TEAM ONE

CRAG -- 隊伍一

META COMPHREHENSIVE RAG BENCHMARK STARTER KIT

Brian J Chan
Chao-Ting Chen
Jui-Hung Cheng
Kuei-Chung Chen

TEAM ONE

Problem Statement

Given 1. some queries
2. web pages (likely) relevent to the queries

Goal: implement a RAG system, providing useful information to answer the question without adding any hallucination

TEAM ONE

Paper review

2024 KDD Cup CRAG Workshop

- 2024 KDD Cup CRAG Workshop: UM6P Team Technical Report
- Winning Solution For Meta KDD Cup’ 24
- KDD Cup Meta CRAG 2024 Technical Report: Three-step Question-Answering Framework
- Honest AI: Fine-Tuning "Small" Language Models to Say "I Don’t Know", and Reducing Hallucination in RAG

TEAM ONE

Paper review

2024 KDD Cup CRAG Workshop: UM6P Team Technical Report

- Cross encoding
- Query classification

TEAM ONE

Paper review

Winning Solution For Meta KDD Cup’ 24

- Finetune RAG specific model.

TEAM ONE

Paper review

KDD Cup Meta CRAG 2024 Technical Report: Three-step Question-Answering Framework

- RAG Approach Enhanced by Category Classification with BERT
- Chain of Thought
- Voting

TEAM ONE

Paper review

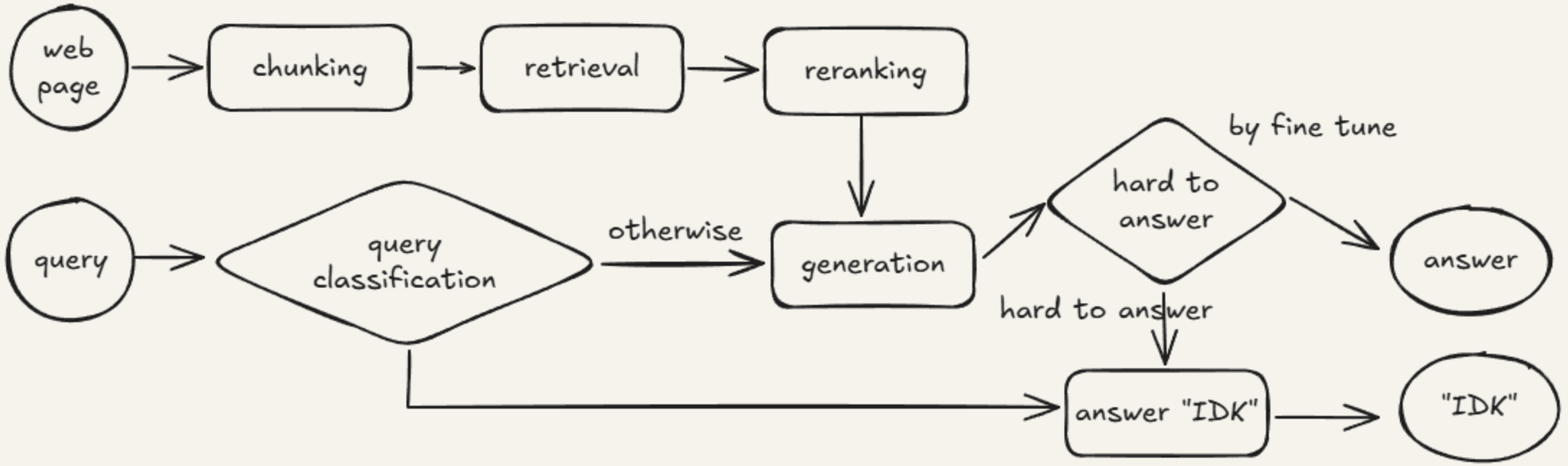
Honest AI: Fine-Tuning "Small" Language Models to Say "I Don’t Know", and Reducing Hallucination in RAG

Devide the Dataset into Easy Question or Hard Question

- If question_type is “comparison” or “false_premise”, **Keep the answer**
- If the answer is “yes” / “no” / “True” / “False”, **Keep the answer**
- otherwise, Replace answer with “**I don’t know**”

