaluation of Naive RAG, MAR, RAPTOR, and Hybrid Model.
- Integration with vector stores and generative models for response synthesis.
- Focus on modularity and extensibility.

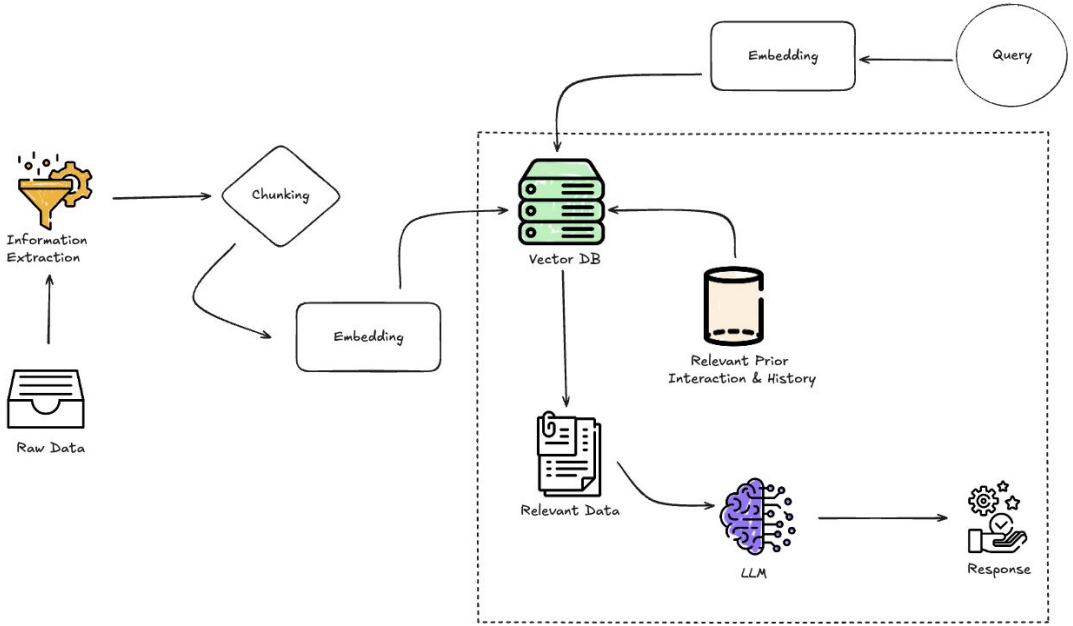
Hyperparameter	Value/Default	Description
chunk_size	1000	The maximum size of text chunks for splitting documents.
chunk_overlap	115	The number of overlapping characters between consecutive text chunks.
embedding_dim	1024	Dimension of the embeddings generated by the model.
batch_size	16 (embedding), 50 (indexing)	Batch size for document processing during embedding generation or indexing.
max_documents	Derived dynamically	The maximum number of documents stored, based on space and model embedding size.
avg_document_size	1000	Average size of documents in bytes, used for storage estimation.

Naive RAG



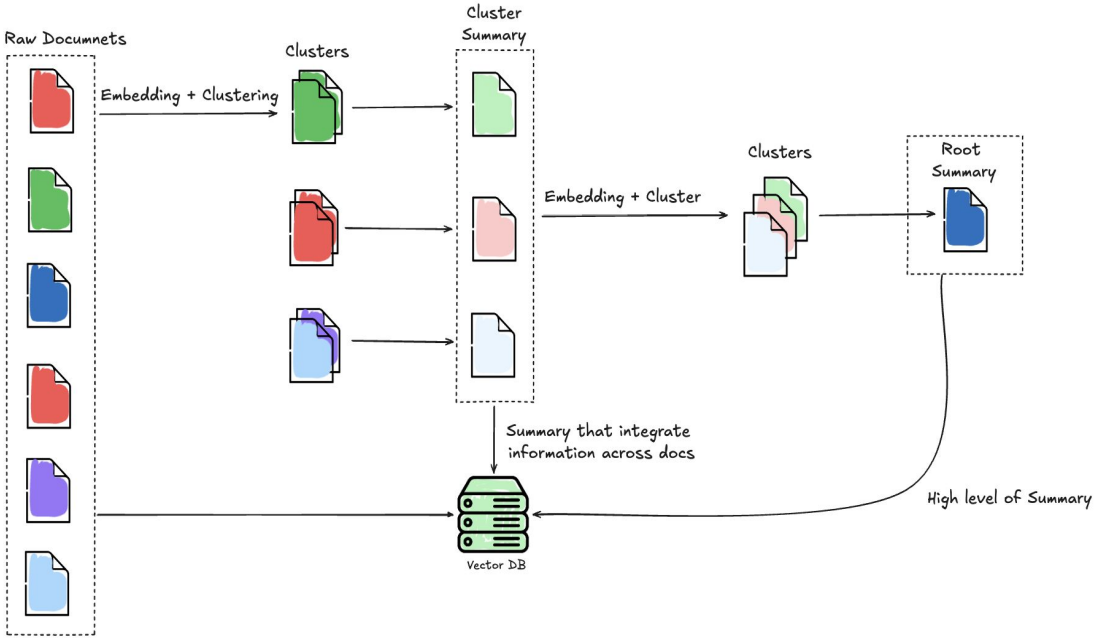
- Baseline RAG framework.
- Straightforward retrieval-to-generation pipeline.
- Effective for static, simple queries but limited in dynamic adaptability.

