THE UNIVERSITY of EDINBURGH

Université de Montréal

UNIVERSITY OF AMSTERDAM

Conversational Search: Foundations, Large Language Models, and Agents

Chuan Meng, Fengran Mo, Mohammad Aliannejadi, Jeff Dalton and Jian-Yun Nie

Tutorial at the 48th European Conference on Information Retrieval (ECIR 2026)

29 March 2026

Presenters

Chuan Meng
The University of Edinburgh

Fengran Mo
Université de Montréal

Mohammad Aliannejadi
University of Amsterdam

Jeff Dalton
The University of Edinburgh

Jian-Yun Nie
Université de Montréal

Supplementary materials

The materials, including slides, are available on the tutorial website at https://convsearch.github.io/

Overview

Part I: Fundamentals of Conversational Search [90 min]

- Introduction to conversational search
- Conversational search basic paradigms
- Conversational retrieval-augmented generation
- Mixed initiative
- Personalization

Part II: Emerging Topics in the Agent Era [90 min]

- Agentic search
- Agentic conversational search
- Proactive conversational agents
- Domain-specific applications

Part III: Summary and Future Directions

Part 1

Fundamentals of Conversational Search

Conversational Search

General goal: Conversational search aims to identify relevant documents to satisfy users' complex information needs through multi-turn interactions.

Conversational Search v.s. **Ad-hoc Search**:

- ➢ Multi-turn interaction v.s. Single-turn search
- ➢ Natural language based query v.s. Keyword based query
- ➢ Flexible interface and return forms v.s. Fixed page links return

Comparison between Conversational and Ad-hoc Search

what is information retrieval *Query 1* ---▸ tell me some of its famous scholars *Query 2*

Wikipedia
https://en.wikipedia.org › wiki › Information_retrieval

Information retrieval

Information retrieval is the science of **searching for information in a document**, searching for documents themselves, and also searching for the metadata that ...

This list of the greatest scholars includes **Angela Davis, Sigmund Freud, Cornel West, Nicolaus Copernicus**, and more. From reputable, prominent, and well known scholars to the lesser known scholars of today, these are some of the best professionals in the scholar field.

Ranker
https://www.ranker.com › People

List of Famous Scholars - Ranker

Ad-hoc Search

Sources

What Is Information Retrieval? - Coveo	Information retrieval - Wikipedia	Information Retrieval: Everything You Need to Know When ... - Alooba	
coveo · 1	en.wikipedia · 2	alooba · 3	View 5 more

Answer

Information retrieval (IR) is a crucial field within computer science and information science that focuses on the process of identifying and retrieving information from a collection of resources in response to a user's information need. This process can involve various types of data, including text documents, images, audio, and metadata.

Sources

Online edition (c)©2009 Cambridge UP	[PDF] The History of Information Retrieval Research	Information retrieval - Wikipedia	
nlp.stanford · 1	clir-publications.... · 2	en.wikipedia · 3	View 5 more

Answer

Some famous scholars who have made significant contributions to the field of information retrieval include:

Early Pioneers

- **Calvin Mooers** - Coined the term "information retrieval" in 1948/1950 [1] [3].
- **Hans Peter Luhn** - Began work on mechanized punch card-based systems for searching chemical compounds in the late 1940s and published "Auto-encoding of documents for information retrieval" in 1959 [3].
- **Gerard Salton** - Began work on IR at Harvard in the early 1960s, later moved to Cornell. Published "Automatic Information Organization and Retrieval" in 1968 [3].

Influential Researchers

- **Karen Spärck Jones** - Finished her thesis "Synonymy and Semantic Classification" at Cambridge in 1964 and continued work on computational linguistics as it applies to IR [3].

Conversational Search

Why Conversational Search is Important?

- **Natural Interaction** - feel like talking to a human
- **Context Awareness** - understand follow-up queries and refine results
- **Handles Complex Queries** - support clarification, refinement, and reasoning
- **Improves User Experience:**
  - reduces the need of query reformulation
  - friendly for non-technical users
  - delivers more precise, personalized results
- Etc.

Introduction for Conversational Search

Introduction

User queries in conversational search

- Context-dependent query
  - Query: How many rings does he have? (what rings? who is he?)
- Ambiguous query
  - Query: What is the price of apple? (fruit or any apple products)
- Topic-Switch
  - Previous Query: When was the byzantine empire born? (Topic: History)
  - Query: What is its famous tourist places now? (Topic: Tourism)
- Etc.

Introduction

Conversational search systems capacity

- ➢ Context-dependent query ⇒ Understand real search intent via context modeling
- ➢ Ambiguous query ⇒ Search intent clarification (Mixed Initiatives)
- ➢ Topic-Switch ⇒ Context denoising via turn relevance/usefulness
- ➢ Etc.

Introduction

Conversational search systems capacity

- Understand real search intent via context modeling
  - Q1: Who is the best player in NBA so far? R1: Michael Jordan.
  - Q2: How many rings does he have?
  - ⇒ How many NBA championship rings does Michael Jordan have?
- Search intent clarification (Mixed Initiatives)
- Context denoising via turn relevance/usefulness
- Etc.

Introduction

Conversational search systems capacity

- Understand real search intent via context modeling
- Search intent clarification (Mixed Initiatives)
  - What is the price of apple here?
  - ⇒ Are you requesting for the price of apple fruit or any digital products from apple company?
- Context denoising via turn relevance/usefulness
- Etc.

Introduction

Conversational search systems capacity

- Understand real search intent via context modeling
- Search intent clarification (Mixed Initiatives)
- Context denoising via turn relevance/usefulness
  - Q1: When was the byzantine empire born? (Relevant)
  - Q3: Which battle or event marked the fall of this empire?
  - Q5: Can you name some of major cities in Turkey? (Relevant)
  - Current Query: Were any of these cities associated with the first empire you were discussing?

Introduction

User-system interactions in conversational search

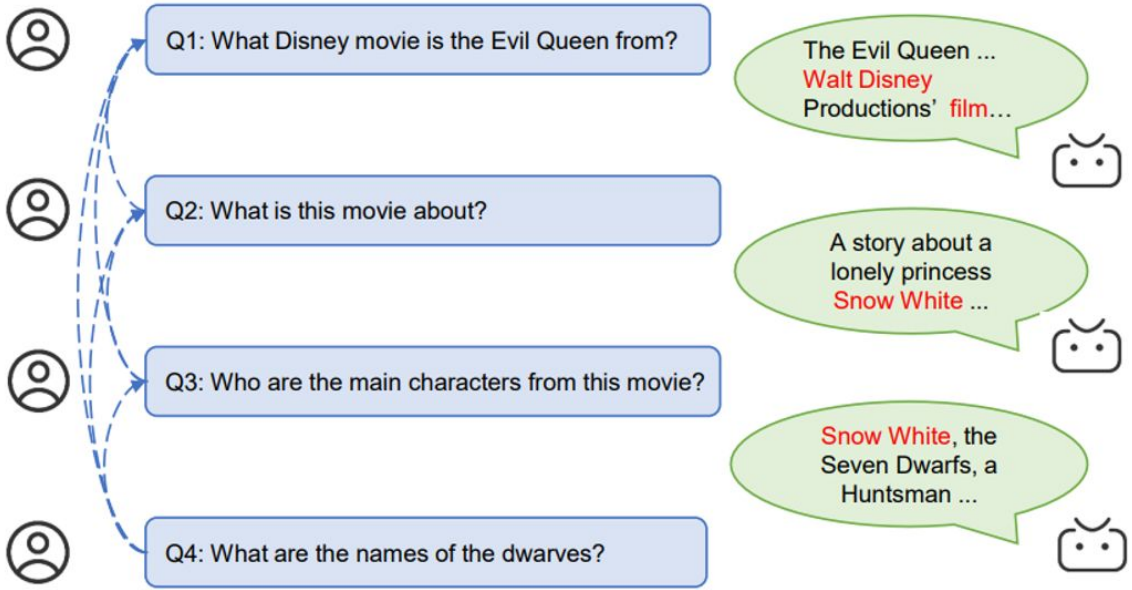
- Context-dependent query ⇔ Understand real search intent
- Ambiguous query ⇔ Search intent clarification (Mixed Initiatives)
- Topic-Switch ⇔ Context denoising via turn relevance/usefulness
- Etc.

The goal is to understand and satisfy users' complex information needs under multi-turn natural language based conversations with flexible input and interface.

Problem Definition

Task Formulation of Conversational Search

Given history context $H^k = \{q_k, r_k\}_{k=1}^{n-1}$, find the relevant passage p_i^* for the current query q_i, from a large collection C. (Then, generate the final response on top of the retrieval.)



Widely Used Datasets

From NLP community

- TopiOCQA [1], QReCC [2], INSCIT [3], CORAL [4], etc.

From IR community

- TREC CAsT 2019-2023 [5] and TREC iKAT 2023-2025 [6]
- OR-QuAC [7], ProCIS [8]
- Etc.

[1] TopiOCQA: Open-domain Conversational Question Answering with Topic Switching. Adlakha et al. TACL 2022.
[2] Open-Domain Question Answering Goes Conversational via Question Rewriting. Anantha et al. NAACL 2021.
[3] InSCIt: Information-Seeking Conversations with Mixed-Initiative Interaction. Wu et al. TACL 2023.
[4] CORAL: Benchmarking Multi-turn Conversational Retrieval-Augmentation Generation. Cheng et al. NAACL 2024.
[5] https://github.com/daltonj/treccastweb
[6] https://www.trecikat.com/
[7] Open-retrieval conversational question answering. Qu et al. SIGIR 2020.
[8] ProCIS: A benchmark for proactive retrieval in conversations. Samarinas et al. SIGIR 2024.

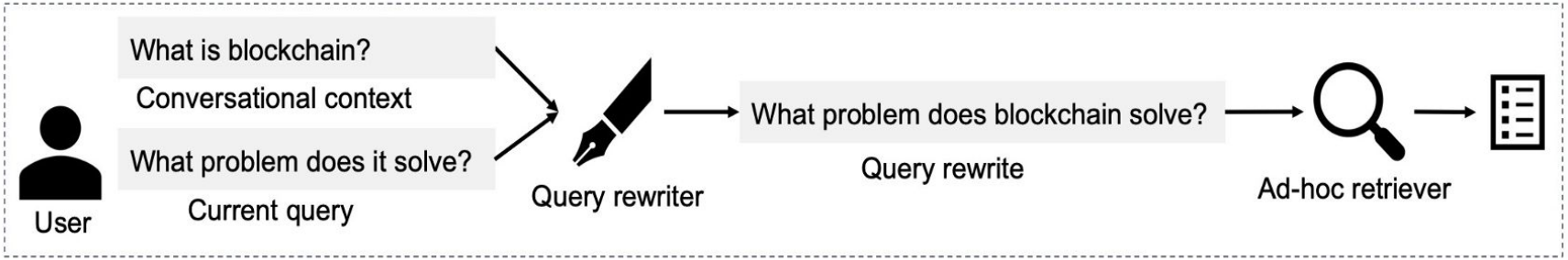
Two Paradigms to achieve Conversational Search

1. Conversational Query Rewriting
2. Conversational Dense Retrieval

Two Conversational Search Paradigms

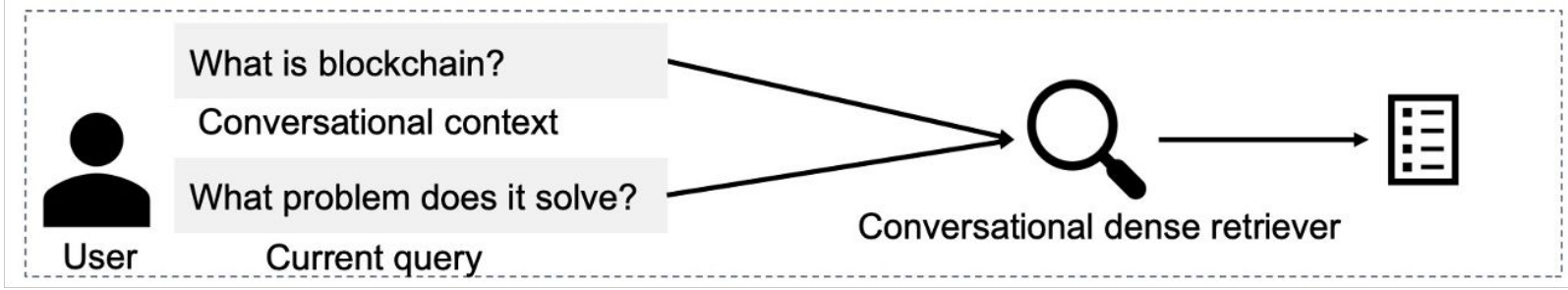
Conversational Query Rewriting (CQR)

➢ Idea: Transform a context-dependent query into an explicit rewritten query.



Conversational Dense Retrieval (CDR)

➢ Idea: Obtain a conversational dense retriever with contextual representation.



Conversational Query Rewriting

Conversational query rewriting methodologies in literature:

Approaches of earlier studies:

- Selecting useful terms from historical context.
- Rewriting context-dependent query to mimic human-rewritten one.
- Leveraging search task signals for rewriter model training.

Under large language models (LLMs) era:

- Prompting LLMs to directly rewrite context-dependent query.
- Leverage LLMs to generate better rewritten query as training signals.

Conversational Query Rewriting

Selecting useful terms from historical context

➢ **Idea**: Context from the conversational history can be used to arrive at a better expression of the current turn query [1].

Turn	Query
1	who formed **saosin**?
2	when was the **band** founded?
3	what was their **first** album?
4	when was the album released?
	resolved: when was saosin 's first album released?

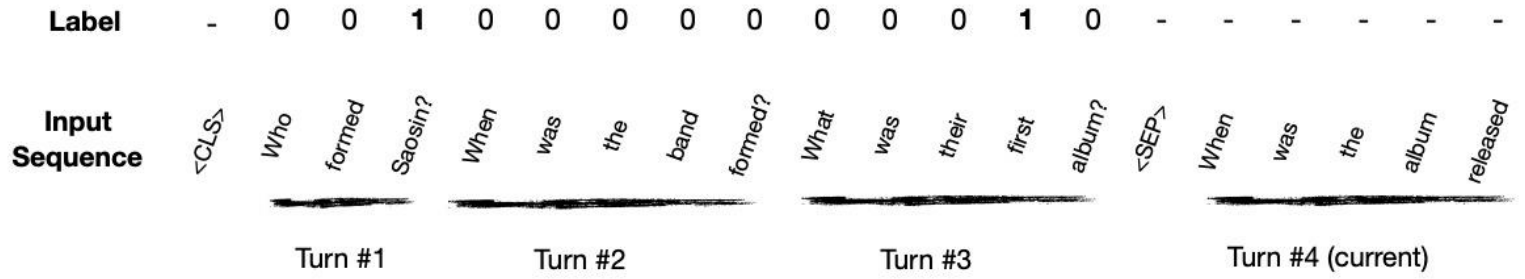
Relevant passage to turn #4: The original lineup for **Saosin**, consisting of Burchell, Shekoski, Kennedy and Green, was formed in the summer of 2003. On June 17, the **band** released their **first** commercial production, the EP Translating the Name.

[1] Query resolution for conversational search with limited supervision. Voskarides et al. SIGIR 2020.

Conversational Query Rewriting

Selecting useful terms from historical context

- **Challenge**: The token-level usefulness annotations are unavailable.
- [1,2,3] propose to generating token-level pseudo relevant labels and use them to train a binary classifier or selector to select useful terms in the context.



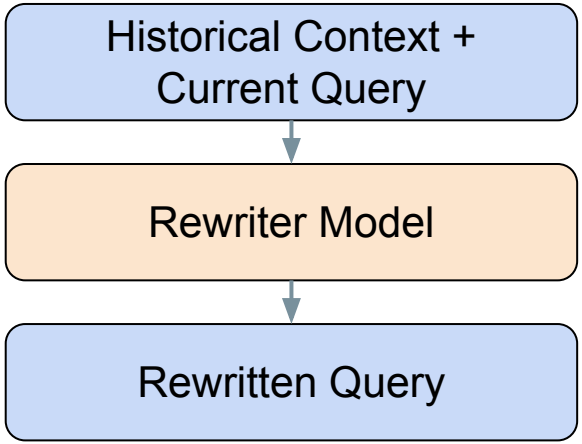
- The selected relevant terms could act as query expansion, but could be noisy.

[1] Query resolution for conversational search with limited supervision. Voskarides et al. SIGIR 2020.
[2] Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting. Lin et al. TOIS 2021.
[3] Contextualized Query Embeddings for Conversational Search. Lin et al. EMNLP 2021.

Conversational Query Rewriting

Rewriting context-dependent query to mimic human-rewritten one

➢ **Idea**: [1,2,3,4] Train a generative rewriter via the pairs of context and rewrites.



Turn	Conversational Queries
Q_1	Tell me about the Bronze Age collapse.
Q_2	What is the evidence for it?
Q_3	What are some of the possible causes?
Manual Query Rewrites	
Q_2^*	What is the evidence for **the Bronze Age collapse**?
Q_3^*	... the possible causes **of the Bronze Age collapse**?

➢ **Cons**: Cannot optimize with downstream search task and rely on manual labels.

[1] Few-shot generative conversational query rewriting. Yu et al. SIGIR 2020.
[2] Question rewriting for conversational question answering. Vakulenko et al. WSDM 2021.
[3] A Comparison of Question Rewriting Methods for Conversational Passage Retrieval. Vakulenko et al. ECIR 2021.
[4] Explicit query rewriting for conversational dense retrieval. Qian et al. EMNLP 2022.

Conversational Query Rewriting

Leveraging search task signals for rewriter model training

- **Idea**: [1,2,3,4] enhance the learning of rewriter with search task signals.
- **Approach**: Contain two optimization parts, query generation and search signals in the training objective. The search signals could be formulated as representation fine-tuning [3,4] or reinforcement learning [1,2].

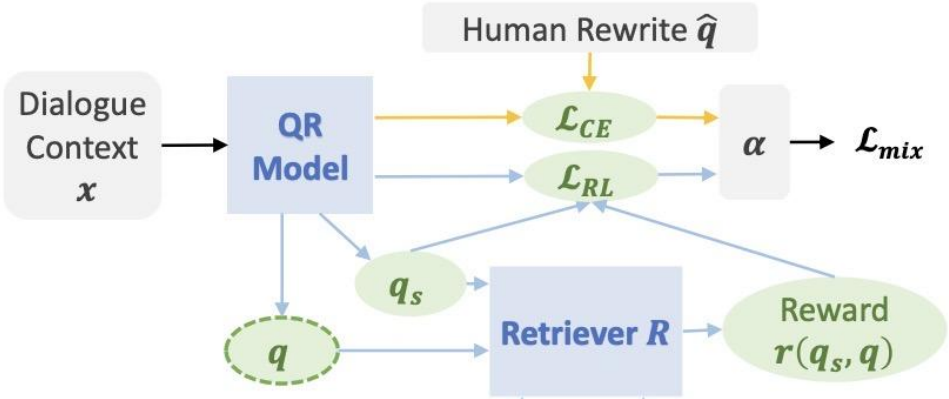
$$L_{Final} = L_{q-gen} + \alpha \cdot L_{search}$$

[1] CONQRR: Conversational Query Rewriting for Retrieval with Reinforcement Learning. Wu et al. EMNLP 2022.
[2] Reinforced Question Rewriting for Conversational Question Answering. Chen et al. EMNLP 2022.
[3] ConvGQR: Generative Query Reformulation for Conversational Search. Mo et al. ACL 2023.
[4] Search-Oriented Conversational Query Editing. Mao et al. ACL 2023.

