UvA-IRLab at iKAT25: Exploring Learned Sparse Retrieval and Query Rewriting for Personalized Conversational QA

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

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

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

Abstract

The TREC interactive Knowledge Assistant Track (iKAT) 2025 is the third edition of the iKAT shared task. It focuses on developing conversational assistants that can adapt their responses using personal user knowledge from a Personal Textual Knowledge Base (PTKB). This year's edition also introduces a new interactive task that evaluates systems using a user simulator. Since query rewriting is an effective way to handle conversational context, we study the use of Large Language Models (LLMs) as query rewriters. In our runs, we generate multiple query aspects using the MQ4CS framework and frontier LLMs (GPT-4.1), as well as open-source LLMs finetuned for the task (Llama-8B). We also strengthen the approach with SPLADE-based sparse retrieval and cross-encoder reranking. Finally, we also explore a rewrite-free technique, based on learned sparse retrieval (LSR) using the DiSCo model. Our results show that multi-aspect query generation improves performance when paired with strong retrieval and reranking models. They also suggest that LLM-based query rewriting can support better personalization in conversational search.

Keywords

conversational search, personalization, retrieval augmented generation, query understanding, neural sparse retrieval

1 Introduction

TREC iKAT 2025. Conversational search (CS) has seen important developments recently, along with a need for systems that can engage in meaningful, context-aware interactions [6]. The TREC interactive Knowledge Assistant Track (iKAT) 2025 specifically addresses these needs for the third consecutive year [2, 4, 5], by developing conversational assistants capable of discussing a wide range of topics with users. These assistants are not only designed to engage in dialogue but also to support users in making informed decisions by providing contextually relevant, personalized responses and dynamically ranked information within each conversational exchange. The TREC iKAT achieves this by leveraging two main sources of knowledge: a broad knowledge collection and a Personal Text Knowledge Base (PTKB), which contains user-specific information. This PTKB captures individual preferences, constraints and prior experiences of the user, all in a textual format. By integrating user-centric data, the track makes the retrieval and response generation processes significantly more challenging, as systems must adapt to each user's unique context and preferences. TREC iKAT includes several subtasks: passage ranking, PTKB classification, and response generation. The track also proposes different submission formats: (1) an automatic task, where all evaluation is done offline, (2) a generation-only task with fixed retrieval, and (3) an interactive task, which includes a Large Language Model (LLM) as user simulator to explore the dialog-level behavior of systems.

Multi-Aspect Query Generation. Recent studies in conversational search have demonstrated that LLMs are effective in tackling such tasks [1, 16–19], disambiguating the user's latest utterance in the conversation. Building on this, our submitted runs explore LLM-based multi-aspect query rewriting techniques to resolve conversational ambiguities. Our work leverages the recent *MQ4CS* framework [1] with frontier and open-source LLMs.

Rewrite-Free Conversational Search. Another line of work in conversational search focuses on efficient retrieval using rewrite-free methods [8, 21]. We also try in our submission to use the existing *DiSCo* model [15], built upon a learned sparse retrieval (LSR) model [7, 20] to tackle the task.

Contributions. Through our runs, we demonstrate that LLMs are highly effective when used for query rewriting in this task. At the same time, we show that pipelines without an LLM-based rewriting step can still achieve competitive performance, while being much more efficient. Finally, we report strong results with LSR models.

2 Task Description

Personalization, Retrieval and Generation. As in the previous two years, iKAT 2025 provides user utterances, conversation history, and Personal Text Knowledge Base (PTKB) as input for each conversational turn. The conversation history includes the user's previous utterances and gold responses from the system. TREC iKAT 2025 benchmark includes 188 turns across 17 topics, of which 45 turns and 12 topics were assessed. Each topic was paired with a unique user persona (i.e., PTKB) of 20 descriptive textual statements on average per persona. Passage ranking is done on the ClueWeb-iKAT collection, with each conversational turn evaluated independently. The iKAT subtasks are as follows:

(1) *Passage Ranking*: aims to retrieve and rank the relevant passages from the given collection in response to each user utterance.