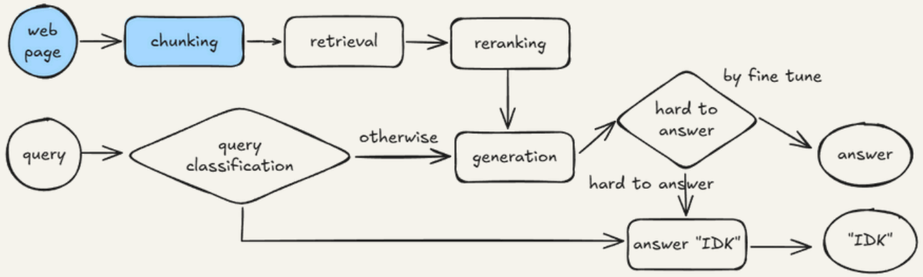
TEAM ONE

Our Solution



TEAM ONE

Preprocessing



1. BeautifulSoup4

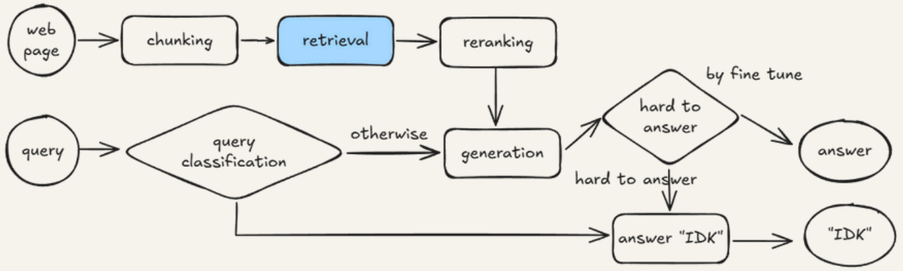
   - Parsing HTML to string

2. chunk: cross encoding

   - Dividing into smaller pieces

TEAM ONE

