TREC iKAT 2025: The Interactive Knowledge Assistance Track Overview

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

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

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

ABSTRACT

Conversational information seeking has evolved rapidly in the last few years with the development of large language models (LLMs) providing the basis for interpreting and responding in a naturalistic manner to user requests. iKAT emphasizes the creation and research of conversational search agents that adapt responses based on the user's prior interactions and present context, maintaining a long-term memory of user-system interactions. This means that the same question might yield varied answers, contingent on the user's profile and preferences. The challenge lies in enabling conversational search agents (CSA) to incorporate personalized context to guide users through the relevant information effectively. iKAT's third year introduced an interactive conversation task, attracting seven teams and a total of 47 runs. Most of the runs leveraged LLMs in their pipelines, for single or multiple query rewriting, some also adopted agentic pipelines.

1 INTRODUCTION

As conversational retrieval systems become increasingly ubiquitous, it is more important than ever to expand the boundaries of research into the development and evaluation of these systems. Although significant contributions have been made to the field in recent years [7, 22], adapting responses of these systems to individual user preferences and characteristics remains an open and important issue. The TREC Interactive Knowledge Assistance Track (iKAT), built on top of foundational work of the TREC Conversational Assistance Track (CAsT) [20], aims at advancing research in personalized conversational information access and how to evaluate these systems. Personalized information access means that the same question might yield different answers, depending on the user's profile and preferences, which may have been established through past interactions. To illustrate personalized information access, we have outlined a scenario in Figure 1 in which different users with different backgrounds each ask the same conversational system which cow's milk alternatives they should try. The following three users participate in this scenario:

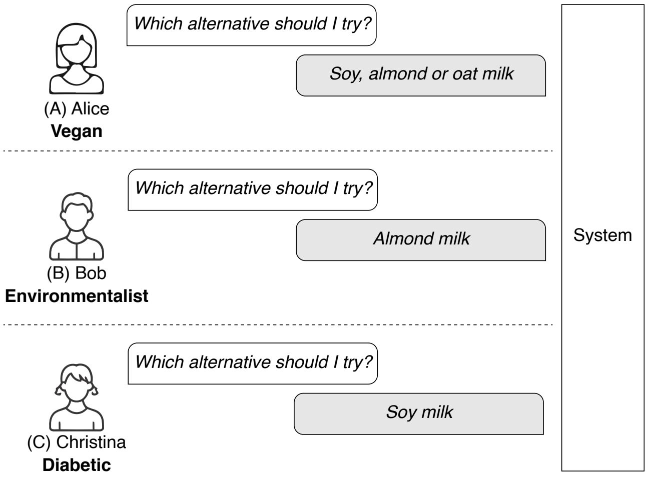


Figure 1: Example interaction for each of the three users, Alice, Bob, and Christina, when searching for alternatives to cow's milk using a conversational retrieval system that takes into account the users' individual personas.

- (A) Alice wants to try milk alternatives because she follows a vegan diet.
- (B) Bob wants to switch to milk alternatives because he is concerned about the environmental impact of cow's milk.
- (C) Christina wants to switch to milk alternatives because she was diagnosed with lactose intolerance. She also has diabetes and a nut allergy.

TREC'25, November 2025,

Given Alice, Bob, and Christina's personas, their conversation with the system evolves and develops in distinct ways. What is relevant to Alice may not necessarily be relevant to Bob or Christina, and vice versa. Consequently, by the end of their conversation, what they have learned about, what they have understood, and what they have decided regarding milk alternatives varies significantly, reflecting their personalized contexts. In order to provide personalized information access, a system needs to learn about the users' backgrounds, preferences, and characteristics. In Figure 2, two example conversations highlight how a conversational retrieval system can incorporate user information from a prior conversation (here about "finding a university") into another conversation (here about "finding a vacation destination").

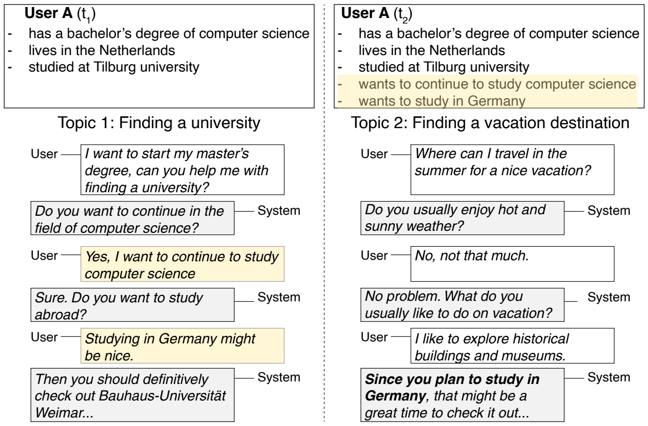
The challenge lies in enabling conversational search agents (CSA) to incorporate this personalized context to guide users through the relevant information effectively. iKAT also emphasizes decisional search tasks [25], where users sift through data and information to weigh up options in order to reach a conclusion or perform an action. These tasks, prevalent in everyday information-seeking decisions—be it related to travel, health, or shopping—often revolve around a subset of high-level information operators where queries or questions about the information space include: finding options, comparing options, and identifying the pros and cons of options. Given the different personas and their information need (expressed through the sequence of questions), diverse conversation trajectories will arise because the answers to these similar queries will vary substantially.

In iKAT's debut year [3, 4], we emphasized these tailored information needs by accounting for a person's knowledge, objectives, tastes, and limitations. To represent their personas, we used a Personal Text Knowledge Base (PTKB) to encapsulate both the task contexts and user specifics. The information requirements encompassed multifaceted tasks, including research, planning, and decision-making processes. Key research questions revolved around:

(1) **Personal Contexts**: How efficiently can an agent navigate various personal contexts, leading to distinct, relevant conversations?
(2) **Personalization**: Can agents adeptly modify conversational feedback based on the user's knowledge?
(3) **Elicitation**: Are agents proficient in drawing out pertinent persona information to customize discussions?
(4) **Dependent Relevance**: Can agents effectively employ context and prior responses to foster relevant conversations?

In Year 2 [2, 5], we continued this line by developing topics that go beyond the complexities of Year 1 by: (1) including more complex and ambiguous PTKB statements, and (2) testing the systems' ability not to answer the user's questions when the answer does not exist. The primary challenge in the track was to deliver a relevant and informative response given the user's PTKB by summarizing insights from various passages.

The Year 2 of iKAT also focused on novel approaches for evaluation with *nugget-based evaluation* and human-written gold answers, with in-depth comparisons of automatic and manual methods for evaluation. Year 2 also introduced a *dynamic pooling* strategy to increase the pool depth of assessed passages.


User A (t_1)
- has a bachelor's degree of computer science
- lives in the Netherlands
- studied at Tilburg university

Topic 1: Finding a university

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

System: *Do you want to continue in the field of computer science?*

User: *Yes, I want to continue to study computer science*

System: *Sure. Do you want to study abroad?*

User: *Studying in Germany might be nice.*

System: *Then you should definitively check out Bauhaus-Universität Weimar...*

User A (t_2)
- has a bachelor's degree of computer science
- lives in the Netherlands
- studied at Tilburg university
- wants to continue to study computer science
- wants to study in Germany

Topic 2: Finding a vacation destination

User: *Where can I travel in the summer for a nice vacation?*

System: *Do you usually enjoy hot and sunny weather?*

User: *No, not that much.*

System: *No problem. What do you usually like to do on vacation?*

User: *I like to explore historical buildings and museums.*

System: ***Since you plan to study in Germany**, that might be a great time to check it out...*

Figure 2: Two dialogues between a user (User A) and a conversational system on the topics of "finding a university" and "finding a vacation destination". In the dialogue at timestamp one (t_1), the system adapts its responses to the user's persona and identifies new persona traits. In the second dialogue at timestamp two (t_2), the original and newly identified persona traits are taken into account in the system's responses.

In Year 3, we expanded iKAT in various aspects. First, we introduced the concept of having multiple conversations by the same user. In this year's conversations, the current information could depend on utterances from a previous dialogue in the user's history. This is more in line with the everyday use of conversational systems, where an information need could be dependent on previous dialogues. Furthermore, we introduced a completely new subtask, namely, *interactive response generation*. In this subtask, we assigned a user simulator to each participating team via an API. Therefore, each system was assessed based on having an actual dialogue with a simulator. Given that each team could have a very different dialogue compared to other teams, we designed a novel question-based evaluation paradigm and asked NIST assessors to rate the quality of the dialogues based on grading rubric questions.

2 TRACK, TASKS, DATA, AND RESOURCES

In the following, we provide a detailed explanation of the task, data, and resources.

2.1 Track and Tasks

The task for participants in our track is to develop a personalized conversational search agent. These agents receive a context of prior utterances, retrieve relevant passages (*passage ranking*), and produce a relevant response (*response generation*), taking into account the user's persona. The persona is represented as a PTKB, a set of sentences in natural language that reflect the user's preferences or characteristics (e.g., "I want to increase my protein intake"). The sentences are assumed to be collected from previous conversations of the user with the system. A successful system should select from the PTKB statements those that are relevant to the current user utterance and context. This task is defined as a series of binary classification tasks (*statement classification*) over all statements in the PTKB. Besides an offline submission type, where users' utterances and contexts are static and taken from the test collection, iKAT offered an interactive submission format where conversations evolved naturally and in real time through the interaction of participant systems with simulated users (*interactive response generation*).

To sum up, the track includes the following subtasks:

- **PTKB Statement Classification:** The relevant statements from the PTKB for answering the current user utterance should be determined in this step. Given the context of the conversation and user utterance, the system classifies the statements from PTKB based on their relevance.
- **Passage Ranking:** Given the current user utterance, the context of the conversation, and the PTKB, the system retrieves relevant passages from the collection and ranks them based on their relevance.
- **Response Generation:** Given the current user utterance, the context of the conversation, and the PTKB statements of the user, a response should be generated that is fluent, satisfies the information need, and does not contain extraneous or redundant information. To achieve that, relevant passages should be retrieved such that the response can be considered a generative or abstractive summary of the relevant passages.
- **Interactive Response Generation:** Analogous to the response generation task, the task is to generate informative and relevant responses based on the current user utterance, context, and PTKB statements. However, for this task, participant systems interact in real time with a simulated user to submit conversations that are the subject of evaluation.

