

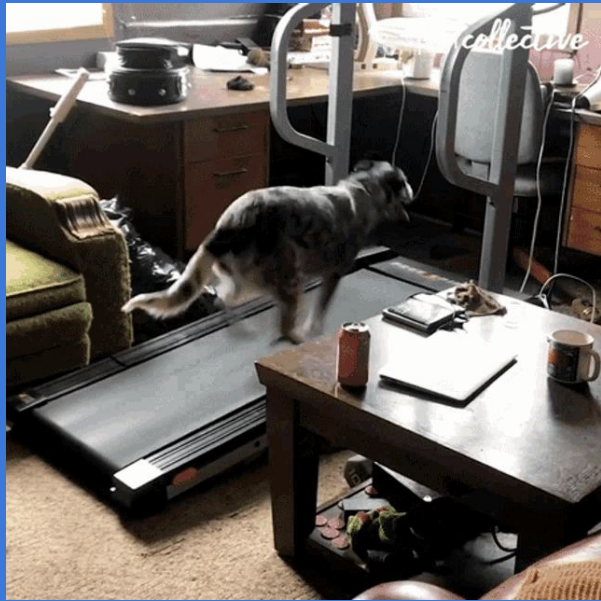
Providing Factual Information with Dual Neural Knowledge

Xin Luna Dong, Meta Reality Labs

8/2025

This talk does not represent the company's point of view

Question We Asked in 2024



*Is Correct, Relevant
Information Closer
or Farther Away?*

My Favorite Example

Google search results for "what college is the sister college of trinity college oxford". The search bar shows the query and the "AI Mode" tab is selected. The results include an "AI Overview" section with two images of Trinity College, Cambridge. A red thumbs-down icon is overlaid on the text: "Christ Church, Oxford is the sister college of Trinity College, Cambridge. These two colleges have a long-standing relationship and are known to participate in an annual tradition involving student exchanges, formal dinners, and other activities."

Google search results for "what college is the sister college of trinity college oxford". The search bar shows the query and the "AI Mode" tab is selected. The results include a "Quick results from the web" section with two snippets: "List of Oxbridge sister colleges - Wikipedia" and "Trinity College, Cambridge - Wikipedia". A green thumbs-up icon is overlaid on the text: "Trinity College, Oxford has a formal sister college relationship with Churchill College, Cambridge." Below this, a red thumbs-down icon is overlaid on the text: "While Trinity College, Cambridge is a historically significant institution, it is actually Christ Church, Oxford that shares a sister college relationship with Trinity College, Cambridge. This connection is rooted in their shared founding year of 1546 by King Henry VIII."

My Favorite Example (cont')

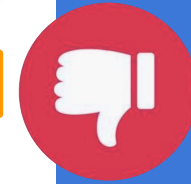
ChatGPT 4o ▾

↑ Share ...

what college is the sister college of trinity college oxford

The sister college of **Trinity College, Oxford** is **Trinity College, Cambridge**.

This pairing follows the tradition of linking colleges between Oxford and Cambridge, often with historical or thematic similarities, including name and foundation history.



search the web and answer the question

The official sister college of **Trinity College, Oxford** is **Churchill College, Cambridge**

[en.wikipedia.org](#) +13 . This pairing is part of Oxbridge's long-standing tradition of inter-university "twinning" arrangements, often allowing for reciprocal college stays, dinners, and shared events. [history.queens.cam.ac.uk](#) [academickids.com](#)



Another Fun Example—Myself

ChatGPT 4o ▾

↑ Share

tell me about luna xin dong

Xin Luna Dong—often publishing as “XinLuna Dong”—is a leading researcher in **knowledge graphs, data integration**, and their intersection with **machine learning and intelligent assistants**. Here's an overview:



Publications & Thought Leadership

- Authored 100+ papers, is **h-index 59**, with nearly **16k citations** Google Scholar
- Wrote the book *Big Data Integration* and co-edited important works in data quality and fusion mlconf.com +4
- Regular keynote and keynote-presenter at VLDB, SIGMOD, KDD, WSDM, and MLconf mlconf.com +2



In Her Own Words

- From an **ACM Bytecast**, she emphasizes making info access effortless for **non-technical users and small businesses**, and leveraging **RAG and CRAG benchmarks** learning.acm.org



Harder Aggregation Example

ChatGPT 4o ▾

↑ Share

tell me about ballet shows in nyc in the upcoming month

Here are the major ballet highlights in New York City over the next month (late July through August 2025):

📅 Upcoming Ballet Events Schedule

Date Range	Event	Venue	Note
Now–mid August	American Ballet Theatre (Swan Lake, etc.)	Metropolitan Opera House	Tickets from ~\$48 New York Theatre Guide + New York Post
Jul 29–Aug 2	BAAND Together Dance Festival	Koch Theater	Multi-company festival, tickets ~\$35 davidkochtheater.com Dance NYC
August (exact TBD)	Hong Kong Ballet – <i>The Butterfly Lovers</i>	Lincoln Center (likely Koch)	Dates to be confirmed TDF
Aug 4	NYCB single tickets release for Fall	David H. Koch Theater	Performances from Sep 16 onward nycballet.com nycballet.com



tell me about white snake at lincoln center

Here's the full lowdown on **Lady White Snake** (白蛇), the Chinese dance-ballet drama at Lincoln Center:

🤖 What It Is

A large-scale contemporary dance-drama by **Shanghai Grand Theatre**, retelling the classic Chinese folk tale *The Legend of the White Snake*. Combining Western ballet, Chinese classical and modern dance—with martial arts and multimedia design—it's directed by Zhou Ke and choreographed by Wang Peixian. Artistic vision is led by former San Francisco Ballet principal **Tan YuanYuan**, featuring an original score by Xu Zhong

[davidkochtheater.com](#) +15

Missing Answer



📅 Dates & Venue

David H. Koch Theater, Lincoln Center

- **Saturday, July 26:** 1:30 PM (Cast Moon) & 7:30 PM (Cast Luck)
- **Sunday, July 27:** 1:30 PM (Cast Moon) [Shore Fire Media +4](#)

This marks its **New York premiere**, part of *Summer for the City – Shanghai Day*

[Lincoln Center](#) +10

Another Aggregation Example

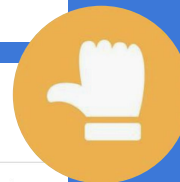
ChatGPT 4o Share

how many CS conferences have been hosted in Toronto in the past 40 years? What are they?

I couldn't locate a definitive list of *all* computer science (CS) conferences held in Toronto over the past 40 years—but here are the major, recurring ones I did identify:

Summary Table

Conference	First Held in Toronto	Recurrence	Focus Area
CASCON	1991	Annual	Applied CS, industry-academia
ACM KDD	2025	Rotating-Annual?	Data mining / ML
CCCG	2025	Rotating?	Computational geometry
WADS	2025	Rotating?	Algorithms & data structures
CP (with XCSP3 competition)	2023	Rotating?	Constraint programming
Various ICB/ICSPS-style general CS conferences	2025	Rotating global	Cybersecurity, AI, software eng.



Harder Reasoning Example

ChatGPT 4o ▾ ↑ Share ...

For papers regarding improving conversation truthfulness, which venues shall I submit to?

For research on **improving conversation truthfulness**, especially in the context of dialogue systems, LLMs, or conversational AI, consider submitting to top-tier **NLP, AI, or ML** venues that actively publish in these areas. Here's a breakdown of the most relevant venues:

Decision Flow

- **Theory-heavy or model-focused truthfulness?** → NeurIPS / ICLR / ICML
- **Empirical and language-specific analysis?** → EMNLP / ACL / NAACL
- **Dialogue system-specific or conversational agents?** → SIGDIAL / INLG
- **Early or exploratory work?** → Workshops at ACL/EMNLP

can I submit RAG papers to KDD?

Yes, you **can submit Retrieval-Augmented Generation (RAG) papers to KDD**, but with **conditions**—your paper must emphasize **data mining, knowledge retrieval, or large-scale ML system aspects**, not just language modeling.



Where Are We in This Journey? —A Quantitative Answer

Round 1: Completed Round 1b: Completed Phase 2: Completed

#rag #llm #knowledge_retrieval #question_answering_systems #generative_ai #knowledge_graph

∞ Meta KDD Cup 2024

CRAG: Comprehensive RAG Benchmark

🏆 31,500

By  Meta

👁 169.7k 👤 3050 👥 384 🚀 6323

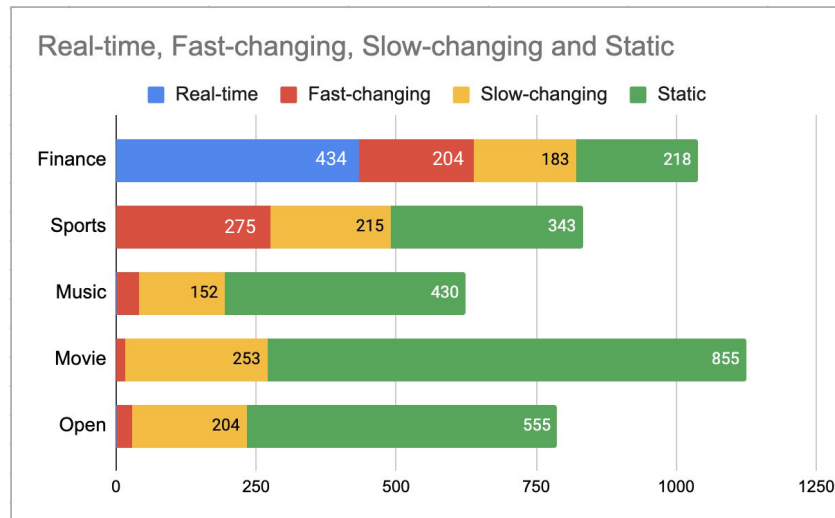
❤️ 94 Share



The background features a light blue gradient with various icons floating around, including people, documents, and speech bubbles. On the right side, there are three wooden bookshelves filled with colorful, geometric books. Below the shelves is a small illustration of a library building with a sign that says 'LIBRARY'.

Rich and Insightful Question-Answer Set

- 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

Accessible Retrieval Content

- 220K webpages: 50 webpages for each question from BraveAPI web search
- Mock KG: 2.6M entities, 30:1 signal-to-noise ratio
- Mock APIs: 38 mock APIs

Reliable Tasks and Evaluation

- Task 1: Answer generation over top-5 web search results—**Answer Summarization**
- Task 2: + Mock-KG Search API—**Structured Search, Answer Selection**
- Task 3: + 50 web search results—**Search Ranking**

LLM-only vs. Straightforward RAGs

Domain	Solution	Accuracy	Hallucination	Missing	Factuality =Accuracy-Halluci..
Llama-3 (70B)	LLM Only	32%	29%	39%	3%
	Straightforward Task 3	41%	32%	28%	9%
GPT4 Turbo	LLM Only	34%	14%	53%	20%
	Straightforward Task 1	36%	28%	36%	8%
	Straightforward Task 2	41%	25%	34%	16%
	Straightforward Task 3	44%	30%	26%	14%

RAG may improve
accuracy

KG w. crisper retrieval
results helps

RAG may also bring
more hallucinations

KDD Cups Winning Solutions



Big jump!!

Tasks	Auto-eval			Manual-eval
	GPT-4	Llama 3	KDDCup Winning team	
Task 1	8%	5%	29%	30%
Task 2	16%	8%	30%	32%
Task 3	13%	9%	31%	36%

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, et al. CRAG–Comprehensive RAG Benchmark. NeurIPS, 2024.

Xiao Yang, Kai Sun, et al. KDD Cup CRAG competition: Systems, Findings and Learnings. IEEE Data Engineering Bulletin, 2024.

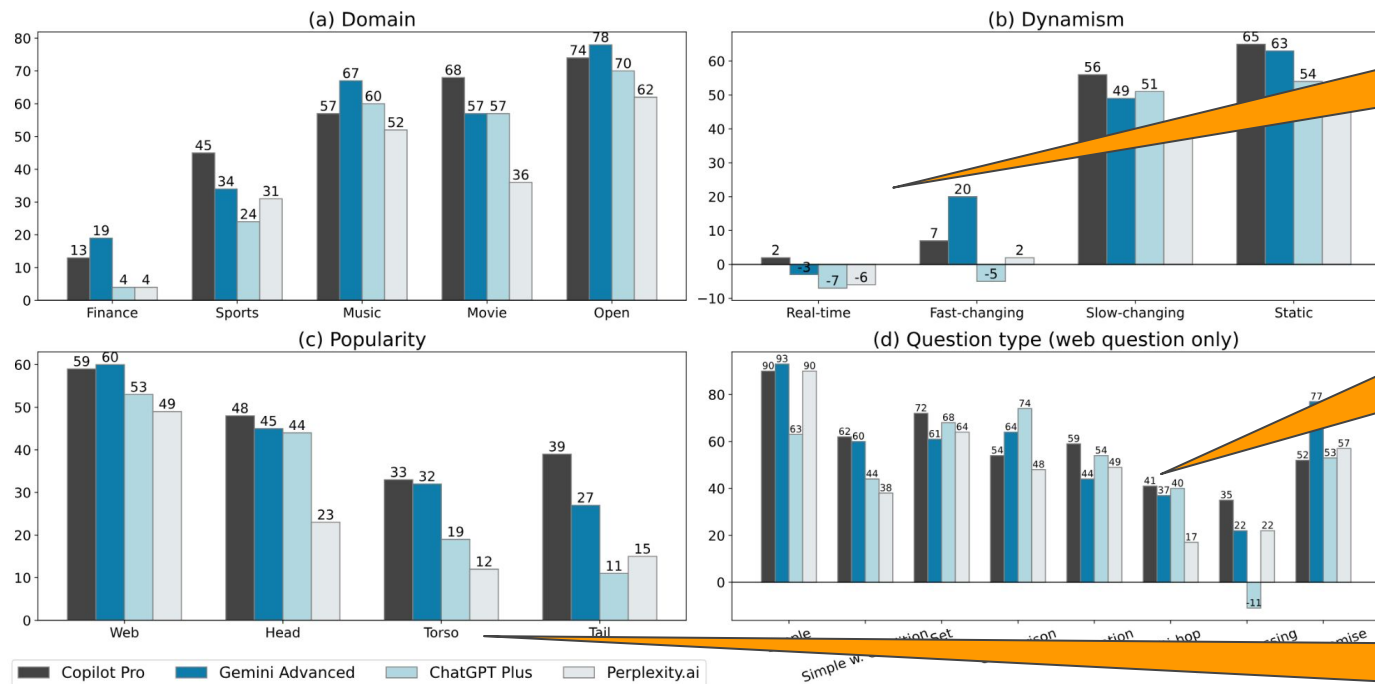
State-of-the-Art Industry Solutions (2024)

System	Perfect	Acceptable	Incorrect	Missing	Factuality	Latency (s)
Copilot Pro	63%	12%	18%	8%	50.4%	11.6
Gemini Advanced	61%	10%	17%	13%	49.5%	5.2
ChatGPT Plus (4o)	60%	13%	25%	2%	42%	6.2
Meta Wearables	53%	10%	16%	22%	41%	3.4
Perplexity.ai	56%	9%	25%	10%	35%	4.6
Notes: 1. Manual annotations. 2. Retrieval by the SOTA solutions themselves.						

Perfect < 63%. Still a big gap to fill

Different latency-quality tradeoffs

SOTA Industry Solutions on Diff Dimensions



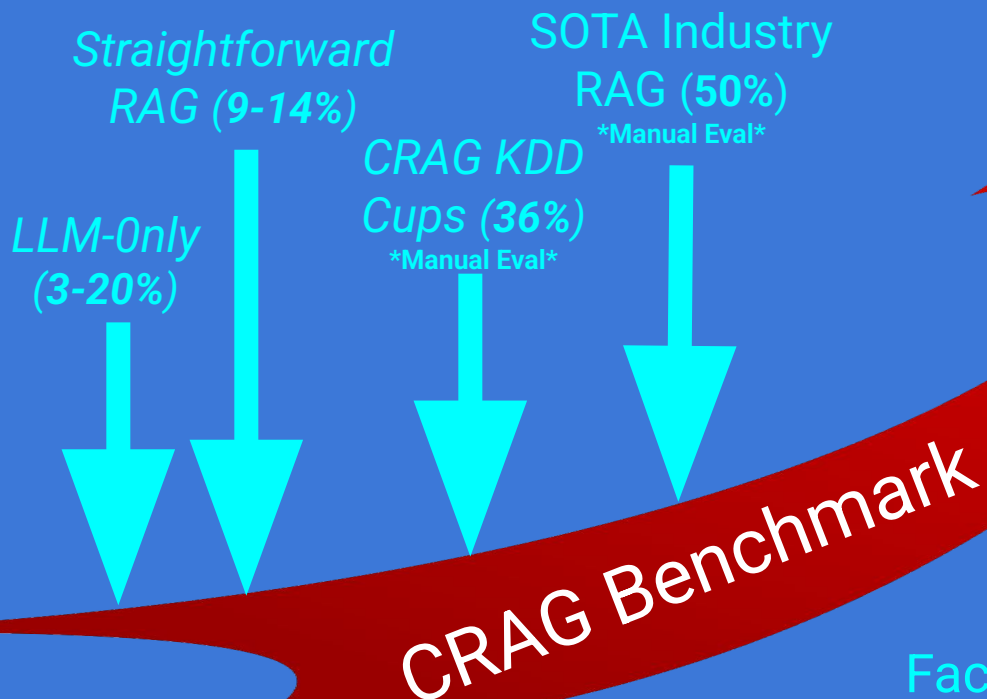
Improvements needed for **fast-changing** facts

Improvements needed for **complex** questions

Improvements needed for **torso/tail** questions

Figure 3: SOTA systems human-eval scores (in percentage) across different dimensions.

