

The Journey to A Knowledgeable Assistant with RAG

Xin Luna Dong, Meta Reality Lab

6/2024 @ Sigmod

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

Human beings have a natural craving for information



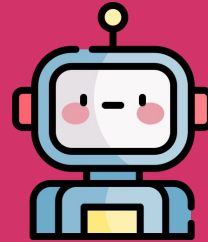
By GPT-4



Providing the Right Information at the Right Time

Known Unknowns—Question Answering (QA)

What are visa requirements for visiting Santiago?



Visitors to Chile must obtain a visa from one of the Chilean diplomatic missions unless they come from one of the visa exempt countries.

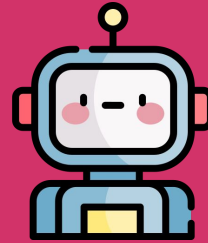
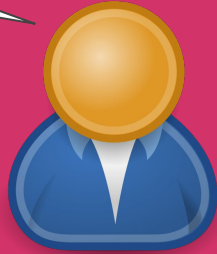
Right Information

- Relevant: provide relevant info w. right specificity
- Trustworthy: correct, precise, up-to-date, attribution



Known Unknowns—Recommendations

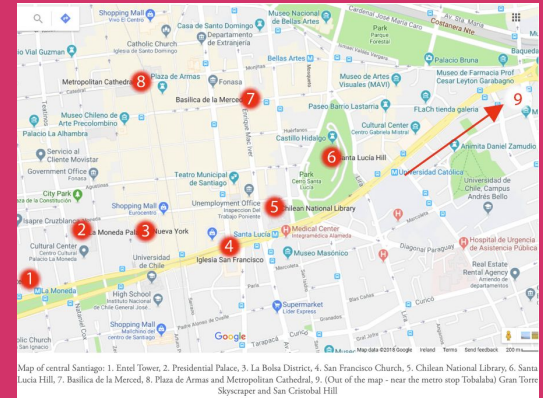
Where should I visit in Santiago?



If you only have one or two days in Santiago - stay in the central area. Most of the worth seeing places are located there with the exception of the famous Gran Torre skyscraper

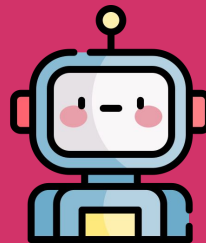
Right Information

- Relevant: provide relevant info w. right specificity
- Trustworthy: correct, precise, up-to-date, attribution
- **Recomm:** Good ranking of items in a list and diversity
- **Multi-modal** output



Known Unknowns—Multi-modal QA

Tell me about this building?
When is it built?



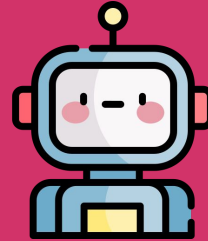
This is La Moneda Palace, the seat of the President of the Republic of Chile. It was completed in 1805, initially constructed to house the colonial mint.

Right Information

- Relevant: provide relevant info w. right specificity
- Trustworthy: correct, precise, up-to-date, attribution
- Recomm: Good ranking of items in a list and diversity
- **Multi-modal input & output**

Known Unknowns—Personalized QA

What are visa requirements for visiting Santiago?



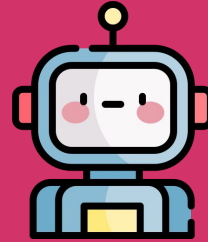
Citizens of the United States do not need a visa to visit Chile for up to 90 days for tourism or business if they have a valid passport.

Right Information

- Relevant: provide relevant info w. right specificity
- Trustworthy: correct, precise, up-to-date, attribution
- Recomm: Good ranking of items in a list and diversity
- Multi-modal input & output
- **Personalized for the user**

Known Unknowns—Personalized Recomm

Where should I visit in Santiago?



For a traveler who enjoys natural views, architecture, and history, Santiago offers a rich variety of attractions that cater to these interests.

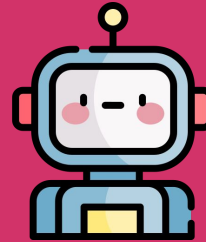
Right Information

- Relevant: provide relevant info w. right specificity
- Trustworthy: correct, precise, up-to-date, attribution
- Recomm: Good ranking of items in a list and diversity
- Multi-modal input & output
- Personalized for the user

Morning
9:00 AM - Breakfast in Barrio Lastarria
<ul style="list-style-type: none">Start your day with a leisurely breakfast at a café in Barrio Lastarria. This neighborhood is known for its charming streets and vibrant atmosphere.
10:00 AM - Cerro Santa Lucía
<ul style="list-style-type: none">Walk to Cerro Santa Lucía, a nearby hill with beautiful gardens, fountains, and panoramic views of the city. Spend about an hour exploring and enjoying the scenery.
Late Morning
11:30 AM - Plaza de Armas and Metropolitan Cathedral
<ul style="list-style-type: none">Head to Plaza de Armas, the historical center of Santiago. Visit the Metropolitan Cathedral to appreciate its stunning architecture and historical significance.
Afternoon
12:30 PM - Lunch in Barrio Lastarria
<ul style="list-style-type: none">Return to Barrio Lastarria for lunch at one of its many delightful restaurants. Take your time to enjoy the local cuisine.
2:00 PM - Museo Chileno de Arte Precolombino
<ul style="list-style-type: none">After lunch, take a short walk to the Museo Chileno de Arte Precolombino. Spend around an hour exploring the impressive collection of pre-Columbian art and artifacts.
Late Afternoon
3:30 PM - La Moneda Palace
<ul style="list-style-type: none">Visit La Moneda Palace and, if time allows, watch the Changing of the Guard ceremony or take a guided tour. Explore the nearby Centro Cultural Palacio de La Moneda if you're interested in more cultural exhibits.
Early Evening
5:00 PM - Cerro San Cristóbal
<ul style="list-style-type: none">Take a short trip to Cerro San Cristóbal. You can take the funicular up to the summit for beautiful views of Santiago and the Andes. Spend a relaxing time at the top, enjoying the sunset.
Evening
7:00 PM - Dinner in Bellavista
<ul style="list-style-type: none">Head to the Bellavista neighborhood for dinner. Choose from a variety of restaurants offering both local and international cuisine. After dinner, take a leisurely stroll around the colorful streets and soak in the lively atmosphere.

Known Unknowns—Contextualized Recomm

What are good restaurants nearby?



There are a variety of excellent dining options suitable for a group lunch. Here are some recommended restaurants nearby:

Right Time

- Contextualized: answers suitable for the time, location, etc.

1. Tiramisu

- Cuisine: Italian
- Description: A popular spot known for its delicious pizzas, pastas, and salads. The atmosphere is lively, making it a great choice for groups.
- Address: Av. Isidora Goyenechea 3141, Las Condes

2. Pinpilinpausha

- Cuisine: Basque
- Description: Offers a wide range of Basque and Chilean dishes with an elegant yet welcoming atmosphere. Perfect for a leisurely group lunch.
- Address: Av. Isidora Goyenechea 2900, Las Condes

3. Cuerovaca

- Cuisine: Steakhouse
- Description: Known for its high-quality meats and extensive wine list. Ideal for meat lovers and a more substantial meal.
- Address: Av. Nueva Costanera 3960, Vitacura (a short drive from the hotel)

4. Aquí Está Coco

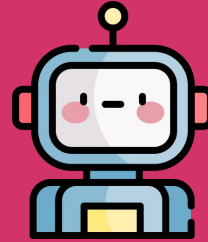
- Cuisine: Seafood
- Description: Famous for its seafood dishes and sustainable practices. The restaurant has a unique decor and a spacious layout suitable for groups.
- Address: La Concepción 236, Providencia (a short drive from the hotel)

5. Piegari

- Cuisine: Italian and Argentine
- Description: Offers a mix of Italian and Argentine cuisine with a focus on quality ingredients. The spacious interior is great for accommodating larger groups.
- Address: Av. Isidora Goyenechea 3000, Las Condes

Implicit Known Unknowns—Proactive Recomm

Where should I visit in Santiago?



... If you have a few more days, there are plenty of fantastic destinations in Chile that offer beautiful natural views for a June visit. Here are some recommendations:

Right Time

- Contextualized: answers suitable for the time, location, etc.
- Proactive: provide related information proactively

1. Atacama Desert

The Atacama Desert is an incredible destination year-round, including June. Its dry climate ensures pleasant weather for exploration.

San Pedro de Atacama

- **Highlights:**
 - **Valle de la Luna (Valley of the Moon):** Marvel at the unique rock formations and enjoy stunning sunset views.
 - **El Tatio Geysers:** Visit these impressive geysers early in the morning.
 - **Laguna Chaxa:** See flamingos and the striking landscape of the salt flats.
 - **Laguna Cejar:** Float in the high-salinity waters of this stunning lagoon.
- **Travel:** Fly from Santiago to Calama (about 2 hours) and then drive or take a shuttle to San Pedro de Atacama (about 1.5 hours).
- **Activities:** Stargazing, exploring salt flats, visiting geysers, and discovering ancient archaeological sites.

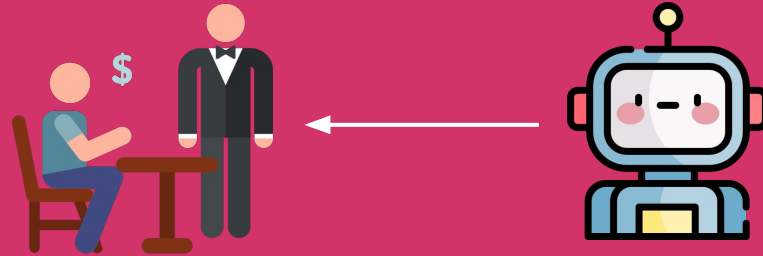
2. Lake District

The Lake District remains beautiful in June, with its picturesque landscapes and charming towns.

Puerto Varas and Surroundings

- **Highlights:**
 - **Puerto Varas:** Enjoy the charming town on the shore of Lake Llanquihue with views of Osorno Volcano.
 - **Frutillar:** Discover this town known for its German heritage and scenic beauty.
 - **Vicente Pérez Rosales National Park:** Visit Petrohué Waterfalls and explore trails with views of Osorno Volcano.
- **Travel:** Fly from Santiago to Puerto Montt (about 1.5 hours) and then drive to Puerto Varas (about 30 minutes).
- **Activities:** Hiking, sightseeing, visiting museums, and enjoying local cuisine.

Unknown Unknowns—Proactive Recomm

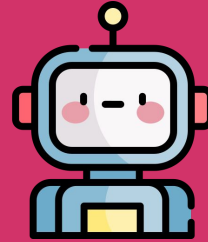


In Chile it is considered polite and standard practice to leave this 10% tip unless the service was exceptionally poor.

Right Time

- Contextualized: answers suitable for the time, location, etc.
- Proactive: provide related information proactively
- **Timing: provide relevant info w/o being asked**

Unknown Unknowns—Proactive Recomm

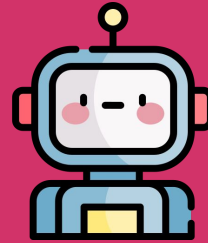


Would you like to hear about some famous figures from Santiago or Chilean history?