2.2 Topics

For iKAT 2025, conversations on 17 test topics were developed. Novel in this year's edition is that the same user (i.e., with the same set of PTKB statements) participates in multiple conversations. A total of 9 personas were created, each of whom participated in one or two conversations. Conversations from the same persona were designed in a way that users reveal additional information about themselves in the first conversation, which becomes relevant in the second conversation. All the topics, conversations, and PTKB statements were manually created by the organizers. Additional statistics about the test topics can be found in Table 1.

Topic Creation. For the manual creation process of topics, conversations, and PTKB statements, comprehensive guidelines have been compiled. 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 include a checklist to ensure the quality of the topics. These criteria include quality assurance in terms of the persona, turn, and conversation trajectory. We used a mix of LLM-assisted and personal exploration for each topic and discussed them in our meetings to improve the narrative. The topics that did not meet the quality criteria were discarded. Each topic was reviewed and refined by at least two other experts. The topic-creation process included the following steps:

(1) Generation of the user's PTKB for a given conversation/topic.
(2) Formulating the user utterances for each turn.
(3) Identifying relevant PTKB statements.
(4) Retrieving the relevant passages using the searcher tool provided to annotators (iKAT searcher).
(5) Employing GPT-based relevance assessment to estimate the number of relevant passages within the corpus.
(6) Formulating the responses of the system.

In generating the PTKB, we took great care to ensure that only high-level personal information was included, and any personally identifiable information was not included to ensure the privacy of the contributors.

Table 1: TREC iKAT 2025 collection statistics for manually constructed conversational data.

Overall Statistics	
Total Dialogues	17
Total Topics	17
Total Turns	188
Avg. Turns per Dialogue	11
Personalization	
Unique Users	9
Total PTKB Statements	337
Total Relevant PTKBs	134
Avg. PTKBs per Dialogue	20
Avg. Relevant PTKBs per Dialogue	8
Gold Responses	
Total Responses	45
Avg. length (words)	163
Total "nuggets"	2 241
Avg. "nugget" length (words)	30

2.3 Passage Collection

Given the size of ClueWeb22-B, we reused the same subset as in Years 1 and 2. No changes were made to the collection to ensure comparability across years. For clarity, we briefly summarize how this subset was originally constructed. The subset was created by manually inspecting document domains in ClueWeb22-B. We selected domains to maintain diversity and removed those deemed irrelevant. The resulting collection contains 116 838 987 passages, as distributed by CMU.

Document segmentation is the same as in previous years, following the methodology used in the TREC Deep Learning track for MS MARCO:

(1) Each document was initially shortened to a length of 10 000 characters.
(2) A sliding window approach was then used, where we took 10 consecutive sentences as a single passage.
(3) After these 10 sentences, we moved the window by 5 sentences (i.e., a 5-sentence stride) to create the next passage.

We provided the following resources to the participants:

(1) Python scripts that were used to segment the passages.
(2) Segmented passages along with MD5 hashes.
(3) Pyserini index of the collection.
(4) ir_datasets access to the collection and indexes.

(5) Learned sparse index of the collection and script to build a dense index.

2.4 Baselines

The organizers provided seven baselines for the *passage ranking* and *response generation* tasks, with diverse retrieval systems (sparse, dense, and learned sparse), as well as various query rewriting and reranking strategies. The baselines are detailed below:

(1) **orga-bm25-nopersonal**: Employs GPT-4.1-mini for query rewriting, BM25 for retrieval, MiniLM[1] for reranking, and Llama-70B for answer generation without PTKB personalization. The top 100 results are reranked, and the top 5 are used for answer generation.
(2) **orga-bm25-personal**: Similar to (1), with integration of PTKB in query rewriting and answer generation. This run is used as the generation-only run.
(3) **orga-ance-norerank**: Uses GPT-4o-mini for query rewriting, ANCE for dense retrieval, no reranking, and Llama-70B for answer generation.
(4) **orga-splade-norerank**: Uses GPT-4o-mini for query rewriting, SPLADE for sparse retrieval, no reranking, and Llama-70B for answer generation.
(5) **orga-ance-llama70b**: Uses Llama-70B for query rewriting, ANCE for dense retrieval, MiniLM for reranking, and Llama-70B for answer generation.
(6) **orga-splade-llama70b**: Uses Llama-70B for query rewriting, SPLADE for sparse retrieval, MiniLM for reranking, and Llama-70B for answer generation.
(7) **orga-bm25-human**: Uses manually rewritten queries, BM25 for retrieval, MiniLM for reranking, and Llama-70B for answer generation.

Additional baselines were released for the *interactive response generation* task, some of which mirror those used in the response generation task to support cross-task comparison.

(1) **orga-gpt41mini-bm25-minilm-llama70b**: This run is the pipeline used for the automatic baseline *orga-bm25-personal*, but applied for the interactive task.
(2) **orga-gpt41mini-bm25-minilm-llama70b-nopersonal**: This run is the pipeline used for the automatic baseline *orga-bm25-nopersonal*, but applied for the interactive task.
(3) **orga-llama70b-bm25-minilm-llama70b**: This run is using Llama-70B for rewriting and PTKB integration, and then uses BM25 and MiniLM for ranking, with Llama-70B for the final answer using RAG.
(4) **orga-llama8b-bm25-minilm-llama8b-v2**: This run is a minimal version of (3) with Llama-8B as the backbone LLM.
(5) **orga-no-no-no-gpt41mini**: This run applied GPT-4.1 mini directly on the conversation history to produce an answer. It does not use any retrieval for grounding.
(6) **orga-no-no-no-llama70b**: Similar to (5), this run uses Llama-70B for answer generation without grounding.
(7) **orga-no-no-no-llama8b-v2**: This run uses Llama-8B for answer generation without grounding.

2.5 User Simulation

In Year 3 of `iKAT`, the interactive response generation task required the development of an approach to simulate users of conversational systems. To ensure fair assessment across participants' systems, a user simulation approach was designed that combines the reproducibility of Cranfield-style evaluation with the flexibility and interactivity of user simulation.

To ensure that each conversation on a specific topic followed a comparable, predefined trajectory across all participants' systems, a sequence of grading rubric questions was extracted from the test conversations. A *rubric question* represents a concise question that reflects a specific aspect of a broader information need [9] (e.g., "what is the impact of roasting on coffee taste?" on the topic of "how to make good coffee"). For each conversation in the test set, we extracted at least one sequence of five rubrics that best capture the course of the conversation. These sequences of rubrics were given to the user simulator to guide the utterance generation.

The user simulator was designed to follow a simple protocol: it generates an utterance based on a given rubric, evaluates the relevance of the system's response with respect to that rubric, and, if the response is sufficiently relevant, proceeds to the next rubric. If the response is not relevant enough, the simulator provides feedback and asks again. Finally, after all rubrics were addressed, the user simulator closes the conversation with a farewell.

For the utterance generation, we prompt an LLM by providing the current state of the dialogue context, all PTKB statements of the given user, and the current rubric in focus. To favor user utterances that align with the given rubric, the simulator considers multiple output sequences and selects the one that is semantically most similar to the rubric. This is quantified by cosine similarity between Sentence-BERT embeddings [24] of the rubric and the user utterance.

The system's response to the user utterance is then assessed by prompting an LLM with the same prompt that was used for the grading of rubric questions in the RUBRIC score approach [9]. Given the rubric and the system response, the LLM is asked to assign a number on a six-point Likert (0–5) scale to quantify whether the system response answers the rubric question, where "5" means that the response is highly relevant, complete, and accurate, and "0" entails that the response is not relevant or complete at all. If this score exceeds the threshold of 3, the user simulator proceeds to the next rubric. If the threshold is not met, the user simulator is asked to provide feedback and rephrase its request. This feedback and rephrase loop could be repeated up to three times before the user simulator was guided to proceed to the next rubric.

Participants of the interactive response generation task could interact with the user simulator through a stateful API that was developed by the organizers. Participants could use this API to debug their systems with the help of an installed debugging simulator that was implemented by an LLM with fewer parameters (Gemma 3 with 4 billion parameters [13]). The run submission was also handled by the API, but the simulator was equipped with a much larger LLM (GPT 4.1 [19]).

[1] `ms-marco-MiniLM-L6-v2` on HuggingFace.

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

Total Passages assessed	5 650
Fails to meet (0)	2 227
Slightly meets (1)	1 288
Moderately meets (2)	1 288
Highly meets (3)	606
Fully meets (4)	241

3 EVALUATION

3.1 Manual Assessment and Evaluation

PTKB Statement Relevance Assessment. The relevance of PTKB statements for each turn was assessed using two distinct sets of evaluations, created by the organizers and by NIST assessors. During topic creation, the organizers annotated each turn in terms of their provenance to PTKB statements and included these labels in the released topic files. To ensure annotation quality, each turn was assigned to at least two organizers. In cases of disagreement, a third organizer was involved to make the final decision based on majority voting.

Moreover, during the assessment of passage relevance, NIST assessors evaluated the relevance of PTKB statements to each turn. The assessment pool for NIST assessors was smaller than that of the organizers: while the organizers evaluated all turns, NIST assessors only assessed those turns selected for passage relevance. Only turns annotated as "personalized" by both NIST and the organizers were retained, and PTKB performance was reported based on the assessments from both groups.

Passage Retrieval Assessment. The NIST assessors judged the relevance of the passages based on the methodology and scale established in CAsT. A subset of 116 out of 218 turns was selected to be judged by NIST assessors. Among the unassessed turns were responses that were either clarifications (e.g., "Do you have any dietary requirements?") or were responses to utterances that were too generic and hence would return too many relevant documents (e.g., "I'm traveling to California, do you have any suggestions?"). A pool of 20 575 passages was created and manually evaluated. On average, 177 passages were judged for each turn. More detailed statistics of the collected data and judgments can be found in Table 1. The number of turns per dialogue and the number of turns evaluated per dialogue are shown in Figure 3.

