SIGIR 2025
Padova
ITALY

Conversational Query Understanding Evaluation

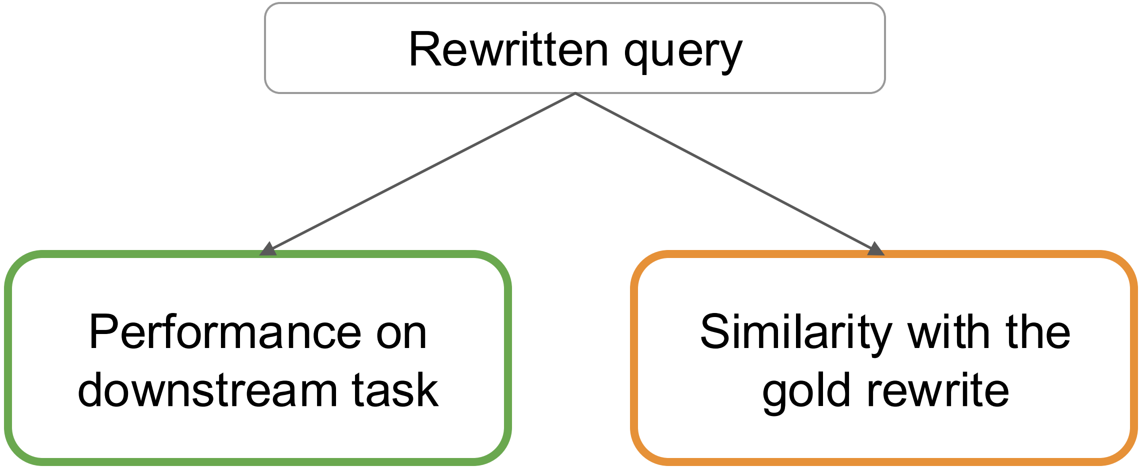
Zahra Abbasiantaeb
University of Amsterdam

Agenda

- Evaluation
  - **Evaluation paradigms**
  - Datasets
- LLM-based Relevance assessment
  - Document relevance assessment
  - Response generation assessment

Query Understanding Evaluation

Two different paradigms



Similarity with Gold Response

Assess how similar the generated query is to a human-written rewrite

- Compare the model-generated query with the human-written rewrite
    - Text similarity metrics: BLEU, ROUGE, Exact Match.
    - Machine translation metrics: METEOR

Lin, Sheng-Chieh, et al. "Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting." *ACM TOIS* (2021).

Similarity with Gold Response

Assess how similar the generated query is to a human-written reformulation.

Limitations:

- There are multiple valid ways to rewrite a query.
- Relying on a single gold reference can penalize correct but diverse reformulations.
    - This may misrepresent the model's actual effectiveness.

Lin, Sheng-Chieh, et al. "Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting." *ACM TOIS* (2021).

Similarity with Gold Response

Assess how similar the generated query is to a human-written reformulation.

Examples:

- Different wording:
    - “Restaurants near me open now”, “places to eat nearby that are currently open”
- Paraphrase or synonym:
    - “cheap hotels in San Francisco”, “affordable accommodations in SF”
- Minor changes:
    - “how to reset iPhone”, “how do I reset my iPhone”

Lin, Sheng-Chieh, et al. "Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting." *ACM TOIS* (2021).

End-to-end Evaluation

Assess how effective the generated query is in improving performance on the downstream task.

- Use the rewritten query as input to the downstream task
- Measure the performance of the downstream task using task specific metrics

End-to-end Evaluation

Assess how effective the generated query is in improving performance on the downstream task.

Downstream task:

- Passage Retrieval (PR)
    - Metrics: MRR, MAP, Precision@K, Recall@k.
- Question Answering (QA)
    - Exact match, F1 score.

Agenda

- Evaluation
  - Evaluation paradigms
  - **Datasets**
- LLM-based Relevance assessment
  - Document relevance assessment
  - Response generation assessment

QReCC

- Large-scale dataset of multi-turn question answering
- Manually collected conversations
    - Question, answer, and rewrite by the same annotator
- Answers are spans from the text
- Seed topics from Natural Questions (NQ), CAsT 2019, QuAC
- Includes unanswerable questions to simulate real-world scenario
    - 9% of the questions
- Average of 6 questions per dialogue
- Supports three tasks:
    - Query rewrite (QR), passage retrieval (PR), reading comprehension (RC)

Raviteja Anantha et al. “Open-Domain Question Answering Goes Conversational via Question Rewriting”. In NAACL (2021), ACL.

Query Rewriting Types in QReCC

- **Insertion**: Adding missing context
  - "What are some of the main types" -> "What are some of the main types **of yoga**"
- **Removal**: Eliminating redundant context
  - "Can you tell me about the C++ language **mentioned**" -> "Can you tell me about the C++ language"
- **Replacement**: swap vague reference
  - "Does **it** help in reducing stress" -> "Does **Yoga** help in reducing stress"
- **Copy**: No rewrite needed
  - "What are common poses in Kundalini Yoga" -> "What are common poses in Kundalini Yoga"

Raviteja Anantha et al. "Open-Domain Question Answering Goes Conversational via Question Rewriting". In NAACL (2021), ACL.

