A Tutorial on Graph RAG & Meta KDD Cup

2025.5.29
Data Systems Lab

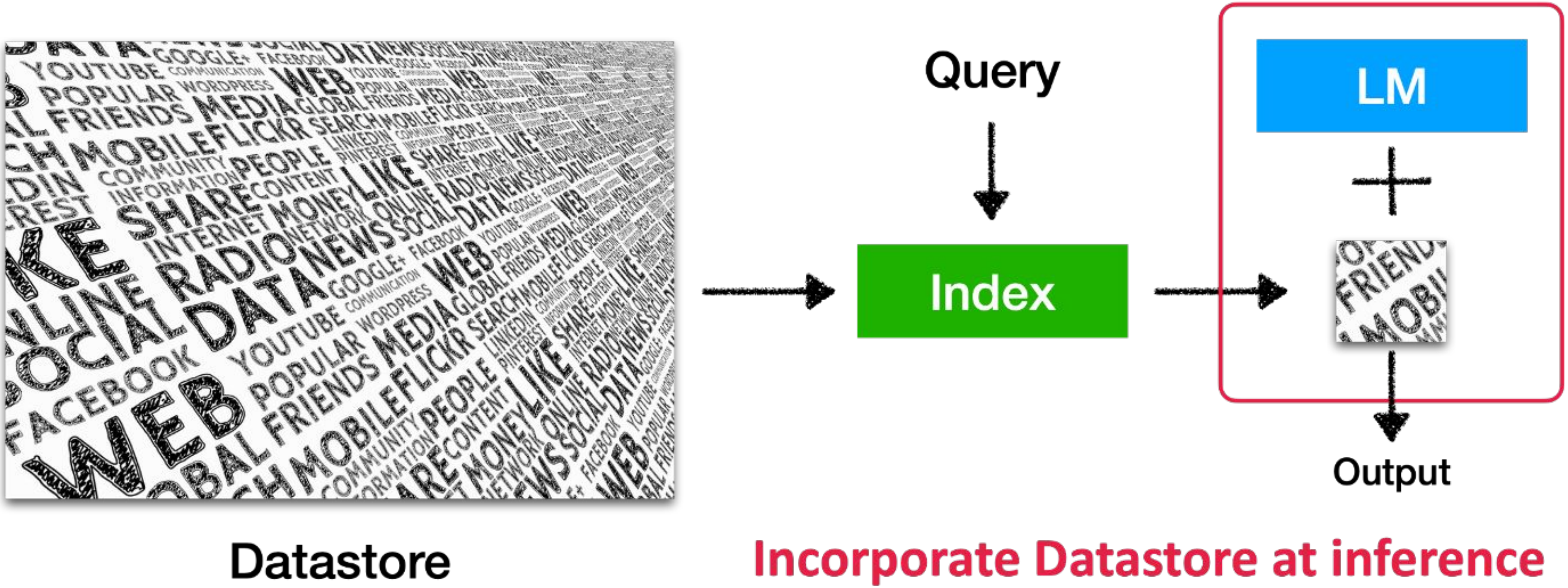
Overview

- Introduction
  - Retrieval-augmented Generation (RAG)
  - Graph RAG
- Meta KDD Cup
  - Benchmark
  - Baseline Approach

Overview

- Introduction
  - Retrieval-augmented Generation (RAG)
  - Graph RAG
- Meta KDD Cup
  - Benchmark
  - Baseline Approach

Retrieval-augmented Generation (RAG)



Incorporate Datastore at inference

Slides adapted from the ACL-23 tutorial:
Retrieval-based Language Models and Applications

RAG Mitigates Weakness of Large Language Models

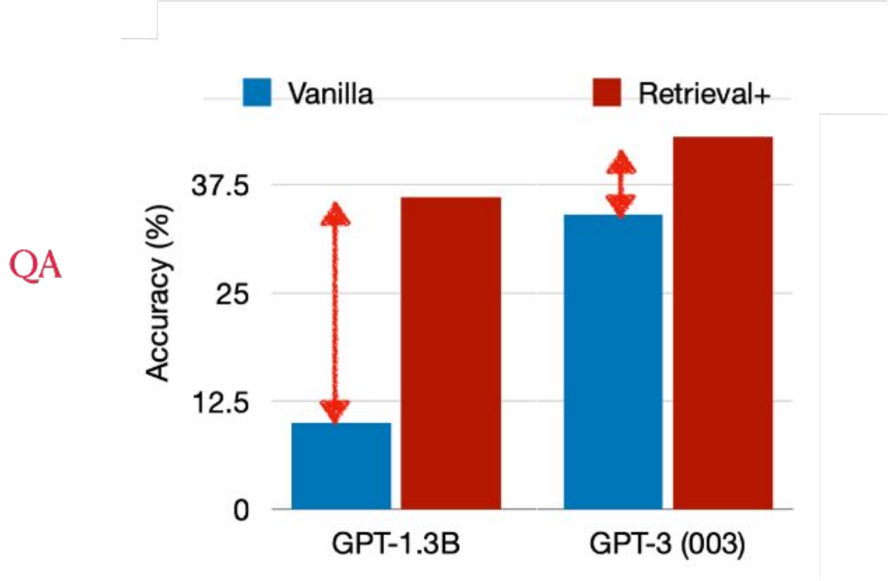
Hallucinations

Costs of adaptations

Copyright / privacy

Large parameter size

Significant improvements across model scale, with larger gain with smaller LM



Mallen*, Asai* et al., When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories (ACL, Best Video; Oral) 2023.

RAG Mitigates Weakness of Large Language Models

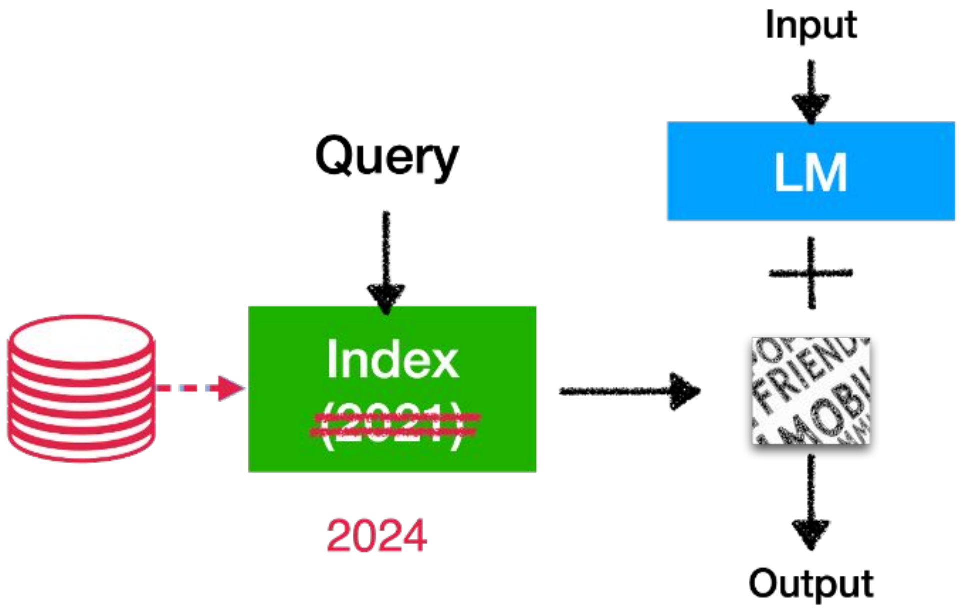
Hallucinations

Costs of adaptations

Copyright / privacy

Large parameter size

Replacing Datastore (Index) for adaptations without training



Izacard*, Lewis* et al., Atlas: Few-shot Learning with Retrieval Augmented Language Models. JMLR 2023.

RAG Mitigates Weakness of Large Language Models

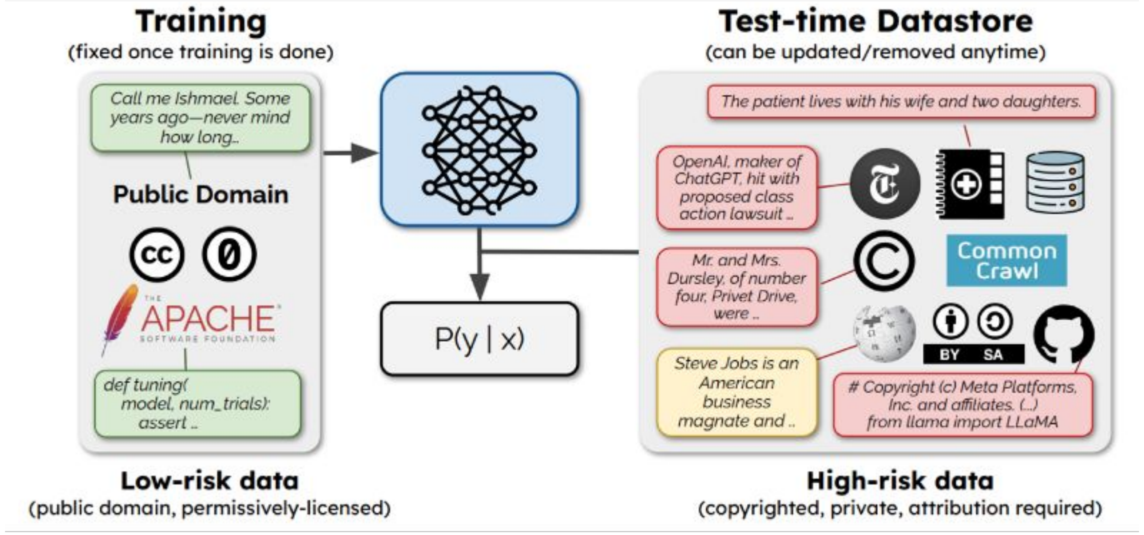
Hallucinations

Costs of adaptations

Copyright / privacy

Large parameter size

Segregating copyright-sensitive data from pre-training data


Training
(fixed once training is done)
Call me Ishmael. Some years ago—never mind how long..
Public Domain
APACHE SOFTWARE FOUNDATION
def tuning(model, num_trials): assert ..
Low-risk data
(public domain, permissively-licensed)
P(y | x)
Test-time Datastore
(can be updated/removed anytime)
The patient lives with his wife and two daughters.
OpenAI, maker of ChatGPT, hit with proposed class action lawsuit ..
Mr. and Mrs. Dursley, of number four, Privet Drive, were ..
Steve Jobs is an American business magnate and ..
Common Crawl
BY SA
# Copyright (c) Meta Platforms, Inc. and affiliates. (...) from llama import LLaMA
High-risk data
(copyrighted, private, attribution required)