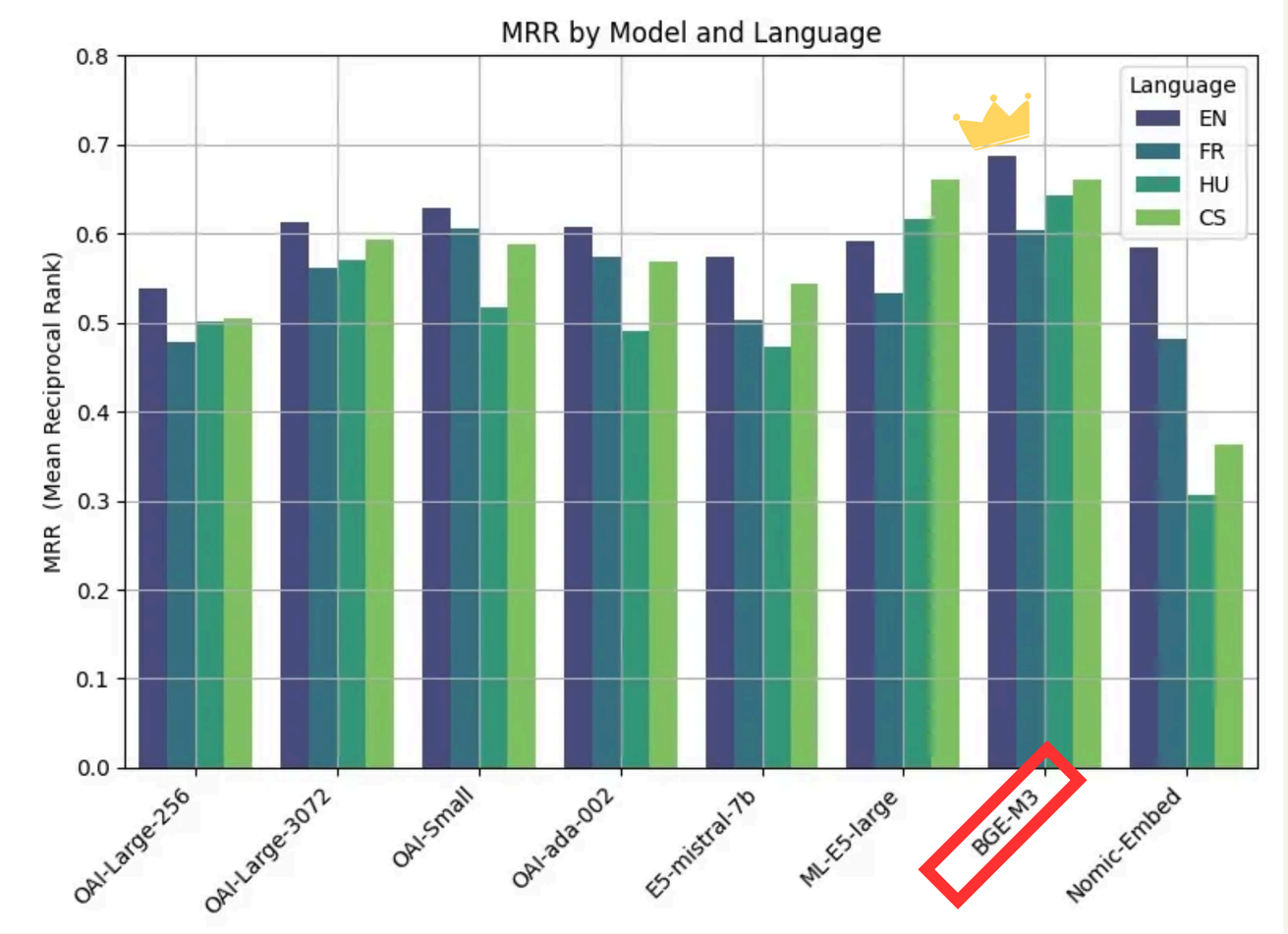
Retrival : BGE-M3



BGE-M3 is better than

- Openai text-embedding
- Microsoft E5-mistral-7b
- Microsoft ML-E5-large
- Nomic-embed-text-v1
- Jina Embedding .

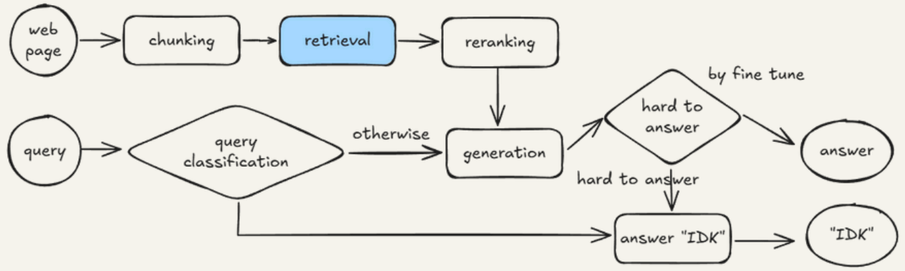
in English、Czech(捷克語)、
French、Hungarian(匈牙利語)