Conversational Query Rewriting

Leveraging search task signals for rewriter model training

➢ **Approach**: The search signals could be formulated as representation fine-tuning [3,4] or reinforcement learning [1,2].



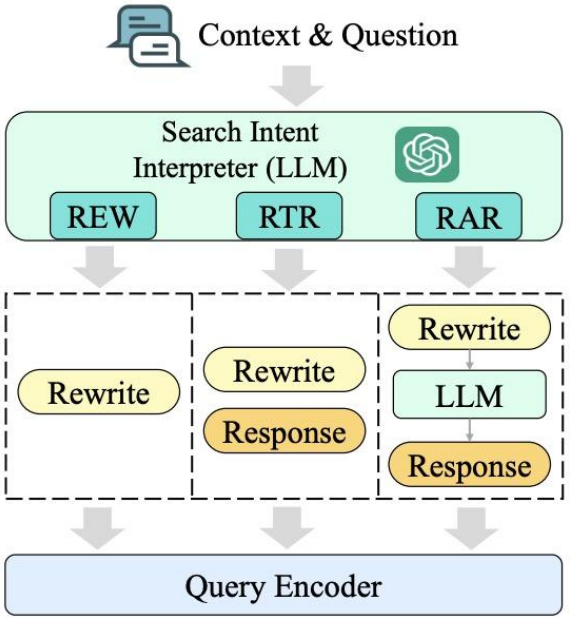
➢ **Pros**: Optimizing query generation toward search task.

[1] CONQRR: Conversational Query Rewriting for Retrieval with Reinforcement Learning. Wu et al. EMNLP 2022.
[2] Reinforced Question Rewriting for Conversational Question Answering. Chen et al. EMNLP 2022.
[3] ConvGQR: Generative Query Reformulation for Conversational Search. Mo et al. ACL 2023.
[4] Search-Oriented Conversational Query Editing. Mao et al. ACL 2023.

Conversational Query Rewriting

Prompting LLMs to directly rewrite context-dependent query

- **Idea**: Leveraging LLMs' conversation understanding and text generation capacity to grasp users' contextual search intent [1].
- **Approach**: Design prompts from various aspects [2,3] to generate query.
  - LLM4CS [1]: generate different types of queries and then aggregate them.



[1] Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search. Mao et al. EMNLP 2023.
[2] Enhancing Conversational Search: Large Language Model-Aided Informative Query Rewriting. Ye et al. EMNLP 2023.
[3] CHIQ: Contextual History Enhancement for Improving Query Rewriting in Conversational Search.. Mo et al. EMNLP 2024.

Conversational Query Rewriting

Prompting LLMs to directly rewrite context-dependent query

- **Observation**: LLM-based query rewriting could obtain much better results [1] compared to SLM-based query rewriter [2,3].
- **Limitations**:
  - High inference cost by calling LLMs (multiple times) for each query.
  - The rewritten query might still contain noise and cannot generalize.

System	CAsT-19			CAsT-20			CAsT-21		
	MRR	NDCG@3	R@100	MRR	NDCG@3	R@100	MRR	NDCG@3	R@100
T5QR	0.701	0.417	0.332	0.423	0.299	0.353	0.469	0.330	0.408
ConvGQR	0.708	0.434	0.336	0.465	0.331	<u>0.368</u>	0.433	0.273	0.330
LLM4CS	**0.776**†	**0.515**†	**0.372**†	**0.615**†	**0.455**†	**0.489**†	**0.681**†	**0.492**†	**0.614**†

[1] Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search. Mao et al. EMNLP 2023
[2] Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. Lin et al. arXiv 2020
[3] ConvGQR: Generative Query Reformulation for Conversational Search. Mo et al. ACL 2023.

Conversational Query Rewriting

Leverage LLMs to generate better rewritten query as training signals

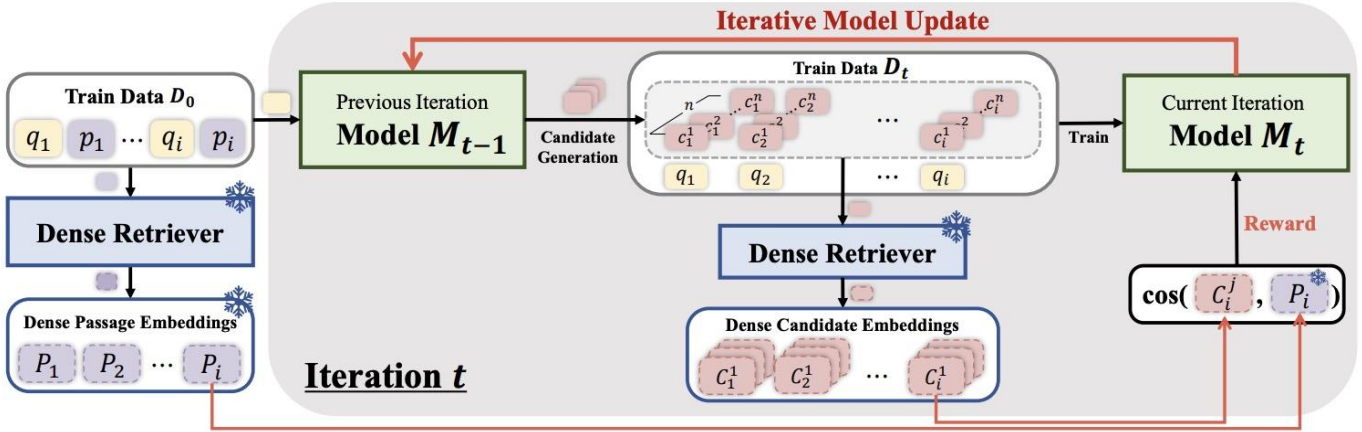
- **Assumption**: The human-rewritten query might be sub-optimal [1] as a search query.
- **Motivation**: Leverage small LM for query rewriting to reduce latency.
- **Idea**: Use LLMs to generate better pseudo query with qualified signal (e.g., relevance judgment [2,3], search reward [4,5]) for model training, similar to knowledge distillation from LLMs.

[1] ConvGQR: Generative Query Reformulation for Conversational Search. Mo et al. ACL 2023.
[2] IterCQR: Iterative Conversational Query Reformulation without Human Supervision. Jang et al. NAACL 2023.
[3] CHIQ: Contextual History Enhancement for Improving Query Rewriting in Conversational Search.. Mo et al. EMNLP 2024.
[4] ADACQR: Enhancing Query Reformulation for Conversational Search via Sparse and Dense Retrieval Alignment. Lai et al. COLING 2024.
[5] Adaptive Query Rewriting: Aligning Rewriters through Marginal Probability of Conversational Answers. Zhang et al. EMNLP 2024.

Conversational Query Rewriting

Leverage LLMs to generate better rewritten query as training signals

➢ **Approach**: [1] iteratively update training signals and model based on LLM multi-rounds generated signals.



[1] IterCQR: Iterative Conversational Query Reformulation without Human Supervision. Jang et al. NAACL 2023.

Conversational Query Rewriting

Summary of CQR paradigm:

- **Pros**: Can re-use any existing retrievers and has good interpretability with explicit rewritten query.
- **Cons**: Cannot directly optimize with downstream search task and the rewriter model training rely on available annotations as supervision signals.
- **Open question**:
  - Does LLM already solve conversational query rewriting?
  - How to deal with instruction-following style long query in LLM era?

Q & A

Conversational Dense Retrieval

Conversational dense retrieval methodologies in literature:

- Explicit and implicit context denoising
- Data augmentation for query-documents relevance judgments
- Leveraging more conversational signals for dense retrieval training
- Generative LLM-based conversational dense retrieval

Conversational Dense Retrieval

Assumption: Not all historical turn are relevant for the current turn search [1,2].

Explicit context denoising

➢ **Idea**: First developing some mechanisms to identify the useful/relevant historical context and then use the context to enhance dense retrieval [1,2].

Implicit context denoising

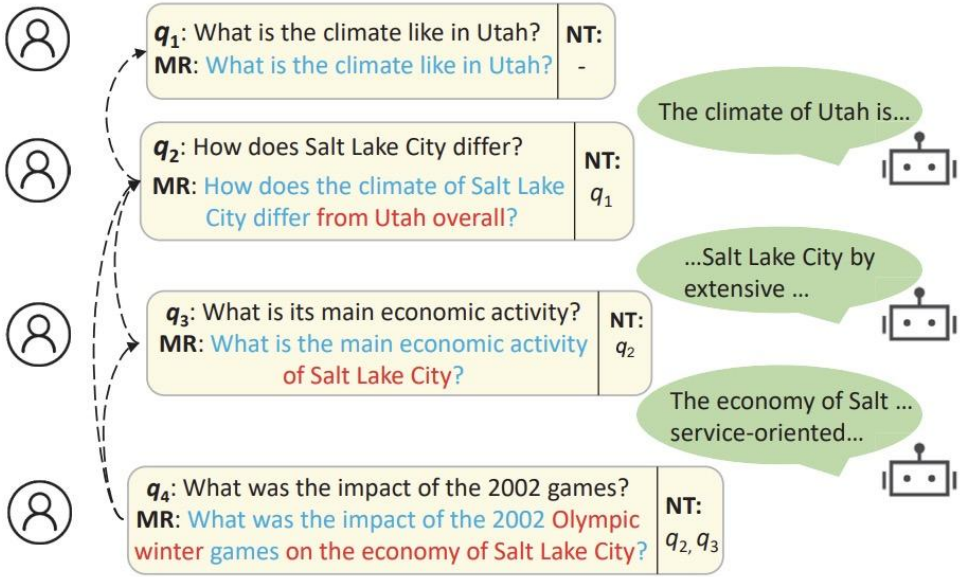
➢ **Idea**: Enable the dense retriever to implicitly identify and pay less attention to noisy/irrelevant historical context [3,4].

[1] Curriculum contrastive context denoising for few-shot conversational dense retrieval. Mao et al. SIGIR 2022.
[2] Learning to relate to previous turns in conversational search. Mo et al. SIGKDD 2023.
[3] Learning denoised and interpretable session representation for conversational search. Mao et al. WWW 2023..
[4] History-aware conversational dense retrieval.. Mo et al. ACL 2024.

Conversational Dense Retrieval

Explicit and implicit context denoising

- **Key challenge**: Turn relevance annotation is unavailable.
- Human-annotated turn relevance based on topic information [1].
- **Cons**:
  - Judgments are subjective.
  - Cannot scaling-up.

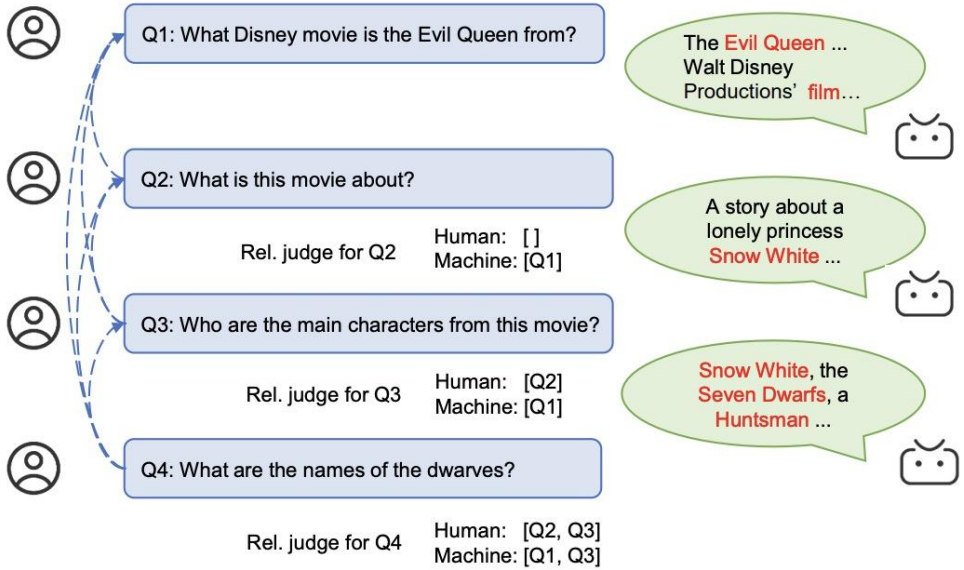


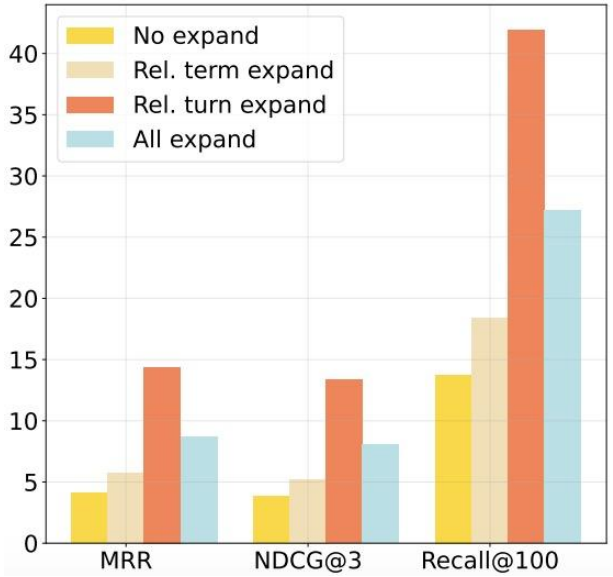
[1] Curriculum contrastive context denoising for few-shot conversational dense retrieval. Mao et al. SIGIR 2022.

Conversational Dense Retrieval

Explicit and implicit context denoising

- [1,2] conducts pseudo labeling for the context based on the impact on retrieval results of a candidate turn or term, which is used to expand the query.
- Example: If $Score(q_n) < Score(q_n * q_i)$, we assume q_i is relevant to q_n.





[1] Learning to relate to previous turns in conversational search. Mo et al. SIGKDD 2023.
[2] History-aware conversational dense retrieval.. Mo et al. ACL 2024.

Conversational Dense Retrieval

Data augmentation for query-documents relevance judgments

- **Idea**: Generating more query-document relevance judgments to address the data scarcity issue [1,2] — conversational search systems are not widely deployed.
- **Key challenge**:
  - The conversation session should be consistent and aligned with query-documents relevance judgments [2,3].
  - The distribution between generated data and evaluated benchmark [4].

[1] Dialog inpainting: Turning documents into dialogs. Dai et al. ICML 2022.
[2] Convtrans: Transforming web search sessions for conversational dense retrieval. Mao et al. EMNLP 2022.
[3] ConvSDG: Session Data Generation for Conversational Search. Mo et al. WWW 2024 @LLM4IR.
[4] Generalizing conversational dense retrieval via llm-cognition data augmentation. Chen et al. ACL 2024.

Conversational Dense Retrieval

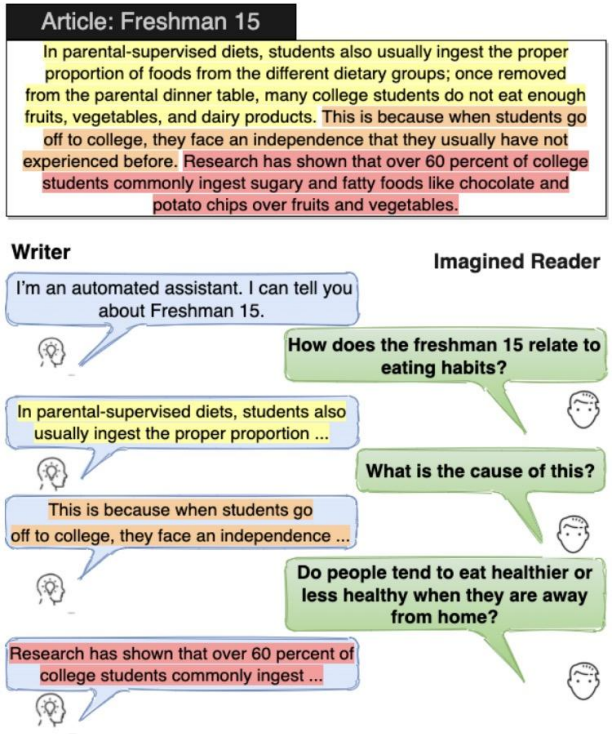
Data augmentation for query-documents relevance judgments

- **Solutions for generating conversational search session**:
  - From documents to simulate a user-system interaction [1].
  - From session search data to reuse relevance judgments [2].
  - From existing conversational search session by rewriting each turn [3].
  - From existing conversational search session to enhance diversity [4].

[1] Dialog inpainting: Turning documents into dialogs. Dai et al. ICML 2022.
[2] Convtrans: Transforming web search sessions for conversational dense retrieval. Mao et al. EMNLP 2022.
[3] ConvSDG: Session Data Generation for Conversational Search. Mo et al. WWW 2024 @LLM4IR.
[4] Generalizing conversational dense retrieval via llm-cognition data augmentation. Chen et al. ACL 2024.

Conversational Dense Retrieval

Simulation from a passage [1]



Transfer from session search [2]

Search Session Example

Web: William Shakespeare

Conv: Who is William Shakespeare?

Intent: *William Shakespeare*

Web: William Shakespeare poems

Conv : Can you show some his poems?

Intent: *William Shakespeare poems*

Web: When WS wrote hamlet

Conv : When he wrote hamlet?

Intent: *Hamlet's finished time*

Rewrite from existing data [3]

Generated Augmented Query

q_1: What is a physician's assistant?

q_1': What does the term "physician's assistant" mean?

...

q_{12}: How much longer does it take to become a doctor after being an NP?

q_{12}': How much additional time is required to become a physician after becoming a nurse practitioner (NP)?

[1] Dialog inpainting: Turning documents into dialogs. Dai et al. ICML 2022.
[2] Convtrans: Transforming web search sessions for conversational dense retrieval. Mao et al. EMNLP 2022.
[3] ConvSDG: Session Data Generation for Conversational Search. Mo et al. WWW 2024 @LLM4IR.