Gold Response and Gold Nuggets Assessment. Together with the passage retrieval assessment, NIST assessors extracted substrings of relevant passages as *gold nuggets*. Overall, 2 241 nuggets were extracted from the relevant passages, with an average length of 30 words. NIST assessors also wrote gold responses based on extracted nuggets, relevant passages and PTKB. In total, 45 turns have nuggets and gold answers that can be used for the evaluation of the response generation task.

Interactive Response Generation Assessment. For the interactive response generation task, a human evaluation of the submitted runs was conducted. NIST assessors evaluated system responses using two levels of criteria: rubric level and dialogue level.

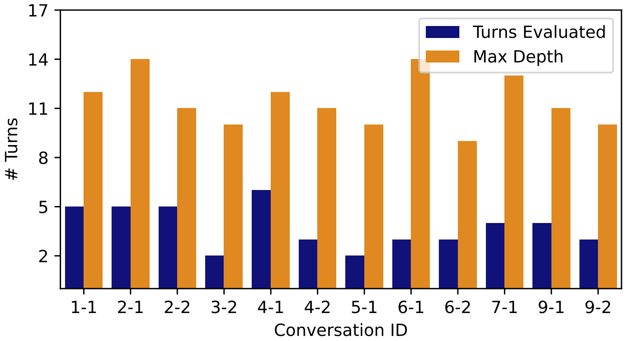


Figure 3: Number of turns evaluated per dialogue in the final judgment pool vs. the maximum depth of each topic.

Rubric-Level Manual Evaluation. Each system response to a user utterance in the interactive response generation task was evaluated at the rubric level. For each set of turns associated with a given rubric, system responses were assessed based on the following criteria: (i) *engagement* ("To what extent does the system encourage the user to engage with it?"), (ii) *relevance* ("Given the conversation history, user's latest request, and user's PTKB, how would you assess the relevance and usefulness of the generated response to the user's request?"), and (iii) *overall quality* ("Considering all factors (relevance, engagement, and other factors that were not covered by our questions), what is the overall quality and utility of the system's performance?").

Dialogue Level Manual Evaluation. In addition to the rubric-level evaluation, the overall quality of the dialogues was assessed using several dialogue-level criteria: (i) *mixed-initiative strategies* ("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?"), (ii) *personalization* ("How well does the system tailor the dialogue to the specific user persona provided?"), (iii) *information flow* ("How well does the system maintain a coherent and logical flow of information throughout the conversation?"), (iv) *trustworthiness* ("How likely would you be to believe that past and future answers of this system are factually correct? "), and (v) *user satisfaction* ("Considering the entire dialogue, how satisfied would you be as the user and how successful was the interaction?").

Each criterion in both rubric-level and dialogue-level evaluations was rated on a five-point Likert scale (1–5), where "5" indicates the highest level of performance for the respective criterion. Additionally, NIST assessors rated their confidence in the ratings assigned on a 3-point scale (low, medium, high). The complete guidelines provided to NIST assessors for the interactive response generation assessment will be publicly released.

3.2 Evaluation Metrics

Statement Classification Task. The PTKB statement classification task was evaluated at both the turn and conversation level using set-based metrics, including precision, recall, and F1-Measure.

Passage Ranking Task. The runs were evaluated across two dimensions, given the ranking for each topic turn: (i) the ranking depth and (ii) the turn depth. For ranking depth, we focused on early positions 3 and 5 for the conversational scenario, under the assumption that the top k results would be used to formulate the response. For turn depth, we evaluated the run performance at the n-th conversational turn. Performing well on later turns indicates a better ability to understand the preceding context. We used the mean nDCG@5 as the main evaluation metric, with all conversational turns averaged using uniform weights. We also measured the turn-depth measure based on nDCG@5, with the per-turn nDCG@5 scores averaged at depth (n). In addition to the nDCG metrics (nDCG, nDCG@3, and nDCG@5), we also calculated Precision@10 (P@10), Recall, Recall@10, and mean Average Precision (mAP), where again, we averaged over all turns.

Response Generation Task. The evaluation for response generation in this cycle relied on written gold responses from NIST assessors, as well as extracted nuggets (pieces of information from relevant passages). First, a diverse set of surface-based metrics is applied to the predicted responses, including Rouge and F1. Then, semantic-based metrics were employed using BEM [8], followed by LLMEval with both GPT-4.1 [12] and SOLAR-10.7B-Instructv1.0 [14] as answer equivalence evaluators. SOLAR has been shown to provide a good compromise between parameter size (efficiency) and effectiveness [23].

Finally, we evaluated the coverage of generated responses relative to human-written nuggets. For each turn, GPT-4.1 determined how many gold nuggets were entailed from the predicted response, following a Nugget-to-Response (NtR) evaluation [2]. This evaluation allow the calculation of recall metrics for nugget coverage.

Interactive Response Generation Task. The evaluation for the interactive response generation task was based on the manual assessments conducted by the NIST assessors (see Section 3.1). For rubric- and dialogue-level assessments, individual scores are aggregated, taking an assessor's confidence into account. For rubric-level assessments, we consider the overall quality weighted by assessor confidence as the score. The dialogue-level score was calculated as the rated user satisfaction, weighted by the assessor's confidence. The system-level score represents the harmonic mean between the rubric-level and dialogue-level scores.

In addition to the evaluation of the generated responses, we assess the given citations for a response according to retrieval metrics. To this end, we employ the relevance judgments from the response generation task, pool them by rubric, and assess the cited passages on a rubric level.

4 PARTICIPANTS

The iKAT offline task received 30 run submissions (auto and gen_-only) from seven groups shown in Table 3. The organizers provided seven runs (auto) as baselines for comparison. iKAT's new interactive task received 16 runs together with 7 baselines.

Figure 4: The prompt designed for the evaluation of generated responses, with GPT4.1 and SOLAR-10.7B-Instruct-v1.0.

You are an evaluation tool. Just answer by {Yes} or {No}. Here is a question, a golden answer, and an AI-generated answer. Judge whether the AI-generated answer is correct according to the question and gold answer. Answer with {Yes} or {No}.
Question: {*question*}
Gold Answer: {*golden answer*}
Generated Answer: {*predicted response*}
Response: {Yes/No}

Figure 5: The prompt designed for nugget coverage evaluation. For turns associated with multiple gold nuggets, the LLM is invoked separately for each nugget-answer pair, ensuring one binary decision (yes/no) per nugget.

# **Instruction:** *You are an automatic evaluator of answers for an information need. You will be given a gold fact (nugget) and a model answer. Your task is to determine whether the answer fully covers the nugget, meaning that it correctly includes the nugget's content and meaning (not partially or incorrectly). If in doubt, respond "no".*
# **Gold Fact:** {*gold nugget*}
# **Model Answer:** {*predicted response*}
# **Please answer the following:**
Does the Model Answer fully cover the Gold Fact? Respond strictly with "yes" or "no", without any explanation or punctuation.
# **Answer** (yes/no):

Most teams used a multi-step pipeline consisting of the following: (1) PTKB statement relevance prediction; (2) conversational rewriting (most incorporating the previous canonical responses as well as predicted relevance PTKB statements) and conversational query expansion; (3) retrieval using a traditional or dense IR model; and (4) multi-stage passage re-ranking with neural language models fine-tuned for point-wise (mono) and pairwise (duo) ranking. We categorize such submissions as *automatic* runs. Table 3 lists the submissions from the teams, as well as their pipelines.

4.1 Participant Runs

Table 3 provides an overview of the participant runs, and below we include a summary of each, starting with **automatic** runs:

(1) **auto_REnpr_npsg20_thru03_d3c5**: Employs CHIQ-AD [18] and LLM4CS [17] for query rewriting, with the resulting rewrites scored by AdaRewriter [15]. It performs retrieval using SPLADE on both rewrite sets and combines the results via Reciprocal Rank Fusion (RRF). The retrieved passages are reranked using cross-encoder model[2]. GPT-4o-mini summarizes the top-k documents and the PTKBs to generate the final response.
(2) **auto_ori_npsg20_thru03_d3c5**. Same as (1); variant settings only.

[2] https://huggingface.co/naver/trecdl22-crossencoder-debertav3

Table 3: Participants and their submitted runs in the automatic (auto), generation only (gen_only) and interactive subtasks.

Run ID	Type
organizers	
orga-bm25-human	auto
orga-splade-norerank	auto
orga-splade-llama70b	auto
orga-bm25-personal	auto
orga-ance-llama70b	auto
orga-ance-norerank	auto
orga-bm25-nopersonal	auto
orga-gpt41mini-bm25-minilm-llama70b	interactive
orga-gpt41mini-bm25-minilm-llama70b-nopersonal	interactive
orga-llama8b-bm25-minilm-llama8b-v2	interactive
orga-llama70b-bm25-minilm-llama70b	interactive
orga-no-no-no-gpt41mini	interactive
orga-no-no-no-llama8b-v2	interactive
orga-no-no-no-llama70b	interactive
cfd	
auto_npr1_npsg20_thru03_d3c5	auto
auto_npr_npsg20_thru03_d3c5	auto
auto_ori_npsg20_thru03_d3c5	auto
auto_REnpr_npsg20_thru03_d3c5	auto
gen-only_npsg13_thru0_d4c5	gen_only
gen-only_npsg20_thru03_d3c5	gen_only
cfda-adarewriter-chiq-llm4cs-splade	interactive
cfda-chiq-llm4cs-splade-rrf	interactive
grillab	
grilllab-larf-finetuned-10-rounds	auto
grilllab-larf-finetuned	auto
grilllab-larf-finetuned-rankllm	auto
grilllab-larf-finetuned-22-rounds	auto
grilllab-agentic-gpt4.1	interactive
grilllab-agentic-gpt4.1-larf	interactive
grilllab-agentic-gpt4.1-larf-v2	interactive
grilllab-larf-fine-tuned-judge	interactive
genaius	
genaius-genonly-summary-gpt4o	gen_only
genaius-genonly-full-gpt4o	gen_only
genaius-full-rewrite	interactive
genaius-summary-rewrite	interactive
guidance	
cosine-orconvqa	auto
agg_true-qrec-mse	auto
agg_false-qrec-mse	auto
genonly_clariftop10	gen_only
ucsc	
UCSC-base-trained-MiniLM	auto
UCSC-base-ensemble	auto
UCSC-SIMRAG-trained-MiniLM	auto
UCSC-SIMRAG-ensemble	auto
ucsc-base-dynamicPTKB-trainedReranker	interactive
ucsc-SIMRAG-guidelineQuery-dynamicPTKB-trainedReranker	interactive
ucsc-SIMRAG-keywordQuery-dynamicPTKB-ensembleReranker	interactive
ucsc-SIMRAG-keywordQuery-dynamicPTKB-trainedReranker	interactive
usiir	
usiir_run1	gen_only
usiir_run2	gen_only
uva	
disco-qrecc	auto
mq4cs-llamaft-splade	auto
mq4cs-gpt41-bm25	auto
mq4cs-gpt41-splade	auto
nuggets-ptkb	gen_only
nuggets-noptkb	gen_only
uva-gpt5-bm25-debertav3-gpt5	interactive
uva-gpt5-bm25-debertav3-gpt5mini-nopersonal	interactive
uva-gpt5mini-bm25-debertav3-gpt5mini	interactive
uva-gpt5mini-no-no-gpt5mini	interactive