https://arxiv.org/abs/2402.03216v3

https://huggingface.co/BAAI/bge-m3

TEAM ONE



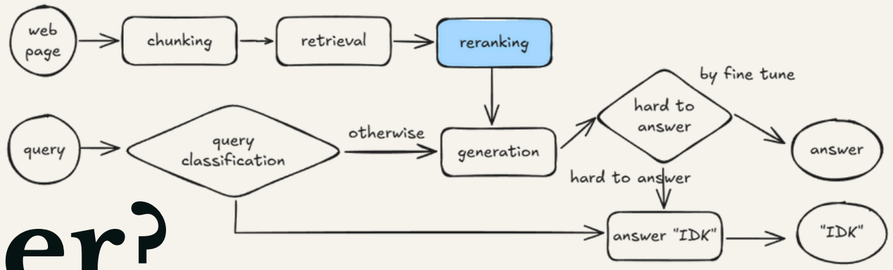
Retrival : BGE-M3

Baseline

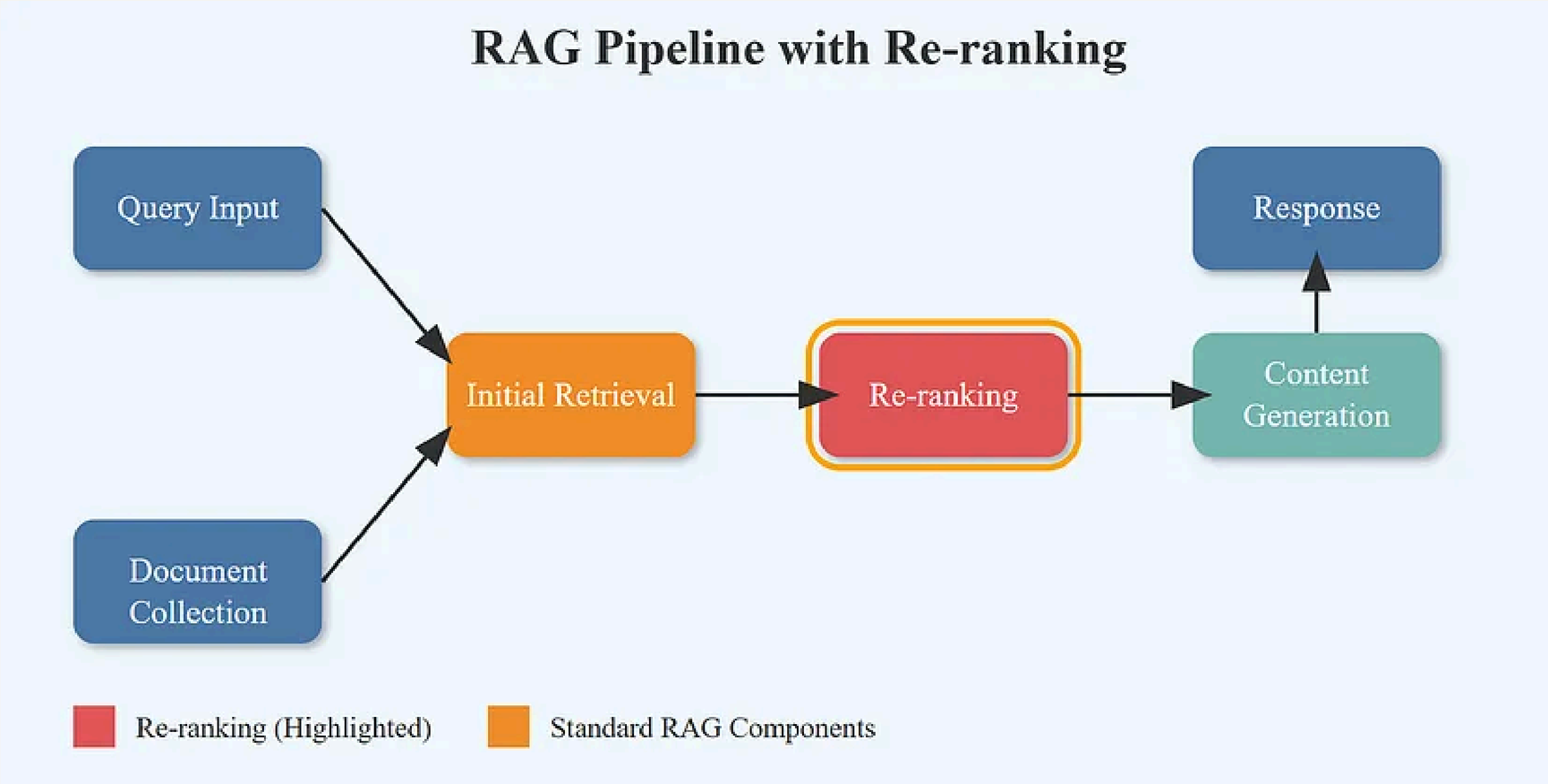
Model	BAAI/bge-m3	sentence-transformers/all-MiniLM-L6-v2
Exact AC	0.0	0.0
AC	0.2	0.1
Hallucination	0.3	0.2
Miss	0.5	0.7

10 questions with top-3

TEAM ONE



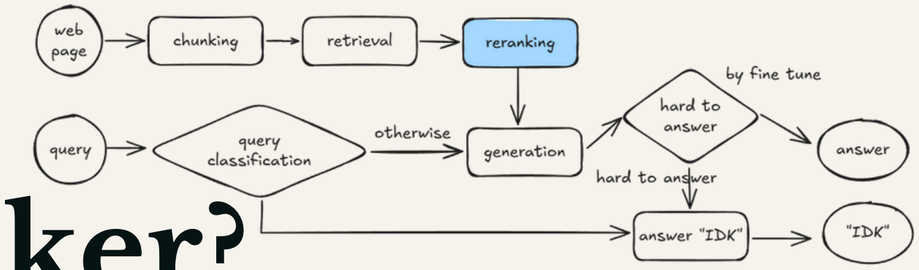
Why using Reranker?



