SIGIR 2025
Padova
ITALY

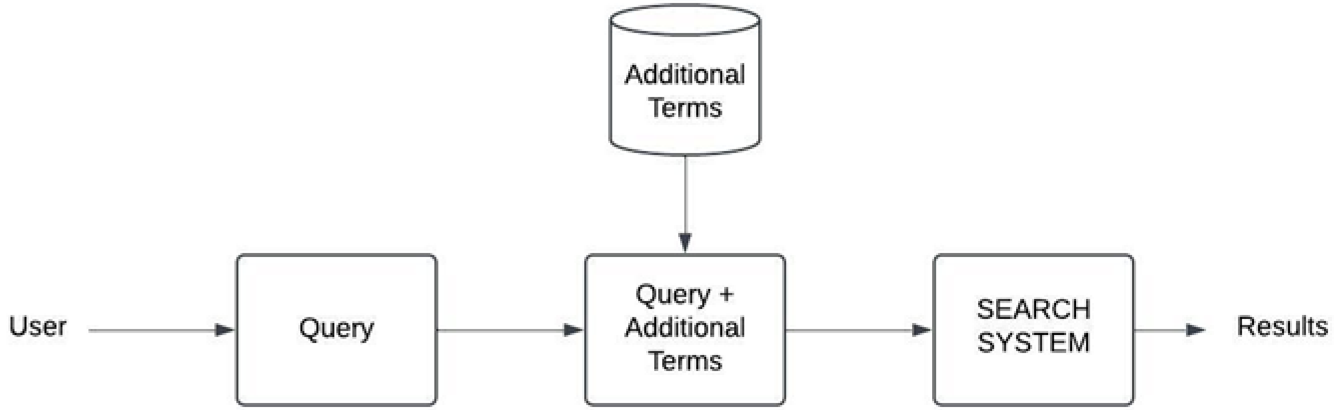
LLM-based Query Enhancement

Yifei Yuan

ETH Zürich, University of Copenhagen

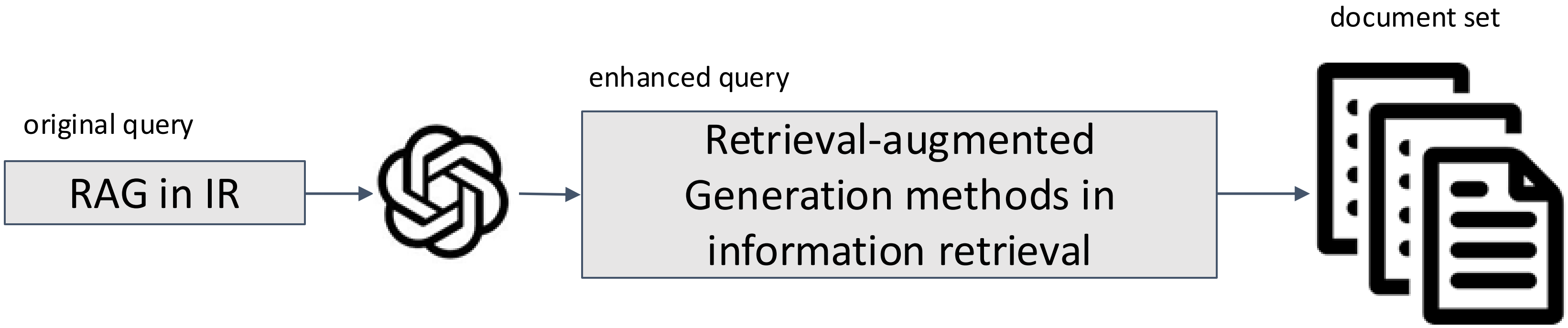
Introduction

Query Enhancement: enhances the query based on pseudo-relevance feedback or external knowledge sources, given that search queries are often **short, ambiguous, or lack necessary background information**.



Introduction

Large Language Models (LLMs) have seen a growing interest in the Information Retrieval (IR) community in recent years. They exhibit several properties, including the ability to answer questions and generate text, that make them powerful tools.

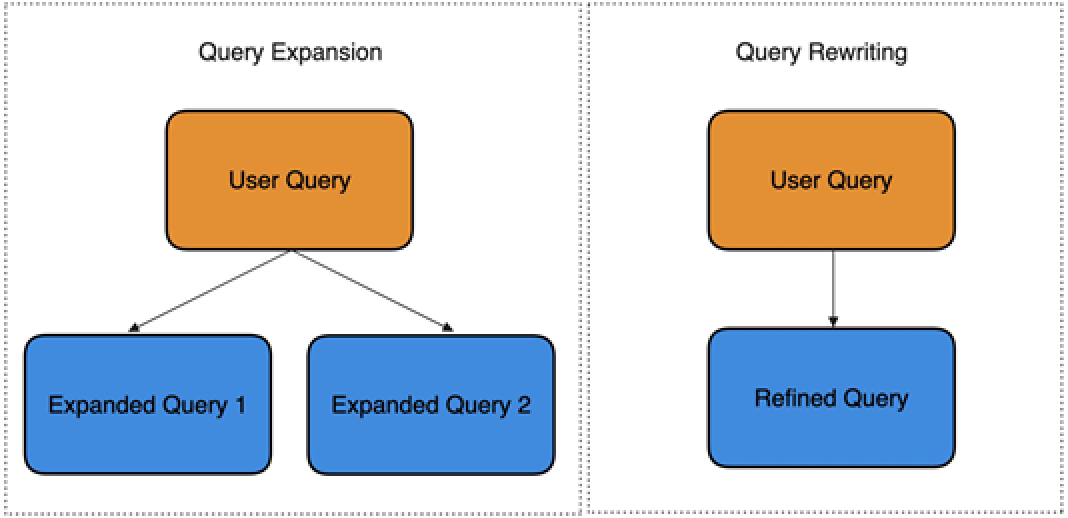


Taxonomies

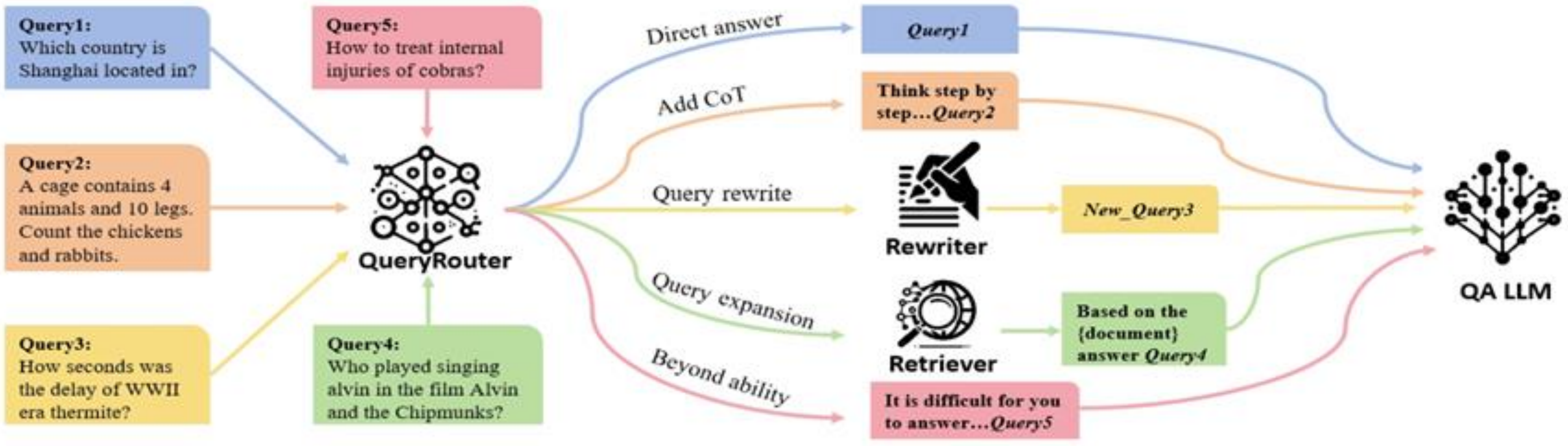
Taxonomies

- Resolving ambuity
  - Query expansion
  - Query clarification
  - Query suggestion
  - Query rewrite