Min* and Gururangan* et al., SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore. ICLR 2024

RAG Mitigates Weakness of Large Language Models

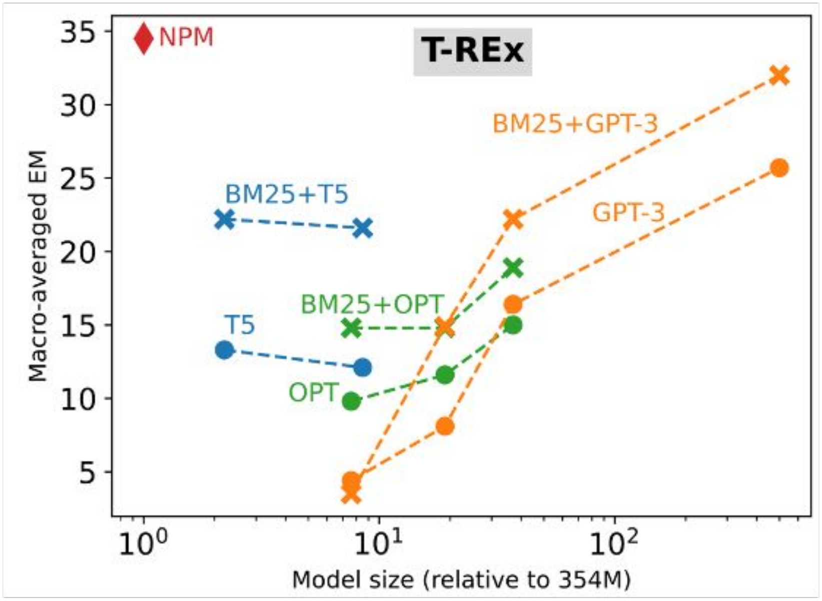
Hallucinations

Costs of adaptations

Copyright / privacy

Large parameter size

Models with much less parameters can outperforms much larger models!



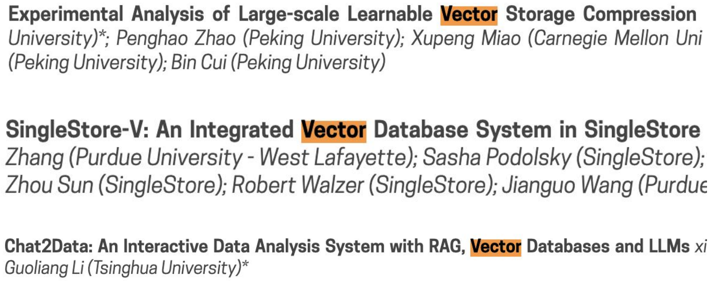
Min et al., Nonparametric Masked Language Modeling. Findings of ACL 2023.

One of the Hottest Topics in VLDB 2024: (Graph)RAG

Vector Databases

A2 Panel Vector Databases: What's Really New and What's Next?

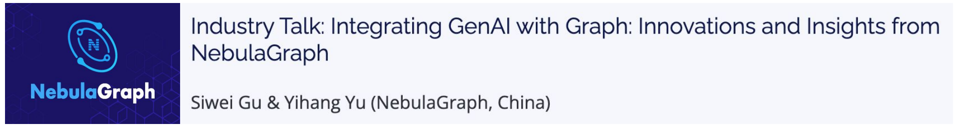
A1 Data management and support for ML/AI


Experimental Analysis of Large-scale Learnable Vector Storage Compression
University); Penghao Zhao (Peking University); Xupeng Miao (Carnegie Mellon Uni (Peking University); Bin Cui (Peking University)*

SingleStore-V: An Integrated Vector Database System in SingleStore
Zhang (Purdue University - West Lafayette); Sasha Podolsky (SingleStore); Zhou Sun (SingleStore); Robert Walzer (SingleStore); Jianguo Wang (Purdue

Chat2Data: An Interactive Data Analysis System with RAG, Vector Databases and LLMs *xi Guoliang Li (Tsinghua University)**

Graph Databases



Industry Talk: Integrating GenAI with Graph: Innovations and Insights from NebulaGraph

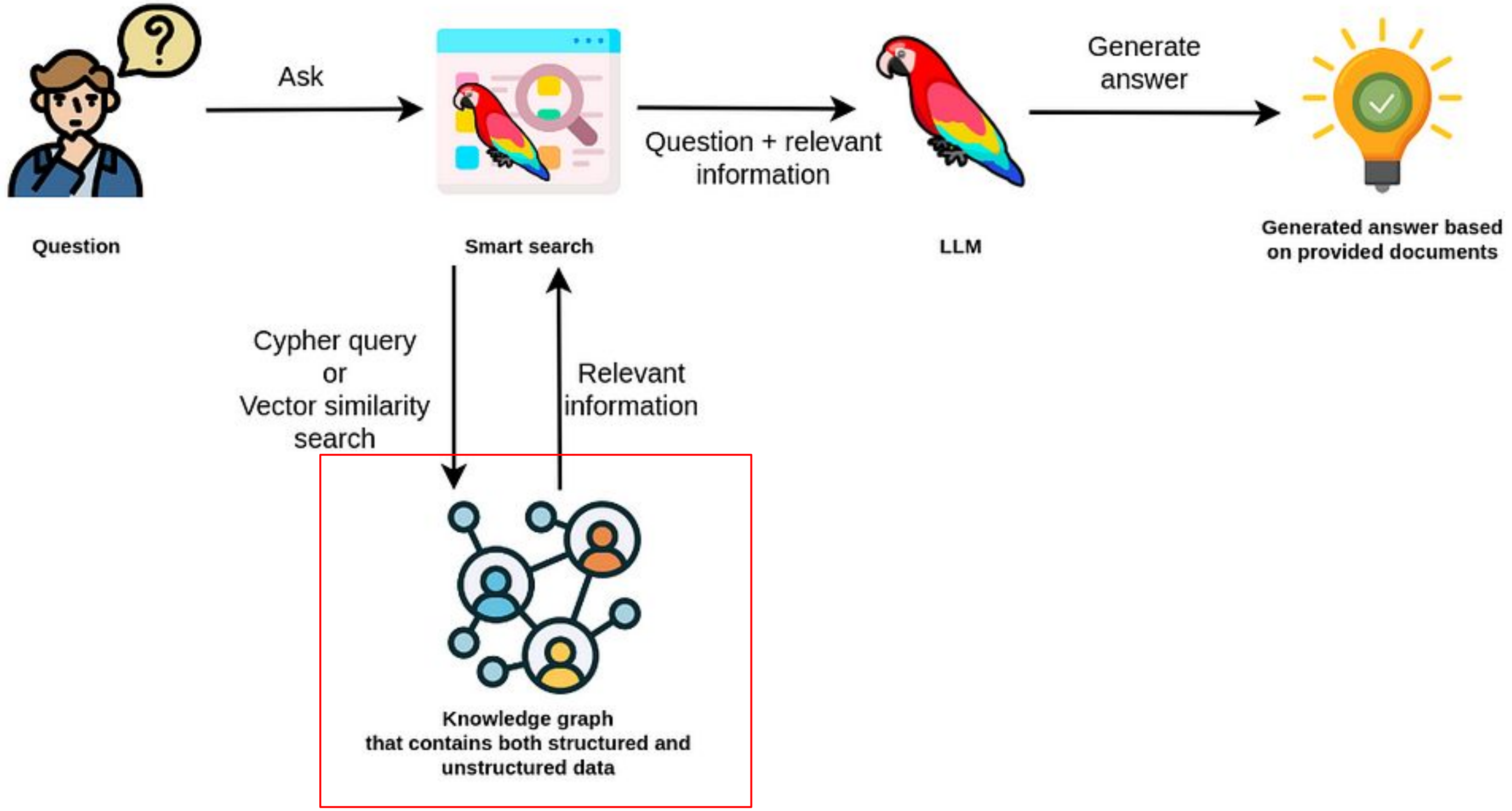
Siwei Gu & Yihang Yu (NebulaGraph, China)

