uva-irlab-conv at SemEval-2026 Task 8: Multi-Turn RAG with Learned Sparse Retrieval and Listwise Reranking

Simon Lupart **Kidist Amde Mekonnen**
Zahra Abbasiantaeb **Mohammad Aliannejadi**
University of Amsterdam
Amsterdam, The Netherlands
{s.c.lupart, k.a.mekonnen, z.abbasiantaeb, m.aliannejadi}@uva.nl

Abstract

This report describes our participation in SemEval-2026 Task 8 on multi-turn retrieval and question answering. The task evaluates conversational systems across four domains (finance, cloud documentation, government, Wikipedia), and includes unanswerable queries where the available collection does not contain sufficient evidence to produce a complete response. We propose a multi-turn retrieval-augmented generation (RAG) pipeline that combines learned sparse retrieval with LLM–based reranking and generation. Using sparse retrieval as the primary retrieval method, we leverage its strong generalization across domains. In addition, we make use of the long-context capabilities of LLMs for conversational query rewriting, pointwise and listwise reranking, and generating the final response, each conditioned on the full conversational history. This multi-step design enables effective integration of conversational context throughout retrieval and generation, improving robustness across domains.

1 Introduction

SemEval-2026 Shared Task 8 is based on the Multi-Turn RAG (MTRAGEval) benchmark (Rosenthal et al., 2026b,a), which evaluates conversational search and question answering (QA) systems across multi-turn interactions. The benchmark focuses on complex information needs arising in dialogue, where conversational context evolves and may introduce topic shifts, ambiguity, and noisy historical information that systems must resolve (Laban et al., 2026; Radlinski and Craswell, 2017). In addition, the task is motivated by the increasing role of LLMs as information access interfaces (Chatterji et al., 2025; Dalton et al., 2022). Systems are required not only to retrieve and synthesize evidence across multiple turns, but also to recognize and appropriately handle unanswerable or underspecified queries when insufficient supporting information is available. This includes the ability to signal uncertainty or request clarification when needed.

Our approach follows a multi-stage retrieval-augmented generation (RAG) pipeline combining conversational query rewriting, learned sparse retrieval (LSR), and LLM-based reranking. We first rewrite the latest user utterance into a standalone query using the full conversation history (Elgohary et al., 2019). The rewritten query is then used for retrieval with LSR (Zamani et al., 2018; Nguyen et al., 2023), which integrates neural semantic representations with lexical sparsity to provide robust cross-domain generalization (Formal et al., 2021). An initial reranking stage selects a set of candidate passages, which are subsequently refined through LLM-based listwise reranking that leverages the full conversational context for finer-grained evidence selection (Sun et al., 2023). Finally, the top-ranked passages are used to generate responses in a zero-shot RAG setting (Gao et al., 2023).

Participation in the shared task provides several insights. Our system demonstrates strong retrieval effectiveness, *ranking 2nd out of 38 teams* with an nDCG@5 score of 0.5475. Performance varies across domains, with the finance domain proving the most challenging and the Wikipedia-based domain the easiest. In contrast, generation performance is lower (*23/26 and 20/29*), primarily because our submission does not explicitly model unanswerable scenarios. Nevertheless, qualitative analysis shows that the system remains faithful when evidence from the collection is insufficient.

2 Background

Conversational search and QA have been studied through several shared tasks (Dalton et al., 2020; Aliannejadi et al., 2024; Abbasiantaeb et al., 2025; Gohsen et al., 2025) and offline datasets (Anantha

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2972–2980

et al., 2021; Adlakha et al., 2022), but most existing resources focus primarily on general domains (where LLMs can often rely on parametric knowledge) or focus on different aspects of the evaluation (e.g. personalization). In contrast, the MTRAG benchmark emphasizes domain-specific collections (FiQA (Maia et al., 2018), ClapNQ (Rosenthal et al., 2025), and two novel collections on Cloud documentation and government, all in English). It also explicitly evaluates unanswerable scenarios and response faithfulness, placing stronger requirements on retrieval and grounding (Es et al., 2024). The organizers defined three sub-tasks. **Task A** focuses on conversational search. **Task C** evaluates response generation in a standard RAG setting, where answers are generated from retrieved passages. Finally, **Task B** serves as an oracle RAG, where the response generation system has access to gold-labeled passages.

Query rewriting is a dominant paradigm in conversational QA, reformulating context-dependent utterances into standalone queries compatible with standard retrieval models (Vakulenko et al., 2021), further enhanced now with the use of LLMs (Mao et al., 2023; Mo et al., 2024; Lupart et al., 2025b,c). Unified approaches that jointly model retrieval and reasoning have also been proposed (Mo et al., 2025; Abbasiantaeb et al., 2024; Lupart et al., 2025a; Mo et al., 2026), but they typically require substantial supervised data. We therefore adopt a zero-shot rewriting and RAG-based strategy. More complex RAG strategies could also be combined with query decomposition and self-reflection over retrieved passages (Yao et al., 2022; Asai et al., 2023; Trivedi et al., 2023), but we did not explore them in our submission.