Where Are We in This Journey? —A Quantified Answer

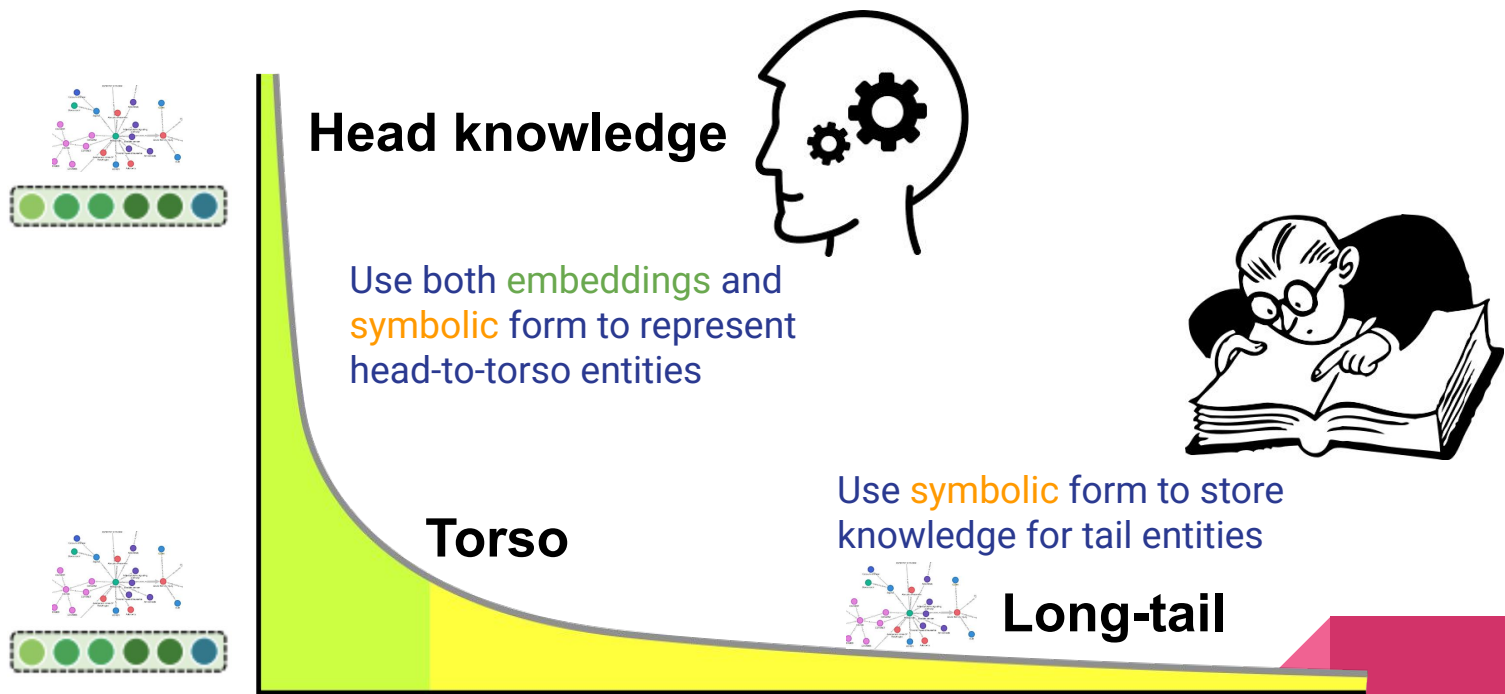


Factuality = Corr% - Incorr%



Dual Neural Knowledge As a Path to Factuality

Dual Neural Knowledge



Dual Neural Knowledge

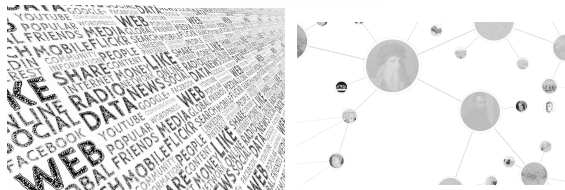


Offline
Integration
through
Pre-training

Internalized head-to-torso
knowledge *within* LLMs

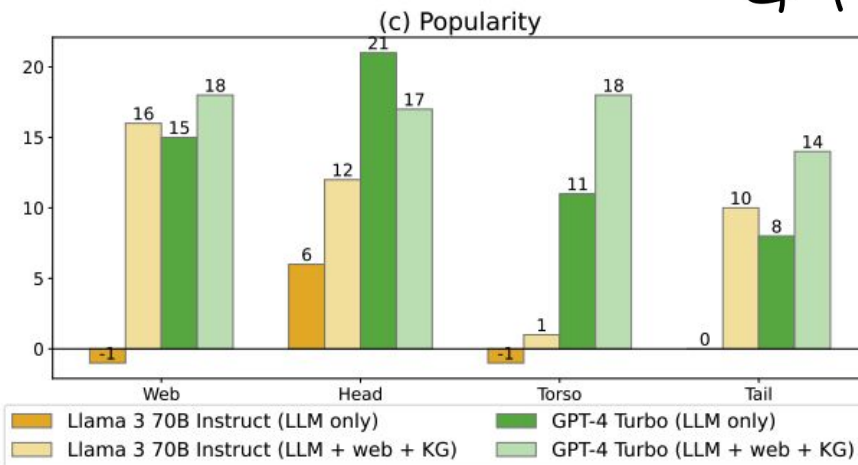


Runtime
Integration
through RAG



Symbolic tail or dynamic
knowledge as is *in the wild*

LLM-only (Closed-book)



Low factuality EVEN for
head entities

Dual Neural Knowledge

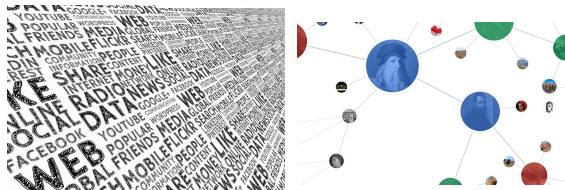


Offline
Integration
through
Pre-training

Internalized head-to-torso
knowledge *within LLMs*



Runtime
Integration
through RAG

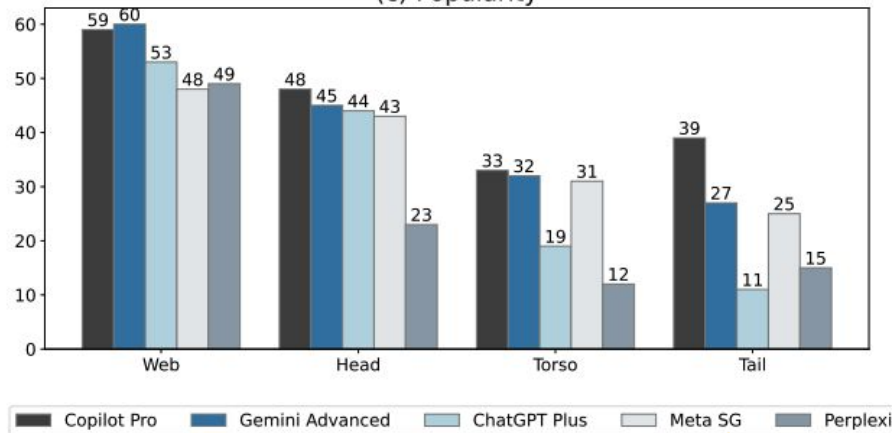


Symbolic tail or dynamic
knowledge as is *in the wild*

RAG (Open-book)



(c) Popularity



Still **low factuality** EVEN with
external knowledge

Dual Neural Knowledge

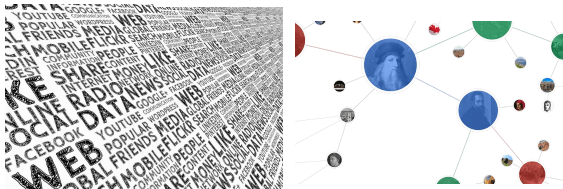


Offline
Integration
through
Pre-training

Internalized head-to-torso
knowledge *within LLMs*



Runtime
Integration
through RAG



Symbolic tail or dynamic
knowledge as is *in the wild*

Question 3. How to internalize more factual knowledge into LLMs?

**Pre-train
/
Post-train**

Question 1. How to know when to rely on what?

Post-train

Question 2. How to leverage external data in RAG more effectively?

Post-train



1. Choosing Between Internalized Neural Knowl. vs. Symbolic Knowledge

How to Know When to Rely on What?

All of the following questions have a never-changing unique answer. Some are head (popular), others are tail (less popular). Decide if you can reliably answer the questions without resorting to external knowledge sources. Return yes for those you are confident to answer, and no for others. Here are the questions.

1. How tall is the Statue of Liberty
2. How many bones are in a human adult's body
3. When was George Washington. (US president born)
4. Who's the author of the book "Seed of Evil"



Metamate · GPT-4

2 steps >

1. Yes, the height of the Statue of Liberty is a well-known fact.
2. Yes, the number of bones in a human adult's body is a standard piece of knowledge.
3. Yes, George Washington's birth date is a widely known historical fact.
4. No, the author of the book "Seed of Evil" is not a commonly known fact.

How to Know When to Rely on What?

- Does an LLM **know what it knows**?
- Can we teach LLMs to **refrain from hallucinating**?
- What is the optimal strategy for **RAG triggering**?

RQ1. Does an LLM Know What It Knows?

Experiments on Three Factuality Benchmarks

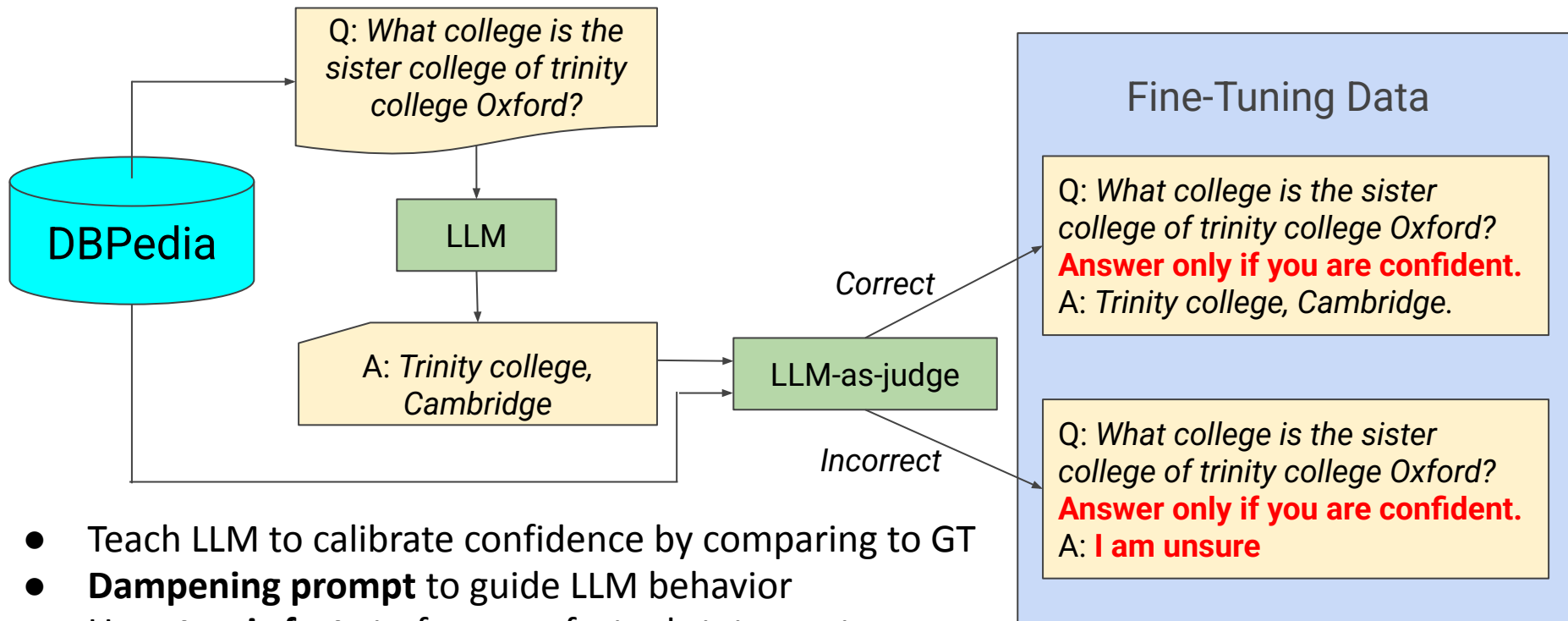


Harder for nuanced facts
about popular entities

*Smaller models are even
more confident*

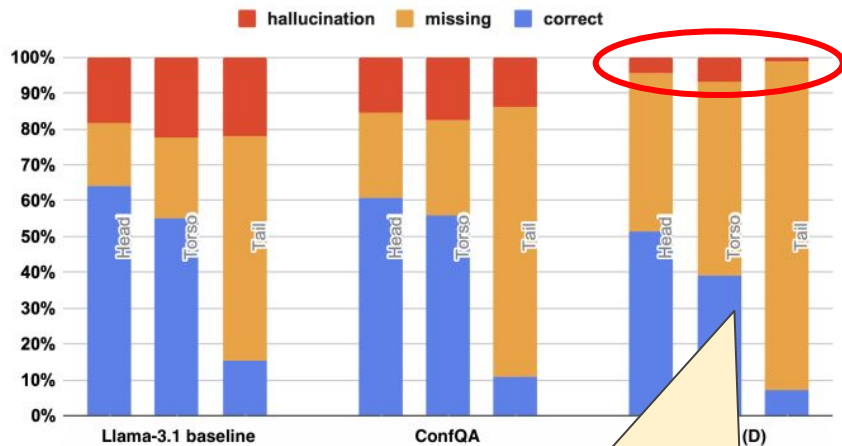
LLMs tend to be
over-confident

RQ2. Can We Teach LLMs to Refrain from Hallucinating?



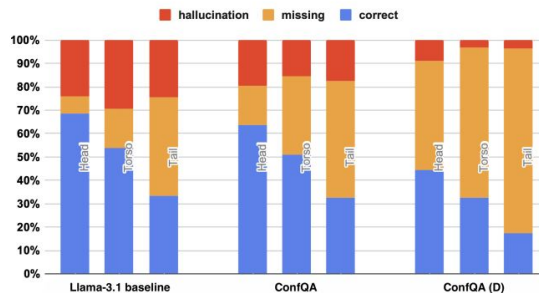
- Teach LLM to calibrate confidence by comparing to GT
- **Dampening prompt** to guide LLM behavior
- Use **atomic facts** to focus on factual statements

RQ2. Can We Teach LLMs to Refrain from Hallucinating?

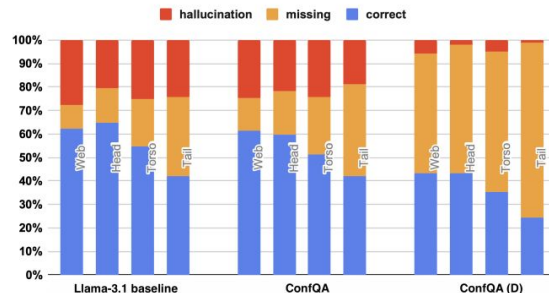


(b) IMDB

1. Dampener is critical in training. At inference, with the dampener prompt, hallucination reduced to below 5%

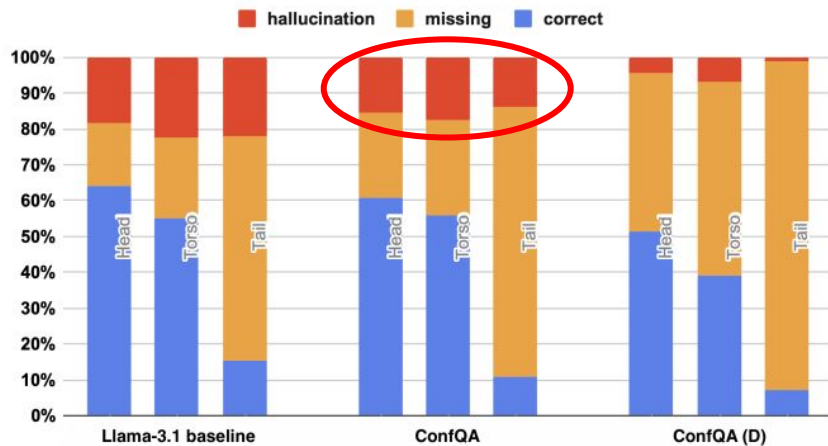


(a) DBPedia



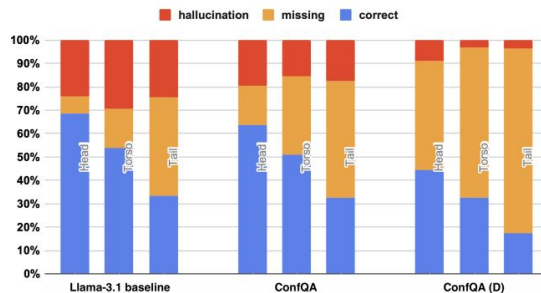
(c) CRAG

RQ2. Can We Teach LLMs to Refrain from Hallucinating?

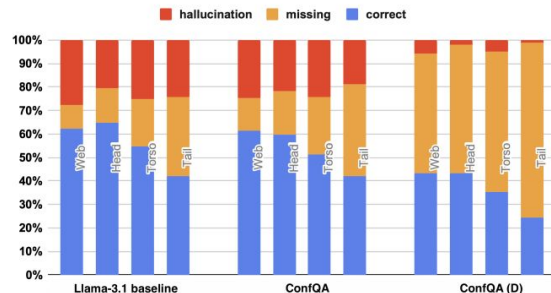


IMDB

2. Without the dampener prompt in inference, similar correct% and mild hallucination% reduction

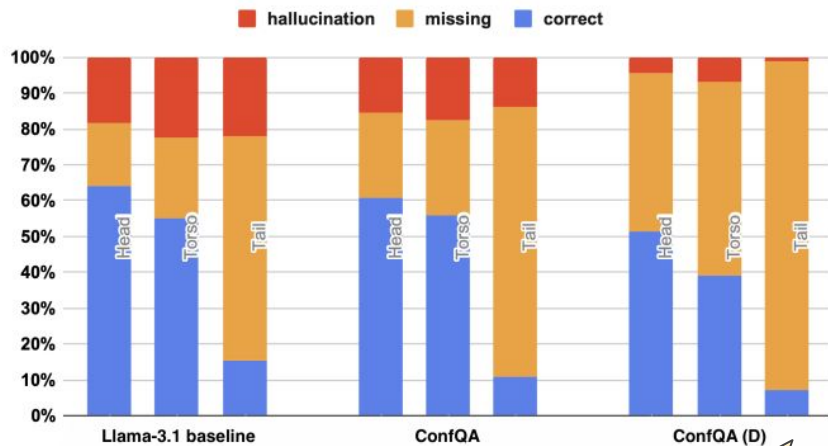


(a) DBPedia



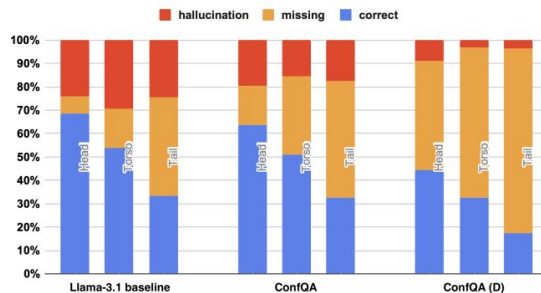
(c) CRAG

RQ2. Can We Teach LLMs to Refrain from Hallucinating?