LLM+KG Workshop

neo4j

Microsoft

GraphRAG Overview


Ask
Question + relevant information
Generate answer
Question
Smart search
LLM
Generated answer based on provided documents
Cypher query or Vector similarity search
Relevant information
Knowledge graph that contains both structured and unstructured data

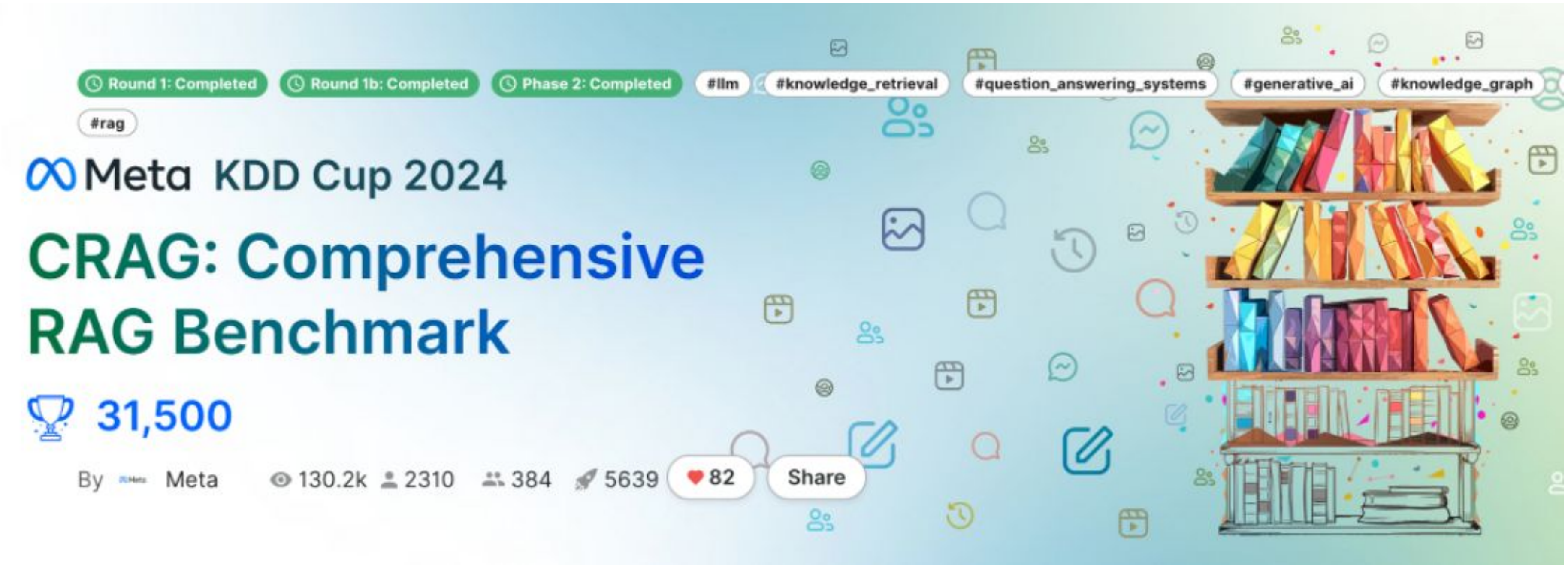
GraphRAG Overview (From Neo4j Document)

Overview

- Introduction
  - Retrieval-augmented Generation (RAG)
  - Graph RAG
- Meta KDD Cup
  - Benchmark
  - Baseline Approach

Meta KDD Cup 2024

- 2300+ participants, 384 teams, 5600+ submissions
- CRAG benchmark was listed in HuggingFace "Daily papers". Hugging Face
- Our team was the only Korean team to receive an award, achieving First Place in the Comparison Question category for Tasks 1, 2, and 3!!



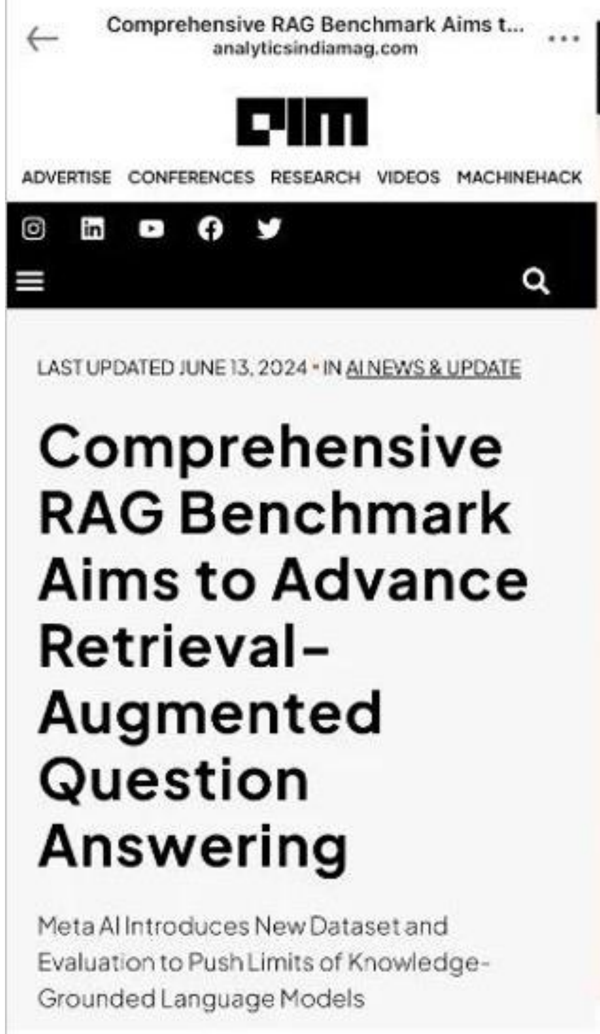
Round 1: Completed Round 1b: Completed Phase 2: Completed #llm #knowledge_retrieval #question_answering_systems #generative_ai #knowledge_graph #rag

Meta KDD Cup 2024

CRAG: Comprehensive RAG Benchmark

31,500

By Meta 130.2k 2310 384 5639 82 Share

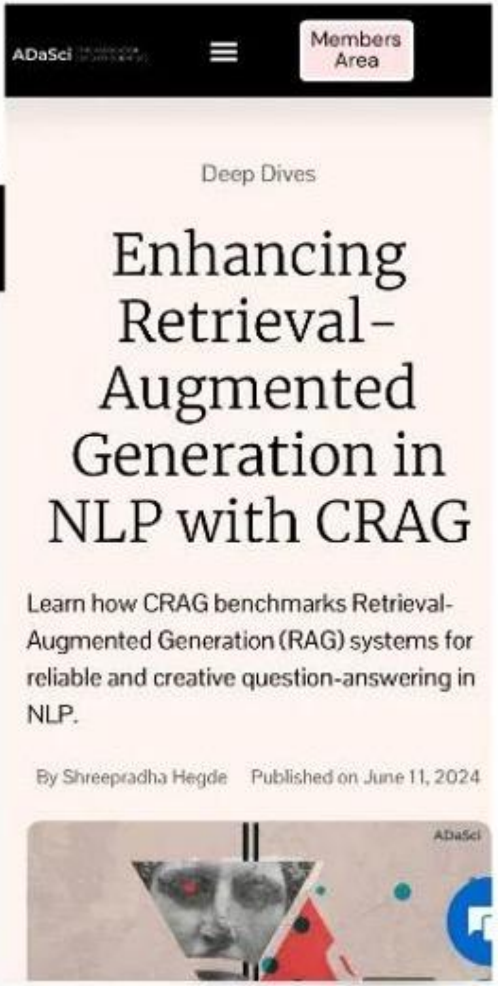


←
Comprehensive RAG Benchmark Aims t...
analyticsindiamag.com
•••
AIM
ADVERTISE CONFERENCES RESEARCH VIDEOS MACHINEHACK

LAST UPDATED JUNE 13, 2024 • IN AI NEWS & UPDATE

Comprehensive RAG Benchmark Aims to Advance Retrieval-Augmented Question Answering

Meta AI Introduces New Dataset and Evaluation to Push Limits of Knowledge-Grounded Language Models

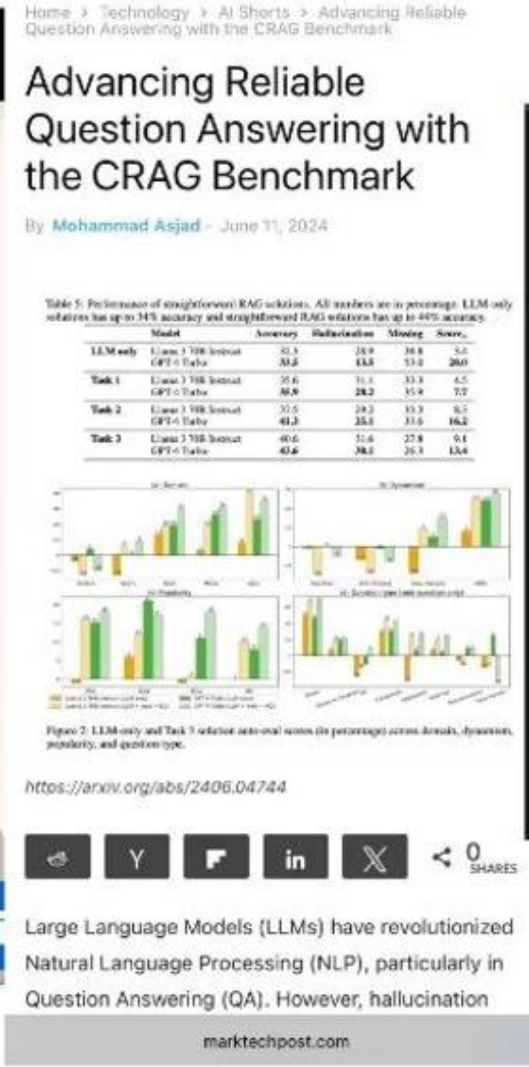


Deep Dives

Enhancing Retrieval-Augmented Generation in NLP with CRAG

Learn how CRAG benchmarks Retrieval-Augmented Generation (RAG) systems for reliable and creative question-answering in NLP.