*Equal contribution.

TREC'25, November 2025, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Experimental results on the official evaluation set. Bold and underlined indicate the best and second-best results across both Automatic offline evaluation with threshold ≥ 1 for relevance.

Runs	LLM	Retrieval	Reranker	nDCG@5	nDCG	P@20	Recall@1k	mAP
(1) MQ4CS-gpt41-bm25	GPT-4.1	BM25	DebertaV3	0.4053	0.4126	0.4756	0.4481	0.2139
(2) MQ4CS-gpt41-splade	GPT-4.1	SPLADE	DebertaV3	**0.4897**	**0.5357**	**0.5789**	0.6101	**0.3099**
(3) MQ4CS-llamaft-splade	Llama-8B	SPLADE	DebertaV3	0.3680	0.3911	0.4756	0.4413	0.2045
(4) DiSCo-qrecc-norerank	Rewrite-Free	SPLADE	–	0.3042	0.3926	0.3544	0.5524	0.1683
(o2) orga-splade-norerank	GPT-4o-mini	SPLADE	–	0.3498	0.4719	0.4000	**0.6760**	0.2150
(o3) orga-splade-llama70b	Llama-70B	SPLADE	MiniLM	0.3676	0.4816	0.4122	0.6721	0.2320
(o1) orga-bm25-personal	GPT-4o-mini	BM25	MiniLM	0.3902	0.3779	0.4178	0.4308	0.1871

To understand the user query for each turn, the system must reason over relevant PTKBs and the conversation history.

(2) *Response Generation*: The goal of this task is to provide a natural answer to the user utterance, following the flow of the conversation, and based on the retrieved passages. The responses must be grounded from the retrieved passages from the collection. Hence, the responses must be generated using at least one passage, referred to as "provenance", from the source collection. The organizer also proposed a reference generation-only submission task, where retrieval was fixed by the organizer and only the generation was made by participant teams.

(3) *PTKB Classification*: the PTKB consists of several statements. The goal of this task is to classify the statements within the PTKB as relevant/irrelevant for each conversational turn.

Offline and Interactive Mode. Besides, the organizers proposed different types of task: first the automatic offline task (including the three subtasks all at once), then an interactive task, where users are simulated via LLM, and with whom participants interacted through an API. This task is much harder, as an error at the beginning of the conversation can directly cause misunderstandings in subsequent user utterances. It differs from the offline task, where each turn is re-written with gold past history. The retrieval model must thus correctly understand the history, removing noise from it much more consistently.

3 Methodology

3.1 Passage Ranking

We explore two families of models in our runs, (1) a *multi-aspect query rewriting* approach, leveraging LLMs for disambiguation; (2) and a *rewrite-free* method that directly creates representations from the whole conversation without any rewriting step.

Multi-Aspect Query Generation (MQ4CS). We rely heavily on LLMs to disambiguate the conversational context and persona of the user. In particular, we use the MQ4CS framework proposed by [1], generating multiple queries each covering different aspects of the information need. MQ4CS proposes breaking the user information need into multiple queries rather than forming a single query rewrite that includes the complete information need. This allows for better coverage of the collections, similar to query expansion, but also enables a decomposition of the query into several sub-pieces of simpler information needs. Queries are generated for each user utterance by reasoning over both the context and personal information about the user provided in PTKB. Retrieval and re-ranking are applied to each query independently and normalized with min-max normalization before being merged.

We extend the original work on MQ4CS that uses term-based and dense retrieval with learned sparse retrieval, similar to the run we submitted during the last year [14]. Our runs use SPLADE, which adapts the MLM pre-trained layer of BERT to predict term importance and token expansion within the vocabulary space [7].

Rewrite-Free Conversational Search (DiSCo). We also experiment with DiSco [15], an efficient conversational search approach that avoids LLM-based rewriting by directly encoding the dialogue into the embedding space. DiSCo learns to capture user intent from the full conversation through knowledge distillation and integrates LSR via a SPLADE model. Such rewrite-free approaches are more efficient, as they avoid the sequential generative decoding step required by the LLM for rewriting.

