Revisiting the Solution of Meta KDD Cup 2024 : CRAG

Jie Ouyang, Yucong Luo, **Mingyue Cheng***, Daoyu Wang, Shuo Yu, Qi Liu, and Enchong Chen

State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China (USTC)

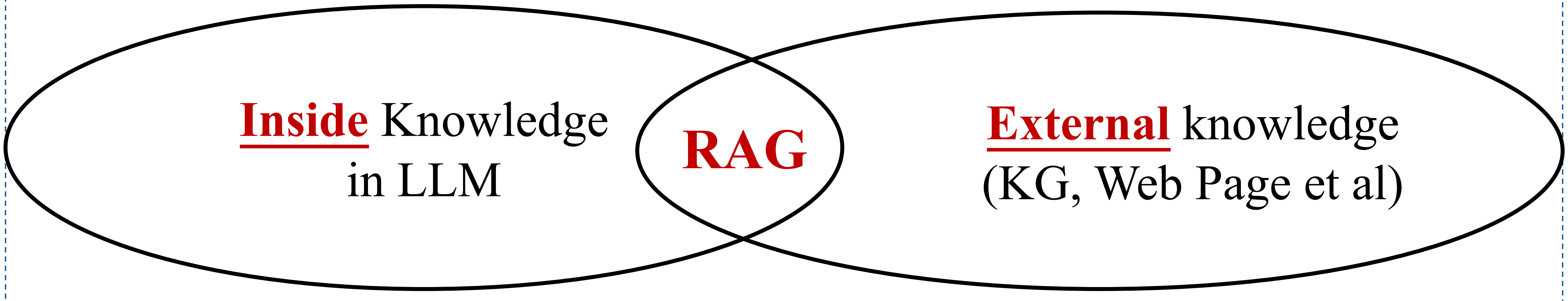
Team APEX

for Task 2&3

Background

Problem Definition

➢ **LLM Hallucination**
- **Dynamic** (facts changing over time)
- **Diverse** (torso and tail facts)



Background

Main Challenges

➢ **Challenges of Retrieval-Augmented Generation**

① Knowledge Indexing: GraphRAG

② Knowledge Retrieve: sparse, dense or hybrid retrieval

③ LLM Reasoning: Chain of –thought (CoT), In-context Learning (ICL)

