TREC iKAT 2025: A Test Collection for the Offline and Interactive Evaluation of Conversational Search

Zahra Abbasiantaeb*
University of Amsterdam
Amsterdam, The Netherlands
z.abbasiantaeb@uva.nl

Simon Lupart*
University of Amsterdam
Amsterdam, The Netherlands
s.c.lupart@uva.nl

Marcel Gohsen
Bauhaus-Universität Weimar
Weimar, Germany
marcel.gohsen@uni-weimar.de

Nailia Mirzakhmedova
Bauhaus-Universität Weimar
Weimar, Germany
nailia.mirzakhmedova@uni-weimar.de

Johannes Kiesel
GESIS - Leibniz Institute for the Social Sciences
Cologne, Germany
johannes.kiesel@gesis.org

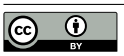
Jeffrey Dalton
University of Edinburgh
Edinburgh, Scotland, UK
jeff.dalton@ed.ac.uk

Mohammad Aliannejadi
University of Amsterdam
Amsterdam, The Netherlands
m.aliannejadi@uva.nl

Abstract

Conversational search agents, especially with the advent of large language models, have developed into useful tools to satisfy complex information needs of their users. Former research has shown that personalization (i.e., adaptation of agent responses to the preferences and traits of the user) can increase the relevance and perceived answer quality of these systems even further. However, developing accurate personalization methods typically requires rich datasets, both in terms of user profiles and complex conversations, for which only a few resources are publicly available. Over the past three years, the goal of the TREC Interactive Knowledge Assistance Track (iKAT) has been to bridge this gap. In organizing this shared task, we have developed a collection of complex information needs and associated conversations, made to challenge today's conversational agents and thus highlight aspects in need of further research. In this paper, we present the resources made for iKAT 2025, focusing on multi-session conversations (i.e., multiple dialogues per user), dynamically evolving user models, mixed-initiative dialogues, and large-scale human and automatic assessments. In addition to manually designed user profiles and conversations, the test collection for 2025 also contains dialogues between participating systems and our user simulators. All the resources are publicly available in our repository, including the system evaluation both as a result of the offline (i.e., test collection-based) and interactive tasks (i.e., user simulation-based), as well as their source code and model weights, to foster future research in this direction.[1]

*The authors contributed equally to this work.
[1]https://github.com/irlabamsterdam/ikat2025



This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGIR '26, Melbourne, VIC, Australia
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808591

CCS Concepts

• **Information systems** → **Test collections**.

Keywords

Conversational Information Seeking, Retrieval-Augmented Generation, Evaluation, Test Collection, Interactive Evaluation

ACM Reference Format:
Zahra Abbasiantaeb, Simon Lupart, Marcel Gohsen, Nailia Mirzakhmedova, Johannes Kiesel, Jeffrey Dalton, and Mohammad Aliannejadi. 2026. TREC iKAT 2025: A Test Collection for the Offline and Interactive Evaluation of Conversational Search. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3805712.3808591

1 Introduction

Over the past years, large language models (LLMs) have revolutionized the field of conversational information access, transforming heuristic conversational systems into intelligent agents that are capable of comprehending complex information needs and collaborating with their users [13, 22, 33, 53]. As a result, the conversational interaction paradigm has been adopted by the general population as a helpful mode of information retrieval [46].

Personalization (i.e., adapting search results based on a learned user profile) has been researched for decades in the domain of ad hoc search and has proven its potential to improve the relevance of search results for individual users [18, 45]. Due to this potential, personalization found its way into the field of conversational search, where similar positive effects have been observed [38]. However, developing personalization methods requires datasets that contain user profiles and associated conversations. These datasets are scarce since they also pose a privacy problem, as user profiles by their nature contain personal information.

Among the goals of the TREC Interactive Knowledge Assistance Track (iKAT) shared task [5] is the development of reusable test collections and resources for personalized conversational search.

Over three years, the iKAT organizers have assembled a set of curated conversational test collections [2, 3, 7] that cover aspects such as personalization, mixed initiative (e.g., clarifying questions to elicit information needs), longer conversational context, and complex information needs.

This resource paper introduces the latest addition to the pool of iKAT test collections, created in the context of the iKAT 2025 shared task [8]. The resource contains 17 manually designed long-context conversations on different topics with 188 turns. Each conversation is based on one of nine distinct user profiles modeled as Personal Text Knowledge Base (PTKB). For 45 user turns, the resource contains relevance assessments of passages (5,650 qrels) and persona statements, gold-level responses, gold-level information nuggets, grading rubric questions, and query rewrites that are manually created and assessed by the task organizers and by expert NIST assessors. This rich data is additionally augmented with automatic assessments to increase the size of the dataset. Moreover, this dataset contains all conversations that resulted from 26 participant runs in two subtasks: next response prediction for a given conversation (offline task) and responding to simulated users (interactive task). Conversations created during the interactive task contain assessments by NIST assessors on three turn-level and five conversation-level quality dimensions. Finally, our resource contains implementations of baselines and reproducibility assets, such as the used prompts for user simulations and automatic assessments.

The presented resource, as well as our analyses, provide insights into the intricacies of evaluation with Cranfield-style test collections compared to evaluation with user simulation for conversational search systems. This resource aims to facilitate investigations into the behavior of search systems under the two evaluation setups, both using automatic and human annotations. We believe that, besides benefiting research into personalization, this resource can contribute to the development of user simulation into a reproducible evaluation paradigm that becomes widely adopted in the information retrieval community [23].

2 Related Work

Interactive conversational search (CS) has long been an important direction in the field of information retrieval. Early efforts such as the TREC Interactive Track (1998–2002) [39] and the TREC Dynamic Domain Track (2015–2017) [49–51] focused on iterative passage retrieval with relevance feedback across multiple rounds. The TREC Conversational Assistance Track (CAsT) [15–17, 42] test collections target passage retrieval for CS, providing multi-turn topics together with a pooled set of manually assessed passages for evaluation. In four editions (2019–2022), CAsT progressively increases conversational dependency and realism. CAsT 2019 features fully scripted conversations with predefined turns, while CAsT 2020 introduces sequential dependence between user utterances and system responses. CAsT 2021 further strengthens contextual dependencies and adds richer interaction signals such as user feedback, and CAsT 2022 incorporates mixed-initiative behaviors and multi-path dialogues, enabling more realistic and flexible conversational flows. Building on CAsT, the TREC iKAT incorporated personalization into CS by modeling users with explicit personas [6]. Systems are required to reason over conversational history, user profiles, and external knowledge sources to address complex, multi-turn information needs. The iKAT 2023 collections provided resources for passage retrieval and response generation in personalized CS. The iKAT 2024 [2] collection introduced gold responses and nuggets as additional resources for response evaluation. Also, it proposed CONE-RAG, a nugget-based framework for calculating nugget recall and precision metrics. However, most existing benchmarks evaluate systems in a static manner, scoring individual responses independently rather than assessing behavior across an entire interaction. Such setups only partially reflect real-world usage, where systems must adapt to evolving user feedback and maintain coherence over multiple turns. To address this limitation, the TREC iKAT 2025 collection introduces an interactive evaluation framework.