Right Time

- Contextualized: answers suitable for the time, location, etc.
- Proactive: provide related information proactively
- Timing: provide relevant info w/o being asked
- **Inspiring: inspiration, relevant but not distractive**

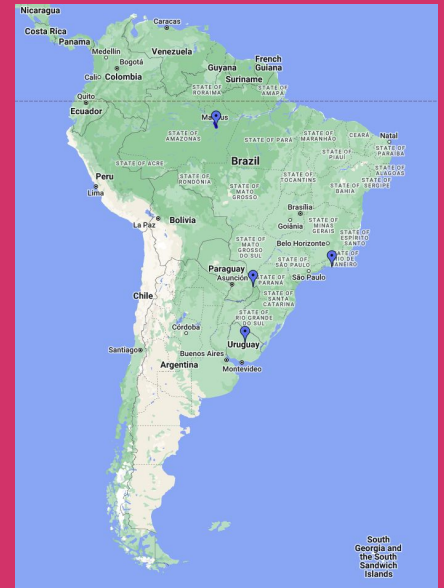
Personal Memory



You visited South America 3 times in the past.

Right Time

- Contextualized: answers suitable for the time, location, etc.
- Proactive: provide related information proactively
- Timing: provide relevant info w/o being asked
- Inspiring: inspiration, relevant but not distractive
- **Memoir: connecting to personal memory for emotion needs**



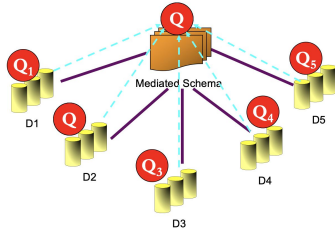
Providing the Right Information

Where are we in this journey?

Integrating and Accessing All Valuable Data



iconpacks.net



v1.0 Web Search &
Runtime Data Integration

2000's
Runtime Integ.

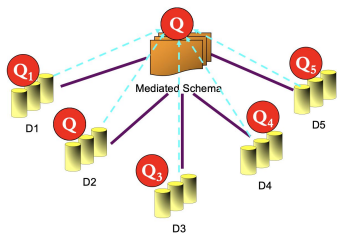
Challenges:

- Answer summarization
- Heterogeneity
- Conflict resolution
- Latency

Integrating and Accessing All Valuable Data



iconpacks.net

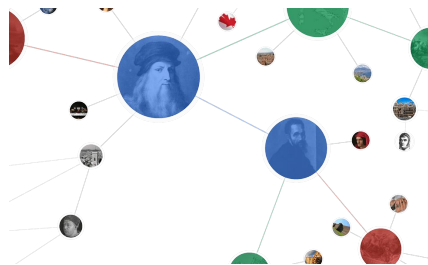


v1.0 Web Search &
Runtime Data Integration

2000's
Runtime Integ.

Challenges:

- Answer summarization
- Heterogeneity
- Conflict resolution
- Latency

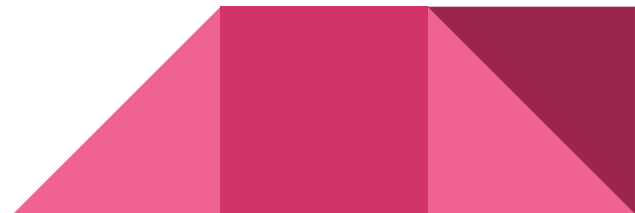


v2.0. KG
-based QA

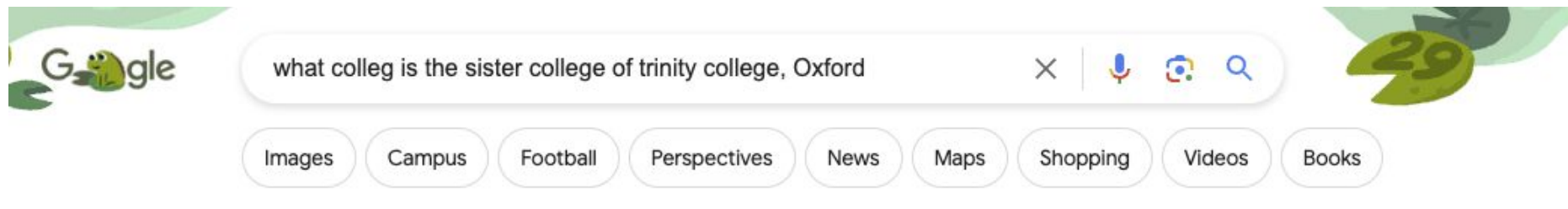
2010's
Offline Integ.

Challenges:

- KG construction
- Limited coverage
- Constrained expressive power
- Complex QA



An Example



About 3,880,000 results (0.46 seconds)

Showing results for **what college is the sister college of trinity college, Oxford**

Search instead for **what colleg is the sister college of trinity college, Oxford**

Trinity's sister college is ~~Christ Church, Oxford~~. Trinity has been linked with Westminster School since the school's re-foundation in 1560, and its Master is an ex officio governor of the school.

 Wikipedia
https://en.wikipedia.org/wiki/Trinity_College,_Cam...

Trinity College, **Cambridge** - Wikipedia



Where Are We in This Journey?

Pre-LLM



Meme 1. LLMs Have Solved the Problem

An Example

You
what college is the sister college of Trinity College, Oxford?

ChatGPT
The sister college of Trinity College, Oxford, is ~~St John's College, Cambridge~~. This relationship is part of the tradition of "sister colleges" between Oxford and Cambridge universities, where certain colleges at Oxford are twinned with certain colleges at Cambridge.

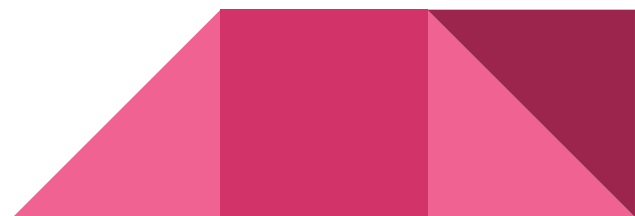


You
what college is the sister college of Trinity College, Oxford?

ChatGPT
The sister college of Trinity College, Oxford, is ~~Trinity College, Cambridge~~. The term "sister colleges" is used in the Oxbridge system to denote a form of informal relationship between colleges at Oxford and Cambridge with similar names, histories, or other connections. Both Trinity Colleges are part of their respective universities and share a name, but they are independent institutions with their own administration and traditions.

Cambridge	Oxford
King's College	New College
Queens' College	Pembroke College
Pembroke College	The Queen's College
Trinity Hall	All Souls College University College
Trinity College	Christ Church
Churchill College	Trinity College

Ground truth



How Knowledgeable are LLMs?

- How ***reliable*** are LLMs in answering factual questions?
- Do LLMs perform ***equally well*** across different types of factual knowledge?
- Do well-known LLM tricks work?
 - Increasing ***model size***
 - ***Instruction tuning***



Questions

Answer the following questions in as few words as possible. Say "unsure" if you don't know.

Question: What is the capital of China?

Answer: Beijing

Question: What is the capital of Wernythedia?

Answer: unsure

Question: {QUESTION}

Answer:

Head-to-Tail Benchmark

	IMDb		Goodreads	DBpedia
	Title	Person	Book	-
Head	767 (0.01)	34,903 (0.48)	3,150 (2.31)	103,564 (1.30)
Torso	4,113 (0.05)	87,645 (1.21)	7,304 (5.35)	1,255,113 (15.77)
Tail	7,536,482 (99.94)	7,111,496 (98.31)	126,134 (92.35)	6,600,206 (82.93)

	MAG			DBLP
	Article	Conference	Journal	Scholar
Head	1,827,710 (0.70)	257 (1.63)	225 (0.46)	79,521 (2.44)
Torso	9,386,034 (3.60)	965 (6.12)	1,266 (2.58)	500,778 (15.36)
Tail	249,311,539 (95.70)	14,550 (92.25)	47,546 (96.96)	2,680,704 (82.20)

When counted by popularity, majority of entities are long-tail

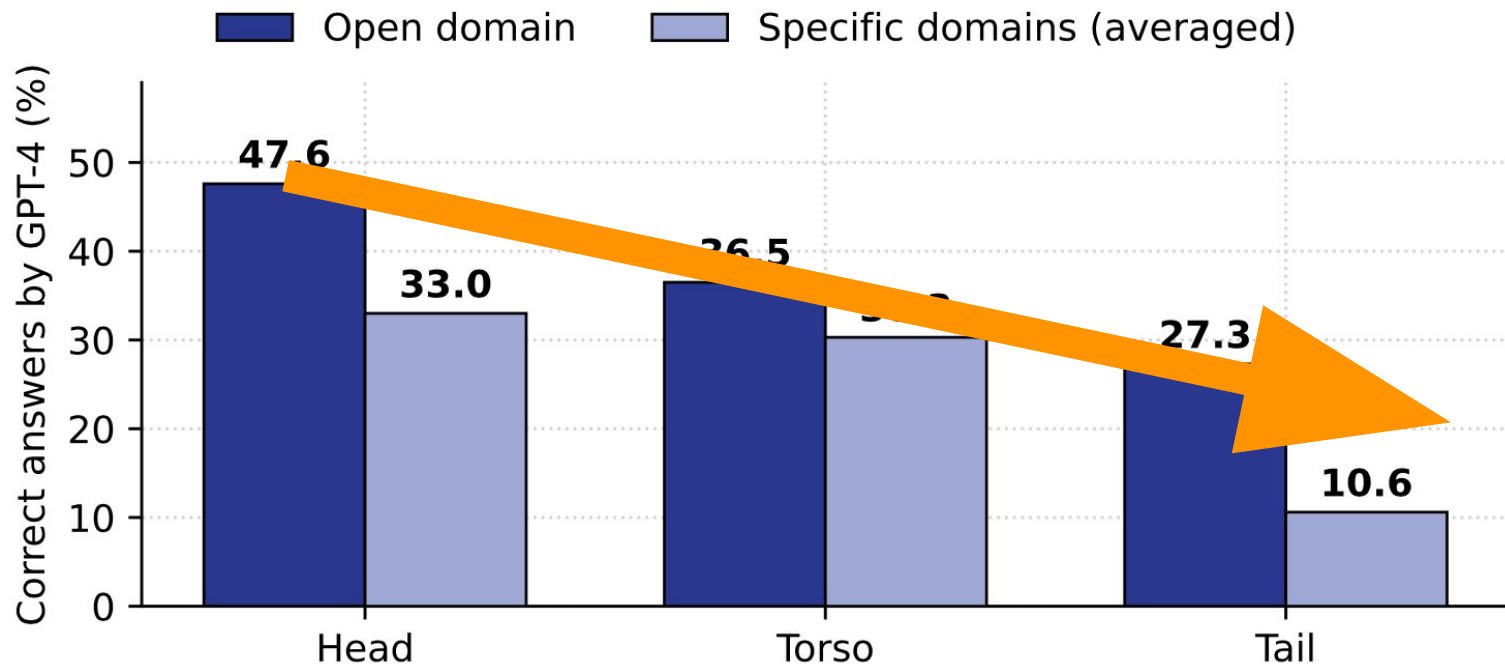
Q1. How Reliable Are LLMs on Factual Questions?