Conversational Dense Retrieval

Leveraging more conversational signals for dense retrieval training

- **Idea**: Using additional signal mined from conversational scenarios for dense retriever training, e.g., rewritten query, conversational hard negatives.
- [1,3] leverage rewritten query and relevance judgment for model training.

$$\mathcal{L} = -\log \frac{\exp\left(q_n^s \cdot d_n^+\right)}{\exp\left(q_n^s \cdot d_n^+\right) + \sum_{d_n^- \in D} \exp\left(q_n^s \cdot d_n^-\right)} + \mathrm{MSE}(q_n^s, q_n')$$

- [2,4] mine additional hard negatives from historical turns as contrastive samples.
  - From conversational query rewriting model [2]
  - From irrelevant historical turns' positive documents [4]

[1] Few-shot conversational dense retrieval. Yu et al. SIGIR 2021.
[2] Saving dense retriever from shortcut dependency in conversational Search. Kim et al. EMNLP 2022.
[3] Aligning Query Representation with Rewritten Query and Relevance Judgments in Conversational Search. Mo et al. CIKM 2024.
[4] History-aware conversational dense retrieval.. Mo et al. ACL 2024.

Conversational Dense Retrieval

Generative LLM-based conversational dense retrieval

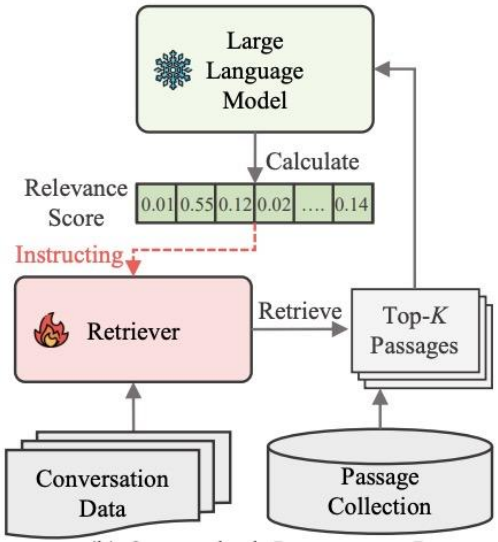
- **Idea**: Using the powerful LLM with high capacity to facilitate conversational dense retriever fine-tuning.
- [1,4] leverage the semantic feature distilled from LLM to improve the conversational dense retriever fine-tuning based on small language models.
- [2,3] use LLM as backbone to fine-tune for retrieval and conversation tasks.

[1] InstructoR: Instructing Unsupervised Conversational Dense Retrieval with Large Language Models. Jin et al. EMNLP 2023.
[2] ChatRetriever: Adapting Large Language Models for Generalized and Robust Conversational Dense Retrieval. Mao et al. EMNLP 2024.
[3] UniConv: Unifying Retrieval and Response Generation for Large Language Models in Conversations. Mo et al. ACL 2025.
[4] DiSCo: LLM Knowledge Distillation for Efficient Sparse Retrieval in Conversational Search. Lupart et al. SIGIR 2025.

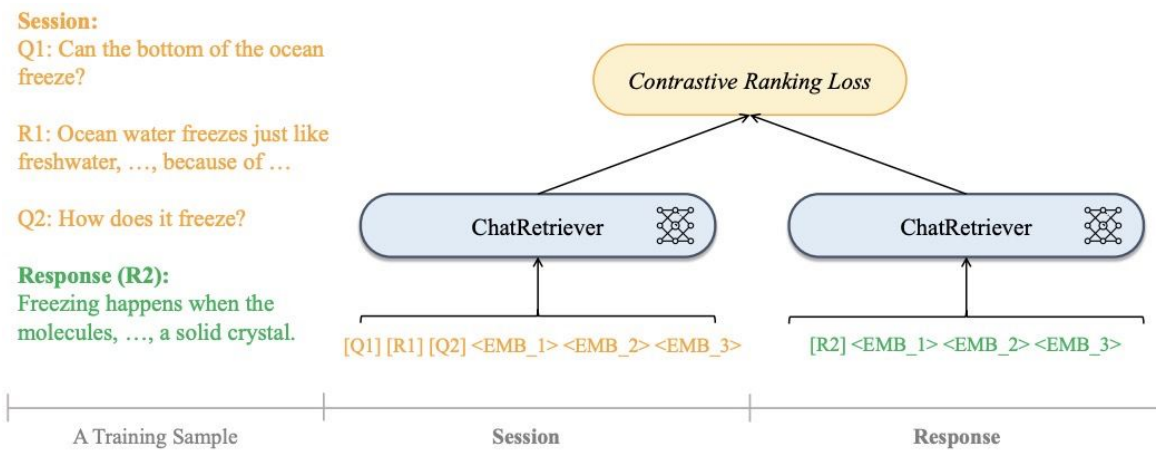
Conversational Dense Retrieval

Generative LLM-based conversational dense retrieval

Distill features from LLM [1]



Use LLM as backbone to fine-tune for retrieval and conversation tasks [2]



More details in Part II

[1] InstructoR: Instructing Unsupervised Conversational Dense Retrieval with Large Language Models. Jin et al. EMNLP 2023.
[2] ChatRetriever: Adapting Large Language Models for Generalized and Robust Conversational Dense Retrieval. Mao et al. EMNLP 2024.

Conversational Dense Retrieval

Summary:

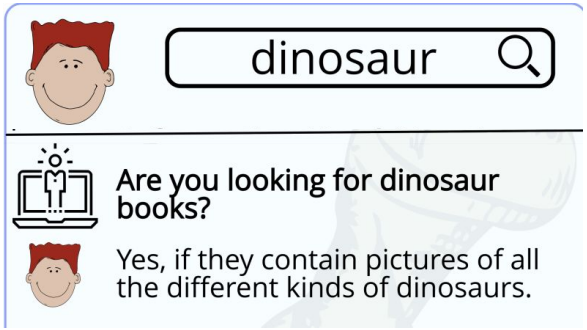
- **Pros**: Direct optimize with conversational session to obtain representation.
- **Cons**: Data scarcity problem and de-noising requirement for the input context.
- **Open question**:
  - How to improve efficiency and generalizability?
  - How to mine more conversational signals for better representation?

Q & A

Mixed Initiative

Mixed Initiative

- What is mixed initiative?
  - User and system can both take the initiative at different times in a conversation [1]
  - System can take the initiative to ask clarifying questions, elicit user preferences, ask for feedback, provide suggestions
  - User satisfaction has been reported to increase when prompted with system-initiatives, e.g., clarifications [2]



[1] Radlinski et al. A Theoretical Framework for Conversational Search. CHIIR 2017.
[2] Kiesel et al. Toward voice query clarification. SIGIR 2018.

Mixed Initiative

- Scope for mixed initiatives
  - What
    - Clarifying question selection/generation
  - When
    - Clarification need prediction
    - System initiative prediction

Mixed Initiatives

- Scope for mixed initiatives
  - **What**
    - **Clarifying question selection/generation**
  - When
    - Clarification need prediction
    - System initiative prediction

Mixed Initiative

- Clarifying question selection
  - [1] releases the Qulac dataset, where each query is associated with a set of human-generated questions
    - Retrieve a set of questions for a given query, and then select the best question by a BERT-based model (NeuQS)
    - Adding selected question improves document retrieval quality
  - [2] releases a larger dataset, ClariQ

Method	Qulac-T Dataset				
	MRR	P@1	nDCG@1	nDCG@5	nDCG@20
OriginalQuery	0.2715	0.1842	0.1381	0.1451	0.1470
σ-QPP	0.3570	0.2548	0.1960	0.1938	0.1812
LambdaMART	0.3558	0.2537	0.1945	0.1940	0.1796
RankNet	0.3573	0.2562	0.1979	0.1943	0.1804
NeuQS	**0.3625***	**0.2664***	**0.2064***	**0.2013***	**0.1862***
WorstQuestion	0.2479	0.1451	0.1075	0.1402	0.1483
BestQuestion	0.4673	0.3815	0.3031	0.2410	0.2077

[1] Aliannejadi et al. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. SIGIR 2019.
[2] Aliannejadi et al. Building and Evaluating Open-Domain Dialogue Corpora with Clarifying Questions. EMNLP 2021.

Mixed Initiative

- Clarifying question generation
  - Selecting clarifying questions from a human-generated question set does not generalize well in real-world scenarios; training data is scarce
  - [1] learns to generate clarifying questions
    - Mine question templates from query reformulation data from Bing
    - Generate training data by selecting and filling out question templates
    - Train a sequence-to-sequence model on the data

(1) What do you want to know about QUERY?
(2) What do you want to know about this QUERY_ENTITY_TYPE?
(3) What ASPECT_ENTITY_TYPE are you looking for?
(4) Whom are you looking for?
(5) Who are you shopping for?

[1] Zamani et al. Generating Clarifying Questions for Information Retrieval. WWW 2020.

Mixed Initiative

- Clarifying question generation
  - [1,2] finetunes BART, while [3] fine-tunes GPT-2
  - [3] argues that more semantic guidance is needed
    - Fine-tine GPT-2 conditioned on a facet and the user query
      - facet [SEP] user query [BOS] →clarifying question [EOS]
  - [4] extracts facets from documents retrieved by the user query

Initial request	Tell me about kiwi	
Facet terms	information fruit	biology bird
Template-based	Are you interested in information fruit?	Are you interested in biology bird?
Q-GPT (only query)	Are you looking for kiwi clothing?	Are you looking for kiwi reviews?
QF-GPT (+facet)	Are you interested in kiwi fruit?	Are you interested in kiwi birds?

[1] Guo et al. Abg-CoQA: Clarifying Ambiguity in Conversational Question Answering. AKBC 2021.
[2] Lee et al. Asking Clarification Questions to Handle Ambiguity in Open-Domain QA. EMNLP 2023.
[3] Sekulić, et al. Towards Facet-Driven Generation of Clarifying Questions for Conversational Search. ICTIR 2021.
[4] Sekulić, et al. Exploiting Document-Based Features for Clarification in Conversational Search. ECIR 2022.

Mixed Initiative

- Clarifying question generation
  - Heavily relies on human-annotated data; open-ended generation often yields non-clarifying text, and ignores given facets
  - [1] uses GPT-2 for zero-shot clarifying question generation
    - Template-initiated generation: use clarifying question templates as the starting text of the generation
    - Facet-constrained generation: use facet words as constraints during generation decoding

Templates	Generated clarifying question
would you like to	would you like to [take pictures of]
do you want to	do you want to [see pictures of them]
are you interested in	are you interested in [taking pictures of them]
are you looking for	are you looking for [pictures of South Africa]

For the query "I am looking for information about South Africa."

[1] Wang et al. Zero-shot Clarifying Question Generation for Conversational Search. WWW 2023.

Mixed Initiative

- Clarifying question generation
  - [1,2] use LLMs with better capabilities
    - Apply few-shot [1,2] and chain-of-thought (CoT) prompting [1,2] to improve performance

---

Few-shot w/ CoT: Given a query, answer the question or ask a clarifying question. The response should start with "The answer is" or "The clarifying question is".

Question: Who played michael myers in rob zombie's movie?
Output: In Rob Zombie's "Halloween" films, the role of Michael Myers was primarily played by Tyler Mane. Mane portrayed the iconic horror character in both Rob Zombie's "Halloween" (2007) and its sequel, "Halloween II" (2009). Therefore, the question is not ambiguous. The answer is Tyler Mane.

Question: Give me some Mother's Day gift ideas.
Output: there could be underlying ambiguities depending on the interests of the specific mother in question, the budget, and the giver's relationship to the mother. Without this context, a broad range of ideas could be considered appropriate. Therefore, the question is ambiguous. The clarifying question is: What are the interests or hobbies of the mother, and is there a particular budget range for the gift?

Question: <Question>

---

[1] Zhang et al. CLAMBER: A Benchmark of Identifying and Clarifying Ambiguous Information Needs in Large Language Models. ACL 2024.
[2] Deng et al. Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration. EMNLP 2023.

Mixed Initiative

- Clarifying question generation
  - Previous work with CoT prompting overlooks clarification-specific aspects
  - [1] Integrates ambiguity types in CoT prompting to improve clarifying question generation

Ambiguity Type	Definition
Semantic	The query is semantically ambiguous for several common reasons: it may include homonyms; a word in the query may refer to a specific entity while also functioning as a common word; or an entity mentioned in the query could refer to multiple distinct entities.
Generalize	The query focuses on specific information; however, a broader, closely related query might better capture the user's true information needs.
Specify	The query has a clear focus but may encompass too broad a research scope. It is possible to further narrow down this scope by providing more specific information related to the query.

[1] Tang et al. Clarifying Ambiguities: on the Role of Ambiguity Types in Prompting Methods for Clarification Generation. SIGIR 2025.

Mixed Initiative

- Clarifying question generation
  - Previous work with CoT prompting overlooks clarification-specific aspects
  - [1] Integrates ambiguity types in CoT prompting to improve clarifying question generation

```
Given a query in an information-seeking system, generate a clarifying question that you
think is most appropriate to gain a better understanding of the user's intent. The ambiguity
of a query can be multifaceted, and there are multiple possible ambiguity types:
<AT definitions>
Before generating the clarifying question, provide a textual explanation of your reasoning
about which types of ambiguity apply to the given query. Based on these ambiguity types,
describe how you plan to clarify the original query.
<query>
```

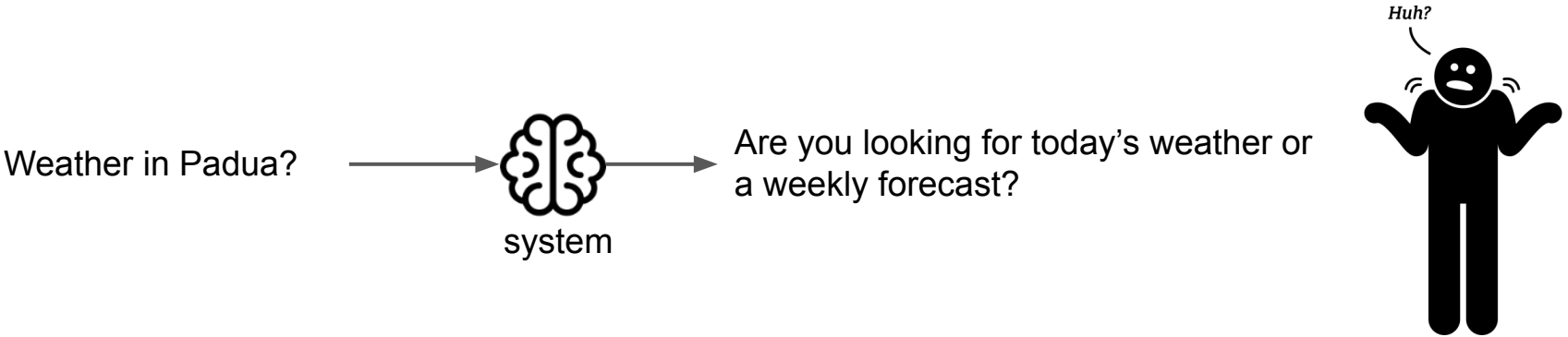
[1] Tang et al. Clarifying Ambiguities: on the Role of Ambiguity Types in Prompting Methods for Clarification Generation. SIGIR 2025.

Mixed Initiatives

- Scope for mixed initiatives
  - What
    - Clarifying question selection/generation
  - **When**
    - Clarification need prediction
    - System initiative prediction

Mixed Initiatives

- Why timing matters in taking initiative
  - Initiative-taking carries the risk of offending or overwhelming users, which can lower the overall user experience [1,2]



[1] Wang et al. Controlling the Risk of Conversational Search via Reinforcement Learning. WWW 2021.
[2] Wang et al. Simulating and Modeling the Risk of Conversational Search. TOIS 2022.

Mixed Initiatives

- Scope for mixed initiatives
  - What
    - Clarifying question selection/generation
    - Conversation contextualisation/interest anticipation
  - **When**
    - **Clarification need prediction**
    - System initiative prediction

Mixed Initiatives

- Clarification need prediction
  - [1,2,3] fine-tune pre-trained language models on human-annotated data
    - E.g., given the user query, [1] fine-tunes a model to output 1 (no need for clarification) to 4 (clarification is necessary)

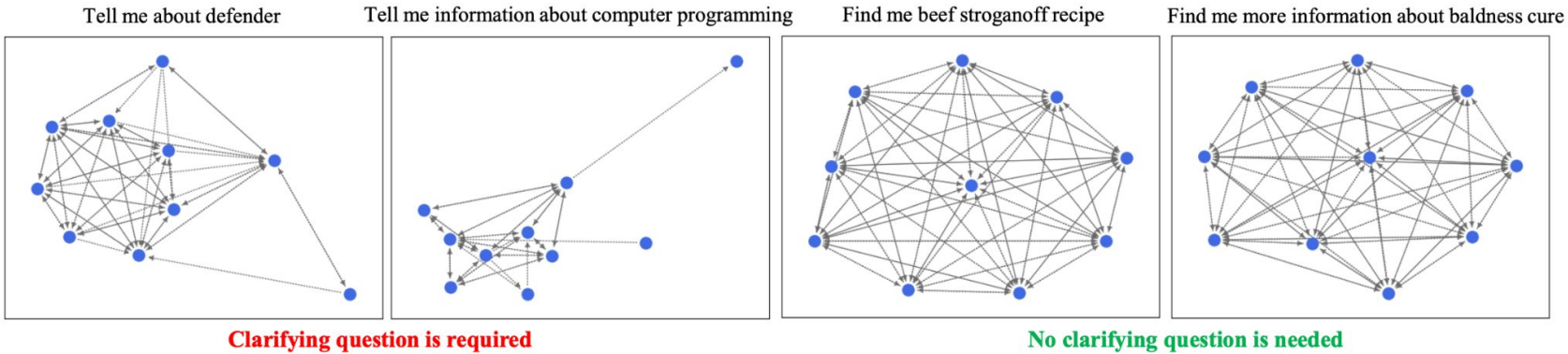
Model		Precision	Recall	F1-Measure	MSE
RoBERTa-based	dev	0.6039	0.5600	0.5551	0.6200
	test	**0.5981**	**0.6557**	**0.6070**	**0.5409**
BART	dev	0.7008	0.7000	0.6976	0.5200
	test	0.4813	0.4754	0.4756	0.7705
BERT-based	dev	0.5218	0.4800	0.5000	0.8200
	test	0.3931	0.4918	0.4253	0.6557

Results from [1] on clarification need prediction using ClariQ

[1] Aliannejadi et al. Building and Evaluating Open-Domain Dialogue Corpora with Clarifying Questions. EMNLP 2021.
[2] Guo et al. Abg-CoQA: Clarifying Ambiguity in Conversational Question Answering. AKBC 2021.
[3] Lee et al. Asking Clarification Questions to Handle Ambiguity in Open-Domain QA. EMNLP 2023.