Example Question from QReCC


```
{
  "Context": [
    "What are the pros and cons of electric cars?",
    "Some pros are: They're easier on the environment. Electricity is cheaper than gasoline.
  ],
  "Question": "Tell me more about Tesla",
  "Rewrite": "Tell me more about Tesla the car company.",
  "Answer": "Tesla Inc. is an American automotive and energy company based in Palo Alto, Cal
  "Answer_URL": "https://en.wikipedia.org/wiki/Tesla,_Inc.",
  "Conversation_no": 74,
  "Turn_no": 2,
  "Conversation_source": "trec"
}
```

Datasets

Dataset	# Dialogues	# Questions	Task	No-Answer Questions	PR Collection	Training set
QReCC	13.7 K	81 K	PR, QR, RC	Yes (9%)	Common Crawl & Wayback (54 M passages)	Yes

PR: Passage Retrieval QR: Query Rewriting RC: Reading Comprehension

TopioCQA

- Large scale dataset for information seeking
- Conversations are collected by two annotators as questioner and answerer
- Seed topics from Natural Questions (NQ) dataset
- Answers are in free-form text
  - Not spans from the text
- Answers have rationals
  - Rational is a span of the text
- Topic shifts per each conversation
  - Average of 4 wikipedia pages per each conversation
- No manual query rewrite

Adlakha, Vaibhav, et al. "Topiocqa: Open-domain conversational question answering with topic switching." *TACL* (2022): 468-483.

Example Question from TopioCQA


```
{
    "Context": [
        "when will the new dunkirk film be released on dvd",
        "18 December 2017",
        "what is this film about?",
        "Dunkirk evacuation of World War II"
    ],
    "Conversation_no": 1,
    "Turn_no": 3,
    "Question": "can you mention a few members of the cast?",
    "is_nq": false,
    "Answer": "Fionn Whitehead, Tom Glynn-Carney, Jack Lowden, Harry Styles",
    "Topic": "Dunkirk (2017 film)",
    "Topic_section": "Introduction",
    "Rationale": "Its ensemble cast includes Fionn Whitehead, Tom Glynn-Carney, Jack Lowden, Harry Styles,
    "Additional_answers": [
        {
            "Answer": "Fionn Whitehead, Tom Glynn-Carney, Jack Lowden",
            "Topic": "Dunkirk (2017 film)",
            "Topic_section": "Introduction",
            "Rationale": "Its ensemble cast includes Fionn Whitehead, Tom Glynn-Carney, Jack Lowden, Harry
        },
```

Datasets

Dataset	# Dialogues	# Questions	Task	No-Answer Questions	PR Collection	Training set
QReCC	13.7 K	81 K	PR, QR, RC	Yes (9%)	Common Crawl & Wayback (54 M passages)	Yes
TopioCQA	3,920	50,574	PR, QA	Yes	Wikipedia dump, (5.9 M passages)	Yes

PR: Passage Retrieval QR: Query Rewriting RC: Reading Comprehension QA: Question Answering

CANARD

- Human written rewrites of QuAC dataset
  - Entire development set and a sample of train set

- QuAC is a conversational QA dataset
  - Conversations are collected by two humans
  - Student and teacher
  - Wikipedia pages
  - Reading comprehension dataset

Elgohary, Ahmed et al. "Can You Unpack That? Learning to Rewrite Questions-in-Context." *EMNLP-IJCNLP*. 2019.
Eunsol, Choi et al. “QuAC: Question Answering in Context”. EMNLP. 2018.

Example Question from CANARD


```
"History": [
              "Ara Parseghian",
              "First national title",
              "When did ara parseghian win his first title.",
              "In 1966,"
            ],
"Question": "what was their record for that year?",
"Rewrite": "what was Ara Parseghian's record for 1966?",
"QuAC_dialog_id": "C_4ae4e1bbf2534dd18304f05d7f88a440_0",
"Question_no": 2
```

Datasets

Dataset	# Dialogues	# Questions	Task	No-Answer Questions	PR Collection	Training set
QReCC	13.7 K	81 K	PR, QR, RC	Yes (9%)	Common Crawl & Wayback (54 M passages)	Yes
TopioCQA	3,920	50,574	QA, PR	Yes	Wikipedia dump, (5.9 M passages)	Yes
CANARD	-	40,527	QR	-	-	Yes