3.2 Response Generation

Retrieval Augmented Generation. For response generation, we rely on the simple yet effective Retrieval-Augmented Generation (RAG) pipeline using LLMs on top of retrieved passages [13]. In this pipeline, the model is given the user utterance, context of the conversation, PTKB of the user, and the top passages retrieved from our passage ranking. Each submission follows this pipeline, with a different top ranking. We use a similar prompt to Abbasiantaeb et al., including all PTKB within the prompt. This approach was effective for the last year of iKAT.

Nugget-based Generation. For the gen-only runs, we adopt a nugget-based generation strategy. We first extract the relevant nuggets from the top 20 retrieved passages and then generate the final answer conditioned on these nuggets. This approach aims to produce a complete response with broad coverage, ensuring that different angles and aspects of the required information are addressed. This approach is inspired by the recent work on nugget-based response generation in single-turn QA [11].

3.3 PTKB Classification

In both previous tasks (Ranking and Response Generation), all PTKBs are included in the model's input. We thus entirely rely on LLMs to filter and extract relevant information from the user in

Table 2: 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**	-	-	-
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
organizers	orga-bm25-nopersonal	0.8000	0.6889	0.1099	0.1507	0.2937	0.2500
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							
uva	nuggets-noptkb	**0.9111**	**0.8222**	**0.1070**	0.1721	0.3026	0.2537
uva	nuggets-ptkb	0.8667	0.7778	0.1030	**0.1923**	0.3052	0.2524
organizers	orga-bm25-personal	0.7727	0.6222	0.0924	0.1849	**0.3110**	**0.2676**

different contexts. However, to better understand and make the model's reasoning more explicit, we also prompted the model to classify relevant statements from the PTKB list. In particular, we follow the zero-shot approach proposed by [3] for iKAT 2023, where an LLM is prompted to return the relevant PTKB given the user utterance and context of the conversation, with an adaptation for classification.

As an alternative approach, one could only use those filtered PTKB for the context modeling and response generation tasks, breaking down the personalization task.

4 List of Submissions

We provide below the list of submitted runs with detailed descriptions for each subtask. We also link model checkpoints we use.

Automatic Runs:

(1) *MQ4CS-gpt41-bm25*: This run uses the MQ4CS framework, generating multiple queries (ϕ=5, GPT-4.1) for our first-stage BM25 retrieval. For reranking, we rerank each top 1000 with a DebertaV3 cross-encoder, and aggregate the five ranking lists with a min-max normalization. Both PTKB classification and response generation models use GPT-4.1 as LLM with the top 5 retrieved passages. All prompts are zero-shot.

(2) *MQ4CS-gpt41-splade*: Similar to the first run, this submission relies on MQ4CS. The entire pipeline is similar except for retrieval, which uses SPLADE. Reranking uses DebertaV3. Response generation and PTKB classification use the same method as in the previous run, with GPT-4.1.

(3) *MQ4CS-llamaft-splade*: For this run, we used Llama-8B query rewrites, using a Llama-8B model finetuned for query rewriting. This approach is similar to the one described in the original MQ4CS paper. Reranking is still the same. Response generation and PTKB classification use the same method as in the previous runs, with GPT-4.1.

(4) *disco-qrecc-norerank*: This run uses the open-source DiSCo model without query rewriting. It is not reranked and does not use PTKBs.

Generation Only Runs:

(5) *nuggets*: This run takes the top retrieved passages and decomposes them into nuggets, the decomposition also acts as a filtering step on relevant parts of each retrieved passage. Then it uses these nuggets to formulate a response.

(6) *nuggets-noptkb*: This run is similar but doesn't integrate the PTKB in both the nugget's extraction and the response generation.

Interactive Runs:

(7) *uva-gpt5mini-bm25-debertav3-gpt5mini* : This interactive run uses a single GPT-5 mini rewrite for both retrieval and reranking. It relies on BM25 and DebertaV3. Response generation and PTKB classification are added afterward from GPT-5 mini.