By Shreepradha Hegde Published on June 11, 2024



Home > Technology > AI Shorts > Advancing Reliable Question Answering with the CRAG Benchmark

Advancing Reliable Question Answering with the CRAG Benchmark

By Mohammad Asjad - June 11, 2024

Table 5: Performance of straightforward RAG solutions. All numbers are in percentage. LLM-only solutions has up to 34% accuracy and straightforward RAG solutions has up to 44% accuracy.

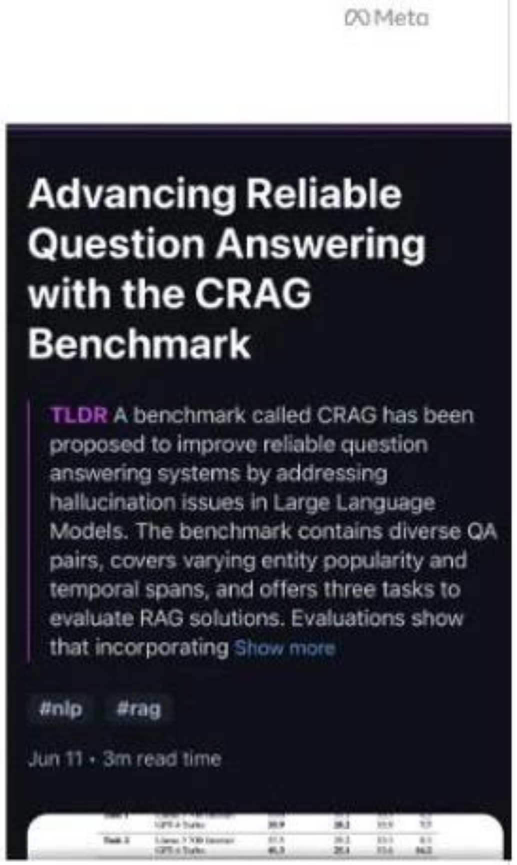
	Model	Accuracy	Hallucination	Missing	Score$_a$
LLM only	Llama 3 70B Instruct	32.3	28.9	38.8	3.4
	GPT-4 Turbo	**33.5**	**13.5**	53.0	**20.0**
Task 1	Llama 3 70B Instruct	35.6	31.1	33.3	4.5
	GPT-4 Turbo	**35.9**	**28.2**	35.9	**7.7**
Task 2	Llama 3 70B Instruct	37.5	29.2	33.3	8.3
	GPT-4 Turbo	**41.3**	**25.1**	33.6	**16.2**
Task 3	Llama 3 70B Instruct	40.6	31.6	27.8	9.1
	GPT-4 Turbo	**43.6**	**30.1**	26.3	**13.4**

Figure 2: LLM-only and Task 3 solution auto-eval scores (in percentage) across domains, dynamism, popularity, and question type.

https://arxiv.org/abs/2406.04744

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP), particularly in Question Answering (QA). However, hallucination

marktechpost.com



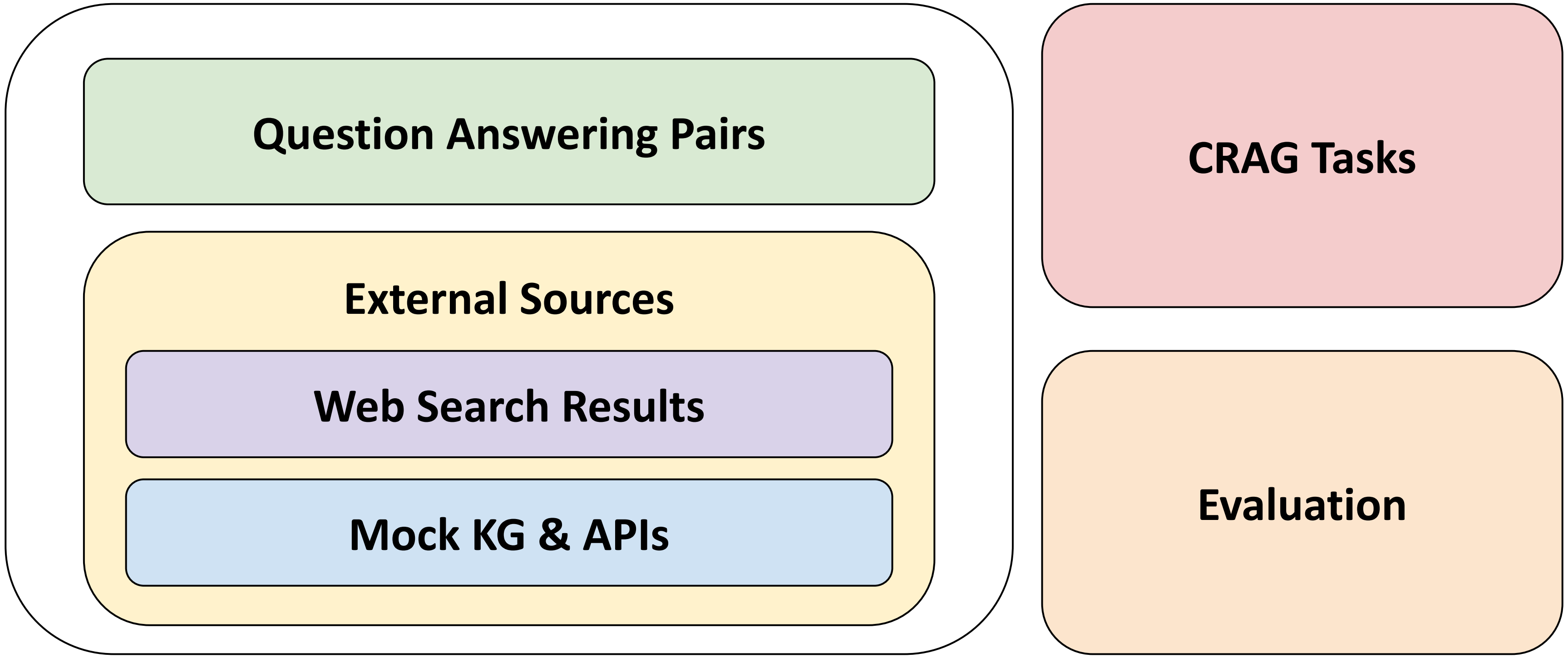
Advancing Reliable Question Answering with the CRAG Benchmark

TLDR A benchmark called CRAG has been proposed to improve reliable question answering systems by addressing hallucination issues in Large Language Models. The benchmark contains diverse QA pairs, covers varying entity popularity and temporal spans, and offers three tasks to evaluate RAG solutions. Evaluations show that incorporating Show more