PR: Passage Retrieval QR: Query Rewriting RC: Reading Comprehension

TREC CAsT

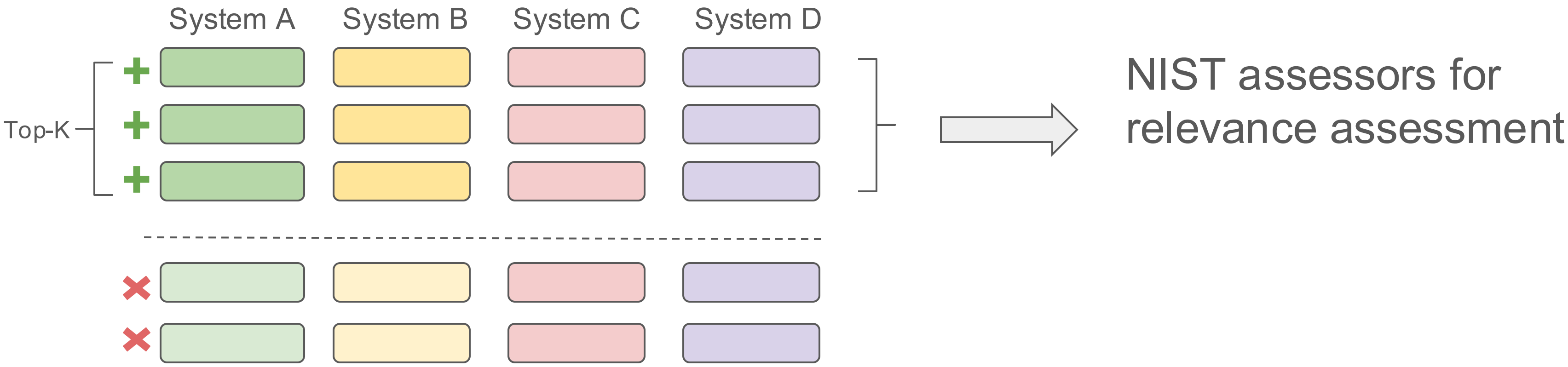
- Conversational search dataset
  - Ran each year from 2019 - 2022
- Small-size dataset with only test set
- Mixed-initiative conversations
  - Both the system and user can take initiative
- Types of system-initiative turns:
  - Clarifying question : ‘**Are you using Latex or Word?**’
  - Preference elicitation: ‘**What movie genre are you interested in?**’
  - Feedback: **‘Here are several options ….. What do you think about the first movie?’**

Dalton, Jeffrey et al. "TREC CAsT 2019: The conversational assistance track overview." *TREC*. 2019.
Dalton, Jeffrey et al. "Cast 2020: The conversational assistance track overview." *TREC*. 2021.
Owoicho, Paul, et al. "TREC CAsT 2022: Going Beyond User Ask and System Retrieve with Initiative and Response Generation." *TREC*. 2022.

TREC CAsT

TREC style pooling and relevance assessment

- Relevance with scale of 5 (0-4)



TREC CAsT

Dataset	2019	2020	2021	2022
Canonical Answer	No	Yes	Yes	Yes
Initiative	Single	Single	Single	Mixed
Manual Rewrite	No	Yes	Yes	Yes
System Turn Types	Inform	Inform	Inform	Inform, clarification, elicitation,feedback
User Turn Types	Question	Question	Question, revealment, feedback	Question, revealment, feedback
Tasks	PR	PR	PR	PR, RG, MI

PR: Passage Retrieval RG: Response Generation MI: Mixed-initiative

TREC CAsT

Dataset	# Dialogues	# Turns	Pool size	Collection	Collection size
CAsT 2019	20	173	29,571	MSMARCO, TREC CAR (Wikipedia) V2.0	~ 38 M
CAsT 2020	25	216	40,451		
CAsT 2021	26	239	19,334	MSMARCO (V1.0), WAPO (V4.0), Wikipedia (KLIT)	9.2 M
CAsT 2022	17	163	43,027	MSMARCO (V2.0), WAPO (V4.0), Wikipedia (KLIT)	17 M

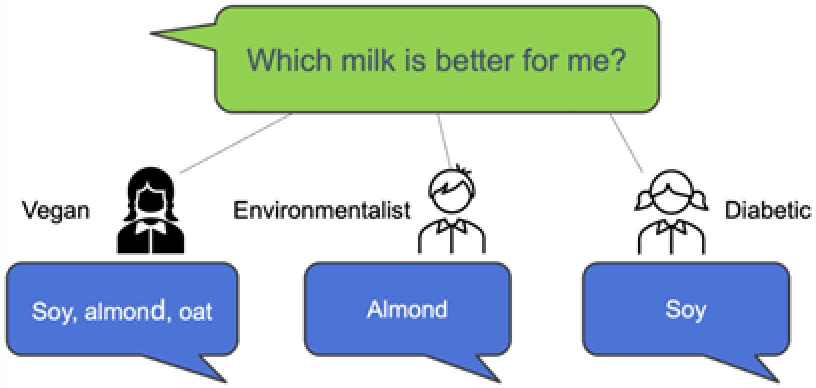
TREC CAsT

Dataset	# Dialogues	# Turns	[illegible]	[illegible]	Collection size
CAsT 2019	20	173	[illegible]	…), …CAR (…kipedia) V2.0	~ 38 M
CAsT 2020	25	216	40…		
CAsT 2021	26	239	19,334	MSMARCO (V1.0), WAPO (V4.0), Wikipedia (KLIT)	9.2 M
CAsT 2022	17	163	43,027	MSMARCO (V2.0), WAPO (V4.0), Wikipedia (KLIT)	17 M