In parallel to TREC, some works extended and built new conversational search datasets. QReCC [9] and TopiOCQA [4] are two conversational search datasets grounded in external collections. However, QReCC, built from web documents, is known to exhibit annotation bias toward BM25, which was used during construction. In contrast, TopiOCQA relies on Wikipedia and constructs conversations via hyperlink navigation, a design later followed by CORAL [12]; but typically provides a limited number of relevant passages per turn, limiting coverage of alternative evidence. More recent datasets extend specific aspects of the task: INSCIT [48] introduces mixed-initiative interactions; MultiDoc2Dial [21] focuses on multi-document reasoning in domain-specific settings but with a relatively small collection; and FaithDial [19] emphasizes faithfulness rather than information-seeking behavior. Recent work on multi-turn RAG has also proposed new datasets and evaluation settings [25, 44], including an ongoing SemEval shared task [43] that emphasizes unanswerable questions and automatic evaluation. Despite this progress, most existing resources rely on shallow relevance judgments and offline, pre-constructed conversations, limiting their ability to capture interactive search dynamics. Conversely, real-world conversational datasets derived from LLM logs [36, 54] provide richer interaction patterns but lack grounding to a fixed document collection.

3 TREC iKAT 2025 Test Collection

The primary objective of TREC iKAT is to advance the field of personalized conversational information access. A personalized conversational search system must leverage knowledge about the user to address their information needs during a conversation in order to provide more relevant results. TREC iKAT decomposes personalized conversational search into three distinct problems. First, for a given turn in a conversation and a user profile, the system has to identify knowledge from the profile relevant to the information need underlying the turn. For iKAT, each user profile is modeled as a set of natural language statements about the user ("Personal Text Knowledge Base", PTKB). Second, continuing from the first problem, the system has to retrieve passages from a given collection that are relevant to the information need and user profile. Third, continuing from the second problem, the system has to generate a complete, correct, and personalized text response based on the passages and user profile. To assess the system's ability to solve these tasks, a test collection requires at least conversations, user

profiles (as sets of PTKBs), and relevance assessments of individual PTKB statements and passages from a collection.

3.1 Conversation Collection Development

We compiled comprehensive guidelines for the manual creation process of topics, conversations, and PTKB statements. The guidelines include a detailed and step-by-step procedure for topic creation and a thorough explanation of the points that should be considered during the process. In addition, the guidelines included a checklist to ensure the quality of the topics. These criteria include quality assurance terms at the persona, turn, and conversation level. We used a mix of LLM-assisted development and personal exploration methodology for each topic and discussed them in regular meetings to improve the coherence of the narrative. The topics that did not meet the quality criteria were re-created by another curator. Each topic developed was reviewed and refined by at least two other experts. The process of creating a conversation for a topic was carried out by performing the following steps:

(1) Creating an imaginary user persona by defining their topic-specific PTKB (possibly extending an existing one).
(2) Formulating a user utterance in accordance with the topic and user profile.
(3) Identifying relevant PTKB statements.
(4) Employing GPT-based relevance assessments to estimate the number of relevant passages for each turn.
(5) Retrieving relevant passages using a rudimentary retrieval pipeline over our passage collection (iKAT searcher).
(6) Formulating the response of the system.
(7) Starting over from (2) until satisfied.

The PTKB statements were designed to contain information such as a user's interests (e.g., "I enjoy Stephen King's novels."), demographics (e.g., "I'm a software engineer."), or recent experiences (e.g., "I spent last summer in Brazil."). In generating the PTKB, we took great care to ensure that only high-level personal information was included. While some commonalities between expert and imaginary users were desirable to ensure some authenticity of conversations, PTKBs were checked not to include personally identifiable information.

The TREC iKAT 2025 dataset contains 17 conversations over distinct topics with a total number of 188 turns. Each turn includes a user utterance, a canonical system response, and a set of PTKB statements deemed relevant for the given query. The users in these conversations reflect nine different user profiles, where each of those contains an average number of 19.8 PTKB statements. Consequently, "users" participate in multiple conversations, of which the order is predetermined. Personal properties and preferences of the users that are revealed in a conversation and are not yet part of the PTKB are added to the profile of the user for a later conversation. These can become relevant in the following conversation to simulate dynamic user models in conversational systems. On average, user profiles were extended by 2.6 PTKB statements from their first conversation to their last in the dataset. Additional statistics on the presented resource are provided in Table 1.

Table 1: TREC iKAT collection statistics (2023-2025) for manually constructed conversational data.

	2025	2024	2023
Overall Statistics			
Total Dialogues	17	17	36
Total Topics	17	15	21
Total Turns	188	218	427
Avg. Turns per Dialogue	11	13	12
Personalization			
Unique Users	9	17	36
Total PTKB Statements	337	288	363
Total Relevant PTKBs	134	175	246
Avg. PTKBs per Dialogue	20	17	10
Avg. Relevant PTKBs per Dialogue	8	10	7
Gold Responses			
Total Responses	45	62	–
Avg. length (words)	163	22	–
Total "nuggets"	2 241	2 279	–
Avg. "nugget" length (words)	30	32	–

Table 2: Retrieved passage assessment statistics for participant and baseline runs in the TREC iKAT 2025 task performed by human assessors (NIST) and LLM (GPT-4.1).

	Human	GPT-4.1
Total Passages assessed	5 650	38 809
Fails to meet (0)	2 227	9 884
Slightly meets (1)	1 288	22 078
Moderately meets (2)	1 288	5 644
Highly meets (3)	606	883
Fully meets (4)	241	320

We selected a subset of 45 especially interesting turns from the developed conversations for relevance assessment by NIST assessors. Excluded turns comprise, among others, generic, clarification, elicitation, or feedback turns.