- Interactive query refinement



Taxonomies



Chen et al., "When and How to Augment Your Input: Question Routing Helps Balance the Accuracy and Efficiency of Large Language Models" (Naacl findings '25)

Query Clarification

Proves to be efficient for ambiguous queries with multiple answers

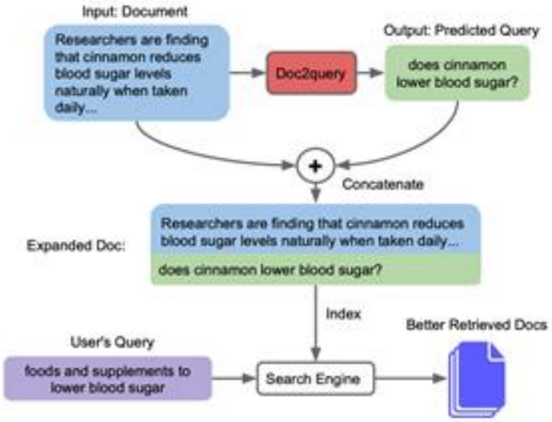
LLMs must learn to clarify the query by identifying user's intent

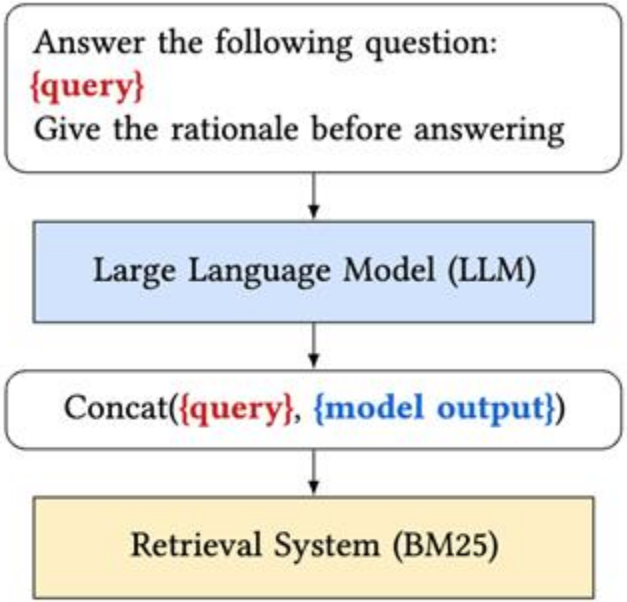
- **User Request:** an initial user request in the conversational form, e.g. *What is Fickle Creek Farm*, with a label reflects if clarification is needed ranged from 1 to 4;
- **Clarification questions:** a set of possible clarifying questions, e.g. *do you want to know the location of fickle creek farm*;
- **User Answers:** each questions is supplied with a user answer, e.g. *no i want to find out where can i purchase fickle creek farm products*.

Query Expansion

Query expansion augments a user's original query with additional terms or phrases to improve retrieval performance.

Can be divided into **internal expansion** and **external expansion**.

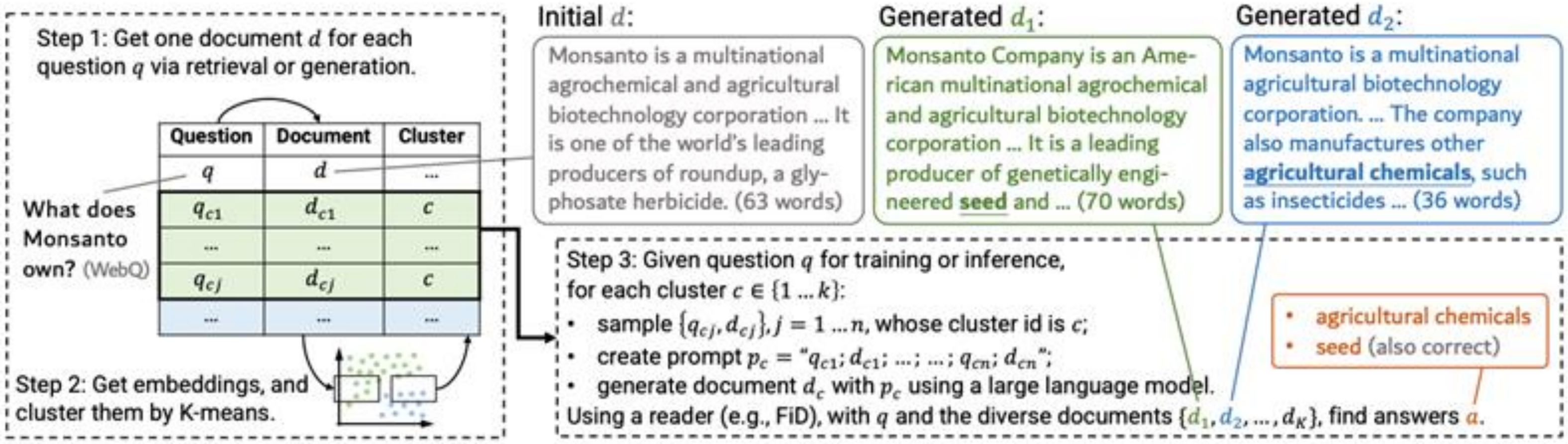




Jagerman et al., "Query Expansion by Prompting Large Language Models" (Arxiv'23)
Nogueira et al, "Document Expansion by Query Prediction" (EMNLP'19)

Internal Query Expansion

It focuses on maximizing the value of existing information in the original query or the used LLM **without relying on external knowledge sources**

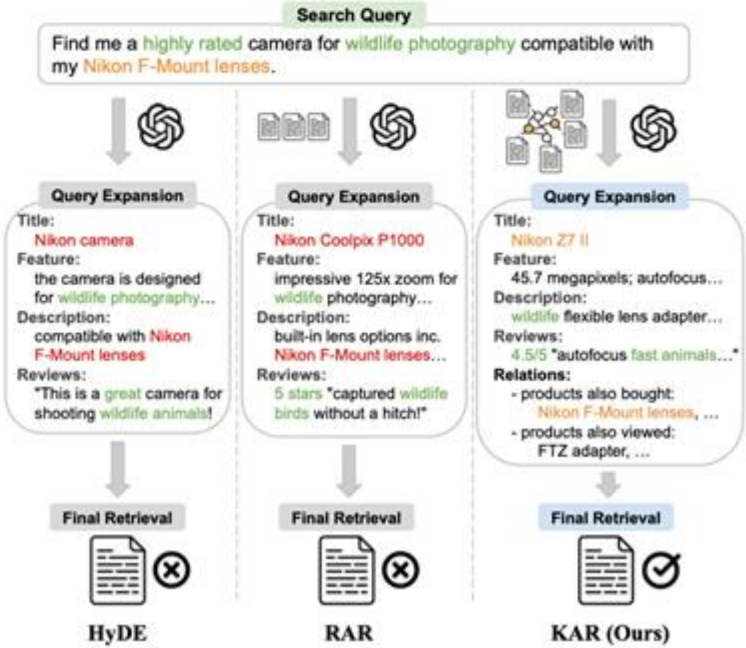


Yu et al., "Generate rather than Retrieve: Large Language Models are Strong Context Generators" (ICLR'23)

External Query Expansion

It introduces supplementary data from outside sources (e.g., Web or Knowledge base) to fill gaps, provide additional context, or broaden the scope of the content

