THE UNIVERSITY *of* EDINBURGH

Edinburgh NLP
University of Edinburgh
Natural Language Processing

Conversational Search: From **Fundamentals** to **Frontiers** in the **LLM** Era

Chuan Meng
University of Edinburgh

23 Oct 2025

About me

Chuan Meng

- Postdoc, EdinburghNLP at University of Edinburgh (since 1 Sep 2025)
  - working with Dr. Jeff Dalton

- PhD, University of Amsterdam (Oct 2021 to Jun 2025)
  - supervised by Prof. Maarten de Rijke, Dr. Mohammad Aliannejadi

- Former Applied Scientist Intern, Amazon

- Research in IR & NLP, with a focus on Agentic Information Access

Overview

Part I: Fundamentals of Conversational Search [25 min]

- Introduction to conversational search
- Conversational search paradigms
- Mixed initiatives
- Personalization

Part II: Emerging Topics in the LLM Era [20 min]

- Conversational retrieval-augmented generation
- Automatic evaluation using LLM judges
- Agentic conversational search

Part III: Summary and Future Directions [5 min]

Part 1
Fundamentals of Conversational Search

Introduction for Conversational Search

Comparison between Conversational and Ad-hoc Search

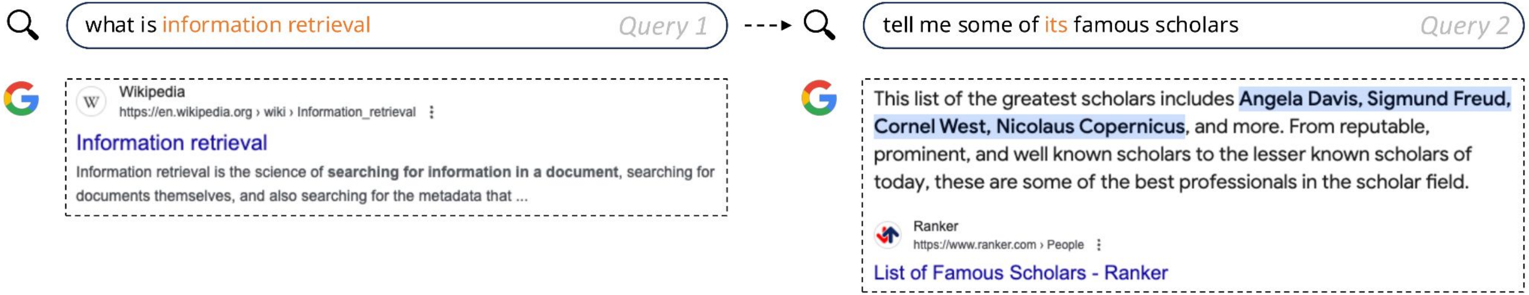
General goal: Conversational search aims to identify relevant documents to satisfy users' complex information needs through multi-turn interactions.

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

Ad-hoc Search



what is information retrieval Query 1

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 ...

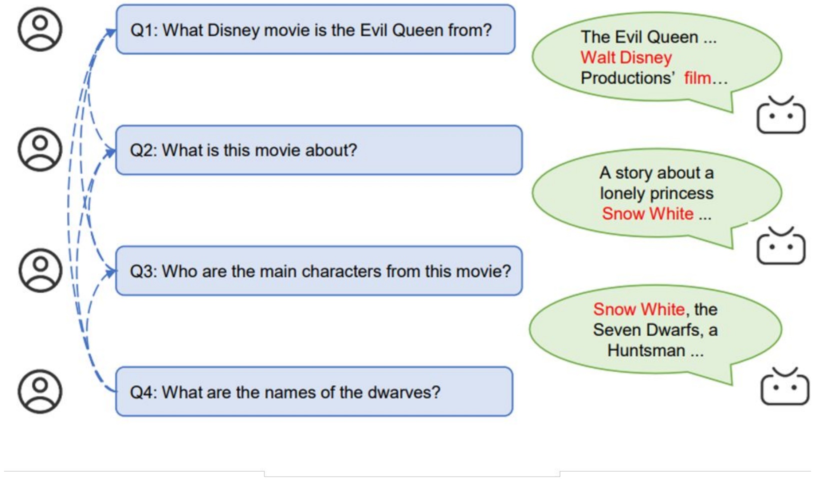
tell me some of its famous scholars Query 2

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

Conversational Search



Q1: What Disney movie is the Evil Queen from?

The Evil Queen ... Walt Disney Productions' film...

Q2: What is this movie about?

A story about a lonely princess Snow White ...

Q3: Who are the main characters from this movie?

Snow White, the Seven Dwarfs, a Huntsman ...

Q4: What are the names of the dwarves?

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

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)
  - Current 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

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?

Widely Used Datasets

From NLP community

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

From IR community

- ➢ TREC CAsT 2019-2022 [5] and TREC iKAT 2023-2024 [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.

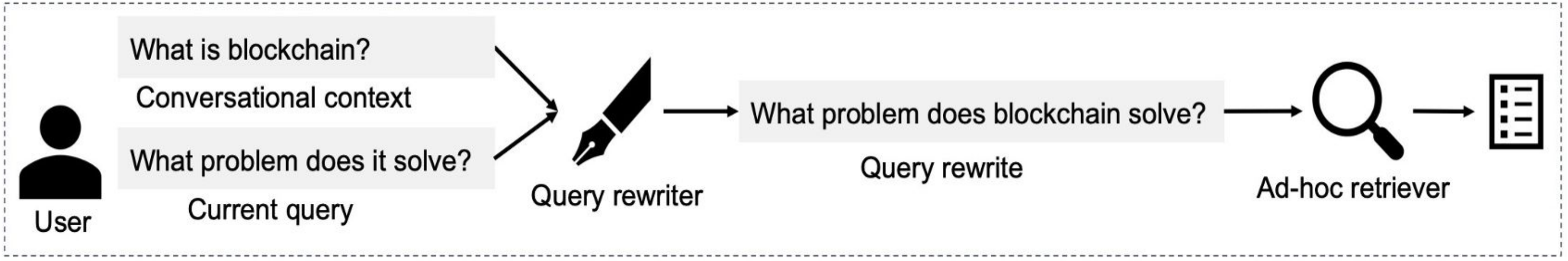
Two Paradigms to achieve Conversational Search

1. Conversational Query Rewriting
2. Conversational Dense Retrieval

Two Conversational Search Paradigms

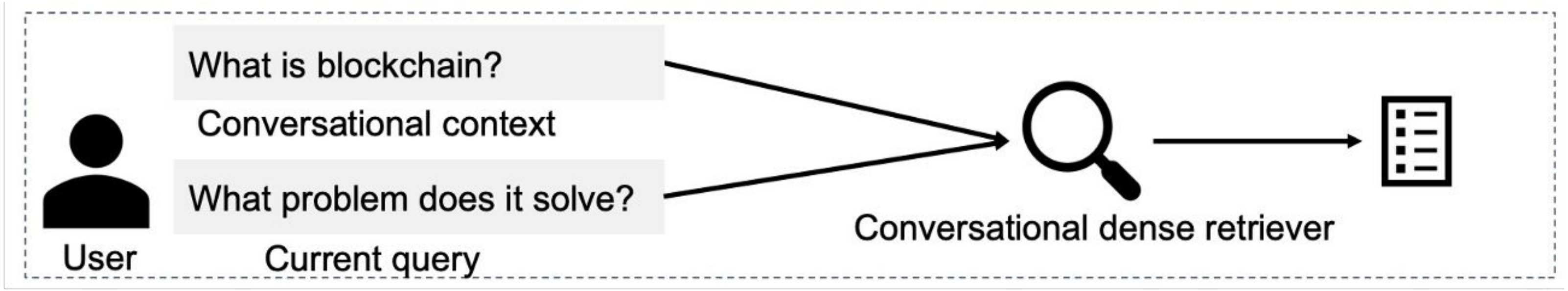
Conversational Query Rewriting (CQR)

➢ 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 methods 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:

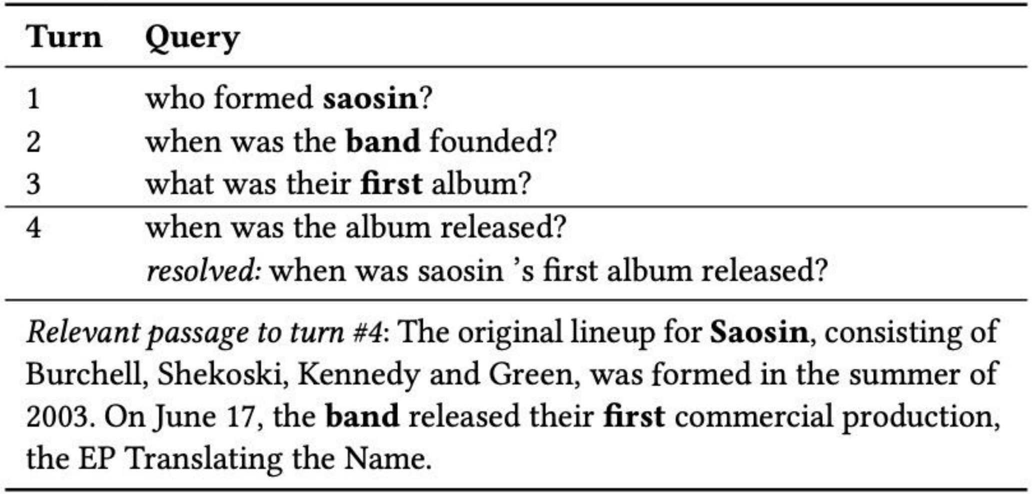
➢ Prompting LLMs to directly rewrite context-dependent query.

➢ 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?

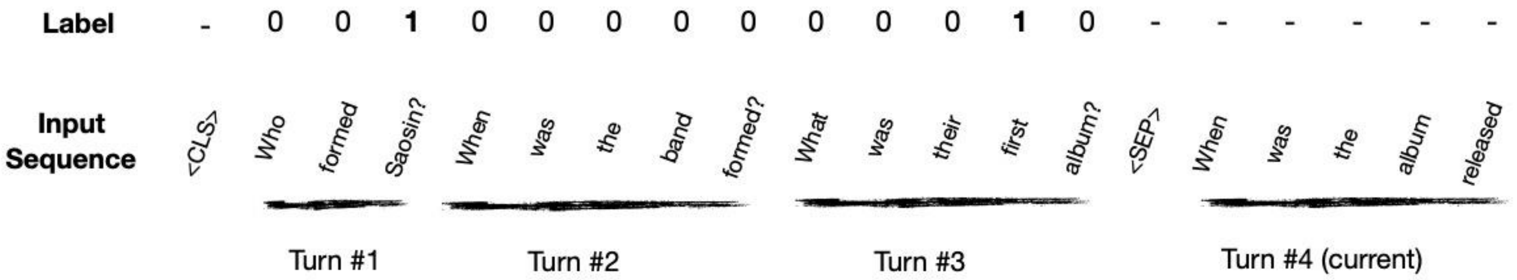
Relevant passage to turn #4: The original lineup for **Saosin**, consisting of Burchell, Shekoski, Kennedy and Green, was formed in the summer of 2003. On June 17, the **band** released their **first** commercial production, the EP Translating the Name.