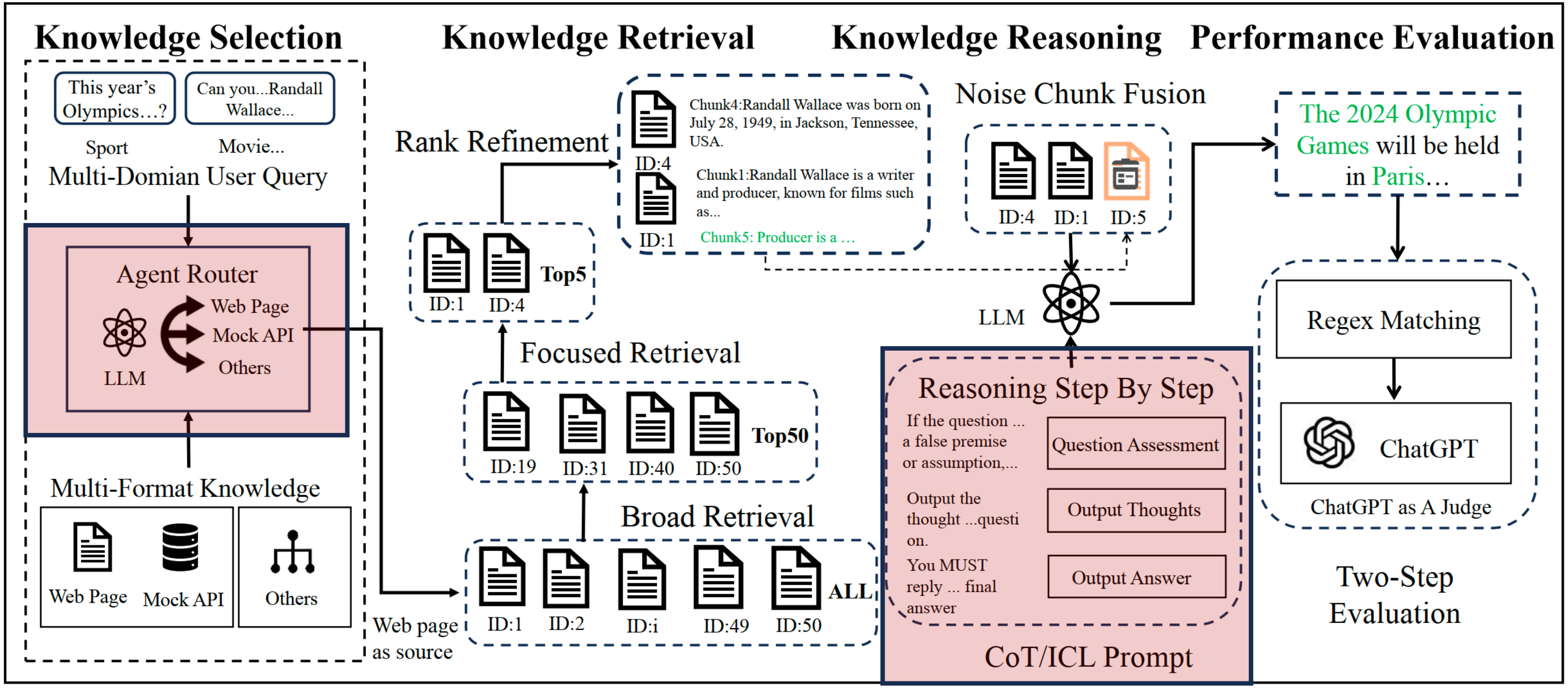
CRAG - Comprehensive RAG Benchmark

Xiao Yang*[1], Kai Sun*[1], Hao Xin*[3], Yushi Sun*[3], Nikita Bhalla[1], Xiangsen Chen[4], Sajal Choudhary[1], Rongze Daniel Gui[1], Ziran Will Jiang[1], Ziyu Jiang[4], Lingkun Kong[1], Brian Moran[1], Jiaqi Wang[1], Yifan Ethan Xu[1], An Yan[1], Chenyu Yang[4], Eting Yuan[1], Hanwen Zha[1], Nan Tang[4], Lei Chen[3,4], Nicolas Scheffer[1], Yue Liu[1], Nirav Shah[1], Rakesh Wanga[1], Anuj Kumar[1], Wen-tau Yih[2], and Xin Luna Dong[1]

[1]Meta Reality Labs, [2] FAIR, Meta, [3] HKUST, [4] HKUST (GZ)

Methodology