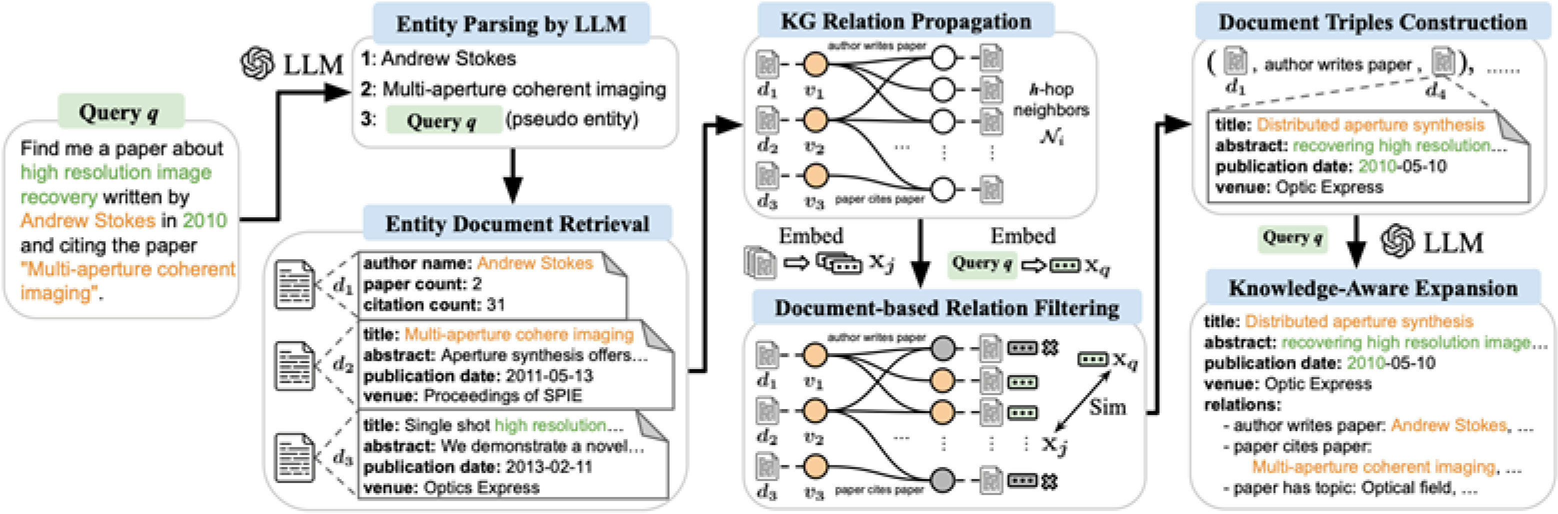
Sources: WordNet, ontologies (e.g. DBpedia), search logs, etc...



Xia et al., "Knowledge-Aware Query Expansion with Large Language Models for Textual and Relational Retrieval" (NAACL'25)

External Query Expansion

It introduces supplementary data from outside sources (e.g., Web or Knowledge base) to fill gaps, provide additional context, or broaden the scope of the content

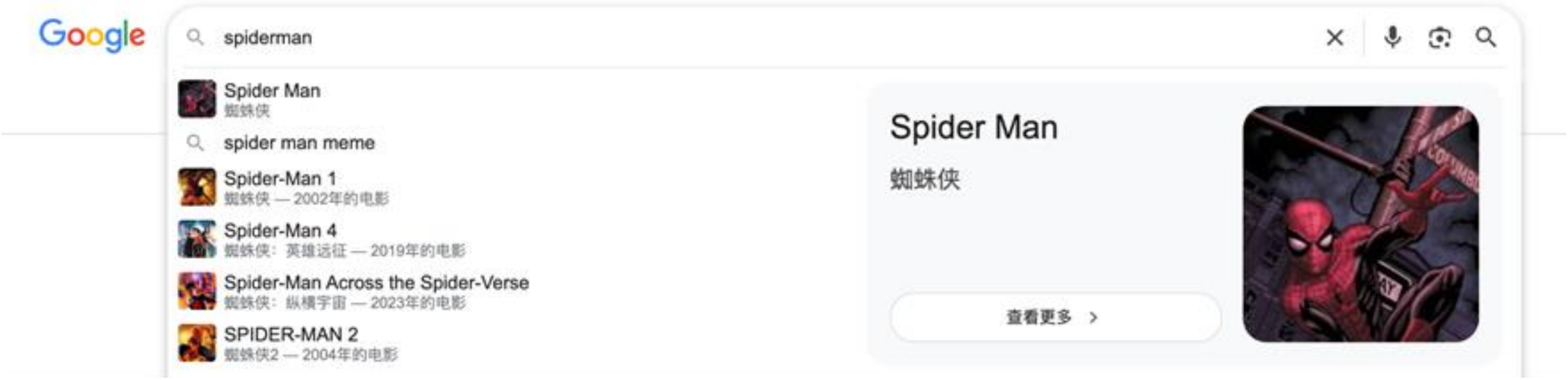


Pros and Cons?

Query Suggestion

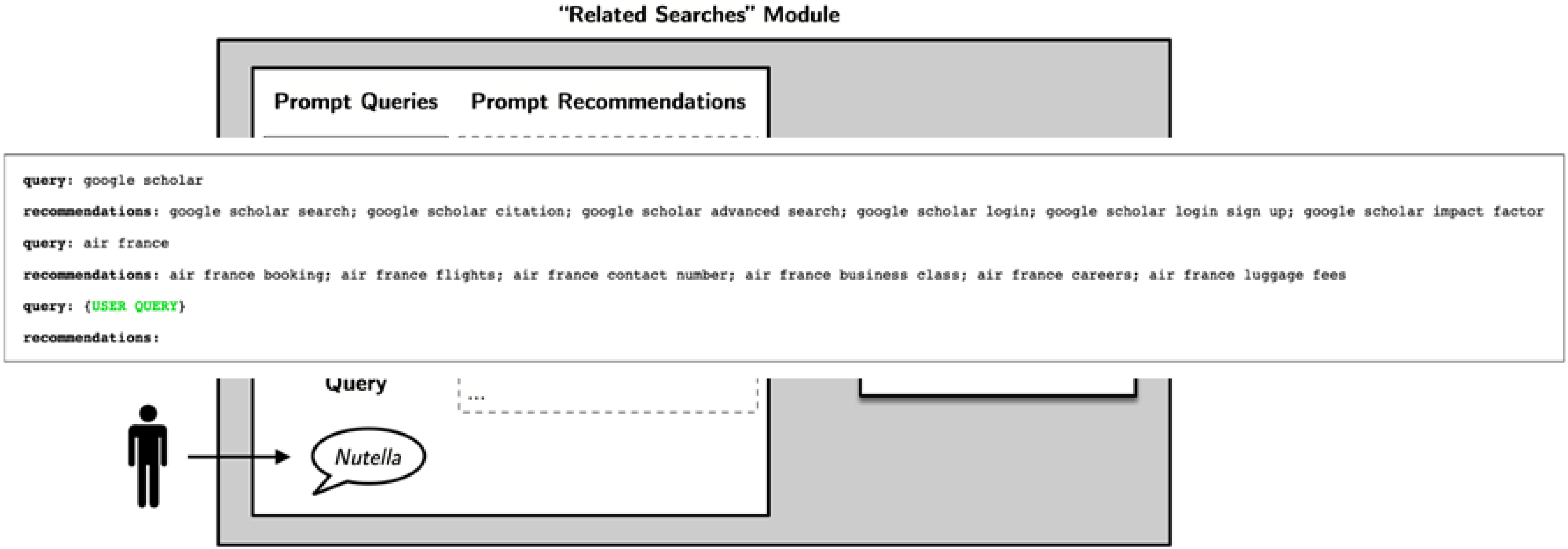
Works by suggesting alternative or additional queries based on what the user has typed so far.

Can be formulated as a recommendation problem.



Query Suggestion

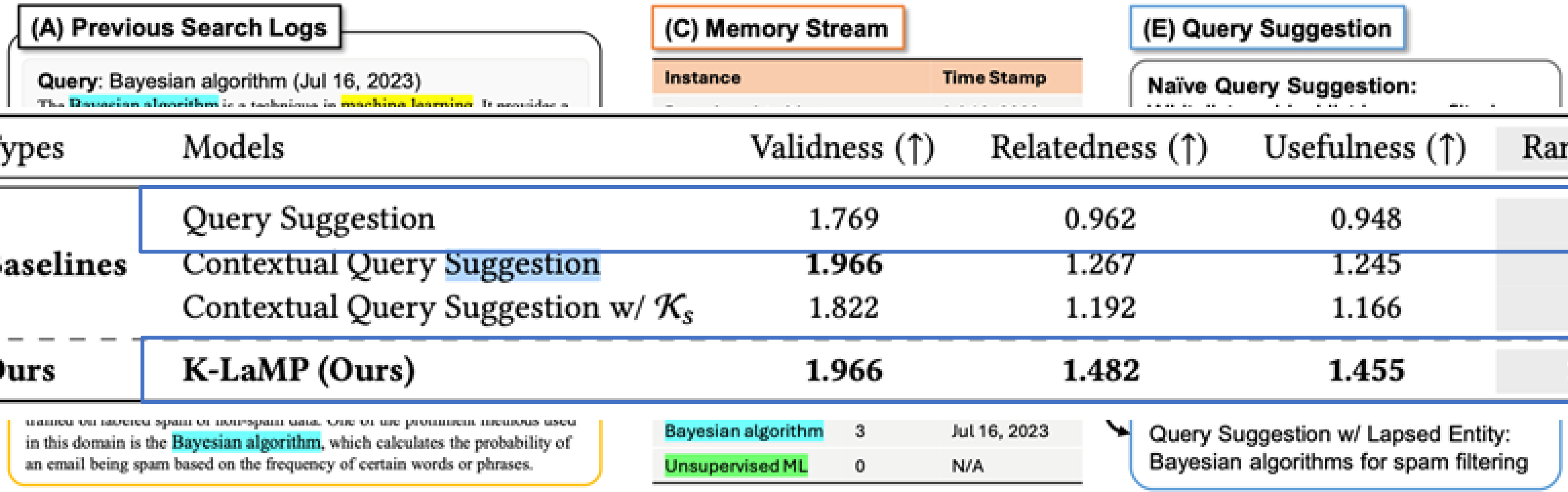
LLMs can generate the recommendation suggestions with simple prompting