(3) **auto_npr_npsg20_thru03_d3c5**. Same as (1); AdaRewriter trained on query rewriting datasets (QReCC [6], InfoCQR [27]).

(4) **auto_npr1_npsg20_thru03_d3c5**. Same as **(3)** with minor variant settings.

(5) **grilllab-larf-finetuned**. A multi-stage BM25 pipeline where an LLM iteratively refines retrieval. The model first rewrites the user query using dialogue and PTKB context, retrieves initial passages, then expands coverage by identifying similar documents and generating new synthetic queries to address missing information. Each retrieval stage includes filtering and reranking, and the final response is produced by the LLM from the top-ranked passages.

(6) **grilllab-larf-finetuned-10-rounds**. Same as (5), iterated up to 10 rounds or until no new passages are retrieved.

(7) **grilllab-larf-finetuned-22-rounds**. Same as (5), iterated for 22 rounds.

(8) **grilllab-larf-finetuned-rankllm**. Same as (5), with final reranking done with RankGPT [11].

(9) **mq4cs-gpt41-splade**: This run employs the MQ4CS framework [1] to generate five query rewrites using GPT-4.1. Retrieval is done using SPLADE, followed by reranking with DeBERTa-v3. Finally, response generation is done with GPT-4.1 on the top-5.

(10) **mq4cs-gpt41-bm25**. Same as (9); with BM25 as retriever and MiniLM as reranker.

(11) **mq4cs-llamaft-splade**: Same as (9); employs a fine-tuned LLaMA model to generate a contextualized query.

(12) **disco-qrecc-norerank**: Employs the DiSCo model [16] in a zero-shot setting for direct retrieval, without applying query rewriting. No reranking step is performed. The final response is generated using a nugget-based approach with GPT-4.1.

(13) **cosine-orconvqa**: Uses CoSPLADE [10] (with SPLADEv3 as backbone) trained on the OrConvQA [21] dataset using cosine loss with a history size of 1. It doesn't employ query rewriting or reranking. The top-10 results are summarized, with GPT-4.1 handling PTKB integration, CQ detection, answer generation.

(14) **agg_true-qrec-mse**: Same as (13), but uses CoSPLADE [10] with NeoBERT, optimized with QReCC MSE loss, and aggregates context over the full dialogue history.

(15) **agg_false-qrec-mse**: Same as (14), but aggregates context only over the current turn.

Table 4: Passage ranking evaluation of submitted runs. Evaluation at retrieval cutoff of 1000, with >=1 as relevance threshold.

Group	Run ID	nDCG@3	nDCG@5	nDCG	P@20	Recall@20	Recall	mAP
grilllab	grilllab-larf-finetuned-rankllm	0.5941	0.5843	0.5070	0.5867	0.2173	0.5273	0.3041
grilllab	grilllab-larf-finetuned	0.5489	0.5427	0.4991	0.5711	0.2137	0.5245	0.2943
grilllab	grilllab-larf-finetuned-10-rounds-10-rounds	0.5394	0.5375	0.5025	0.5711	0.2130	0.5334	0.2953
grilllab	grilllab-larf-finetuned-22-rounds	0.5346	0.5267	0.5048	0.5656	0.2134	0.5412	0.2955
cfda	auto_ori_npsg20_thru03_d3c5	0.5239	0.5040	0.4116	0.5522	0.2037	0.3721	0.2375
uva	mq4cs-gpt41-splade	0.4916	0.4897	0.5357	0.5789	0.2073	0.6101	0.3099
ucsc	UCSC-SIMRAG-ensemble	0.4787	0.4789	0.5462	0.5656	0.2033	0.6648	0.3232
ucsc	UCSC-base-ensemble	0.4816	0.4721	0.5624	0.5622	0.2019	0.7016	0.3252
ucsc	UCSC-base-trained-MiniLM	0.4562	0.4705	0.5264	0.4544	0.1677	0.7016	0.2551
ucsc	UCSC-SIMRAG-trained-MiniLM	0.4502	0.4392	0.5116	0.4800	0.1758	0.6648	0.2592
cfda	auto_REnpr_npsg20_thru03_d3c5	0.4458	0.4326	0.3731	0.5056	0.1778	0.3619	0.2133
cfda	auto_npr1_npsg20_thru03_d3c5	0.4153	0.4160	0.3638	0.4900	0.1786	0.3588	0.2046
uva	mq4cs-gpt41-bm25	0.4180	0.4053	0.4126	0.4756	0.1675	0.4481	0.2139
cfda	auto_npr_npsg20_thru03_d3c5	0.4135	0.4050	0.3593	0.4900	0.1775	0.3546	0.2091
organizers	orga-bm25-personal	0.4074	0.3902	0.3779	0.4178	0.1498	0.4308	0.1871
organizers	orga-bm25-nopersonal	0.3739	0.3797	0.3779	0.4378	0.1572	0.4307	0.1950
organizers	orga-ance-llama70b	0.3937	0.3783	0.4245	0.4044	0.1427	0.5267	0.2187
uva	mq4cs-llamaft-splade	0.3612	0.3680	0.3911	0.4756	0.1701	0.4413	0.2045
organizers	orga-splade-llama70b	0.3753	0.3676	0.4816	0.4122	0.1470	0.6721	0.2320
organizers	orga-ance-norerank	0.3692	0.3510	0.3872	0.3989	0.1414	0.5258	0.1714
organizers	orga-splade-norerank	0.3555	0.3498	0.4719	0.4000	0.1545	0.6760	0.2150
uva	disco-qrecc	0.3087	0.3042	0.3926	0.3544	0.1347	0.5524	0.1683
organizers	orga-bm25-human	0.3021	0.2988	0.2955	0.3411	0.1336	0.3514	0.1492
guidance	agg_true-qrec-mse	0.3161	0.2985	0.3692	0.3167	0.1233	0.5185	0.1690
guidance	cosine-orconvqa	0.2715	0.2765	0.3609	0.3233	0.1225	0.5205	0.1549
guidance	agg_false-qrec-mse	0.2483	0.2411	0.3409	0.2922	0.0943	0.5140	0.1378

(16) **UCSC-base-trained-MiniLM**: Employs SPLADE for retrieval, followed by reranking with a MiniLM model fine-tuned on TREC-derived data. The final response is generated based on the top-20 retrieved passages.
(17) **UCSC-base-ensemble**: Same as (16), but uses an ensemble reranking strategy utilizing DeBERTaV2/V3, ALBERT, ELECTRA, and RoBERTa, aggregated via min-max normalization.
(18) **UCSC-SIMRAG-trained-MiniLM**: Employs PTKB-aware keyword rewriting by LLM, which is then validated and regenerated by SIM-RAG [26] (using a TREC-tuned T5 model). Retrieval is done with SPLADE, followed by reranking using MiniLM with a complete base-prompt query. The generation is done based on top-20 passages.
(19) **UCSC-SIMRAG-ensemble**: Same as (18), but replaces the single model reranker with the 5-model ensemble (aggregated via min-max normalization).

In addition, there were nine **generation only** runs:

(1) **gen-only_npsg20_thru03_d3c5**: This approach employs classification of PTKB statements. It summarizes the top-20 passages but retains the top-3 passages in their unsummarized form for the final generation with GPT-4o-mini.
(2) **gen-only_npsg13_thru0_d4c5**: Same pipeline as (1), but summarizes the top-13 passages and retains the top-4 passages in their unsummarized form.
(3) **genaius-genonly-summary-gpt4o**: GPT-4o generation based on the top-5 passages (as-is) combined with PTKB information. It conditions the response on a summary of the conversation history.
(4) **genaius-genonly-full-gpt4o**: Same as (3); but uses full conversation history.
(5) **genonly_clariftop10**: A GPT-4.1 pipeline that detects clarification needs. It uses the top-10 documents, summarizing each one prior to response generation.
(6) **nugget-ptkb**: A GPT-4.1 pipeline that judges the top-10 retrieved passages and extracts information "nuggets" to generate an answer. PTKBs are incorporated into all steps (judgment, extraction, and generation).
(7) **nugget-noptkb**: Same as (6); without using the PTKBs.
(8) **usiir_run1**: Qwen3-8B generation conditioned on a ranked list of documents, without access to PTKBs.
(9) **usiir_run2**: Same as (8); but with access to PTKBs.