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

Conversational Query Rewriting

Selecting useful terms from historical context

➢ [1,2,3] train a binary classifier or selector to select useful terms in the context



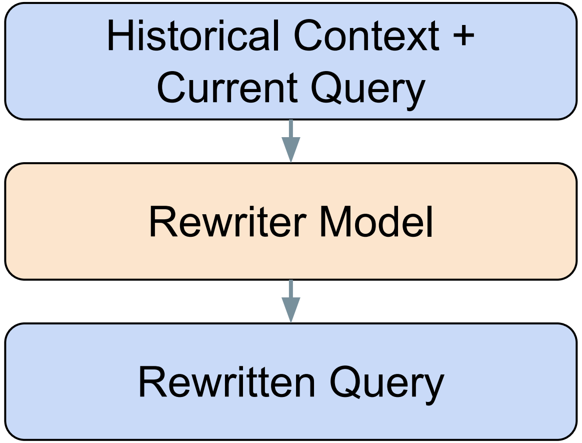
Label	-	0	0	1	0	0	0	0	0	0	0	0	1	0	-	-	-	-	-	-
Input Sequence	<CLS>	Who	formed	Saosin?	When	was	the	band	formed?	What	was	their	first	album?	<SEP>	When	was	the	album	released

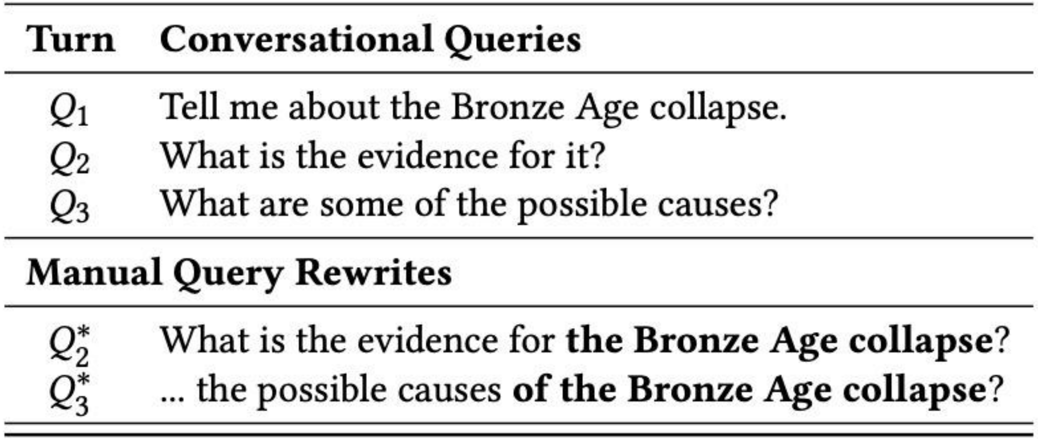
[1] Query resolution for conversational search with limited supervision. Voskarides et al. SIGIR 2020.
[2] Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting. Lin et al. TOIS 2021.
[3] Contextualized Query Embeddings for Conversational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

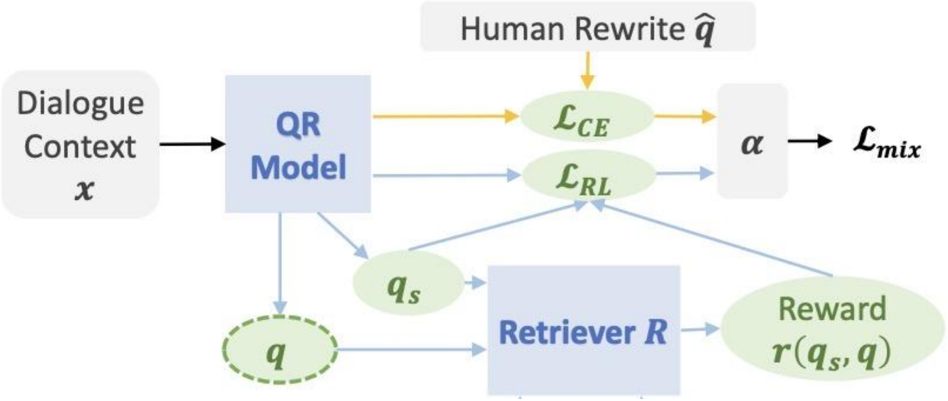
- ➢ **Cons**: rely heavily on manual labels and cannot optimize for search performance

[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

➢ **Approach**: The search signals could be used to train rewriters via fine-tuning [3,4] or reinforcement learning [1,2].

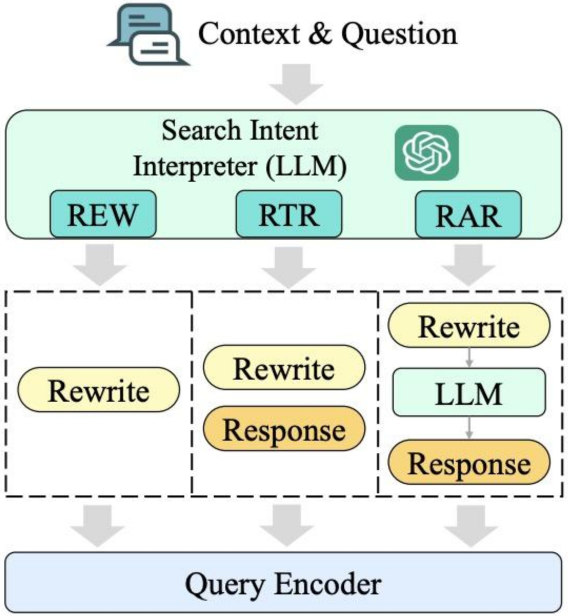


[1] CONQRR: Conversational Query Rewriting for Retrieval with Reinforcement Learning. Wu et al. EMNLP 2022.
[2] Reinforced Question Rewriting for Conversational Question Answering. Chen et al. EMNLP 2022.
[3] ConvGQR: Generative Query Reformulation for Conversational Search. Mo et al. ACL 2023.
[4] Search-Oriented 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 [1,2,3]
  - ○ [1] generates different types of queries and then aggregate them
- ➢ Limitation: High inference cost by calling LLMs



[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

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.

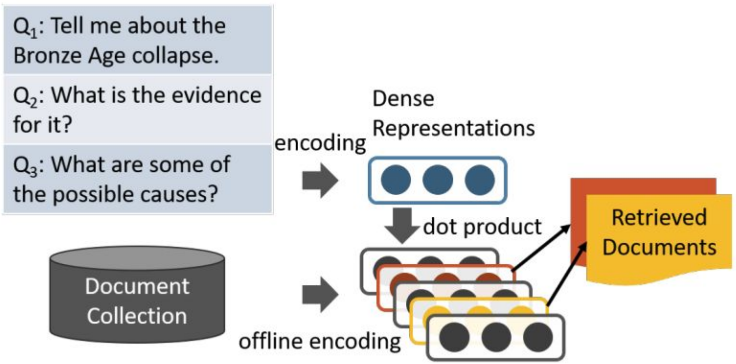
[1] ConvGQR: Generative Query Reformulation for Conversational Search. Mo et al. ACL 2023.
[2] IterCQR: Iterative Conversational Query Reformulation without Human Supervision. Jang et al. NAACL 2023.
[3] CHIQ: Contextual History Enhancement for Improving Query Rewriting in Conversational Search.. Mo 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.

Q & A

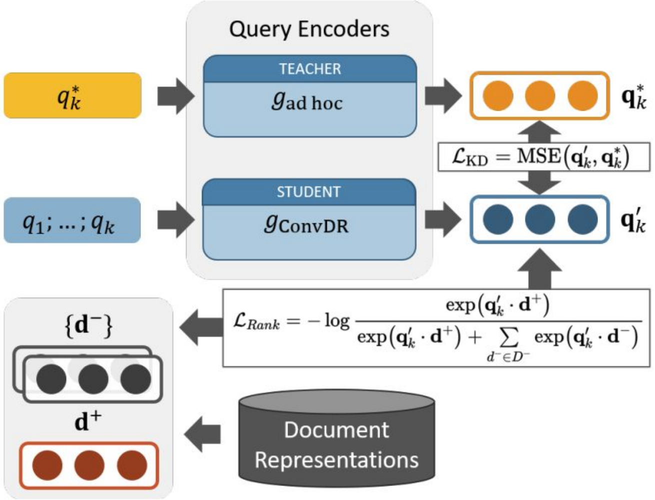
Conversational Dense Retrieval

Teacher-student framework

- [1] learns a student query encoder (fed with raw conversational query) to mimic the embeddings from a teacher encoder (self-contained human rewritten 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?
encoding
Dense Representations
dot product
Retrieved Documents
Document Collection
offline encoding



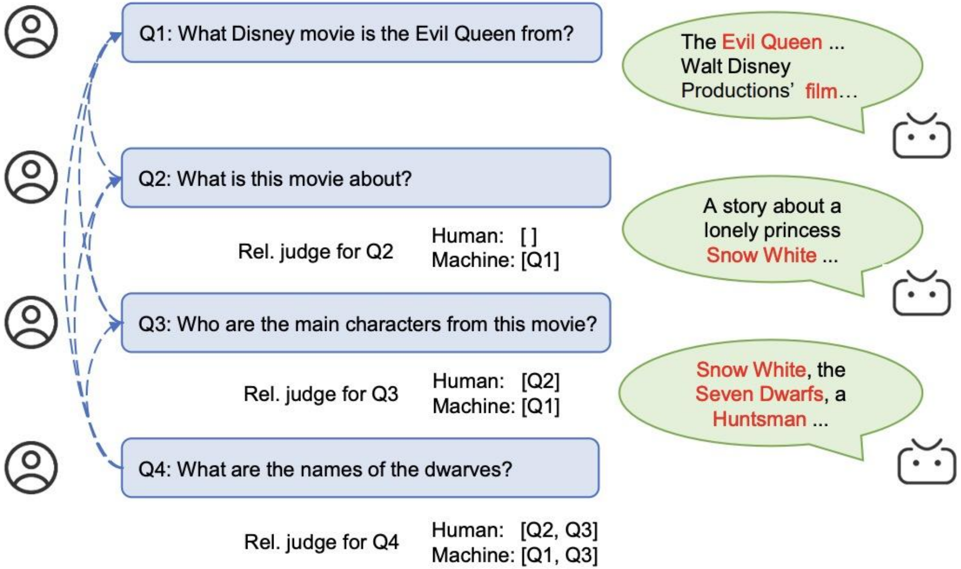
Query Encoders
TEACHER
$g_{\text{ad hoc}}$
STUDENT
g_{ConvDR}
q_k^*
$q_1; \dots ; q_k$
$\mathbf{q}_k^*$
$\mathbf{q}_k'$
$\mathcal{L}_{\text{KD}} = \text{MSE}(\mathbf{q}_k', \mathbf{q}_k^*)$
$\mathcal{L}_{Rank} = -\log \frac{\exp(\mathbf{q}_k' \cdot \mathbf{d}^+)}{\exp(\mathbf{q}_k' \cdot \mathbf{d}^+) + \sum_{d^- \in D^-} \exp(\mathbf{q}_k' \cdot \mathbf{d}^-)}$
$\{\mathbf{d}^-\}$
$\mathbf{d}^+$
Document Representations

[1] Few-Shot Conversational Dense Retrieval. Yu et al. SIGIR 2021.