Bacciu et al., "Generating Query Recommendations via LLMs" (IR-RAG @ SIGIR'2024)

Query Suggestion

LLM with knowledge injection boosts the performance further



Types	Models	Validness (↑)	Relatedness (↑)	Usefulness (↑)	Ranking (↓)
Baselines	Query Suggestion	1.769	0.962	0.948	2.736
	Contextual Query Suggestion	**1.966**	1.267	1.245	2.415
	Contextual Query Suggestion w/ $\mathcal{K}_s$	1.822	1.192	1.166	2.654
Ours	**K-LaMP (Ours)**	**1.966**	**1.482**	**1.455**	**2.160**

Baek et al, WWW 2024, Knowledge-Augmented Large Language Models for Personalized Contextual Query Suggestion

Baek et al., "Knowledge-Augmented Large Language Models for Personalized Contextual Query Suggestion" (WWW '24)

LLM-based Query Rewrite

Query rewrite transforms a user's context-dependent query into a fully-specified version that can be understood independently of the surrounding context.

Two key challenges:

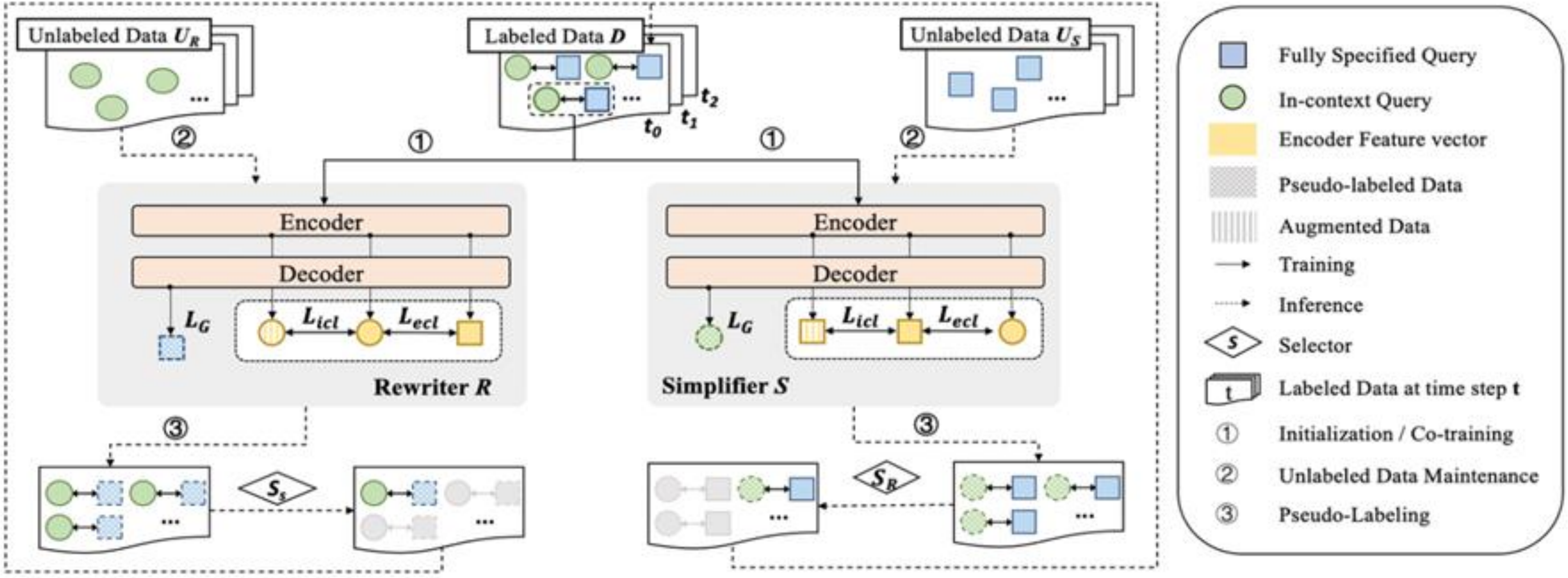
- Coreference resolution
- Ellipsis completion

Conversation Contexts	
$\mathcal{Q}_1$	What can you tell me about ***Beyoncé's*** voice ?
$\mathcal{A}_1$	Her tone and timbre as particularly distinctive...
$\mathcal{Q}_2$	What are some other facts about *her* voice ?
$\mathcal{A}_2$	The New York Times commented her voice is "velvety yet tart"...
$\mathcal{Q}_3$	What else ?
$\mathcal{A}_3$	Other critics praises she was "capable of punctuating any beat".
Query Rewrites	
$\mathcal{Q}_2^*$	What are some other facts about *Beyoncé's* voice ?
$\mathcal{Q}_3^*$	What else *can you tell me about Beyoncé's voice* ?

Yu et al., "Few-Shot Generative Conversational Query Rewriting" (SIGIR '20)

LLM-based Query Rewrite

Weakly-supervised training can be adopted for the difficulty to find gold labels



Yuan et al., "CO3: Low-resource Contrastive Co-training for Generative Conversational Query Rewrite" (COLING '24)

LLM-based Query Rewrite

LLMs show promise in generating high-quality rewrites **especially** in the low-resource setting

	Model	BLEU-1	BLEU-2	ROUGE-1	ROUGE-2	ROUGE-L	EM	NDCG@3
	Original	72.50	66.17	79.71	65.66	79.66	18.65	30.40
	Allen Coref	79.37	74.29	86.04	76.72	85.94	36.13	43.59
Zero-shot	GQR	16.02	10.63	27.37	13.13	27.29	1.47	12.56
	GPT-2	15.41 (15.45)	10.54 (10.40)	27.17 (28.46)	12.42 (12.86)	26.75 (28.12)	1.17 (1.86)	11.32 (11.56)
	MS MARCO	35.19 (34.62)	19.90 (19.73)	31.06 (29.93)	13.18 (13.21)	30.41 (29.39)	0.93 (0.93)	16.90 (14.32)
	Rule Based	82.49 (79.31)	74.29 (72.30)	82.92 (82.93)	71.03 (70.53)	81.55 (81.86)	25.87 (26.81)	43.72 (43.25)
	CO3	**83.94* (80.91)**	**75.36* (73.37)**	**84.08* (83.08)**	**72.32* (71.31)**	**82.94* (82.02)**	**27.91* (27.04)**	**45.72* (44.67)**
	L-CO3	89.42†	77.31†	89.06†	74.90†	85.26†	30.55†	48.90†
Few-shot	Seq2Seq	72.11	62.47	78.75	65.61	78.02	6.45	20.42
	GQR	84.84	78.80	87.42	77.93	86.40	40.82	47.28
	GPT-2	84.61 (83.20)	78.62 (77.00)	87.27 (85.52)	77.86 (75.79)	86.25 (84.66)	40.79 (35.89)	46.74 (43.28)
	Rule Based	85.71 (82.35)	79.66 (76.23)	88.08 (85.91)	78.71 (75.97)	86.97 (85.09)	40.79 (36.13)	49.21 (46.76)
	Self-Learn	85.12 (**83.53**)	79.73 (77.51)	88.22 (86.82)	79.36 (76.90)	87.38 (85.91)	43.12 (38.23)	49.24 (46.53)
	CO3	**85.87*** (83.42)	**80.24* (78.14)**	**89.04* (86.95)**	**80.08* (77.48)**	**87.92* (86.36)**	**44.05* (40.79)**	**50.43* (48.26)**
	L-CO3	90.05†	86.47†	93.26†	85.28†	92.43†	49.07†	56.22†