3.2 PTKB Relevance Assessments

The relevant PTKB statements for all 188 conversational turns are annotated by us as track organizers during the topic development phase. Each turn is annotated by one annotator and reviewed by at least one other annotator. Moreover, NIST assessors assessed the relevance of the full set of PTKB statements to the 45 selected turns (see above; relevance scale 0 to 4) to cross-check our own assessment.

3.3 Passage Relevance Assessments

To maintain reusability across all collections of iKAT, we base our relevance assessments on the same passage collection as iKAT 2023 and 2024. We continue using the collection originally introduced in iKAT 2023, which consists of a passage-segmented subset of the ClueWeb22 corpus [40] containing 16,838,987 passages. Researchers can obtain access to the full collection by signing the data-sharing agreement provided by CMU. In addition to the raw passages, we also release pre-built Pyserini [28] (BM25) and SPLADE indices for the collection to facilitate fast and effective retrieval experiments.

Dense indexing using ANCE-style representations was also examined in prior reproducibility work, with detailed commands and procedures on how to use it [34].

The pool of passages selected for human assessment by NIST assessors for the 45 selected turns (see above) is derived from submitted runs and baselines, with depths of up to 30. We use a dynamic pooling strategy where all top 10 of each run are manually assessed by NIST, and additional passages within the top 11–30 are manually assessed if an LLM-as-judge provided positive relevance (>1). This enables filtering out numerous irrelevant passages and allows for deeper pooling. The relevance of passages was assessed on a scale of 0–4, following the evaluation paradigm of TREC CAsT [15]. In total, 5,650 qrels assessments were created by NIST assessors. Additionally, our resource also includes LLM-based relevance judgments following the methodology and prompting approach of Abbasiantaeb et al. [3]. We assess the relevance of 38,809 query-passage pairs using GPT-4.1 as our base LLM. The distribution of relevance assessments across the five relevance levels can be seen in Table 2.

3.4 Gold Information Nuggets

Information nuggets are atomic facts for which an assessor can make a binary decision whether a text (e.g., a response) contains this fact [47]. For simplicity, we use passage spans as nuggets. We instruct the NIST assessors to select spans of relevant passages as nuggets of information. The NIST assessors extracted a total number of 2,241 nuggets of information from relevant passages for 45 turns. The average length of these gold nuggets is 30.4 words. This reservoir of information nuggets enables the calculation of metrics such as nugget recall and nugget precision [2].

We also extract nuggets from passages with relevance score higher than 0 by an LLM-based approach following Abbasiantaeb et al. [2]. We extract a total number of 8,607 information nuggets from 3,423 qrels over 45 turns with this automatic approach.

3.5 Gold Responses

Every turn in the developed conversations includes a user turn and grounded canonical response (cf. Section 3.1), and for the 45 selected turns the collection also contains one so-called gold response as reference for computing similarity-based evaluation metrics. The canonical responses are intended to mimic realistic systems for a realistic conversation flow and hence contain misunderstandings or irrelevant information that reflect the inherent ambiguity in web search. However, fair evaluation of systems must be based on optimal responses. Therefore, NIST assessors provided for each of the 45 selected turns additionally one "gold" response created from about 50 gold information nuggets that are on average extracted from the respective relevant passages (cf. Section 3.4).

3.6 Grading Rubric Questions

To enable rubric-based evaluations as in the iKAT 2025 interactive task, we include a sequence of manually derived rubric questions from the test conversations. A grading rubric according to the methodology of Farzi and Dietz [20] can be represented as a concise question for which a system response can be judged whether the question was answered or not. The goal of grading rubrics is to break down the complex problem of deciding whether a given piece of text is relevant to a query into asking whether individual subtopics of an information need are sufficiently addressed by a response. For example, if an information need is about how to start tattooing as a hobby, a reasonable grading rubric question would be about "What equipment is needed to start tattooing?". For the interactive task, these rubrics also guide the simulation process. Specifically, our simulator generates user utterances according to the current rubric and progresses to the next one once the previous rubric has been satisfied. In total, we manually derived 92 rubric questions from the 17 conversations in our dataset.

3.7 Participants' Runs and Baselines

TREC iKAT 2025 featured two subtasks in which participants were asked to provide next system utterances that continue conversations from the test collection (offline task), or to interact with online simulated users (interactive tasks) through an HTTP-API (Sim.API [24]). All the conversations, retrieved passages, and PTKB assessments of the participant systems and baselines across both subtasks are included in our resource to allow for reproducing our results and for further meta-analyses. Individual technical reports were also submitted by each participant with more details [7, 11, 26, 29, 31, 41, 52].

The offline task included automatic and generation-only run categories. In the automatic task, participants had to build complete conversational retrieval systems. In the generation-only runs, participants were provided with a ranked list of passages and just had to generate textual responses based on these passages. In the offline task, we received a total of 19 automatic runs, 9 generation-only runs, and 7 baseline added runs. The baselines use different retrieval models, including lexical (BM25), sparse (SPLADE), and dense (ANCE) retrieval models and include variants with and without personalization based on PTKB statements. This can be used to study the impact of personalization on the performance of the models. The runs submitted by participants cover a good variety of state-of-the-art and strong query rewriting models such as LLM4CS [35], MQ4CS [1], DiSCo [32], CHIQ-AD [37], and novel multi-stage retrieval pipelines. More details about the submissions and baselines are provided in the track overview paper [8].

In the interactive task, we received 16 runs with an additional seven baseline submissions which include different LLMs such as GPT-4.1-mini, or Llama 3 variants with 70 billion or 8 billion parameters for answer generation. Among the baselines, some of them directly use the LLM for response generation without grounding the response to retrieved passages. Across all participant submissions, the systems combine LLMs with retrieval-augmented and agentic pipelines in different ways. LLMs are widely used for query rewriting and expansion, including full or summarized history rewrites, multi-query generation, and extraction of additional knowledge from the conversation. For retrieval, both lexical (BM25) and neural sparse methods (SPLADE) are used, sometimes fused or supported by specialized search tools. Reranking is performed with neural cross-encoders and trained or ensemble rerankers. Most runs use modern LLMs for answer generation over retrieved passages, while some adopt agentic setups that dynamically decide when and how to search, and a few rely on LLMs alone without retrieval.