Memory-Augmented Retrieval (MAR)



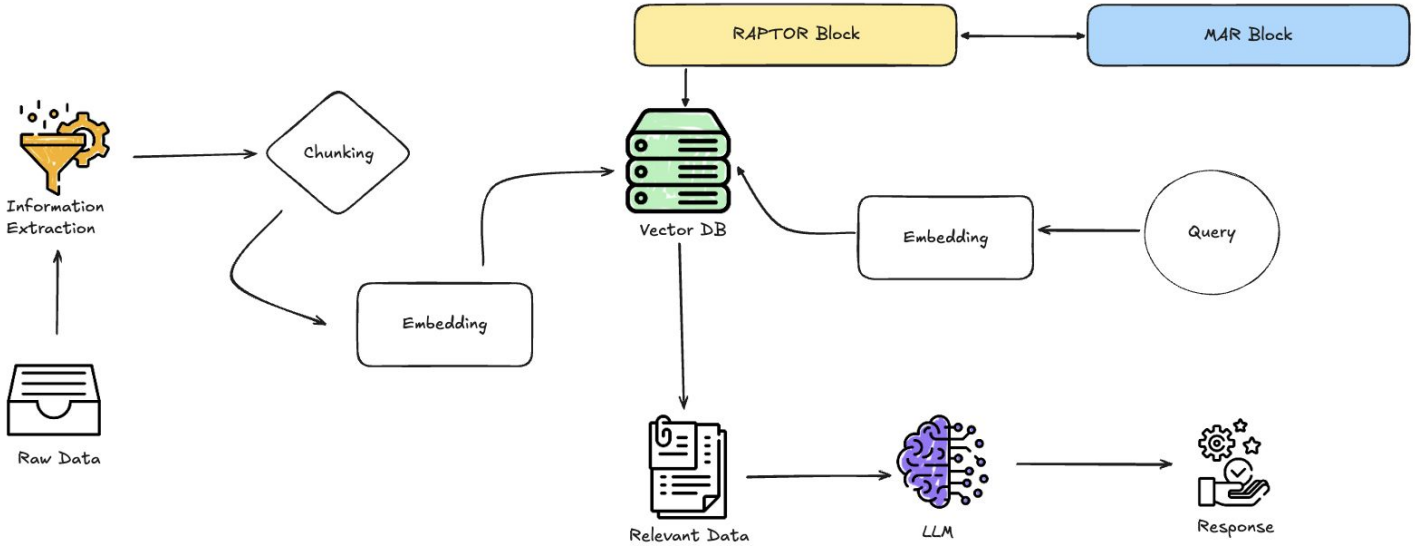
- Adds a memory database for recurring queries.
- Efficient for static tasks, limited for evolving contexts.

Recursive Abstractive Processing for Tree-Organized Retrieval (RAPTOR)



- Hierarchical document organization for multi-hop reasoning.
- Semantic clustering enhances contextual synthesis.
- Limited by lack of memory.

Hybrid Model (MAR + RAPTOR)