Furthermore, there were sixteen runs for the **interactive** task:

(1) **cfda-adarewriter-chiq-llm4cs-splade**: Rewrites queries with AdaRewriter (CHIQ-AD+LLM4CS), retrieves passages via SPLADE, reranks top-2000 with DeBERTaV3 cross-encoder, then generates answers with GPT4o.
(2) **cfda-chiq-llm4cs-splade-rrf**: Generates 2 rewrites each with CHIQ-AD and LLM4CS, retrieves passages with SPLADE, fuses results using RRF, then generates response with GPT-4o.
(3) **genaius-full-rewrite**:
FullHistory-RewriteUtterance-BM25-GPT4o
(4) **genaius-summary-rewrite**:
SummaryHistory-RewriteUtterance-BM25-GPT4o

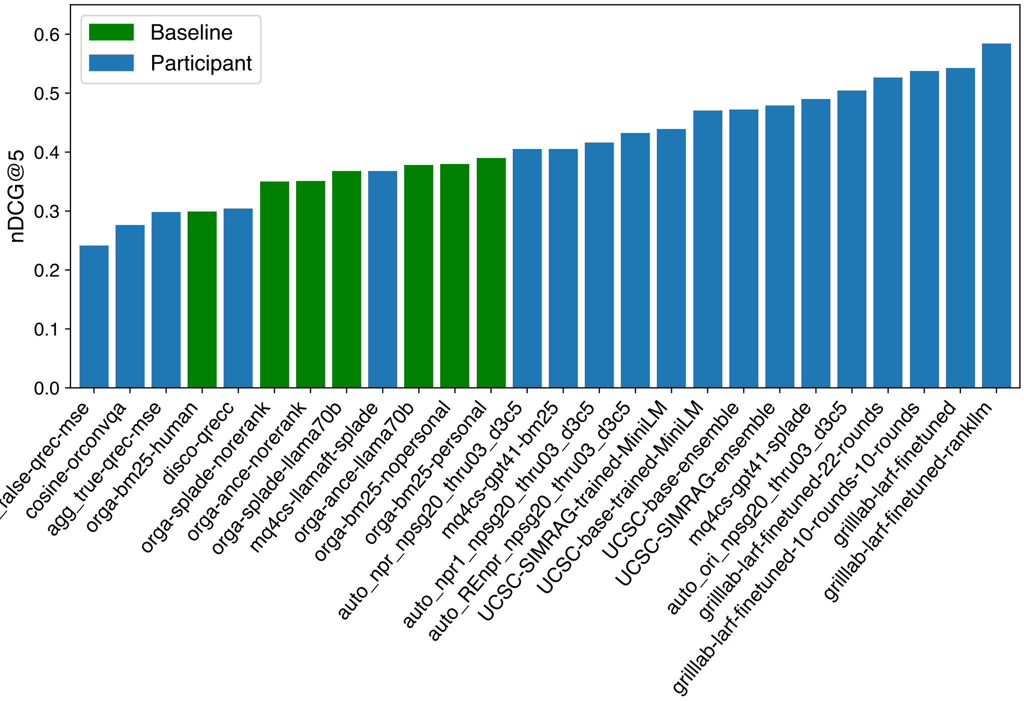


Figure 6: Performance of all automatic runs in terms of nDCG@5 on the passage ranking task.

(5) **grilllab-agentic-gpt4.1**: Agentic pipeline where an LLM decides an action to take based on the conversation context. Agent has access to search engine as a tool.
(6) **grilllab-agentic-gpt4.1-larf**: Agentic pipeline where an LLM decides an action to take based on the conversation context, manual orchestration is minimal. Agent has access to a LARF based search engine as a tool
(7) **grilllab-agentic-gpt4.1-larf-v2**: Agentic pipeline where an LLM decides an action to take based on the conversation context, manual orchestration is minimal. Agent has access to a LARF based search engine as a tool.
(8) **grilllab-larf-fine-tuned-judge**: Three stage pipeline where a candidate pool is retrieved, the candidate pool is expanded with doc2query techniques, and the pool is refined with an LLM. Each stage has a sub-step where the candidate pool is assessed with respect to the query and filter out passages that are not relevant, followed by a reranking step.
(9) **ucsc-base-dynamicPTKB-trainedReranker**: LLMs are utilized to extract additional relevant PTKB statements from the conversation. Query generation is done using the base query, and the reranking is based on a model trained on previous TREC data.
(10) **ucsc-SIMRAG-guidelineQuery-dynamicPTKB-trained Reranker**: This approach utilizes LLMs to extract additional relevant PTKB statements from the conversation. It then utilizes a trained SIMRAG model to obtain a good guideline-generated retrieval query, along with a good reranking query. The reranking is based on a model trained on previous TREC data.
(11) **ucsc-SIMRAG-keywordQuery-dynamicPTKB-ensemble Reranker**: This approach utilizes LLMs to extract additional relevant PTKB statements from the conversation. It then utilizes a trained SIMRAG model to obtain a good keyword retrieval query, along with a good reranking query. The reranking is based on an ensemble of five rerankers, reranking the top 100.
(12) **ucsc-SIMRAG-keywordQuery-dynamicPTKB-trained Reranker**: This approach utilizes LLMs to extract additional relevant PTKB statements from the conversation. It then utilizes a trained SIMRAG model to obtain a good keyword retrieval query, along with a good reranking query. The reranking is based on a model trained on previous TREC data.
(13) **uva-gpt5mini-bm25-debertav3-gpt5mini**: The approach uses a RAG pipeline, using first GPT-5 mini as a query rewriter on the conversation history and the PTKBs, followed by BM25 for retrieval and DeBERTa-v3 for reranking, then it uses GPT-5 mini on the top retrieved passages and PTKBs to generate the final answer.
(14) **uva-gpt5-bm25-debertav3-gpt5mini-nopersonal**: This run is similar to *uva-gpt5mini-bm25-debertav3-gpt5mini* but without the PTKB integration.
(15) **uva-gpt5mini-no-no-gpt5mini**: This run doesn't use retrieval, but uses GPT-5-mini as query rewriter and then for the answer generation.
(16) **uva-gpt5-bm25-debertav3-gpt5**: This run is similar to *uva-gpt5mini-bm25-debertav3-gpt5mini* but uses GPT-5 instead of GPT-5 mini.

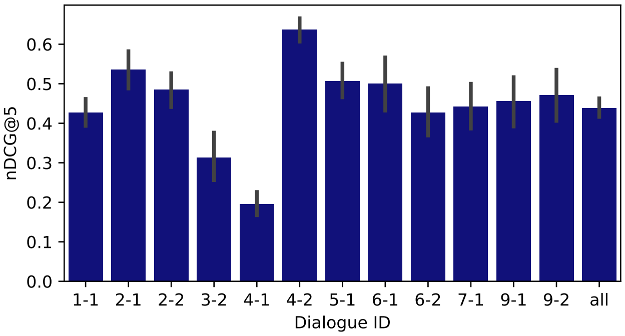


Figure 7: nDCG@5 aggregated for each topic across all runs on the passage ranking task (mean and std across runs).

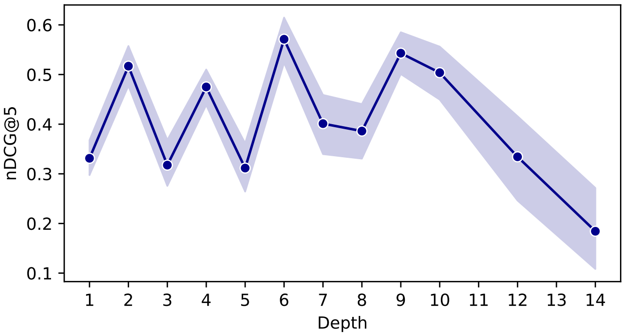


Figure 8: nDCG@5 at varying conversation turn depths on the passage ranking task (mean and std across runs).

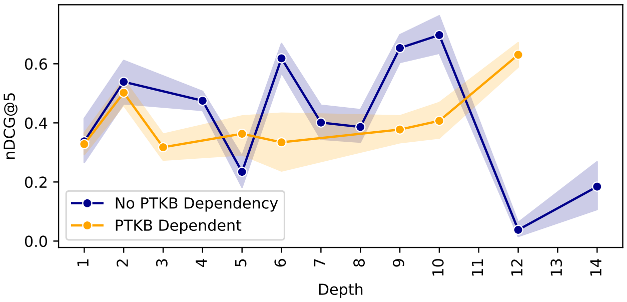


Figure 9: nDCG@5 at varying conversation turn depths on the passage ranking task, for turns that depend on PTKB statements vs. those that do not (mean and std across runs).

5 RESULTS

5.1 Passage Ranking

5.1.1 Overall results. Table 4 lists the performance of the automatic runs in terms of all the evaluation metrics. We see the dominance of models that leverage GPT-* in their pipeline on top of the list, followed by models that leverage Llama. Figure 6 compares the performance of all the automatic runs in terms of nDCG@5, where the baseline runs are colored in green.

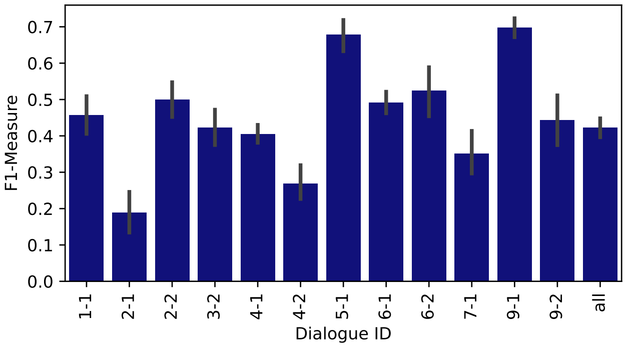


Figure 10: F1-Measure on PTKB relevance prediction, for each topic across all runs (mean and std across runs).

5.1.2 Performance per dialogue. Figure 7 reports the average run performance per topic. We see in particular that while the runs perform well for some topics, they fail to perform well for others. In particular, we find topics 4-2 ("Visiting New York") and 2-1 ("Making good coffee") to be the easiest, while 4-1 ("Preparing for a run") and 3-2 ("New activities") to be the most difficult ones.