3.8 Conversation Quality Assessments

Conversations that were submitted to the interactive task were manually assessed by NIST assessors based on various rubric- and conversation-level quality dimensions.

On a rubric-level (cf. Section 3.6), which can cover from one to three turns, the systems are assessed based on (R1) engagement, i.e., to what extent does the system encourage the user to engage with it?; (R2) relevance, i.e., given the conversation history and the user's PTKB, how would one assess the relevance and usefulness of the generated response to the user's request?; and (R3) overall quality, i.e., considering all factors (relevance, engagement, and other factors that were not covered by these questions), what is the overall quality and utility of the system's performance?.

In addition, the entire conversation is evaluated based on the following criteria: (C1) mixed-initiative strategies, i.e., how well does the system employ a portfolio of different proactive actions (e.g., asking clarifying questions, asking follow-up questions, asking for feedback, etc.) throughout the conversation?; (C2) personalization, i.e., how well does the system tailor the dialogue to the specific user persona provided?; (C3) information flow, i.e., how well does the system maintain a coherent and logical flow of information throughout the conversation?; (C4) trustworthiness, i.e., how likely would one be to believe that past and future answers of the system are factually correct?; and (C5) user satisfaction, i.e., considering the entire dialogue, how satisfied would one be as the user and how successful was the interaction?.

For each of these quality dimensions, thorough annotation guidelines were developed among the organizers. Each of the eight criteria is rated on a scale from 1 to 5, where a 5 indicates the highest level of performance for the given criterion. Additionally, NIST assessors rate their confidence in the given annotation by selecting one of three levels: low, medium, or high. Responses generated by participant teams are assessed by both NIST assessors and an LLM acting as a judge. We release the complete set of annotations and guidelines provided to the NIST assessors, as well as the prompts designed for using the LLM as a judge. Furthermore, we provide the scripts for automatic evaluation of the response quality as a resource to the community.

4 Results & Analyses

We report results for both the offline and interactive tasks to validate the proposed resources, show that they support meaningful evaluation, and document the current state-of-the-art. The results also provide a useful reference for the performance of existing models, serving as points for comparison for future methods.

4.1 Offline Task Evaluation

Passage Retrieval. We report the passage retrieval performance of participating runs and baselines using nDCG@5, Precision@20, Recall@20, Recall@1000, and mean Average Precision (mAP) and baselines in Table 3. We use the qrels provided by NIST assessors to calculate retrieval performance. The best-performing run based on nDCG@5 and recall achieves an nDCG@5 of 0.5843 and a recall of 0.7016. In contrast, the best run in iKAT 2024 achieved an nDCG@5 of 0.494 and a recall of 0.783 [2, 30]. This difference indicates that the iKAT 2025 test topics are more challenging in terms of the recall metric. One might argue that this discrepancy arises from differences in dynamic pooling strategies. As in iKAT 2024 dynamic pooling was applied to passages ranked between 6–30 while in iKAT 2025 it was applied to passages ranked between 11–30. However, the statistics show iKAT 2024 contains a total of 5,649 relevant passages across 116 turns (48.7 relevant passages per turn), whereas iKAT 2025 contains 2,135 relevant passages across 45 turns (47.5 relevant passages per turn). This shows that both benchmarks have nearly the same number of relevant passages per turn.

Table 3: Automatic evaluation of passage retrieval results. Evaluation is done at retrieval cutoff of 1000.

Run	nDCG@5	P@20	Recall@20	Recall	mAP
gril-o4	**0.5843**	**0.5867**	**0.2173**	0.5273	0.3041
gril-o2	0.5427	0.5711	0.2137	0.5245	0.2943
gril-o3	0.5375	0.5711	0.2130	0.5334	0.2953
gril-o1	0.5267	0.5656	0.2134	0.5412	0.2955
cfda-o1	0.5040	0.5522	0.2037	0.3721	0.2375
uva-o2	0.4897	0.5789	0.2073	0.6101	0.3099
ucsc-o4	0.4789	0.5656	0.2033	0.6648	0.3232
ucsc-o1	0.4721	0.5622	0.2019	**0.7016**	**0.3252**
ucsc-o3	0.4705	0.4544	0.1677	**0.7016**	0.2551
ucsc-o2	0.4392	0.4800	0.1758	0.6648	0.2592
cfda-o2	0.4326	0.5056	0.1778	0.3619	0.2133
cfda-o3	0.4160	0.4900	0.1786	0.3588	0.2046
uva-o3	0.4053	0.4756	0.1675	0.4481	0.2139
cfda-o4	0.4050	0.4900	0.1775	0.3546	0.2091
orga-o2	0.3902	0.4178	0.1498	0.4308	0.1871
orga-o1	0.3797	0.4378	0.1572	0.4307	0.1950
orga-o7	0.3783	0.4044	0.1427	0.5267	0.2187
uva-o1	0.3680	0.4756	0.1701	0.4413	0.2045
orga-o4	0.3676	0.4122	0.1470	0.6721	0.2320
orga-o3	0.3510	0.3989	0.1414	0.5258	0.1714
orga-o6	0.3498	0.4000	0.1545	0.6760	0.2150
uva-o4	0.3042	0.3544	0.1347	0.5524	0.1683
orga-o5	0.2988	0.3411	0.1336	0.3514	0.1492
guida-o2	0.2985	0.3167	0.1233	0.5185	0.1690
guida-o1	0.2765	0.3233	0.1225	0.5205	0.1549
guida-o3	0.2411	0.2922	0.0943	0.5140	0.1378

Response Evaluation. We automatically evaluate the quality of generated responses using gold responses and extracted gold nuggets (see Table 4). We use the gold response to calculate surface-level similarity metrics (ROUGE, F1) and semantic-level evaluation metrics, e.g. BEM (BertScore [10]) and LLMEval (LLM-as-Judge). For LLMEval we use GPT-4.1 and SOLAR-10.7B-Instruct-v1.0 [27] as backbone LLMs to judge how much the two answers are equivalent. In addition, the coverage of generated responses according to the human-extracted nuggets is assessed using the Nugget-to-Response (NtR) metric [2]. According to Tables 4 and 3, runs with higher retrieval effectiveness generally also produce more effective responses. Moreover, the runs with the highest nugget recall (i.e., `gril-ox`) are based on multiple rounds of query generation [41]. Also, the highest nugget recall achieved by participant runs is 0.25, indicating that generated responses cover, on average, only 25% of the gold nuggets. In contrast, the best-performing model in iKAT

Table 4: Automatic evaluation of response generation. Best is sorted according to Nugget-to-Response (NtR) recall.