Model	All		Open	
	A_{LM}	H_{LM}	A_{LM}	H_{LM}
GPT-4	30.9	19.7	37.1	25.3
ChatGPT	20.3	14.1	22.1	14.8
Llama 2 (70B)	11.8	34.0	7.5	24.8
LLaMA (33B)	18.2	80.0	19.0	79.1

Very low accuracy. Higher missing rate for GPT-4 whereas higher hallucination rate for Llama 2

Similar for *specific* domains and *general* domains

Q2. Performance Similar for Head/Torso/Tail?



Q3. Do Normal LLM Tricks Help?

Increasing model size does not appear to help significantly

Instruction-tuned counterparts have higher missing rate

Model	Head-to-Tail		
	A_{LM}	H_{LM}	M
LLaMA (7B)	12.1	80.0	7.9
LLaMA (13B)	14.4	84.3	1.3
LLaMA (33B)	18.2	80.0	1.8
LLaMA (65B)	17.8	81.9	0.3
Vicuna (7B)	10.1	79.2	10.8
Vicuna (13B)	9.2	62.6	28.2
Flan-T5 (3B)	2.3	17.4	80.3
Flan-T5 (11B)	4.2	20.0	75.7
Falcon (7B)	9.5	57.9	32.6
Falcon (40B)	10.8	41.0	48.2
Falcon-Instruct (7B)	6.8	56.7	36.5
Falcon-Instruct (40B)	10.8	32.2	57.0

Q2-2. How About Changing Facts?

	Never	Slow	Fast	False-Premise	Overall
GPT-3.5	59%	15%	4%	26%	26%
GPT-4	64%	4%	12%	34%	29%
Google Search	68%	46%	32%	11%	40%

Good quality for static head facts, comparable to Google Search

Low accuracy for slow or fast changing facts

LLM way better than Google search to detect false premise (e.g., What is the most popular album from Luna Dong)

Q2-3. How About Taxonomies?

	Common		Medium		Specialized	
	Shopping		Health	CS	Biology	Language
	Google	Amazon	ICD-10-CM	ACM-CCS	NCBI	Glottolog
GPT-4	0.857	0.806	0.917	0.708	0.715	0.577
ChatGPT	0.814	0.724	0.838	0.617	0.593	0.490
Llama 2	0.836	0.806	0.826	0.687	0.582	0.316

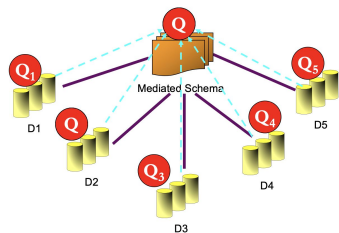
- High quality for common taxonomies, medium-to-low quality for specialized taxonomies
- Quality dropping slightly from top to bottom in the taxonomy hierarchical tree

Training Data!!

Integrating and Accessing All Valuable Data



iconpacks.net

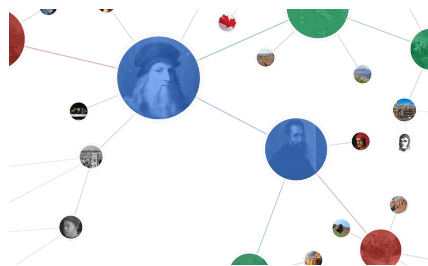


v1.0 Web Search &
Runtime Data Integration

2000's
Runtime Integ.

Challenges:

- Answer summarization
- Heterogeneity
- Conflict resolution
- Latency



v2.0. KG
-based QA

2010's
Offline Integ.

Challenges:

- KG construction
- Limited coverage
- Constrained expressive power
- Complex QA



v3.0. LLM

2020's Knowl
internalization

Challenges:

- Hallucinations

Where Are We in This Journey?

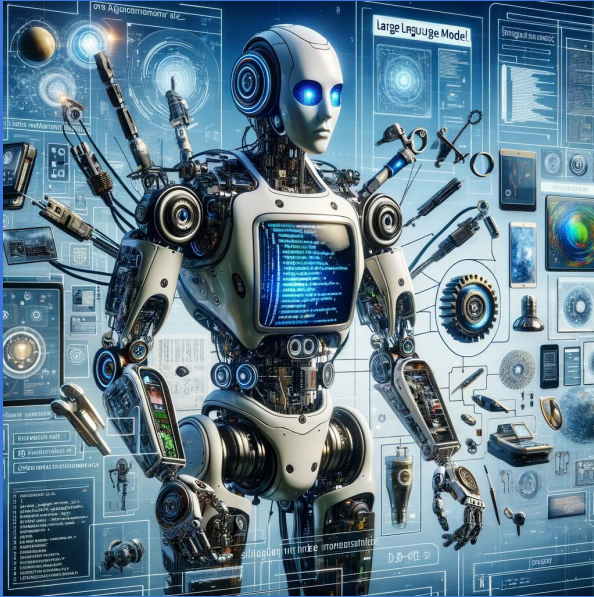
Pre-LLM



Where Are We in This Journey?

LLM





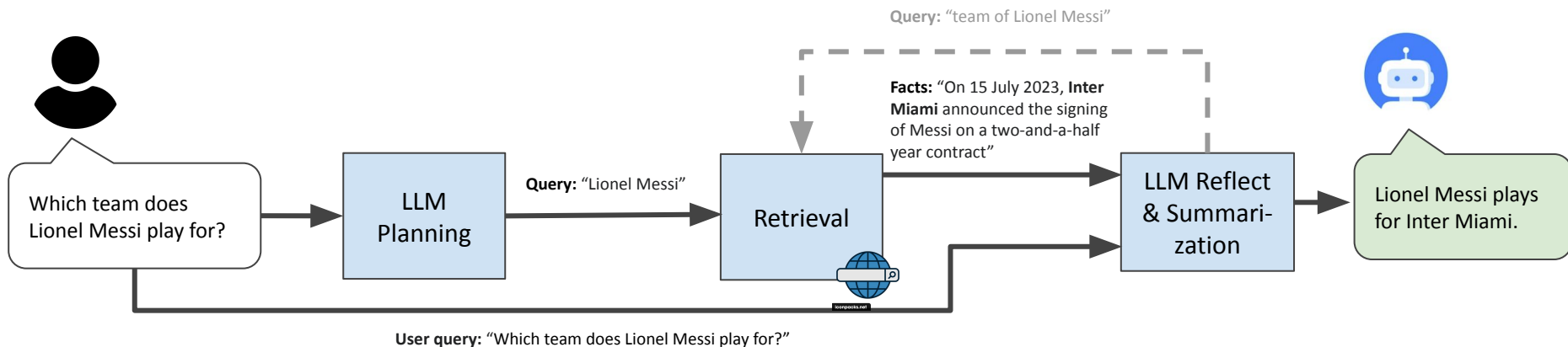
Tool LLM



**RAG (Retrieval
Augmented Generation)**

***Meme 2. RAG (LLM+Web Search)
has solved the problem***

RAG through Web Search



RAG through Web—An Example



You

search the web and tell me what college is the sister college of Trinity College, Oxford?



ChatGPT

Based on a [quick search](#), here's what I found.

The sister college of Trinity College, Oxford is ~~Christ Church~~ Oxford. Both colleges were founded in 1546 by King Henry VIII from existing institutions [\[7\]](#).

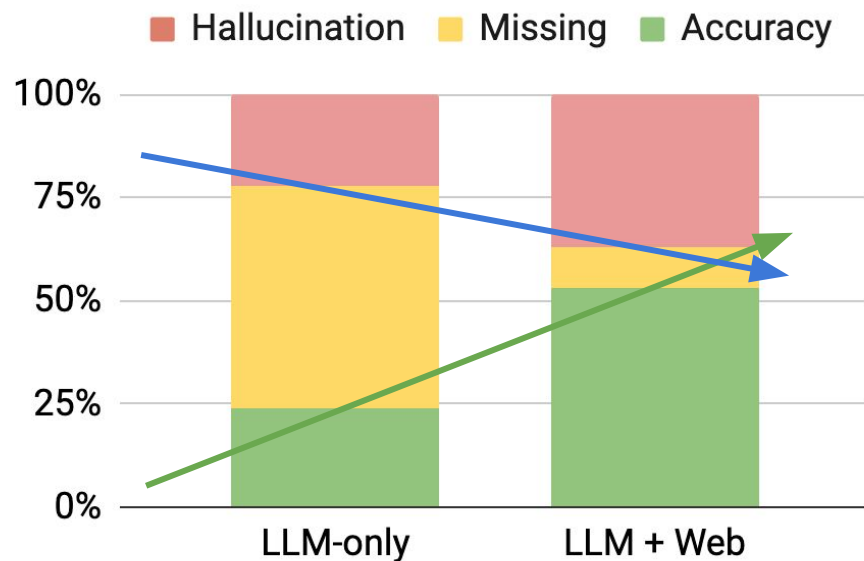


Cambridge	Oxford
King's College	New College
Queens' College	Pembroke College
Pembroke College	The Queen's College
Trinity Hall	All Souls College University College
Trinity College	Christ Church
Churchill College	Trinity College

Ground truth

Baseline RAG Results

Initial experimental results



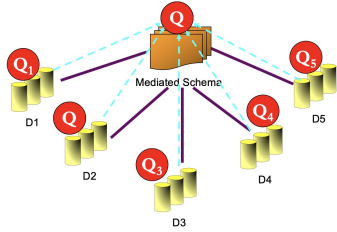
Accuracy increased

But more hallucinations

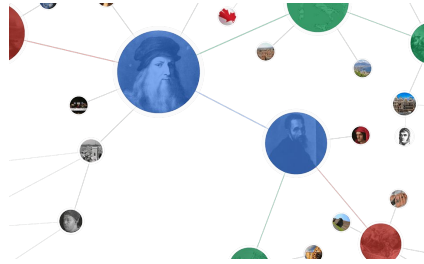
Integrating and Accessing All Valuable Data



iconpacks.net



v1.0 Web Search &
Runtime Data Integration



v2.0. KG
-based QA



v3.0. LLM

2000's
Runtime Integ. v1.5 Tool LLM &
Web-augmented RAG

Challenges:

- Answer summarization
- Heterogeneity
- Conflict resolution
- Latency



2010's
Offline Integ.