Relevance assessment is done at the level of documents rather than passages

TREC iKAT

- Next generation of CAsT
- Personalized conversational search benchmark
  - TREC iKAT 2023 and 2024 benchmarks are released
- Covers three tasks
  - Passage retrieval (PR), PTKB classification, response generation (RG)
- More complex
  - Compare different items
  - Mixed-initiative turns
  - Topic shifts



Aliannejadi, Mohammad, et al. "Trec iKAT 2023: A test collection for evaluating conversational and interactive knowledge assistants." *SIGIR*. 2024.

TREC iKAT

Persona of the user is shown with Personalized Text Knowledge Base (PTKB)

- Collected from previous interactions of the user with system
- Is available at the beginning of the conversation

Persona_1:{

1. I'm 26 years old;
2. I have bachelor degree of computer science from Tilburg university;
3. I liked these courses during my bachelor's: data structure, algorithm, data mining, and artificial intelligence;
4. I didn't like computer architecture and logical circuits courses;
5. I live in the Netherlands;
6. I worked as a web developer for 2 years;
7. My bachelor GPA is 5.6;
8. My TOEFL score is 91.}

TREC iKAT

The collection includes:

- Static PTKB for the user in each conversation
    - Dynamic PTKB for iKAT 2025
- Static predefined conversation trajectories
- Canonical grounded answers
    - Gold response for iKAT 2024
- Fixed document collection and indices
    - Subset of ClueWeb 22-B, 17M passages
- Nuggets of information for iKAT 2024
    - CONE-RAG: a nugget-based pipeline for evaluation of answer generation

Abbasiantaeb, Zahra, et al. "Conversational Gold: Evaluating Personalized Conversational Search System using Gold Nuggets." SIGIR (2025)

TREC iKAT

Dataset	# Dialogues	# Turns / assessed	Pool size	Collection
iKAT 2023	25	326 / 176	26,159	Subset of ClueWeb 22-B with ~ 17M passages
iKAT 2024	17	218 / 116	20,575	

Agenda

- Evaluation
  - Evaluation paradigms
  - Datasets
- LLM-based Relevance assessment
  - **Document relevance assessment**
  - Response generation assessment

Motivation

Judgement holes are unassessed documents in benchmarks

- Judgement holes are considered as irrelevant
- More judgement holes can make the existing benchmarks **less reliable**

The new systems that did not contribute to the pooling are in disadvantage

- Can retrieve new relevant documents

Motivation

Why should we use LLMs for relevance assessment?

- Cheap
- Fast
- Consistent
- Accessible
- Scalable

- Blackbox
- Unknown biases
- Hallucination
- Lack of diversity
- Models evaluate models

Motivation

Existing research has shown High rank correlation between using judgments by LLMs and Human on **ad-hoc search**

Relevance assessment is more challenging in conversational search (CS)

- Relevance depends on the **user query**, **previous responses**, and **personal preferences**
- The information need can be addressed differently, based on the interpretation of the system

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large Language Models can Accurately Predict Searcher Preferences.
Sean MacAvaney and Luca Soldaini. 2023. One-Shot Labeling for Automatic Relevance Estimation.

Reusability of CS Benchmarks

Hypothesis: As conversation evolves, the systems retrieve a much larger and diverse array of documents

- Leads to having bigger holes
- Limits the reusability

Simulate the case of having a new system

Calculate the reusability with two metrics:

- ϕ : Number of holes for the new system
- $\phi+$: Number of relevant holes for the new system

Abbasiantaeb, Z. et al. Improving the Reusability of Conversational Search Test Collections. ECIR 2025.