Overall Pipeline



Methodology

Knowledge Selection

Routing is a crucial component of RAG systems, especially in real-world QA scenarios. In practical applications, RAG systems frequently incorporate multiple data sources.

In response to the specific characteristics of the questions in the CRAG Challenge, we designed two specialized routers: the **Domain Router** and the **Dynamism Router**.

Domain Router

(SequenceClassifier)
(Llama3-8B-Instruct)
(LORA)

LLM
Rather than LM

Domain&Dynamic

Dynamism Router

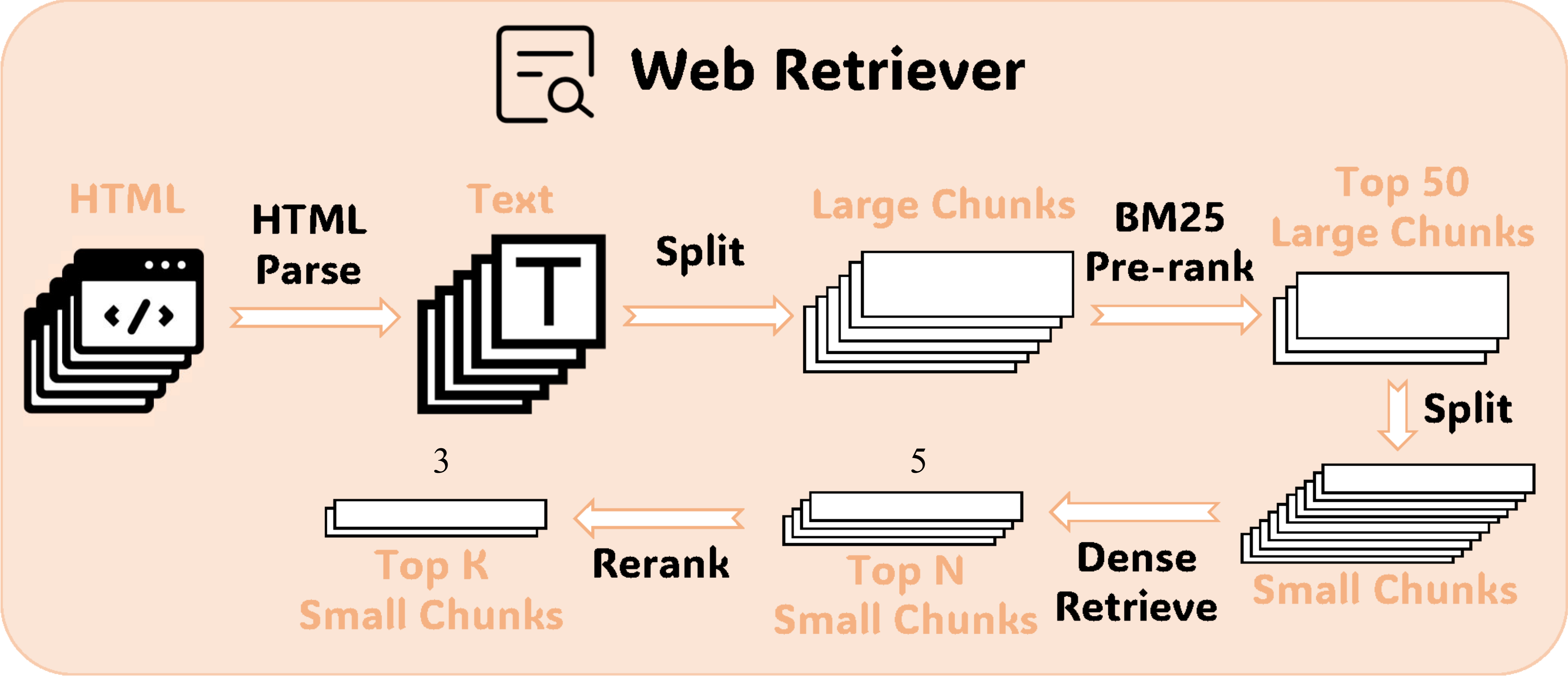
(SequenceClassifier)
(Llama3-8B-Instruct)
(LORA)