Conversational Dense Retrieval

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.



Q1: What Disney movie is the Evil Queen from?

The Evil Queen ... Walt Disney Productions' film…

Q2: What is this movie about?

Rel. judge for Q2 Human: [ ] Machine: [Q1]

A story about a lonely princess Snow White ...

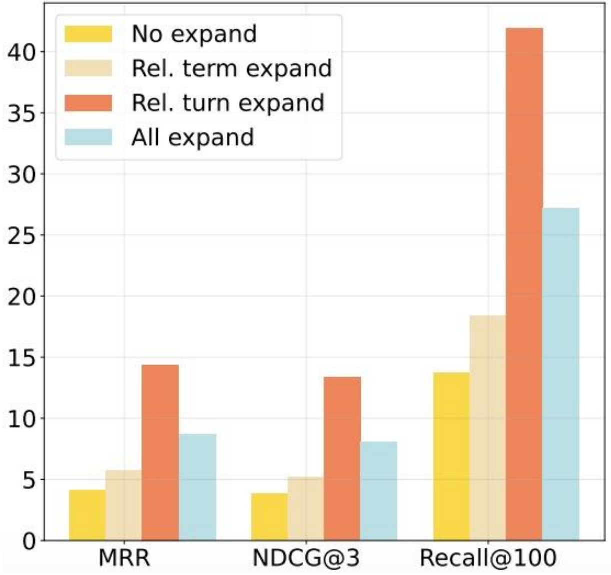
Q3: Who are the main characters from this movie?

Rel. judge for Q3 Human: [Q2] Machine: [Q1]

Snow White, the Seven Dwarfs, a Huntsman ...

Q4: What are the names of the dwarves?

Rel. judge for Q4 Human: [Q2, Q3] Machine: [Q1, Q3]



[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.

Summary

Conversational query rewriting

- **Pros**: Can re-use any existing retrievers and has good interpretability with explicit rewritten query.
- **Cons**: Cannot directly optimize for ranking performance; and the rewriter model training rely on available annotations as supervision signals.

Conversational dense retrieval

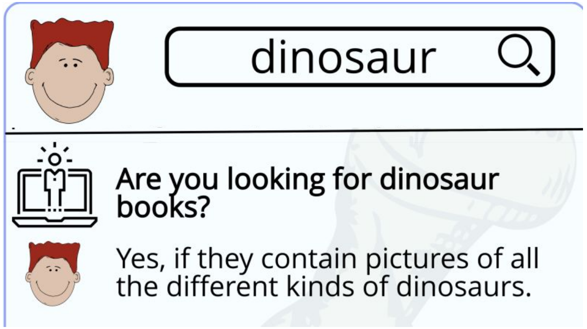
- **Pros**: Direct optimize with conversational session to obtain representation in an end-to-end way
- **Cons**: Data scarcity problem and de-noising requirement for the input context.

Q & A

Mixed Initiatives

Mixed Initiatives

- What is mixed initiative?
  - User and system can both take the initiative at different times in a conversation [1]
  - System can take 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]



dinosaur

Are you looking for dinosaur books?

Yes, if they contain pictures of all the different kinds of dinosaurs.

[1] Radlinski et al. A Theoretical Framework for Conversational Search. CHIIR 2017.

[2] Kiesel et al. Toward voice query clarification. SIGIR 2018.

Mixed Initiatives

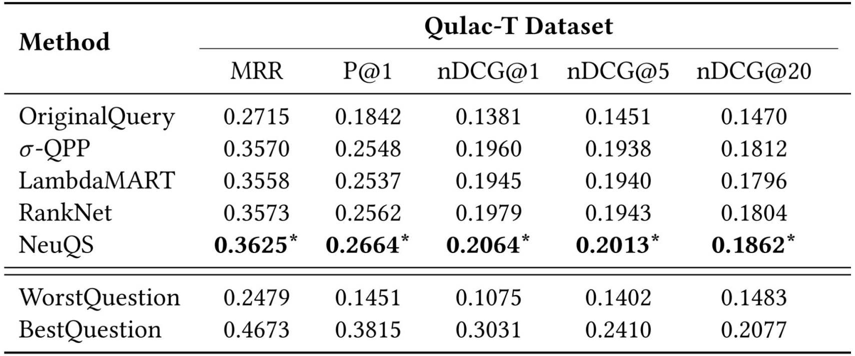
- Scope for mixed initiatives
    - What
        - Clarifying question selection/generation
        - Conversation contextualisation/interest anticipation
    - When
        - Clarification need prediction

Mixed Initiatives

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

Mixed Initiatives

- 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



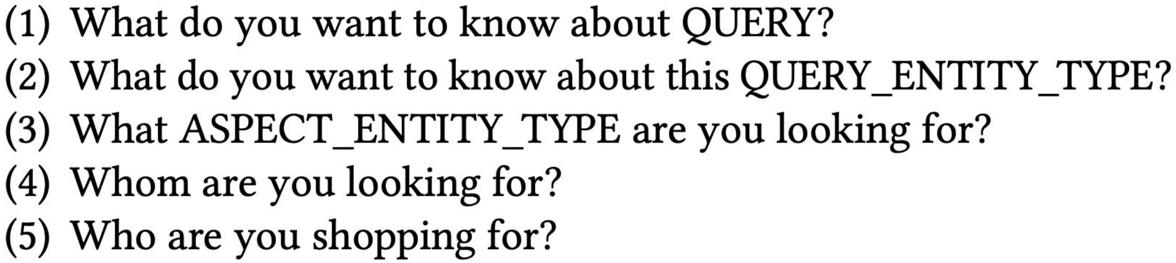
Method	**Qulac-T Dataset**				
	MRR	P@1	nDCG@1	nDCG@5	nDCG@20
OriginalQuery	0.2715	0.1842	0.1381	0.1451	0.1470
σ-QPP	0.3570	0.2548	0.1960	0.1938	0.1812
LambdaMART	0.3558	0.2537	0.1945	0.1940	0.1796
RankNet	0.3573	0.2562	0.1979	0.1943	0.1804
NeuQS	**0.3625*** 	**0.2664***	**0.2064***	**0.2013***	**0.1862***
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 Initiatives

- 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

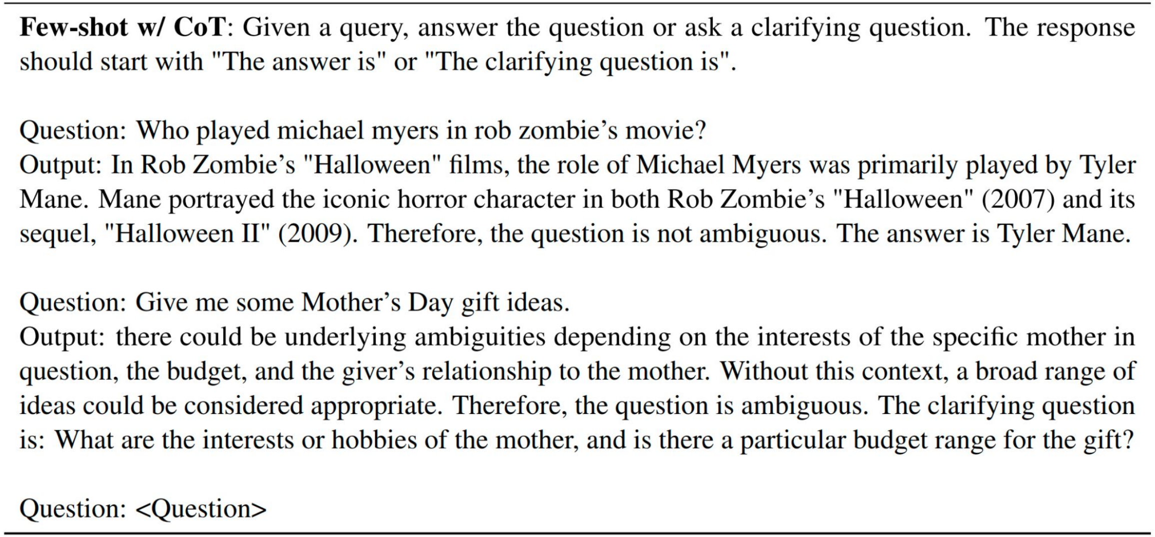


(1) What do you want to know about QUERY?
(2) What do you want to know about this QUERY_ENTITY_TYPE?
(3) What ASPECT_ENTITY_TYPE are you looking for?
(4) Whom are you looking for?
(5) Who are you shopping for?

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

Mixed Initiatives

- 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?