Yuan et al., "CO3: Low-resource Contrastive Co-training for Generative Conversational Query Rewrite" (COLING '24)

LLM-based Query Rewrite

Instruction-tuning also proves to be an efficient way for rewriting the query

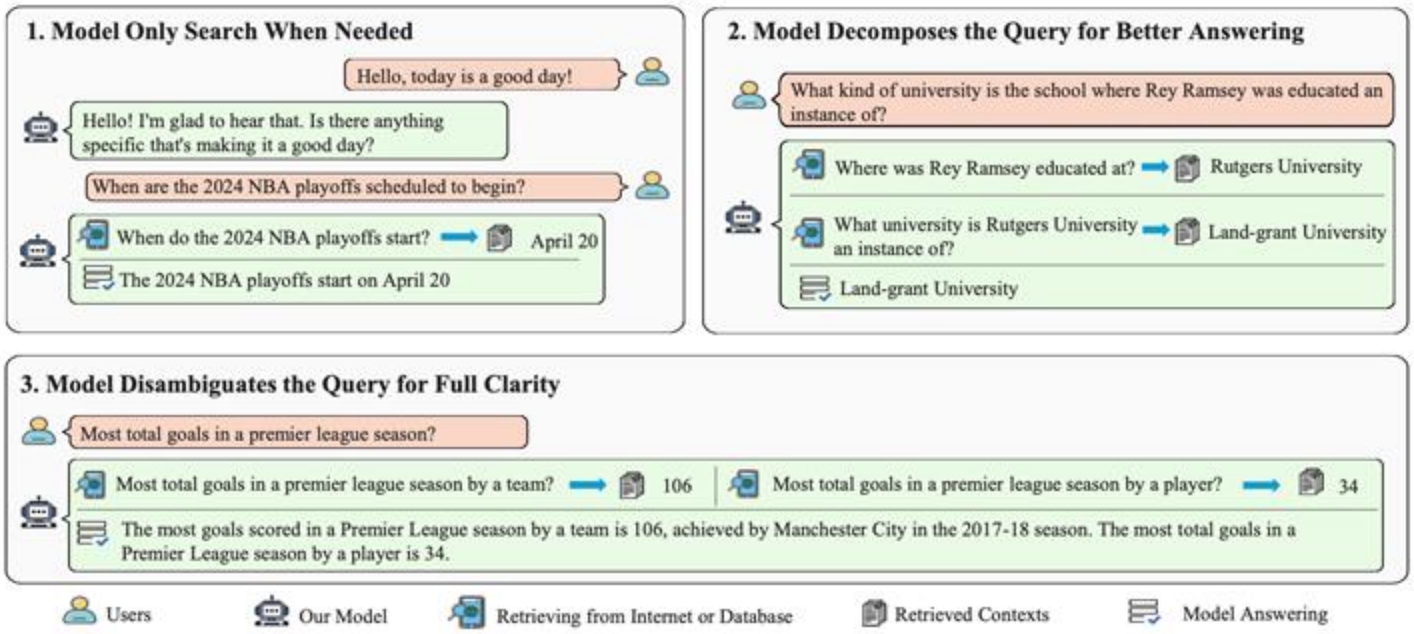
	Query	QReCC (8209)			QuAC-Conv (6396)			NQ-Conv (1442)			TREC-Conv (371)		
		MRR	MAP	R@10	MRR	MAP	R@10	MRR	MAP	R@10	MRR	MAP	R@10
Sparse (BM25)	Original	9.30	8.87	15.50	9.29	8.84	15.20	9.06	8.64	15.14	10.30	10.27	22.10
	Human	39.81	38.45	62.65	40.32	38.98	62.90	40.78	39.05	**63.80**	**27.34**	**27.04**	**53.77**
	T5QR	33.67	32.50	53.68	34.04	32.90	53.83	34.24	32.66	53.92	25.23	24.96	50.13
	ConQRR	38.30	-	60.10	39.50	-	61.60	37.80	-	58.00	19.80	-	43.50
	ConvGQR	44.10	-	64.40	-	-	-	-	-	-	-	-	-
	RW(ZSL)	42.63	41.31	60.46	45.43	44.11	63.20	36.43	34.81	54.69	18.50	18.26	35.58
	RW(FSL)	46.96	45.53	65.57	49.81	48.38	68.28	41.51	39.71	60.13	19.02	18.86	39.89
	ED(Self)	**49.39**	**47.89**	**67.01**	**53.01**	**51.52**	**70.46**	41.57	39.69	59.63	17.43	17.08	36.25
	ED(T5QR)	47.93	46.40	66.25	50.67	49.18	68.84	**42.69**	**40.64**	60.67	21.04	20.79	43.26
Dense (GTR)	Original	12.12	11.49	18.74	11.34	10.69	17.79	13.11	12.57	19.49	21.76	21.11	32.08
	Human	43.15	41.27	66.12	40.67	38.92	64.59	**54.01**	**51.25**	**73.13**	**43.74**	**42.98**	**65.23**
	T5QR	37.67	35.93	58.65	35.51	33.88	57.23	46.95	44.47	64.43	38.94	38.16	60.51
	ConQRR	41.80	-	65.10	41.60	-	65.90	45.30	-	64.10	32.70	-	55.20
	ConvGQR	42.00	-	63.50	-	-	-	-	-	-	-	-	-
	RW(ZSL)	40.64	38.95	62.28	40.12	38.48	62.47	44.85	42.57	63.58	33.26	32.88	54.09
	RW(FSL)	43.89	42.09	66.45	43.50	41.78	66.87	48.60	46.12	68.10	32.37	31.79	52.65
	ED(Self)	**44.99**	**43.19**	**67.34**	**45.21**	**43.48**	**68.30**	47.64	45.20	67.27	30.91	30.48	51.03
	ED(T5QR)	44.76	42.90	66.64	44.29	42.50	66.65	49.67	47.12	69.22	33.90	33.43	56.47

- In-context demonstration helps
- LLM improvement is more obvious in dense retrieval

Ye et al., "Enhancing Conversational Search: Large Language Model-Aided Informative Query Rewriting" (EMNLP Findings '23)

Interactive Query Refinement

Interactive query refinement generally allows the system to rewrite, disambiguate, decompose a query when interacting with a user

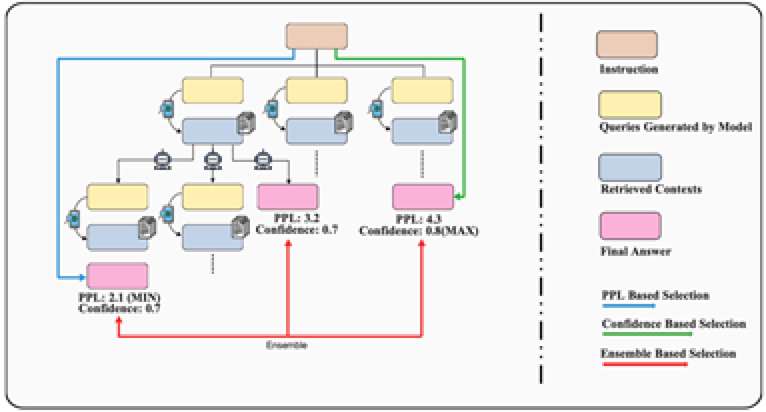


Interactive Query Refinement

Adopts a tree-decoding strategy for selecting the best way for refinement

LLM serves as an efficient way for refining the query and incorporating RAG

Model	HOTPOTQA	2WIKI	MUSIQUE	AVG.
	Proprietary LLM			
GPT-3.5-TURBO				
+ Chain-of-Thought	58.6	43.9	32.3	44.9
+ Chain-of-Note	52.5	34.1	24.6	37.1
GPT-4				
+ Chain-of-Thought	71.4	70.1	50.3	63.9
+ Chain-of-Note	72.4	58.3	44.1	58.3
	Baseline without Retrieval			
LLama2-7B (Zero Shot)	6.6	16.0	3.0	8.5
LLama2-7B-Chat (Zero Shot)	3.6	7.9	1.8	4.4
LLama2-7B (SFT on Multi Hop QA)	34.7	34.2	6.8	25.2
LLama2-7B (SFT on No Augmented Set)	35.6	30.8	6.7	24.4
	Baseline with Retrieval			
LLama2-7B (Zero Shot)	16.7	18.7	7.4	14.3
LLama2-7B-Chat (Zero Shot)	2.8	3.5	1.8	2.7
LLama2-7B (SFT on Multi Hop QA)	37.5	32.3	7.9	25.9
LLama2-7B (SFT on No Augmented Set)	43.5	28.8	9.1	27.1
RQ-RAG (Ours)	**62.6** (19.1↑)	**44.8** (16.0↑)	**41.7** (32.6↑)	**49.7** (22.6↑)