Methodology

Knowledge Retrieve

➢ Web Pages

Task 3

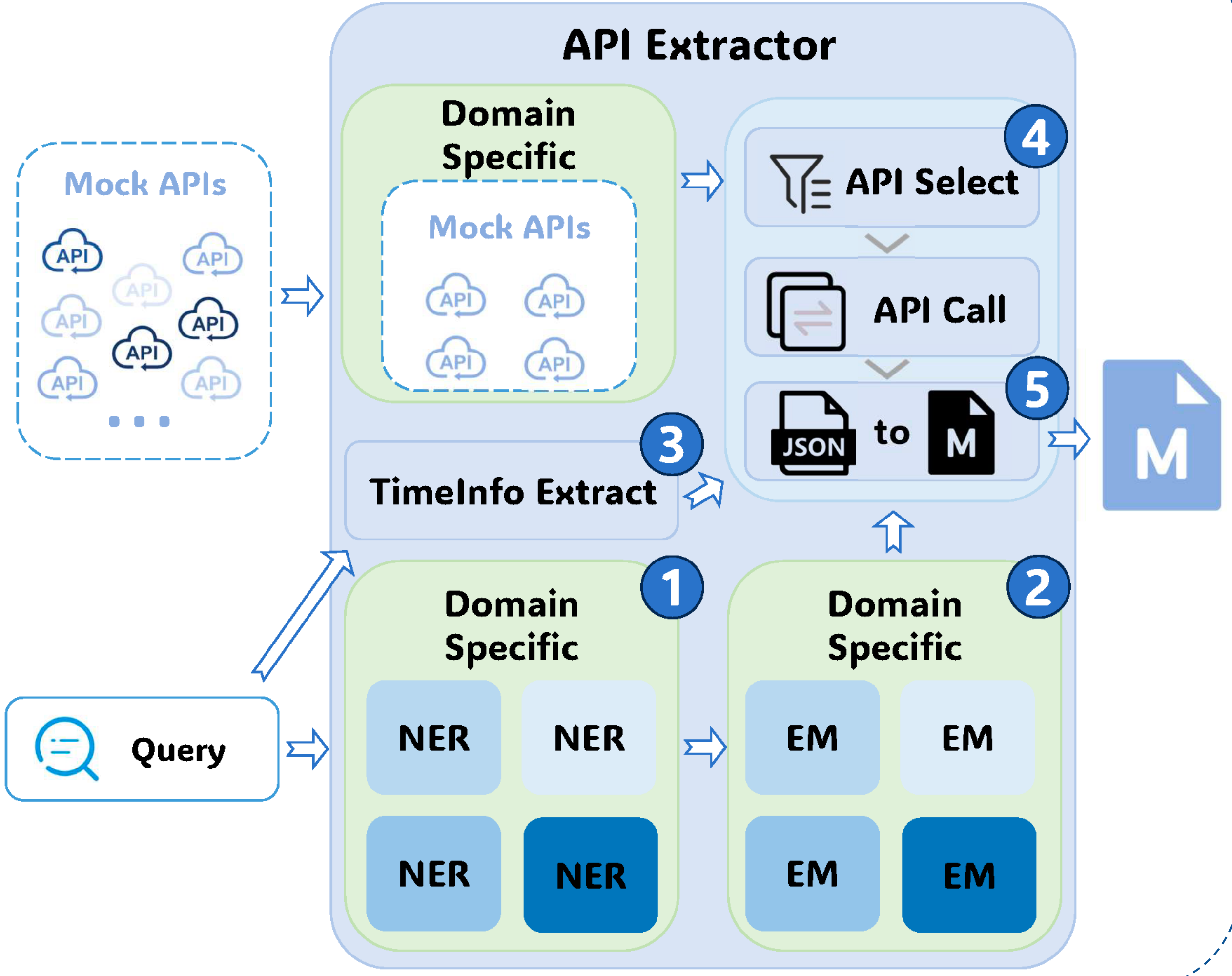


Methodology

Task 2&3

Knowledge Retrieval

- **Mock APIs**
  - **NER (query)**
    - LLM
    - Plain Text (Not JSON)
  - **Entity Match (EM)**
    - Exact Match
    - BM25 + Rules
  - **Time Information Extraction**
    - Regex
    - pytz + datetime
  - **API Post-processing**
  - **Json to Markdown**



Methodology

Knowledge Reasoning

- **Chain of Thought (CoT)**
  - Think Step by Step!
- **In-context Learning**
  - Few-shot examples (e.g., **false premises)**
  - **Adaptive** for different domains
- **Post-Processing**
  - **Domain&Dynamism** Specific
  - Reject complex numerical calculations (I don't know)

Denoising **Reasoning**

Few-shot Example I

What's the latest score for OKC's game today?

There is no game for OKC today.

Few-shot Example II

How many times has Curry won the NBA dunk contest?

Steph Curry has never participated in the NBA dunk contest.

Experiments

Experimental Settings

Components	Our choice
HTML Parser	*Newspaper3K*
Embedding Model	*BAAI/bge-m3*
Rerank Model	*BAAI/bge-m3-v2-reranker*
LLM	*Llama3-70B-Instruct (GPTQ)*

Task 2:

Building on Task 1， we concatenate the references retrieved from HTML with those retrieved from the Mock API.

Task 3:

Building on Task 2, we first use **BM25** to select the **50** most relevant passages, and then apply **embedding model** to narrow it down to the **5** most relevant ones.

Experiments

Experimental Results

➢ Overall Performance

Table 1: Overall Preformance of our solutions on all 3 Tasks.

	Score(%)	Accuracy(%)	Hallucination(%)	Missing(%)
LLM Only	-7.29	28.01	35.30	36.69
Direct RAG	-6.78	34.79	41.58	**23.63**
Task 1	11.82	29.98	18.16	51.86
Task 2	31.22	46.75	**15.54**	37.71
Task 3	**31.66**	**48.21**	16.56	35.23

Experiments

Experimental Results

➢ Ablation Study for Major Strategies

Table 2: Ablation Study for Major Strategies Employed in the System.

		Score(%)	Accuracy(%)	Hallucination(%)	Missing(%)	Time Cost(s)
Task 2	w/o Rerank	29.17	43.54	14.37	42.09	-
	w/o EntityMatch	21.44	32.31	10.87	56.82	-
	w/o TimeInfoExtract	18.45	28.45	**9.99**	61.56	-
	w/o Fewshot&CoT	25.53	52.08	26.55	**21.37**	-
	w/o Fewshot	27.13	51.35	24.22	24.43	-
	w/o CoT	28.52	**53.32**	24.80	21.88	-
	Ours	**31.22**	46.75	15.54	37.71	-
Task 3	w/o Prerank	29.53	44.34	**14.81**	40.85	68.17
	Ours	**31.66**	**48.21**	16.56	**35.23**	**5.96**

Conclusion

Our Perspectives

The quality of knowledge source is significant

- Traditional QA evaluation often overlooks hallucinations.
- Future focus: Assessing models' cognitive abilities using methods from human cognition research.

How to retrieve relevant knowledge as context is the core of RAG

- Manual rules for API matching may fall short in real-world usage.
- Future focus: Developing universal methods for selecting and calling APIs, processing results effectively.

The capacity of LLM can be roughly divided into denosing and reasoning

- Denoising: reducing the noise from numerous context.
- Reasoning: extracting useful knowledge from the limited context.

Teaching the LLM know what they do not know is very important.

- Evaluation of hallucination is very vital.

Thank You for Your Attention!

Mingyue Cheng
mycheng@ustc.edu.cn
https://mingyue-cheng.github.io/