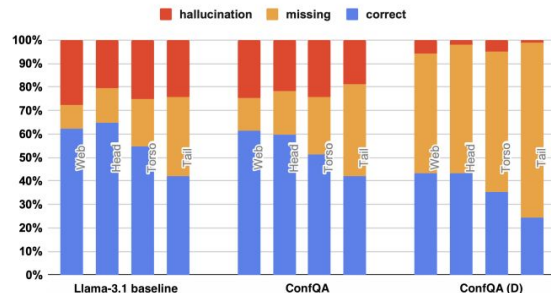


(b) IMDB

3. Suppress more for long-tail facts.

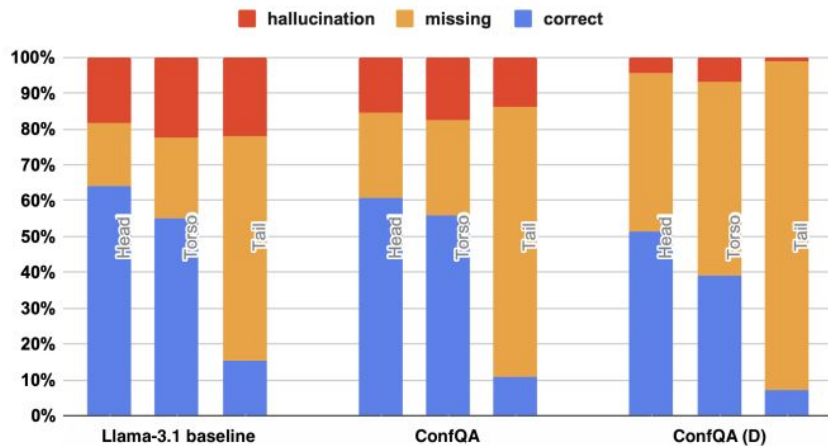


(a) DBPedia



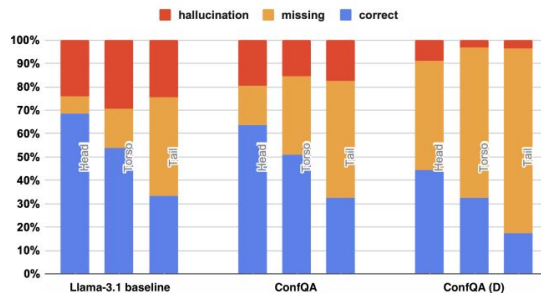
(c) CRAG

RQ2. Can We Teach LLMs to Refrain from Hallucinating?

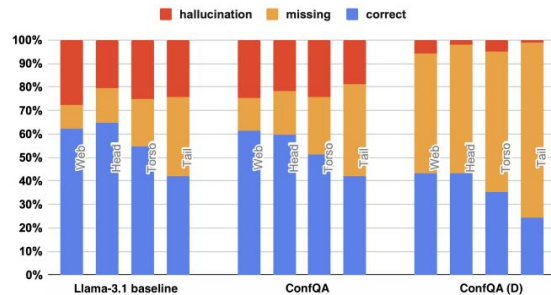


(b) IMDB

4. Transfer well from DBPedia to IMDB. However, training data from MMLU (w. non-factual questions) significantly lower accuracy



(a) DBPedia



(c) CRAG

RQ2. Can We Teach LLMs to Refrain from Hallucinating?

Model	Long Fact				Alpaca Fact				Biography			
	Prec	Rec	F1	Miss	Prec	Rec	F1	Miss	Prec	Rec	F1	Miss
Llama3.1	64.5	65.4	64.3	0	62.3	71.0	63.8	0	35.4	40.3	37.1	0
RAG (Llama3.1) (Yu et al., 2022)	71.7	74.6	72.7	0	65.8	74.3	66.0	0	44.9	48.1	43.8	0
ConfQA	67.0	67.7	66.7	0.8	62.2	71.1	63.8	0.4	42.0	46.5	42.6	12.6

Table 4 ConfQA improves precision and recall for long-form answer generation.

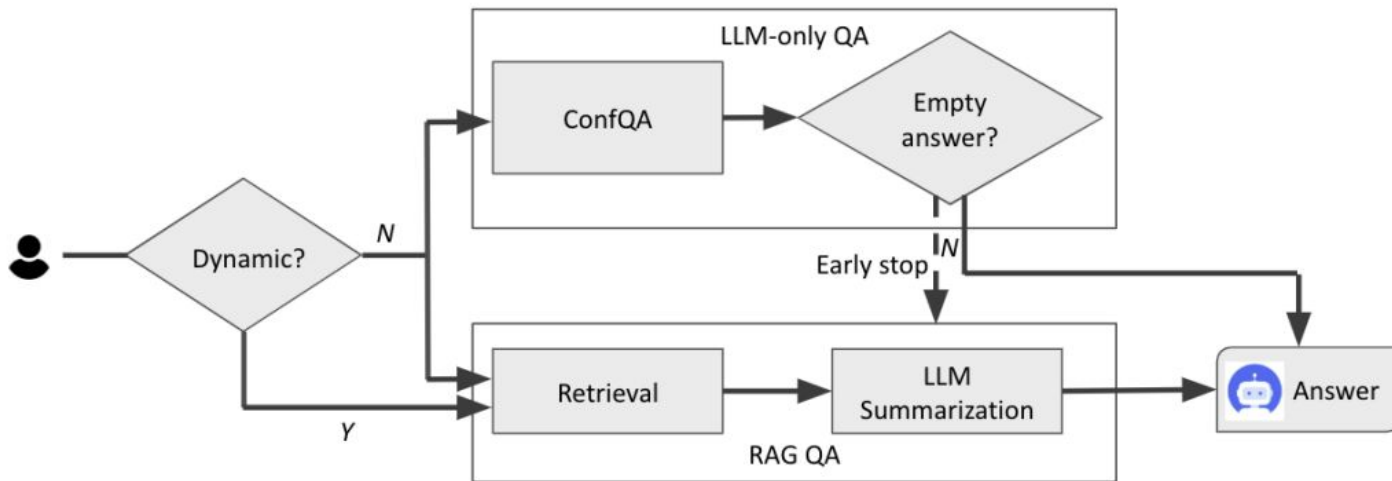
5. Transfer well to long-form answers w. higher quality, and no regression on other tasks

6. Feeding GT only will teach LLMs to hallucinate

Model	MMLU (5-shot)	MMLU-Pro
Llama3.1	82.7	66.3
ConfQA	82.8	65.4

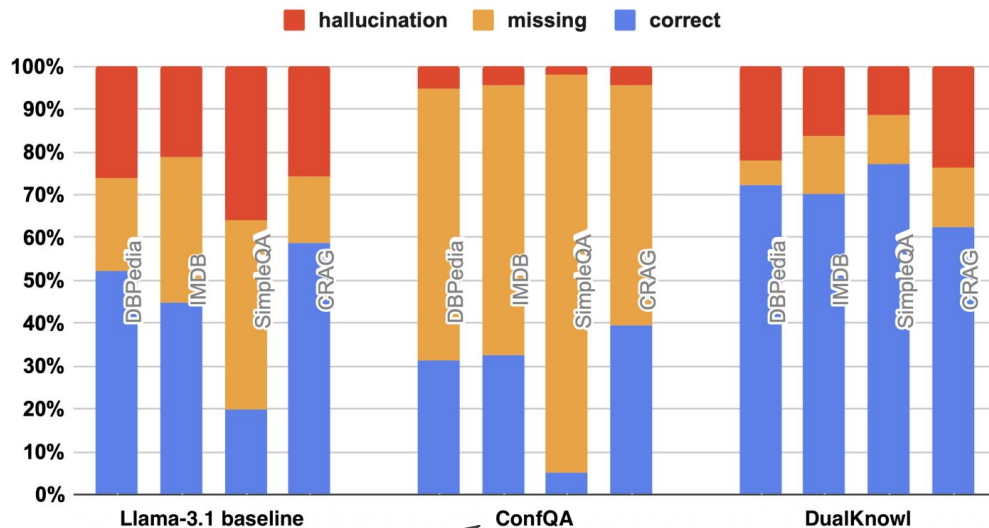
Table 5 ConfQA does not regress on MMLU.

RQ3. What Is Optimal Strategy for RAG Triggering?



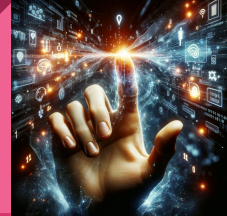
- Trigger LLM for **dynamic** questions and **unsure** answers
- **Parallel** invocations RAG and ConfQA to cut latency

RQ3. What Is Optimal Strategy for RAG Triggering?



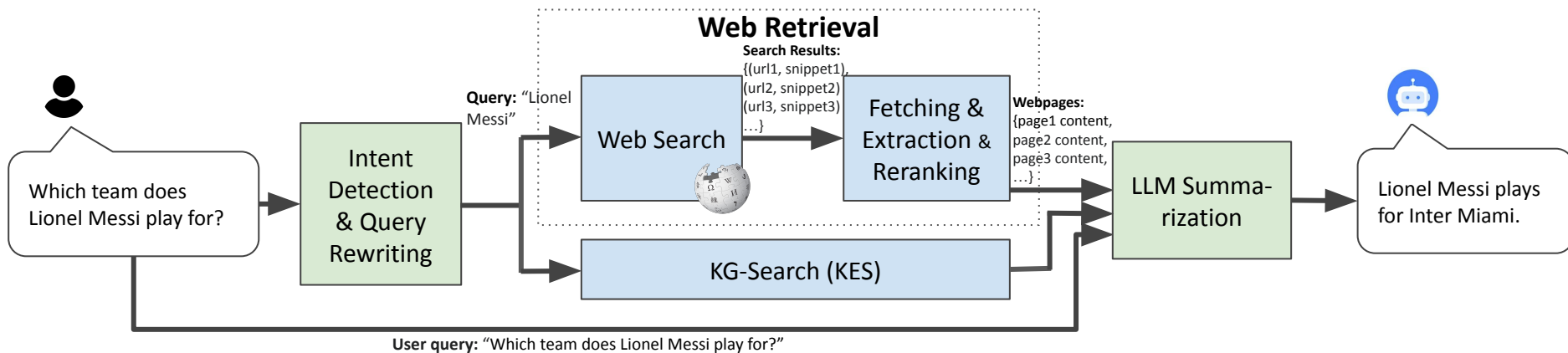
ConfQA reduces hallucination to below 5% on various benchmarks

RAG based on ConfQA increases accuracy by 45%, w. reasonable latency



2. Effectively Leveraging Symbolic Knowl. for RAG

Overview of Federated RAG



RAG Accuracy = Retrieval Recall x Summarization Accuracy

Tradeoff Between Retrieval Recall & Summarization Accuracy

Data sources	Recall	Accuracy
KB	0.952	0.908
Wiki Infobox	0.980	0.872
Wiki Page	0.990	0.845

Increased
recall

Decreased accuracy

✎ 66 languages
[Read](#) [View source](#) [View history](#) [Tools](#)

From Wikipedia, the free encyclopedia

"Facebook, Inc." redirects here. For the social media, see Facebook.

Meta Platforms, Inc.,^{[19][16]} doing business as **Meta**,^{[17][18]} and formerly named **Facebook, Inc.**, and **TheFacebook, Inc.**,^[19] is an American multinational technology conglomerate based in Menlo Park, California. The company owns and operates Facebook, Instagram, Threads, and WhatsApp, among other products and services.^[20] It is considered one of the Big Five American information technology companies, alongside Google's parent company Alphabet, Amazon, Apple, and Microsoft.

In addition to Facebook, Instagram, Threads and WhatsApp, Meta has also acquired Oculus (which it has integrated into Reality Labs), Mapillary, CTRL-Labs, and a 9.99% stake in Jio Platforms; the company additionally endeavored into non-VR hardware, such as the discontinued *Meta Portal* smart display line and presently partners with Luxottica through the Ray-Ban Stories series of smart glasses.^{[21][22]} Despite endeavors into hardware, the company relies on advertising for a vast majority of its revenue, and in 2022 amounted to 97.5 percent of it.^[12]

On October 28, 2021, the parent company of Facebook changed its name from Facebook, Inc., to Meta Platforms, Inc., "to reflect its focus on building the metaverse".^[23] According to Meta, the term "metaverse" refers to the integrated environment that links all of the company's products and services.^{[24][25][26]}

History

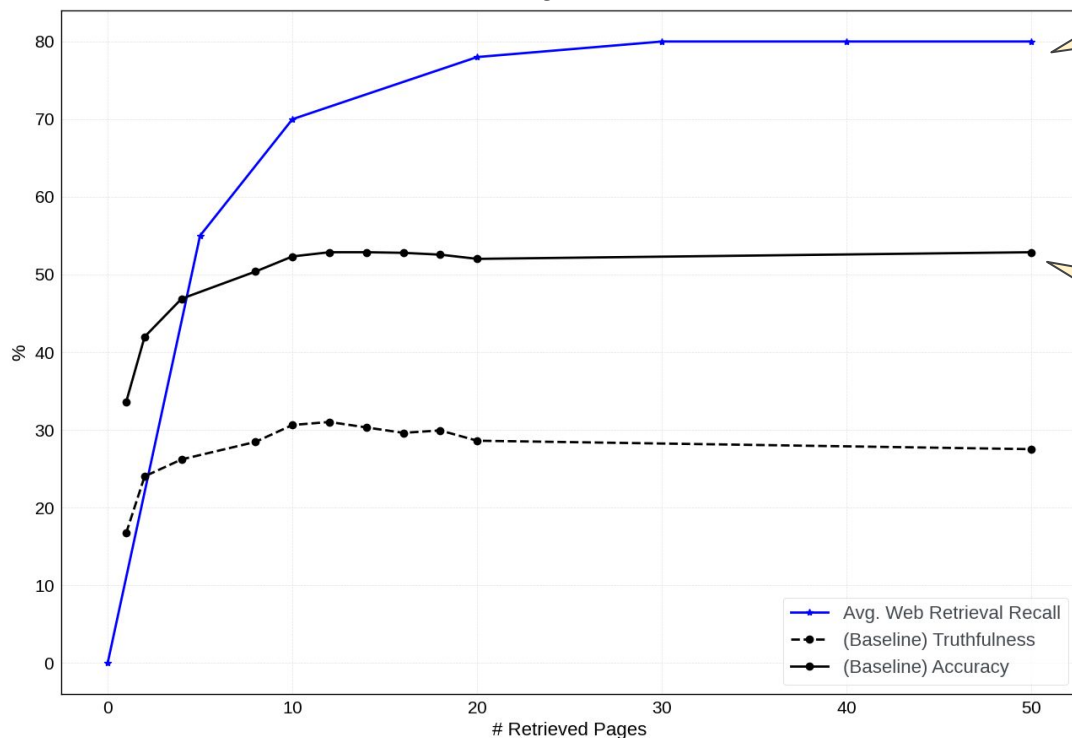
Further information: History of Facebook and Initial public offering of Facebook

Facebook filed for an *initial public offering* (IPO) on January 1, 2012.^[27] The preliminary prospectus stated that the company sought to raise \$5 billion, had 845 million monthly active users, and a website accruing 2.7 billion hits and comments daily.^[28] After the IPO, Zuckerberg would retain a 22% ownership share in Facebook and would own 57% of the voting shares.^[29]

Underwriters valued the shares at \$38 each, valuing the company at \$104 billion, the largest valuation to date for a newly public company.^[30] On May 16, one day before the IPO, Facebook announced it would sell 25% more shares than originally planned due to high demand.^[31] The IPO raised \$16 billion, making it the third-largest in US history (slightly ahead of AT&T Wireless and behind only General Motors and Visa). The stock price left the company with a higher market capitalization than all but a few U.S. corporations—surpassing heavyweights such as Amazon, McDonald's, Disney, and Kraft Foods—and made Zuckerberg's stock worth \$19 billion.^{[32][33]} *The New York Times* stated that the offering overcame questions about Facebook's difficulties in attracting advertisers to transform the company into a "must-own stock". Jimmy Kimmel joked that Zuckerberg was "the first person who sold his entire net worth to go public."

Tradeoff Between Retrieval Recall & Summarization Accuracy

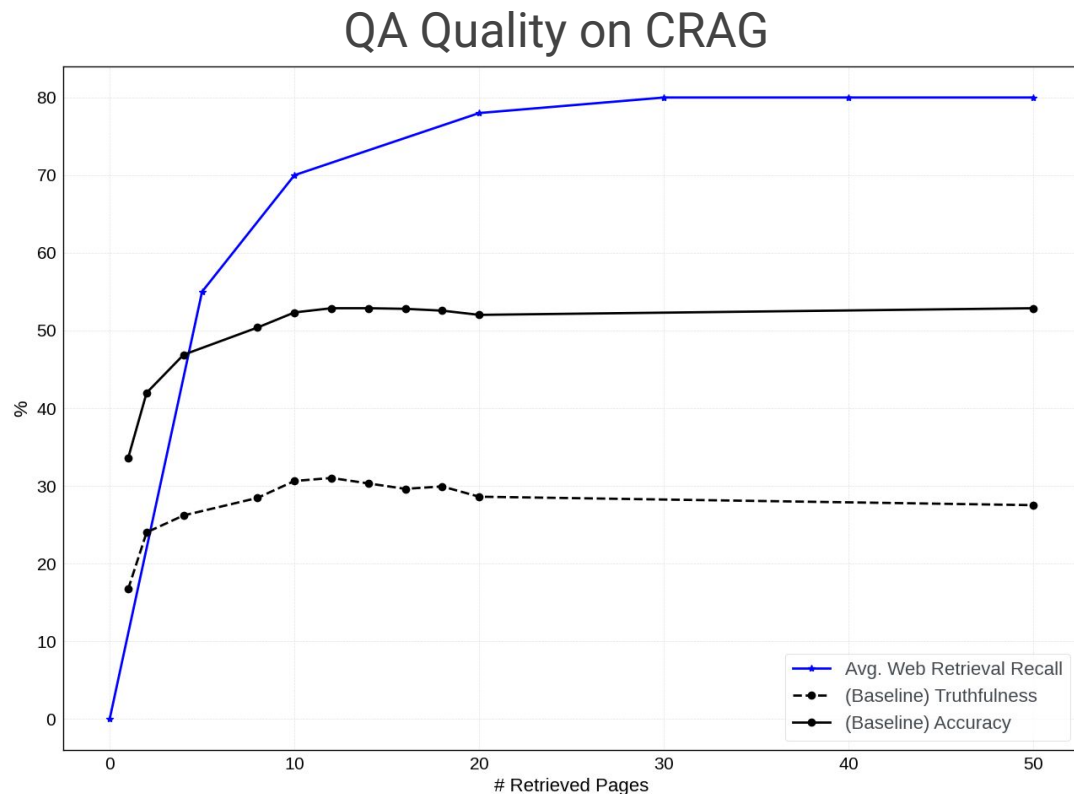
QA Quality on CRAG



Retrieval recall decides the ceiling of RAG accuracy

LLM cannot yet leverage richer but more noisier retrievals

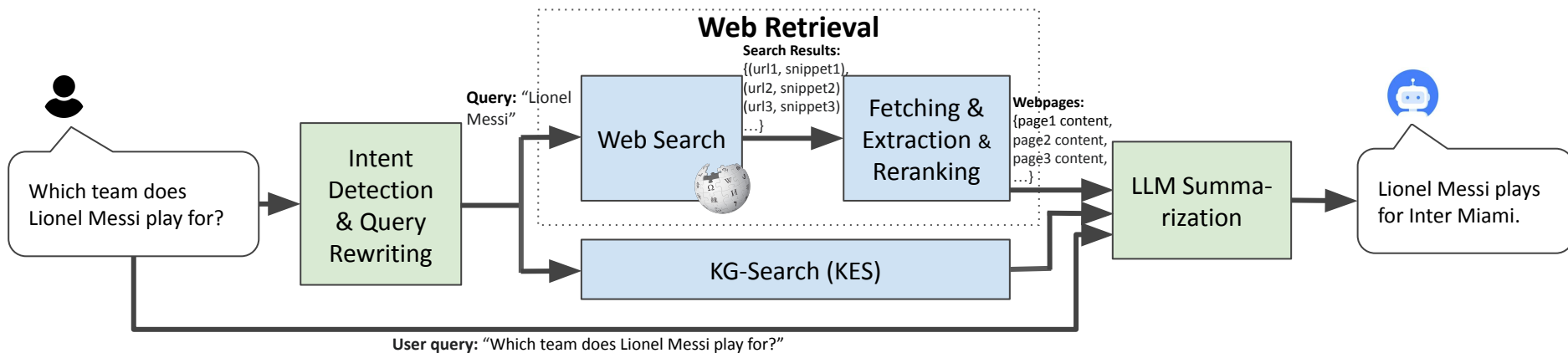
Tradeoff Between Retrieval Recall & Summarization Accuracy