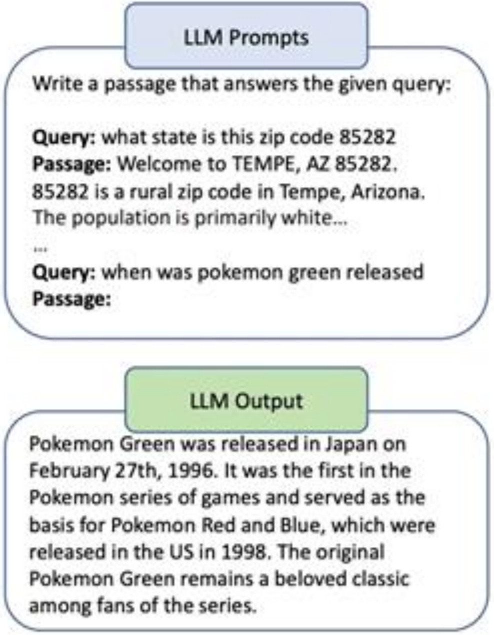
Chan et al., "RQ-RAG: Learning to Refine Queries for Retrieval Augmented Generation" (COLM '24)

Common Techniques

Few-shot Prompting

Pseudo data (e.g. queries, documents) can be synthesized by prompting LLMs in the few-shot manner.

Synthesized data can be attached and added with the original query for better retrieval results.



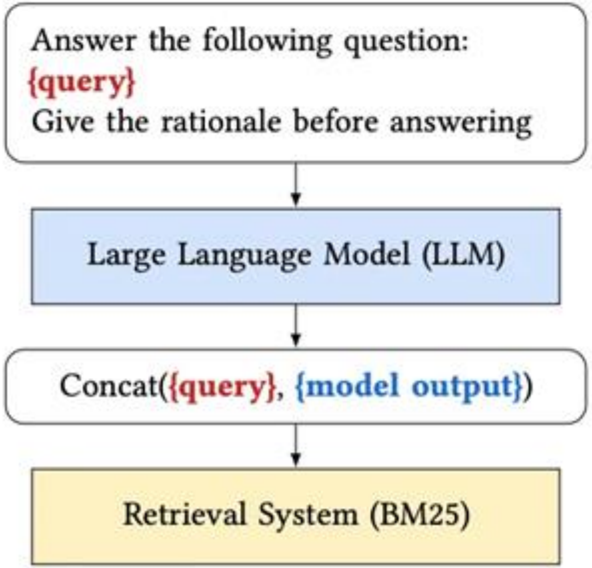
SimLM (Wang et al., 2023)	✓	41.1	87.8	98.7	71.4	69.7
+ query2doc	✓	$\mathbf{41.5}^{+0.4}$	$88.0^{+0.2}$	$\mathbf{98.8}^{+0.1}$	$72.9^{+1.5}$	$71.6^{+1.9}$
$E5_{base}$ + KD (Wang et al., 2022)	✓	40.7	87.6	98.6	74.3	70.7
+ query2doc	✓	$\mathbf{41.5}^{+0.8}$	$\mathbf{88.1}^{+0.5}$	$98.7^{+0.1}$	$\mathbf{74.9}^{+0.6}$	$\mathbf{72.5}^{+1.8}$

Wang et al., "Query2doc: Query Expansion with Large Language Models" (EMNLP '23)

CoT Prompting

Adding CoT generally improves the performance in most metrics!

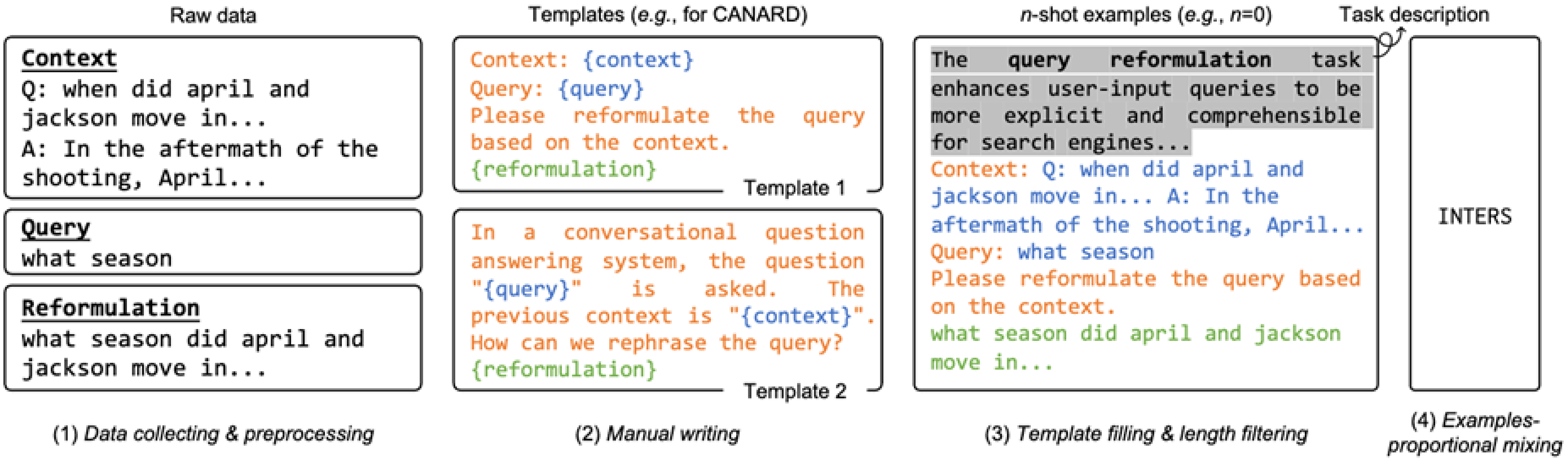
Dataset	Metrics	Direct answer	Add CoT
AmbigNQ	Acc(%)	42.65	44.60
	TCA(s)	**1.98**	5.26
	TAI(s/%)	NaN	1538.50
Hotpot	Acc(%)	31.75	38.00
	TCA(s)	**2.74**	5.72
	TAI(s/%)	NaN	416.14
MMLU-STEM	EM	44.20	65.34
	TCA(s)	**2.13**	3.90
	TAI(s/%)	NaN	273.69
PopQA	Acc(%)	32.70	32.80
	TCA(s)	**1.39**	6.78
	TAI(s/%)	NaN	3534.20



Jagerman et al., "Query Expansion by Prompting Large Language Models" (Arxiv '23)