Run	LLMeval		NtR	BEM	F1	ROUGE-1
	SOLAR	GPT-4.1				
Gold Response	-	-	**0.251**	-	-	-
gril-o1	0.911	0.777	**0.250**	0.175	0.285	0.210
gril-o2	**0.977**	0.800	0.241	0.171	0.288	0.213
gril-o3	0.955	0.800	0.220	0.188	0.283	0.207
ucsc-o1	0.909	0.822	0.190	0.168	**0.317**	0.252
ucsc-o2	0.888	0.755	0.187	0.171	0.313	0.255
ucsc-o3	0.888	0.822	0.186	**0.199**	0.312	0.253
ucsc-o4	0.933	**0.844**	0.178	0.183	0.316	0.253
gril-o4	0.955	0.822	0.176	0.164	0.287	0.212
uva-o1	0.911	0.777	0.151	0.151	0.303	0.256
uva-o2	0.911	**0.844**	0.149	0.182	0.313	0.258
uva-o3	0.844	0.822	0.142	0.161	0.298	0.250
uva-o4	0.866	0.822	0.130	0.146	0.306	0.259
cfda-o1	0.911	0.733	0.114	0.144	0.299	0.254
guida-o1	0.777	0.666	0.111	0.155	0.270	0.229
orga-o1	0.800	0.688	0.109	0.150	0.293	0.250
guida-o2	0.755	0.711	0.107	0.142	0.280	0.232
guida-o3	0.800	0.644	0.104	0.161	0.271	0.229
cfda-o2	0.955	0.800	0.101	0.162	0.292	0.248
cfda-o3	0.955	0.777	0.099	0.151	0.286	0.246
cfda-o4	0.911	0.822	0.097	0.147	0.290	0.247
orga-o2	0.772	0.622	0.092	0.184	0.311	**0.267**
orga-o3	0.822	0.622	0.086	0.170	0.296	0.258
orga-o4	0.733	0.666	0.083	0.149	0.301	0.265
orga-o5	0.822	0.622	0.078	0.181	0.296	0.264
orga-o6	0.755	0.555	0.061	0.156	0.283	0.242
orga-o7	0.777	0.688	0.057	0.152	0.299	0.262

2024 achieved a nugget recall of 0.395 [2]. The drop in recall suggests that the iKAT 2025 topics are more challenging, as systems capture less relevant information in their responses. Since the top-performing runs from iKAT 2024, along with other strong and newer systems, were also evaluated in 2025, the lower recall is unlikely due to weaker models. Instead, it indicates that the 2025 topics are inherently more complex and harder to cover comprehensively.

4.2 Interactive Evaluation.

Response Quality. We report response quality based on evaluations conducted by NIST assessors in Table 5. From the submitted runs, we first identify `orga-i1` that uses direct inference with GPT-4.1-mini without retrieval. As can be seen, GPT-4.1-mini achieves the highest performance on most rubric-level and conversation-level metrics, but at the cost of trustfulness, where it is below most other runs. From all the runs `gril-i2` is the one in comparison that achieves the highest trustfulness and uses RAG with passages retrieved from the collection. Some of the runs were also designed to compare the impact of personalization, (`uva-i3` vs `uva-i4` and `orga-i3` vs `orga-i7`). Comparing these runs shows the importance of using personal information, which improves the overall quality of the generated response in terms of most of the metrics.

Table 5: Evaluation of runs in the interactive task based on human assessments. On the rubric level, assessors evaluated engagement (R1, Eng.), relevance (R2, Rel.), and quality (R3, Qual.). On the dialogue level, assessors rated mixed-initiative strategies (C1, Mix.), personalization (C2, Pers.), information flow (C3, Flow), trustworthiness (C4, Trust), and user satisfaction (C5, Sat.).

Run	Rubric-level			Conversation-level				
	Eng.	Rel.	Qual.	Mix.	Pers.	Flow	Trust	Sat.
orga-i1	**4.28**	**4.40**	**4.41**	3.00	**4.59**	**4.88**	2.88	**4.29**
cfda-i1	3.60	4.28	4.37	2.53	2.94	4.65	3.88	4.24
genai-i1	3.88	4.24	4.32	3.00	3.71	4.71	3.88	**4.29**
genai-i2	3.91	4.23	**4.41**	3.12	3.71	4.65	4.06	4.12
cfda-i2	3.53	4.20	4.28	2.53	2.94	4.76	3.82	4.06
gril-i1	3.87	4.01	4.16	**4.00**	3.65	4.29	4.06	3.88
uva-i1	3.62	3.95	3.91	2.82	3.65	3.94	3.59	4.06
gril-i2	3.68	3.91	3.95	3.71	3.59	4.12	**4.29**	4.06
uva-i2	3.51	3.90	3.90	3.12	3.71	3.94	2.94	3.88
gril-i3	3.74	3.79	3.97	3.82	3.47	4.35	4.24	3.76
orga-i2	3.41	3.73	3.80	2.29	2.82	3.65	3.53	3.88
ucsc-i1	3.34	3.63	3.74	2.18	2.76	4.18	3.65	3.53
uva-i3	3.02	3.63	3.65	2.53	2.18	3.71	3.65	3.82
orga-i3	3.36	3.59	3.62	2.71	3.18	3.94	3.53	3.65
ucsc-i2	3.37	3.59	3.70	2.53	2.82	4.53	3.76	3.65
orga-i4	3.57	3.53	3.53	2.24	3.47	3.71	2.76	3.29
ucsc-i3	3.18	3.48	3.60	2.12	2.82	3.94	3.76	3.71
orga-i5	3.28	3.47	3.49	2.59	3.47	3.53	2.88	3.35
ucsc-i4	3.26	3.45	3.53	2.00	2.88	3.94	3.53	3.47
orga-i6	3.21	3.38	3.39	2.41	3.41	3.41	3.65	3.41
uva-i4	2.76	3.27	3.27	2.06	3.12	2.76	3.41	3.29
orga-i7	3.15	3.26	3.36	2.18	2.24	3.47	3.00	2.88
gril-i4	2.99	2.96	3.14	2.47	3.65	3.47	3.94	3.41

Table 6: Correlation of system ranking using LLM and NIST assessment for interactive evaluation on rubric-level (R*X*) and conversation-level critera (C*X*) introduced in Section 3.8. Correlation is computed by Spearman's (ρ) and Kendall's rank correlation (τ) coefficients. The best correlation per coefficient among the two assessor LLMs is shown bold.