https://medium.com/@sahin.samia/what-is-reranking-in-retrieval-augmented-generation-rag-ee3dd93540ee

TEAM ONE

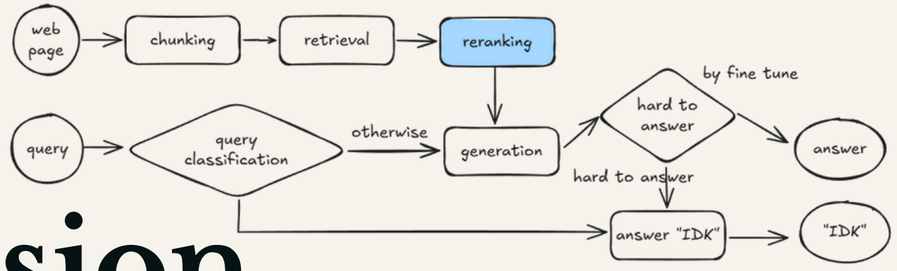
How we use Reranker?



Attempt

- LLM Reranker (OpenAI)
- Sentence Transformer Reranker **(cross-encoder/ms-marco-MiniLM-L-2-v2)**
- Flag Embedding BGE M3 Reranker **(BAAI/bge-reranker-large)**

TEAM ONE



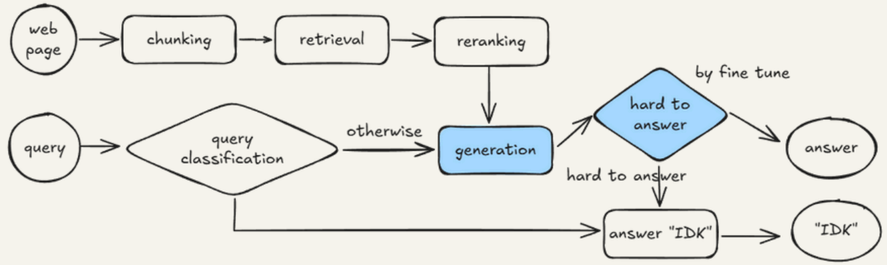
Reranker Comparision

Model	LLM Reranker	Sentence Transformer	BGE Reranker
Exact AC	0.0	0.0	0.0
AC	0.0	0.1	0.2
Hallucination	0.6	0.2	0.3
Miss	0.4	0.7	0.5

10 questions with top-3

TEAM ONE

Finetune Dataset

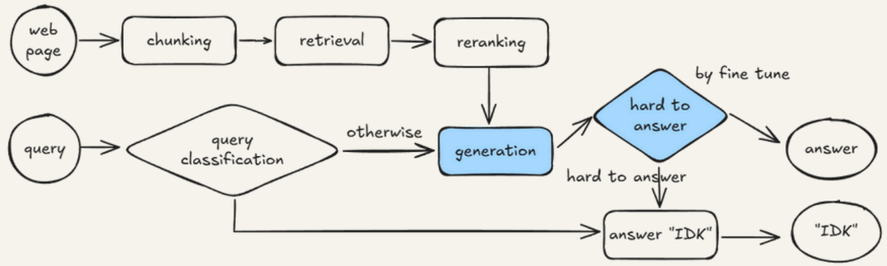


<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are a helpful and honest assistant. Please, respond concisely and truthfully in 70 words or less.Now is {time}<|eot_id|><|start_header_id|>user<|end_header_id|>
Context information is below.
<DOC>
{A}
</DOC>
<DOC>
{B}
</DOC>
Answer:<|eot_id|><|start_header_id|>assistant<|end_header_id|>