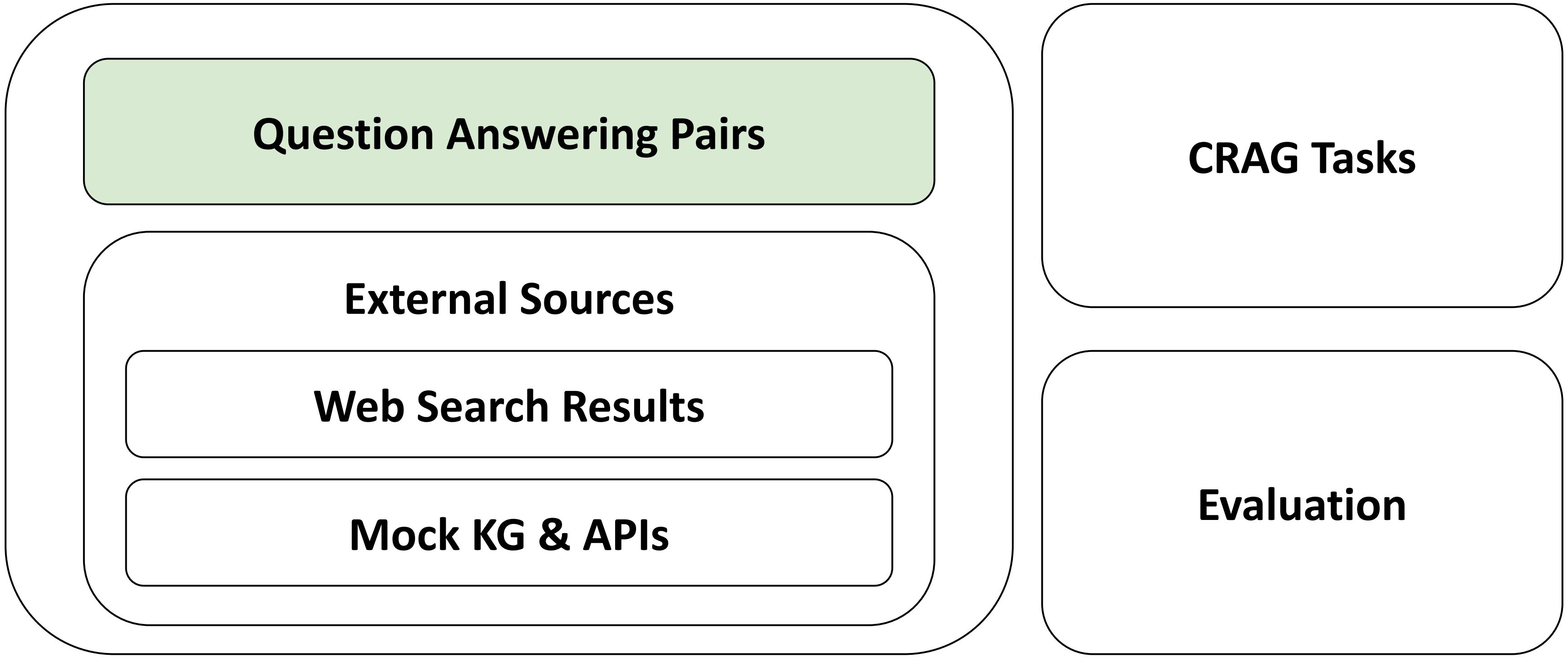
#nlp #rag

Jun 11 • 3m read time

CRAG Benchmark Overview



CRAG Benchmark Overview



Question Answering Pairs

External Sources

Web Search Results

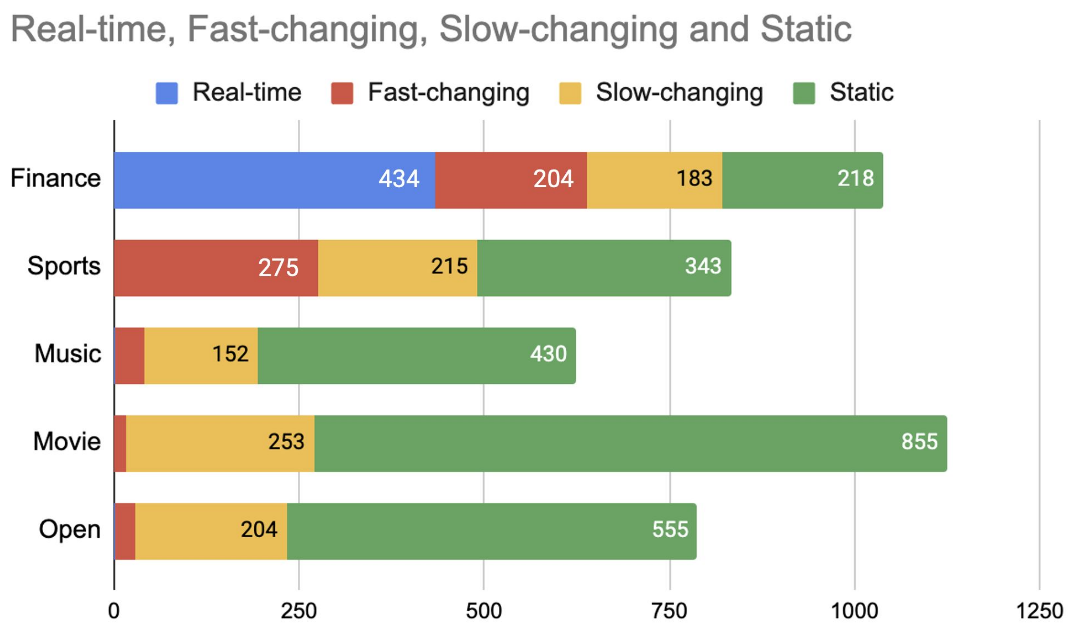
Mock KG & APIs

CRAG Tasks

Evaluation

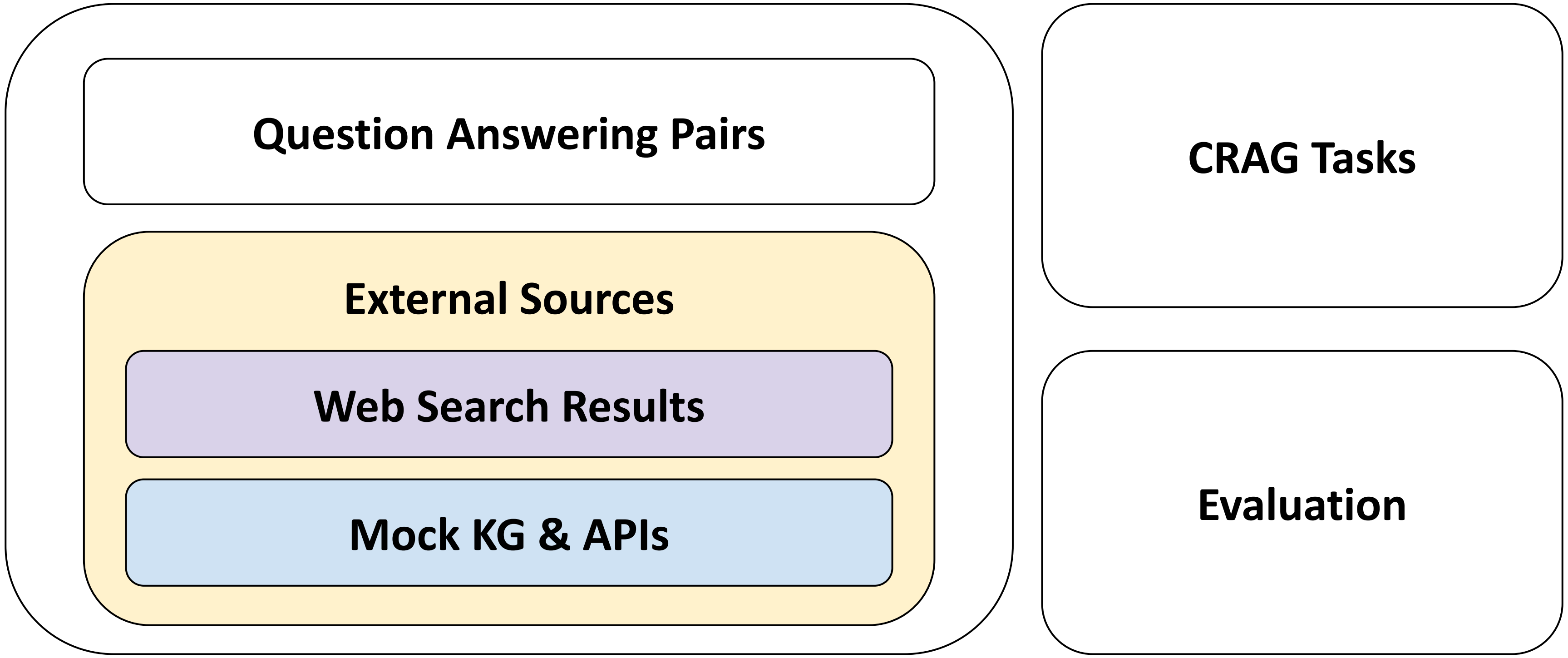
Question Answering Pairs

- 4400+ QA pairs from 5 domains (Finance, Sports, Music, Movie, Encyclopedia)
- Questions for static, slow-changing, fast-changing, and real-time information
- Questions for head, torso, and tail entities
- Simple-fact questions and complex questions



Total	Simple	Simple w. Cond	Set	Comparison	Aggregation	Multi-hop	Post-Processing	False Premise
4409	1205	689	403	546	489	382	180	525

CRAG Benchmark Overview



Question Answering Pairs

External Sources

Web Search Results

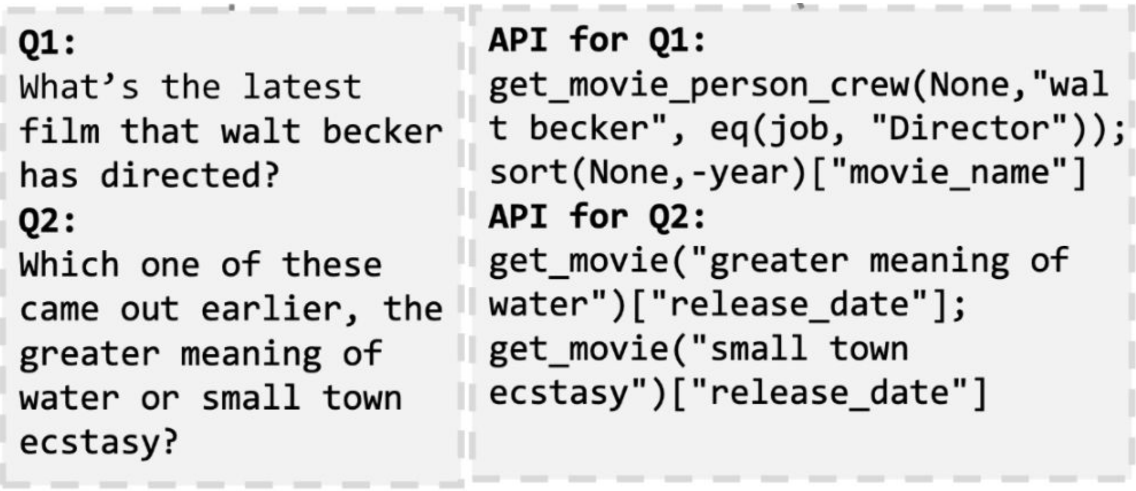
Mock KG & APIs

CRAG Tasks

Evaluation

External Sources

- Web Search Results: 50 webpages for each question from BraveAPI web search
- Mock KG & APIs
  - Mock KG: 2.6M entities
  - Mock APIs: 38 mock APIs



Q1:
What’s the latest film that walt becker has directed?
Q2:
Which one of these came out earlier, the greater meaning of water or small town ecstasy?