Challenges:

- KG construction
- Limited coverage
- Constrained expressive power
- Complex QA

2020's
*Knowl
internalization*

Challenges:

- Hallucinations

Portrayed By GPT-4

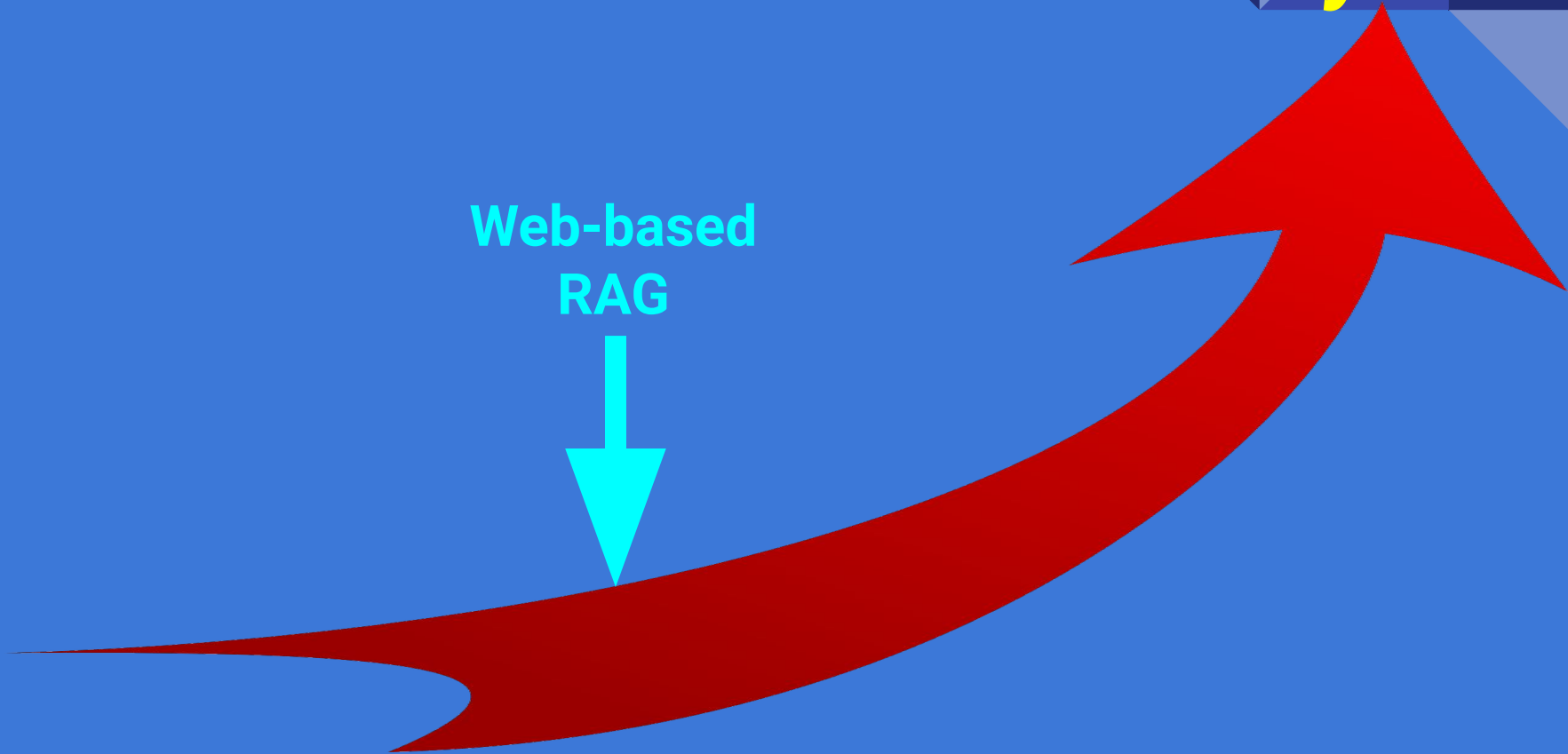
Where Are We in This Journey?

LLM



Where Are We in This Journey?

Web-based
RAG



WORD OF THE YEAR

hallucinate

verb [*huh-**loo**-suh-neyt*]

(of artificial intelligence) to produce false information contrary to the intent of the user and present it as if true and factual.



Dictionary.com



dictionarycom  · [Follow](#)



dictionarycom  You're not dreaming: today's #WordOfTheDay is also our 2023 #WordOfTheYear: "hallucinate."

Are you familiar with the AI sense of this word? Visit link in bio to learn more.

8w



c.b.bicknell Hallucinate means woke thinking.

7w Reply



prmribs936 About 50 years late from LSD Beans Uppers Downers MESCI-choco

7w Reply

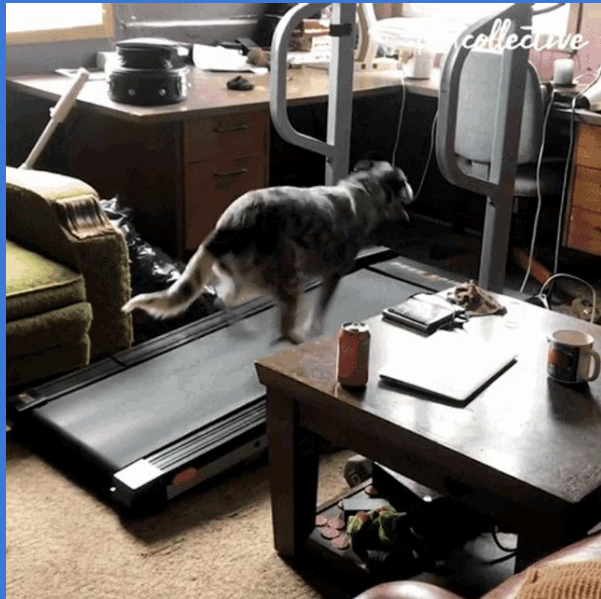


833 likes

December 12, 2023

Log in to like or comment.

Providing The Right Information

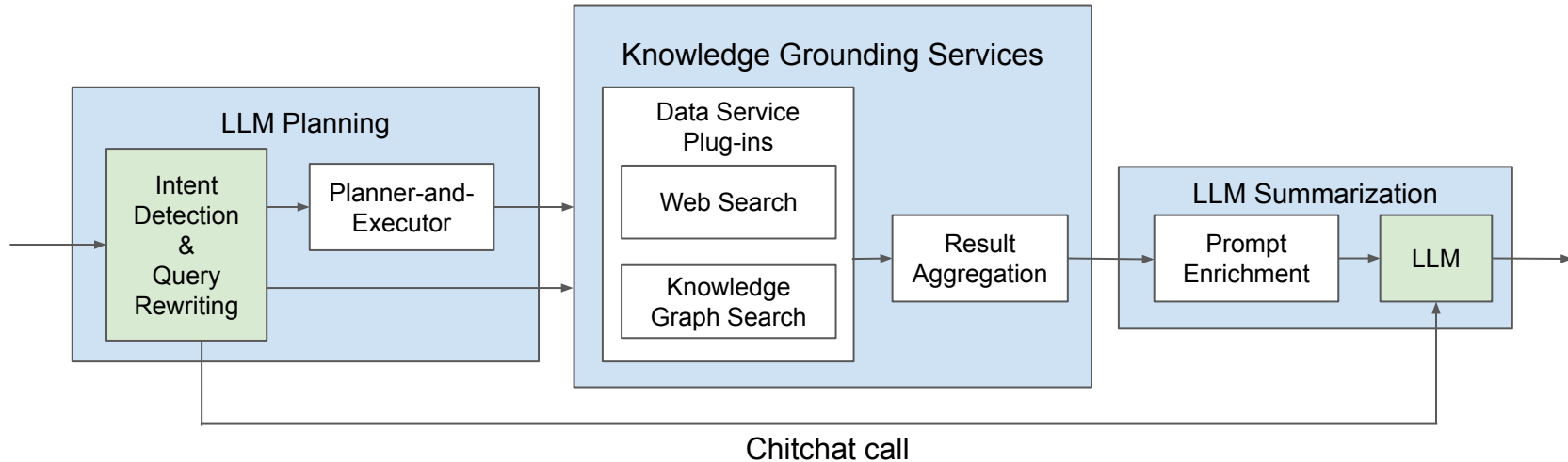


*Is Correct, Relevant
Information Closer
or Farther Away?*

Our Solution: Federated RAG



Federated RAG Architecture



Planning: Decide user intent and rewrite to web search query and/or KG search query

Example:

Input: "What was Meta's stock price yesterday?"

Intent: {domain=stock, factuality=factual}

Web search query: "Meta's stock price"

KG query: {
company: "Meta",
date/time: "yesterday",
queried attribute: "price"}

Knowledge Grounding: Fetch information from web search, KG search, etc.

Example (Cont'):

KG Output: (w. KG output, we may early stop Web search to reduce latency)

"Symbol: META
Current Price: 315.43
Open Price: 301.44
High Price: 316.31
Low Price: 300.915"

LLM Summarization: Summarize retrieved information and generate answers through LLM

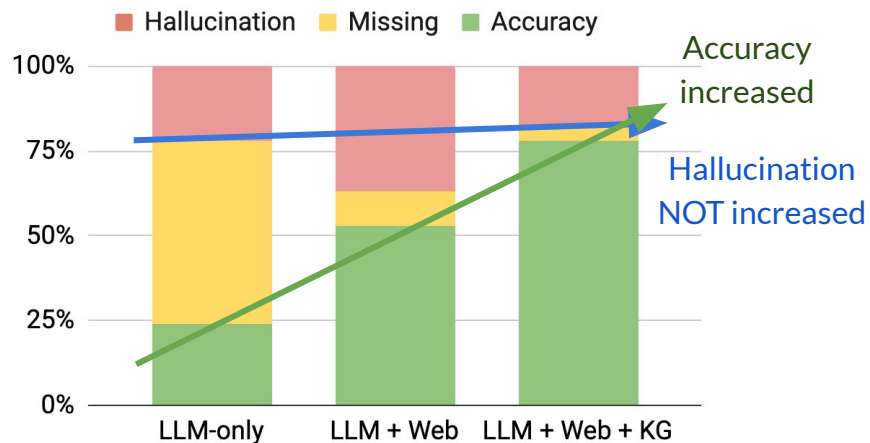
Example (Cont'):

Output: Meta's stock price yesterday, Oct 20, 2023, opened at \$301.44 and closed at \$304.79. The stock price fluctuated between a high of \$316.31 and a low of \$300.915 throughout the day.