{i don't know || answer }<|eot_id|><|end_of_text|>

TEAM ONE

Finetune Dataset

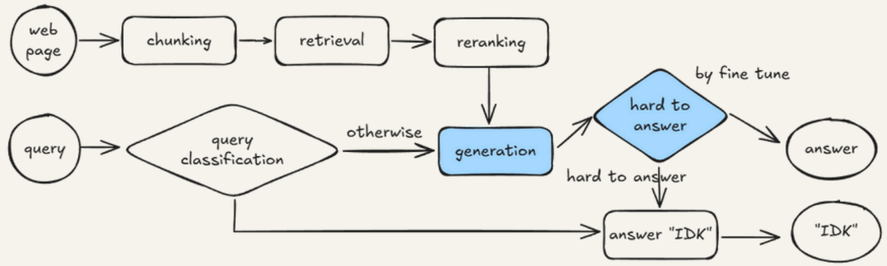


Retrieve 是否 Miss，由GPT4等大模型做判斷
是否答對，也由 GPT4 做判斷

	Retrieve Hit	Retrieve Miss
Ans Correct	直接放進 Dataset	直接放進 Dataset
Ans Wrong	修正答案再放進 Dataset	答案改為 IDK 再放進 Dataset

TEAM ONE

Finetune

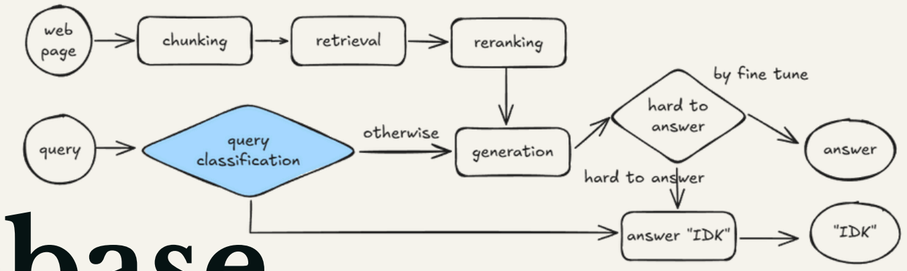


Model	Llama3.1 8B Instruction
Exact AC	0.15
AC	0.16
Hallucination	0.02
Miss	0.8

10 record

TEAM ONE

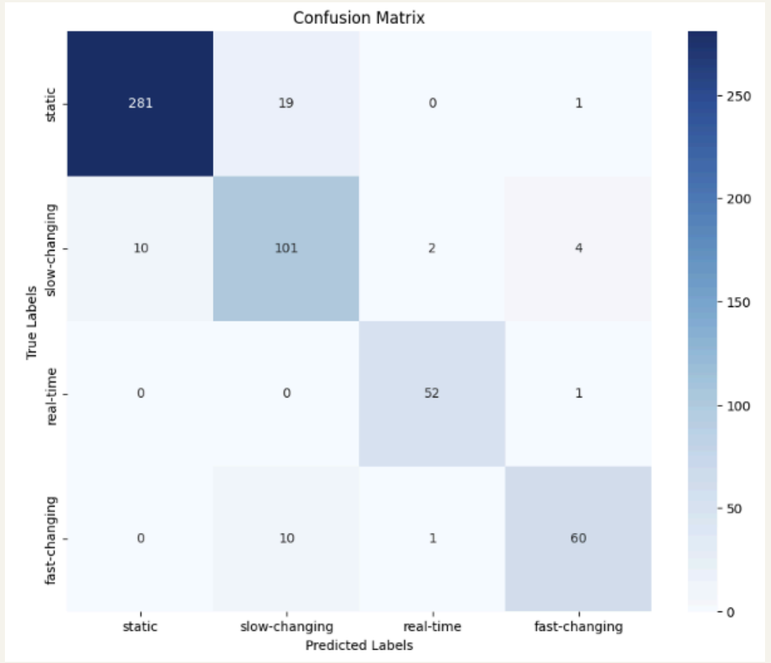
Classify: distilbert-base



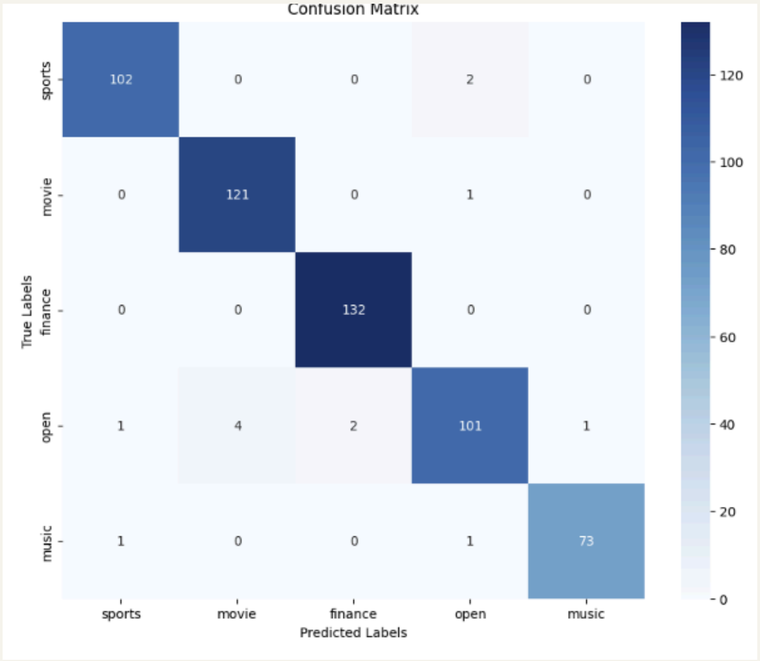
Classify query according to the following three criteria:

- Domain:
  - Movie, Music, etc
- Static_or_dynamic
- Question_type:
  - Simple, Aggregation, false premise, etc