Gap1. Retrieval recall

Gap 2. LLM summarization capability

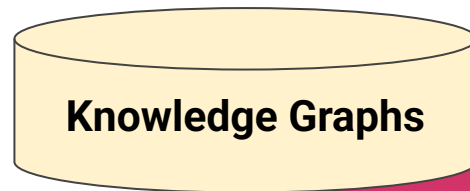
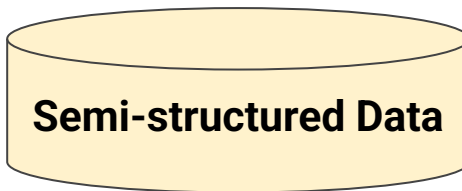
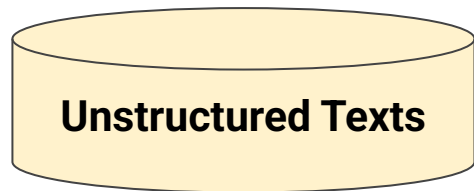
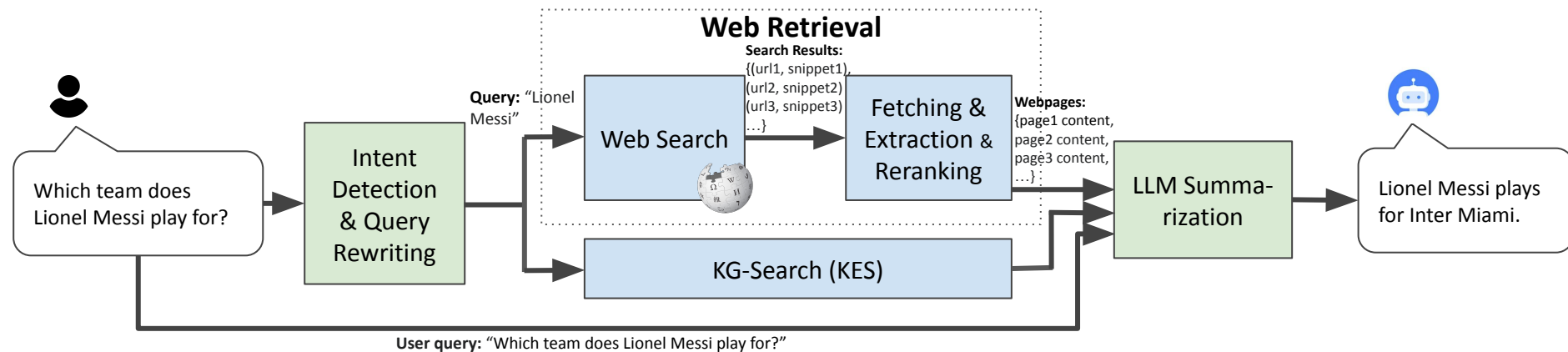
Overview of Federated RAG



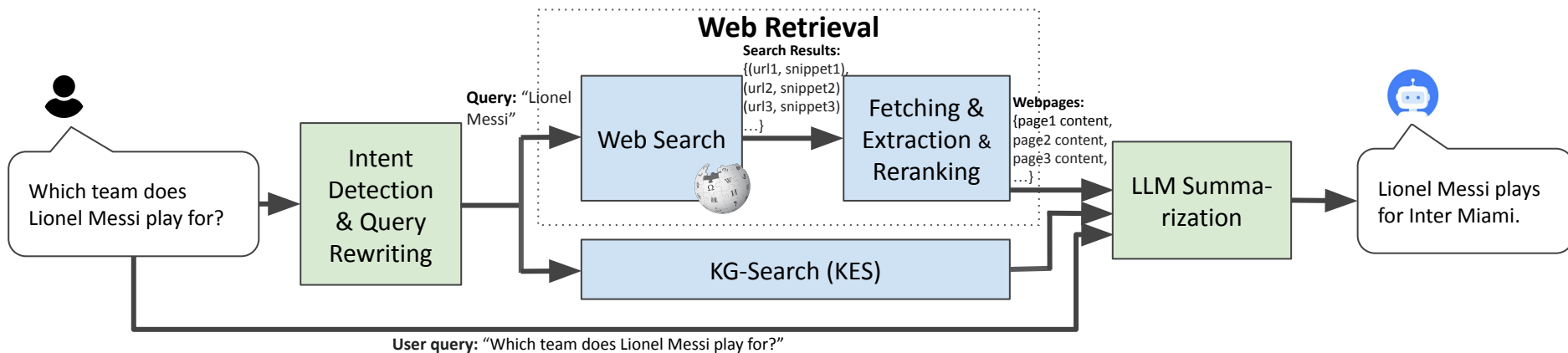
RAG Accuracy = Retrieval Recall x Summarization Accuracy

- Search recall improvement
- Retrieval precision improvement
- Complex-question decomposition
- Summarization improvement

Overview of Federated RAG



Overview of Federated RAG

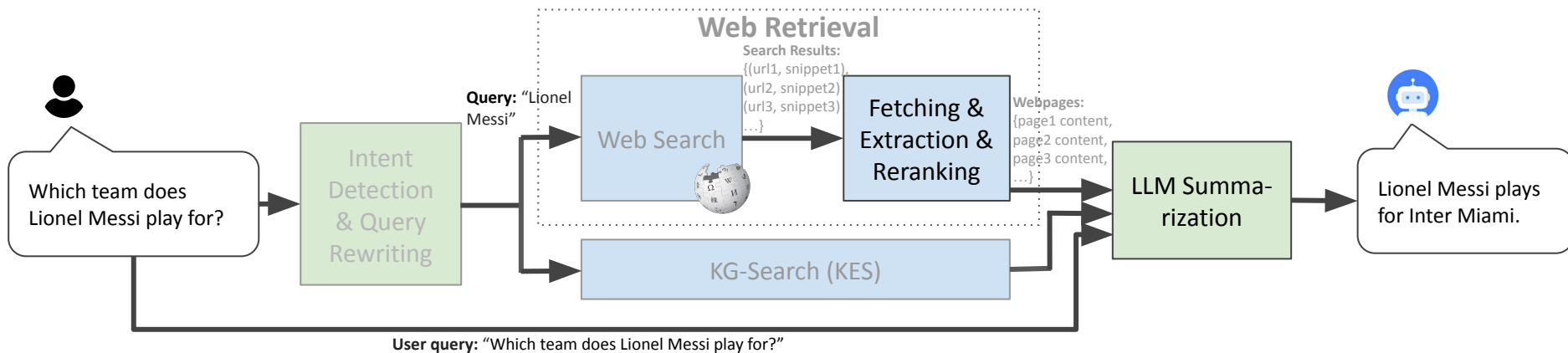


Unstructured Texts

Semi-structured Data

Knowledge Graphs

RAG on Knowledge Graphs

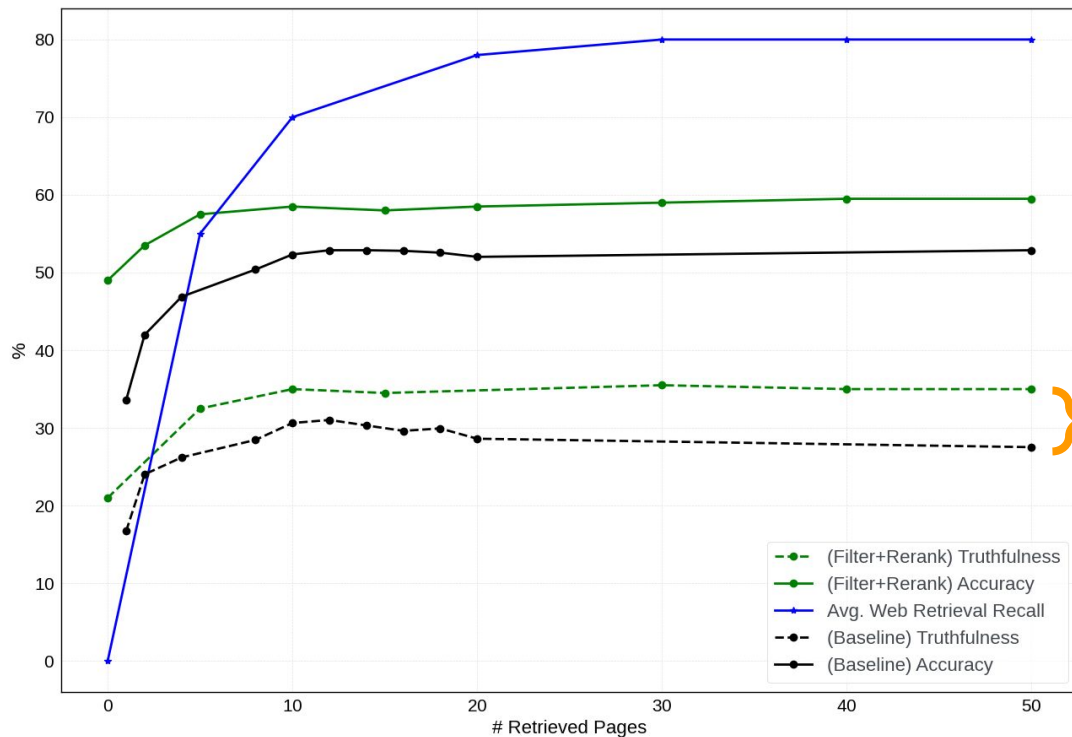


RAG Accuracy = Retrieval Recall x Summarization Accuracy

- Retrieval precision improvement
- Summarization improvement

III. RAG on Texts: Improving Retrieval Precision

QA Quality on CRAG



Passage-level filtering and re-ranking helps, but not enough

III. RAG on Texts: Increasing Robustness Against Distractions

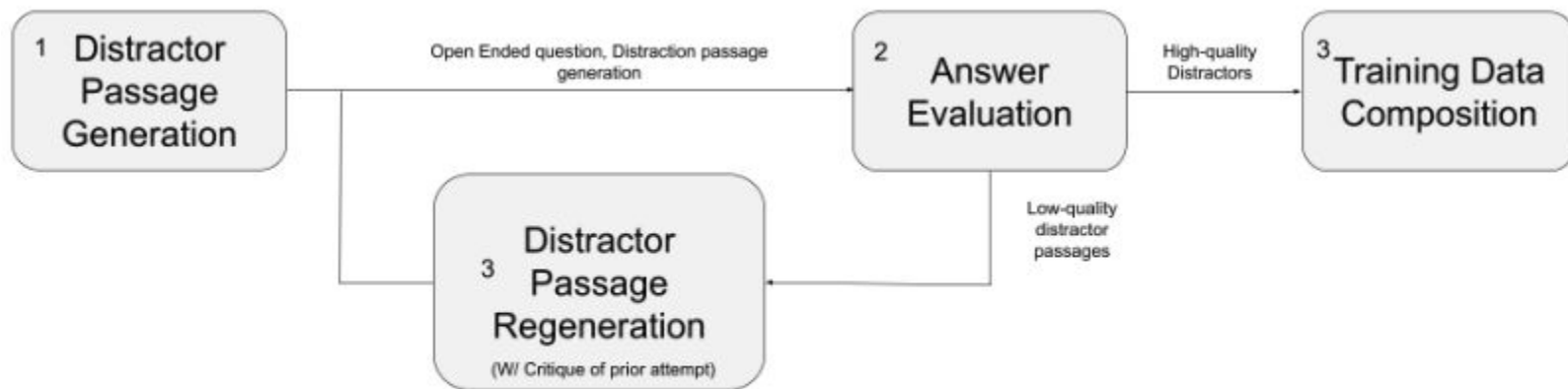


Figure 2: Overview of the synthetic distractor generation process.

III. RAG on Texts: Improving CoT w. Strategization

Q: What is the current distance between the Earth and the Moon?

Strategy:

- Step 1: Identify the most relevant documents that contain information about the distance between the Earth and the Moon.
- Step 2: Determine which of these documents provide real-time or the most current data on the Moon's distance from Earth.
- Step 3: Extract the specific information regarding the current distance between the Earth and the Moon from the most relevant and up-to-date document.

Reasoning:

- Step 1: Upon reviewing the provided documents, it's clear that documents [1], [2], [4], [5], [7], [9], and [10] all discuss the distance between the Earth and the Moon. However, to find the current distance, we need documents that either provide real-time data or are updated frequently.
- Step 2: Documents [4] and [5] seem to offer more dynamic information, with [4] explicitly mentioning that it sources the Moon's distance from Earth via the Horizon JPL API, which is updated daily. Document [5] provides a calculator for moon distances but doesn't seem to offer real-time data directly within the provided text. Document [4] stands out because it mentions an update as of "December 28, 2024," which is closer to the current date of "Thursday, January 9, 2025," than any other document.
- Step 3: Extracting the specific information from document [4], it states that as of December 28, 2024, "The Moon is 246,728 miles (396,985 kilometers) away from Earth."

A: As of December 28, 2024, the Moon is 246,728 miles away from Earth.

III. RAG on Texts: Improving CoT w. Strategization

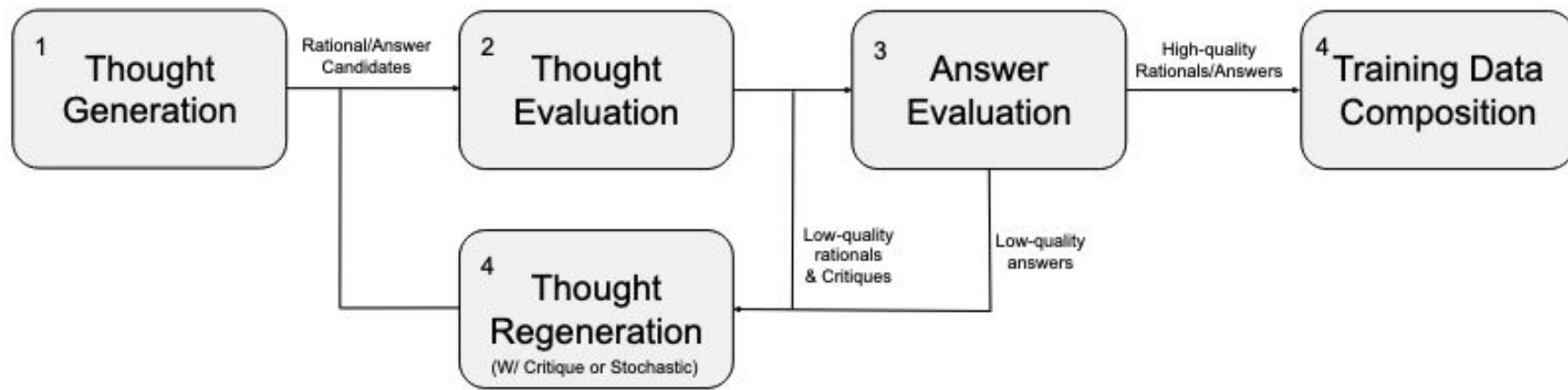


Figure 4: Overview of the iterative synthetic CoT generation process.

III. RAG on Texts: Improving CoT w. Strategization

Benchmark	Baseline	NaiveSFT	STaR (Zelikman et al., 2022)	LLMQuoter (Bezerra et al., 2025)	RAFT (Zhang et al., 2024a)	PRISM-RAG (This Work)
Avg. on 12 benchmarks	78.4%	73.1%	72.9%	80.1%	80.9%	83.8%

Method	Accurate	Hallucinated	Missing	Factual
Baseline	59.1%	24.9%	16.0%	34.2%
PRISM-RAG	62.1%	22.9%	15.1%	39.2%
•Distractor	59.3%	23.2%	17.6%	37.0%
•Strategization	62.4%	23.2%	12.3%	36.1%

Table 3: Ablation study using the CRAG dataset for the proposed fine-tuning method based on distractor resilience and dynamic strategization.