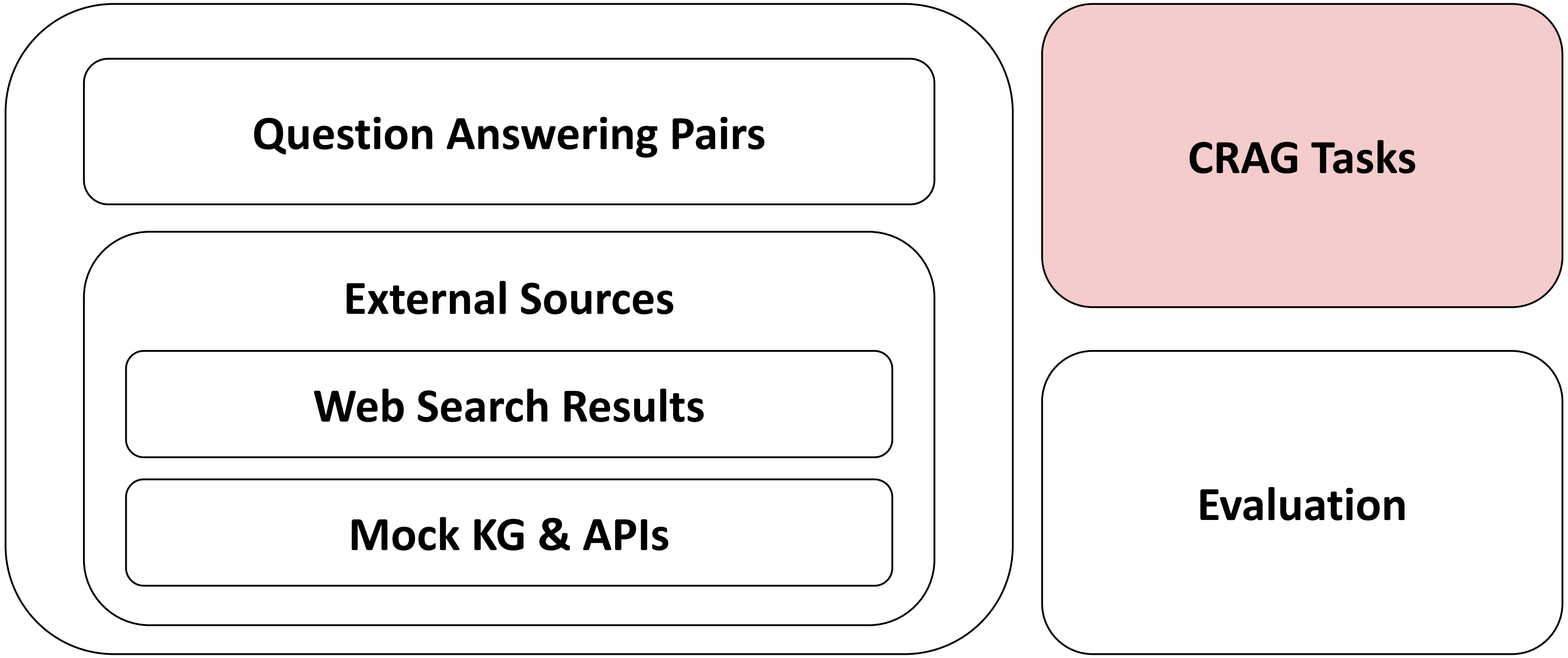
API for Q1:

```
get_movie_person_crew(None,"walt becker", eq(job, "Director"));
sort(None,-year)["movie_name"]
```

API for Q2:

```
get_movie("greater meaning of water")["release_date"];
get_movie("small town ecstasy")["release_date"]
```