5.1.3 Performance at different depths. Figure 8 reports the performance of all runs (median or better) at varying conversation turns in terms of nDCG@5. We also report the performance at different depths, separating the turns that depend on PTKB provenance in Figure 9. Our intuition is that the PTKB statement ranking step will introduce additional difficulty and error in the pipeline and consequently the runs exhibit lower performance. However, this was not always the case, and in most cases, PTKB dependence led to lower performance. Unlike CAsT and iKAT Year 1, we see that the models do not necessarily perform best in the first turns. Interestingly, we see an upward performance trend in deeper turns, with the peak performance at depth 14. This is an interesting phenomenon that needs further investigation. When looking at the personalized turns in Figure 9, we observe a different trend, showing that PTKB-dependent turns generally become more challenging as the dialogue progresses. This is also corroborated by the turn-level PTKB performance presented in Figure 11.

5.2 PTKB Provenance

5.2.1 Overall results. As previously described, we evaluated the submissions for the PTKB statement ranking task based on two relevance judgments, namely, assessed by the NIST assessors, as well as the organizers. We report the results based on NIST assessments in Table 5, and the results based on the organizers' assessment in Table 6 in terms of all evaluation metrics. We report the results only on the intersection of turns that are deemed to be personalized by both NIST assessors and organizers. Unlike Year 1, we do not see a high agreement between the two tables in the relative order of the submissions.

Table 5: Performance of automatic runs on the PTKB provenance task based on NIST assessment.

Group	Run ID	Precision	Recall	F1-Measure
guidance	agg_true-qrec-mse	0.8411	0.4932	0.5226
guidance	cosine-orconvqa	0.7983	0.4823	0.5168
guidance	agg_false-qrec-mse	0.7966	0.4517	0.4969
grilllab	grilllab-larf-finetuned	0.7896	0.4073	0.4928
grilllab	grilllab-larf-finetuned-22-rounds	0.7699	0.3980	0.4807
grilllab	grilllab-larf-finetuned-rankllm	0.7726	0.3914	0.4778
grilllab	grilllab-larf-finetuned-10-rounds-10-rounds	0.7259	0.3791	0.4587
organizers	orga-bm25-human	0.6998	0.2887	0.3882
uva	disco-qrecc	0.6998	0.2887	0.3882
uva	mq4cs-gpt41-splade	0.6998	0.2887	0.3882
organizers	orga-bm25-personal	0.6998	0.2887	0.3882
organizers	orga-ance-norerank	0.6998	0.2887	0.3882
organizers	orga-splade-llama70b	0.6998	0.2887	0.3882
organizers	orga-splade-norerank	0.6998	0.2887	0.3882
organizers	orga-bm25-nopersonal	0.6998	0.2887	0.3882
uva	mq4cs-gpt41-bm25	0.6998	0.2887	0.3882
organizers	orga-ance-llama70b	0.6998	0.2887	0.3882
uva	mq4cs-llamaft-splade	0.6998	0.2887	0.3882
cfda	auto_REnpr_npsg20_thru03_d3c5	0.6500	0.2244	0.3054
cfda	auto_npr_npsg20_thru03_d3c5	0.5963	0.2021	0.2795
cfda	auto_npr1_npsg20_thru03_d3c5	0.5907	0.2060	0.2786
cfda	auto_ori_npsg20_thru03_d3c5	0.5974	0.1927	0.2690

Table 6: Performance of automatic runs on the PTKB provenance task based on the organizers’ assessment.

Group	Run ID	Precision	Recall	F1-Measure
grilllab	grilllab-larf-finetuned	0.7407	1.0000	0.8100
grilllab	grilllab-larf-finetuned-10-rounds-10-rounds	0.7123	1.0000	0.7965
grilllab	grilllab-larf-finetuned-rankllm	0.7074	1.0000	0.7875
grilllab	grilllab-larf-finetuned-22-rounds	0.7030	1.0000	0.7872
uva	disco-qrecc	0.5639	0.7333	0.5769
uva	mq4cs-gpt41-splade	0.5639	0.7333	0.5769
uva	mq4cs-llamaft-splade	0.5639	0.7333	0.5769
organizers	orga-bm25-human	0.5639	0.7333	0.5769
organizers	orga-bm25-personal	0.5639	0.7333	0.5769
organizers	orga-splade-norerank	0.5639	0.7333	0.5769
uva	mq4cs-gpt41-bm25	0.5639	0.7333	0.5769
organizers	orga-ance-norerank	0.5639	0.7333	0.5769
organizers	orga-ance-llama70b	0.5639	0.7333	0.5769
organizers	orga-splade-llama70b	0.5639	0.7333	0.5769
organizers	orga-bm25-nopersonal	0.5639	0.7333	0.5769
guidance	agg_true-qrec-mse	0.4714	0.8778	0.5352
cfda	auto_REnpr_npsg20_thru03_d3c5	0.4907	0.6148	0.5022
guidance	agg_false-qrec-mse	0.4271	0.8778	0.5009
guidance	cosine-orconvqa	0.4461	0.8556	0.4994
cfda	auto_npr_npsg20_thru03_d3c5	0.4815	0.5593	0.4855
cfda	auto_npr1_npsg20_thru03_d3c5	0.4722	0.5593	0.4744
cfda	auto_ori_npsg20_thru03_d3c5	0.4704	0.5037	0.4627

5.2.2 Performance per dialogue. Using the organizers’ assessments, in Figure 10 we plotted the mean performance of all the submissions (median and better) in terms of F1-Measure, aggregated on each topic. While we observed a reasonably high performance for all the topics, we find topic 4 to be the most challenging for this task, and 14 to be among the easiest ones.

5.2.3 Performance at different depths. Using the organizers’ assessments, in Figure 11 we plot the mean performance of all the submissions (median and better) in terms of F1-Measure, at varying conversation depths. We noticed a high variance in the performance of different models when the higher conversation depths.

Table 7: Automatic and Generation-only evaluation of response generation. Best is sorted according to Nugget recall.

Group	Run ID	LLMeval		Nugget recall	BEM	F1	ROUGE-1
		SOLAR	GPT-4.1				
NIST	gold-human-response	-	-	**0.2509**	-	-	-
grilllab	grilllab-larf-finetuned-22-rounds	0.9111	0.7778	**0.2507**	0.1759	0.2858	0.2103
grilllab	grilllab-larf-finetuned	**0.9778**	0.8000	0.2418	0.1715	0.2887	0.2138
grilllab	grilllab-larf-finetuned-10-rounds-10-rounds	0.9556	0.8000	0.2203	0.1887	0.2830	0.2079
ucsc	UCSC-base-ensemble	0.9091	0.8222	0.1902	0.1682	**0.3174**	0.2528
ucsc	UCSC-SIMRAG-trained-MiniLM	0.8889	0.7556	0.1870	0.1717	0.3136	0.2555
ucsc	UCSC-base-trained-MiniLM	0.8889	0.8222	0.1867	**0.1990**	0.3126	0.2530
ucsc	UCSC-SIMRAG-ensemble	0.9333	**0.8444**	0.1787	0.1836	0.3168	0.2538
grilllab	grilllab-larf-finetuned-rankllm	0.9556	0.8222	0.1768	0.1649	0.2871	0.2120
uva	mq4cs-llamaft-splade	0.9111	0.7778	0.1510	0.1513	0.3038	0.2561
uva	mq4cs-gpt41-splade	0.9111	**0.8444**	0.1490	0.1822	0.3134	0.2585
uva	mq4cs-gpt41-bm25	0.8444	0.8222	0.1423	0.1616	0.2981	0.2501
uva	disco-qrecc	0.8667	0.8222	0.1308	0.1468	0.3068	0.2595
cfda	auto_ori_npsg20_thru03_d3c5	0.9111	0.7333	0.1145	0.1444	0.2990	0.2541
guidance	cosine-orconvqa	0.7778	0.6667	0.1111	0.1552	0.2707	0.2291
organizers	orga-bm25-nopersonal	0.8000	0.6889	0.1099	0.1507	0.2937	0.2500
guidance	agg_true-qrec-mse	0.7556	0.7111	0.1073	0.1426	0.2804	0.2324
guidance	agg_false-qrec-mse	0.8000	0.6444	0.1041	0.1612	0.2713	0.2299
cfda	auto_REnpr_npsg20_thru03_d3c5	0.9556	0.8000	0.1018	0.1626	0.2929	0.2481
cfda	auto_npr1_npsg20_thru03_d3c5	0.9556	0.7778	0.0991	0.1516	0.2869	0.2467
cfda	auto_npr_npsg20_thru03_d3c5	0.9111	0.8222	0.0972	0.1479	0.2907	0.2471
organizers	orga-bm25-personal	0.7727	0.6222	0.0924	0.1849	0.3110	**0.2676**
organizers	orga-ance-norerank	0.8222	0.6222	0.0863	0.1702	0.2969	0.2580
organizers	orga-splade-llama70b	0.7333	0.6667	0.0830	0.1497	0.3010	0.2651
organizers	orga-bm25-human	0.8222	0.6222	0.0780	0.1813	0.2966	0.2645
organizers	orga-splade-norerank	0.7556	0.5556	0.0614	0.1565	0.2838	0.2425
organizers	orga-ance-llama70b	0.7778	0.6889	0.0576	0.1523	0.2995	0.2625
Generation only							
cfda	gen-only_npsg13_thru0_d4c5	**0.9333**	0.8000	**0.1195**	0.1641	**0.3136**	**0.2689**
uva	nuggets-noptkb	0.9111	**0.8222**	0.1070	0.1721	0.3026	0.2537
guidance	genonly_clariftop10	0.7778	0.6889	0.1041	0.1650	0.2799	0.2306
uva	nuggets-ptkb	0.8667	0.7778	0.1030	**0.1923**	0.3052	0.2524
genaius	genaius-genonly-full-gpt4o	0.8889	0.7556	0.0999	0.1672	0.2827	0.2485
cfda	gen-only_npsg20_thru03_d3c5	0.9111	0.7778	0.0978	0.1524	0.3065	0.2552
genaius	genaius-genonly-summary-gpt4o	0.8667	0.7556	0.0811	0.1407	0.2750	0.2500
usiir	usiir_run2	0.4000	0.2222	0.0510	0.1267	0.1877	0.1708
usiir	usiir_run1	0.6000	0.3111	0.0508	0.1186	0.1916	0.1740

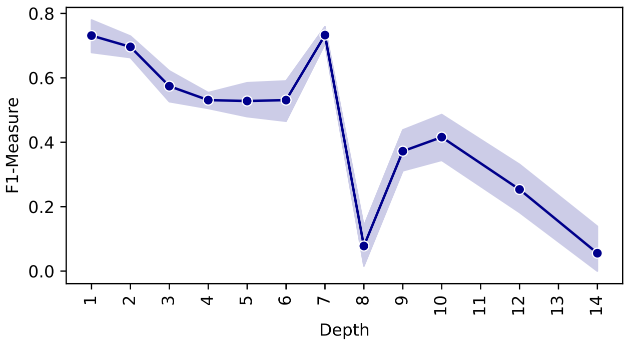


Figure 11: F1-Measure on PTKB relevance prediction at varying conversation turn depths (mean and std across runs).