Mixed Initiatives

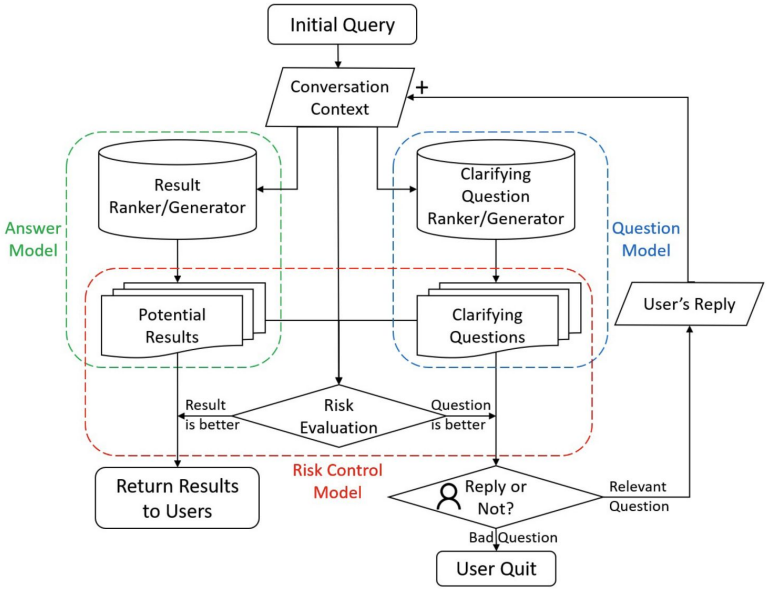
- Clarification need prediction
  - Existing studies rely on small-scale and costly human-annotated data
  - [1] proposes an unsupervised method, assuming that less ambiguous queries retrieve more coherent results
    - It builds a graph from retrieved items using context similarity, and uses graph connectivity as an ambiguity signal



[1] Arabzadeh et al. Unsupervised Question Clarity Prediction Through Retrieved Item Coherency. CIKM 2022.

Mixed Initiatives

- Clarification need prediction
  - Without using any human-annotated data, [1,2] train a model by reinforcement learning (RL), with rewards from a rule-based simulator



	Relevant	Irrelevant
Answer	Answer Reciprocal Rank	
Ask	r_{cq}	p_{cq}

Policy table from [1,2]

[1] Wang et al. Controlling the Risk of Conversational Search via Reinforcement Learning. WWW 2021.
[2] Wang et al. Simulating and Modeling the Risk of Conversational Search. TOIS 2022.

Mixed Initiatives

- Clarification need prediction
  - [1,2] use few-shot and CoT prompting
    - Model clarification need prediction, clarifying question generation and answer generation jointly [1,2]

Few-shot w/ CoT: Given a query, answer the question or ask a clarifying question. The response should start with "The answer is" or "The clarifying question is".

Question: Who played michael myers in rob zombie's movie?
Output: In Rob Zombie's "Halloween" films, the role of Michael Myers was primarily played by Tyler Mane. Mane portrayed the iconic horror character in both Rob Zombie's "Halloween" (2007) and its sequel, "Halloween II" (2009). Therefore, the question is not ambiguous. The answer is Tyler Mane.

Question: Give me some Mother's Day gift ideas.
Output: there could be underlying ambiguities depending on the interests of the specific mother in question, the budget, and the giver's relationship to the mother. Without this context, a broad range of ideas could be considered appropriate. Therefore, the question is ambiguous. The clarifying question is: What are the interests or hobbies of the mother, and is there a particular budget range for the gift?

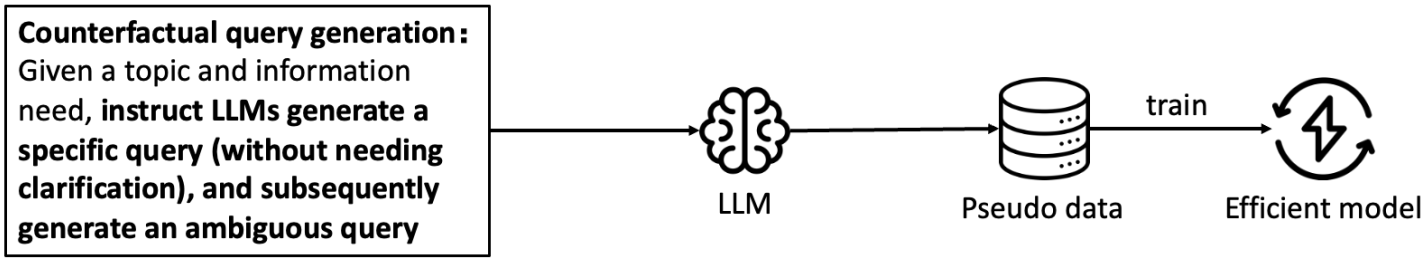
Question: <Question>

---

[1] Zhang et al. CLAMBER: A Benchmark of Identifying and Clarifying Ambiguous Information Needs in Large Language Models. ACL 2024.
[2] Deng et al. Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration. EMNLP 2023.

Mixed Initiatives

- Clarification need prediction
  - LLMs are inefficient, and training smaller models still relies on costly human-annotated data
  - [1] uses LLMs to generate data, and train efficient models (e.g., BERT) on the generated data
    - Propose counterfactual query generation mechanism, which is more effective than seperate generation
      - Efficient models trained on pseudo data outperform zero-shot/few-shot LLMs



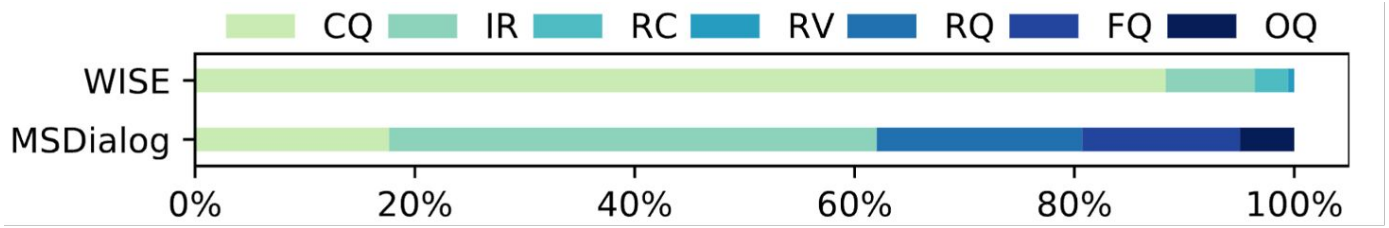
[1] Lu et al. Zero-Shot and Efficient Clarification Need Prediction in Conversational Search. ECIR 2025.

Mixed Initiatives

- Scope for mixed initiatives
  - What
    - Clarifying question selection/generation
    - Conversation contextualisation/interest anticipation
  - **When**
    - Clarification need prediction
    - **System initiative prediction**

Mixed Initiatives

- System initiative prediction (SIP)
  - Existing studies take a narrow view of system initiative, focusing mainly on clarification and ignoring other actions [1]

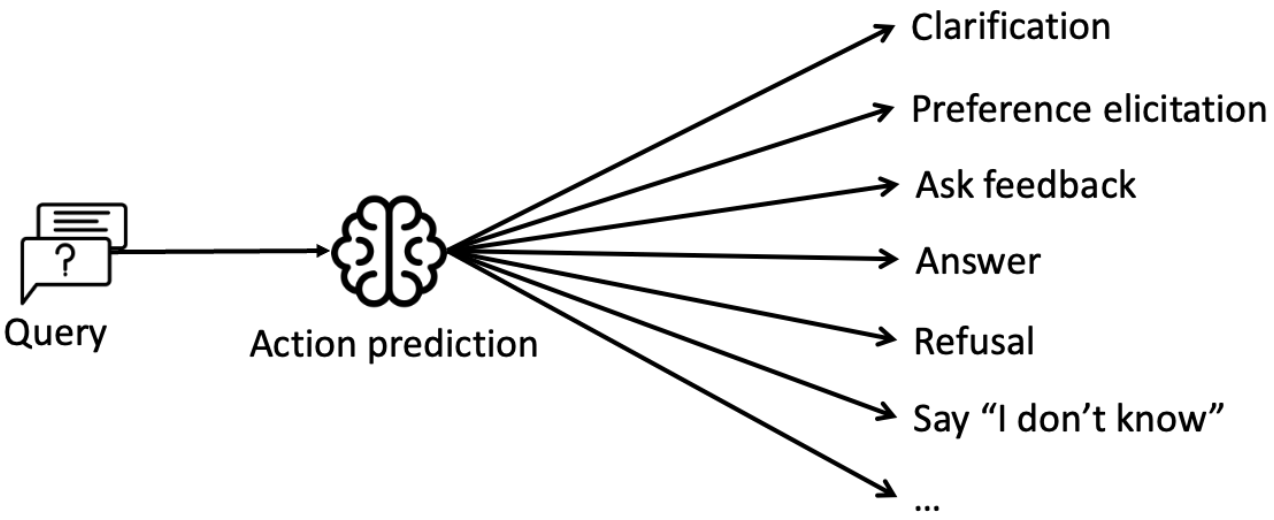


- **CQ**: clarifying question
- **IR**: information request
- **RV**: revise
- **RC**: recommendation
- **OQ**: original question
- **RQ**: repeat question
- **FQ**: Follow-up question

[1] Meng et al. System Initiative Prediction for Multi-turn Conversational Information Seeking. CIKM 2023.

Mixed Initiatives

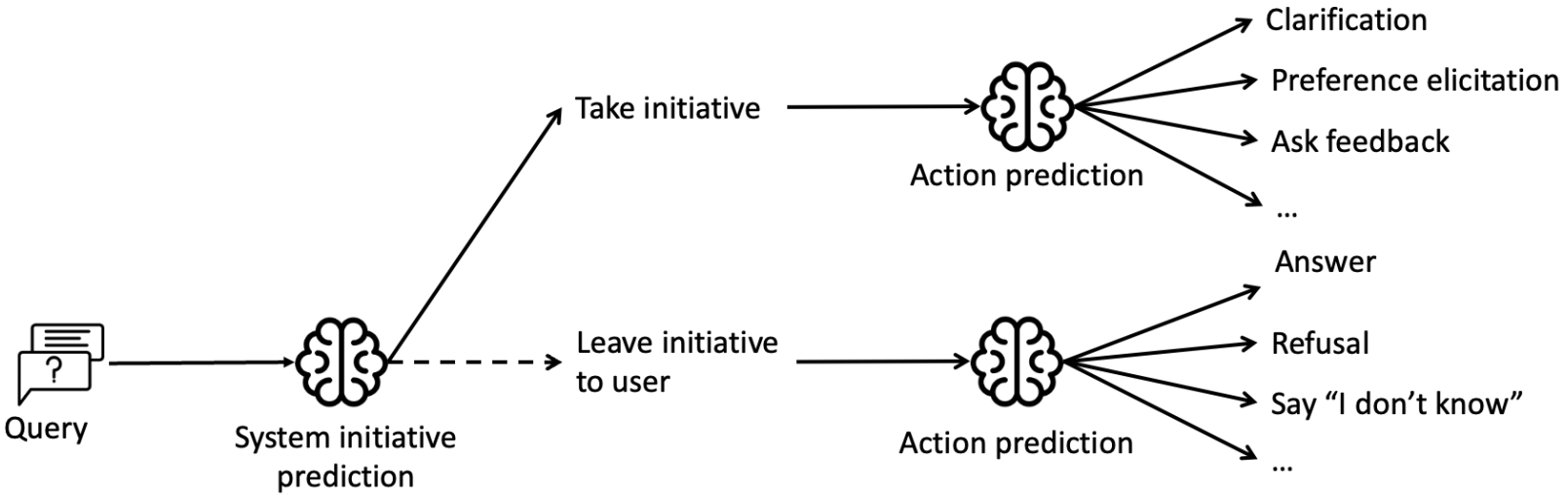
- System initiative prediction (SIP)
  - Directly predicting a system action from a large action space is challenging [1]



[1] Meng et al. System Initiative Prediction for Multi-turn Conversational Information Seeking. CIKM 2023.

Mixed Initiatives

- System initiative prediction (SIP)
  - [1] proposes SIP
    - Model SIP and action prediction into sequential steps
    - SIP-aware action prediction leads to improved effectiveness



[1] Meng et al. System Initiative Prediction for Multi-turn Conversational Information Seeking. CIKM 2023.

Mixed Initiatives

- System initiative prediction (SIP)
  - [1]'s empirical analysis reveals structural dependencies in SIP:
    - *System is more likely to take the initiative immediately after the user has taken the initiative in a conversation*
    - *System is less likely to take the initiative once again if the system has already taken the initiative before*
  - [1] models SIP with CRF, a probabilistic graphical method
    - outperform LLMs and exhibit transparency

Methods	MSDialog (%)			
	F1	Precision	Recall	Accuracy
LLaMA-7B	60.22	60.40	60.13	62.15
LLaMA-13B	62.54	62.73	63.21	62.99
LLaMA-33B	58.11	58.24	58.53	58.76
LLaMA-65B	55.30	62.33	60.44	55.93
BERT	60.17	60.25	60.12	61.86
Ours	**65.37**	**65.79**	**65.19**	**67.23***

[1] Meng et al. System Initiative Prediction for Multi-turn Conversational Information Seeking. CIKM 2023.

Conversational retrieval-augmented generation (RAG)

What are the new features of conversational search in the era of LLM?

Conversational Search in the LLM Era

User behaviour for information-seeking shift in the LLM Era:

- ➢ Interact with LLM application via natural language (Context Modeling)
- ➢ Refine their information needs (Query rewriting and Mix-initiative)

New features:

- ➢ Expect to get (customized) **final response** instead of browsing websites
- ➢ Most of the users have no idea about the used applications based on generative models and cannot distinguish them with search engine **(Truthfulness)**.
- ➢ Interactive information accessing provides **more context and user information.**
- ➢ Etc.

Conversational Search in the LLM Era

User behaviour for information-seeking shift in the LLM Era:

- ➢ Interaction with LLM application via natural language (Context Modeling)
- ➢ Refine their information needs (Query rewriting and Mix-initiative)

Ne

- ➢
- ➢ els

and distinguish them with search engine **(Truthfulness)**.

- ➢ Interactive information accessing provides **more context and user information.**
- ➢ Etc.

Question: How should the goals and paradigms of conversational search shift correspondingly in the LLM era?

Generating Response in Conversational Search

Conversational retrieval-augmented generation (RAG)

- Single turn RAG v.s. Conversational (Multi-turn) RAG
- Leveraging historical information for conversational RAG
- Integrating search model with LLMs in conversations

Generating Response in Conversational Search

Single turn RAG [1]

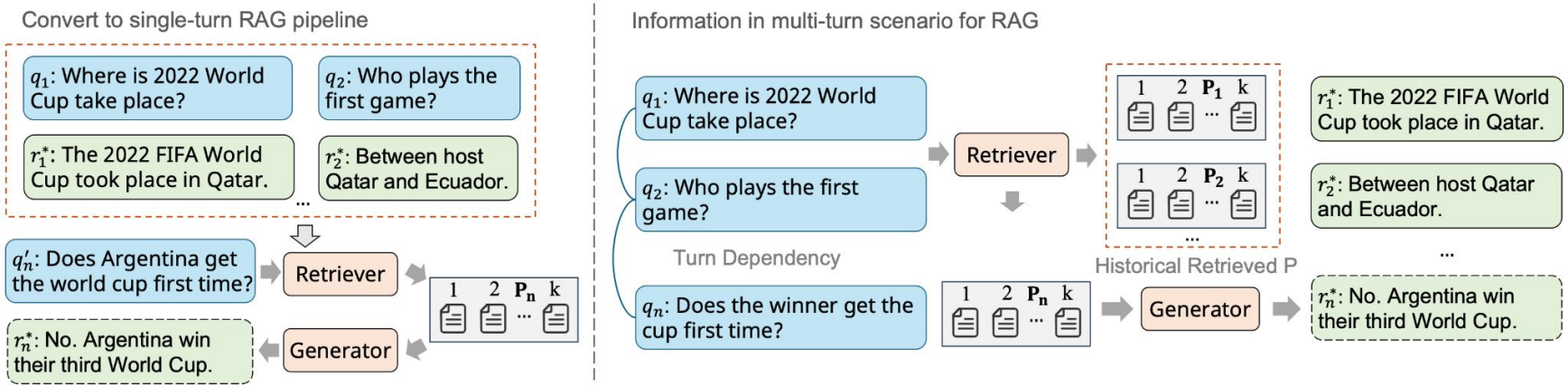
- **Trend**: LLMs can direct reply users' question with their parametric knowledge.
- **Challenge**: LLMs would still generate plausible but incorrect responses for some given queries when their internal knowledge is out-of-date.
- **Goal**: Incorporate the retrieved up-to-date information for generation.
- **Paradigm**: Generate response for a query on top of retrieved information.

[1] Retrieval-augmented generation for knowledge-intensive nlp tasks. Lewis et al. NIPS 2020.

Generating Response in Conversational Search

Single turn RAG [1] v.s. Conversational (Multi-turn) RAG [2]

- **Feature**: More available information beyond query-response pairs.
- **Existing paradigm**: Convert multi-turn into single-turn.
- **Question**: Could we improve the system performance by multi-turn information?



[1] Retrieval-augmented generation for knowledge-intensive nlp tasks. Lewis et al. NIPS 2020.
[2] CORAL: Benchmarking Multi-turn Conversational Retrieval-Augmentation Generation. Cheng et al. NAACL 2024.

Generating Response in Conversational Search

Conversational retrieval-augmented generation (RAG)

- Leveraging historical information for conversational RAG
  - **Idea**: The historical information (not limited to query-response pairs) could enhance the effectiveness [1,3], efficiency [2], and truthfulness [1] of the current turn RAG.

[1] CORAL: Benchmarking Multi-turn Conversational Retrieval-Augmentation Generation. Cheng et al. NAACL 2024.
[2] Learning When to Retrieve, What to Rewrite, and How to Respond in Conversational QA. Roy et al. EMNLP 2024.
[3] Conv-CoA: Improving Open-domain Question Answering in Large Language Models via Conversational Chain-of-Action. Pan 2024.

Generating Response in Conversational Search

Leveraging Historical Information for conversational RAG

- **Effectiveness** and **truthfulness** in conversational RAG
  - [1] proposes a conversational RAG benchmark with passage retrieval, response generation, and citation labeling grounding on passage IDs.

Question3: Who received the Governors Award and why were they chosen?
Response3: The Governors Award was presented to Tyler Perry and the Perry Foundation "in recognition of their unparalleled contributions to shaping the television medium" and for their "inclusion, engagement, employment and other philanthropic initiatives".[65233][65234] The award was moved to the main telecast from its usual presentation at the Creative Arts Emmys.[65235]
Golden Retrieval Passage IDs: 65233, 65234, 65235
Golden Rewrite: Who received the Governors Award at the 72nd Primetime Emmy Awards and for what reasons?
URL: https://en.wikipedia.org/wiki/72nd_Primetime_Emmy_Awards

...