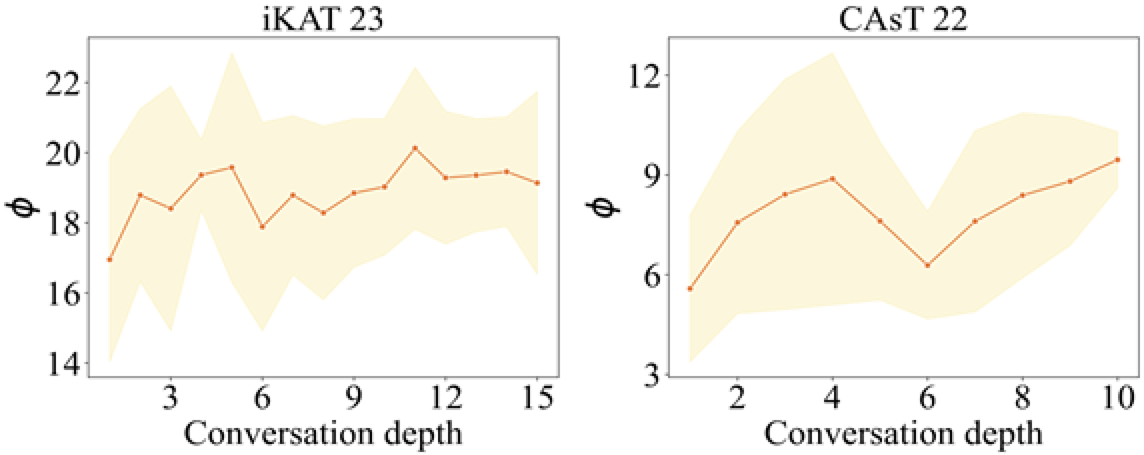
Reusability of CS Benchmarks

The pool becomes more diverse by increasing depth

- Average of ϕ is **18.55** for iKAT and **7.61** for CAsT

Factors influencing the reusability

- Depth of pooling
- Number and diversity of systems
- Complexity of user queries

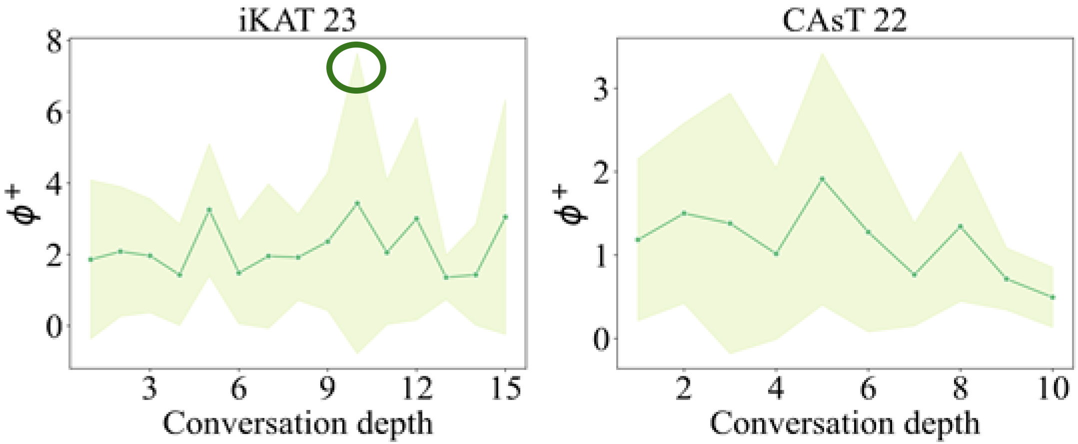


Abbasiantaeb, Z. et al. Improving the Reusability of Conversational Search Test Collections. ECIR 2025.

Reusability of CS Benchmarks

Average of ϕ+ is **decreasing**

- Leads to an unfair assessment
- Average and Std. Dev. of ϕ+ for iKAT is higher than CAsT
- There is a team with 8 missing relevant judgments in iKAT



Abbasiantaeb, Z. et al. Improving the Reusability of Conversational Search Test Collections. ECIR 2025.

Agreement of LLMs with Human

Use different LLMs to do relevance assessment

Compute the agreement with human

Findings:

- Fine-tuned LLMs have high agreement
- Few-shot GPT model has a very low agreement
    - GPT tends to give higher scores compared to human
- One-shot prompting has higher agreement than two- and zero-shot prompting

Abbasiantaeb, Z. et al. Improving the Reusability of Conversational Search Test Collections. ECIR 2025.

Rank Correlation using LLM and Human Judgment

Use judgments by LLM and human to rank systems

Compute the correlation between rankings

Findings:

- Fine-tuned Flan-T5 has the highest rank correlation compared to Llama and GPT
- Gap between rank correlation using GPT and fine-tuned Llama is smaller
    - However GPT tends to assign higher scores,
    - Relative ranking of scores by GPT are more consistent with human

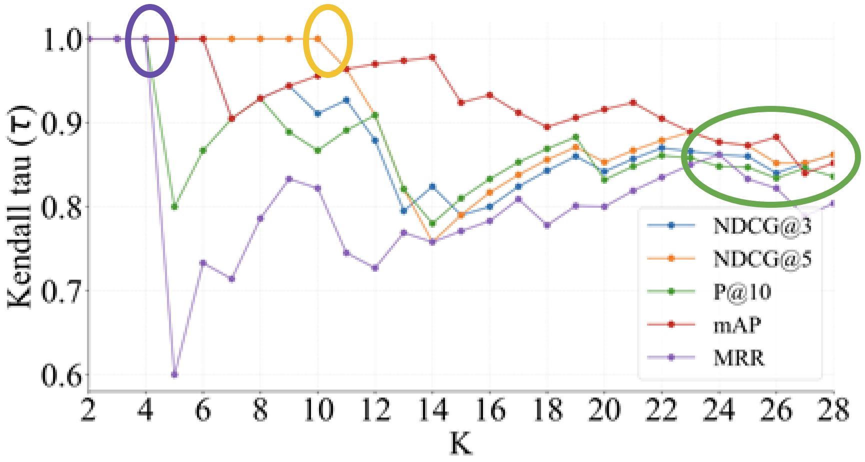
Abbasiantaeb, Z. et al. Improving the Reusability of Conversational Search Test Collections. ECIR 2025.