Criterion	GPT4.1-mini		GPT5-mini	
	ρ	τ	ρ	τ
R1. Engagement	**0.495**	**0.368**	0.252	0.099
R2. Relevance	0.500	0.360	**0.580**	**0.431**
R3. Quality	**0.710**	**0.502**	0.349	0.249
C1. Mixed-initiative strategies	**0.686**	**0.534**	0.625	0.470
C2. Personalization	0.566	0.423	**0.609**	**0.447**
C3. Information flow	0.564	0.399	**0.717**	**0.510**
C4. Trustworthiness	**0.546**	**0.415**	0.418	0.304
C5. User satisfaction	**0.608**	**0.431**	0.426	0.336

LLMs vs NIST Assessors. We use the same guidelines provided to NIST assessors and design prompts for evaluating the quality of the generated responses and dialogues in the interactive task [2]. We use two different LLMs for our assessments and report the

[2]We provide the prompts and the script used for evaluation of responses by LLM in our public repository.

correlation between the ranking of systems using NIST and LLM assessments in Table 6. We use `GPT-4.1-mini` with `temperature=1` and `GPT-5-mini` with `verbosity=low` parameters. Overall, both LLMs show moderate to strong positive correlations with human judgments, indicating that LLM-based evaluation can reasonably approximate manual assessments. Among the two models, `GPT-4.1-mini` demonstrates stronger alignment on rubric-level overall quality metric (Spearman ρ = 0.710), and achieves consistently high correlations across most dialog-level metrics, including mixed-initiative, trustworthiness, and overall satisfaction. In contrast, `GPT-5-mini` shows higher correlations on relevance and certain dialog-level metrics such as personalization and information flow, but its agreement is less consistent across other dimensions. We provide the LLM-based evaluation scripts as part of the presented resource.

5 Conclusion

Our mission at iKAT has been, is, and will remain to establish resources and methods for the development of personalized, conversational search systems. In line with this objective, this paper presents the latest addition to the iKAT collection, which was developed for the TREC iKAT shared task in 2025. Similar to the iKAT 2023 and 2024 collections, iKAT 2025 includes the required resources for assessing passage retrieval performance and response generation quality using both surface- and nugget-based metrics. In addition, the new iKAT dataset contains conversations that are handcrafted by the organizers or result from baseline and participant runs in offline and interactive tasks addressing personalization with a focus on dynamic user profiles. The dataset includes various human and LLM-driven evaluations of these conversations including rubric and dialogue level annotations of response quality made by expert NIST assessors. Finally, we provided an analysis of the correlation between LLM-based and NIST judgments to assess the effectiveness of using LLMs for system evaluation.

Although iKAT will no longer be hosted at TREC, we will continue our mission for advancing the field of personalized conversational information access by hosting iKAT 2026 in the context of the 10th Workshop on Search-Oriented Conversational Artificial Intelligence (SCAI'26) [14] hosted at SIGIR. While the central goal remains the same, iKAT 2026 will focus on evaluating personalized conversational search systems with simulated users and aims to establish foundations for how these evaluations can be conducted.

Acknowledgments

This research was supported in part by the Informatics Institute (IvI), University of Amsterdam, and in part by the Swiss National Science Foundation (SNSF) grant 215742, entitled "Personality and Conversational Information Access (PACINO)."

The authors acknowledge the peoples of the Woi Wurrung and Boon Wurrung language groups of the eastern Kulin Nation on whose unceded lands ACM SIGIR 2026 was hosted. We pay our respects to their Elders past and present, and extend that respect to all Aboriginal and Torres Strait Islander peoples today and their continuing connection to land, sea, sky, and community.

The authors thank Ian Soborrof for his support in designing and running the human annotation tasks, as well as the NIST assessors for annotating the data. Finally, we thank the TREC iKAT 2025 participants for submitting their work and presenting them in the TREC conference.

References

[1] Zahra Abbasiantaeb, Simon Lupart, and Mohammad Aliannejadi. 2026. Generating Multi-Aspect Queries for Conversational Search. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2026 - Volume 1: Long Papers, Rabat, Morocco, March 24-29, 2026*. Association for Computational Linguistics, 8201–8217.

[2] Zahra Abbasiantaeb, Simon Lupart, Leif Azzopardi, Jeffrey Dalton, and Mohammad Aliannejadi. 2025. Conversational Gold: Evaluating Personalized Conversational Search System Using Gold Nuggets. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*. ACM, 3455–3465. doi:10.1145/3726302.3730316

[3] Zahra Abbasiantaeb, Chuan Meng, Leif Azzopardi, and Mohammad Aliannejadi. 2025. Improving the Reusability of Conversational Search Test Collections. In *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part I (Lecture Notes in Computer Science)*. Springer, 196–213. doi:10.1007/978-3-031-88708-6_13

[4] Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain Conversational Question Answering with Topic Switching. *Trans. Assoc. Comput. Linguistics* 10 (2022), 468–483. doi:10.1162/TACL_A_00471

[5] Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffery Dalton, and Leif Azzopardi. 2023. TREC iKAT 2023: The Interactive Knowledge Assistance Track Overview. (2023).

[6] Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffery Dalton, and Leif Azzopardi. 2023. TREC iKAT 2023: The Interactive Knowledge Assistance Track Overview. In *The Thirty-Second Text REtrieval Conference Proceedings (TREC 2023), Gaithersburg, MD, USA, November 14-17, 2023 (NIST Special Publication, Vol. 500-xxx)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec32/papers/Overview_ikat.pdf

[7] Mohammad Aliannejadi, Zahra Abbasiantaeb, Simon Lupart, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. 2024. TREC iKAT 2024: The Interactive Knowledge Assistance Track Overview. In *Proceedings of the Thirty-Third Text REtrieval Conference, TREC 2024, Gaithersburg, MD, USA, November 18-22, 2024 (NIST Special Publication)*. National Institute of Standards and Technology (NIST).