Question10: Who performed during the in Memoriam Segment, and which notable figures were mentioned?
Response10: The annual in Memoriam Segment featured H.E.R. performing "Nothing Compares 2 U" on piano and electric guitar .[65284][65285]...
Golden Retrieval Passage IDs: 65284, 65285, 65286, 65287
Golden Rewrite: Who performed during the in Memoriam Segment at the 72nd Primetime Emmy Awards, and which notable figures were mentioned?
URL: https://en.wikipedia.org/wiki/72nd_Primetime_Emmy_Awards

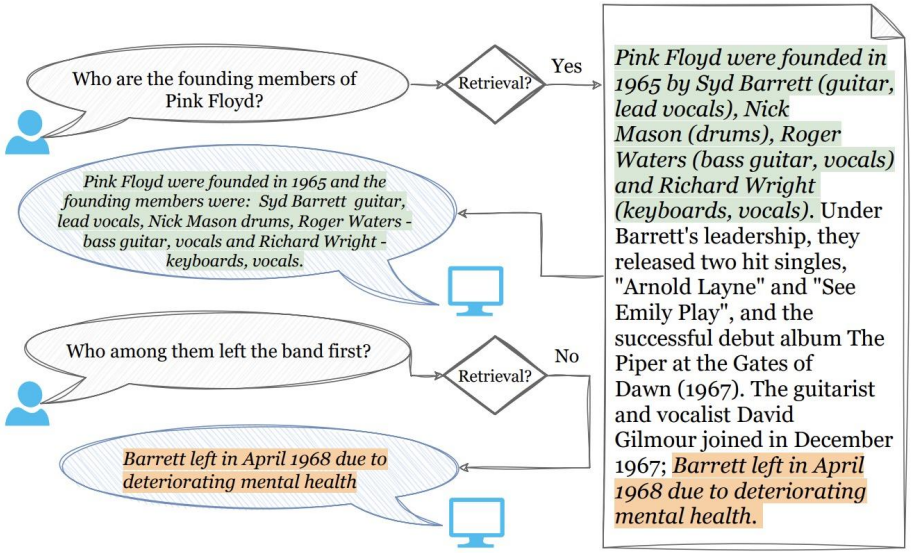
[1] CORAL: Benchmarking Multi-turn Conversational Retrieval-Augmentation Generation. Cheng et al. NAACL 2024.

Generating Response in Conversational Search

Conversational retrieval-augmented generation (RAG)

➢ Leveraging historical information for **efficient** conversational RAG

- **Idea:** Reducing the system latency by judging whether the required passages have already been retrieved in history before calling retriever for searching [1].
- **Challenge**: When to retrieve?

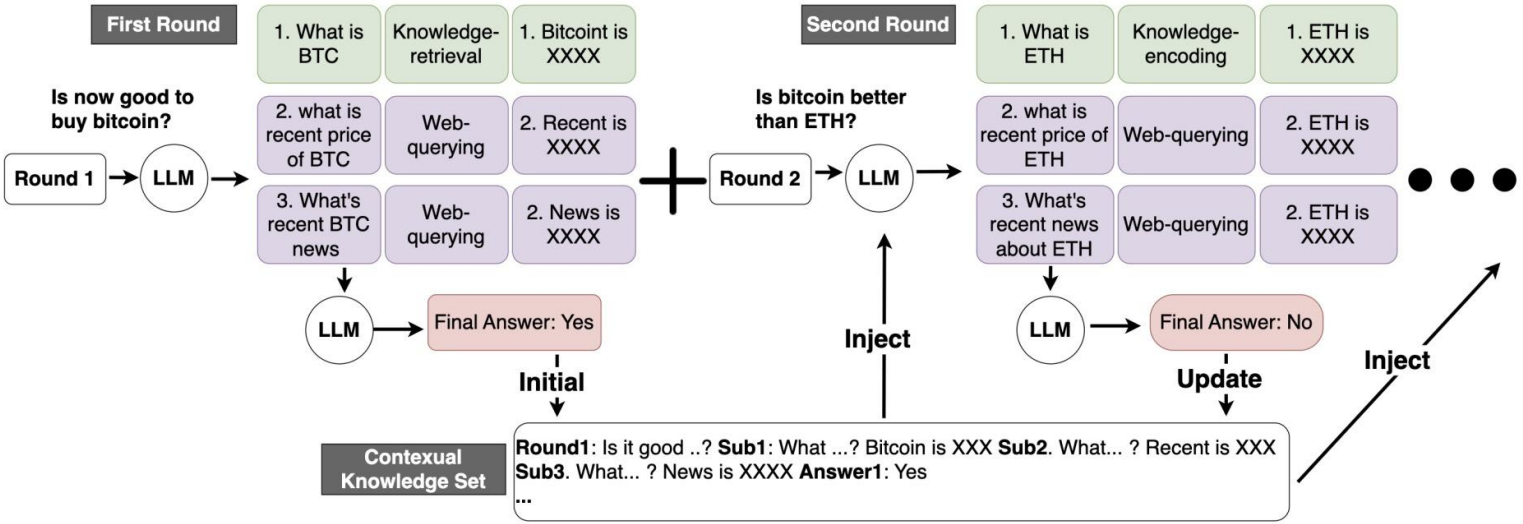


[1] Learning When to Retrieve, What to Rewrite, and How to Respond in Conversational QA. Roy et al. EMNLP 2024.

Generating Response in Conversational Search

Conversational retrieval-augmented generation (RAG)

- Leveraging historical information for conversational RAG
  - **Idea**: [1] maintain a contextual set from history to answer later turns.

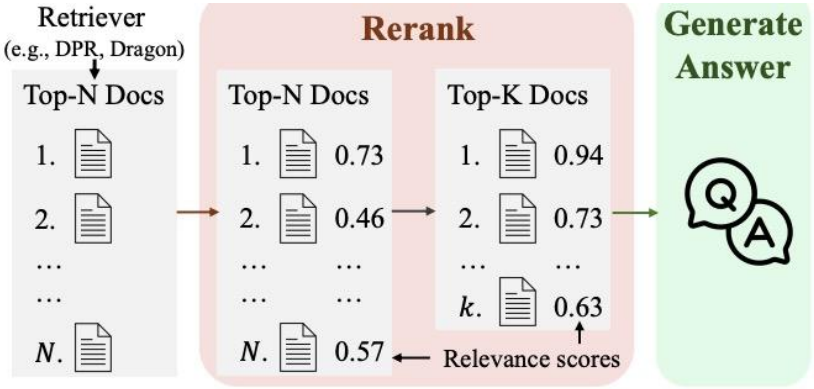


[1] Conv-CoA: Improving Open-domain Question Answering in Large Language Models via Conversational Chain-of-Action. Pan 2024.

Generating Response in Conversational Search

Integrating search model with LLMs in conversations

- A unified model can reduce model maintenance cost [1] and risk of discrepancy (e.g., the utilization of the search results for generation [2]).



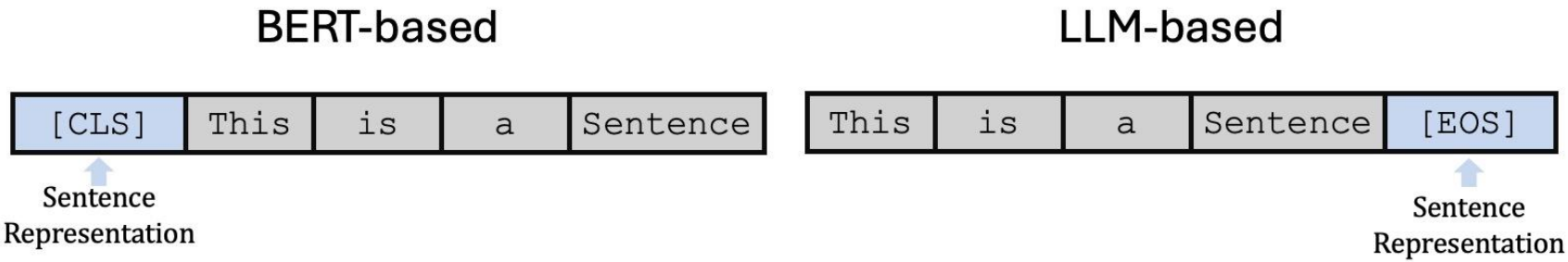
- The intrinsic knowledge of LLMs could be used for ranking and response generation via a unified model [1].

[1] RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs. Yu et al. NIPS 2024.
[2] Evaluating Retrieval Quality in Retrieval-Augmented Generation.Salemi et al. SIGIR 2024.

Generating Response in Conversational Search

Integrating search model with LLM by developing unified model

➢ SLM (e.g., BERT) as retriever [1] v.s. LLM (e.g., LLaMA) as retriever [2].



➢ The success of LLM-based retriever [2] shows the feasibility for adapting it to conversational scenarios [3].

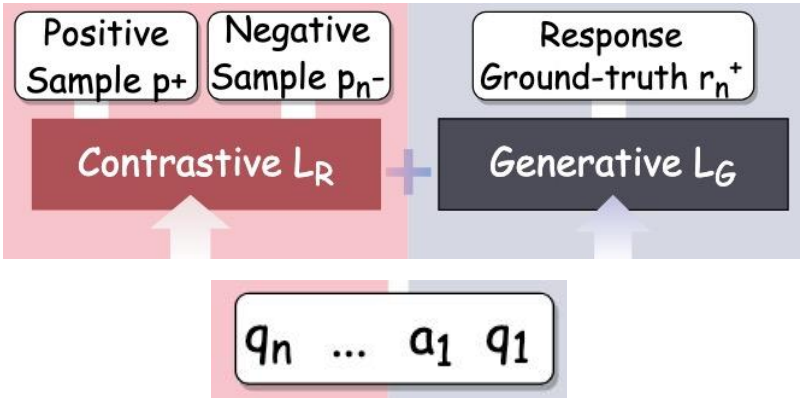
[1] Dense Passage Retrieval for Open-Domain Question Answering. Karpukhin et al. EMNLP 2020.
[2] Fine-tuning llama for multi-stage text retrieval. Ma et al. SIGIR 2024
[3] ChatRetriever: Adapting Large Language Models for Generalized and Robust Conversational Dense Retrieval. Mao et al. EMNLP 2024.

Generating Response in Conversational Search

Search model integrated with LLM by a unified model in conversations

- **Three crucial abilities**: conversational understanding, retrieval, generation.
- [1,2,3] unify a retriever/re-ranker with a generator by accommodating the training objective to keep the retrieval/ranking and response generation ability.

System	Conv.	Ret.	Gen.
RepLLaMA (Ma et al., 2024)	✗	✓	✗
E5 (Wang et al., 2024)	✗	✓	✗
ChatRetriever (Mao et al., 2024a)	✓	✓	✗
RankRAG (Yu et al., 2024)	✓	✗	✓
ChatQA (Liu et al., 2024)	✓	✗	✓
GRIT (Muennighoff et al., 2024)	✗	✓	✓
UniConv (Mo et al., 2025b)	✓	✓	✓

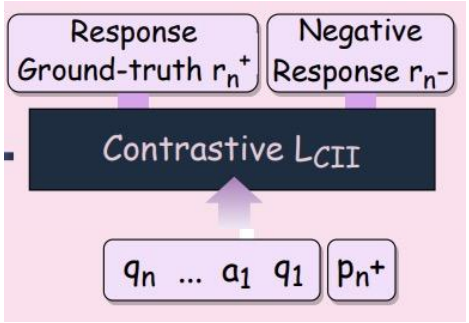


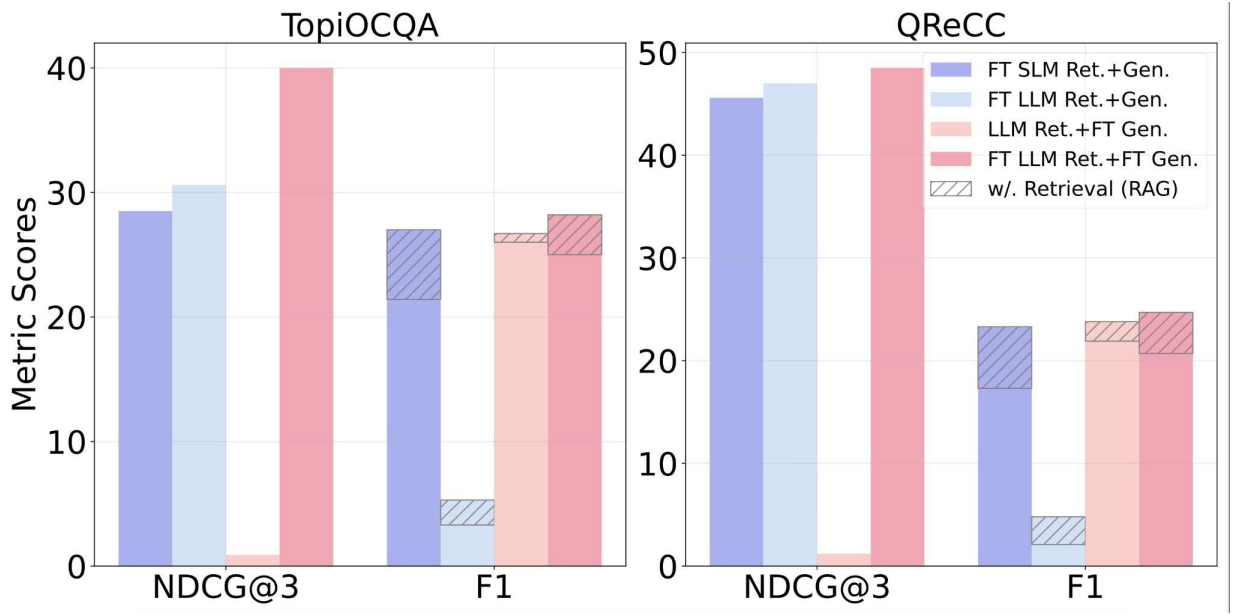
[1] RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs. Yu et al. NIPS 2024.
[2] OneGen: Efficient One-Pass Unified Generation and Retrieval for LLMs. Zhang et al. EMNLP 2024.
[3] UniConv: Unifying Retrieval and Response Generation for Large Language Models in Conversations. Mo et al. ACL 2025.

Generating Response in Conversational Search

Search model integrated with LLM by a unified model in conversations

- **Key points**: Maintain the generation ability and extend with the capability of retrieval and search intent understanding in conversational sessions during training [1].





[1] UniConv: Unifying Retrieval and Response Generation for Large Language Models in Conversations. Mo et al. ACL 2025.

Generating Response in Conversational Search

Summary:

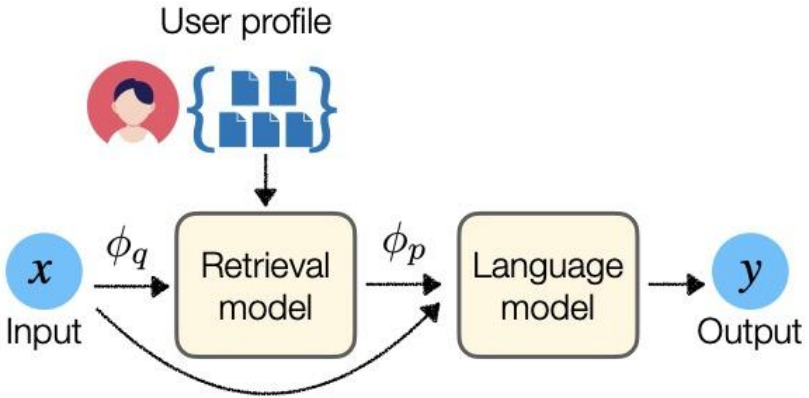
- **Conclusion**: The useful information from historical turns can improve system performance from different perspectives.
- **Key Challenge**: Identify the useful information from super noisy history.
- **Open questions**:
  - How to better leverage historical information for conversational RAG?
  - How to make the system more efficient with large models?
  - How to evaluate the generated response (in conversational scenario)?

Q & A

Personalized Conversational Search

Personalized Conversational Search

- **Goal**: Satisfy users' complex information needs based on users' profiles and preference through multi-turn interactions.
- **Assumption**: The same query turn from different users may correspond to different search intents, thus yielding different results.
- **User information**: Profile, historical preference, click/interactive behaviour.
- **General Paradigm**:



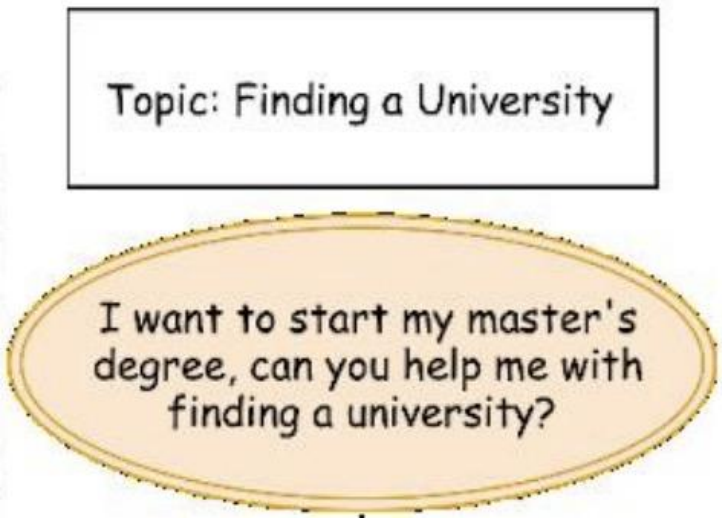
Personalized Conversational Search - TREC iKAT

Incorporating explicit user profile into query rewriting

- User profile in natural language format as Personal Text Knowledge Base [1,2].
- **Sub-task**: (1) PTKB selection, (2) Personalized retrieval in conversations.

PTKB 1: [1. I have bachelor degree of computer science from Tilburg university
2. I live in the Netherlands
3. I worked as a web developer for 2 years
.....]

PTKB 2: [1. I cannot withestand the temperature below -12 C
2. I'm from the Netherlands
3. I'm moving to Canada to study master
4. I have bachelor degree of computer science
.....]



[1] TREC iKAT 2023: The Interactive Knowledge Assistance Track Overview. Aliannejadi et al. TREC 2023.
[2] Conversational Gold: Evaluating Personalized Conversational Search System using Gold Nuggets. Abbasiantaeb et al. SIGIR 2025.

