Multi-Agent Open-Domain Question Answering with Cross-Source Reranking Project Outline

IRE (Information Retrieval and Extraction)
IIIT Hyderabad

Team Name - Chicken sure-ma

Yash Bhaskar
Aditya Raghuvanshi
Pranit Khanna

Problem Statement:

We aim to develop a multi-agent open-domain question-answering (ODQA) system that retrieves and synthesizes information from diverse sources, including web searches, large language models like Llama 3, and vision models for multi-modal retrieval. Using datasets like KILT, Natural Questions, HotspotQA, TriviaQA, and ELI5, the system will employ a cross-source reranking model to select the most accurate answers. The project will focus on both context-free and context-based scenarios, ensuring scalability and reliability across various retrieval methods.

Scope of the Project:

The project aims to develop a sophisticated ODQA pipeline that leverages multiple specialized agents for retrieving and synthesizing information from diverse sources. The agents will operate on various datasets and information sources, including web searches (e.g., Google), specific websites, large language models (LLMs) like Llama 3, and vision models for multi-modal retrieval. The project will also integrate multiple retrieval techniques, such as tf-idf, BM25, and embedding-based methods, to enhance the retrieval process. The retrieved information will be processed through a cross-source reranking system, employing techniques like Reciprocal Rank Fusion to ensure the most accurate and relevant answers are prioritized.

The project will specifically focus on integrating and experimenting with datasets from the KILT Benchmark (Meta), and open-domain QA datasets like Natural Questions, HotspotQA, TriviaQA, and ELI5. The pipeline will explore both context-free and context-based scenarios, with an emphasis on scaling retrieval methods with an increasing ratio of irrelevant to relevant documents.

Literature Review:

1. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks : https://arxiv.org/pdf/2005.11401 : BART and RAG
2. KILT: a Benchmark for Knowledge Intensive Language Tasks : https://arxiv.org/pdf/2009.02252
3. A Survey for Efficient Open Domain Question Answering : https://arxiv.org/pdf/2211.07886
4. RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems : https://arxiv.org/pdf/2407.11005
5. Reciprocal Rank Fusion: https://plg.uwaterloo.ca/~gvcormac/cormacksigir09-rrf.pdf
6. Gemini: A Family of Highly Capable Multimodal Models: https://arxiv.org/pdf/2312.11805
7. The Llama 3 Herd of Models: https://arxiv.org/pdf/2407.21783

Dataset Exploration:

The KILT Benchmark Dataset, released by Meta, is a comprehensive collection of datasets designed for various tasks like Entity Linking, Open-Domain QA, and Dialog. It incorporates several well-known datasets, such as AIDA CoNLL-YAGO for entity linking, and datasets like Natural Questions, HotspotQA, TriviaQA, and ELI5 for open-domain QA. These datasets are based on a 2019 Wikipedia dump, providing a rich source of data for various NLP tasks.

For our project, the focus will be on Open-Domain QA, particularly using Natural Questions, HotspotQA, TriviaQA, and ELI5. However, instead of relying on the large 2019 Wikipedia dump for context, the approach will involve using these datasets without the dump (due to the size of the dump being huge). Alternative methods for context retrieval will be explored in the Methodology section.

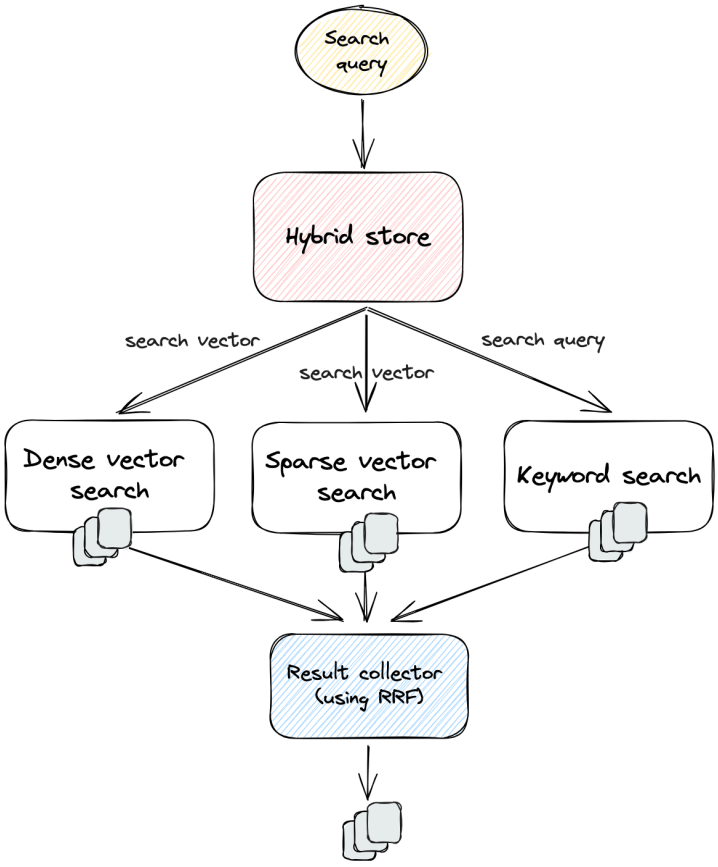
For context-based experiments, we will construct our document collection from the wiki dump by choosing a fraction from it. We will take the validation set for Natural Questions, HotspotQA, TriviaQA, and ELI5 and only get those documents that their answers belong to. But then the number of relevant documents will be equal to the number of questions in the validation dataset. So we will make multiple variants where the ratio of documents that is the context for at least 1 question to irrelevant documents will be 1:0, 1:1, 1:10, and 1:100. This is to see how the retrieval methods scale with an exponential increase in irrelevant documents.