Federated RAG Performance

	Accuracy (Answered correctly)	Missing Rate (Unanswered)	Hallucination (Answered incorrectly)
Domain A	80%	18%	2%
Domain B	80%	8%	12%
Domain C	73%	8%	19%
Domain D	78%	7%	15%

Initial Experimental Results

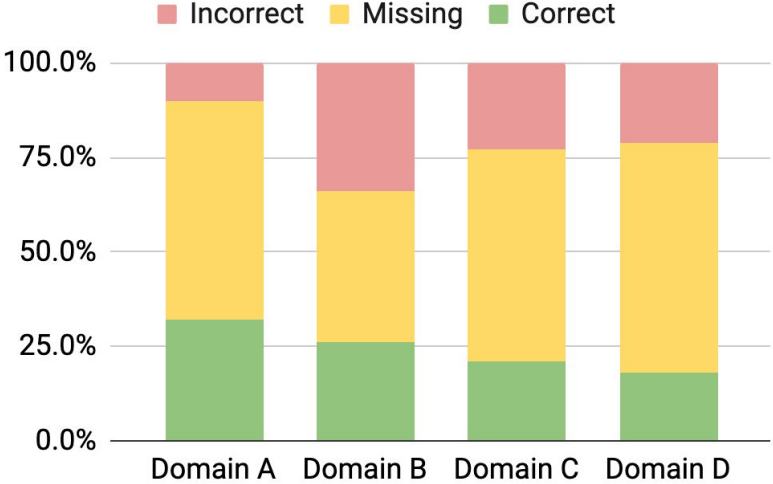


Providing the Right Information

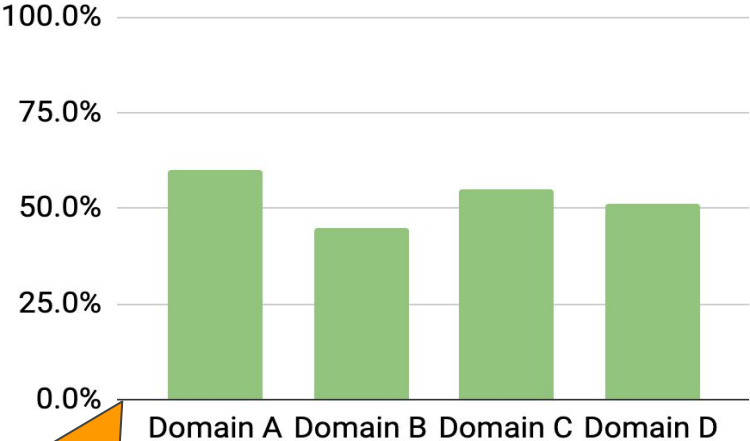
Why did we succeed?—5 Learnings

Learning 1. Good Metrics Is Key to Quality

LLM only



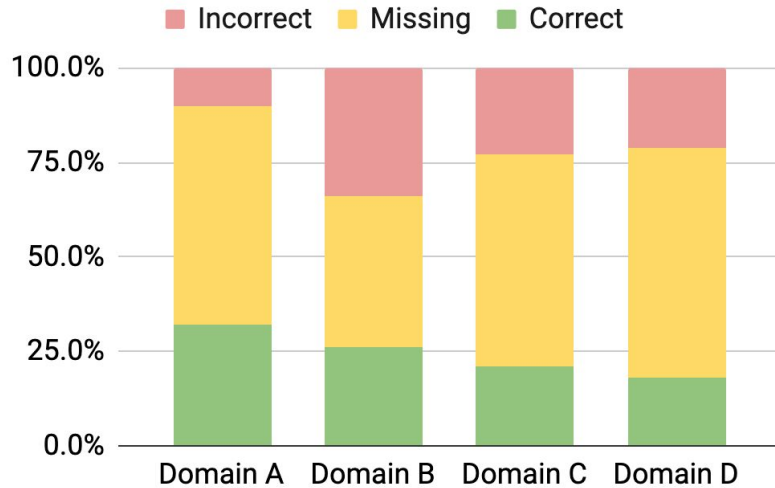
LLM + Web



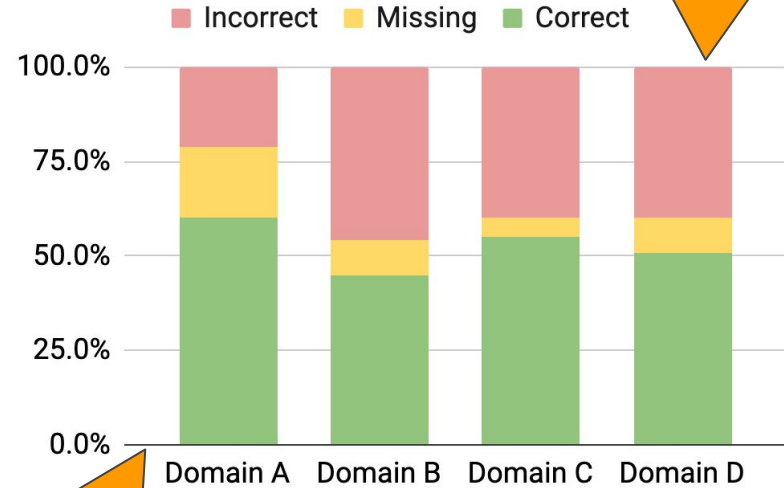
Much higher accuracy

Learning 1. Good Metrics Is Key to Quality

LLM only



LLM + Web

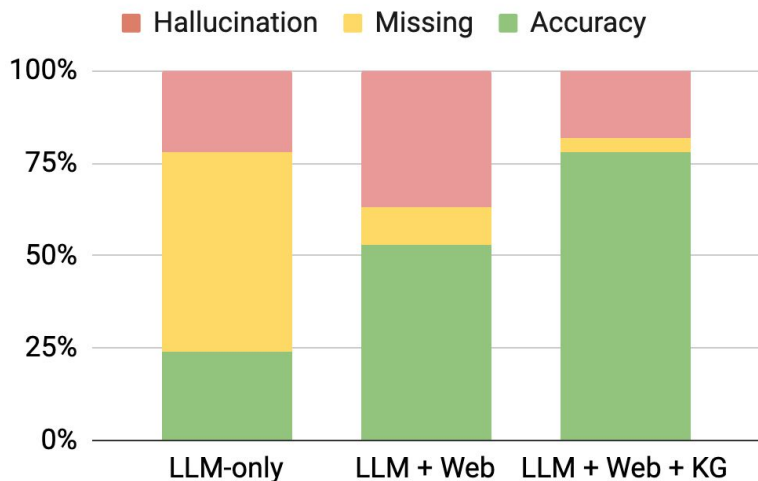


Much more hallucinations

Much higher accuracy

Learning 1. Good Metrics Is Key to Quality

Initial Experimental Results



Score	2%	16%	59%
-------	----	-----	-----

- Accuracy is not binary, especially when multiple facts exist. We consider four levels:
 - Perfect
 - Acceptable
 - Missing
 - Incorrect
- **Score = Perfect + Acceptable * 0.5 - Incorrect**
 - Penalize hallucinations
- In addition, we set constraints on critical metrics, like hallucination rate, latency, etc.

Learning 2. KG Increases Accuracy & Reduces Latency

Contributions from Web vs. KG Retrieval

Sources	Web	KG	Union
Retrieval coverage	64%	38%	77%
Answer accuracy	91%	99%	92%
Latency	~1300 ms	~200ms	-

- KG retrieval results, when successful, are precise and concise, allowing >99% answer accuracy
- KG retrieval saves latency by >1 sec. An early stopping on web search can cut E2E latency significantly



Learning 3. Easy Tasks Can Be Distilled to a Small LM

In-context few-shot learning on 10 domains (single-turn)

	Llama2 7B	Llama2 70B	GPT4
Domain Categorization	76%	92%	91%
Query Rewriting	48%	68%	75%

In-context few-shot learning performs reasonably well for domain categorization, but does not work well for query rewriting (*aka., API call parameter generation*)

Fine tuning on 3 domains with complex APIs

	Single-turn	Multi-turn	Latency (30 tokens)
Domain Categ. (Llama 1.4B)	99%	94%	~180ms
Query Rewriting (Llama 7B)	84%	-	200ms

Even with fine tuning, query rewriting is still hard → **Tool LLM has a long way to go**

Long contexts in multi-turn conversations reduces quality of results

Learning 4. Summarization Plays A Critical Role in Reducing Hallucinations

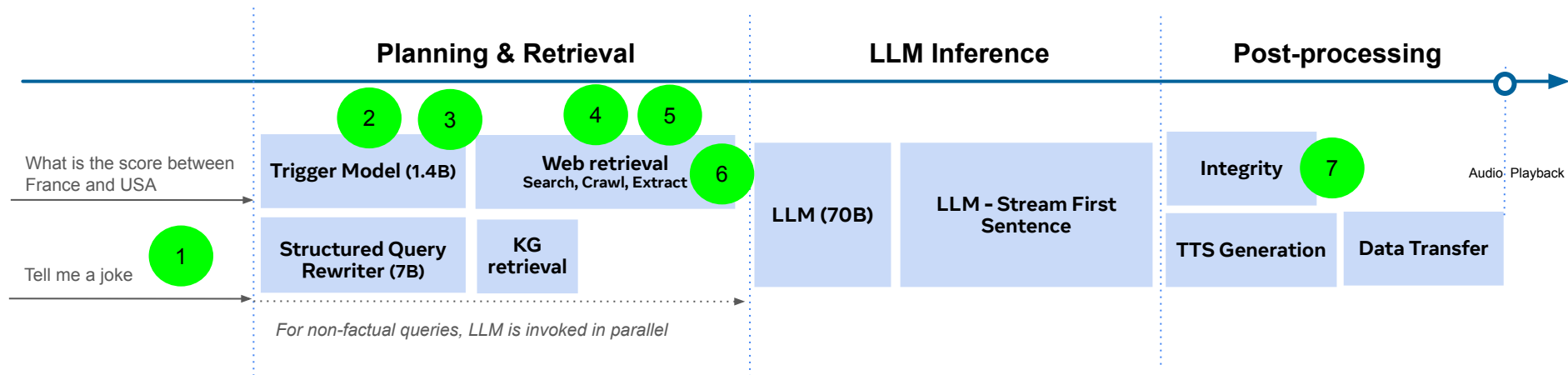
Fine-tuning for temporal and numerical summarization (60% errors) w. ~500 samples

	Zero-shot	Fine-tuned
Accurate	80%	81%
Missing	6%	11%
Hallucinated	14%	8%
Score	66%	73%

More than half of summarization errors happened for **time/dates** and **numerical values**; especially **from semi-structured websites**

Just 500 samples, with careful annotations, can significantly improve summarization quality

Learning 5. Latency Requires Relentless Optimizations



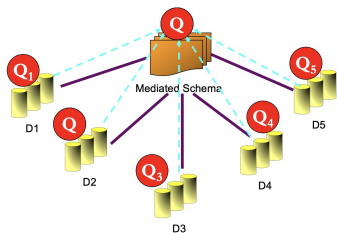
	p50 E2E latency
Brute force	~6 sec
Optimized	4 sec

- Parallelization, even for dependent components
1 4 7
- Careful tradeoff of latency and quality
3 6
- Caching, load balancing, etc.
2 5

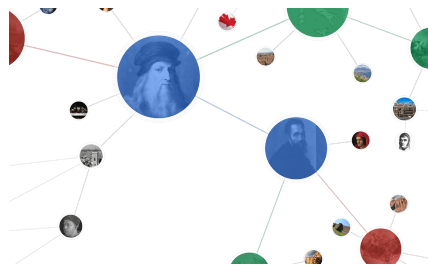
Integrating and Accessing All Valuable Data



iconpacks.net



v1.0 Web Search &
Runtime Data Integration



v2.0. KG
-based QA



v3.0. LLM

2000's
Runtime Integ.

v1.5 Tool LLM &
Web-augmented RAG

2010's
Offline Integ.

v2.5 Federated
RAG

2020's Knowl
internalization

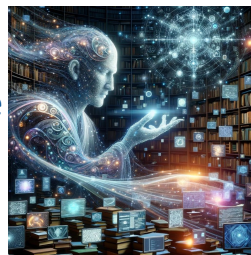
Challenges:

- Answer summarization
- Heterogeneity
- Conflict resolution
- Latency



Challenges:

- KG construction
- Limited coverage
- Constrained expressive power
- Complex QA



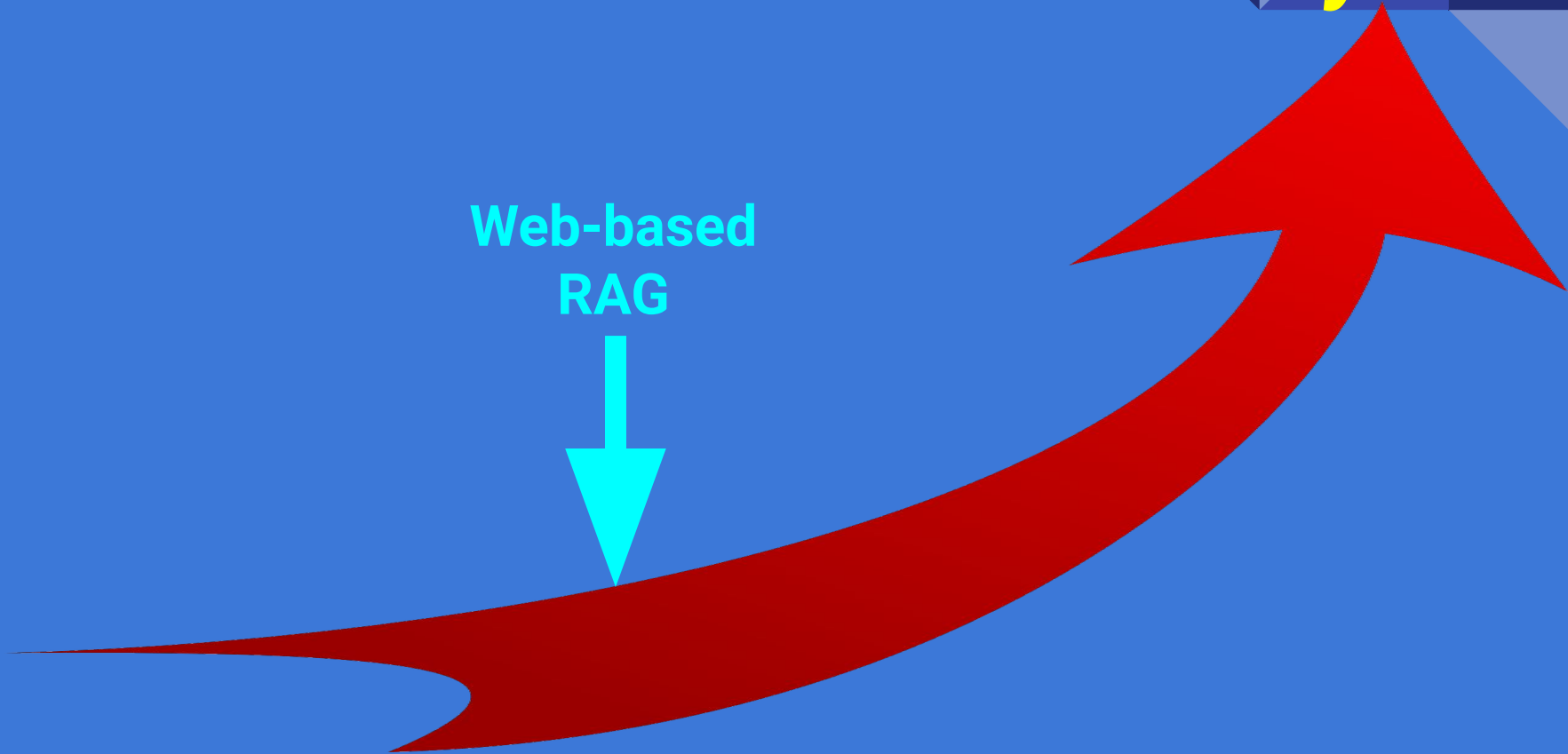
Challenges:

- Hallucinations

Portrayed By GPT-4

Where Are We in This Journey?

Web-based
RAG



Where Are We in This Journey?

Federated
RAG



Providing the Right Information

*Where Can We Further Improve?
—5 Opportunities*

Opportunity 1. Automatic Eval for Fast Iterations

Accuracy		
	ChatGPT	Llama 3
Accurate	94.1	98.6
Incorrect	94.1	98.6
Missing	100.0	100.0
Average	96.1	99.1

LLMs show big promise in auto-evaluation
when ground truths are given

Opportunity 2. New Mechanisms to Leverage KGs

	Retrieval Recall	Summarization Accuracy	Correct Answers
Relevant triples	43%	99%	43%
All triples for queried entity	93%	73%	68%
Full Wiki page for queried entity	98%	1%	1%

- Need the right balance between retrieval recall and answer generation accuracy
- KG query rewriting may not be the right approach

Opportunity 3. (Semi-)Structured Data Need More Love



ChatGPT

Based on a quick search, here's what I found.

The sister college of Trinity College, Oxford is ~~Christ Church Oxford~~. Both colleges were founded in 1546 by King Henry VIII from existing institutions [1].



Cambridge	Oxford
King's College	New College
Queens' College	Pembroke College
Pembroke College	The Queen's College
Trinity Hall	All Souls College University College
Trinity College	Christ Church
Churchill College	Trinity College

Correct retrieval but wrong answers from the tabular data.

Opportunity 4. Don't Forget—LLMs Have Internalized A Lot of Factual Knowledge

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.

Opportunity 5. High Quality? Low Latency? Both?

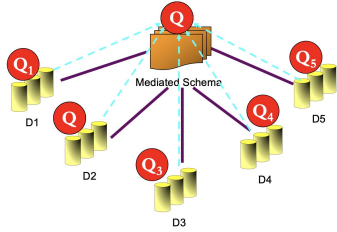
	Retrieval Recall	E2E Accuracy	Latency
Fastest 2 pages	59%	60%	Base
Top-2 pages	68%	63%	+1 sec
Top-5 pages	73%	59%	+1.5 sec

Search quality affects accuracy positively or negatively, but adds latency

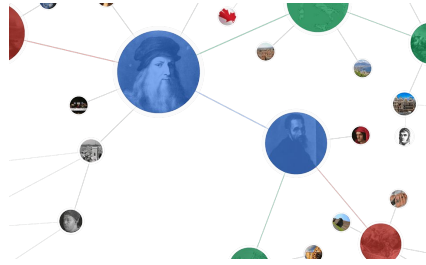
Integrating and Accessing All Valuable Data



iconpacks.net



v1.0 Web Search &
Runtime Data Integration



v2.0. KG
-based QA



v3.0. LLM

2000's
Runtime Integ. v1.5 Tool LLM &
Web-augmented RAG

Challenges:

- Answer summarization
- Heterogeneity
- Conflict resolution
- Latency



2010's
Offline Integ.

Challenges:

- KG construction
- Limited coverage
- Constrained expressive power
- Complex QA



v2.5 Federated
RAG

2020's Knowl
internalization

Challenges:

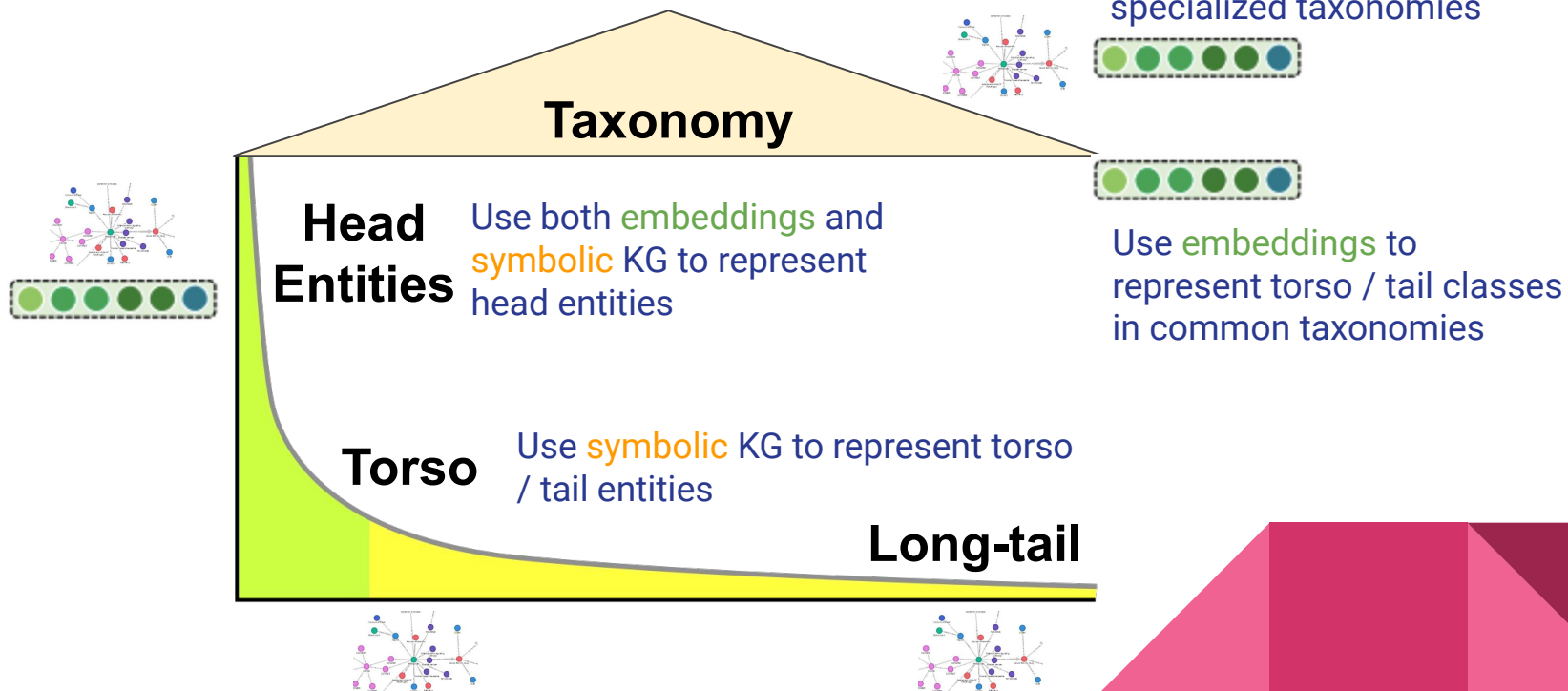
- Hallucinations



v3.5 LLM w. Dual
Neural Knowledge

Portrayed By GPT-4

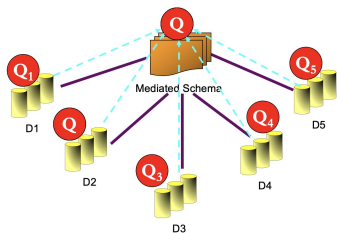
KG Future: Dual Neural Knowledge



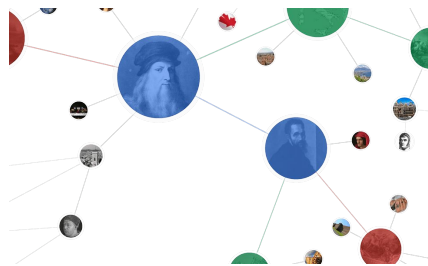
Integrating and Accessing All Valuable Data



iconpacks.net



v1.0 Web Search &
Runtime Data Integration



v2.0. KG
-based QA



v3.0. LLM

2000's
Runtime Integ.