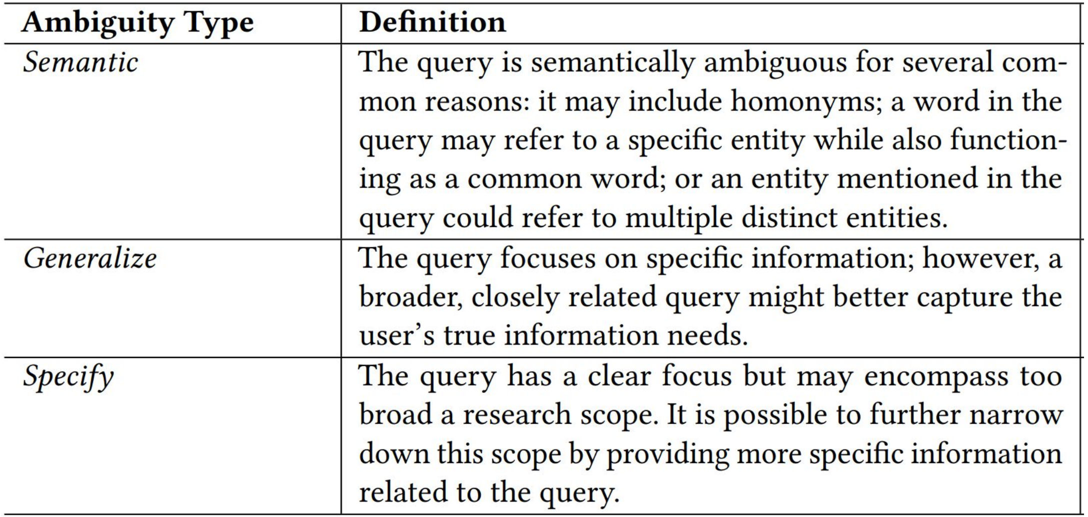
Question: <Question>

[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 Initiatives

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

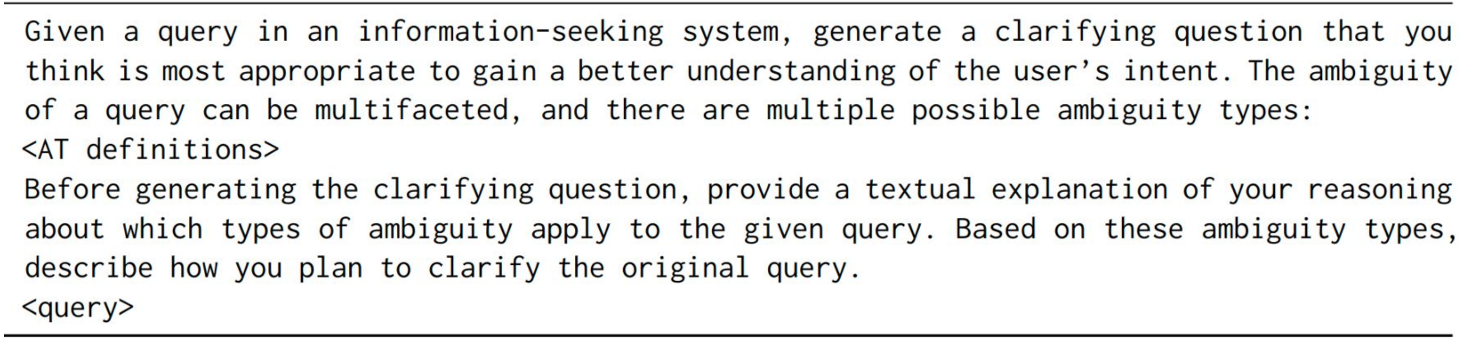


Ambiguity Type	**Definition**
Semantic	The query is semantically ambiguous for several common reasons: it may include homonyms; a word in the query may refer to a specific entity while also functioning as a common word; or an entity mentioned in the query could refer to multiple distinct entities.
Generalize	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 Initiatives