Personalized Conversational Search

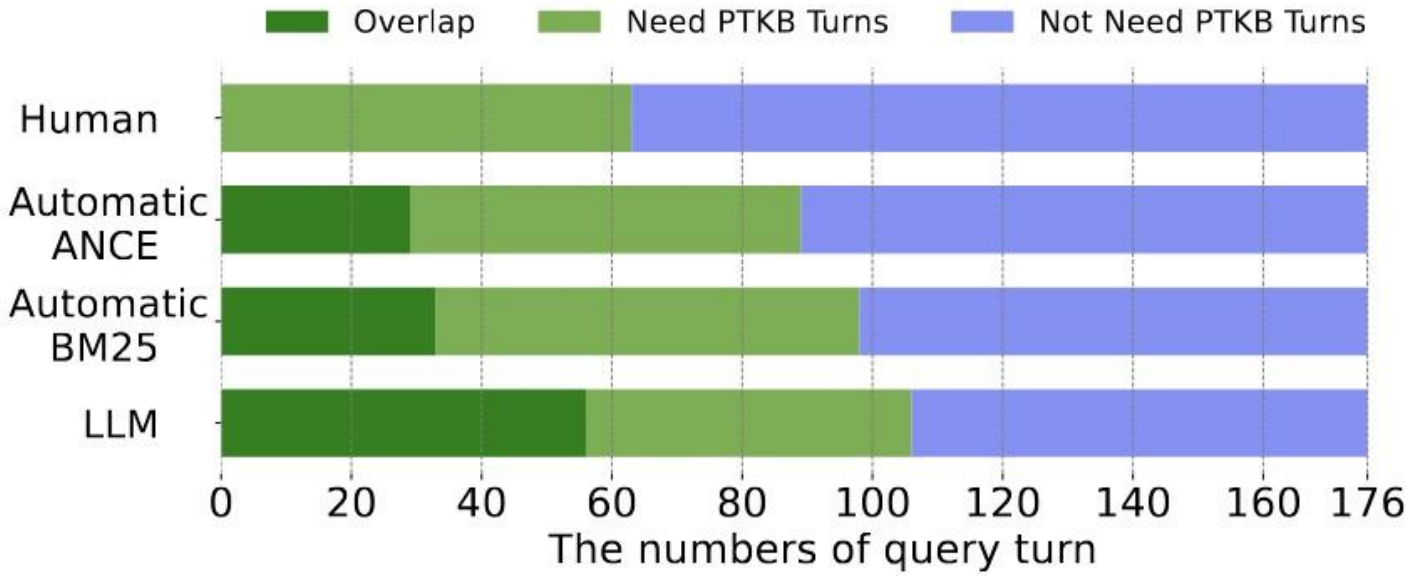
Incorporating explicit user profile into query rewriting

- **Idea**: Determine the relevant pieces from user profile for each query turn and incorporate the selected information into query rewriting as user modeling.
- **Key challenge**: Not all turns require personalization (using user profile).
  - Do I need a visa to travel to Egypt? (Require user information)
  - What are the prices of Egyptian E-visa and on-arrival visa. (Not require)

Personalized Conversational Search

Incorporating explicit user profile into query rewriting

- [1] analyze the potential discrepancies between human labeled relevant pieces and the machine judged ones, when personalization is required.



[1] How to Leverage Personal Textual Knowledge for Personalized Conversational Information Retrieval. Mo et al. CIKM 2024.

Personalized Conversational Search

Incorporating explicit user profile into query rewriting

- **Observation [1]**: If the personalization requirement is not determined well, using all historical turns or the selection judged by LLMs will both hurt the performance compared to without personalized query rewriting.

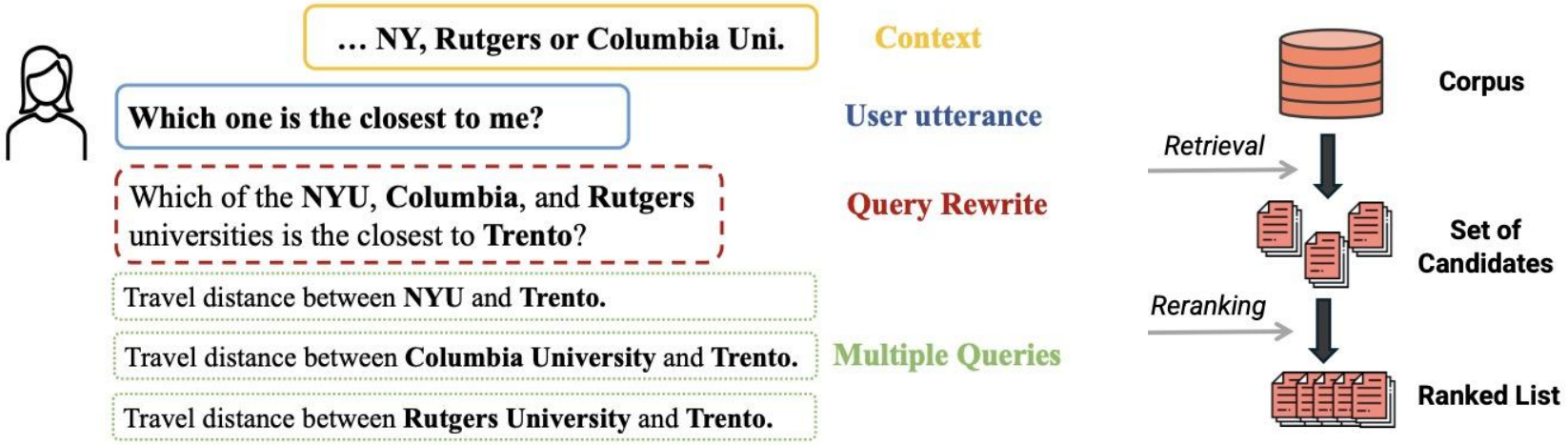
Model	Method	MRR	N@3	N@5	MAP
Evaluate on the whole test set (176 turns)					
BM25	None	**44.35**†	**21.22**†	**20.68**†	**8.91**
	Use all	40.36	19.19	18.84	8.28
	Human	41.65	19.66	19.46	8.82
	Automatic	40.29	19.12	18.87	8.58
	LLM-STR	41.53	18.96	18.09	8.37
	LLM-SAR	36.04	17.48	16.87	8.02
ANCE	None	32.47	14.25	13.73	5.68
	Use all	33.64	15.30	15.09	6.13
	Human	33.63	15.98	15.69	6.16
	Automatic	31.08	14.36	14.01	5.89
	LLM-STR	32.37	15.05	14.02	5.72
	LLM-SAR	31.76	14.78	15.12	5.47

[1] How to Leverage Personal Textual Knowledge for Personalized Conversational Information Retrieval. Mo et al. CIKM 2024.

Personalized Conversational Search

Incorporating explicit user profile into query rewriting

➢ **Idea**: Generating multiple queries with and without user information to cover different aspects and aggregate them to improve retrieval [1].



[1] Generating Multi-Aspect Queries for Conversational Search. Abbasiantaeb et al. arXiv 2024.

Personalized Conversational Search

Incorporating explicit user profile into query rewriting

- Multiple queries retrieval with personalization outperform single query retrieval.
- The LLM might not address personalized query well (answer as expansion hurt).
- How to aggregate the personalized information is important (re-rank hurt).

Method	MAP	R@10	R@100	MRR
GPT4oQR	45.4	66.9	80.9	45.4
T5QR	33.5	50.2	62.5	33.5
ConvGQR	31.1	49.0	63.2	31.1
GPT4o-AQ	42.9	63.1	79.8	42.9
LLM4CS	36.8	57.1	75.9	36.8
$MQ4CS_{ans}$	<u>46.7</u>	<u>70.6</u>	<u>87.0</u>	<u>46.7</u>
$MQ4CS_{ans}$+rerank	43.6	65.5	83.8	43.6
MQ4CS	**47.5**	**72.6**	**87.8**	**47.5**

[1] Generating Multi-Aspect Queries for Conversational Search. Abbasiantaeb et al. arXiv 2024.

Personalized Conversational Search

Leveraging implicit user preference from conversation history

- **Motivation**:
  - Existing studies for single-turn personalized benchmark treats each user utterance as independent [1]
  - The multi-turn conversation focus on modeling interaction structure or dialogue coherence while remaining largely user-agnostic [2].
  - No connection between personalization and conversation.

[1] Lamp: When large language models meet personalization. Salemi et al. ACL 2024.
[2] A Personalized Conversational Benchmark: Towards Simulating Personalized Conversations. Li et al. arXiv 2025.

Personalized Conversational Search

Leveraging implicit user preference from conversation history

- **Idea [1]**: (1) Simulate the conversational context toward the current turn; (2) Construct personalized conversational context according to all historical messages of a specific user as long-term personalized signals; (3) Combine both as condition for personalized generation.

- **Pros**: Standardized conversational personalized generation and benchmarking.
- **Cons**: The condition on historical context navigation is uncontrollable.

[1] A Personalized Conversational Benchmark: Towards Simulating Personalized Conversations. Li et al. arXiv 2025.

Generating Response in Conversational Search

Summary:

- **Conclusion**: Personalization in LLM era with multi-turn interaction is important and require new paradigm to achieve.
- **Key Challenge**: (1) Identify the personalization requirement and injection method and (2) Using user profile in suitable ways.
- **Open questions**:
  - How to modeling and inject personalized signals for various scenarios?
  - How to formulate/evaluate personalization task with LLMs? (user-centric)

Q & A

Discussions

Time for 

Part 2

Emerging Topics in the LLM and Agentic Era

Overview

Part II: Emerging Topics in the Agents Era [90 min]

- Agentic search [30 min]
- Conversational agentic search [15 min]
- Proactive conversational agents [15 min]
- Domain-specific applications [10 min]
- Conclusions and future directions [5 min]
- Discussion [15 min]

Agentic Search

Agentic Search

- What is an “agent”?
  - An agent is an autonomous entity that makes decisions and takes actions on users’ behalf [1,2]
  - The idea of agents traces back to the 1950s with the emergence of symbolic AI [1]

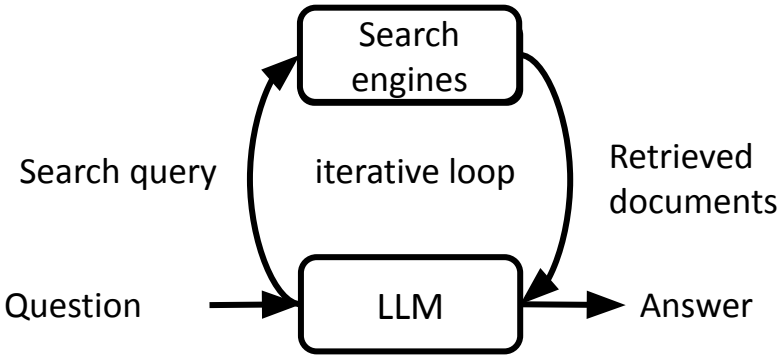
- Typical capabilities of agents [3]
  - Planning
  - Memory - Context management in conversations
  - Tool use - Search engines are a key tool
  - Reflection and refinement
  - Multi-agent collaboration

[1] Shah et al. Agents Are Not Enough. arXiv 2024.
[2] Meng et al. Optimizing Agentic Workflows for Information Access. University of Amsterdam 2025.
[3] White et al. Information Access in the Era of Generative AI. Springer 2025.

How to enable LLMs act as agents that autonomously use search engines to meet users' information needs?

Agentic Search

- Most work follows the **Reason-Act-Reflection (ReAct)** paradigm [1,2]
  - multi rounds of reasoning and search invocations



[1] Jin et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. COLM 2025.
[2] Song et al. R1-Searcher: Incentivizing the Search Capability in LLMs via Reinforcement Learning. arXiv 2025.

Agentic Search

- Search-R1 [1]: Dynamic search tool invocation
  - Search is interleaved with reasoning
  - Train the LLM to conduct reasoning and search calling by RL
    - Use the final answer as a reward signal

Answer the given question. You must conduct reasoning inside <think> and </think> first every time you get new information. After reasoning, if you find you lack some knowledge, you can call a search engine by <search> query </search>, and it will return the top searched results between <information> and </information>. You can search as many times as you want. If you find no further external knowledge needed, you can directly provide the answer inside <answer> and </answer> without detailed illustrations. For example, <answer> xxx </answer>. Question: question.

---

[1] Jin et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. COLM 2025.

Agentic Search

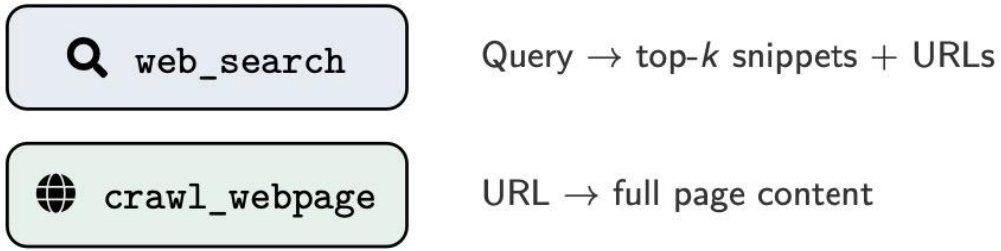
- Search-R1 example [1]

Question: What type of profession does Chris Jericho and Gary Barlow have in common?
Ground Truth: musician
SEARCH-R1:
<think> I need to find the type of profession that Chris Jericho and Gary Barlow have in common. I'll search for it. </think>
<search> Chris Jericho and Gary Barlow profession </search>
<information> Doc 2(Title: "Chris Jericho") ""Do You Wanna Start a War"" (2014) and ""Judas"" (2017) are comprised entirely or predominantly of original compositions. He has also appeared on numerous TV shows over the years, having competed in the 2011 season of ""Dancing With the Stars"", lasting until the sixth week. Prior to that, Jericho hosted the ABC game show ""Downfall"" and the Revolver Golden Gods Awards. Christopher Keith Irvine was born on November 9, 1970, in Manhasset, New York, but grew up in Winnipeg. His father, ice hockey player Ted Irvine, had been playing for the New York Rangers at the time of his birth. ... </information>
<think> I found out that Chris Jericho is a professional wrestler, musician, author, and actor. Now I need to find if Gary Barlow has any of these professions in common. </think>
<search> Gary Barlow profession </search>
<information> Doc 1(Title: "Gary Barlow") hits, twelve number one singles and seven number one albums with Take That. Barlow has received six Ivor Novello Awards from the British Academy of Songwriters, Composers and Authors, including the award for Outstanding Services to British Music. He has sold over 50 million records worldwide. Barlow was born on 20 January 1971 in Frodsham, Cheshire, the second son of Colin (died 15 October 2009) and Marjorie Barlow. In Barlow's autobiography, he relates that his love of music began at an early age, and calls himself ""one of those kids that's forever dancing in front of the TV looking at ... </information>
<think> I found out that Gary Barlow is a singer, songwriter, musician, and record producer. Now I need to find if there is a common profession between Chris Jericho and Gary Barlow. </think>
<search> common profession of Chris Jericho and Gary Barlow </search>
<information> Doc 1(Title: "Gary Barlow") Gary Barlow Gary Barlow (born 20 January 1971) is an English singer, songwriter, musician and record producer. He is best known as the lead singer of the British pop group Take That. Barlow also served as head judge of ""The X Factor UK"" in 2011, 2012 and 2013 and ""Let It Shine"" in 2017. Barlow is one of Britain's most successful songwriters, having written fourteen number one singles and twenty-four top 10 hits. As a solo artist he has had three number one singles, six top 10 singles and two number one albums, and has additionally had seventeen top 5 ... </information>
<think> I found out that Gary Barlow is a singer, songwriter, musician, and record producer. Now I need to find if Chris Jericho has any of these professions in common. </think>
<search> Chris Jericho profession </search>
<information> Doc 1(Title: "Chris Jericho") Chris Jericho Christopher Keith Irvine (born November 9, 1970), better known by the ring name Chris Jericho, is an American-born professional wrestler, musician, author and actor. He currently wrestles for New Japan Pro-Wrestling (NJPW) where he is the current IWGP Intercontinental Champion in his first reign. Jericho also makes sporadic appearances for WWE on the SmackDown brand, with whom he has worked since 1999. Within professional wrestling, he is known for his over-the-top, rock star persona. Jericho is regarded as one of the greatest professional wrestlers of all time. During the 1990s, Jericho performed for the American organizations Extreme Championship ... </information>
<think> I found out that Chris Jericho is a professional wrestler, musician, author, and actor. Now I can compare it to Gary Barlow's professions. I found out that Gary Barlow is a singer, songwriter, musician, and record producer. The common profession between Chris Jericho and Gary Barlow is musician. </think>
<answer> musician </answer>

[1] Jin et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. COLM 2025.

Agentic Search

- Search-R1 [1] showed RL can train search agents. But in real web-based deep research, **tool go beyond a search engine**:

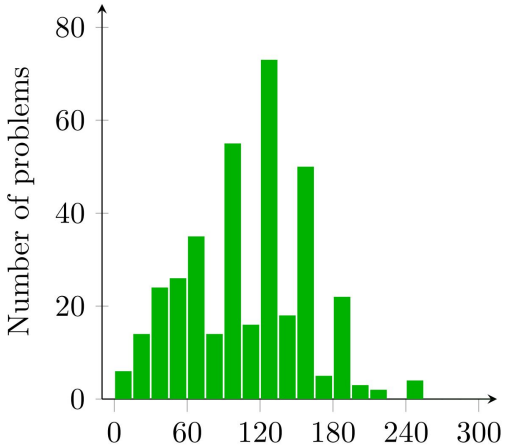


- From Search-R1 to SynPlanResearch-R1 [2]: Snippets alone are often insufficient — the agent must click into pages to get detailed evidence.
  - train agents to use multiple tools effectively

[1] Jin et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. COLM 2025.
[2] Zeng et al. SynPlanResearch-R1: Encouraging Tool Exploration for Deep Research with Synthetic Plans. arXiv 2026.

Agentic Search

- Advanced scenario in agentic search:
  - Deep research aims to answer hard, multi-hop, reasoning-intensive questions that require extensive open-web exploration
    - It is extremely challenging [1]:
      - Humans gave up on 70% of questions after two hours of searching
      - LLMs achieve near-zero accuracy



Human time (with search) to answer the remaining 30% (min)

Model	Accuracy (%)
GPT-4o	0.6
GPT-4o w/ browsing	1.9
GPT-4.5	0.9
OpenAI o1	9.9

[1] Wei et al. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. 2025.

Agentic Search

- An example from [1]:

Question
Identify the title of a research publication
published before June 2023,
that mentions Cultural traditions, scientific processes,
and culinary innovations.

It is co-authored by three individuals:
one of them was an assistant professor in West Bengal
and another one holds a Ph.D.

Answer
The Fundamentals of Bread Making: The Science of Bread

[1] Wei et al. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. 2025.