(8) *uva-gpt5-bm25-debertav3-gpt5*: Similar to the first interactive run, this submission uses RAG, but with GPT-5 as the backbone model.

(9) *uva-gpt5mini-no-no-gpt5mini*: We wanted to explore with this run a direct generation without retrieval, to compare engagement and flow. This run uses GPT-5 mini for direct generation based on conversation history and PTKBs.

(10) *uva-gpt5-bm25-debertav3-gpt5mini-nopersonal*: This run uses retrieval for grounding, and an LLM for the final generation in a RAG fashion; however, no PTKBs are integrated here.

Hyperparameters. Our retrieval mostly relies on SPLADE, using the CoCondenser SelfDistil SPLADE++ checkpoint from HuggingFace[1], also shared by the organizers. For re-ranking, we use

[1] naver/splade-cocondenser-selfdistil

Table 3: 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 dialog 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					Dialog Level							Score
[team]-[rewrite]-[retrieval]-[rerank]-[answer generation]	Eng	Rel	Qual	Conf	Score	Mix	Pers	Flow	Trust	Sat	Conf	Score	
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**
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
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
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-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**
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
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
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
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

a debertav3-large reranker [9, 10], fine-tuned by Lassance and Clinchant for re-ranking[2]. We also rely on the MS MARCO MiniLM cross-encoder[3]. For DiSCO we use the open source checkpoint available on HuggingFace [4].

5 Results

In the section below, we present our retrieval results and main observations from the submitted runs.

5.1 Retrieval Performance

Multi-Aspect Query Generation. Table 1 contains the main retrieval results of our four automatic runs with three baselines from the organizers. Comparing lines (1) and (2), we first see the gains from SPLADE compared to a BM25 backbone retrieval model. This shows that even with query augmentation and several rewrites, a better retrieval backbone can improve performance further. Comparing line (3) that uses a smaller scale LLM for query rewrite with Llama 8B, we observe a substantial drop in performance compared to frontier LLMs in line (2). Lines (1) and (3) show similar performance, suggesting that one can compensate for a poor retrieval model, such as BM25, with a strong LLM reformulation.

Rewrite-Free Conversational Search. Line (4) reports the performance of DiSCo, a SPLADE-based first-stage retrieval model. Since this run does not apply reranking, Recall@1k is the most meaningful point of comparison. It also avoids any LLM-based rewriting, making it considerably more efficient than the other approaches. DiSCo remains competitive with *MQ4CS-Llama 8B (3)* and with *orga-bm25-personal (o1)*. However, despite its strong recall and broad coverage, it still falls short on precision-oriented metrics, highlighting the benefit of query reformulation.

5.2 Response Generation Performance

Table 2 presents the response generation results. Overall, we observe a consistent relationship between retrieval quality and generation quality: systems with stronger first-stage retrieval in the automatic setting tend to achieve higher LLM-based evaluation scores and better nugget coverage. Among our submissions, mq4cs-gpt41-splade provides the best overall performance, achieving the highest GPT-4.1 LLM Eval score as well as the strongest BEM and F1 results, while remaining tied for best on SOLAR. This highlights the benefit of combining high-quality reformulation with effective sparse retrieval.

The *generation-only* runs allow us to focus specifically on the generation component while holding retrieval fixed. All models in this category use the same set of retrieved passages; differences therefore stem solely from how nuggets are extracted and how responses are generated from those nuggets. Both nuggets-noptkb and nuggets-ptkb score competitively on LLM Eval metrics, approaching the performance of several full retrieval-and-generation pipelines. Nugget recall is lower than in the automatic MQ4CS runs (approximately 0.10 vs. 0.15), but the generated responses still achieve F1 and ROUGE-1 scores comparable to other runs. Adding PTKB information leads to a small improvement in BEM but does not consistently improve other metrics.