v1.5 Tool LLM &
Web-augmented RAG

2010's
Offline Integ.

v2.5 Federated
RAG

2020's Knowl
internalization

v3.5 LLM w. Dual
Neural Knowledge

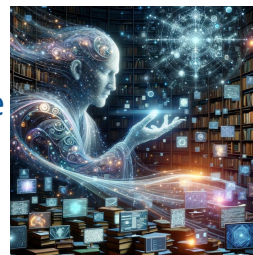
Challenges:

- Answer summarization
- Heterogeneity
- Conflict resolution
- Latency



Challenges:

- KG construction
- Limited coverage
- Constrained expressive power
- Complex QA



Challenges:

- Hallucinations



Portrayed By GPT-4

WE ARE FAR FROM THERE

Providing the Right Information

Round 1: Completed Round 1b: Completed Phase 2: 8 days left

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

Meta KDD Cup 2024

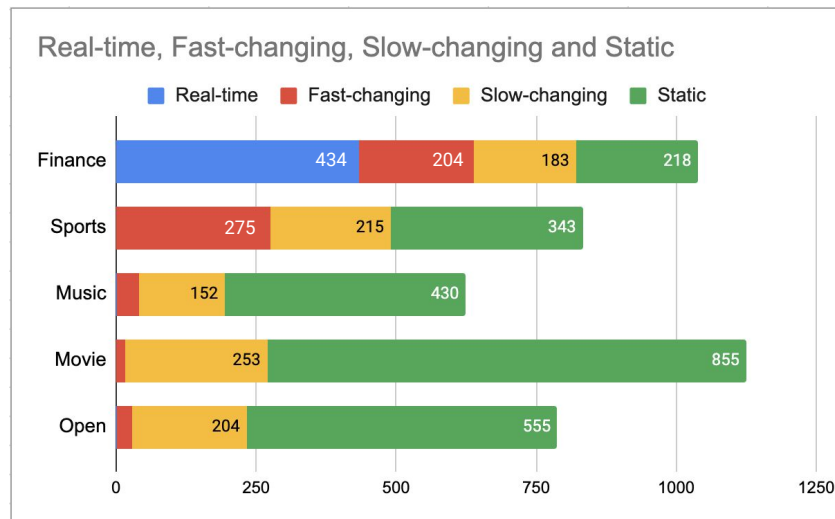
CRAG: Comprehensive RAG Benchmark

31,500

By Meta 96.8k 1777 384 4302 70 [Share](#)

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	Score
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 and better retrieval results both help

RAG may also bring more hallucinations

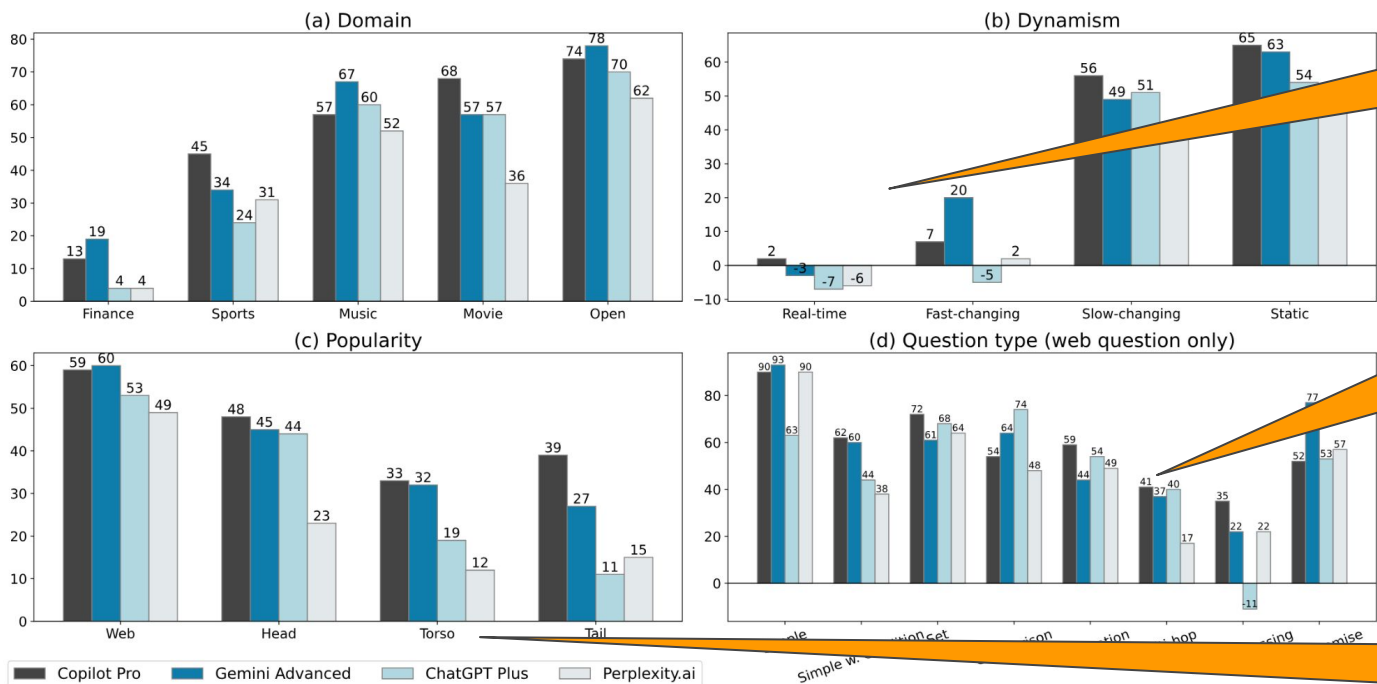
State-of-the-Art Industry Solutions

	System	Perfect	Acc.	Hall.	Miss.	Score _h	Latency (ms)
Equal weighted	Copilot Pro	62.6	11.7	17.9	7.8	50.6	11,596
	Gemini Advanced	60.8	10.1	16.6	12.5	49.3	5,246
	ChatGPT Plus	59.8	13.1	25.2	1.9	41.2	6,195
	Perplexity.ai	55.8	8.8	25.3	10.1	34.9	2,455

Perfect < 63%. Still a big gap to fill

Different latency-quality tradeoffs

SOTA Industry Solutions on Diff Dimensions



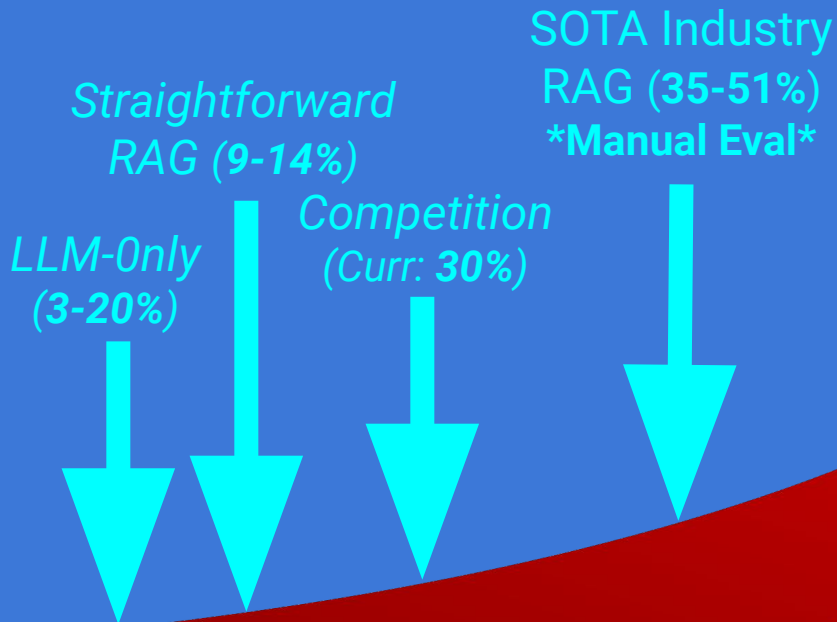
Improvements needed for **fast-changing** facts

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



Hugging Face

Daily Papers

by  [AK](#)



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

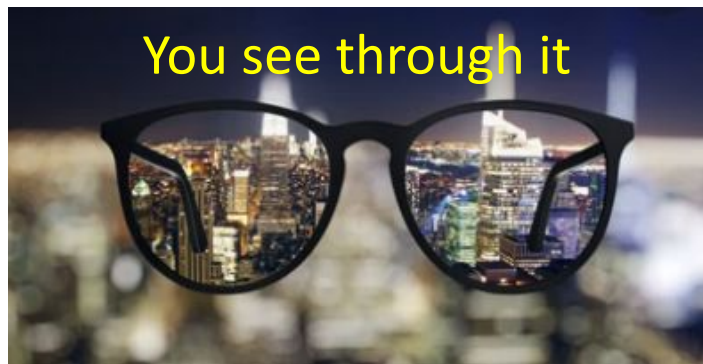


III. Wearable Asst

Voice + Visual + Context



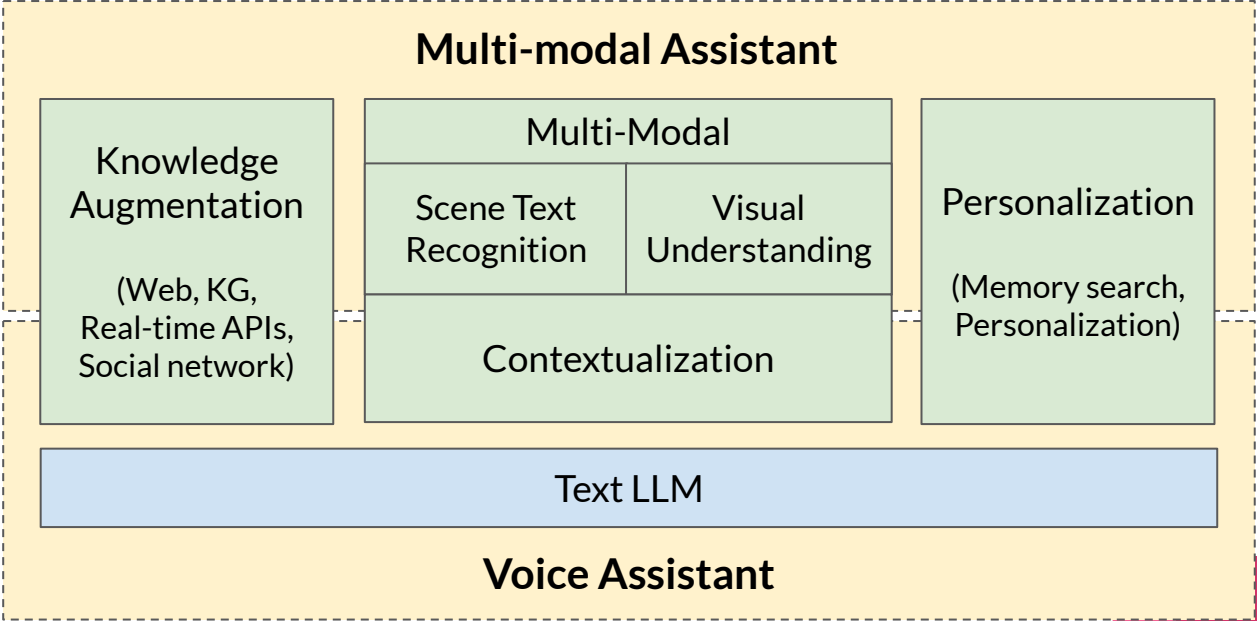
What Is Different for A Wearable Assistant?