Macro accuracy of three criteria: 94.67



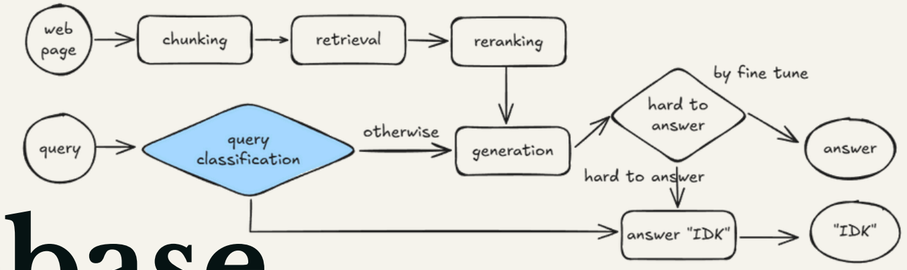
Static_or_dynamic



Domain

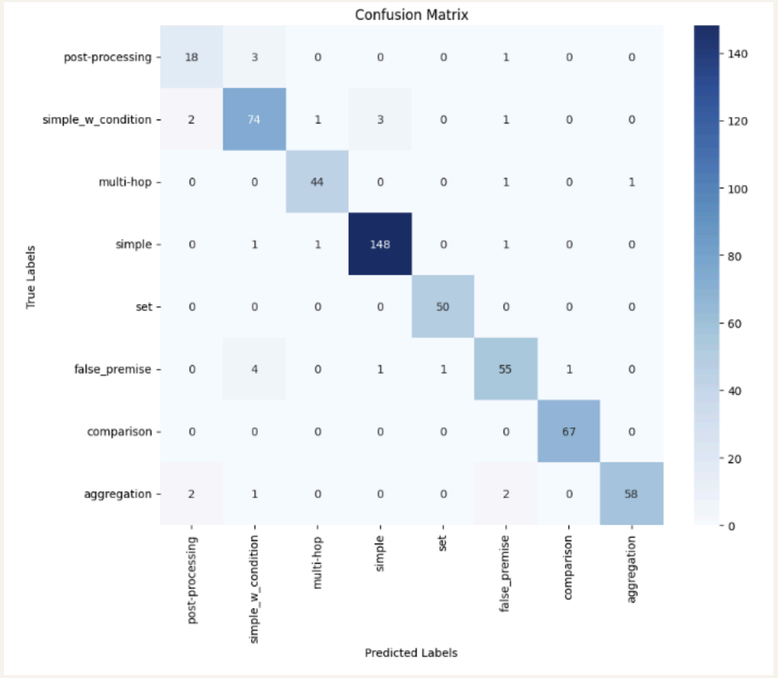
TEAM ONE

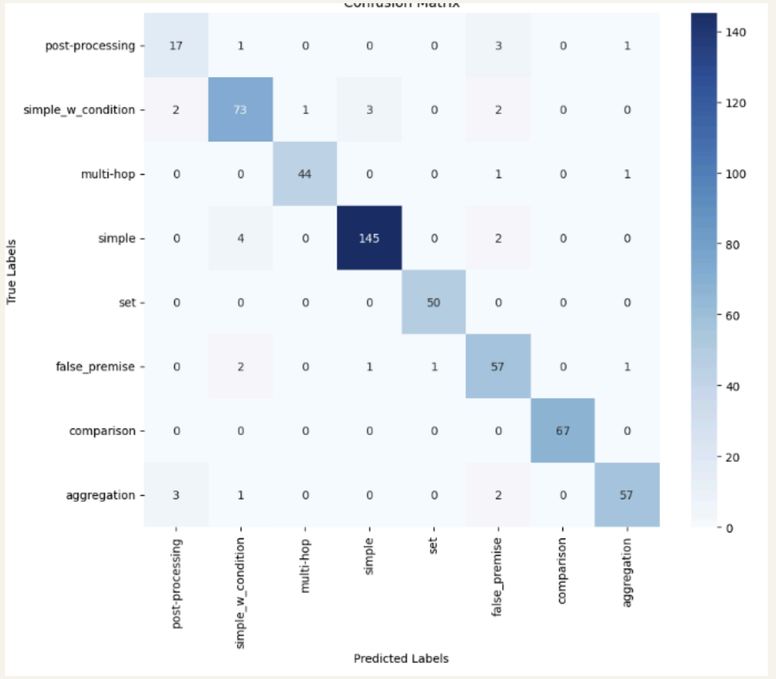
Classify: distilbert-base



Attempting to balance the data -- didn't work

Origin dataset
Accuracy: 0.95
Macro F1: 0.94



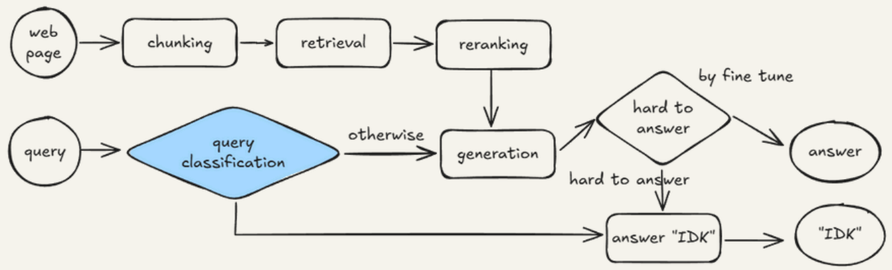


Balanced dataset
Accuracy: 0.94
Macro F1: 0.93

Question_type

TEAM ONE

Query classifier



Model	without classifier	distilbert-base-uncased(strict)	distilbert-base-uncased(loose)
Exact AC	0.25	0.20	0.25
AC	0.34	0.24	0.34
Hallucination	0.09	0.02	0.07
Miss	0.57	0.74	0.59

TEAM ONE

Conclusion (Fine-tuned)

Classifier	Retrieval top-n	Rerank top-k	Score
Strict	**10**	3	Exact AC: 0.24 AC: 0.29 Hallucination: 0.03 Miss: 0.68 Score: 0.26
Loose	**15**	*5*	Exact AC: 0.25 AC: 0.34 Hallucination: 0.07 Miss: 0.59 Score: 0.27

Citation

1. 2024 KDD Cup CRAG Workshop
2. Don't Do RAG: When Cache-Augmented Generation is All You Need for Knowledge Tasks
3. What is Reranking in Retrieval-Augmented Generation (RAG)?
4. BGE Paper : https://arxiv.org/abs/2402.03216v3

TEAM ONE

Contribution

陳肇廷	**Retriever; Reranker; Testing**
鄭睿宏	**Retriever; Reranker**
詹松霖	**Fine tuning; Generation**
陳歸中	**Chunking; Query classification**

TEAM ONE

The End

THANK YOU FOR LISTENING