5.3 Interactive Task Performance

Table 3 reports the human evaluation results for the interactive task. First, note that because this task is interactive (using simulated users), it allows us to evaluate systems at the dialog level, while the offline task was limited to single turns without possible interaction or mixed initiatives. Overall, two main patterns emerge from our submissions. First, the uva-gpt5mini-no-no-gpt5mini run, despite using no retrieval, performs surprisingly well across most rubric- and dialog-level dimensions. Its scores on Relevance, Quality, and Flow are close to those of our full pipeline, and it achieves the strongest Mixed-Initiative score among our runs. The main weakness appears in the Trust dimension, where RAG-based

[2] naver/trecdl22-crossencoder-debertav3

[3] cross-encoder/ms-marco-MiniLM-L-6-v2

[4] slupart/splade-disco-human

systems perform better, suggesting that the model's free-form responses lack factual grounding.

Second, personalization has a clear impact. The `uva-gpt5-bm25-debertav3-gpt5mini-[nopersonal]` run shows a marked drop in the Personalization score, even though other dimensions remain comparable. This aligns with the organizers' baselines, where removing personalization consistently reduces Pers while leaving other metrics mostly stable.

Across all runs, GPT-5–based models lag behind GPT-5-mini variants, particularly on dialog-level metrics (Mix, Flow, Sat). This suggests that GPT-5 may be less aligned with conversational adaptation and natural-language understanding, performing better on agentic or tool-augmented tasks than on multi-turn interaction.

Overall, retrieval and personalization still matter, but the interactive evaluation highlights that model alignment for dialogue plays a central role in this task.

6 Conclusion

This report aims to describe the technical details of our iKAT 2025 submission. Through our submission, we demonstrate how the MQ4CS framework generalizes to iKAT 2025, and show the additional gains from combining it with learned sparse retrieval and more effective cross-encoder models. We also compare it to efficient conversational search models to analyze the trade-off between efficiency and effectiveness.

References

[1] Zahra Abbasiantaeb, Simon Lupart, and Mohammad Aliannejadi. 2024. Generating Multi-Aspect Queries for Conversational Search. *arXiv preprint arXiv:2403.19302* (2024).

[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* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 3455–3465. https://doi.org/10.1145/3726302.3730316

[3] Zahra Abbasiantaeb, Chuan Meng, David Rau, Antonis Krasakis, Hossein A Rahmani, and Mohammad Aliannejadi. 2023. LLM-based Retrieval and Generation Pipelines for TREC Interactive Knowledge Assistance Track (iKAT) 2023. (2023).

[4] Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. 2024. TREC iKAT 2023: A Test Collection for Evaluating Conversational and Interactive Knowledge Assistants. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 819–829. https://doi.org/10.1145/3626772.3657860

[5] Mohammad Aliannejadi, Zahra Abbasiantaeb, Simon Lupart, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. 2024. 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). https://trec.nist.gov/pubs/trec33/papers/Overview_ikat.pdf

[6] Aaron Chatterji, Tom Cunningham, David J. Deming, Zoë Hitzig, Christopher Ong, Carl Shan, and Kevin Wadman. 2025. How People Use ChatGPT. https://cdn.openai.com/pdf/a253471f-8260-40c6-a2cc-aa93fe9f142e/economic-research-chatgpt-usage-paper.pdf

[7] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2288–2292. https://doi.org/10.1145/3404835.3463098

[8] Nam Hai Le, Thomas Gerald, Thibault Formal, Jian-Yun Nie, Benjamin Piwowarski, and Laure Soulier. 2023. CoSPLADE: Contextualizing SPLADE for Conversational Information Retrieval. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I* (Dublin, Ireland). Springer-Verlag, Berlin, Heidelberg, 537–552. https://doi.org/10.1007/978-3-031-28244-7_34

[9] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543* (2021).

[10] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654* (2020).

[11] Weronika Łajewska and Krisztian Balog. 2025. GINGER: Grounded Information Nugget-Based Generation of Responses. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 2723–2727. https://doi.org/10.1145/3726302.3730166

[12] Carlos Lassance and Stéphane Clinchant. 2023. Naver Labs Europe (SPLADE)@ TREC Deep Learning 2022. *arXiv preprint arXiv:2302.12574* (2023).