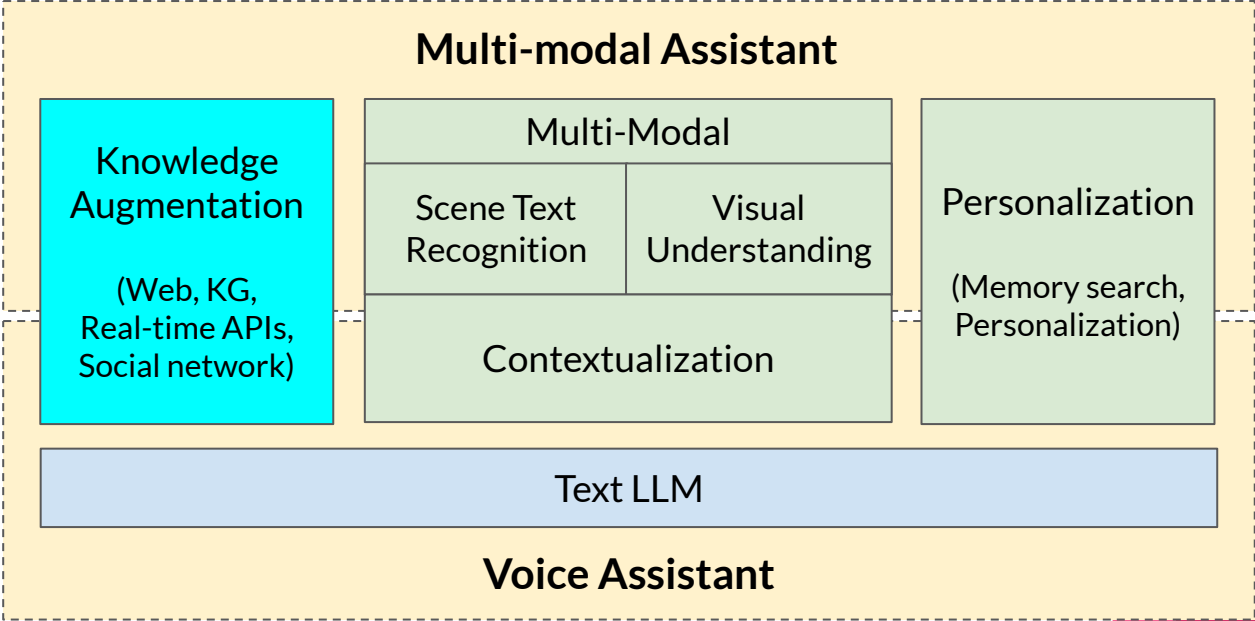
Ray-ban Meta [Video](#)



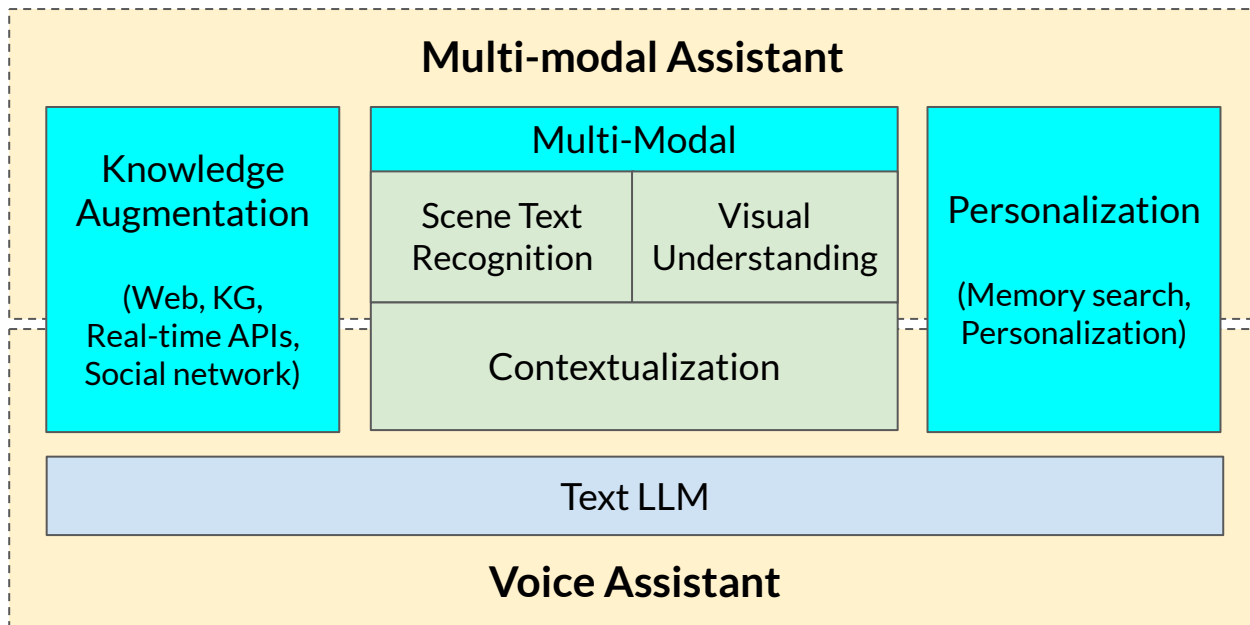
Pillars for Next-Generation Assistant



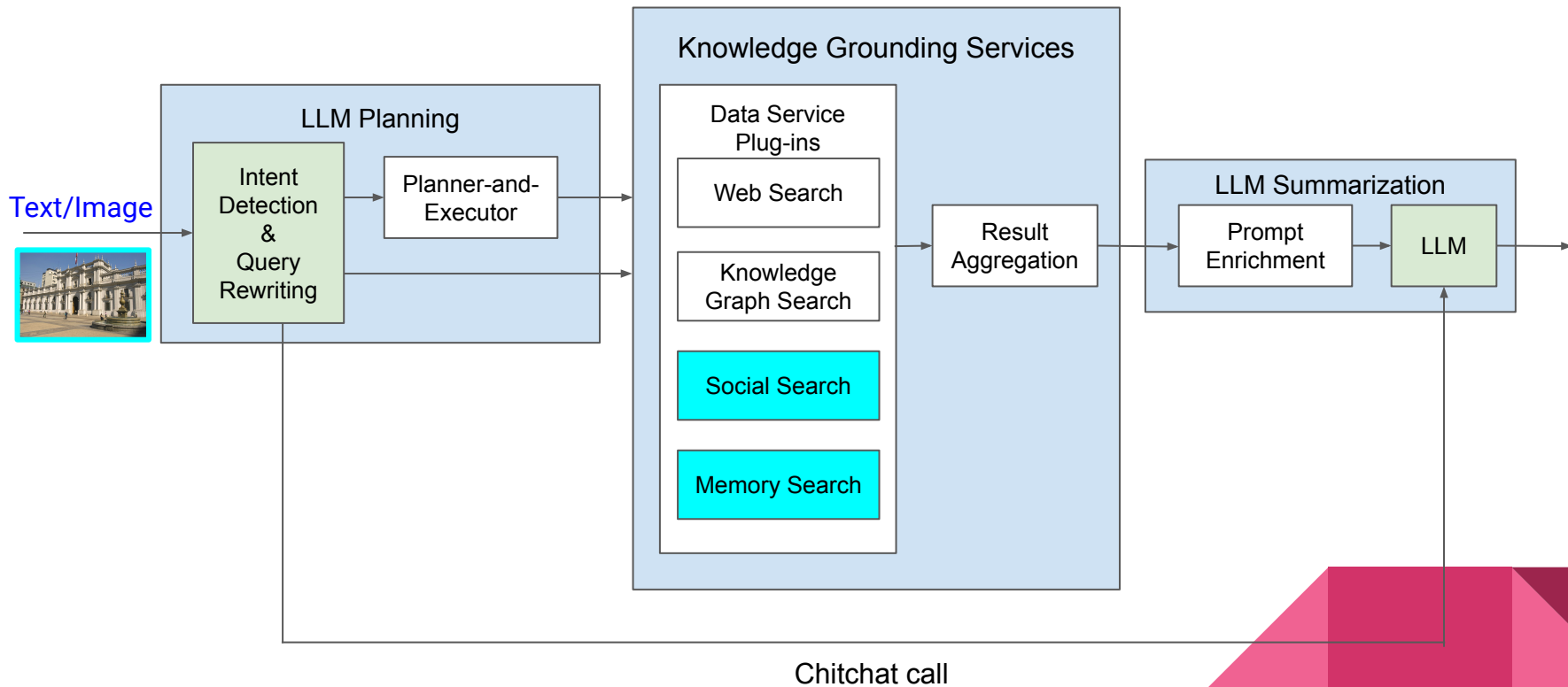
Pillars for Next-Generation Assistant



Similar Tech Extends to Other Pillars



Extending RAG for Memory & Public Search



DISCOVERY VS INVENTION

Research in the GenAI Era

- Discovering capabilities of LLMs
 - LLMs have strengths, but NOT all-powerful
 - Previous intuitions still apply
- Data is still a key in many aspects
 - Data analysis to understand the model
 - Training data to guide the model
 - Efficient access of data to improve the model
- The landscape is changing fast, calling for fast iterations and considerations of universal solutions



Call-for-Action: From Data Manager to Data Provider



Take-Aways

- **1 Mission:** *Providing the right information at the right time*
 - **2 Memes:** LLM alone or LLM+Web has not yet solved the problem
 - **Our solution:** *Federated RAG*
 - **Future KG:** *Dual Neural Knowledge*
 - **Benchmark:** **CRAG** (*Comprehensive RAG*)
 - **Call for action:** from **data manager** to **data provider**
-

Acknowledgement



Kai Sun



Xiao Yang



Ethan Xu



Scott Yih

Meta: Nikita Bhalla, Sajal Choudhary, Rongze Daniel Gui, Annie Dong, Teja Gollapudi, Yin Huang, Ziran Will Jiang, Lingkun Kong, Zhaojiang Lin, Xin Liu, Rahim Manji, Ahmed Mohamed, Brian Moran, Luis Poletto, Abhinandan Prativadi, Akshay Shekher, Jigar Surana, David Xu, Fan Wang, Jiaqi Wang, An Yan, Eting Yuan, Hanwen Zha, Hongyuan Zhan, Vijay Janakiraman, Nicolas Scheffer, Yue Liu, Ahmed Aly, Zhan Zhang, Ling Wang, Nirav Shah, Rakesh Wanga, Anuj Kumar, Jinsong Xu, Heidi Young



Yushi Sun



Hao Xin



Nan Tang



Lei Chen

HKUST: Xiangsen Chen, Ziyu Jiang, Chenyu Yang



Thank You

Q&A?