Instruction Tuning

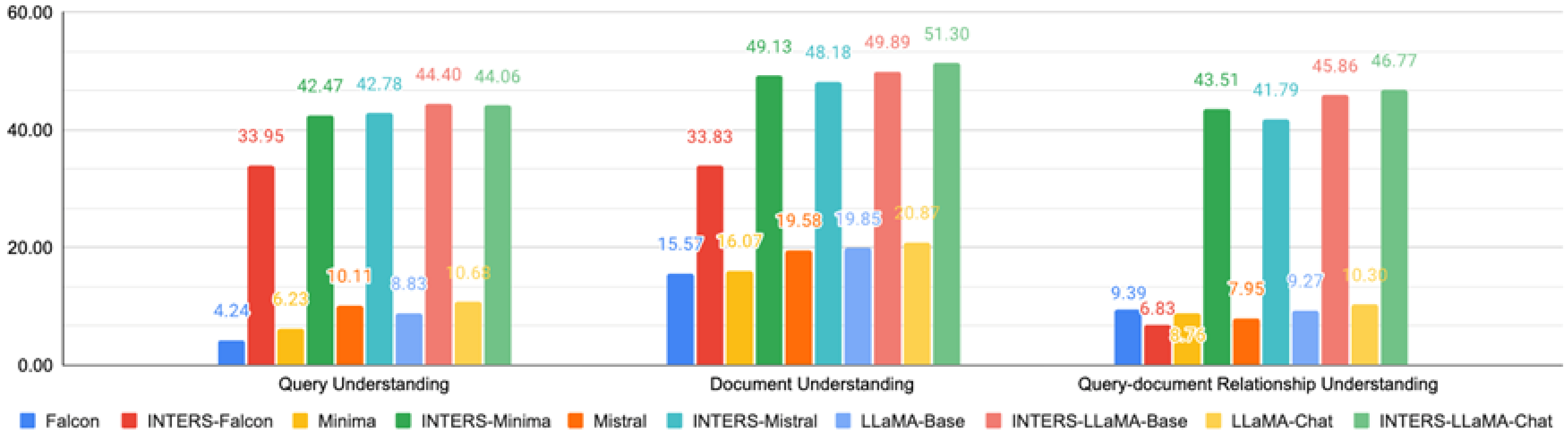
INTERS: an instruction tuning dataset for search tasks



Zhu et al., "INTERS: Unlocking the Power of Large Language Models in Search with Instruction Tuning" (ACL '24)

Instruction Tuning

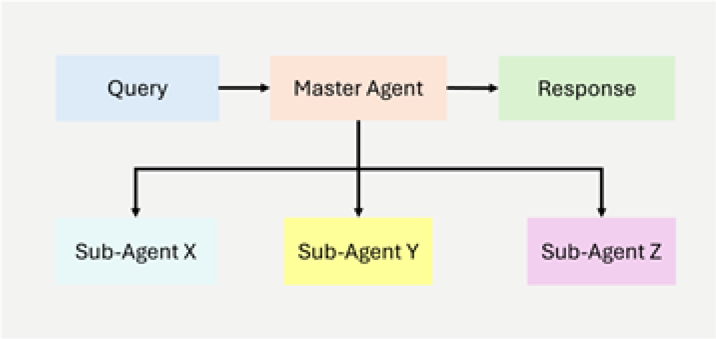
Most LLMs can obtain the capability to solve search tasks through instruction tuning on INTERS.

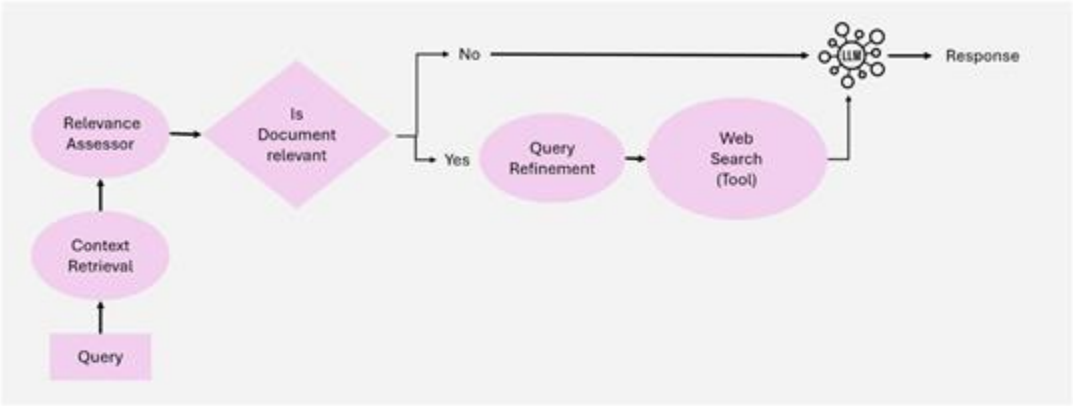


Zhu et al., "INTERS: Unlocking the Power of Large Language Models in Search with Instruction Tuning" (ACL '24)

Agentic RAG

Agentic RAG refers to the use of *agent-like behavior* in Retrieval-Augmented Generation systems, where the system actively plans and executes steps—such as rewriting, refining, or expanding a user's query—to improve retrieval quality and response generation.





Singh et al., "Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG" (Arxiv '54)

Applications

Multimodal Query Rewrite

Definition: Reformulating user queries by leveraging both textual and visual context to improve retrieval relevance.

Challenges:

- How to effectively align multimodal features?
- Matain user intent during rewrite
- Handling ambiguity in multimodal input

Yuan et al., "Mcqueen: a benchmark for multimodal conversational query rewrite" (EMNLP '23)

Multimodal Query Rewrite

Techniques: Cross-modal fusion models for encoding → an encoder-decoder model with pointer mechanism

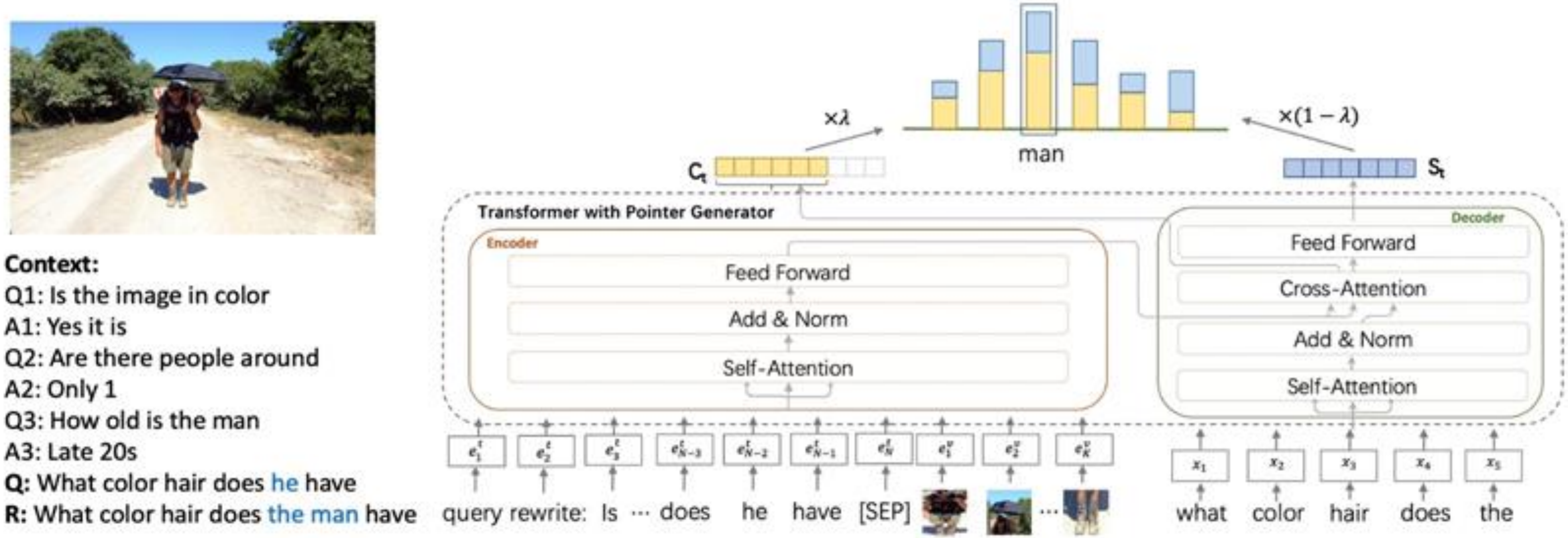


Figure 2: The overall architecture of our model.