Rank Correlation using LLM and Human Judgment

Use the judgments by few-shot GPT

Compute rank correlation considering top K systems

- Top 4 models are ranked correctly -> all metrics
- Top 10 models are ranked correctly -> NDCG@5
- By increasing the K, the rank correlation converges
    - Reliability of LLM judgments for evaluation

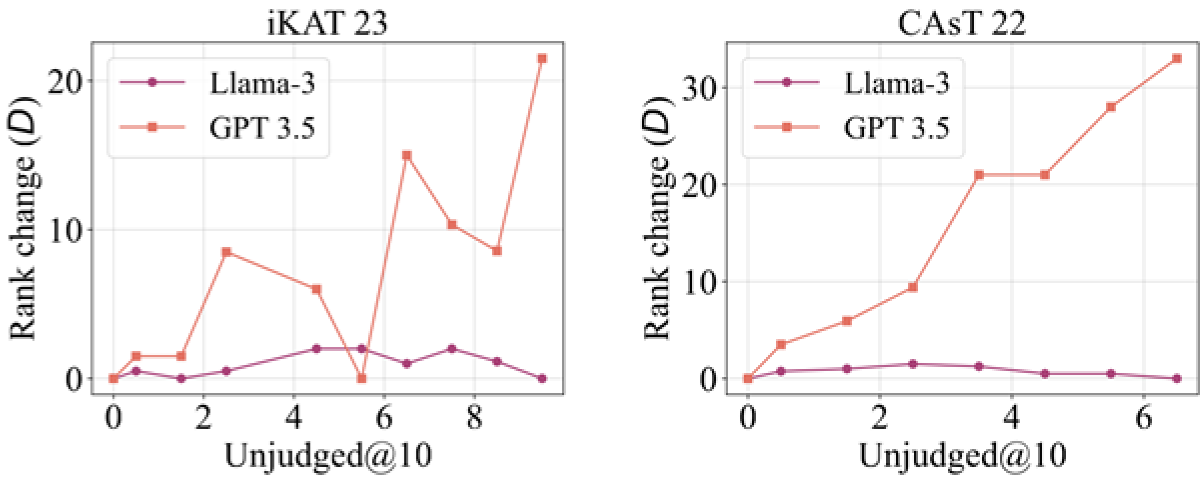


Abbasiantaeb, Z. et al. Improving the Reusability of Conversational Search Test Collections. ECIR 2025.

Filling Judgement Holes with LLM

Simulate the case of having a new system

- Fill the holes using few-shot LLMs
- Few-shot GPT: ranks the new system far from the the original location
- Few-shot Llama: ranks the new system closer to the original location

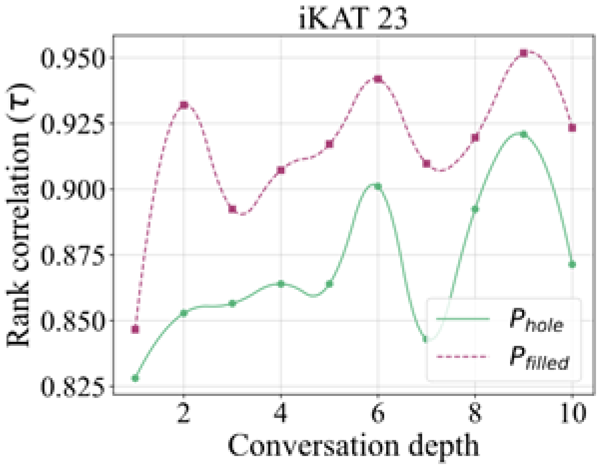


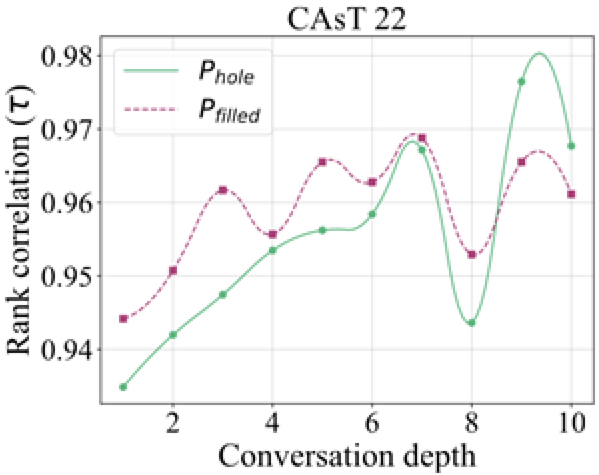
Abbasiantaeb, Z. et al. Improving the Reusability of Conversational Search Test Collections. ECIR 2025.