CRAG Benchmark Overview



Question Answering Pairs

External Sources

Web Search Results

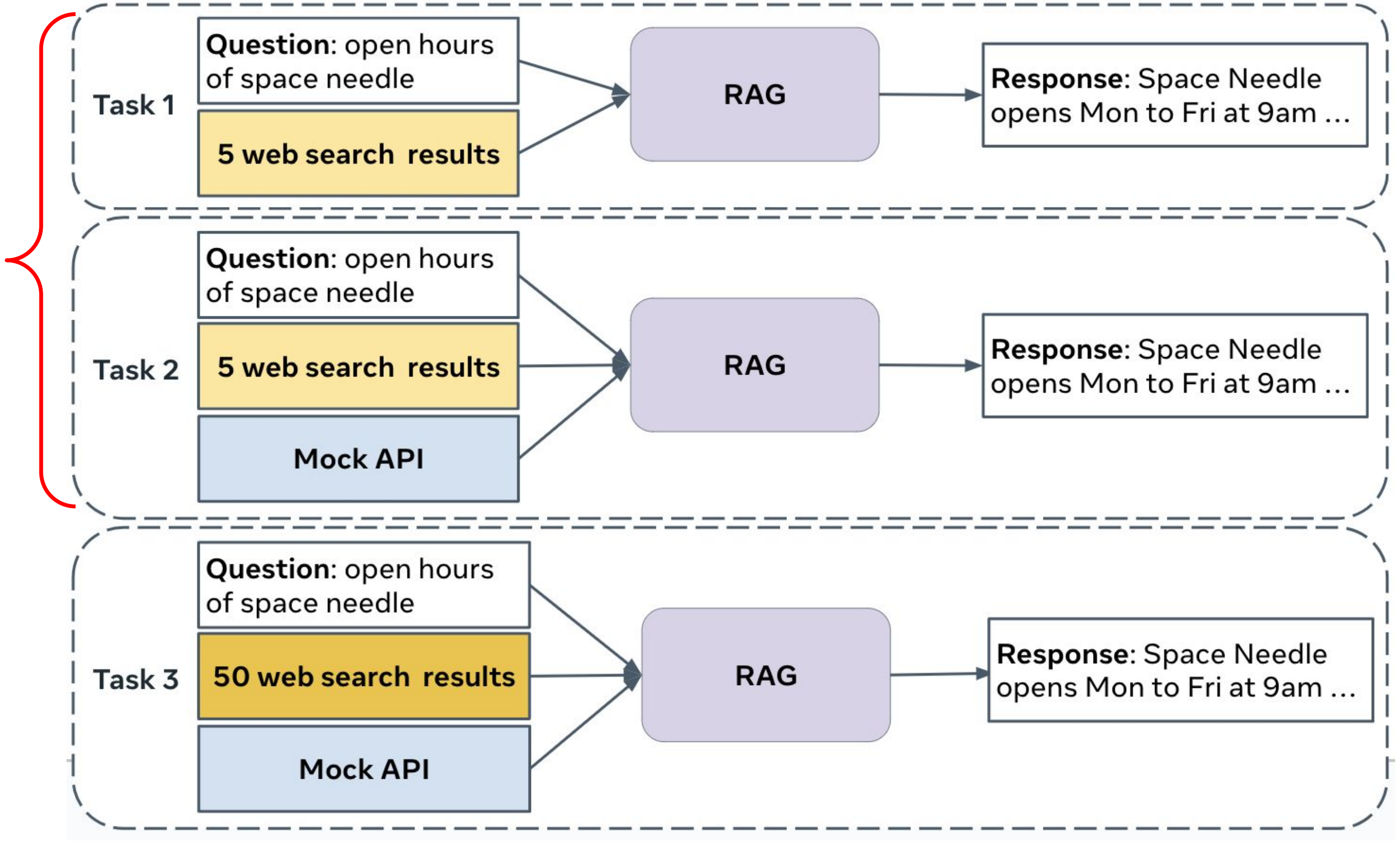
Mock KG & APIs

CRAG Tasks

Evaluation

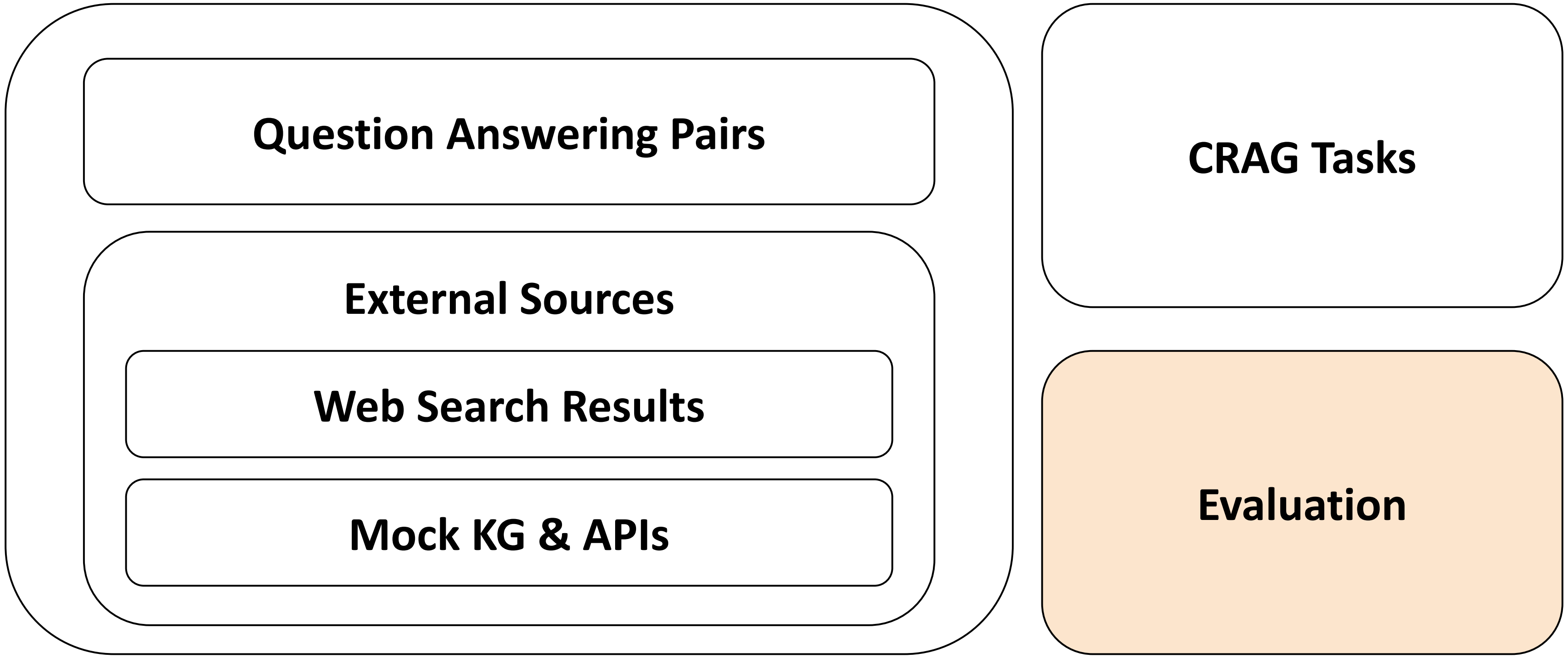
CRAG Tasks

Three tasks build up information gradually to test different capabilities of RAG systems.


Task 1
Question: open hours of space needle
5 web search results
RAG
Response: Space Needle opens Mon to Fri at 9am …
Task 2
Question: open hours of space needle
5 web search results
Mock API
RAG
Response: Space Needle opens Mon to Fri at 9am …
Task 3
Question: open hours of space needle
50 web search results
Mock API
RAG
Response: Space Needle opens Mon to Fri at 9am …

Today’s focus

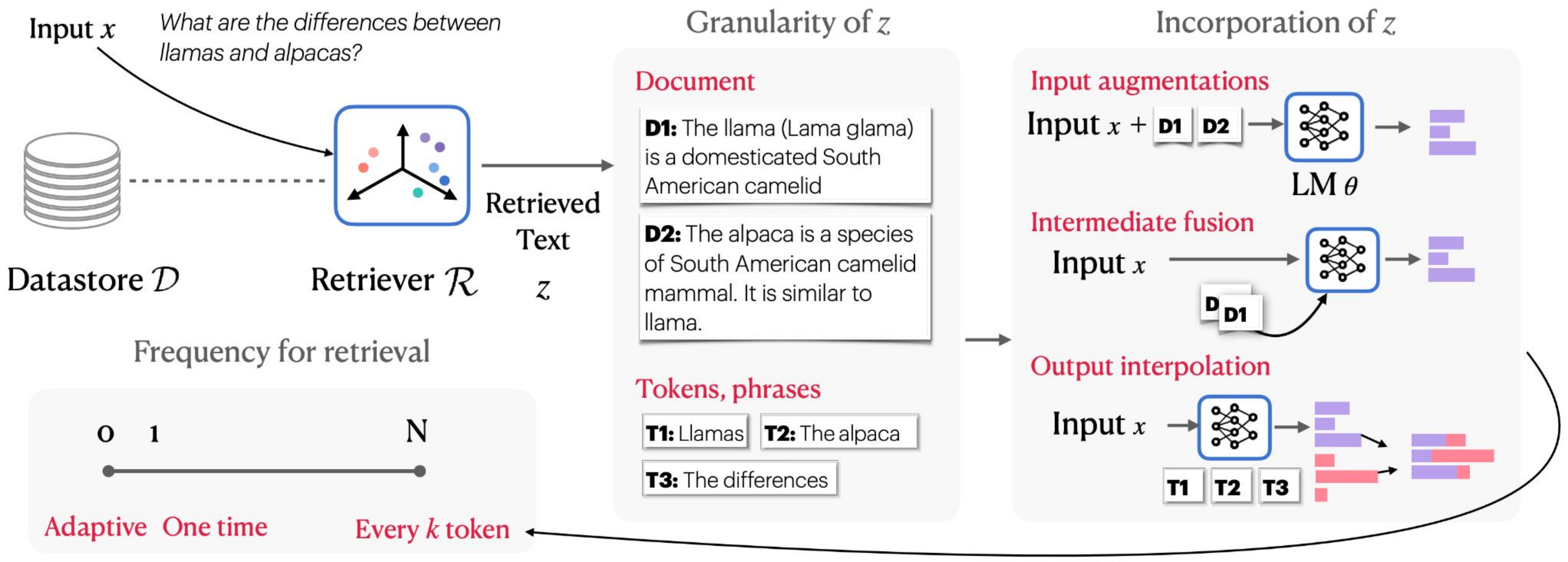
CRAG Benchmark Overview



Evaluation

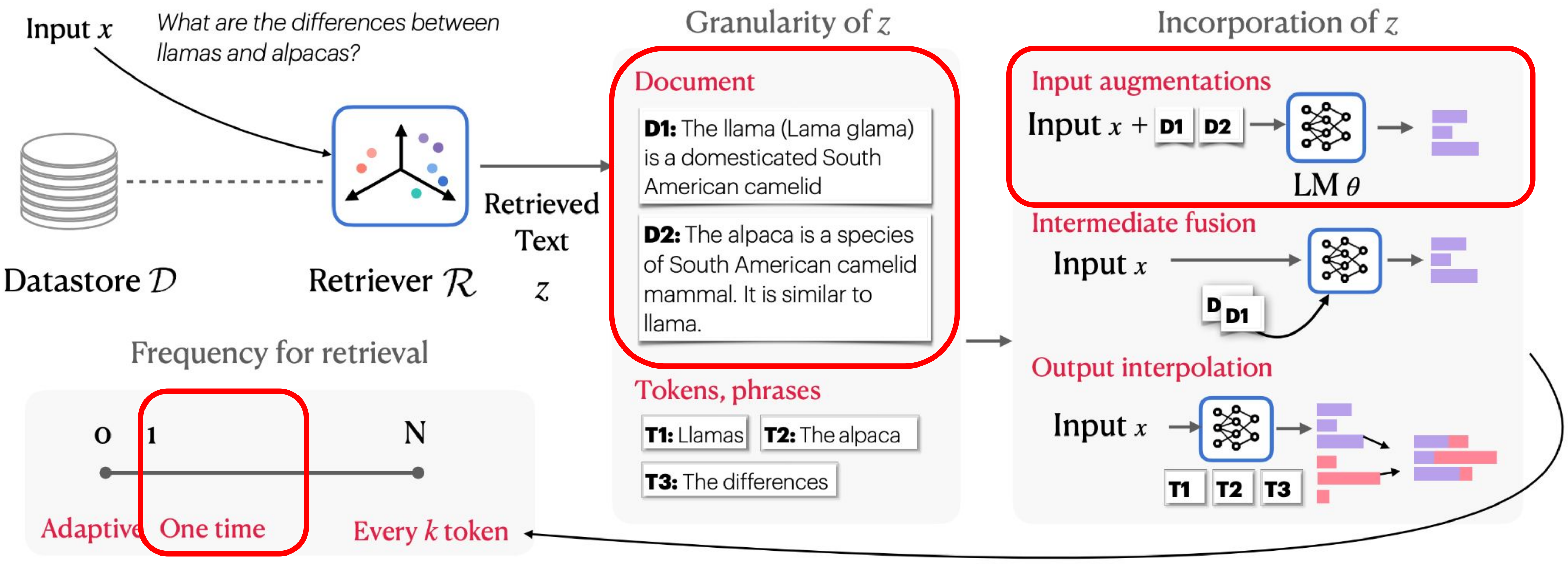
- Metrics: CRAG Score = Exact Accuracy + 0.5 * Accuracy - Hallucination rate
  - **Exact Accuracy:** The percentage of questions for which the generated answer exactly matches the ground truth answer.
  - **Accuracy:** The percentage of questions for which the generated answer is not exact but has the same meaning as the ground truth.
  - **Hallucination:** The percentage of questions for which an incorrect answer was generated.
  - **Missing:** The percentage of questions where the response was "I don’t know."
- Evaluation
  - Auto-eval (with GPT-4)
  - Manual-eval (with Human)

Diverse Architectures of RAG



Asai, Akari, et al. "Reliable, adaptable, and attributable language models with retrieval." arXiv preprint arXiv:2403.03187 (2024).