- 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>
```

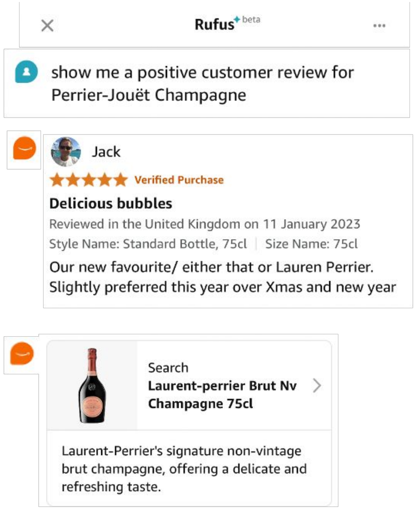
[1] Tang et al. Clarifying Ambiguities: on the Role of Ambiguity Types in Prompting Methods for Clarification Generation. SIGIR 2025.

Mixed Initiatives

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

Mixed Initiatives

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



Rufus beta

show me a positive customer review for Perrier-Jouët Champagne

Jack

★★★★★ Verified Purchase

Delicious bubbles

Reviewed in the United Kingdom on 11 January 2023

Style Name: Standard Bottle, 75cl | Size Name: 75cl

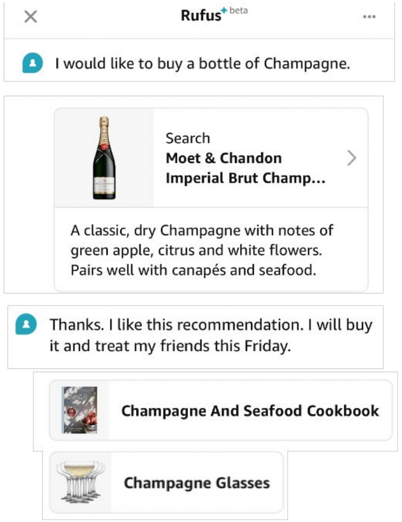
Our new favourite/ either that or Lauren Perrier. Slightly preferred this year over Xmas and new year

Search **Laurent-perrier Brut Nv Champagne 75cl**

Laurent-Perrier's signature non-vintage brut champagne, offering a delicate and refreshing taste.

User might ask: what is Lauren Perrier?

Conversation contextualisation



Rufus beta

I would like to buy a bottle of Champagne.

Search **Moet & Chandon Imperial Brut Champ...**

A classic, dry Champagne with notes of green apple, citrus and white flowers. Pairs well with canapés and seafood.

Thanks. I like this recommendation. I will buy it and treat my friends this Friday.

Champagne And Seafood Cookbook

Champagne Glasses

User might ask: What food should I pair Chamagene with?

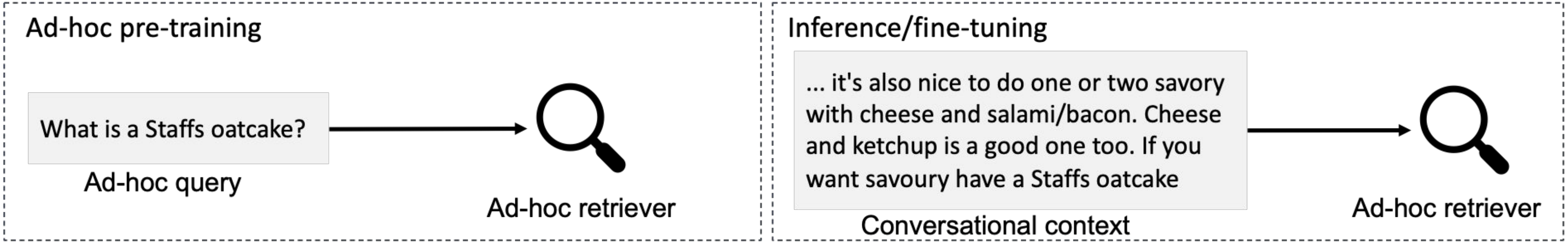
Interest anticipation

[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.

Mixed Initiatives

- 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]



Ad-hoc pre-training

What is a Staffs oatcake?

Ad-hoc query

Ad-hoc retriever

Inference/fine-tuning

... it's also nice to do one or two savory with cheese and salami/bacon. Cheese and ketchup is a good one too. If you want savoury have a Staffs oatcake

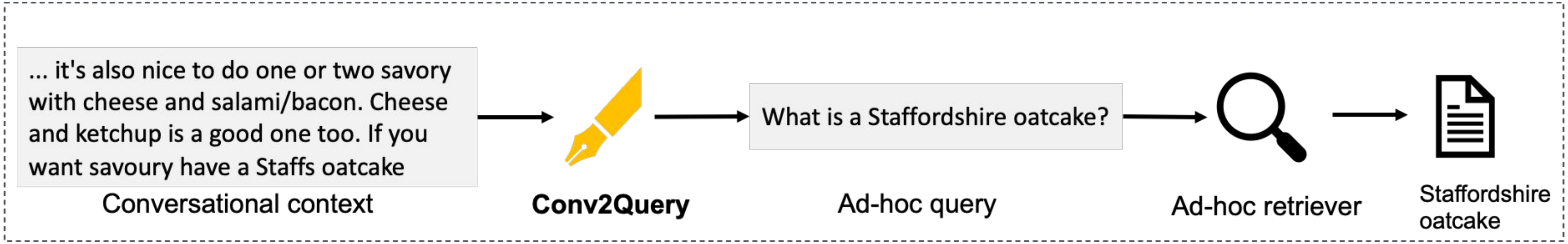
Conversational context

Ad-hoc retriever

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

Mixed Initiatives

- 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


... it's also nice to do one or two savory with cheese and salami/bacon. Cheese and ketchup is a good one too. If you want savoury have a Staffs oatcake
Conversational context
Conv2Query
What is a Staffordshire oatcake?
Ad-hoc query
Ad-hoc retriever
Staffordshire oatcake

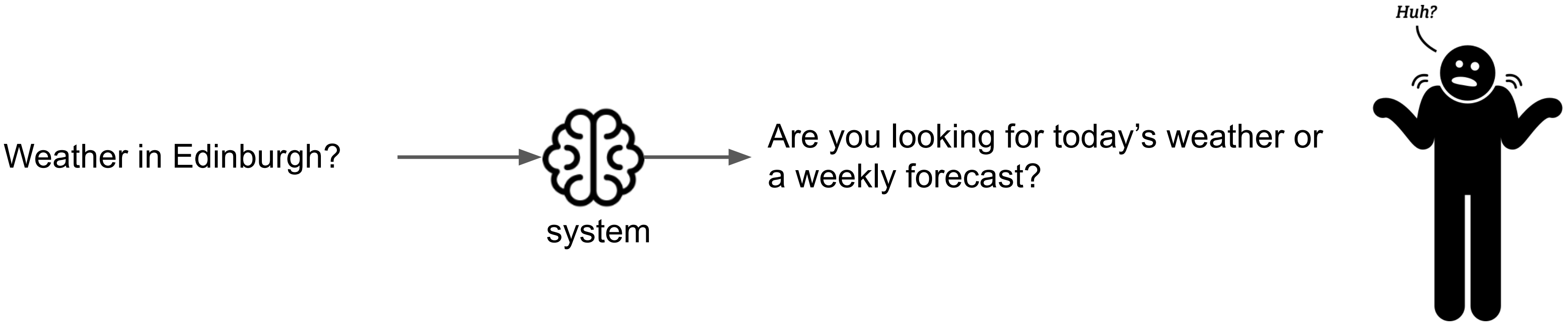
[1] Meng et al. Bridging the Gap: From Ad-hoc to Proactive Search in Conversations. SIGIR 2025.

Mixed Initiatives

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



Weather in Edinburgh?

system

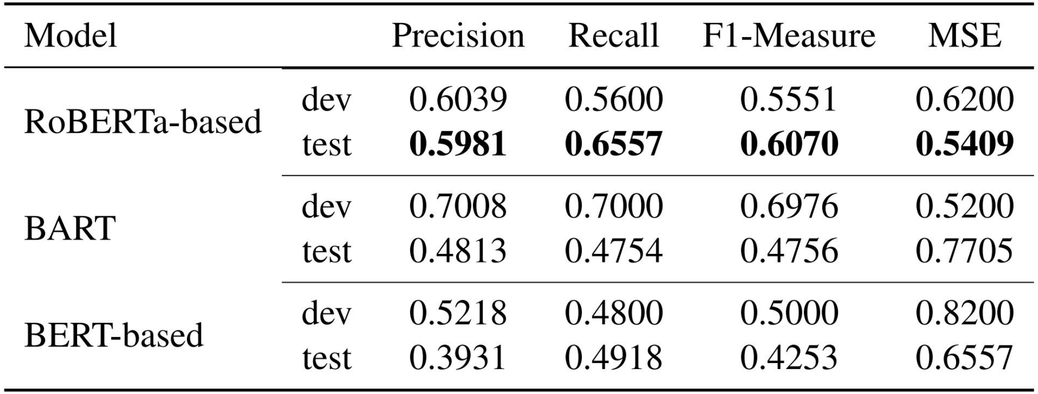
Are you looking for today’s weather or a weekly forecast?

[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
  - [1,2,3] 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)



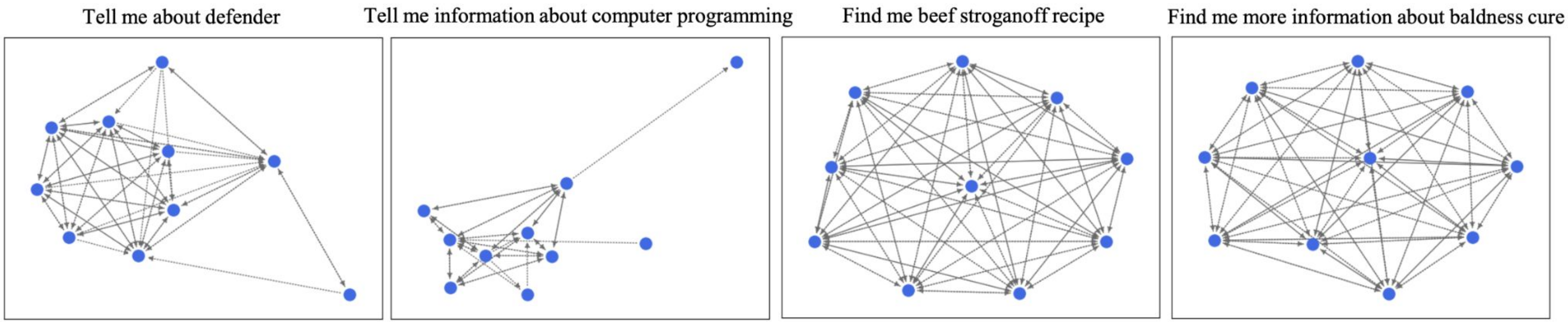
Model		Precision	Recall	F1-Measure	MSE
RoBERTa-based	dev	0.6039	0.5600	0.5551	0.6200
	test	**0.5981**	**0.6557**	**0.6070**	**0.5409**
BART	dev	0.7008	0.7000	0.6976	0.5200
	test	0.4813	0.4754	0.4756	0.7705
BERT-based	dev	0.5218	0.4800	0.5000	0.8200
	test	0.3931	0.4918	0.4253	0.6557

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



Tell me about defender

Tell me information about computer programming

Find me beef stroganoff recipe

Find me more information about baldness cure

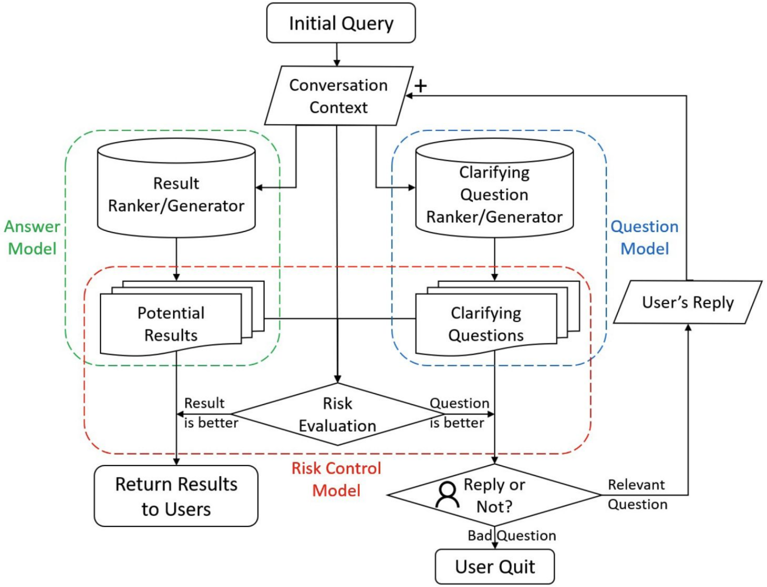
Clarifying question is required

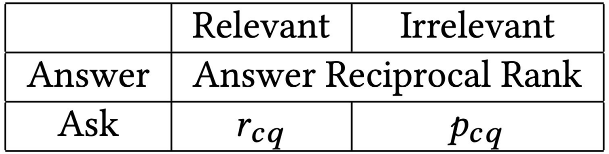
No clarifying question is needed

[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





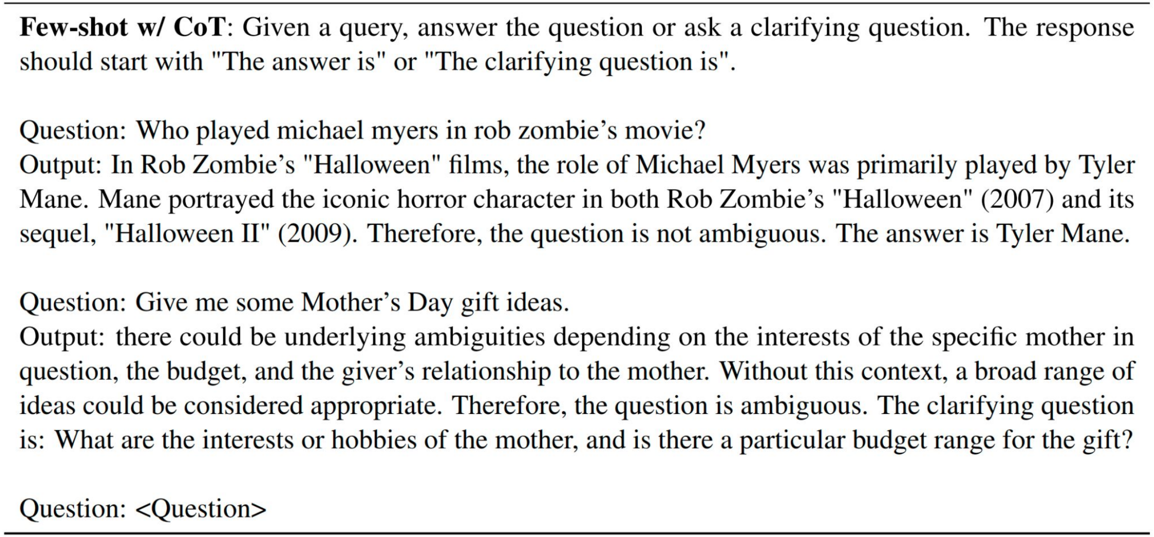
	Relevant	Irrelevant
Answer	Answer Reciprocal Rank	
Ask	r_{cq}	p_{cq}

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.

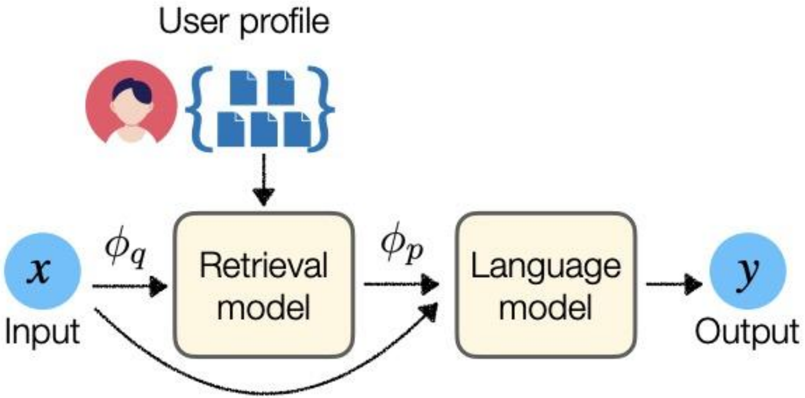
[2] Deng et al. Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration. EMNLP 2023.

Q & A

Personalization

Personalization

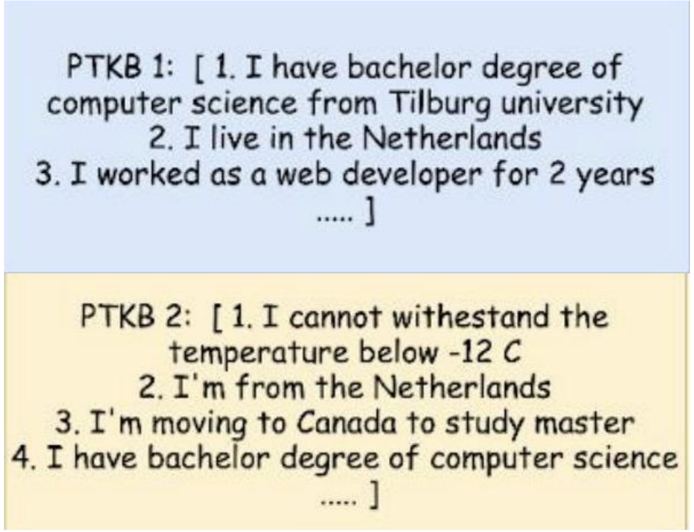
- **Goal**: Satisfy users' complex information needs based on **users' profiles and preference** through multi-turn interactions.
- **Assumption**: The same query turn from different users may correspond to different search intents, thus yielding different results.