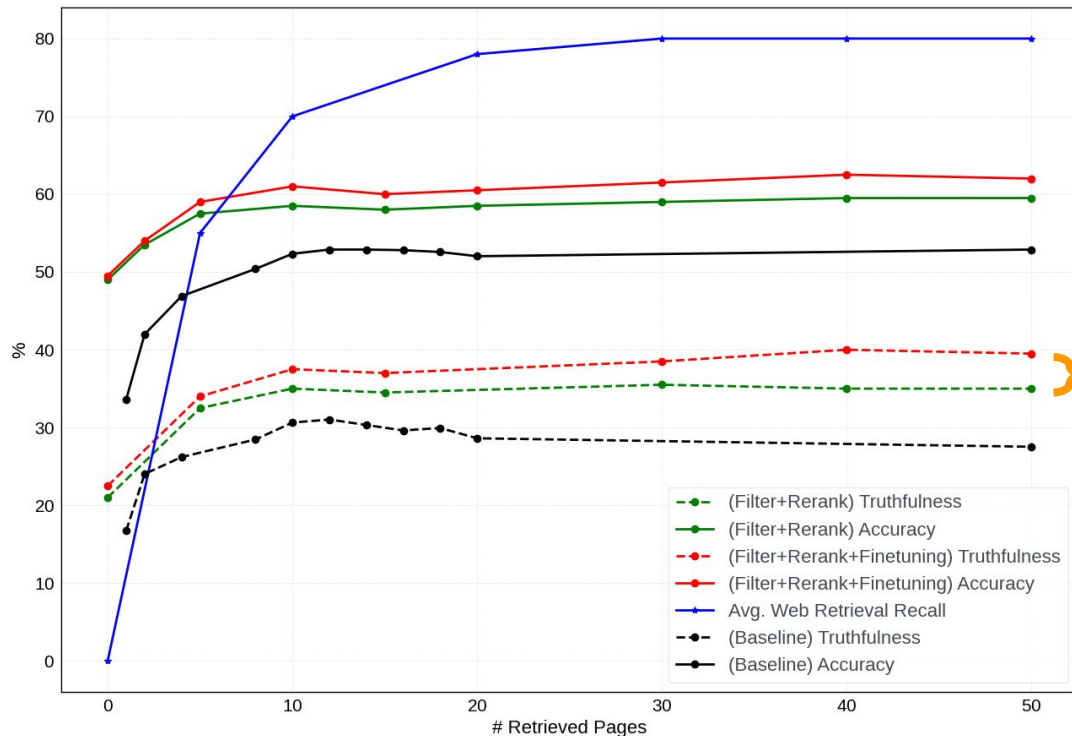
Significant improvements over baseline and sota solutions

Naive SFT only providing QA pairs can even reduce quality

Distraction SFT data reduces hallucinations, whereas strategizing also increases accuracy

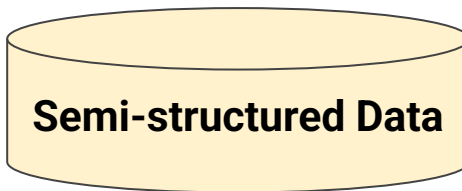
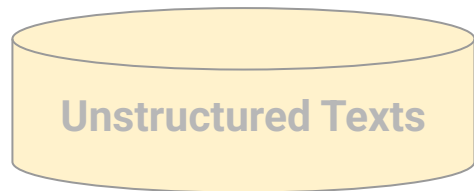
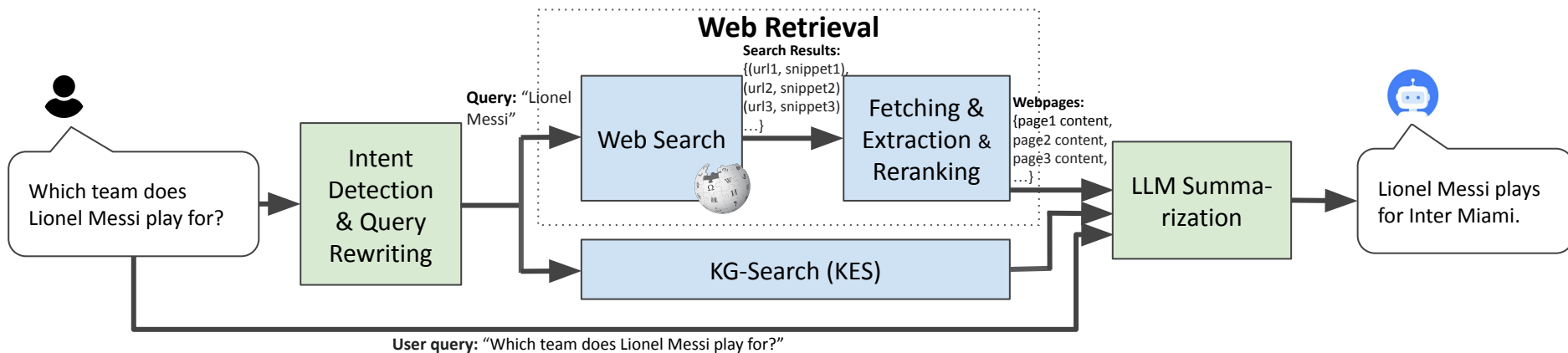
III. RAG on Texts: Improving Summarization Accuracy

QA Quality on CRAG



Our fine-tuned model further improve accuracy and reduce hallucinations





Overview of Federated RAG



II. RAG on Semi-structured Data

Title	Location	Date Posted	Member Organization
Assistant Curator Duke Lemur Center	Durham , NC	Oct 14, 2020	Yes
Deputy Director Lee Richardson Zoo	Garden City , KS	Oct 14, 2020	Yes
Curator of Reptiles, Amphibians and Birds Philadelphia Zoo	Philadelphia , PA	Oct 14, 2020	Yes

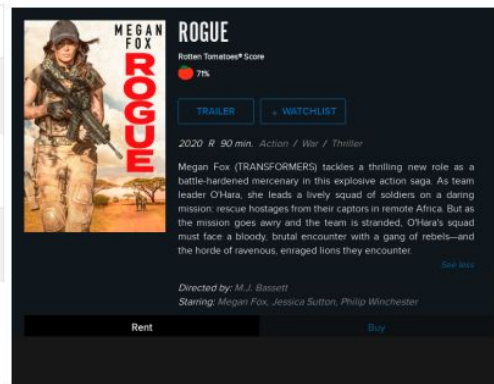
Horizontal Tables (a)

 University of Barcelona Spain Universitat de Barcelona, Gran Via de les Corts Catalanes, 585, 08007 Barcelona		  WORLD RANK 130		 COUNTRY RANK 1	
Foundation year:	1450	Rankings		Rank	Score
Short name:	UB	🏆 World University Ranking		132	77.410
Type:	Public	📊 Teaching Ranking		152	72.388
Students:	48210	🔍 Research Ranking		96	85.397
Faculty:	3020	🌐 International Diversity Ranking		238	67.672
Students/Faculty Ratio:	15.1	💰 Financial Sustainability Ranking		495	56.972
Web-site:	ub.edu	Additional Rankings		Rank	Score
Region:	Europe	Reputation Rankings		130	84.594
Location:	Barcelona	Academic Rankings		145	79.398

Free-form (b)

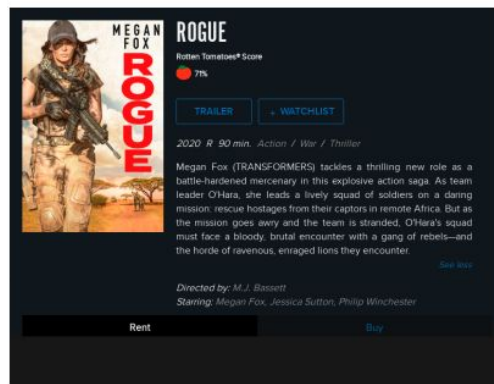
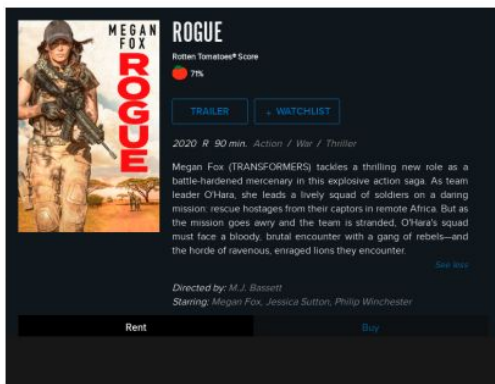
Game details	
Name	Castlevania III: Dracula's Curse
First release date	December 22, 1989
Platform	Nintendo Entertainment System Wii Shop Nintendo 3DS eShop Wii U
Developer	Konami
Publisher	Konami
Genre	Platformer Action-Adventure
Theme	Horror Fantasy
Franchises	Castlevania
Aliases	Castlevania 3: Dracula's Curse Akumajō Densetsu Devil's Castle Legends

(c) Attr-Val Pairs
(Vertical Tables)



(d) Free-form

II. QA vs. Knowledge Extraction on Semi-Data

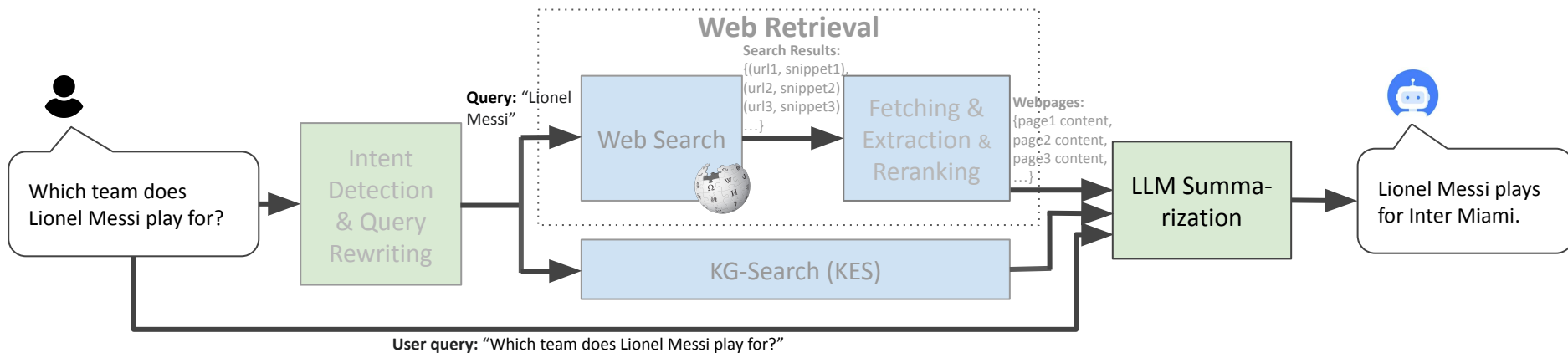


Q: Is the movie Rogue well received at Rotten Tomatoes?

A: Rogue got a medium score of 71% on Rotten Tomatoes

Subject	Predicate	Object
Rogue	Rotten Tomatoes® Score	71%
Rogue	Directed by	M.J. Bassett
Rogue	Starring	Megan Fox Jessica Sutton Philip Winchester

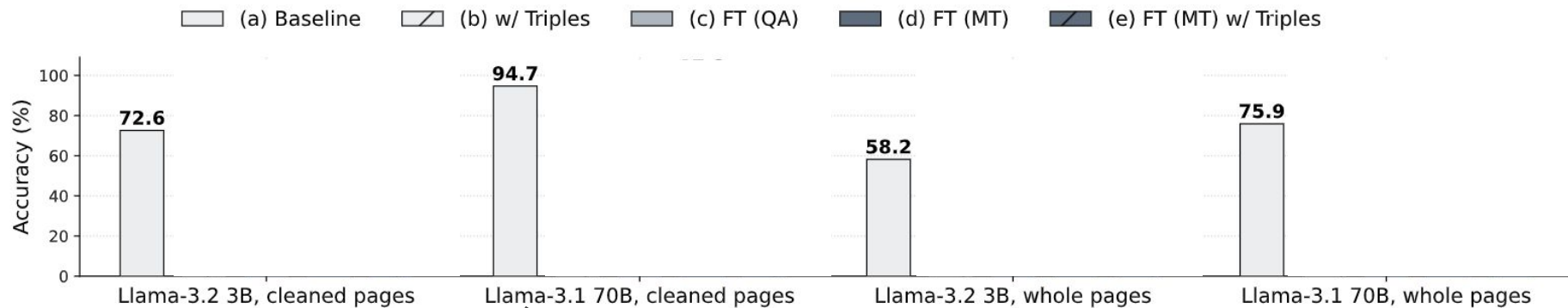
RAG on Semi-Structured Data



RAG Accuracy = Retrieval Recall x Summarization Accuracy

- Summarization improvement

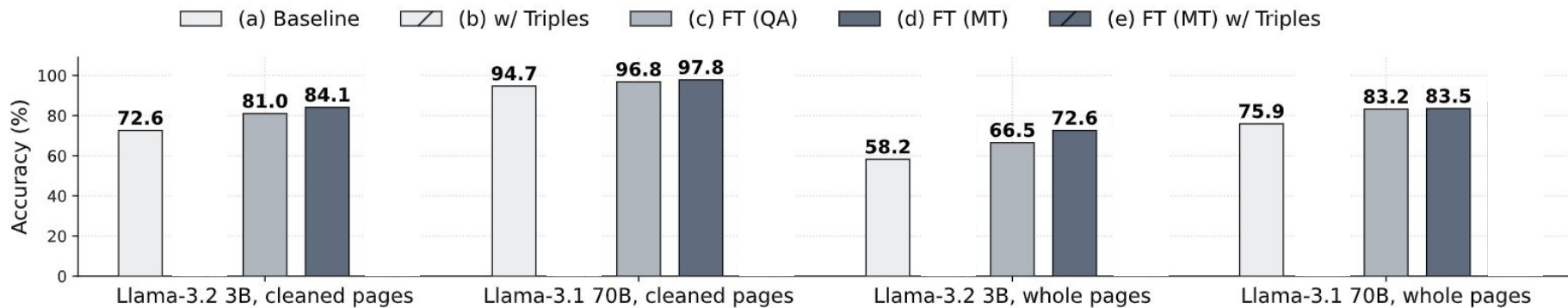
II. RAG on Semi: Current QA quality



Smaller LLMs still have quality gaps, but large LLMs have superior quality on cleaned pages

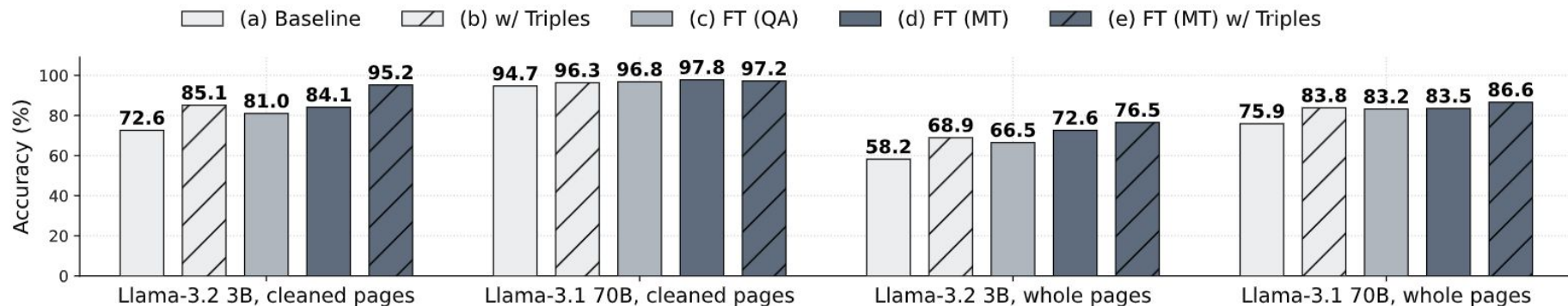
LLMs still struggle with whole HTML pages in the wild

II. RAG on Semi: Extraction as Fine-tuning Tasks



Multi-task fine-tuning with QA and Extraction tasks can significantly improve small QA quality

II. RAG on Semi: Extraction Augmentation



Augmenting the webpages with knowledge triples further improve QA quality. However, it is hard to obtain correct triples.

II. Knowledge Extraction on Semi-structured Data

	Backbone	Setting	Triple-level		
			P _{LM}	R _{LM}	F-1 _{LM}
in-domain	Llama 3.1-70B-Instruct	zero-shot	53.1	71.8	61.0
	Llama 3.1-70B-Instruct	2-shot	88.4	92.0	90.2
	Llama 3.1-70B-Instruct	fine-tuned	71.8	76.3	74.0
	Claude 3.7 Sonnet	2-shot	89.9	92.8	91.3
	GPT-4o	2-shot	93.9	95.5	94.7
out-of-domain	generated scripts	single call	53.1	49.3	51.1
	generated scripts	multiple calls with feedback	75.8	69.9	72.7
	Llama 3.1-70B-Instruct	zero-shot	60.3	74.1	66.5
	Llama 3.1-70B-Instruct	3-shot	69.4	70.1	69.7
	Llama 3.1-70B-Instruct	fine-tuned	72.1	71.0	71.5
	Claude 3.7 Sonnet	3-shot	76.0	77.4	76.7
	GPT-4o	3-shot	76.4	76.9	76.6
	generated scripts	single call	50.9	45.6	48.1
	generated scripts	multiple calls with feedback	63.8	57.9	60.7

Few-shots w. in-domain examples achieves high extraction quality

Knowledge extraction from unseen websites remains difficult

We can teach LLMs to write scripts to extract triples from semi-structured content, even from out-of-domain websites

Table 2: Triple extraction performance. All numbers are in percentage (%).

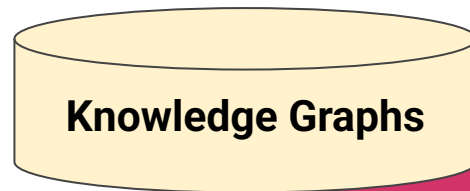
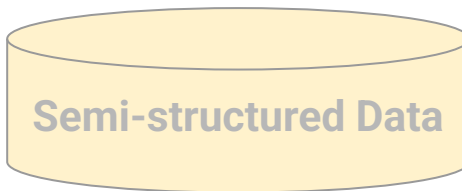
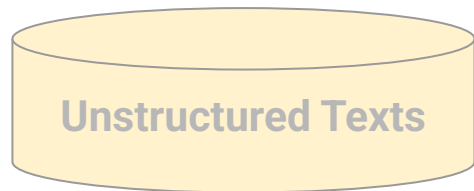
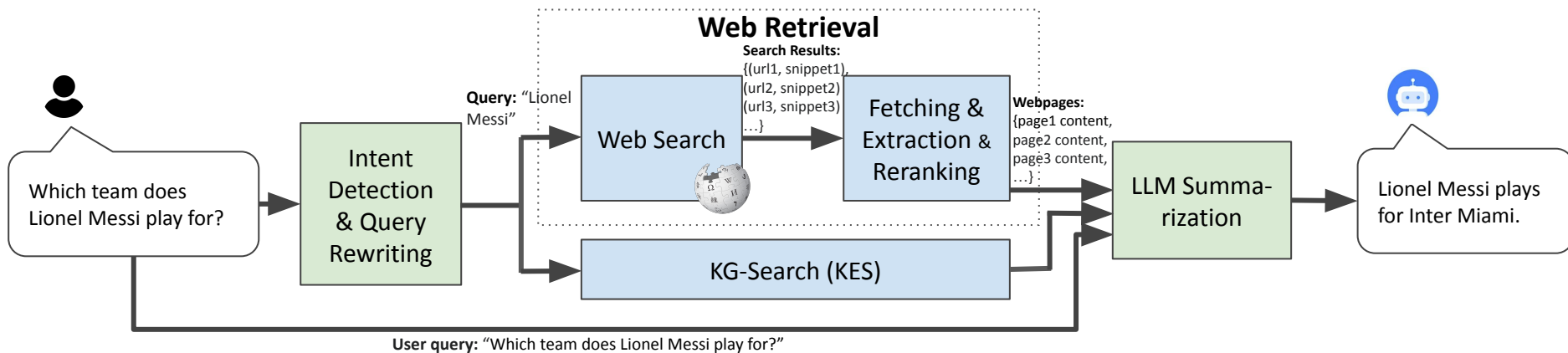
II. RAG on Semi: Extraction Results to Improve QA

Table 6: Zero-shot QA performance in $\text{Accuracy}_{\text{LM}}$ (%) on cleaned pages. Script-extracted triples improve QA quality for 3B models.