LLMs for Reusability

Fill the judgement holes using few-shot Llama

- Compare with leaving the holes unassessed
- Hole filling with Llama increases rank correlation over different depth



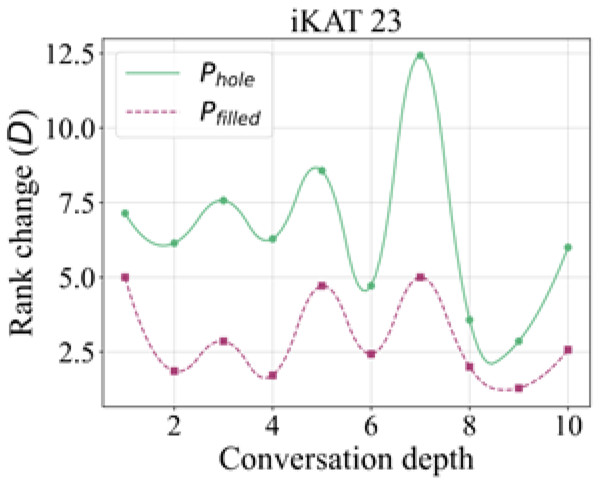


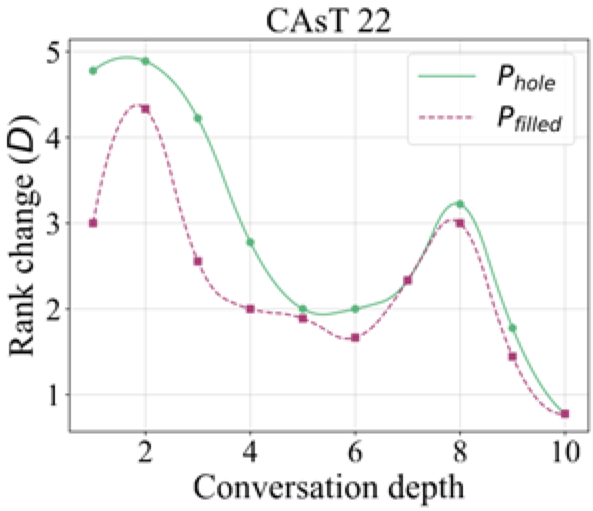
Abbasiantaeb, Z. et al. Improving the Reusability of Conversational Search Test Collections. ECIR 2025.

LLMs for Reusability

Fill the judgement holes using few-shot Llama

- Compare with leaving the judgement holes unassessed
- Rank the missing system closer to the original location





Abbasiantaeb, Z. et al. Improving the Reusability of Conversational Search Test Collections. ECIR 2025.

Summary

- Fine-tuning LLMs can lead to higher agreement with human
  - Lower agreement does not always result in lower rank correlation
- For assessing new systems with large holes either
  - Fill the holes by few-shot Llama
  - Reassess the complete pool using few-shot GPT
- CS test collections show a trend toward less reusability per deeper turns
  - LLM judgments can be used to enhance reusability

Abbasiantaeb, Z. et al. Improving the Reusability of Conversational Search Test Collections. ECIR 2025.

Agenda

- Evaluation
  - Evaluation paradigms
  - Datasets
- LLM-based Relevance assessment
  - Document relevance assessment
  - **Response generation assessment**

Evaluating Retrieval-Augmented Generation (RAG)

Evaluation of the RAG is challenging due to its complexity

- Both retrieved documents and LLM knowledge are used
- Information needs of the existing RAG collections are complex

Surface-based metrics are not ideal

- How complete is the generated response?
- How much is the generated response correct?