- **User information**: Profile, historical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



Personalization

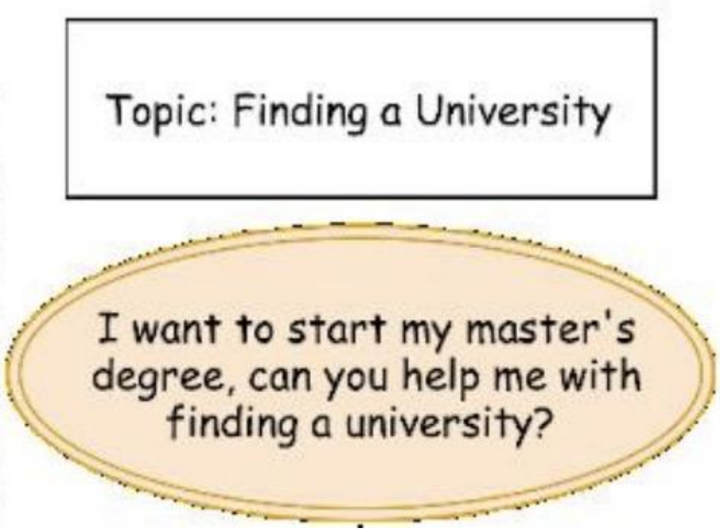
Incorporating explicit user profile into query rewriting

- ➢ 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 withstand 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
.....]



Topic: Finding a University

I want to start my master's degree, can you help me with finding a university?

[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.

Personalization

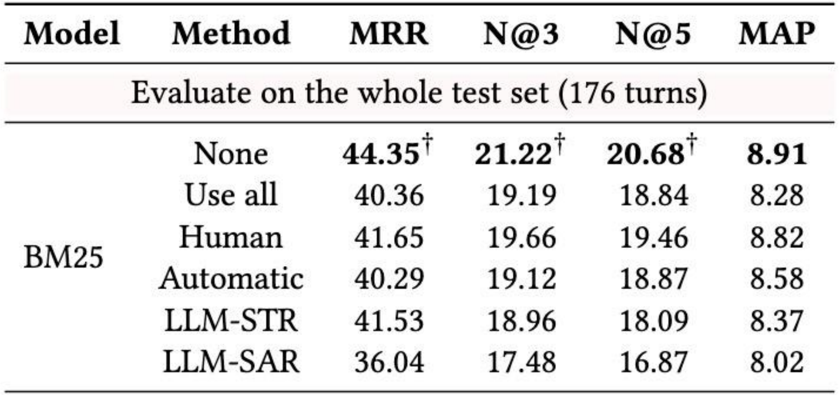
Incorporating explicit user profile into query rewriting

- ➢ **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)

Personalization

Incorporating explicit user profile into query rewriting

- **Observation [1]**: using personal info at a wrong time or using all historical turns 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

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

Q & A

Part 2

Emerging Topics in the LLM Era

Conversational Search in the LLM Era

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

- ➢ User expect to get (customized) **final response** instead of browsing documents; LLMs are good at generating natural responses
- ➢ LLMs can replace humans in automatically evaluating conversational search systems
- ➢ LLMs can be viewed as agents that enable more autonomous systems.

Overview

Part II: Emerging Topics in the LLM Era

- Conversational retrieval-augmented generation
- Automatic evaluation using LLM judges
- Agentic conversational search

Conversational retrieval-augmented generation

Conversational retrieval-augmented generation

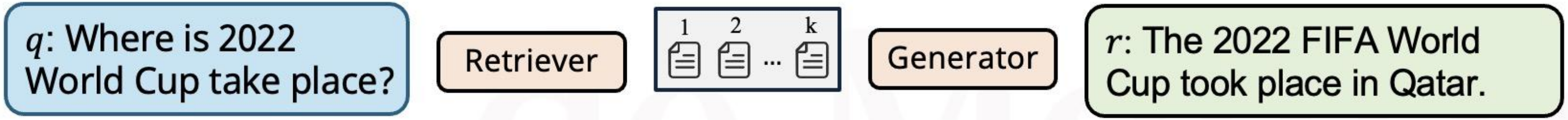
Conversational retrieval-augmented generation (RAG)

- ➢ Single turn RAG v.s. Conversational (Multi-turn) RAG
- ➢ Leveraging historical information for conversational RAG

Conversational retrieval-augmented generation

Single turn RAG [1]

- ➢ **Trend**: LLMs can direct reply users' question with their parametric knowledge.
- ➢ **Challenge**:
  - ○ LLMs often generate plausible but factually incorrect text (hallucination)
  - ○ LLMs' internal knowledge can be out-of-date
- ➢ **Goal**: Incorporate the retrieved up-to-date information for generation.
- ➢ **Paradigm**: Generate response for a query on top of retrieved information.

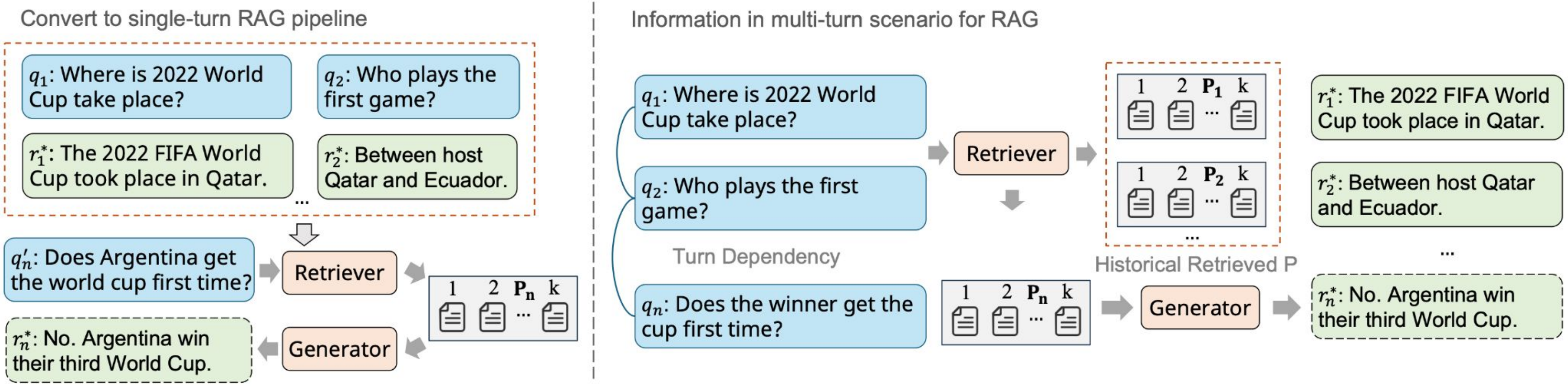

q: Where is 2022 World Cup take place?
Retriever
1
2
...
k
Generator
r: The 2022 FIFA World Cup took place in Qatar.

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

Conversational retrieval-augmented generation

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?


Convert to single-turn RAG pipeline
q_1: Where is 2022 World Cup take place?
q_2: Who plays the first game?
r_1^*: The 2022 FIFA World Cup took place in Qatar.
r_2^*: Between host Qatar and Ecuador.
...
q_n': Does Argentina get the world cup first time?
Retriever
1 2 P_n ... k
Generator
r_n^*: No. Argentina win their third World Cup.
Information in multi-turn scenario for RAG
q_1: Where is 2022 World Cup take place?
q_2: Who plays the first game?
Turn Dependency
q_n: Does the winner get the cup first time?
Retriever
1 2 P_1 ... k
1 2 P_2 ... k
...
Historical Retrieved P
1 2 P_n ... k
Generator
r_1^*: The 2022 FIFA World Cup took place in Qatar.
r_2^*: Between host Qatar and Ecuador.
...
r_n^*: No. Argentina win their third World Cup.

[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.

Conversational retrieval-augmented generation

Conversational retrieval-augmented generation (RAG)

➢ Leveraging historical information for conversational RAG

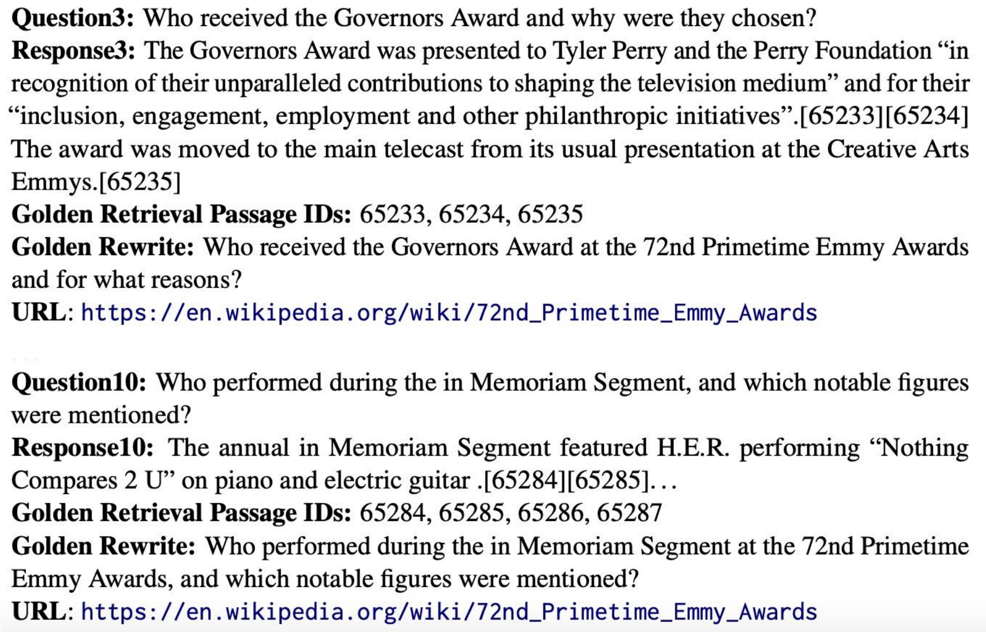
- **Idea**: The historical information (not limited to query-response pairs) could enhance the effectiveness [1,3], efficiency [2], and truthfulness [1] of the current turn RAG.

[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.

Conversational retrieval-augmented generation

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`

...

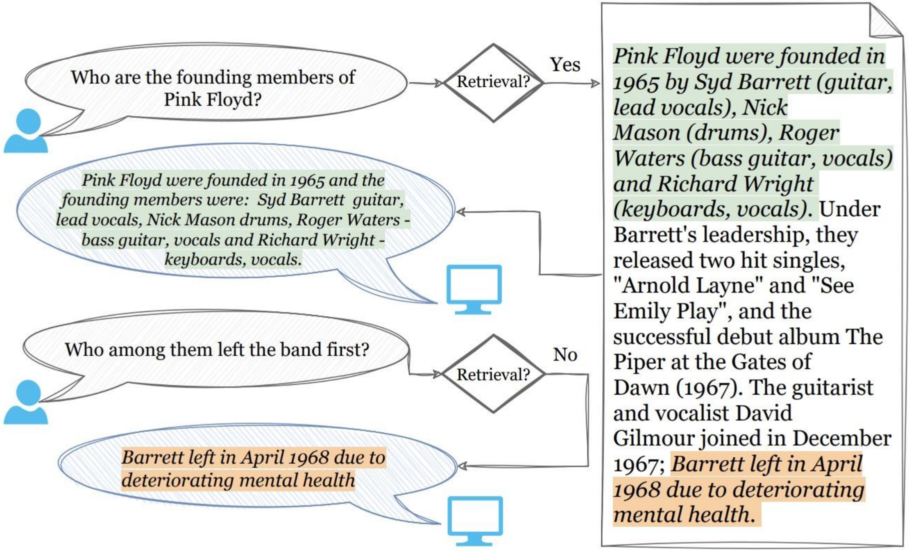
Question10: Who performed during the in Memoriam Segment, and which notable figures were mentioned?