[8] Mohammad Aliannejadi, Simon Lupart, Marcel Gohsen, Zahra Abbasiantaeb, Nailia Mirzakhmedova, Johannes Kiesel, and Jeffrey Dalton. 2025. TREC iKAT 2025: The Interactive Knowledge Assistance Track Overview. In *Proceedings of the 34th Text REtrieval Conference (TREC 2025) (NIST SP 1348)*, Ian Soboroff and George Awad (Eds.). National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, USA, 15 pages. https://pages.nist.gov/trec-browser/trec34/ikat/proceedings/#trec-ikat-2025-the-interactive-knowledge-assistance-track-overview

[9] Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-Domain Question Answering Goes Conversational via Question Rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 520–534. doi:10.18653/v1/2021.naacl-main.44

[10] Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, Tomahto. Beyond Token-level Answer Equivalence for Question Answering Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*. Association for Computational Linguistics, 291–305. doi:10.18653/V1/2022.EMNLP-MAIN.20

[11] Yu-Cheng Chang, Guan-Wei Yeo, Quah Eugene, Fan-Jie Shih, Yuan-Ching Kuo, Tsung-En Yu, Hung-Chun Hsu, Ming-Feng Tsai, and Chuan-Ju Wang. 2025. CFDA & CLIP at TREC iKAT 2025: Enhancing Personalized Conversational Search via Query Reformulation and Rank Fusion. *CoRR* abs/2509.15588. arXiv:2509.15588 doi:10.48550/ARXIV.2509.15588

[12] Yiruo Cheng, Kelong Mao, Ziliang Zhao, Guanting Dong, Hongjin Qian, Yongkang Wu, Tetsuya Sakai, Ji-Rong Wen, and Zhicheng Dou. 2025. CORAL: Benchmarking Multi-turn Conversational Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 1308–1330. doi:10.18653/v1/2025.findings-naacl.72

[13] Sadia Sultana Chowa, Riasad Alvi, Subhey Sadi Rahman, Md. Abdur Rahman, Mohaimenul Azam Khan Raiaan, Md Rafiqul Islam, Mukhtar Hussain, and Sami Azam. 2026. From language to action: a review of large language models as autonomous agents and tool users. *Artif. Intell. Rev.* 59, 2 (2026), 71. doi:10.1007/S10462-025-11471-9

[14] Philipp Christmann, Roxana Petcu, Sneha Singhania, Mohammad Aliannejadi, Marcel Gohsen, and Svitlana Vakulenko. 2026. 10th Workshop on Search-Oriented Conversational Artificial Intelligence (SCAI 2026). In *49th International ACM*

SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26). ACM, New York. doi:10.1145/3805712.3808648

[15] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. CAsT 2019: The conversational assistance track overview. In *Text REtrieval Conference (TREC)*. National Institute of Standards and Technology (NIST).

[16] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2021. TREC CAsT 2021: The Conversational Assistance Track Overview. In *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021 (NIST Special Publication)*. National Institute of Standards and Technology (NIST).

[17] Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. CAsT-19: A Dataset for Conversational Information Seeking. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*. ACM, 1985–1988. doi:10.1145/3397271.3401206

[18] Zhicheng Dou, Ruihua Song, Ji-Rong Wen, and Xiaojie Yuan. 2009. Evaluating the Effectiveness of Personalized Web Search. *IEEE Trans. Knowl. Data Eng.* 21, 8 (2009), 1178–1190. doi:10.1109/TKDE.2008.172

[19] Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. FaithDial: A Faithful Benchmark for Information-Seeking Dialogue. *Transactions of the Association for Computational Linguistics* 10 (2022), 1473–1490. doi:10.1162/tacl_a_00529

[20] Naghmeh Farzi and Laura Dietz. 2024. Pencils Down! Automatic Rubric-based Evaluation of Retrieve/Generate Systems. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2024, Washington, DC, USA, 13 July 2024*. ACM, 175–184. doi:10.1145/3664190.3672511

[21] Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling Dialogues Grounded in Multiple Documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6162–6176. doi:10.18653/v1/2021.emnlp-main.498

[22] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–24.

[23] Marcel Gohsen, Zahra Abbasiantaeb, Mohammad Aliannejadi, Krisztian Balog, Timo Breuer, Jeffrey Dalton, Maik Fröbe, Christin Katharina Kreutz, Andreas Kruff, Simon Lupart, Nailia Mirzakhmedova, Harrisen Scells, Philipp Schaer, Benno Stein, and Johannes Kiesel. 2025. User Simulation in Practice: Lessons Learned from Three Shared Tasks. *SIGIR Forum* 59, 2, 1–18. doi:10.1145/3799914.3799917

[24] Marcel Gohsen, Nailia Mirzakhmedova, Zahra Abbasiantaeb, Johannes Kiesel, Simon Lupart, Jeffrey Dalton, Benno Stein, and Mohammad Aliannejadi. 2026. Sim.API: A Middleware to Simplify the Use of User Simulators for Shared Tasks in Conversational Search. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia*. ACM. doi:10.1145/3805712.3808599

[25] Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. 2025. mtRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems. *Trans. Assoc. Comput. Linguistics* 13 (2025), 784–808. doi:10.1162/TACL.A.19

[26] Ahmed Rayane Kebir, Victor Morand, Pierre-Antoine Lequeu, Zineddine Tighidet, Mitodru Niyogi, Jonah Turner, Rishu Kumar, and Benjamin Piwowarski. 2025. GUIDANCE@TREC iKAT 2025. In *Proceedings of the 34th Text REtrieval Conference (TREC 2025)*. Gaithersburg, Maryland.

[27] Sanghoon Kim, Dahyun Kim, Chanjun Park, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2024. SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*. Association for Computational Linguistics, Mexico City, Mexico, 23–35. doi:10.18653/v1/2024.naacl-industry.3

[28] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. ACM, 2356–2362. doi:10.1145/3404835.3463238

[29] Lili Lu. 2025. USIIR at TREC 2025 iKAT Track. In *Proceedings of the 34th Text REtrieval Conference (TREC 2025)*. Gaithersburg, Maryland.

[30] Simon Lupart, Zahra Abbasiantaeb, and Mohammad Aliannejadi. 2024. IRLab@iKAT24: Learned Sparse Retrieval with Multi-aspect LLM Query Generation for Conversational Search. *CoRR* abs/2411.14739 (2024). arXiv:2411.14739 doi:10.48550/ARXIV.2411.14739

[31] Simon Lupart, Zahra Abbasiantaeb, and Mohammad Aliannejadi. 2025. UvA-IRLab at iKAT25: Exploring Learned Sparse Retrieval and Query Rewriting for Personalized Conversational QA. In *Proceedings of the 34th Text REtrieval Conference (TREC 2025) (NIST SP xxxx)*. Gaithersburg, Maryland.