Intuitively, we see the highest performance at depth 1 as the dialogues are simpler and the performance generally goes down at deeper turns.

5.3 Response Evaluation

We present in Table 7 the results of the response generation task. Overall, the runs with high retrieval performance also perform well on the generation task. Across metrics, we can observe the trend that SOLAR seems to overestimate the quality of the response compared to GPT-4.1. Note that the LLMEval metric prompts LLMs to compare the generated response with the human-written response, and not to assess the generated response based on its internal knowledge. Overall, the recall on nuggets is also low compared to LLMEval. This might be due to the creation process of the nuggets,

Table 8: Evaluation results of runs in the interactive task based on human assessments. On the rubric level, assessor evaluated engagement (Eng), relevance (Rel), quality (Qual), and their confidence in the ratings (Conf). On the dialogue level, assessors rated mixed-initiative strategies (Mix), personalization (Pers), information flow (Flow), trustworthiness (Trust), user satisfaction (Sat), and their confidence in these ratings (Conf).

Run ID	Rubric Level					Dialogue Level							Score
	Eng	Rel	Qual	Conf	Score	Mix	Pers	Flow	Trust	Sat	Conf	Score	
orga-no-no-no-gpt41mini	**0.81**	**0.85**	**0.85**	0.96	**0.84**	0.45	**0.92**	**0.97**	0.47	**0.82**	0.94	**0.78**	**0.79**
genaius-full-rewrite	0.71	0.80	0.82	0.94	0.78	0.46	0.59	0.93	0.72	**0.82**	0.94	**0.78**	0.77
genaius-summary-rewrite	0.72	0.80	**0.85**	0.95	0.81	0.48	0.59	0.91	0.76	0.78	0.90	0.72	0.75
cfda-adarewriter-chiq-llm4cs-splade	0.64	0.81	0.84	0.94	0.78	0.33	0.46	0.91	0.72	0.81	0.92	0.75	0.74
cfda-chiq-llm4cs-splade-rrf	0.63	0.79	0.81	0.95	0.77	0.31	0.41	0.94	0.71	0.76	0.94	0.72	0.74
grilllab-agentic-gpt4.1-larf	0.67	0.73	0.74	0.89	0.68	0.64	0.55	0.78	**0.82**	0.76	0.92	0.72	0.69
uva-gpt5mini-bm25-debertav3-gpt5mini	0.65	0.73	0.73	0.95	0.71	0.42	0.57	0.74	0.65	0.76	0.86	0.69	0.68
grilllab-agentic-gpt4.1-larf-v2	0.71	0.75	0.79	0.90	0.73	**0.81**	0.62	0.82	0.76	0.72	0.84	0.64	0.67
uva-gpt5mini-no-no-gpt5mini	0.62	0.72	0.72	0.92	0.68	0.46	0.57	0.74	0.49	0.72	0.88	0.66	0.66
orga-llama8b-bm25-minilm-llama8b-v2	0.60	0.68	0.70	0.94	0.67	0.25	0.37	0.66	0.63	0.72	0.94	0.68	0.66
uva-gpt5-bm25-debertav3-gpt5mini-nopersonal	0.51	0.66	0.67	0.92	0.63	0.33	0.24	0.68	0.66	0.71	0.92	0.67	0.64
grilllab-agentic-gpt4.1	0.68	0.69	0.74	0.92	0.70	0.65	0.54	0.84	0.81	0.69	0.88	0.63	0.63
ucsc-SIMRAG-keywordQuery-dynamicPTKB-ensembleReranker	0.58	0.65	0.67	**0.97**	0.67	0.24	0.38	0.79	0.66	0.63	0.96	0.61	0.62
ucsc-SIMRAG-guidelineQuery-dynamicPTKB-trainedReranker	0.54	0.60	0.63	0.93	0.61	0.25	0.40	0.74	0.69	0.68	0.94	0.64	0.62
ucsc-SIMRAG-keywordQuery-dynamicPTKB-trainedReranker	0.58	0.64	0.66	0.93	0.64	0.33	0.37	0.88	0.69	0.66	0.94	0.63	0.61
orga-gpt41mini-bm25-minilm-llama70b	0.58	0.65	0.65	0.92	0.62	0.42	0.50	0.74	0.63	0.66	0.94	0.63	0.61
ucsc-base-dynamicPTKB-trainedReranker	0.55	0.60	0.62	0.95	0.59	0.22	0.42	0.74	0.63	0.62	0.92	0.58	0.57
orga-no-no-no-llama70b	0.63	0.63	0.63	0.90	0.59	0.25	0.56	0.68	0.44	0.57	0.86	0.52	0.53
uva-gpt5-bm25-debertav3-gpt5	0.44	0.58	0.58	0.94	0.55	0.19	0.43	0.44	0.60	0.57	0.90	0.52	0.52
orga-llama70b-bm25-minilm-llama70b	0.54	0.59	0.59	0.92	0.56	0.37	0.54	0.60	0.66	0.60	0.82	0.52	0.51
orga-no-no-no-llama8b-v2	0.56	0.61	0.61	0.89	0.57	0.36	0.54	0.63	0.47	0.59	0.80	0.49	0.50
orga-gpt41mini-bm25-minilm-llama70b-nopersonal	0.52	0.55	0.58	0.90	0.53	0.27	0.23	0.62	0.50	0.47	0.84	0.39	0.42
grilllab-larf-fine-tuned-judge	0.49	0.49	0.54	0.74	0.44	0.33	0.59	0.62	0.74	0.60	0.59	0.36	0.38

which include lots of sub-pieces of information from a large set of passages.

5.4 Interactive Task

5.4.1 Human assessment results. Table 8 presents the results of the interactive task using simulated users and human rubric annotations.

Baselines not using RAG (e.g., orga-no-no-no-gpt41mini and all orga-no-no-no-[LLM] variants) show strong engagement and high overall scores, but low trust (<0.50), indicating that humans detect hallucinations.

Overall, we can observe that better retrieval leads to better generation, with comparisons like cfda-adarewriter-chiq-llm4cs-splade that uses splade backbone retrieval in the high ranges of scores.

A few runs without personalization underperform in comparison to those integrating personalization, for example, uva-gpt5-bm25-debertav3-gpt5mini-nopersonal vs. uva-gpt5mini-bm25-debertav3-gpt5mini, or orga-gpt41mini-bm25-minilm-llama70b-nopersonal vs. orga-gpt41mini-bm25-minilm-llama70b, both with personalization scores below 0.25. Also, the orga-no-no-no-gpt41mini run seems to have higher personalization compared to other runs. This could be caused by the LLM not being "distracted" by the retrieved passages and can focus more on the PTKB itself.

The grilllab-agentic-gpt4.1-larf-v2 run has the highest score in terms of mixed-initiative, showing an example of a proactive system, and showing that our simulated user is able to discuss with interactive systems. Also, this same run has a high engagement score, with 0.71, which is in the higher range of scores for engagement.

Considering the runs orga-gpt41mini-bm25-minilm-llama70b and orga-gpt41mini-bm25-minilm-llama70b-nopersonal (denoted as orga-bm25-personal and orga-bm25-nopersonal in the offline task), we can see that the user simulator gives a similar ranking, with the former consistently higher on both tasks. We even observe a more important gap in the interactive task, showing that the simulator might be more expressive in discriminating runs. Overall scores for the interactive task are in the range of 0.90-0.50, while automatic evaluation of the generation showed lower standard deviation between runs, and similarly for the retrieval task.

5.4.2 Passage Relevance Results. Table 9 presents the evaluation results of the cited passages of the participant runs on the basis of pooled qrels from the offline task. Due to this pooling of passages from different runs, the retrieval results appear to be substantially lower than in the offline task. In general, the SIMRAG-based keyword query generation pipeline of ucsc obtained the best nDCG@3 and nDCG@5 performance out of all runs with the guideline query alternative falling further behind. Considering the full set of retrieved documents the base variant of the ucsc runs, obtained the best nDCG and recall scores. Therefore, the runs of the ucsc team seemed to exhibit the most effective retrieval pipelines out of all runs. However, the cfda and grillab runs do not fall far behind in terms of nDCG@3.

Table 9: Evaluation results of document citations based on rubric-level qrels from the offline task.