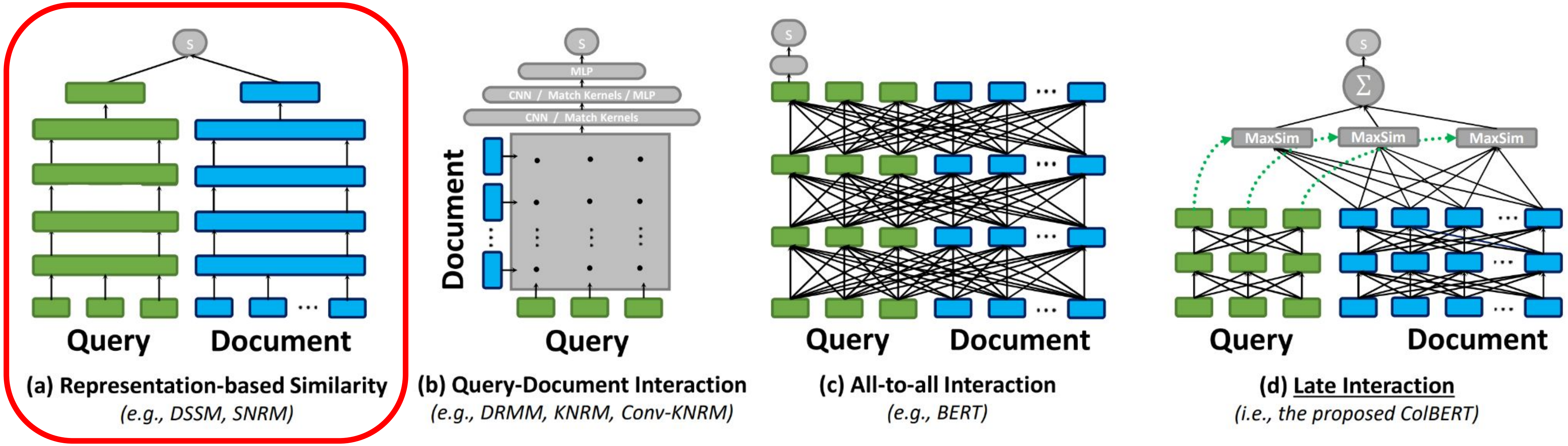
Baseline Approach



Asai, Akari, et al. "Reliable, adaptable, and attributable language models with retrieval." arXiv preprint arXiv:2403.03187 (2024).

Retrieval Module for Web Search Results

- Retriever performs representation-based similarity search.
- It retrieves top-k relevant sentences.



(a) Representation-based Similarity *(e.g., DSSM, SNRM)*

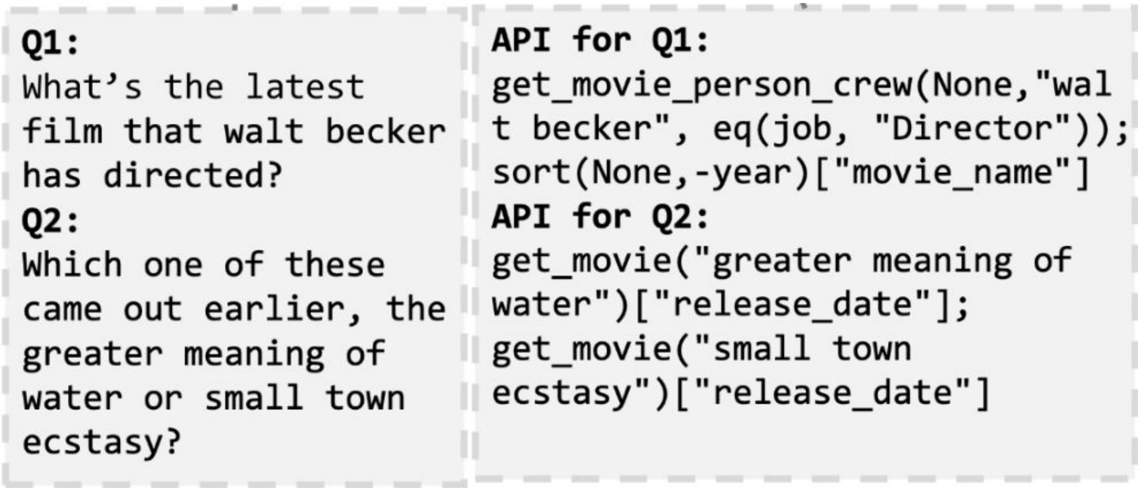
(b) Query-Document Interaction *(e.g., DRMM, KNRM, Conv-KNRM)*

(c) All-to-all Interaction *(e.g., BERT)*

(d) Late Interaction *(i.e., the proposed ColBERT)*

Mock APIs & Mock KG

- Task 2 provides mock APIs to query the provided mock knowledge graph (mock KG).



Q1:
What’s the latest film that walt becker has directed?
Q2:
Which one of these came out earlier, the greater meaning of water or small town ecstasy?

```
API for Q1:
get_movie_person_crew(None,"wal
t becker", eq(job, "Director"));
sort(None,-year)["movie_name"]
API for Q2:
get_movie("greater meaning of
water")["release_date"];
get_movie("small town
ecstasy")["release_date"]
```

- The mock KG, as a structured knowledge base, offers precise information; however, generating an accurate query is essential for retrieving correct answers.

Retrieval Module for Mock KG

- The knowledge graph retrieval module follows these three steps:

    a. The LLM generates the query domain and API arguments.

    b. Based on the generated query domain and API arguments, a decision tree is used to sequentially call the appropriate mock APIs.

    c. The results from these mock API calls are provided to the LLM along with retrieved results from web search results.

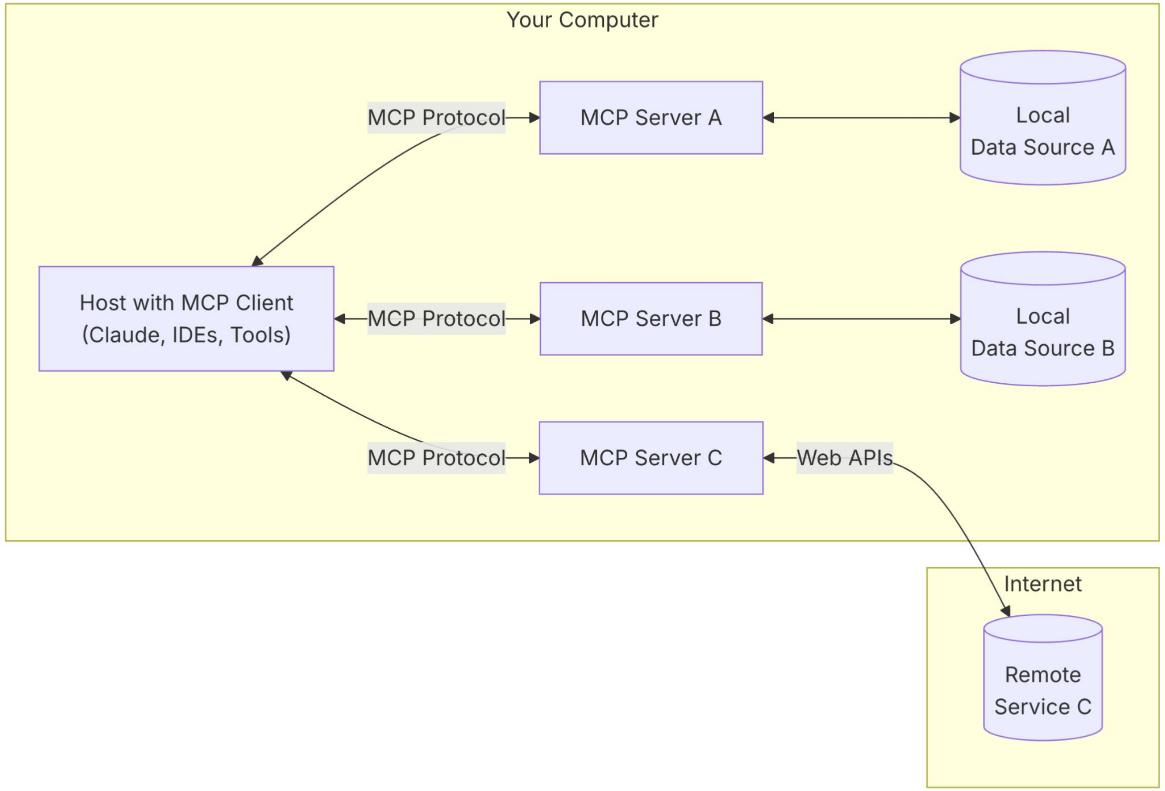
- We aim to go a step further by leveraging the Model Context Protocol (MCP) to upgrade the knowledge graph retrieval module.

Model Context Protocol

- Model Context Protocol (MCP) is an open protocol that standardizes how applications provide context to LLMs.
  - Think of MCP like a USB-C port for AI applications.
  - Just as USB-C provides a standardized way to connect your devices to various peripherals and accessories, MCP provides a standardized way to connect AI models to different data sources and tools.
- MCP helps you build agents and complex workflows on top of LLMs. LLMs frequently need to integrate with data and tools, and MCP provides:
  - A growing list of pre-built integrations that your LLM can directly plug into
  - The flexibility to switch between LLM providers and vendors
  - Best practices for securing your data within your infrastructure

General Architecture

At its core, MCP follows a client-server architecture where a host application can connect to multiple servers:



Appendix