Response10: The annual in Memoriam Segment featured H.E.R. performing "Nothing Compares 2 U" on piano and electric guitar .[65284][65285]. . .
Golden Retrieval Passage IDs: 65284, 65285, 65286, 65287
Golden 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.

Conversational retrieval-augmented generation

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?



Who are the founding members of Pink Floyd?

Pink Floyd were founded in 1965 and the founding members were: Syd Barrett guitar, lead vocals, Nick Mason drums, Roger Waters - bass guitar, vocals and Richard Wright - keyboards, vocals.

Who among them left the band first?

Barrett left in April 1968 due to deteriorating mental health

Pink Floyd were founded in 1965 by Syd Barrett (guitar, lead vocals), Nick Mason (drums), Roger Waters (bass guitar, vocals) and Richard Wright (keyboards, vocals). Under Barrett's leadership, they released two hit singles, "Arnold Layne" and "See Emily Play", and the successful debut album The Piper at the Gates of Dawn (1967). The guitarist and vocalist David Gilmour joined in December 1967; *Barrett left in April 1968 due to deteriorating mental health.*

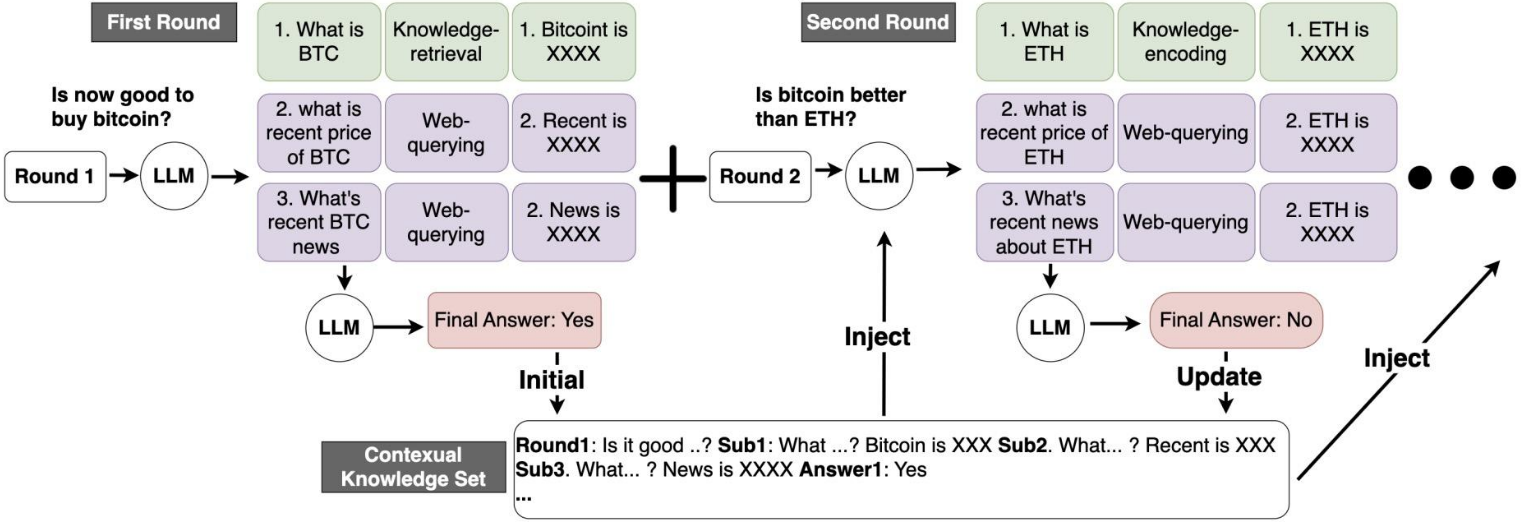
[1] Learning When to Retrieve, What to Rewrite, and How to Respond in Conversational QA. Roy et al. EMNLP 2024.

Conversational retrieval-augmented generation

Conversational retrieval-augmented generation (RAG)

➢ Leveraging historical information for conversational RAG

- ○ **Idea**: [1] maintains 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

Summary:

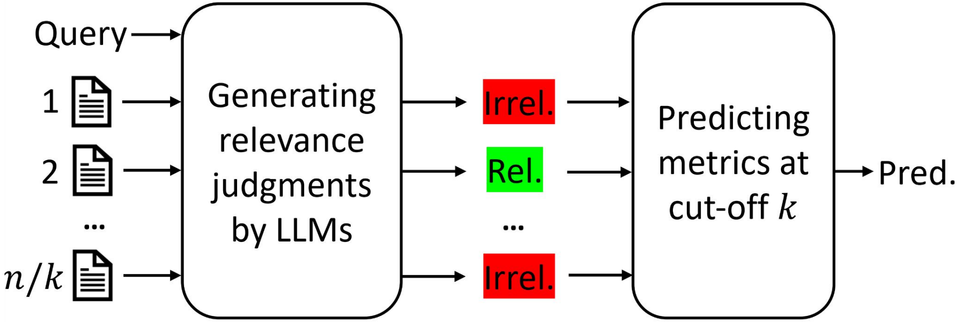
- **Conclusion**: The useful information from historical turns can improve system performance from different perspectives.
- **Key Challenge**: Identify the useful information from super noisy history.
- **Open questions**:
  - How to better leverage historical information for conversational RAG?
  -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

Q & A

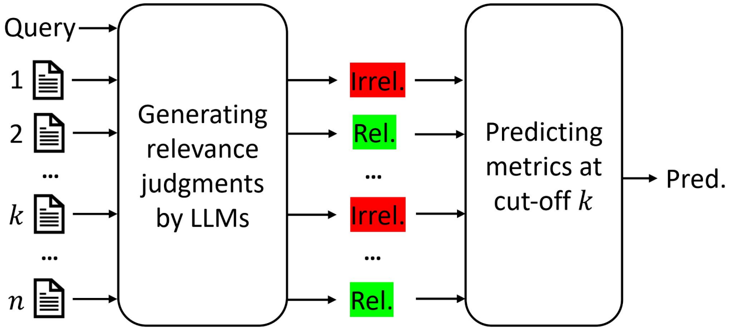
Automatic Evaluation using LLM Judges

Automatic evaluation using LLM judges

- [1] proposes QPP-GenRE, which predicts IR measures using LLM-generated judgments
  - Supports both ad-hoc and conversational search (via query rewrites)
  - [1] devises an approximation strategy for predicting recall-based metrics
    - Only judges the top $\boldsymbol{n}$ items in the ranked list ($\boldsymbol{n} \ll$ total corpus size) to avoid scanning the full corpus


Query
1
2
…
n/k
Generating relevance judgments by LLMs
Irrel.
Rel.
…
Irrel.
Predicting metrics at cut-off k
Pred.

Predicting a precision-based metric

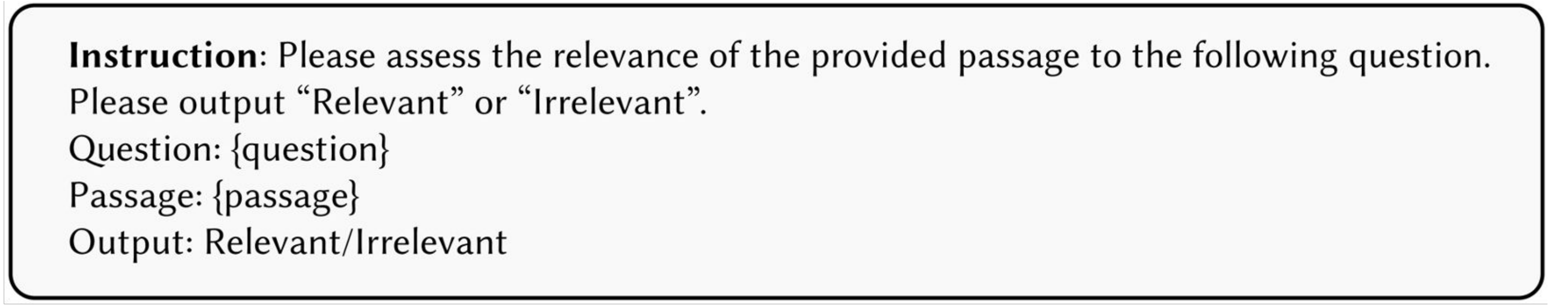

Query
1
2
…
k
…
n
Generating relevance judgments by LLMs
Irrel.
Rel.
…
Irrel.
…
Rel.
Predicting metrics at cut-off k
Pred.

Predicting a metric considering recall

[1] Meng et al. Query Performance Prediction using Relevance Judgments Generated by Large Language Models. TOIS 2025.

Automatic evaluation using LLM judges

- [1] found prompting LLMs for relevance prediction yields limited and unstable performance
- [1] fine-tune LLMs for relevance prediction
    - LLMs: Llama and Mistral families, with sizes ranging from 1B to 70B
    - Fine-tuning method: QLoRA, a parameter-efficient fine-tuning method
    - Training data: human-labeled relevance judgments of MS MARCO

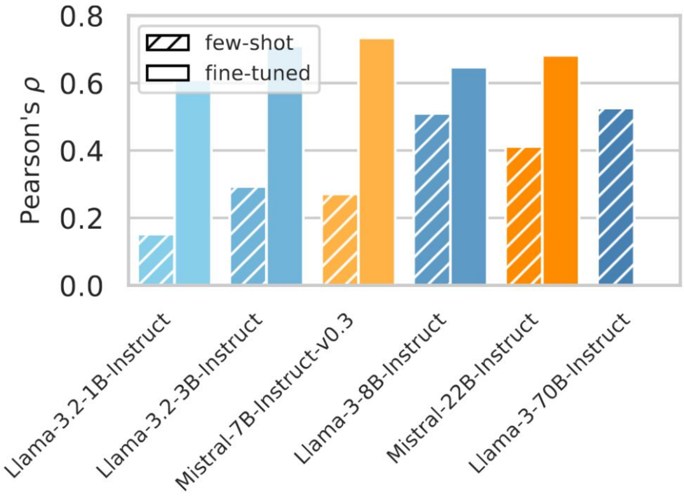


Instruction: Please assess the relevance of the provided passage to the following question. Please output “Relevant” or “Irrelevant”.
Question: {question}
Passage: {passage}
Output: Relevant/Irrelevant

[1] Meng et al. Query Performance Prediction using Relevance Judgments Generated by Large Language Models. TOIS 2025.

Automatic evaluation using LLM judges

- [1] shows that
  - fine-tuning enhances relevance judgment generation and QPP
  - fine-tuning much smaller LLM can yield more effective results than few-shot prompting with much larger models



(a) TREC-DL 19

[1] Meng et al. Query Performance Prediction using Relevance Judgments Generated by Large Language Models. TOIS 2025.

Q & A

Agentic Conversational Search

Agentic Conversational 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