[32] Simon Lupart, Mohammad Aliannejadi, and Evangelos Kanoulas. 2024. DiSCo Meets LLMs: A Unified Approach for Sparse Retrieval and Contextual Distillation in Conversational Search. *CoRR* abs/2410.14609 (2024). arXiv:2410.14609 doi:10.48550/ARXIV.2410.14609

[33] Simon Lupart, Mohammad Aliannejadi, and Evangelos Kanoulas. 2025. Chatr1: Reinforcement learning for conversational reasoning and retrieval augmented question answering. *arXiv preprint arXiv:2510.13312* (2025).

[34] Simon Lupart, Daniël van Dijk, Eric Langezaal, Ian van Dort, and Mohammad Aliannejadi. 2025. Investigating LLM Variability in Personalized Conversational Information Retrieval. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (China) *(SIGIR-AP 2025)*. Association for Computing Machinery, New York, NY, USA, 353–363. doi:10.1145/3767695.3769502

[35] Kelong Mao, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen, and Hongjin Qian. 2023. Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 1211–1225. doi:10.18653/V1/2023.FINDINGS-EMNLP.86

[36] Mihran Miroyan, Tsung-Han Wu, Logan King, Tianle Li, Jiayi Pan, Xinyan Hu, Wei-Lin Chiang, Anastasios N. Angelopoulos, Trevor Darrell, Narges Norouzi, and Joseph E. Gonzalez. 2025. Search Arena: Analyzing Search-Augmented LLMs. *CoRR* abs/2506.05334 (2025). arXiv:2506.05334 doi:10.48550/ARXIV.2506.05334

[37] Fengran Mo, Abbas Ghaddar, Kelong Mao, Mehdi Rezagholizadeh, Boxing Chen, Qun Liu, and Jian-Yun Nie. 2024. CHIQ: Contextual History Enhancement for Improving Query Rewriting in Conversational Search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 2253–2268. doi:10.18653/V1/2024.EMNLP-MAIN.135

[38] Fengran Mo, Yuchen Hui, Yuxing Tian, Zhaoxuan Tan, Chuan Meng, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. 2025. Towards Adaptive Personalized Conversational Information Retrieval. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM 2025, Seoul, Republic of Korea, November 10-14, 2025*. ACM, 2137–2147. doi:10.1145/3746252.3761255

[39] Paul Over. 2001. The TREC interactive track: an annotated bibliography. *Inf. Process. Manag.* 37, 3 (2001), 369–381. doi:10.1016/S0306-4573(00)00053-4

[40] Arnold Overwijk, Chenyan Xiong, Xiao Liu, Cameron VandenBerg, and Jamie Callan. 2022. ClueWeb22: 10 Billion Web Documents with Visual and Semantic Information. arXiv:2211.15848 [cs.IR]

[41] Paul Owoicho and Jeff Dalton. 2025. GRILL Lab at TREC 2025: Agentic Iterative Retrieval and Gap-Aware Refinement for TREC IKAT and TREC RAG. In *Proceedings of the 34th Text REtrieval Conference (TREC 2025)*. Gaithersburg, Maryland.

[42] Paul Owoicho, Jeff Dalton, Mohammad Aliannejadi, Leif Azzopardi, Johanne R. Trippas, and Svitlana Vakulenko. 2022. TREC CAsT 2022: Going Beyond User Ask and System Retrieve with Initiative and Response Generation. In *Proceedings of the Thirty-First Text REtrieval Conference, TREC 2022, online, November 15-19, 2022 (NIST Special Publication, Vol. 500-338)*. National Institute of Standards and Technology (NIST).

[43] Sara Rosenthal, Yannis Katsis, Vraj Shah, Lihong He, Lucian Popa, and Marina Danilevsky. 2026. MTRAG-UN: A Benchmark for Open Challenges in Multi-Turn RAG Conversations. arXiv:2602.23184 doi:10.48550/ARXIV.2602.23184

[44] Sara Rosenthal, Vraj Shah, Yannis Katsis, and Marina Danilevsky. 2026. SemEval-2026 Task 8: MTRAGEval: Evaluating Multi-Turn RAG Conversations. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, San Diego, California.

[45] Mirco Speretta and Susan Gauch. 2005. Personalized Search Based on User Search Histories. In *2005 IEEE / WIC / ACM International Conference on Web Intelligence (WI 2005), 19-22 September 2005, Compiegne, France*. IEEE Computer Society, 622–628. doi:10.1109/WI.2005.114

[46] Johanne R. Trippas. 2024. Conversational Search. In *Information Retrieval: Advanced Topics and Techniques*. ACM Books, Vol. 60. ACM, 285–319. doi:10.1145/3674127.3674135

[47] Ellen M. Voorhees. 2003. Evaluating Answers to Definition Questions. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*, Marti A. Hearst and Mari Ostendorf (Eds.). The Association for Computational Linguistics. https://aclanthology.org/N03-2037/

[48] Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. InSCIt: Information-Seeking Conversations with Mixed-Initiative Interactions. *Trans. Assoc. Comput. Linguistics* 11 (2023), 453–468. doi:10.1162/TACL_A_00559

[49] Grace Hui Yang and Ian Soboroff. 2016. TREC 2016 Dynamic Domain Track Overview. In *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016 (NIST Special Publication,*

Vol. 500-321), Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec25/papers/Overview-DD.pdf

[50] Grace Hui Yang, Zhiwen Tang, and Ian Soboroff. 2017. TREC 2017 Dynamic Domain Track Overview. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017 (NIST Special Publication, Vol. 500-324)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec26/papers/Overview-DD.pdf

[51] Hui Yang, John R. Frank, and Ian Soboroff. 2015. TREC 2015 Dynamic Domain Track Overview. In *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015 (NIST Special Publication, Vol. 500-319)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST).

[52] Suveyda Yeniterzi and Reyyan Yeniterzi. 2025. Evaluating Full Dialogue History vs. Summarized Context for Personalized Knowledge Assistance: Findings from the TREC 2025 iKAT Track. In *Proceedings of the 34th Text REtrieval Conference (TREC 2025)*. Gaithersburg, Maryland.

[53] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. 2024. Building Cooperative Embodied Agents Modularly with Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

[54] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.