Group	Run ID	nDCG@3	nDCG@5	nDCG	P@20	Recall@20	Recall	mAP
ucsc	ucsc-SIMRAG-keywordQuery-dynamicPTKB-ensembleReranker	**0.2766**	**0.2870**	0.2146	**0.3788**	**0.0988**	0.1951	0.1103
ucsc	ucsc-SIMRAG-keywordQuery-dynamicPTKB-trainedReranker	0.2491	0.2365	0.3423	0.2833	0.0741	0.4820	0.1320
cfda	cfda-adarewriter-chiq-llm4cs-splade	0.2367	0.2076	0.0914	0.2061	0.0575	0.0601	0.0349
ucsc	ucsc-base-dynamicPTKB-trainedReranker	0.2351	0.2216	**0.3520**	0.2909	0.0772	**0.5070**	**0.1356**
grillab	grilllab-agentic-gpt4.1-larf	0.2159	0.1855	0.0525	0.0773	0.0234	0.0234	0.0212
grillab	grilllab-agentic-gpt4.1	0.2121	0.1759	0.0530	0.0803	0.0230	0.0230	0.0200
organizers	orga-llama8b-bm25-minilm-llama8b-v2	0.1919	0.1758	0.0436	0.0712	0.0181	0.0181	0.0146
cfda	cfda-chiq-llm4cs-splade-rrf	0.1881	0.1847	0.0761	0.1864	0.0464	0.0464	0.0304
grilllab	grilllab-agentic-gpt4.1-larf-v2	0.1847	0.1732	0.0474	0.0682	0.0212	0.0212	0.0179
ucsc	ucsc-SIMRAG-guidelineQuery-dynamicPTKB-trainedReranker	0.1844	0.1919	0.3351	0.2652	0.0765	0.4883	0.1288
organizers	orga-llama70b-bm25-minilm-llama70b	0.1821	0.1672	0.0434	0.0803	0.0179	0.0179	0.0131
organizers	orga-gpt41mini-bm25-minilm-llama70b-nopersonal	0.1645	0.1717	0.0538	0.1061	0.0296	0.0296	0.0212
genaius	genaius-full-rewrite	0.1568	0.1302	0.0350	0.0500	0.0129	0.0129	0.0109
genaius	genaius-summary-rewrite	0.1323	0.1052	0.0308	0.0409	0.0138	0.0138	0.0118
uva	uva-gpt5-bm25-debertav3-gpt5mini-nopersonal	0.1162	0.1074	0.0310	0.0455	0.0149	0.0149	0.0102
organizers	orga-gpt41mini-bm25-minilm-llama70b	0.1111	0.1131	0.0366	0.0879	0.0238	0.0238	0.0133
uva	uva-gpt5-bm25-debertav3-gpt5	0.0906	0.0991	0.0303	0.0409	0.0140	0.0140	0.0101
uva	uva-gpt5mini-bm25-debertav3-gpt5mini	0.0680	0.0672	0.0203	0.0318	0.0086	0.0086	0.0050
grillab	grilllab-larf-fine-tuned-judge	0.0457	0.0587	0.0287	0.0727	0.0158	0.0211	0.0085
organizers	orga-no-no-no-llama8b-v2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
organizers	orga-no-no-no-llama70b	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
organizers	orga-no-no-no-gpt41mini	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
uva	uva-gpt5mini-no-no-gpt5mini	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

6 CONCLUSION

The third TREC iKAT edition built on the first two years and developed resources for studying personalized conversational information seeking, and added to the community's understanding of the topic. iKAT Y3 introduced a novel interactive task, where the systems would interact with a live simulated user on the same topics that were available to them in the offline task, allowing us to study and compare systems' performance across offline and interactive tasks. The PTKB statement ranking task provided a way for participants to leverage users' personal information into the conversation, and we saw their impact on the interactive task as well, as per human assessment.

7 ACKNOWLEDGMENTS

We are grateful to Jamie Callan's great help and patience in enabling us to provide the participants with a sample of the ClueWeb'22 collection. We are thankful for Ian Soboroff's experience, patience, and persistence in running the assessment process. We are also thankful to Andrew Ramsay for his help throughout the project. Finally, we thank all our participants and annotators.

REFERENCES

[1] Abbasiantaeb, Z., Lupart, S., Aliannejadi, M.: Generating multi-aspect queries for conversational search (28 Mar 2024), http://arxiv.org/abs/2403.19302

[2] Abbasiantaeb, Z., Lupart, S., Azzopardi, L., Dalton, J., Aliannejadi, M.: 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. p. 3455–3465. SIGIR '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3726302.3730316, https://doi.org/10.1145/3726302.3730316

[3] Aliannejadi, M., Abbasiantaeb, Z., Chatterjee, S., Dalton, J., Azzopardi, L.: TREC iKAT 2023: The interactive knowledge assistance track overview. CoRR **abs/2401.01330** (2024)

[4] Aliannejadi, M., Abbasiantaeb, Z., Chatterjee, S., Dalton, J., Azzopardi, L.: TREC iKAT 2023: A test collection for evaluating conversational and interactive knowledge assistants. In: SIGIR. pp. 819–829. ACM (2024)

[5] Aliannejadi, M., Abbasiantaeb, Z., Lupart, S., Chatterjee, S., Dalton, J., Azzopardi, L.: Trec ikat 2024: The interactive knowledge assistance track overview. In: The Thirty-Third Text REtrieval Conference Proceedings (TREC 2024), Gaithersburg, MD, USA, November 15-18, 2024. NIST Special Publication, vol. 1329. National Institute of Standards and Technology (NIST) (2024), https://trec.nist.gov/pubs/trec33/papers/Overview_ikat.pdf

[6] Anantha, R., Vakulenko, S., Tu, Z., Longpre, S., Pulman, S., Chappidi, S.: Open-domain question answering goes conversational via Question Rewriting (10 Oct 2020), http://arxiv.org/abs/2010.04898

[7] Azzopardi, L., Dubiel, M., Halvey, M., Dalton, J.: Conceptualizing agent-human interactions during the conversational search process. In: The second international workshop on conversational approaches to information retrieval (2018)

[8] Bulian, J., Buck, C., Gajewski, W., Boerschinger, B., Schuster, T.: Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. arXiv preprint arXiv:2202.07654 (2022)

[9] Farzi, N., Dietz, L.: Pencils Down! Automatic Rubric-based Evaluation of Retrieve/Generate Systems. In: Oosterhuis, H., Bast, H., Xiong, C. (eds.) Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2024, Washington, DC, USA, 13 July 2024. pp. 175–184. ACM (2024). https://doi.org/10.1145/3664190.3672511, https://doi.org/10.1145/3664190.3672511

[10] Hai, N.L., Gerald, T., Formal, T., Nie, J.Y., Piwowarski, B., Soulier, L.: CoSPLADE: Contextualizing SPLADE for conversational information retrieval (11 Jan 2023), http://arxiv.org/abs/2301.04413

[11] Huang, Y., Chen, Y., Cao, X., Yang, R., Qi, M., Zhu, Y., Han, Q., Liu, Y., Liu, Z., Yao, X., Jia, Y., Ma, L., Zhang, Y., Zhu, T., Zhang, L., Chen, L., Chen, W., Zhu, M., Xu, R., Zhang, L.: Towards large-scale generative ranking (7 May 2025), http://arxiv.org/abs/2505.04180

[12] Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)

[13] Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., et al.: Gemma 3 technical report. CoRR **abs/2503.19786** (2025). https://doi.org/10.48550/ARXIV.2503.19786, https://doi.org/10.48550/arXiv.2503.19786

[14] Kim, D., Park, C., Kim, S., Lee, W., Song, W., Kim, Y., Kim, H., Kim, Y., Lee, H., Kim, J., et al.: Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. arXiv preprint arXiv:2312.15166 (2023)

[15] Lai, Y., Wu, J., Wang, Z., Zhou, D.: AdaRewriter: Unleashing the power of prompting-based conversational query reformulation via test-time adaptation

(28 Oct 2025). https://doi.org/10.48550/arXiv.2506.01381, http://arxiv.org/abs/2506.01381

[16] Lupart, S., Aliannejadi, M., Kanoulas, E.: Disco: Llm knowledge distillation for efficient sparse retrieval in conversational search (2025), https://arxiv.org/abs/2410.14609

[17] Mao, K., Dou, Z., Mo, F., Hou, J., Chen, H., Qian, H.: Large language models know your contextual search intent: A prompting framework for conversational search (12 Mar 2023), http://arxiv.org/abs/2303.06573

[18] Mo, F., Ghaddar, A., Mao, K., Rezagholizadeh, M., Chen, B., Liu, Q., Nie, J.Y.: CHIQ: Contextual history enhancement for improving query rewriting in conversational search. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 2253–2268. Association for Computational Linguistics, Stroudsburg, PA, USA (2024). https://doi.org/10.18653/v1/2024.emnlp-main.135, https://aclanthology.org/2024.emnlp-main.135.pdf

[19] OpenAI: GPT-4 technical report. CoRR **abs/2303.08774** (2023). https://doi.org/10.48550/ARXIV.2303.08774, https://doi.org/10.48550/arXiv.2303.08774

[20] Owoicho, P., Dalton, J., Aliannejadi, M., Azzopardi, L., Trippas, J.R., Vakulenko, S.: Trec cast 2022: Going beyond user ask and system retrieve with initiative and response generation. In: Proceedings of the NIST Text Retrieval Conference, TREC 2022. pp. 1–11 (2023)

[21] Qu, C., Yang, L., Chen, C., Qiu, M., Croft, W.B., Iyyer, M.: Open-retrieval conversational question answering (22 May 2020), http://arxiv.org/abs/2005.11364

[22] Radlinski, F., Craswell, N.: A theoretical framework for conversational search. In: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. pp. 117–126. ACM (2017)

[23] Rau, D., Déjean, H., Chirkova, N., Formal, T., Wang, S., Nikoulina, V., Clinchant, S.: Bergen: A benchmarking library for retrieval-augmented generation. arXiv preprint arXiv:2407.01102 (2024)

[24] Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 3980–3990. Association for Computational Linguistics (2019). https://doi.org/10.18653/V1/D19-1410, https://doi.org/10.18653/v1/D19-1410

[25] Russell-Rose, T., Chamberlain, J., Azzopardi, L.: Information retrieval in the workplace: A comparison of professional search practices. Information Processing & Management **54**(6), 1042–1057 (2018)

[26] Yang, D., Zeng, L., Rao, J., Zhang, Y.: Knowing you don't know: Learning when to continue search in multi-round RAG through self-practicing. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1305–1315. ACM, New York, NY, USA (13 Jul 2025). https://doi.org/10.1145/3726302.3730018, http://dx.doi.org/10.1145/3726302.3730018

[27] Ye, F., Fang, M., Li, S., Yilmaz, E.: Enhancing conversational search: Large language model-aided informative query rewriting. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 5985–6006. Association for Computational Linguistics, Stroudsburg, PA, USA (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.398, http://dx.doi.org/10.18653/v1/2023.findings-emnlp.398