Methodology:

1. Dataset Construction : Construction the mini-Wiki collection

2. Constructing Retrieval System :
    a. Keyword Search (Direct Query & Modified Query)
    b. Sparse Vector Search (Tf-IDF & BM25)
    c. Dense Vector Search
        i. 1 Open Source Embedding and 1 OpenAI Embedding
        ii. Cosine, Euclidean, some other
        iii. (Direct Query & Modified Query)
    d. Convert Documents to images and embed them in the Qwen 2b Vision Model.
    e. Result Collector (Rank Reciprocal Fusion)

3. Answer Generation :
    a. For without Context Dataset
        i. Making Agents for Web Search (Google) and Wikipedia.
        ii. Extracted Information will be fed to a LLM to generate answers.
    b. For with Context Dataset
        i. Top Documents will be fed to a LLM to generate answers.



Challenges:

1. **Data Integration:**
    a. Combining and processing data from multiple, heterogeneous datasets and sources, including the KILT Benchmark and Open Domain QA datasets.
    b. Constructing a mini-Wiki collection for context-based experiments, managing the balance between relevant and irrelevant documents.
2. **Multi-Modal Retrieval:**
    a. Experimenting with converting documents into images and embedding them in vision models like Qwen 2b Vision Model to assess the effectiveness of multi-modal retrieval.
    b. Evaluating how multi-modal approaches can enhance the accuracy and relevance of answers in an open-domain QA setting.
3. **Cross-Source Reranking:**
    a. Developing and optimizing a cross-source reranking system that effectively combines outputs from multiple agents using techniques like Reciprocal Rank Fusion.
    b. Ensuring that the reranking mechanism is scalable and can handle large volumes of data while maintaining the quality of the final answers.
4. **Scalability:**
    a. Ensuring the system can efficiently manage large datasets and high query volumes, especially when scaling the ratio of irrelevant to relevant documents in context-based experiments.
    b. Addressing computational challenges associated with integrating multiple retrieval techniques and running multi-agent systems in parallel.

Tentative Timeline:

20th September: Literature Review & Dataset Exploration

- Conduct an in-depth review of the literature, focusing on recent advances in open-domain question answering (ODQA), multi-agent systems, and cross-source reranking.
- Explore and select the appropriate datasets, including the KILT Benchmark, Natural Questions, HotspotQA, TriviaQA, ELI5, and Semeval 2025 Task 7.
- Begin initial experiments with baseline models to understand the data distribution and challenges.

5th October : Retrieval Agent Development & Initial Pipeline Setup

- Develop specialized retrieval agents that can query diverse sources such as web searches (e.g., Google), specific websites, and large language models (LLMs) like Llama 3.
- Implement the initial version of the ODQA pipeline, integrating these retrieval agents.
- Start constructing the mini-Wiki collection for context-based experiments.

17th October : Reranker Model Development & Multi-Agent System Integration

- Develop the re-ranker model, utilizing techniques like Reciprocal Rank Fusion to aggregate and rank the outputs from multiple agents.
- Integrate the re-ranker model with the multi-agent system, ensuring that the final answers are both accurate and relevant.
- Begin testing the entire pipeline with both context-free and context-based datasets.

31st October : Multi-Modal Retrieval & Advanced Evaluation

- Experiment with multi-modal retrieval by converting documents into images and embedding them in vision models like Qwen 2b Vision Model.
- Evaluate the effectiveness of multi-modal approaches in enhancing the accuracy and relevance of answers in the ODQA setting.
- Conduct comprehensive evaluations across different retrieval techniques and dataset variants.

10th-15th November: Final Integration, Testing, & Refinement

- Finalize the integration of all components, ensuring seamless interaction between retrieval agents, the reranker model, and multi-modal retrieval systems.
- Conduct extensive testing and refinement of the pipeline to address any remaining challenges, including scalability and performance optimization.
- Prepare the final report and project presentation, summarizing key findings, challenges, and future directions.

Project Github:

https://github.com/yash9439/MultiAgent-OpenDomain-QnA-System

References:

https://ai.meta.com/tools/kilt/
https://huggingface.co/datasets/nvidia/ChatRAG-Bench
https://huggingface.co/datasets/rungalileo/ragbench
https://arxiv.org/pdf/2407.11005
https://www.assembled.com/blog/better-rag-results-with-reciprocal-rank-fusion-and-hybrid-search
https://arxiv.org/pdf/2005.11401
https://arxiv.org/pdf/2009.02252
https://arxiv.org/pdf/2211.07886
https://plg.uwaterloo.ca/~gvcormac/cormacksigir09-rrf.pdf
https://arxiv.org/pdf/2312.11805
https://arxiv.org/pdf/2407.21783