Agentic Search

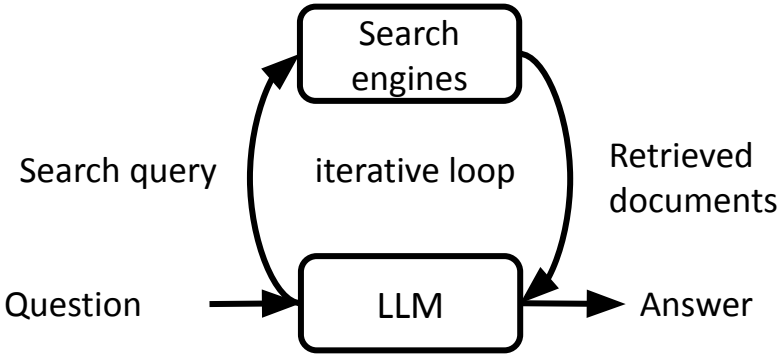
- Search-R1-based methods perform poorly on deep research benchmarks [1].
- LLMs that use web search APIs during pre-training perform more effectively and make more search calls

LLM	Retriever	Accuracy	Recall	Search Calls
gpt-5	Qwen3-Embed-8B	70.12%	78.98%	21.74
gpt-oss-120B-high	Qwen3-Embed-8B	42.89%	52.63%	18.35
SearchR1-32B	Qwen3-Embed-8B	10.36%	10.17%	1.69

[1] Chen et al. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. 2025.

Agentic Search

- [1] studies LLMs pretrained with web search APIs, with access to established text ranking methods in deep research
  - Agent-issued queries follow a web-search-style syntax, favouring lexical/multi-vector retrievers
  - Passage-level units are more effective and token-efficient than document-level units
  - Re-ranking improves performance and reduces search iterations
  - Translating agent queries into natural questions enhances neural ranking

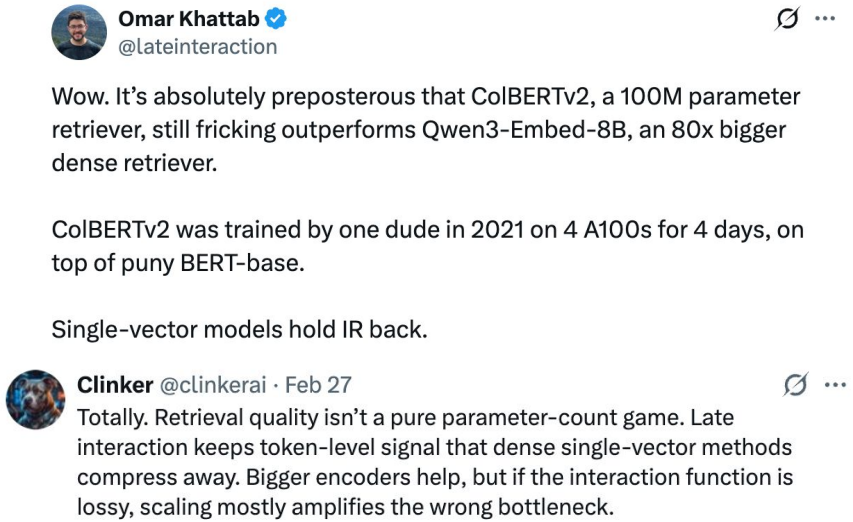


[1] Meng et al. Revisiting Text Ranking in Deep Research. arXiv 2026.

Agentic Search

- Finding from [1]
  - Agent-issued queries typically follow a web-search style with keywords, phrases, and quotation marks for exact matching
  - Lexical retriever BM25 (1994) [1] achieves the best performance in most cases
  - Multi-vector retriever ColBERT (110M) outperforms much larger single-vector retrievers, such as Qwen3-Embed (8B)

Agent	Search query
gpt-oss-20b	“90+7” attendance 61700 “Man United” “4-1” “90+4” “assist” “4–1” “Premier League” “2020”
GLM-4.7-Flash	“90’+4” football match attendance “61,888” football match Stockholm Vienna Prague “goal in the 6th minute” football



Omar Khattab @lateinteraction

Wow. It’s absolutely preposterous that ColBERTv2, a 100M parameter retriever, still fricking outperforms Qwen3-Embed-8B, an 80x bigger dense retriever.

ColBERTv2 was trained by one dude in 2021 on 4 A100s for 4 days, on top of puny BERT-base.

Single-vector models hold IR back.

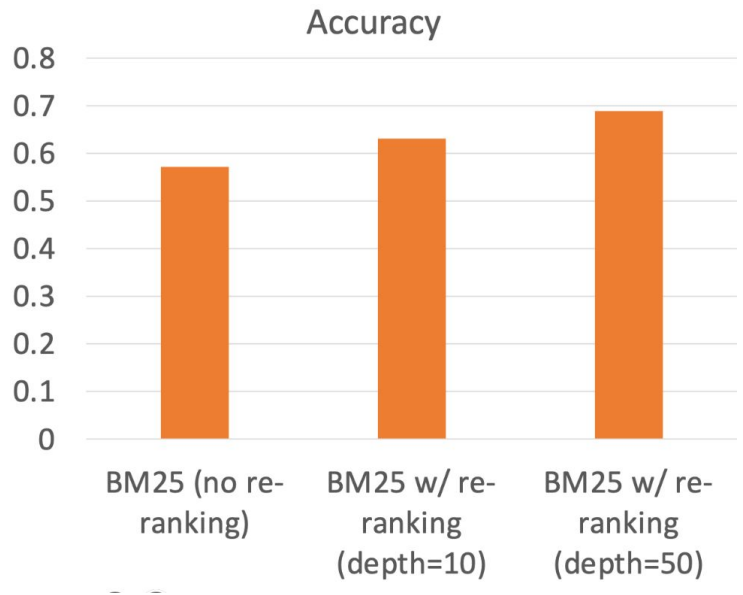
Clinker @clinkerai · Feb 27

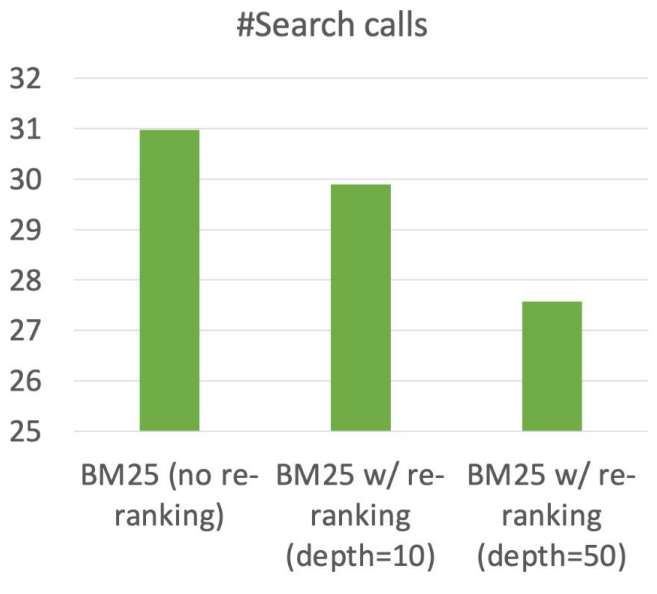
Totally. Retrieval quality isn’t a pure parameter-count game. Late interaction keeps token-level signal that dense single-vector methods compress away. Bigger encoders help, but if the interaction function is lossy, scaling mostly amplifies the wrong bottleneck.

[1] Meng et al. Revisiting Text Ranking in Deep Research. arXiv 2026.

Agentic Search

- Finding from [1]
  - Re-ranking consistently helps and reduces search iterations
  - deeper re-ranking depths amplify the gains



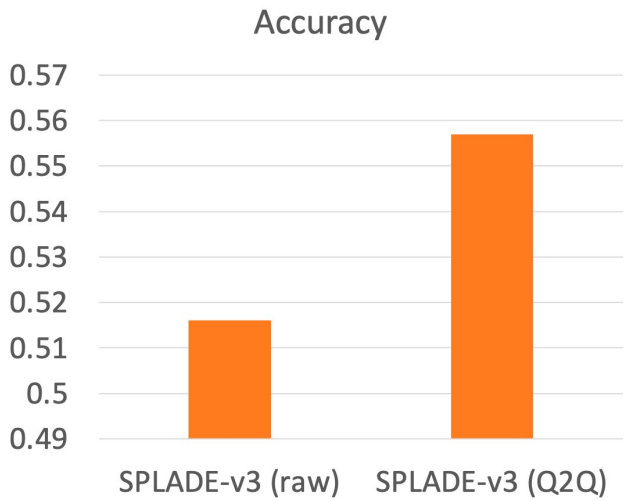


[1] Meng et al. Revisiting Text Ranking in Deep Research. arXiv 2026.

Agentic Search

- Finding from [1]
  - **Query mismatch for neural ranking**: most neural retrievers are trained on natural-language questions, not web-style queries
  - We propose a **Query-to-Question (Q2Q)** method that translates agent-issued queries into natural questions, significantly improving performance.

Method	Search query
Raw query	"61,880" football attendance
Q2Q	What football match had an attendance of 61,880?



[1] Meng et al. Revisiting Text Ranking in Deep Research. arXiv 2026.

Q & A

Agentic Conversational Search

Agentic Conversational Search

Background and Goal

- Previous advanced deep search agentic models are designed for single-turn information accessing tasks, which **might lack the ability to handle multi-turn conversational scenarios.**

- How to enable agentic search in conversational scenario, in terms of multi-turn interaction, context-dependent query understanding, etc?

Conversational Agent with Reasoning

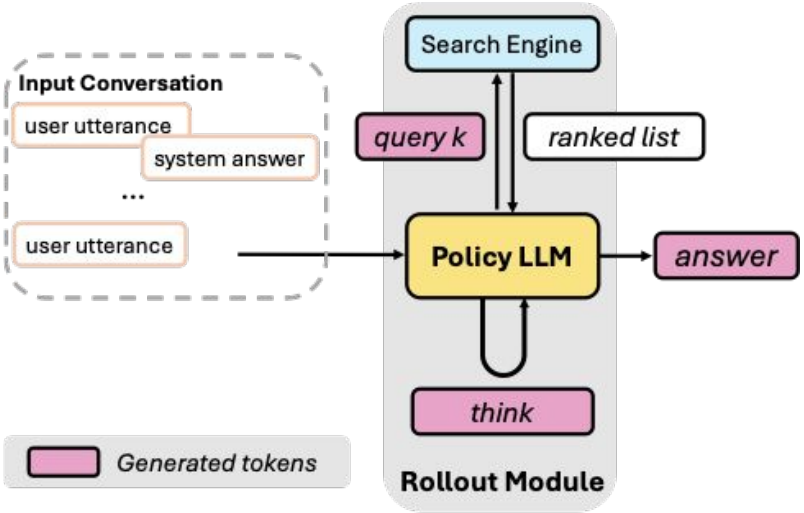
- Previous studies usually follow static “rewrite, retrieve, and generate” workflows with separated models.

- Recent studies [1] and [2] first propose to develop conversational agent with reasoning capacity for multi-turn interactions

[1] Lupart et al. ChatR1: Reinforcement Learning for Conversational Reasoning and Retrieval Augmented Question Answering. arXiv 2025.
[2] Mo et al. Agentic Conversational Search with Contextualized Reasoning via Reinforcement Learning. arXiv 2026.

Agentic Conversational Search

- [1,2] develop agentic conversational search systems that can do multi-rounds of reasoning and search via RL



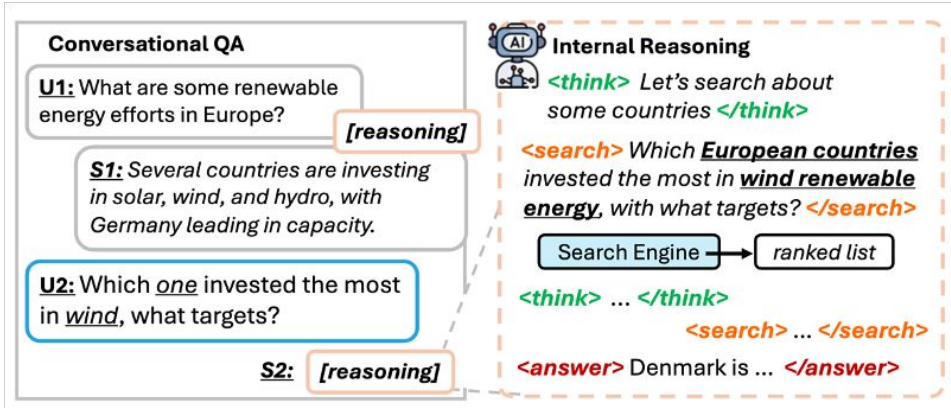
[1] Lupart et al. ChatR1: Reinforcement Learning for Conversational Reasoning and Retrieval Augmented Question Answering. arXiv 2025.
[2] Mo et al. Agentic Conversational Search with Contextualized Reasoning via Reinforcement Learning. arXiv 2026.

Agentic Conversational Search

- [1] find query formulation(<search>...</search>) is more challenging
  - Use the final answer, and **human query rewrite** as reward signals
    - Measure similarity between human- and model-written queries

$$R_{\text{intent}}(Q) = \max_{q^k \in Q} \text{F1}(q^k, q^{rw})$$

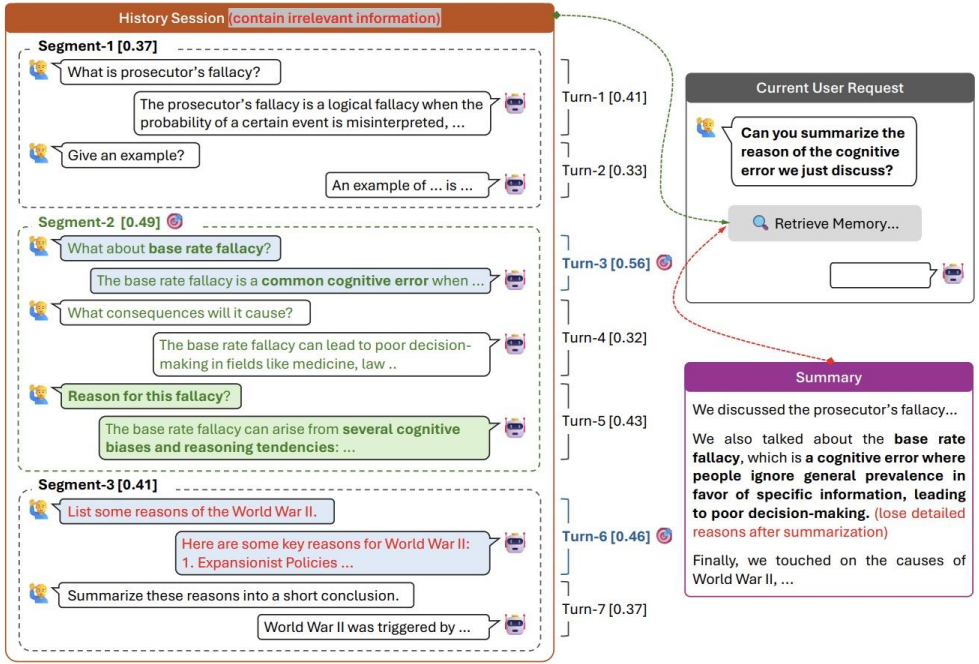
```
You are a helpful assistant tasked with
answering a user query. Your primary goal is
to generate a complete and informative answer.
If the query is ambiguous or refers to
earlier context (e.g., pronouns or ellipsis),
use the conversation history provided below
to resolve it.
- Always perform your reasoning inside <think>
...</think>.
- If external information is needed, use
<search>your query</search>.
- Retrieved documents will appear between
<information>...</information>.
- You may issue multiple search queries if
needed.
- Once you have enough information, provide a
complete answer within <answer>...</answer>.

Conversation context: {context_block}
User query: {last_user_utterance}
```



[1] Lupart et al. ChatR1: Reinforcement Learning for Conversational Reasoning and Retrieval Augmented Question Answering. arXiv 2025.

Agentic Conversational Search

- A recent study [1] systematically investigate the effects of memory granularity on RAG in conversational agents.
- **Observation**: Turn-level, session-level, and summarization-based methods each face challenges.
- **Idea**: Constructing memory bank at segment level by introducing a conversation segmentation model.



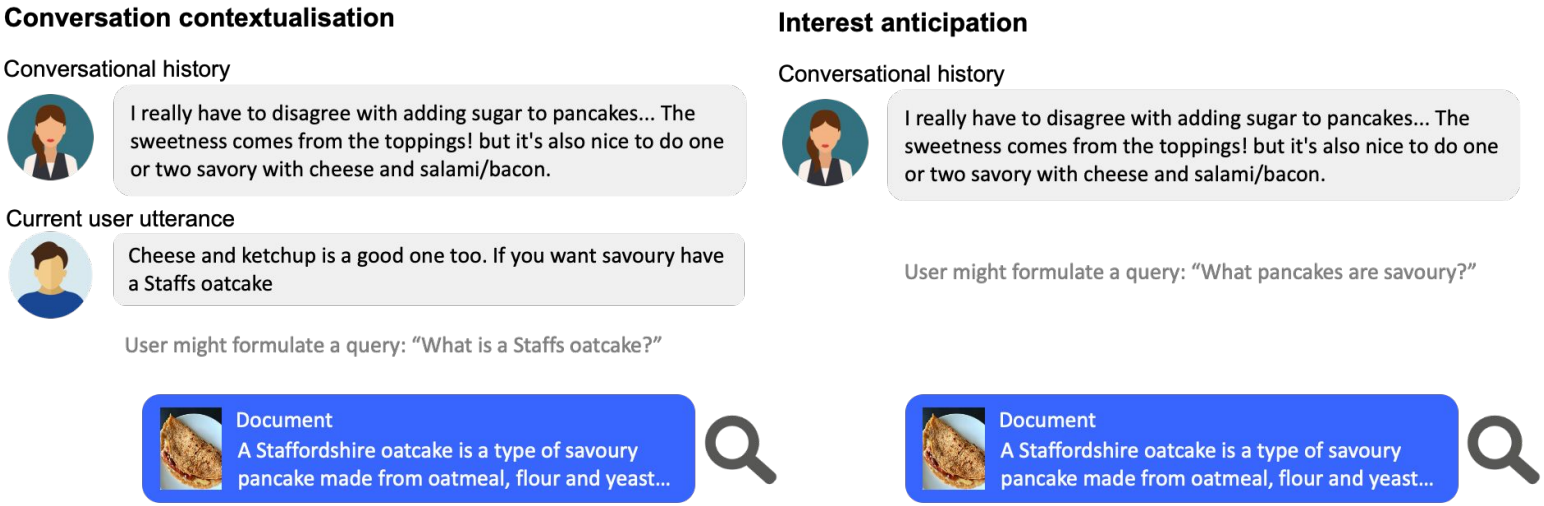
[1] Pan et al. On memory construction and retrieval for personalized conversational agents. ICLR 2025.