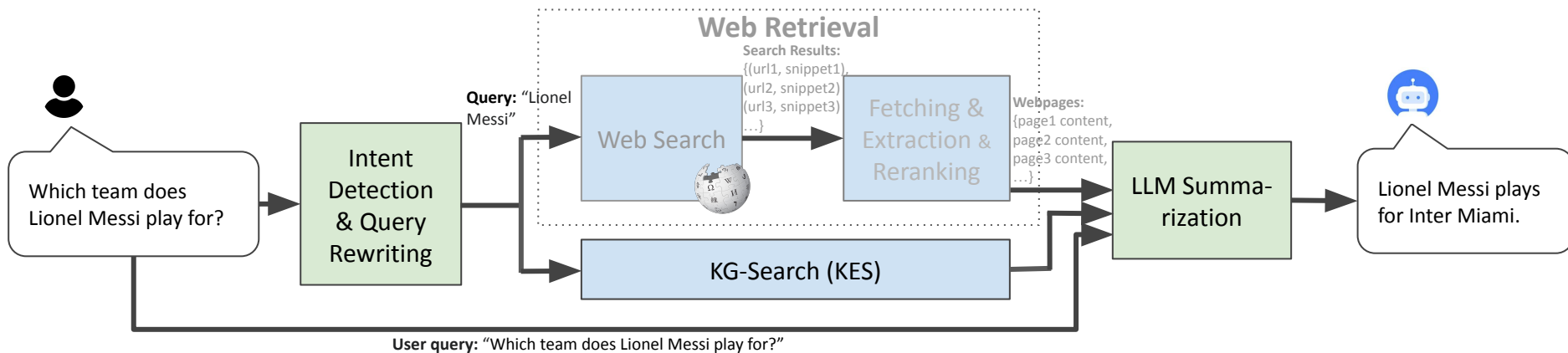
		Additional reference	L-3B	L-70B	Q-3B	Q-72B
in-domain	/		77.1	95.1	81.5	93.5
	Script-extracted triples		80.6	94.9	87.5	94.2
out-of-domain	/		72.6	94.7	81.6	95.1
	Script-extracted triples		77.7	92.9	86.5	94.2

Script-based extraction can improve small LLMs on QA

Overview of Federated RAG



RAG on Knowledge Graphs



RAG Accuracy = Retrieval Recall x Summarization Accuracy

- Search recall improvement
- Retrieval precision improvement
- Summarization improvement

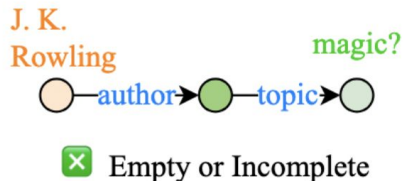
I. RAG on KG: Intuition of KERAG

(a) Natural Language Question:

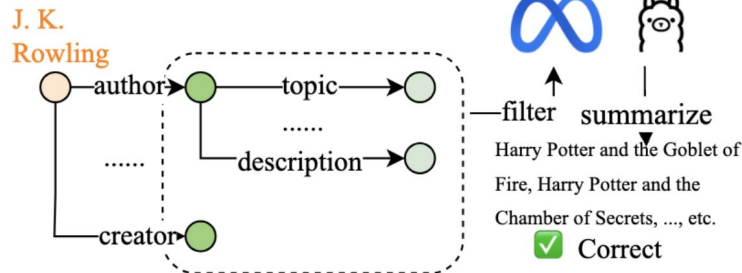
Q: Which **books** written by **J. K. Rowling** are **related to magic**?

(b) Standard SP-based KBQA approach:

```
SELECT ?book
WHERE {
  ?book rdf:type :Book .
  ?book :author :J_K_Rowling .
  ?book :topic :Magic .}
```



(c) Our proposed KERAG approach:

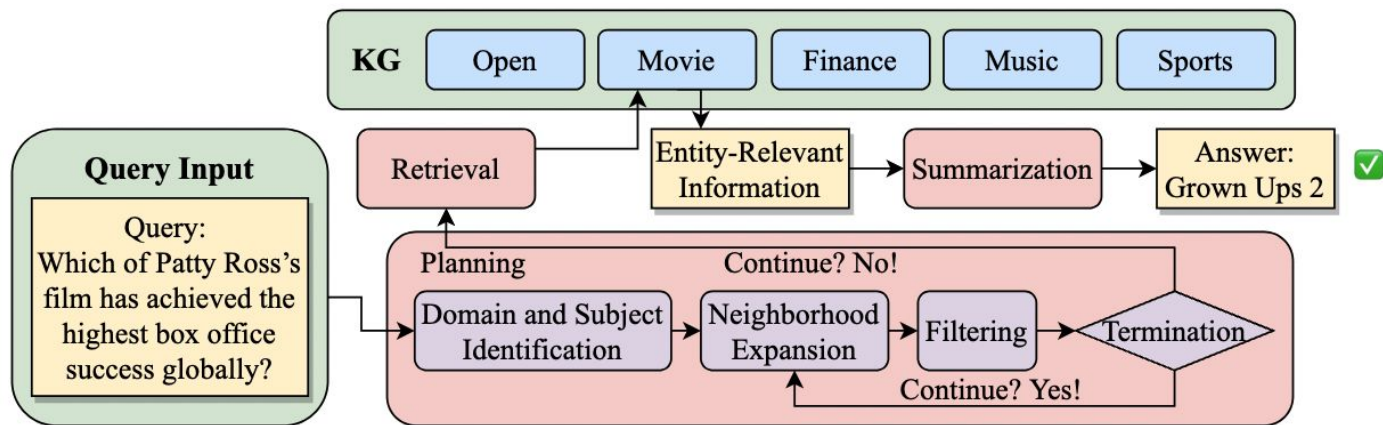


Existing methods: Semantic parsing

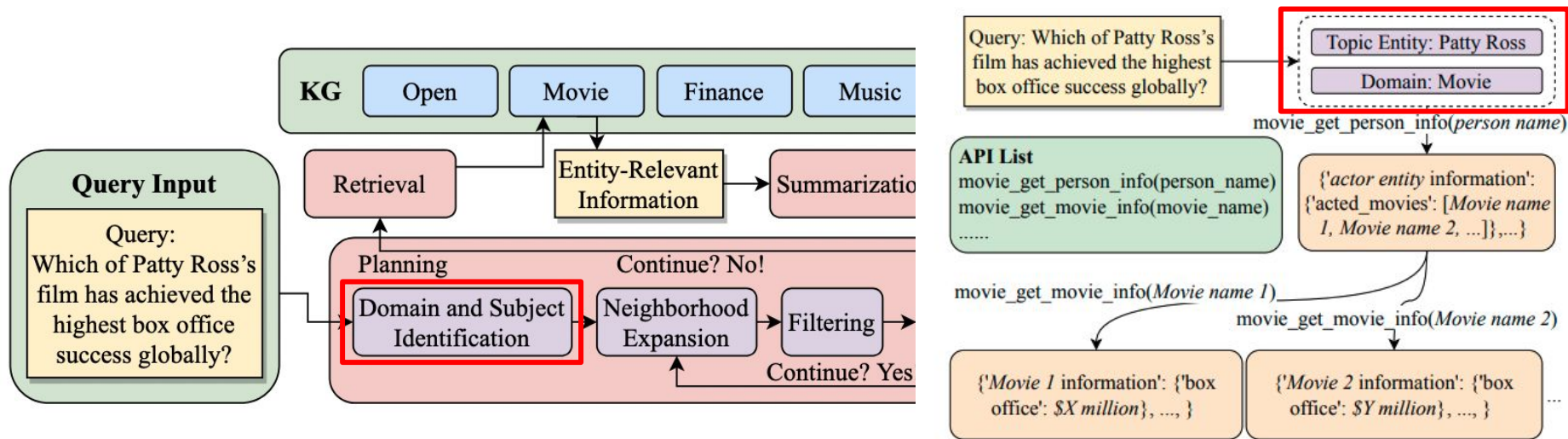
- SPARQL queries
- One or a few paths

Our methods: Identify a subgraph and leverage LLM summarization capability

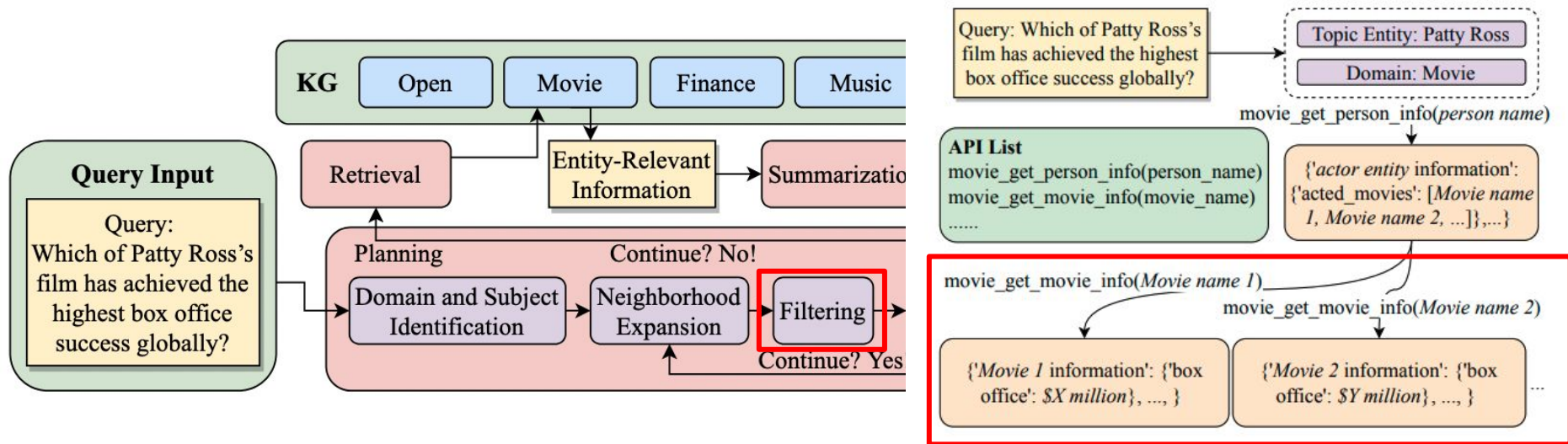
I. RAG on KG: KERAG Solution



I. RAG on KG: KERAG Solution

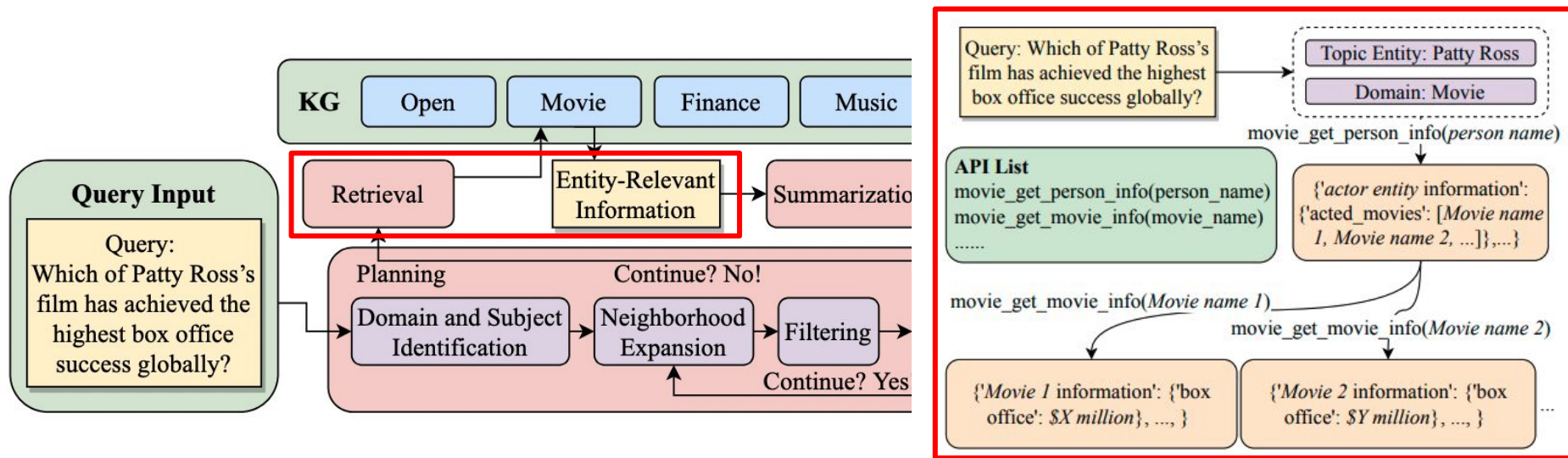


I. RAG on KG: KERAG Solution



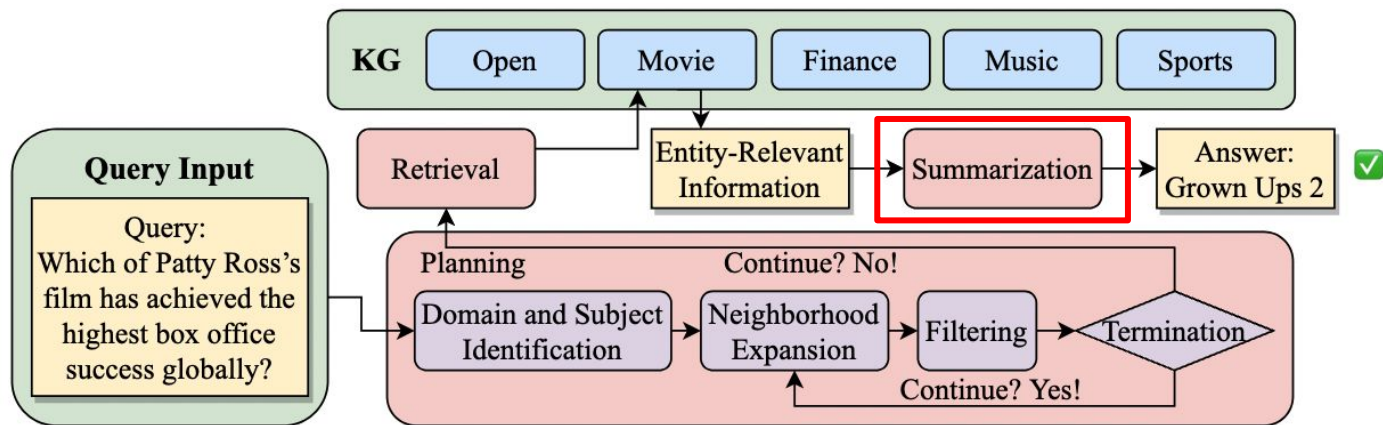
Retrieval **precision** optimization:
Neighborhood filtering to reduce noise

I. RAG on KG: KERAG Solution



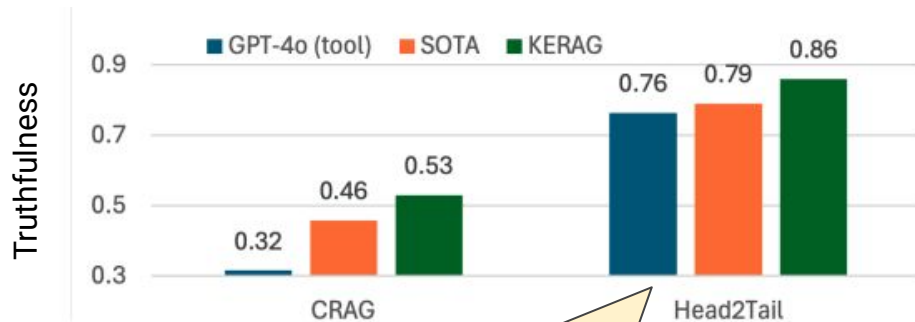
Entity level parsing + predicate level planning
for reduced **latency** & flexibility across different KGs

I. RAG on KG: KERAG Solution



Summarization: CoT-based fine-tuning to better answer complex questions w. reasoning and aggregation

I. RAG on KG: KERAG Results



Higher quality compared to SOTA

Multi-hop improves retrieval recall and QA accuracy, but can also bring more hallucinations

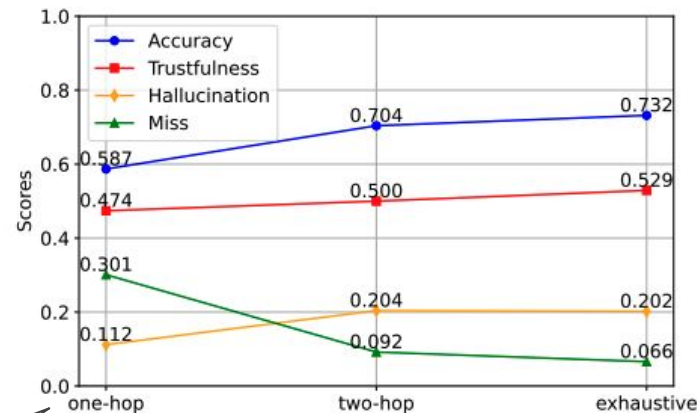
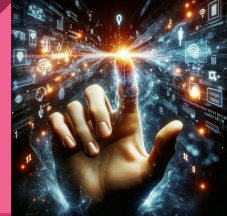


Figure 5: Comparison between one-hop, two-hop, and exhaustive retrieval on CRAG.