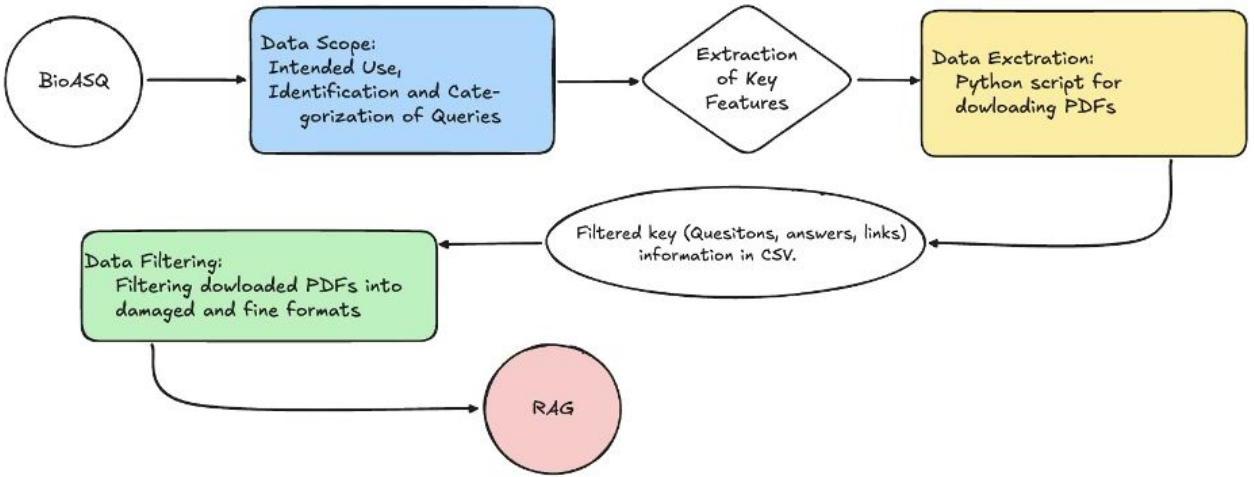


- Combines MAR's memory with RAPTOR's reasoning.
- Handles static, dynamic, and multi-layered queries.

Retrieval System: ChromaDB

- **Role in Retrieval Systems:**
  - Facilitates efficient semantic search through embedding-based matching.
  - Acts as the backbone for storing and retrieving high-dimensional vector representations of documents.
- **Chosen System: ChromaDB**
  - Scalable and low-latency query processing.
  - Supports dense, hybrid (dense + keyword), and multimodal retrieval.
- **Key Features of ChromaDB:**
  - Seamless integration with frameworks like LangChain and modern embedding models.
  - Handles high-dimensional embeddings (e.g., 1024 dimensions).
  - Optimized for hybrid retrieval methods (semantic and keyword-based).
- **Advantages:**
  - Scalability for large datasets.
  - High-speed processing with minimal latency.
  - Compatibility with adaptive and hierarchical retrieval methods.

Dataset Description



- BioASQ biomedical dataset.
- Query types: Static, Dynamic, and Multi-layered.
- Thorough data preprocessing for consistency.

Categorized Biomedical Queries

ID	Question	Category
1	Is Hirschsprung disease a Mendelian or a multifactorial disorder?	M
2	List signaling molecules (ligands) that interact with the receptor EGFR?	S
3	Are long non-coding RNAs spliced?	M
4	Is RANKL secreted from the cells?	S
5	Which miRNAs could be used as potential biomarkers for epithelial ovarian cancer?	M
6	Which acetylcholinesterase inhibitors are used for treatment of myasthenia gravis?	S
7	Has Denosumab (Prolia) been approved by FDA?	D
8	Which are the different isoforms of the mammalian Notch receptor?	S
9	Orteronel was developed for treatment of which cancer?	D
10	Is the monoclonal antibody Trastuzumab (Herceptin) of potential use in the treatment of prostate cancer?	D
11	Which are the Yamanaka factors?	S
12	Where is the protein Pannexin1 located?	S
13	Which currently known mitochondrial diseases have been attributed to POLG mutations?	M

- **Static Queries** involve retrieving straightforward, factual information that relies on pre-existing, often unchanging knowledge.
- **Dynamic Queries** require contextual reasoning and the ability to handle evolving information, such as updates to regulatory statuses or therapeutic applications.
- **Multi-Layered Queries** demand evidence aggregation and multi-hop reasoning, where answers must be synthesized across multiple sources or involve complex reasoning steps.
- A total of **47 PDFs** were used

Frequency of Biomedical Queries

ID	Question	Frequency
1	Is Hirschsprung disease a Mendelian or a multifactorial disorder?	2
2	List signaling molecules (ligands) that interact with the receptor EGFR?	4
3	Are long non-coding RNAs spliced?	4
4	Is RANKL secreted from the cells?	3
5	Which miRNAs could be used as potential biomarkers for epithelial ovarian cancer?	6
6	Which acetylcholinesterase inhibitors are used for treatment of myasthenia gravis?	2
7	Has Denosumab (Prolia) been approved by FDA?	2
8	Which are the different isoforms of the mammalian Notch receptor?	2
9	Orteronel was developed for treatment of which cancer?	3
10	Is the monoclonal antibody Trastuzumab (Herceptin) of potential use in the treatment of prostate cancer?	7
11	Which are the Yamanaka factors?	7
12	Where is the protein Pannexin1 located?	4
13	Which currently known mitochondrial diseases have been attributed to POLG mutations?	2

Embedding Model Evaluation

- High-performance embeddings offer accuracy at higher computational costs.
- Lightweight embeddings balance efficiency and speed.