Q & A

Proactive Conversational Agents

Proactive Conversational Agents

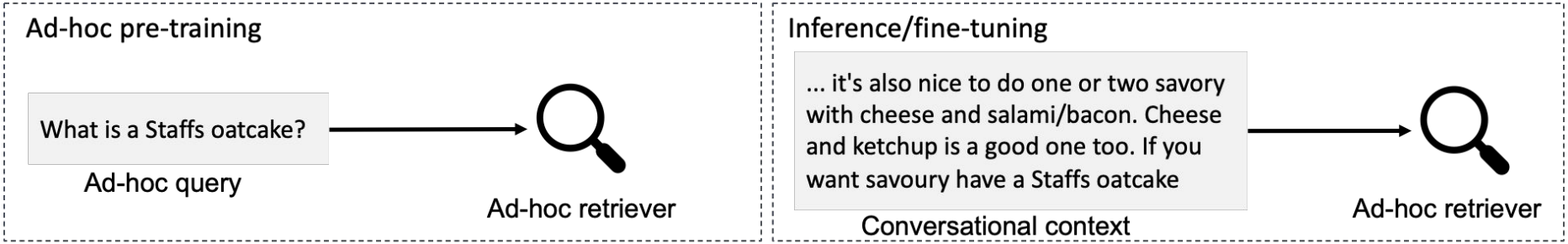
- Conversation contextualisation/interest anticipation
  - [1,2] release datasets targeting:
    - Conversation contextualisation
    - Interest anticipation



[1] Ros et al. Retrieving Webpages Using Online Discussions. ICTIR 2023.
[2] Samarinas et al. ProCIS: A Benchmark for Proactive Retrieval in Conversations. SIGIR 2024.

Proactive Conversational Agents

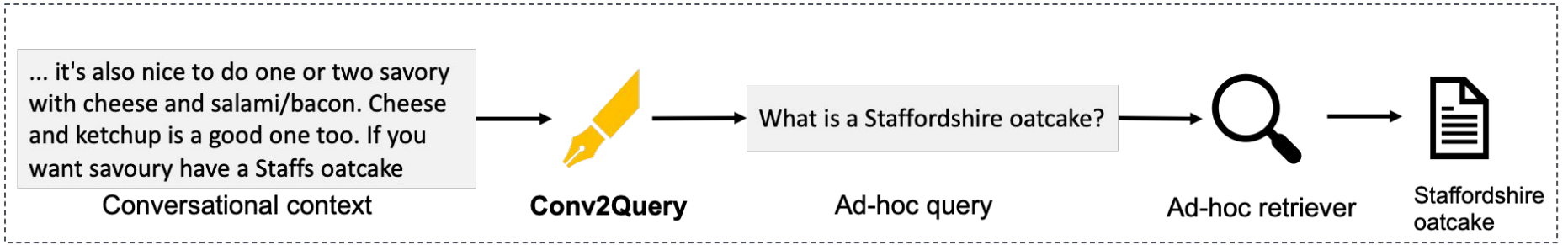
- Conversation contextualisation/interest anticipation
  - Feed raw conversational context to neural retrievers pre-trained on ad-hoc search data
    - Limitation: Input gap between ad-hoc pre-training and inference [1]
  - Further fine-tunes ad-hoc neural retrievers on conversational data
    - Limitation: Input gap between ad-hoc pre-training and fine-tuning [1]



[1] Meng et al. Bridging the Gap: From Ad-hoc to Proactive Search in Conversations. SIGIR 2025.

Proactive Conversational Agents

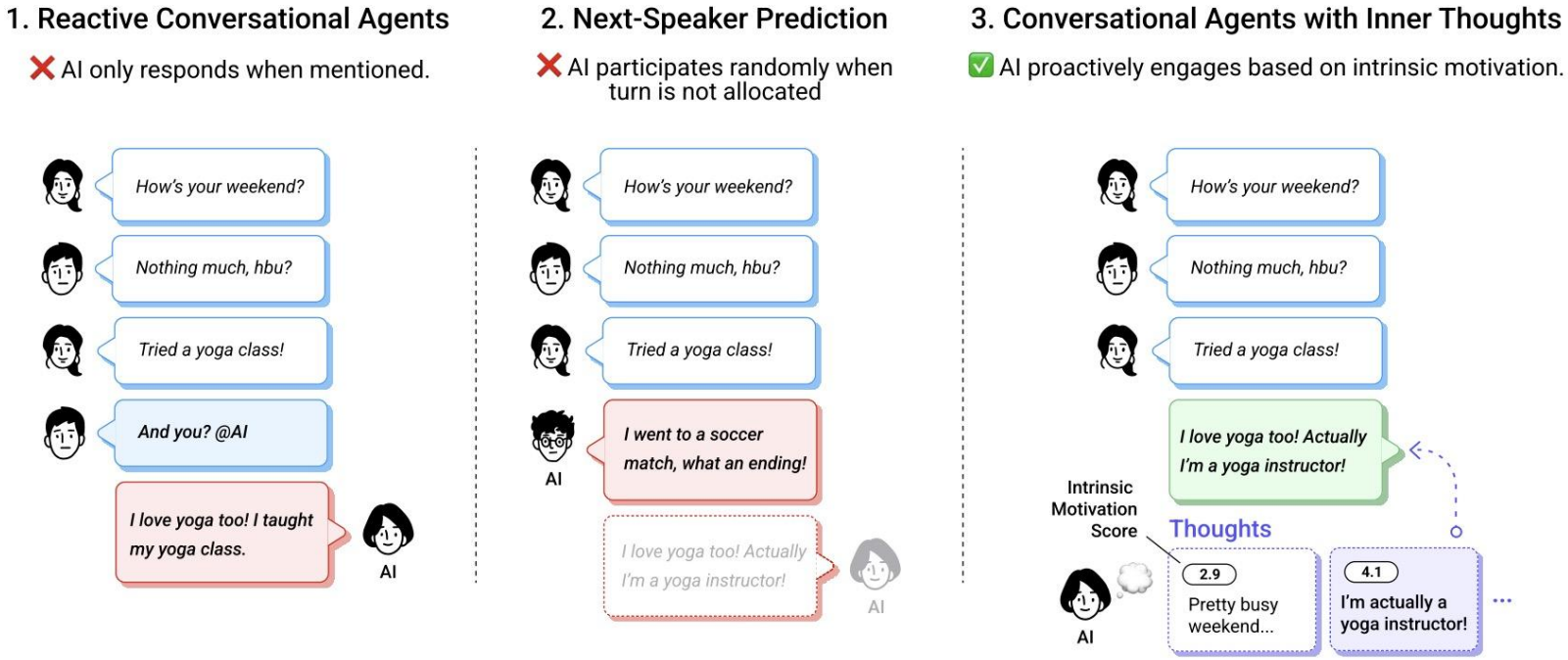
- Conversation contextualisation/interest anticipation
  - [1] proposes Conv2Query
    - Transforms conversational context into ad-hoc queries, which are used to
      - Query off-the-shelf ad-hoc retrievers
      - Further fine-tune ad-hoc retrievers



[1] Meng et al. Bridging the Gap: From Ad-hoc to Proactive Search in Conversations. SIGIR 2025.

Proactive Conversational Agents

- A proactive AI agent should autonomously participate in socially appropriate moments, providing relevant input without requiring explicit cues [1].



[1] Liu et al. Proactive Conversational Agents with Inner Thoughts. CHI 2025.

Proactive Conversational Agents

- The factors influence people's decisions to express or withhold their thoughts during conversations according to user studies [1].

1. Relevance 77 mentions

How much does the thought relate to the party's knowledge, interests, or previous thoughts?

e.g.
✅ *People tend to participate more actively in conversations when they can relate the topic to their own personal experiences.*
❌ *People tend NOT to participate in conversations on topics where they lack knowledge of the ongoing topic.*

2. Information Gap 33 mentions

Does the thought reflect or address an information gap in the conversation?

e.g.
✅ *People tend to participate in conversations when they seek clarification or additional information.*
❌ *People tend NOT to participate when they think the conversation is predictable and uninteresting, based on prior knowledge.*

3. Expected Impact 23 mentions

How significantly might the thought influence the future conversation, if expressed?

e.g.
✅ *People tend to participate when they expect the thought to introduce new topics and discussions and engage other participants' interests.*
❌ *People sometimes withhold thoughts if they feel it will be covered in future conversations.*

4. Urgency 14 mentions

Does the thought need to be expressed immediately?

e.g.
✅ *People tend to participate promptly when they perceive errors or misunderstandings.*
✅ *People tend to participate promptly if their thought, if expressed, will significantly impact the future conversation.*

5. Coherence 30 mentions

Does the thought seem in-place if it is expressed next in the conversation?

e.g.
✅ *People tend to participate when their thought directly follows from and logically builds on the previous speaker's comment or question.*
❌ *People tend NOT to participate when their thought is disconnected from the previous utterance or ignores the context.*

6. Originality 16 mentions

Does the thought provide new and original contribution?

e.g.
✅ *People tend to engage more when their contribution adds fresh, original insights or new information to the conversation.*
❌ *People tend NOT to participate when they notice their thought has already been articulated by others, avoiding repetition.*

7. Balance 33 mentions

Does everyone have a chance to participate in the conversation and not be left out?

e.g.
✅ *People sometimes back-channel to show their presence, interest, attention and a willingness to keep listening.*
❌ *People sometimes refrain from engaging when they notice others who have been less active in the conversation.*

8. Dynamics 30 mentions

Is someone else actively contributing to the conversation or is likely to speak?

e.g.
❌ *People tend to withhold their thoughts when others are in the middle of an intense discussion*
❌ *People tend NOT to participate and wait if they feel someone has not finished their sentence or typing.*

[1] Liu et al. Proactive Conversational Agents with Inner Thoughts. CHI 2025.

Q & A

Domain-Specific Application with Conversational Agents

Domain-Specific Applications

- The goal of agentic conversational search systems is to adapt to specialized knowledge and diverse user intents
- Across domains and settings: multilingual, legal, finance, medical,...

Accuracy matters

Healthcare, Legal, Technical Support

User Diversity

Novice vs. Expert, Low-literacy vs. Specialized

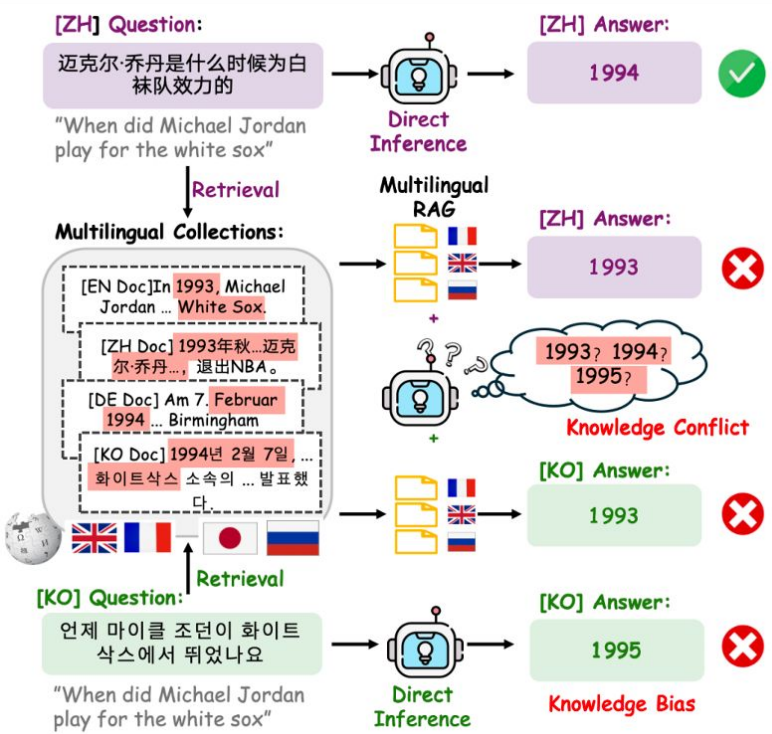
Value

From raw information to actionable decisions

Domain-Specific Applications

- Multilingual RAG [1,2]: LLM answers queries by retrieving evidence from multilingual document collections.

- Challenges:
  - **Knowledge bias**: The same question expressed in different languages may lead to different reasoning paths and answers.
  - **Knowledge conflict**: Conflicting evidence from different languages can mislead the reasoning process.

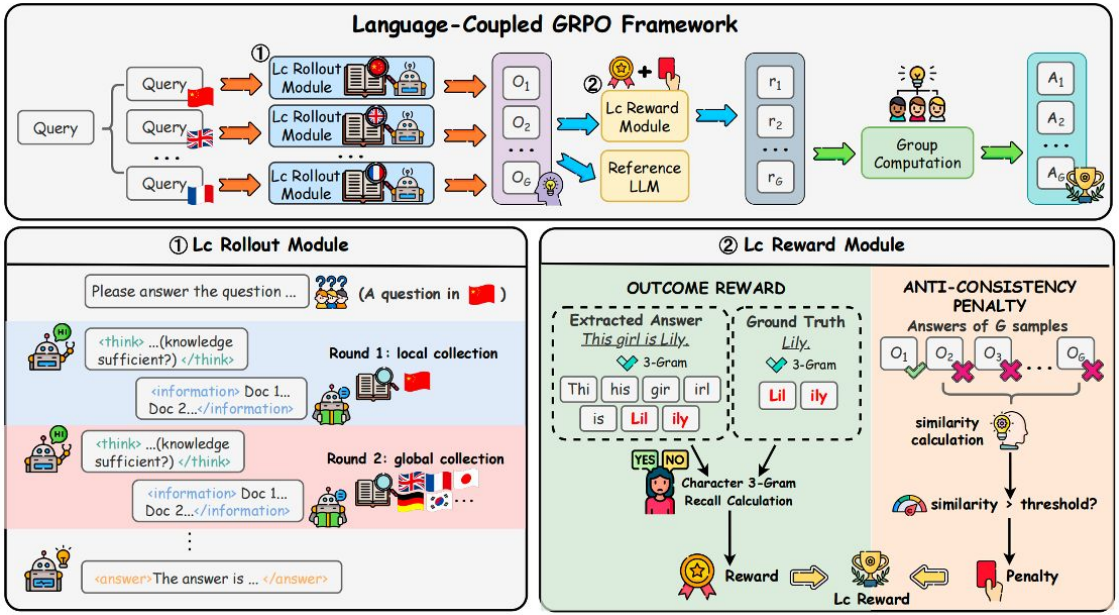


[1] Huang et al. A survey on large language models with multilingualism: Recent advances and new frontiers. AI Review 2026.
[2] Qi et al. Language-Coupled Reinforcement Learning for Multilingual Retrieval-Augmented Generation. arXiv 2026.

Domain-Specific Applications

- LcRL [1]: an RL framework that jointly optimizes multilingual retrieval, reasoning, and reward signals for developing multilingual RAG agents.

- Solutions:
  - Sample semantic equivalent queries in multiple languages.
  - Optimize reasoning trajectories jointly across languages for multilingual retriever calling.



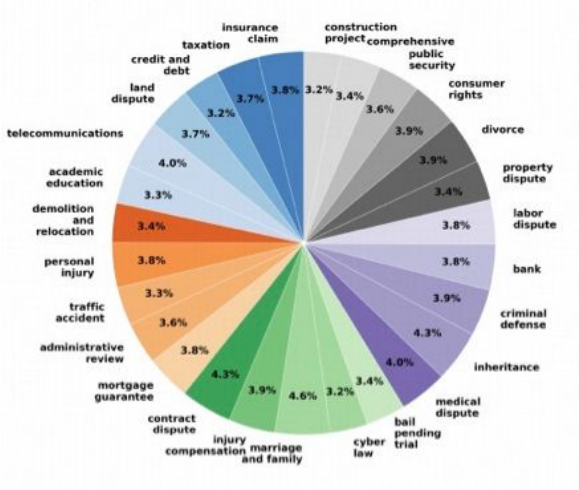
[1] Qi et al. Language-Coupled Reinforcement Learning for Multilingual Retrieval-Augmented Generation. arXiv 2026.

Domain-Specific Applications

Legal consultations often involving progressively unfolding issues

- The relevance of a question to legal knowledge is not simply determined by lexical or semantic similarity
- LexRAG: a benchmark designed for RAG in multi-turn legal consultation conversations.

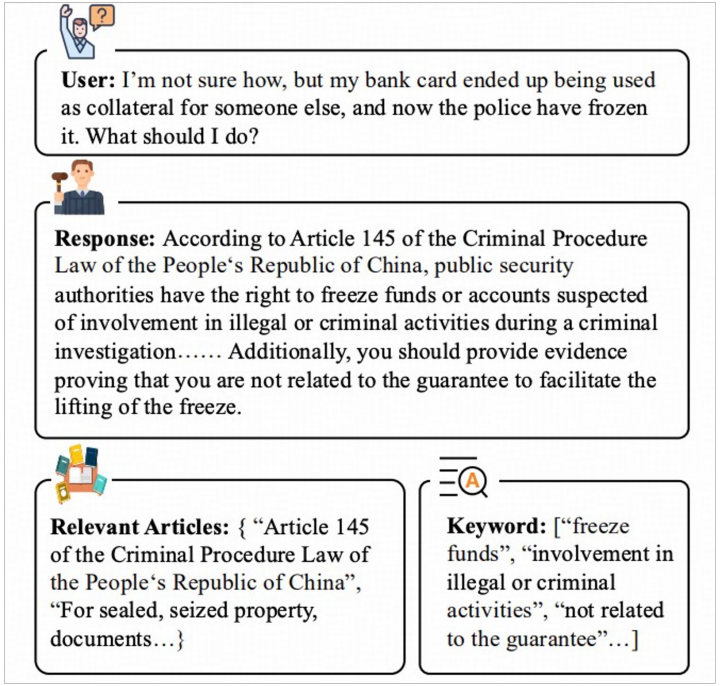
Statistic	#Number
Total Conversations	1,013
Total Queries	5,065
Total Legal Articles	17,728
Avg. Query Length	19.43
Avg. Response Length	165.92
Avg. Relevant Articles per Query	1.09
Avg. Keywords per Query	3.57



[1] Li et al. LexRAG: Benchmarking Retrieval-Augmented Generation in Multi-Turn Legal Consultation Conversation. SIGIR 2025.

Domain-Specific Applications

- **Conversational Knowledge Retrieval**: the RAG system should identify legal articles relevant to the current query while considering the context.
- **Response Generation**: the LLM needs to generate contextually coherent and legally accurate response based on the dialogue history and retrieved legal article.
- **Key difference**: Legal-specific terms, longer context



[1] Li et al. LexRAG: Benchmarking Retrieval-Augmented Generation in Multi-Turn Legal Consultation Conversation. SIGIR 2025.

Q & A

Conclusions and Future Directions

Conclusions and future directions

- We revisited key tasks and concepts in conversational search:
  - The core concepts of conversational search
  - Conversational search paradigms
  - Conversational RAG
  - Mixed-initiative interactions
  - Personalized conversational search
- We explored emerging topics in the era of agents:
  - Agentic search
  - Agentic conversational search
  - Proactive conversational agents
  - Domain-specific applications

Conclusions and future directions

- Future directions
  - Agentic related
    - Efficiency & scalability, running on device
    - Personalization, memory, context management
    - Human-AI collaboration
    - Explainability, transparency & trustworthiness
  - Broader applicability
    - Multilingual and Multimodal scenarios
    - Domain-specific scenarios (financial, legal, medical, etc.)