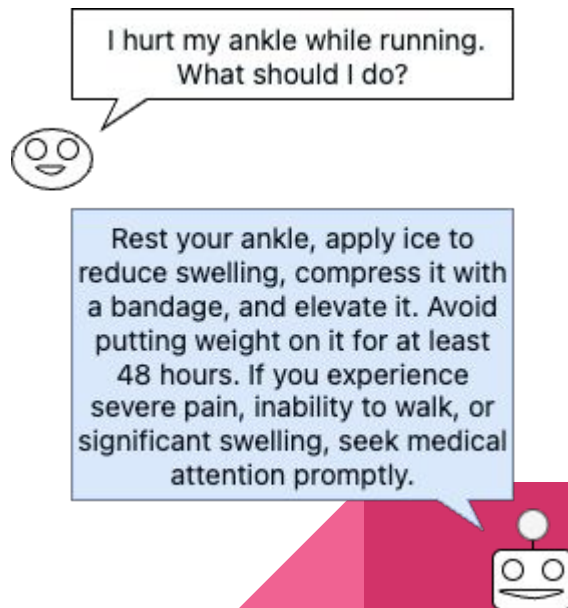


3. How to Internalize More Factual Knowledge to LLMs

Plug-and-Play Memory for Language Models

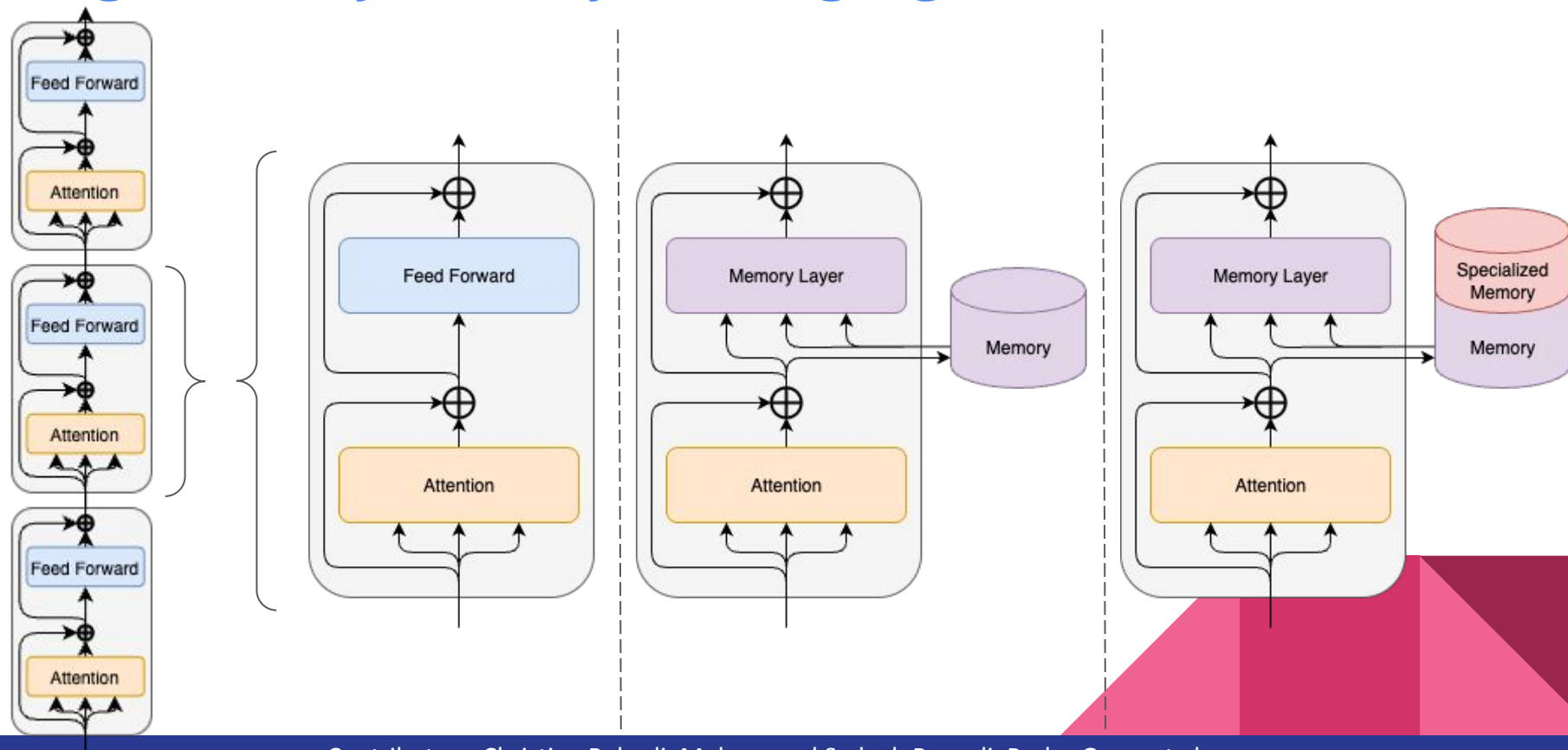


W/o Extended Memory



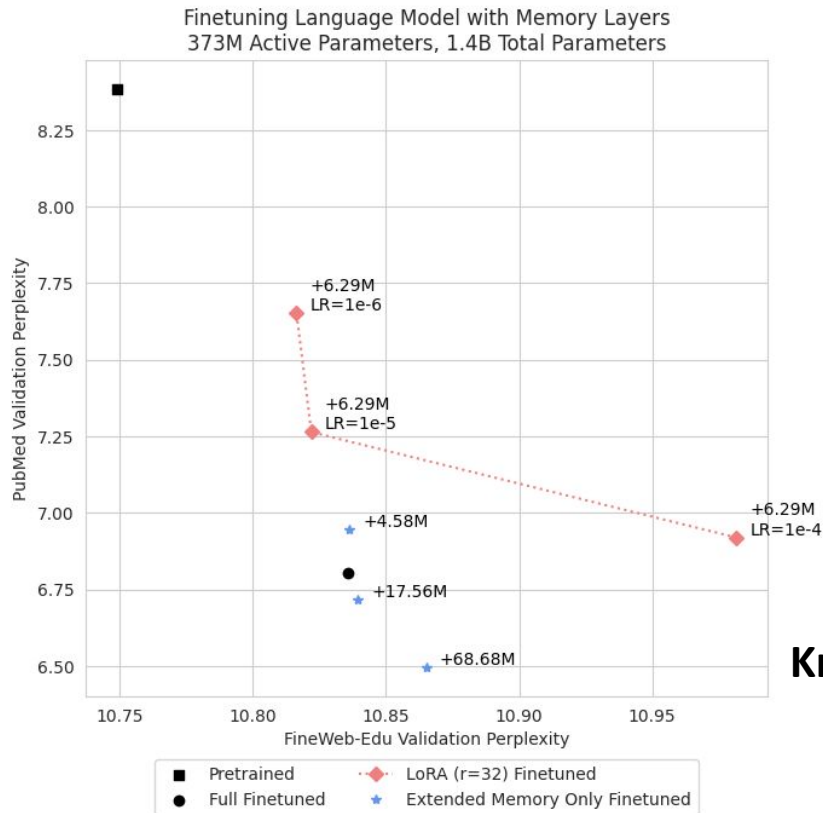
W. Extended Memory

Plug-and-Play Memory for Language Models



Plug-and-Play Memory for Language Models

Additional medical knowledge



Knowledge obtained during pre-train



Putting Everything in Context

***Building Next-Generation Intelligent
Assistants for Wearable Devices***

Stages of Intelligent Assistants

I. Chatbot

Text input



II. Voice Asst

Voice input

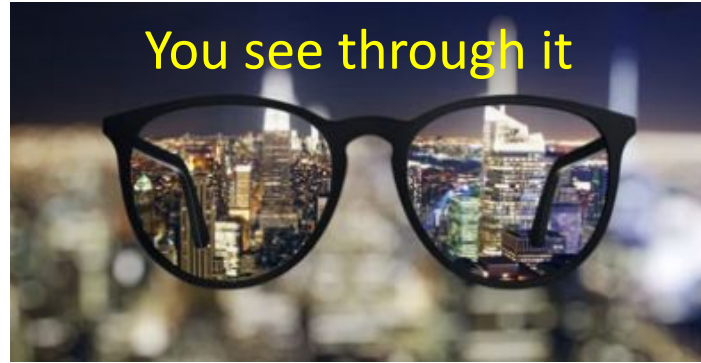


III. Wearable Asst

Voice + Visual + Context



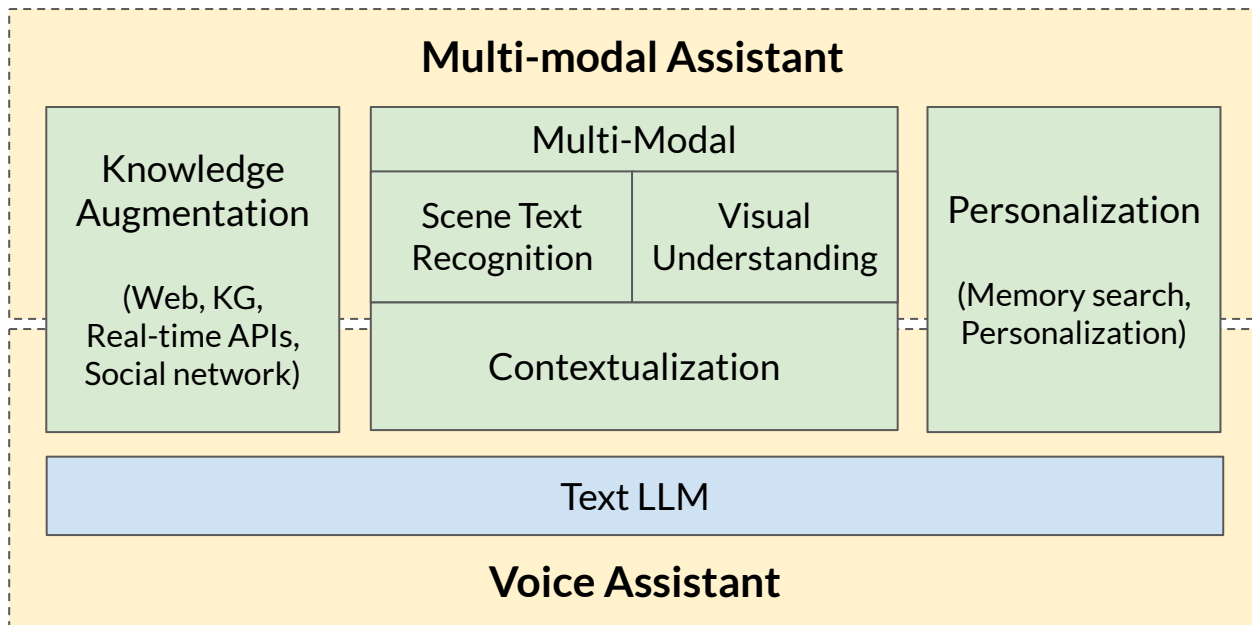
What Is Different for A Wearable Assistant?



Ray-ban Meta Video (0:24, 4:22)

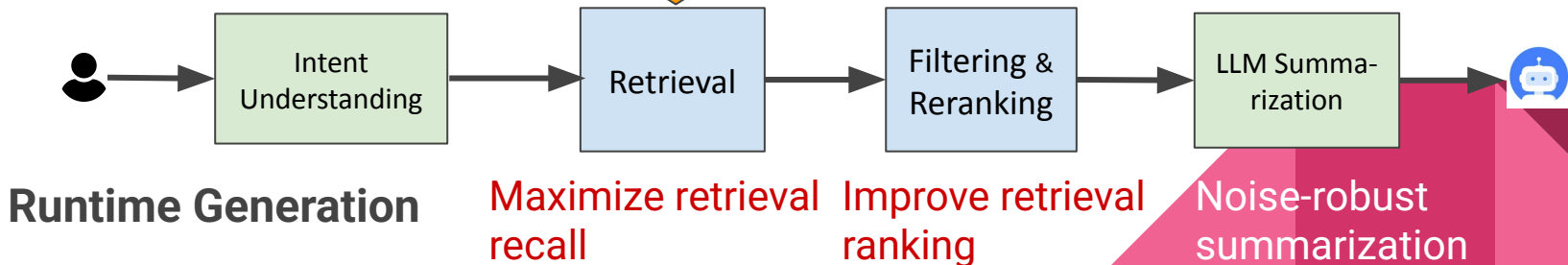
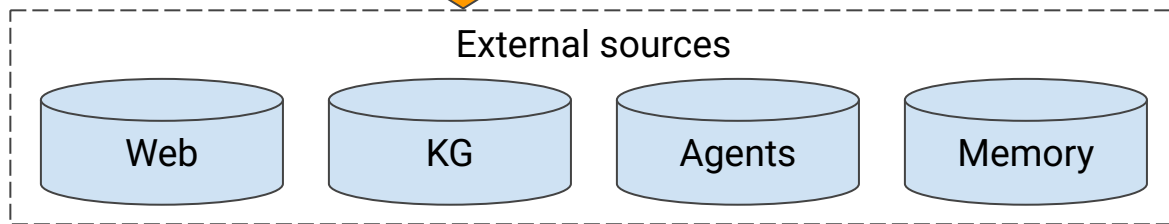
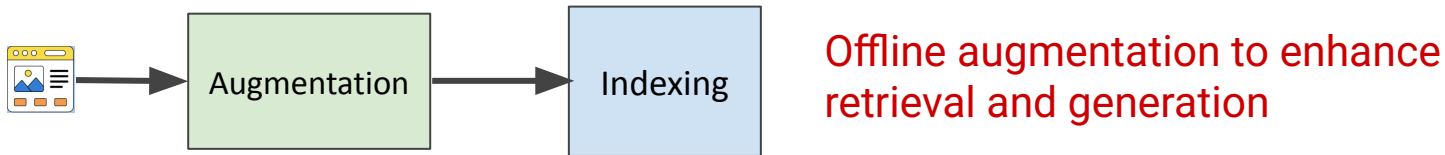


Pillars for Next-Generation Assistant



RAG Pipeline is Fundamental for Various Applications

Offline Augmentation



CRAG IS BACK—CRAG-MM KDDCup'25

🕒 Warm-Up Round: Completed 🕒 Round 1: Completed 🕒 Round 2: Completed

#multi_modal

#multi_turn



CRAG-MM Challenge

Improve RAG with Real-World Benchmarks

🏆 Total Prize Value **USD 33000**



By Meta Meta

👁 113.5k

👤 922

👥 256

🚀 4925

❤ 63

Share

- First benchmark for Wearables use cases
- Multi-modal Multi-turn conversations

CRAG-MM: Example Questions



Simple - recognition

Question: What brand is this?

Answer: The brand is contigo.

Dynamism: static



Comparison

Question: Is this product cheaper on Amazon?

Answer: No, the product costs \$79.96 on amazon which is more expensive than this place.

Dynamism: fast-changing



Multi-hop

Question: how many books has this book's author written?

Answer: Patrick o'Donnell, the author of the unvanquished, has written 13 books in total.

Dynamism: slow-changing



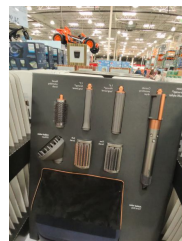
Simple - knowledge

Question: what chemicals are used in this?

Answer:

The active ingredients in Great Value Disinfecting wipes are quaternary ammonium compounds and ethanol.

Dynamism: slow-changing



Aggregation

Question: Which heads can I use for straight hair?

Answer: The Soft smoothing brush and the Firm smoothing brush.

Dynamism: static



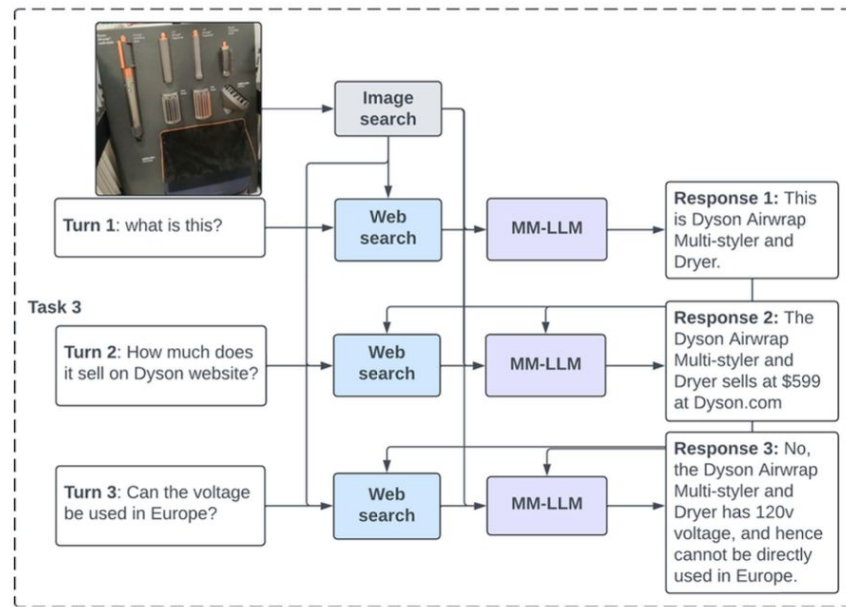
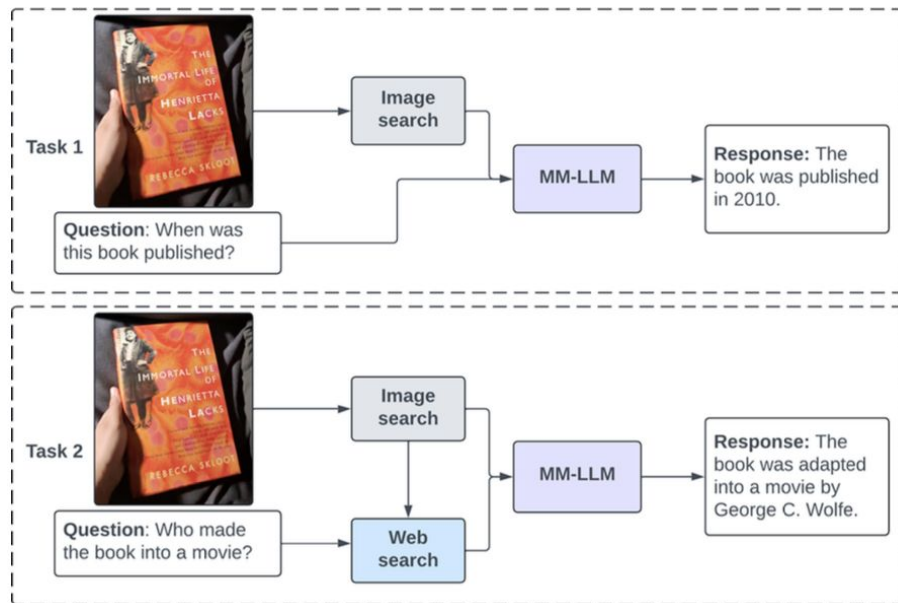
Reasoning

Question: Could a guest tour inside this museum on Christmas?