Nugget-based Response Evaluation

Nugget is a span of the text that carries essential pieces of information

Gold nuggets are collected by human from relevant documents

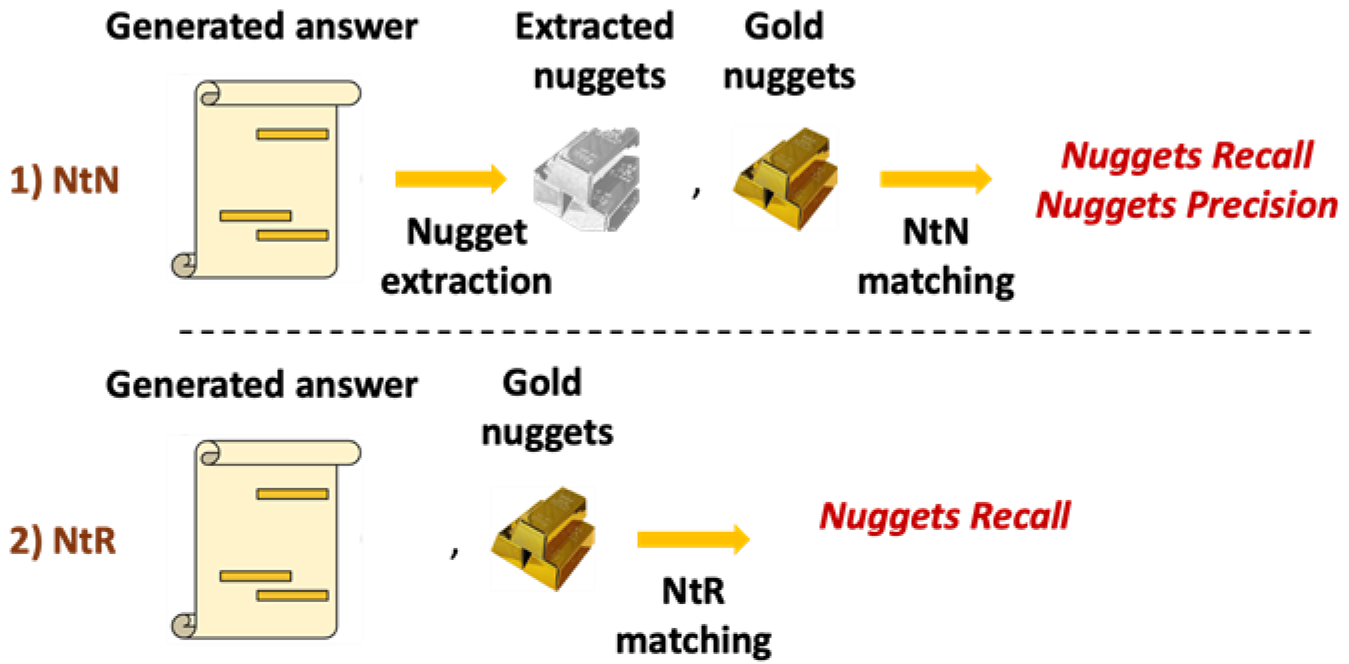
CONE-RAG:

- Extracting nuggets of the response by LLM
- Matching them with the gold nuggets
- Computing the nugget recall and precision

CONE-RAG is evaluate over iKAT 2024 benchmark

Abbasiantaeb, Zahra, et al. "Conversational Gold: Evaluating Personalized Conversational Search System using Gold Nuggets." SIGIR (2025)

Nugget-based Response Evaluation



Abbasiantaeb, Zahra, et al. "Conversational Gold: Evaluating Personalized Conversational Search System using Gold Nuggets." SIGIR (2025)

Nugget-based Response Evaluation

How does the **NtR** matching model compare to human judgments?

- Agreement between humans and LLM

Model	Accuracy	Cohen's κ
GPT4o	0.900	0.610
DeBERTa	0.805	0.247

Abbasiantaeb, Zahra, et al. "Conversational Gold: Evaluating Personalized Conversational Search System using Gold Nuggets." SIGIR (2025)

Nugget-based Response Evaluation

How do the **NtR** and **NtN** matching models compare?

- Correlation between ranking of responses using these methods

Metric	$\mathbb{N}_G$	$\text{Precision}_{\text{NtN}}$		$\text{Recall}_{\text{NtN}}$	
		τ	ρ	τ	ρ
$\text{Recall}_{\text{NtR}}$	Human	0.614	0.786	0.778	0.923
	LLM	0.626	0.781	0.778	0.925

Abbasiantaeb, Zahra, et al. "Conversational Gold: Evaluating Personalized Conversational Search System using Gold Nuggets." SIGIR (2025)

Nugget-based Response Evaluation

How does nugget extraction with LLM compare to human annotations?

- Correlation between ranking of responses using different gold nuggets

$\mathbb{N}_G$	$\mathbb{N}_G$	$\text{Precision}_{\text{NtN}}$		$\text{Recall}_{\text{NtN}}$	
		τ	ρ	τ	ρ
Human	LLM	0.649	0.814	0.731	0.889
Human [D]	LLM [D]	0.649	0.853	0.661	0.832

Abbasiantaeb, Zahra, et al. "Conversational Gold: Evaluating Personalized Conversational Search System using Gold Nuggets." SIGIR (2025)

Nugget-based Response Evaluation

How do nugget recall and precision metrics compare to other metrics?

- Negative correlation with Rouge and Groundedness
- Positive correlation between ndcg@5 and nugget precision and recall
    - Higher correlation using gold nuggets by LLM

Abbasiantaeb, Zahra, et al. "Conversational Gold: Evaluating Personalized Conversational Search System using Gold Nuggets." SIGIR (2025)

Thank you.