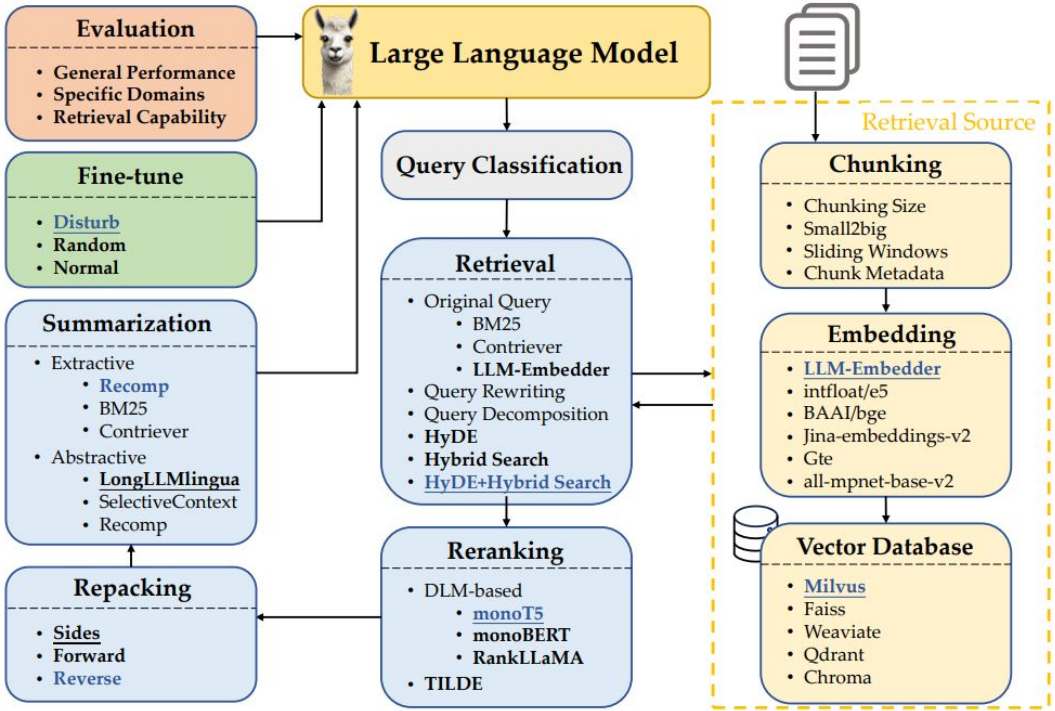
Better Retrieval for Generation

CS6803 - Topics in NLP
Dr. Maunendra Sankar Desarkar

By:
Aaryan - CO21BTECH11001
Abhishek Kumar - AI21BTECH11003

RAG Pipeline



Source: Wang, Xiaohua & Wang, Zhenghua & Gao, Xuan & Zhang, Feiran & Wu, Yixin & Xu, Zhibo & Shi, Tianyuan & Wang, Zhengyuan & Li, Shizheng & Qian, Qi & Yin, Ruicheng & Lv, Changze & Zheng, Xiaoqing & Huang, Xuanjing. (2024). Searching for Best Practices in Retrieval-Augmented Generation. 10.48550/arXiv.2407.01219.

Problem Statement

- While the goal is to improve the overall RAG system, we will specifically focus on **retrieval methods**
- Experiments on efficient fine-tuning of retrievers using specific datasets
- Optimize the retriever's accuracy and overall answer helpfulness

Need for fine-tuning of retriever

- Retriever models are trained on generic datasets from diverse sources
- When dealing with **domain-specific data**, a pre-trained retriever may not retrieve the most relevant documents
- Objective: Improve retrieval accuracy

Parameter-Efficient Fine-Tuning with LoRA

- Reduces the number of trainable parameters in transformer models

$$W \rightarrow W + AB$$
$$\Delta W = A \cdot B$$

- W: weights of original model are **freezed**
- A and B are small low-rank matrices: learnable
- Achieves comparable performance
- Review about other such methods is needed

Source:

```
@inproceedings{hu2022lora,
title={Lo{RA}: Low-Rank Adaptation of Large Language Models},
author={Edward J Hu and yelong shen and Phillip Wallis and Zeyuan Allen-Zhu and Yuanzhi Li and Shean
Wang and Lu Wang and Weizhu Chen},
booktitle={International Conference on Learning Representations},
year={2022},
url={https://openreview.net/forum?id=nZeVKeeFYf9}
}
```

Datasets

From TriviaQA

Question :“`Which American-born Sinclair won the Nobel Prize for Literature in 1930?`”

Gold document: “ `"description": [ "The Nobel Prize in Literature 1930 Sinclair ... The Nobel Prize in Lite….`”

Gold Answer: “`{ "aliases": [ "(Harry) Sinclair Lewis", "Harry Sinclair Lewis", "Lewis, (Harry) ….`”

From HotpotQA

Question :“`When was the institute that owned The Collegian founded?`”

Gold document: “`[ { "idx": 0, "title": "Pakistan Super League", "paragraph_text": "Pakistan Super League (Urdu:`”

Gold Answer: “`1960`”

Metrics Of Comprehensive Evaluation

- Specific Domains

  common -sense reasoning (accuracy), fact checking(accuracy) , open-domainQA(token level F1 score and EM score) , mulithopQA(token level F1 and EM score) , MedicalQA(accuracy) dataset samples

- General performance

  RAG score = simple average of scores from specific domains

- Retrieval Capability

  Five metrics

- Retrieval Capability

  Faithfulness :- factual consistency of generated answer with retrieved context(GPT-4 judge)

  Context Relevancy :- no of relevant sentences / Total sentences retrieved (GPT -4 judge)

  Answer Relevancy :- relevancy of answer wrt query (GPT - 4 judge)

  Answer Correctness :- EM with ground truth

  Retrieval Similarity : - cosine similarity of retrieved document and gold document

Multi-Modal Fine-Tuning (Optional)

- If time permits, we'll explore parameter-efficient fine-tuning (e.g., using LoRA) for multi-modal models
- Fine-tuning retrievers for multi-modal data (text, images) can **greatly enhance** tasks like visual question answering
- Utilize datasets like MMDialog for fine-tuning and evaluation
- Directions: UniMUR, Multimodal Retrieval: Survey

Timelines

10-15 days: literature review

10-15 days: experiments on fine-tuning of retriever

10-15 days: evaluation of retrieval methods and overall RAG pipeline

Extensive documentation at the end

Date :- .....

Appendix

Other Datasets

- Query classification on Databricks-Dolly-15K

  {"instruction": "Why mobile is bad for human", "context": empty}
  {"instruction": "When did Virgin Australia start operating?" , “context” : non-empty}
- DLM ranking - relevance classification (rank on basis of true probabaility)
  “Query :- what is pascal's law in simple terms?” , “labels :- top 1000(x) ranked passages”(MS MARCO dataset)
- Retriever Fine Tuning and Generator Fine Tuning
  NQ , TriviaQA , HotPotQA, 2WikiMultiHopQA , MuSiQue datasets
  Each entry is a (question,golddocument,goldanswer)