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.

[14] Simon Lupart, Zahra Abbasiantaeb, and Mohammad Aliannejadi. 2024. IRLab@ iKAT24: Learned Sparse Retrieval with Multi-aspect LLM Query Generation for Conversational Search. *arXiv preprint arXiv:2411.14739* (2024).

[15] Simon Lupart, Mohammad Aliannejadi, and Evangelos Kanoulas. 2024. DiSCo Meets LLMs: A Unified Approach for Sparse Retrieval and Contextual Distillation in Conversational Search. *arXiv preprint arXiv:2410.14609* (2024).

[16] 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).

[17] 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. https://doi.org/10.1145/3767695.3769502

[18] 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*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1211–1225. https://doi.org/10.18653/v1/2023.findings-emnlp.86

[19] 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. *arXiv preprint arXiv:2406.05013* (2024).

[20] Thong Nguyen, Sean MacAvaney, and Andrew Yates. 2023. A Unified Framework for Learned Sparse Retrieval. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*. Springer, 101–116.

[21] Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-Shot Conversational Dense Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 829–838. https://doi.org/10.1145/3404835.3462856

Table 6: The prompt designed for PTKB classification.

PTKB classification.
I will give you some background information about a user and a conversation between the user and a system. You should tell me which of the background information is relevant for answering the last question of the user. Here is the background information about the user: $\{ptkb\}$ Please just copy the relevant background information to the last user utterance.

Table 7: Prompt for answer generation using nuggets and PTKB.

Answer Generation Prompt.
{nugget_text}
I will give you a conversation between a user and a system, together with background information about the user. Using your own knowledge and the provided nuggets, answer the user's final query. The answer must be at most 200 words. Do not refer to the nuggets or documents explicitly; simply use them to craft the answer.
Background information: $\{ptkb\}$
Conversation: $\{ctx\}$
User query: $\{user\ utterance\}$ **Answer:**

Table 8: Prompt for nugget extraction with PTKB.

Nugget Extraction Prompt.
I will give you a user query, the persona of the user, and a response to the query. Extract all nuggets of information relevant to the query. Each nugget must be an exact span copied verbatim from the response. Write one nugget per line. If no nugget exists, answer only “No nugget.” **User persona:** $\{ptkb\}$
Conversation: $\{ctx\}$
User query: $\{query\}$
Document: $\{doc\}$
Nuggets: (Copy exact spans from the document.)

A Prompts

All our prompts re-use the one proposed within the MQ4CS [1] framework or previous work on iKAT. We only changed the PTKB prompt, to adapt it to the classification task. Table 4 recalls the multiple query generation prompt, Table 5 the RAG prompt for response generation, and Table 6 the prompt for PTKB classification.

We also provide the prompts for the nuggets decomposition and answer generation for the generation only runs in Table 7 and 8.

Table 4: The prompt designed for multiple QD using GPT-4 as a zero-shot learner.

Multiple Query re-writing (QD).
Instruction: *I will give you a conversation between a user and a system. Imagine you want to find the answer to the last user question by searching on Google. You should generate the search queries that you need to search on Google. Please don't generate more than {ϕ} queries and write each query on one line.* # Background knowledge: $\{ptkb\}$ # Context: $\{ctx\}$ # User question: $\{user\ utterance\}$ # Generated queries:

Table 5: The prompt designed for answer generation, using retrieved documents.

Retrieval Augmented Answer Generation (RAG).
Doc1: $\{doc_{(1)}\}$ # Doc2: $\{doc_{(2)}\}$ # Doc3: $\{doc_{(3)}\}$ # Doc4: $\{doc_{(4)}\}$ # Doc5: $\{doc_{(5)}\}$ *# I will give you a conversation between a user and a system. Also, I will give you some background information about the user. You should answer the last utterance of the user by providing a summary of the relevant parts of the given documents. Please remember that your answer shouldn't be more than 200 words.* # Background information about the user: $\{ptkb\}$ # Conversation: $\{ctx\}$ # User query: $\{user\ utterance\}$