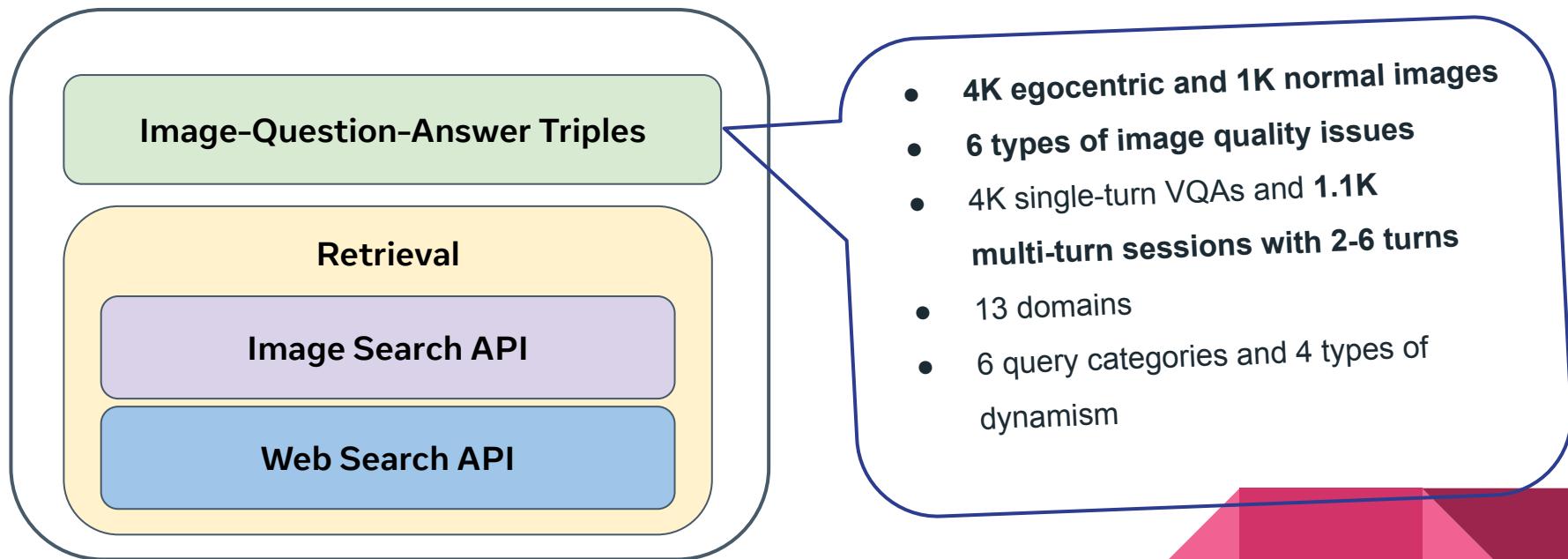
Answer: No, the Indianapolis Firefighters Museum is only open on weekdays and Saturdays from April to October.

Dynamism: slow-changing

CRAG-MM Tasks

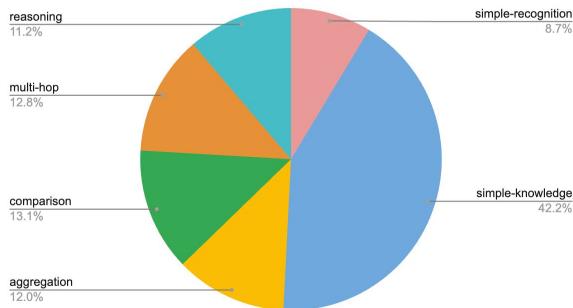


CRAG-MM Benchmark Overview

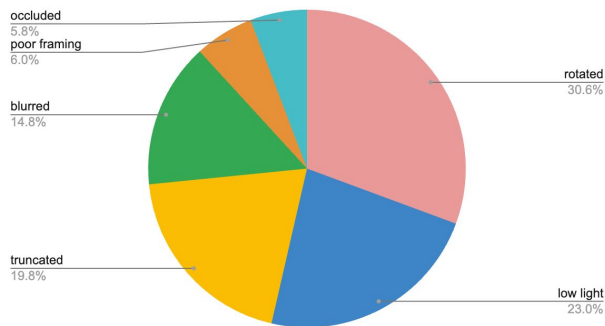


CRAG-MM Benchmark Overview

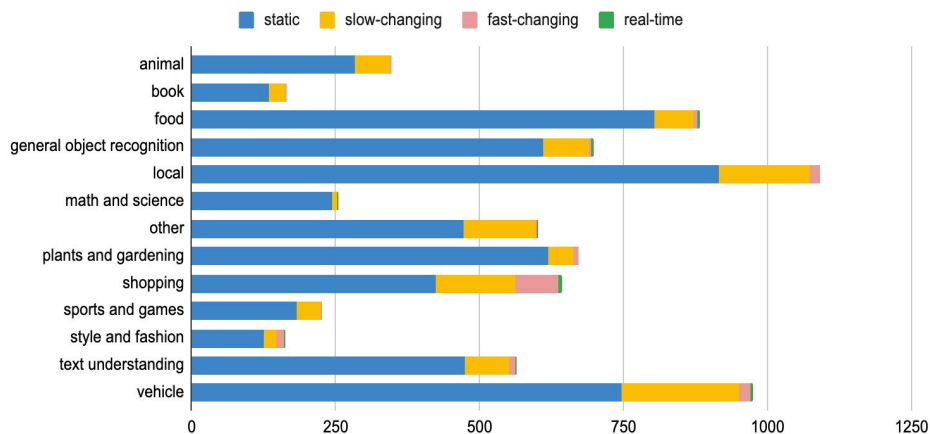
Question Type Distribution



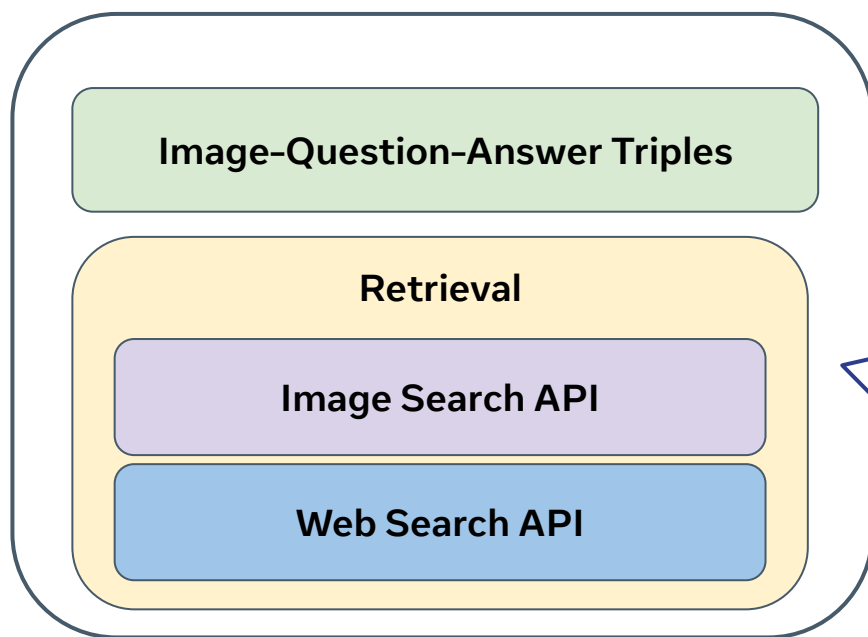
Distribution of Low-quality Images



Dynamism Distribution by Category



CRAG-MM Benchmark Overview



- **Unified Image and Web search API** with semantic search
- **Image KG** with 51k corpus size
- 1.4M web chunks

CRAG-MM Winning Solutions

Tasks	Truthfulness = Perf + Acc * 0.5 - Hallu	Perfect	Acceptable	Missing	Hallucination
Task 1	12.8%	14.7%	1.5%	80.3%	3.4%
Task 2	23.3%	26.8%	3.4%	61.0%	8.8%
Task 3	36.9%	35.9%	2.1%	56.6%	5.5%
All Egocentric Images	21.0%	24.4%	2.1%	65.8%	4.4%

Still a gap from single-turn text RAG results (36%)

Lower quality on ego-centric images

Quite high missing rate

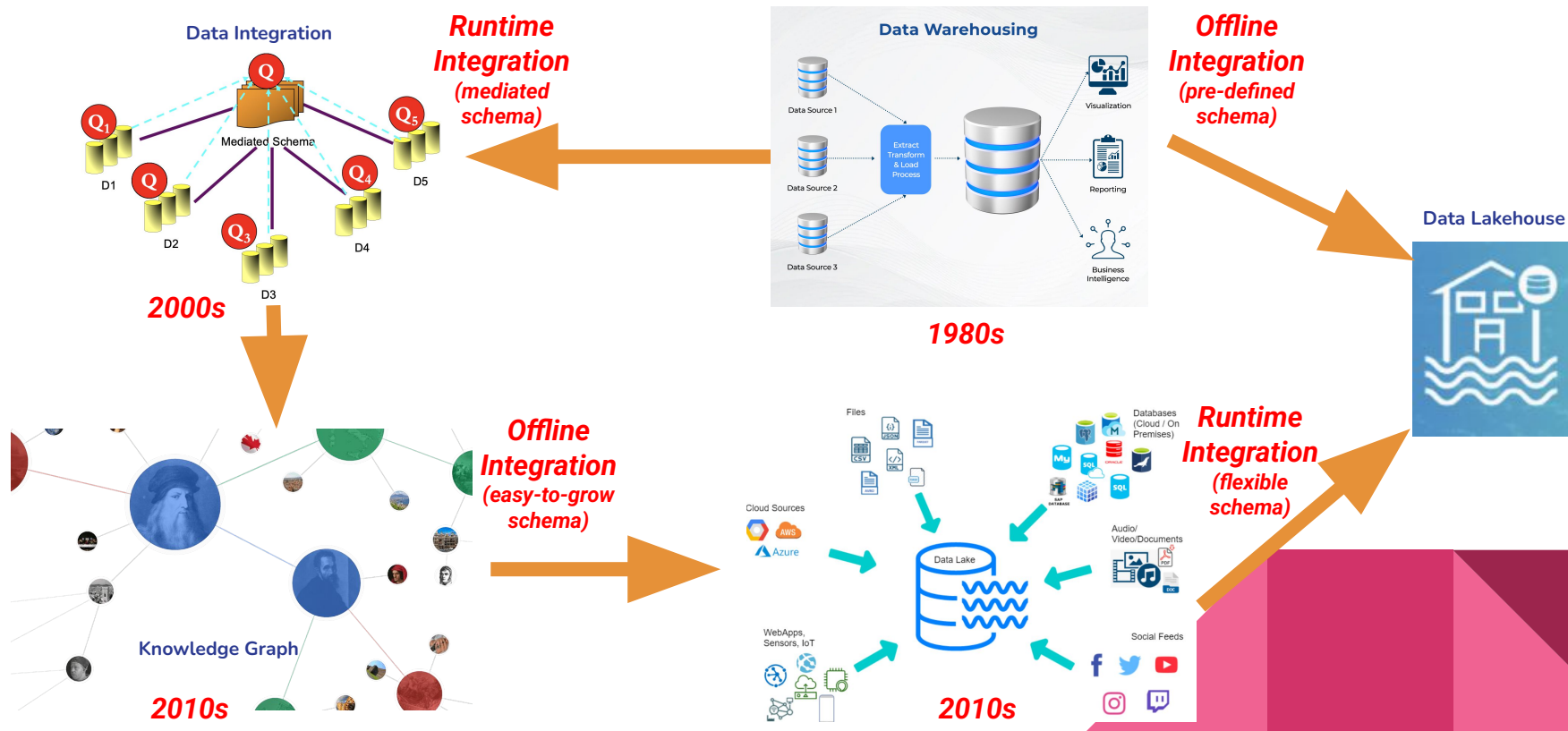
Massive Data Exist in Various Forms

- XML
- JSON
- Key-value stores
- NoSQL (semi-struct.)
- Files

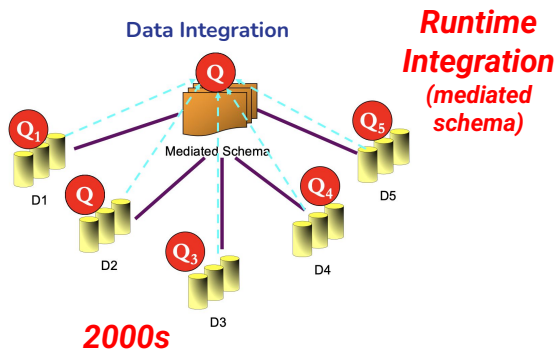


- Relational databases
- Spreadsheets
- Graph databases
- Knowledge graphs

How to Access All Forms of Data in a Holistic Way?

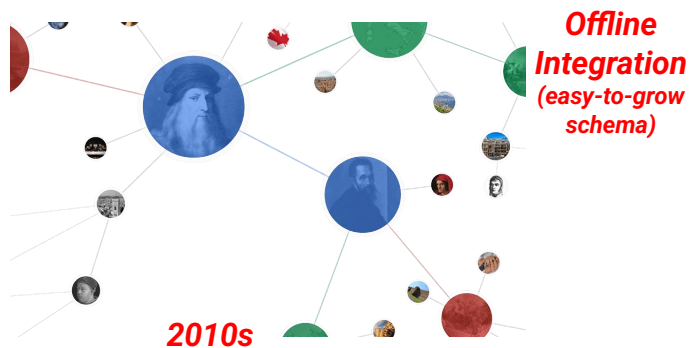


Inherent Challenges Have NEVER Changed



Challenges:

- Answer summarization non-trivial
- Heterogeneity hard-to-resolve
- Conflict resolution challenging
- Latency high



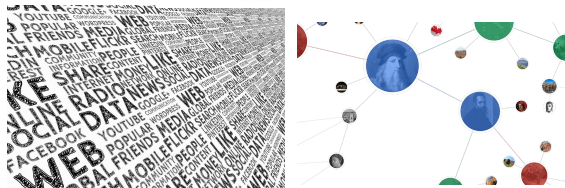
Challenges:

- Offline integration very strenuous
- Limited coverage
- Constrained expressive power
- Complex QA difficult

Dual Neural Knowledge



**Runtime
Integration
through RAG**



**Symbolic knowledge
as is *in the wild***



**Internalized Neural
knowledge *within LLMs***

**Offline
Integration
through
Pre-training**

LLM as the rescue:

- Agentic tool-using, Text2SQL capabilities
- Summarization from multiple sources
- Reasoning capabilities to synthesize data
(no programming needed!)

LLM as the rescue:

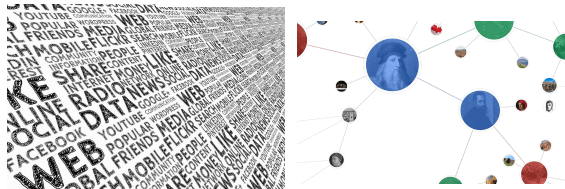
- Min fixed schema
- Max expressive power
- Different modalities in the same neural space

Trend 1. Data Representation Revol.—Dual Neural Knowl.

A.k.a. Do we still need KGs? YES, but more than that



**Runtime
Integration
through RAG**



**Symbolic knowledge
as is *in the wild***



**Internalized Neural
knowledge *within LLMs***

**Offline
Integration
through
Pre-training**

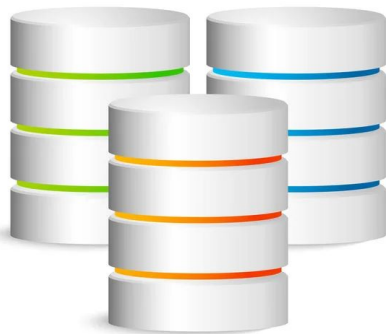
Question 2. How to leverage external data in RAG more effectively?

Question 3. How to know when to rely on what?

Question 1. How to internalize more factual knowledge into LLMs?

Trend 2. Data Integration Revolution—Pay as you go

A.k.a. Do we need to aggressively build KGs? NOT necessarily



~~Use raw data as is;
integration through RAG~~

Runtime Integration Challenges:

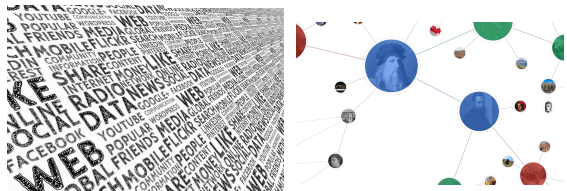
- Answer summarization non-trivial
- Heterogeneity hard-to-resolve
- Conflict resolution hard
- Latency high

Trend 2. Data Integration Revolution—Pay as you go

A.k.a. Do we need to aggressively build KGs? NOT necessarily



*Runtime
Integration
through RAG*



*Pay-as-you-go
Integration*

Question 1. What if we simply buffer answers from previous conversations?

Question 2. When shall data be connected or integrated?

Question 3. Where to store the integrated data, or derived knowledge?

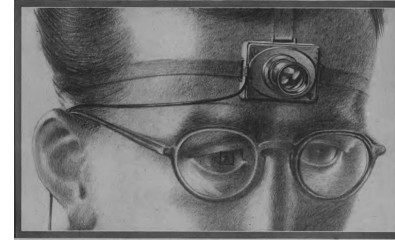
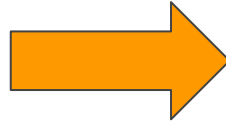
Question 4. How can we teach LLMs data cleaning and curation as teaching coding?

Question 5. How can we connect private (cooperation, personal) data w. public data?

Trend 3. Data Creation Revol.—Second Brain



Manual data creation



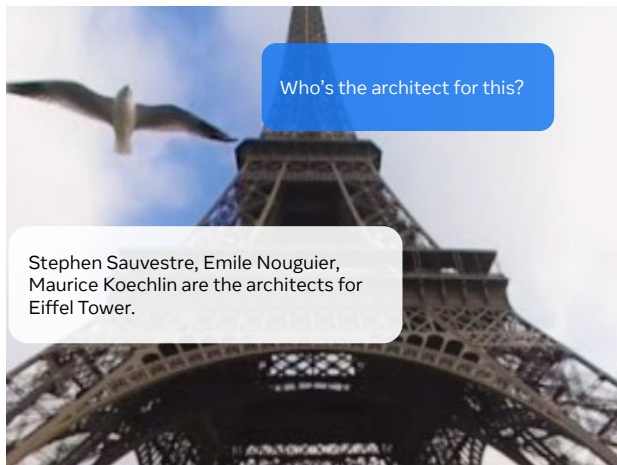
Memex (1945)



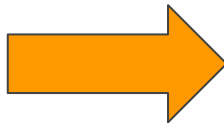
Automatic data recording

Question. If everyone records what's happening from her own perspective, can we piece together recorded data to cover all valuable data?

Trend 4. Data Access Revolution—No-query Info Access



QA



Question 1. When to provide information?

Question 2. What information to provide?

Question 3. How to personalize?

Do We Manage Data?



Do Data Manage Themselves?



Info at the tip of your nose



Data create themselves
Data structure themselves
Data integrate themselves
Data provide themselves

A stylized illustration of a robot with a colorful, segmented body and glowing blue eyes, sitting at a desk and reading an open book. The robot is wearing a dark suit jacket and a white shirt. The background is a vibrant, swirling blue and yellow pattern, reminiscent of J.M.W. Turner's 'Rain, Steam, and Great Bridge' or a modern interpretation of 'The Starry Night'. The robot is holding a pen over the book, and a small white cup sits on the desk. The overall style is a blend of classic art and modern digital art.



Pre-train

Question 3. How to know when to rely on what?

Post-train

Post-train

Symbolic Tail or Dynamic Knowledge