Embedding Model	Top 1 Accuracy (%)	Top 10 Accuracy (%)
bert-large-nli	75.0	87.5
instructor-xl	87.5	87.5
msmacro	87.5	87.5
mxbai-embed-large	87.5	87.5
roberta-base	37.5	87.5
roberta-large	62.5	62.5
dunzhang/stella_en_1.5B_v5	1	80.3
nvidia/NV-Embed-v2	1	80.5

Embedding Resource Utilization

- High-performance embeddings require substantial GPU memory.
- Lightweight models suitable for simpler tasks.

Model Name	GPU Memory Used (GB)	Total Parameters (Millions)	Loaded in 8-bit
mixedbread-ai/mxbai-embed-large-v1	1.25	335.14	No
dunzhang/stella_en_1.5B_v5	3.48	1543.27	Yes
dunzhang/stella_en_1.5B_v5	9.25	1543.27	No
sentence-transformers/all-MiniLM-L6-v2	0.08	22.71	No
meta-llama/Meta-Llama-3-8B-Instruct	10.42	8030.26	Yes

Results Summary

- Hybrid Model achieves balanced performance across query types.
- MAR excels in static recall; RAPTOR in multi-hop reasoning.
- Embedding strategies influence performance and scalability.

- MAR:
  - Answered: 11 out of 13 questions
  - Very Close to Ideal: 7
  - Correct Answer: 6
- RAG:
  - Answered: 11 out of 13 questions
  - Very Close to Ideal: 7
  - Correct Answer: 5
- RAPTOR:
  - Answered: 8 out of 13 questions
  - Very Close to Ideal: 3
  - Correct Answer: 3
- RAPTOR+Memory:
  - Answered: 10 out of 13 questions
  - Very Close to Ideal: 6
  - Correct Answer: 5

Question	Ideal Answer	MAR Answer	RAG Answer	RAPTOR Answer	Hybrid Answer
Is Hirschsprung disease a Mendelian or a multifactorial disorder?	Hirschsprung disease is both Mendelian and multifactorial, depending on the context.	✅ Matches ideal answer	✅ Matches ideal answer	❌ Incorrect	✅ Matches ideal answer
List signaling molecules (ligands) that interact with the receptor EGFR?	The 7 EGFR ligands are EGF, BTC, EPR, HB-EGF, TGF-α, AREG, and EPG.	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer
Are long non-coding RNAs spliced?	Yes, long non-coding RNAs are spliced through the same pathway as mRNAs.	✅ Matches ideal answer	✅ Matches ideal answer	❌ Unclear or incomplete answer	❌ Unclear or incomplete answer
Is RANKL secreted from the cells?	Yes, RANKL is secreted by osteoblasts.	✅ Matches ideal answer	❌ Does not mention secretion	❌ Incomplete answer	✅ Matches ideal answer
Which miRNAs could be used as potential biomarkers for epithelial ovarian cancer?	miR-200a, miR-100, miR-141, miR-200b, miR-200c, miR-203, etc.	✅ Matches ideal answer	❌ Partial match	❌ Partial match	✅ Matches ideal answer
Which acetylcholinesterase inhibitors are used for treatment of myasthenia gravis?	Pyridostigmine and neostigmine.	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer
Has Denosumab (Prolia) been approved by FDA?	Yes, approved by the FDA in 2010.	❌ Not answered	❌ Not answered	✅ Matches ideal answer	❌ Not answered
Which are the different isoforms of the mammalian Notch receptor?	Notch-1, Notch-2, Notch-3, Notch-4.	❌ Not answered	❌ Not answered	❌ Not answered	❌ Not answered
Orteronel was developed for treatment of which cancer?	Castration-resistant prostate cancer.	✅ Matches ideal answer	❌ Not answered	❌ Incorrect	✅ Matches ideal answer
Is the monoclonal antibody Trastuzumab (Herceptin) of potential use in the treatment of prostate cancer?	Controversial, but it can be used in HER2 overexpressing prostate cancer.	✅ Matches ideal answer	✅ Matches ideal answer	❌ Incorrect	❌ Not answered
Which are the Yamanaka factors?	OCT4, SOX2, MYC, and KLF4 transcription factors.	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer
Where is the protein Pannexin1 located?	Localized to the plasma membranes.	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer
Which currently known mitochondrial diseases have been attributed to POLG mutations?	Recessive PEO and MNGIE.	❌ Partial match	❌ Partial match	❌ Partial match	❌ Partial match

Discussion

- Hybrid Model mitigates trade-offs between memory and reasoning.
- Embedding selection critical to balancing accuracy and scalability.
- Limitations include semantic clustering errors and resource demands.

Conclusion

- Hybrid Model: A robust solution for diverse retrieval tasks.
- Combines strengths of MAR and RAPTOR.
- Future directions: Adaptive clustering, multimodal retrieval, and dynamic embeddings.

Acknowledgments

Meet Daxini in help me with evlaution of embeddign models for stella and nvidai as well as calculating Resource Utilization of models

Thank you