  - Tool use
  - 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.

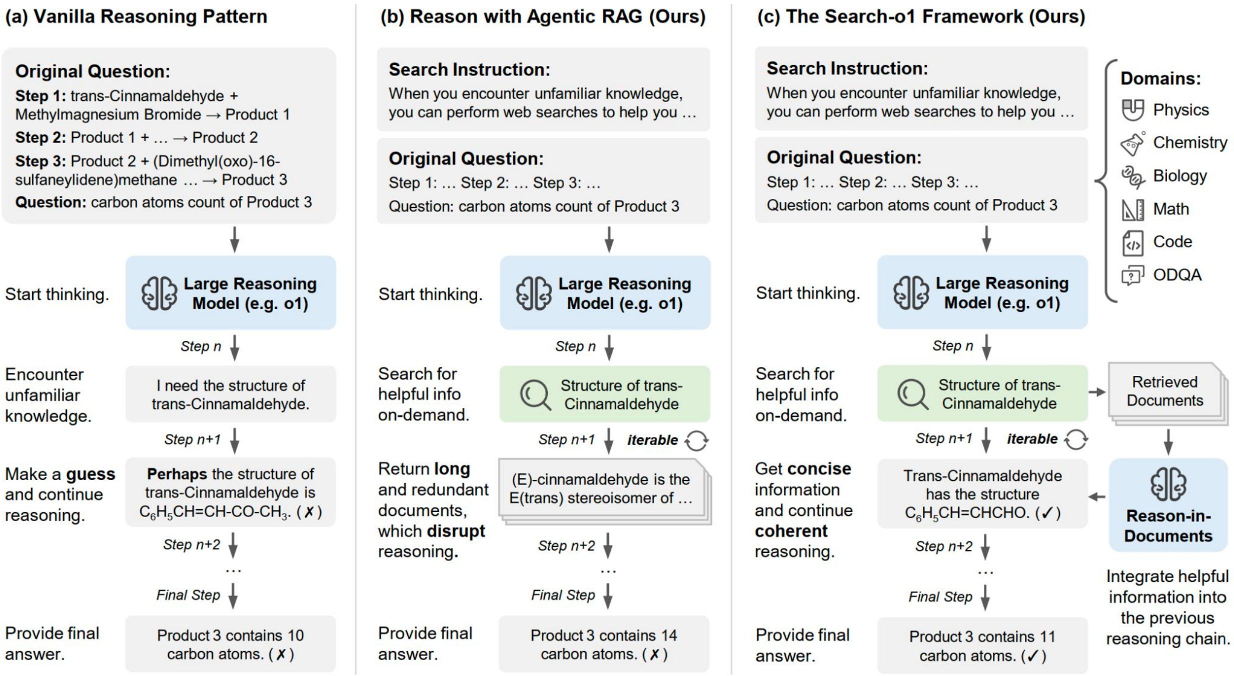
Agentic Conversational Search

- Tool use
  - Search engines are a key tool
  - Recent work explores how LLMs act as agents that autonomously use search engines to meet users' information needs [1,2,3]

[1] Li et al. Search-o1: Agentic Search-Enhanced Large Reasoning Models. arXiv 2025.
[2] Jin et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. COLM 2025.
[3] Song et al. R1-Searcher: Incentivizing the Search Capability in LLMs via Reinforcement Learning. arXiv 2025.

Agentic Conversational Search

- Tool use
  - [1] proposes Agentic RAG and Search-o1, purely based on prompting

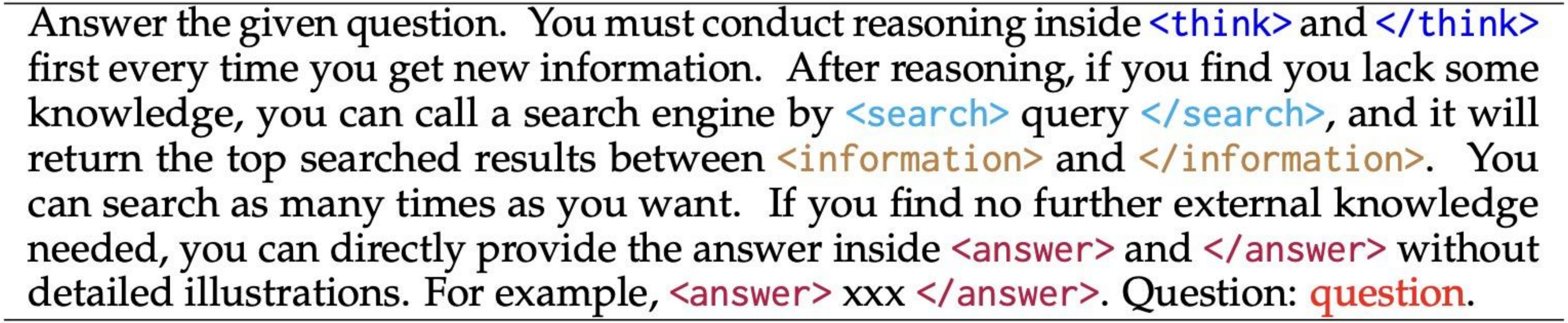


(a) Vanilla Reasoning Pattern
Original Question:
Step 1: trans-Cinnamaldehyde + Methylmagnesium Bromide → Product 1
Step 2: Product 1 + … → Product 2
Step 3: Product 2 + (Dimethyl(oxo)-16-sulfaneylidene)methane … → Product 3
Question: carbon atoms count of Product 3
Start thinking.
Large Reasoning Model (e.g. o1)
Step n
Encounter unfamiliar knowledge.
I need the structure of trans-Cinnamaldehyde.
Step n+1
Make a **guess** and continue reasoning.
Perhaps the structure of trans-Cinnamaldehyde is $C_6H_5CH=CH\text{-}CO\text{-}CH_3$. (✗)
Step n+2
…
Final Step
Provide final answer.
Product 3 contains 10 carbon atoms. (✗)
(b) Reason with Agentic RAG (Ours)
Search Instruction:
When you encounter unfamiliar knowledge, you can perform web searches to help you …
Original Question:
Step 1: … Step 2: … Step 3: …
Question: carbon atoms count of Product 3
Start thinking.
Large Reasoning Model (e.g. o1)
Step n
Search for helpful info on-demand.
Structure of trans-Cinnamaldehyde
Step n+1
iterable
Return **long** and redundant documents, which **disrupt** reasoning.
(E)-cinnamaldehyde is the E(trans) stereoisomer of …
Step n+2
…
Final Step
Provide final answer.
Product 3 contains 14 carbon atoms. (✗)
(c) The Search-o1 Framework (Ours)
Search Instruction:
When you encounter unfamiliar knowledge, you can perform web searches to help you …
Original Question:
Step 1: … Step 2: … Step 3: …
Question: carbon atoms count of Product 3
Domains:
Physics
Chemistry
Biology
Math
Code
ODQA
Start thinking.
Large Reasoning Model (e.g. o1)
Step n
Search for helpful info on-demand.
Structure of trans-Cinnamaldehyde
Retrieved Documents
Step n+1
iterable
Reason-in-Documents
Get **concise** information and continue **coherent** reasoning.
Trans-Cinnamaldehyde has the structure $C_6H_5CH=CHCHO$. (✓)
Step n+2
…
Final Step
Integrate helpful information into the previous reasoning chain.
Provide final answer.
Product 3 contains 11 carbon atoms. (✓)

[1] Li et al. Search-o1: Agentic Search-Enhanced Large Reasoning Models. arXiv 2025.

Agentic Conversational Search

- Tool use
  - [1,2] extend this line of work by applying reinforcement learning to teach LLMs how to effectively use search engines during multi-step reasoning



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.
[2] Song et al. R1-Searcher: Incentivizing the Search Capability in LLMs via Reinforcement Learning. arXiv 2025.

Agentic Conversational Search

- Tool use
  - Future direction: go beyond search engines
    - Use tools to handle broader user needs
      - E.g., for the query “What is the capital of Scotland, and what’s the current weather?”, combine search engines with a weather forecast API

Q & A

Part III: Summary 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
  - Mixed-initiative interactions
  - Personalized conversational search
- We explored emerging topics in the era of LLMs:
  - Conversational RAG
  - Automatic evaluation using LLM judges
  - Agentic conversational search

Conclusions and future directions

- Future directions
  - Agentic related
    - Enhancing reasoning capabilities
    - Reflection and self-correction
    - Tool use beyond traditional document retrieval
  - Broader Applicability
    - Multilingual and Multimodal scenarios
    - Domain-specific scenarios (financial, legal, medical, etc.)
    - Search as an intermediate step in complex tasks (QA, assistance, …)
  - Evaluation

Thank you!

Q & A