Yuan et al., "Mcqueen: a benchmark for multimodal conversational query rewrite" (EMNLP '23)

Multimodal Query Rewrite

	BLEU-2	BLEU-4	ROUGE-2	ROUGE-L	METEOR	EM
Original	57.27	48.44	54.86	73.45	38.28	- | -
AllenNLP Coref	59.72	50.21	56.09	76.07	39.87	8.97 | 96.14
L-Gen	77.14	66.31	74.44	84.68	46.09	41.84 | 57.77
T(E)-Gen	78.03	65.40	75.53	86.80	47.32	41.37 | 56.58
T(L)-Gen	79.30	67.16	76.28	88.14	49.23	42.06 | 58.25
L-Ptr	77.46	69.46	76.07	86.65	48.63	47.87 | 66.72
T(E)-Ptr	79.52	68.26	77.27	88.49	50.10	49.49 | 68.94
T(L)-Ptr	80.12	68.41	78.40	89.54	50.56	48.15 | 69.53
VLBart	90.01	84.28	88.67	93.80	58.87	64.28 | 86.89
VLT5	90.37	84.87	89.23	94.10	59.45	65.62 | 89.95
VLBart-Ptr	90.16	84.52	88.87	93.89	59.26	64.87 | 87.22
VLT5-Ptr	**90.47**	**84.94**	**89.32**	**94.45**	**59.46**	**65.67 | 90.22**

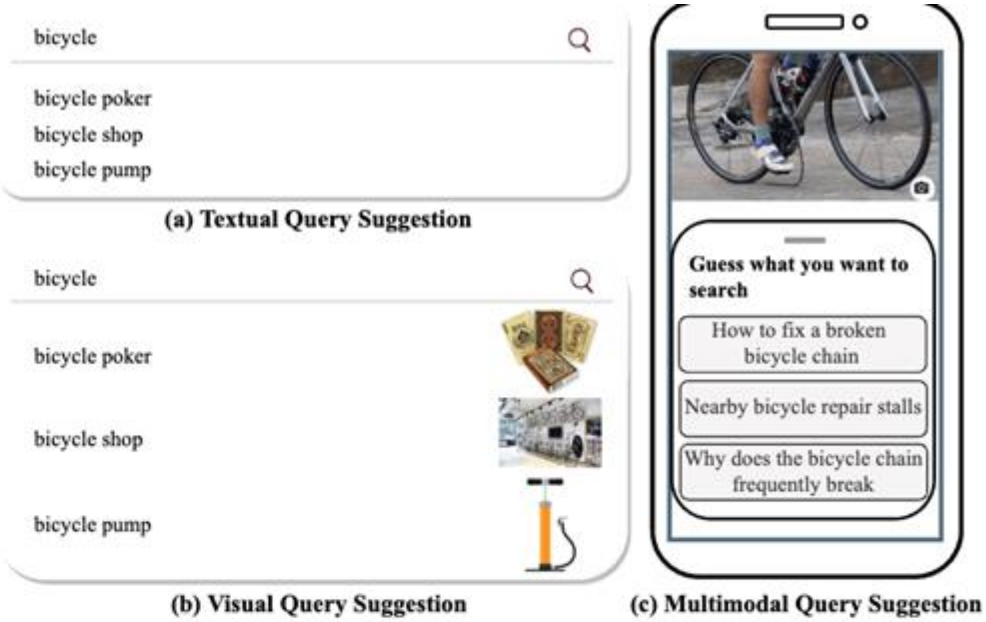
Pointer mechanism does help the results. But more techniques could also be leveraged (e.g. RAG, RLHF, …)

Yuan et al., "Mcqueen: a benchmark for multimodal conversational query rewrite" (EMNLP '23)

Multimodal Query Suggestion

Task: Generate textual query suggestions for image queries

However, predicting user needs from a single user query is challenging.



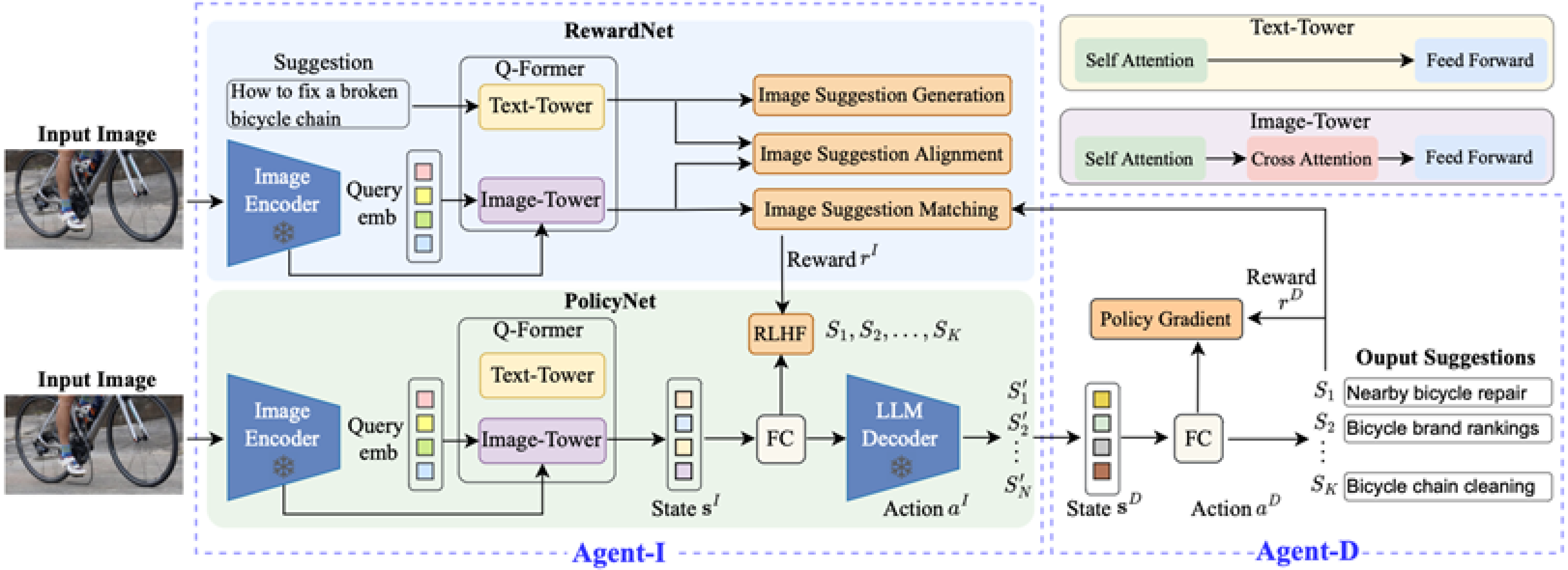
Wang et al., "Multimodal Query Suggestion with Multi-Agent Reinforcement Learning from Human Feedback" (WWW '24)

Multimodal Query Suggestion

Agent-I: generating intentional candidate suggestions

Agent-D: choosing diverse suggestions from the candidates

PolicyNet is trained with RLHF for generating candidate suggestions



Wang et al., "Multimodal Query Suggestion with Multi-Agent Reinforcement Learning from Human Feedback" (WWW '24)

Multimodal Query Suggestion

- Two search engine scenarios:

(1) generation-based

(2) retrieval-based

- δ denotes the accuracy of labeling the generated suggestions
- RL4Sugg works the best with little parameters

Models	#Train/#Total Params	Business Fine-tuned			ImageNet 0-Shot		
		DCG	DIV	δ	DCG	DIV	δ
Flamingo	1.4B/3.4B	0.73	0.25	81.7%	0.67	0.23	80.6%
BLIP-2	104M/3.1B	0.59	0.17	68.3%	0.47	0.18	69.2%
LLaVA	14M/13B	0.60	0.25	73.3%	0.47	0.24	76.5%
RL4Sugg	208M/3.1B	**0.89**	**0.25**	**83.3%**	**0.87**	**0.24**	**86.9%**

Models	#Train/#Total Params	Business Fine-tuned			ImageNet 0-shot		
		PNR	R@1	R@3	PNR	R@1	R@3
CLIP	300M/300M	1.30	0.23	0.33	0.90	0.21	0.32
BLIP-2	104M/3.1B	1.05	0.27	0.60	0.73	0.26	0.58
RL4Sugg	208M/3.1B	**2.80**	**0.63**	**0.83**	**2.17**	**0.58**	**0.74**

Wang et al., "Multimodal Query Suggestion with Multi-Agent Reinforcement Learning from Human Feedback" (WWW '24)

Summary

- **Taxonomy**
  - Query Suggestion/Rewrite/Expansion/…
  - Interactive Query Refinement
  - …
- **Common Techniques**
  - CoT Prompting
  - In-context Learning
  - Instruction Tuning
  - …
- **Applications**