3 System Overview

Our system implements a multi-stage cascading RAG pipeline combining LLM-based conversational query rewriting, learned sparse retrieval, pointwise and LLM listwise reranking, and final response generation. Figure 1 presents the overall architecture, and each component is detailed below. The overall idea is to use a cascading ranking pipeline, with an effective method for retrieval, and then more expensive ranking approaches on the top retrieved passages for a more fine-grained final ranking.

Query Rewriting. We first apply LLM-based conversational query rewriting to transform the context-dependent user utterance into a standalone query. Query rewriting enables the use of standard retrieval models by resolving anaphora and ellipsis while preserving the primary information need expressed in the dialogue. Below is the same example from the MTRAG paper (Rosenthal et al., 2026b):

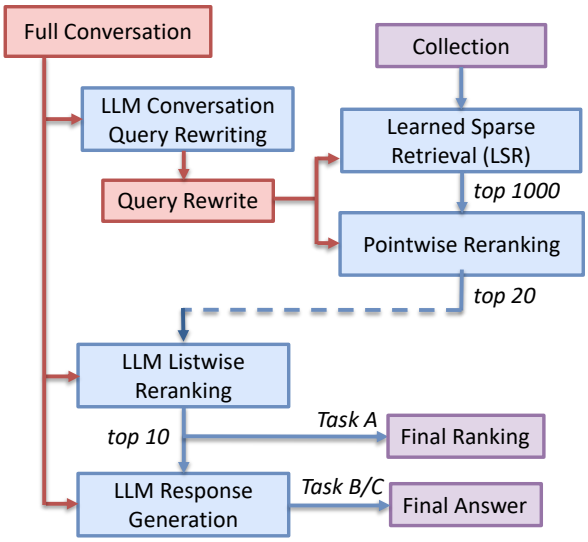

Figure 1: Overview of our submission. Early stages rely on query rewriting, while later stages leverage the full conversation history for more fine-grained processing.

User: Who is the CEO of Apple Inc.?
Agent: The CEO of Apple Inc. is Tim Cook.
User: its address?

[Rewriting] What is the address of Apple Inc?

Retrieval and Initial Reranking. The rewritten query is used for retrieval with the first-stage retrieval model, LION-SP (Zeng et al., 2025), an LSR model trained on an LLM backbone. LION-SP, similar to SPLADE (Formal et al., 2022) combines neural semantic modeling with lexical sparsity, providing strong out-of-domain generalization while remaining compatible with efficient inverted-index search. This reduces the candidate set from the full collection to the top 1000 passages, which is then reranked with a pointwise reranking stage to obtain the top 20 passages.

Listwise Conversational Reranking. While rewriting improves retrieval compatibility, it compresses conversational context and may omit constraints or information introduced in earlier turns. To reintroduce this context, we perform a second reranking stage using LLM-based listwise ranking conditioned on the full dialogue history (Sun et al., 2023). This stage jointly compares candidate pas-

Task A - Retrieval	nDCG		
	@1	@5*	@10
(best baseline) GPT-OSS-20b QR + ELSER	–	0.4795	–
LSR w/ LION-SP-8B (retrieval)	0.4910	0.4841	0.5343
+ Qwen3-Reranker-8B (pointwise)	0.5120	0.5477	0.5921
+ GPT-4.1 Listwise Reranking (†)	0.5331	0.5475	0.5943
(Rank 2 out of 38)			

Table 1: Retrieval performance of our submission on Task A of the MT-RAG SemEval 2026 MTRAG Task 8. We include ablations of the different steps of our pipeline and the best baseline from the organizer (GPT-OSS query rewriting with dense retrieval). (†) denotes our submission ***uva-1***. (*) denotes the official metric used to compare participants.

sages and reorders the top 20 results into a final top 10, enabling finer-grained, context-aware evidence selection prior to generation. We only apply listwise reranking on a shorter retrieved passage list since usual context windows do not allow for including the full 100 or 1000 passages at once. By passing the top 20 we ensure that the model focuses on the most relevant passages, improving precision.

Response Generation. Finally, the top 5 ranked passages are provided to an LLM for response generation in a zero-shot retrieval-augmented generation setting (Gao et al., 2023). Limiting generation to the highest-ranked evidence reduces noise while preserving sufficient contextual coverage. The model conditions on both the selected passages and the full conversation history to produce grounded, context-aware answers.

4 Experimental setup

For retrieval, we use an LSR model, Lion-SP-8B[1] (Zeng et al., 2025). For initial pointwise reranking, we employ Qwen3-Reranker-8B[2] (Yang et al., 2025), a large LLM-based reranker. The second-stage listwise reranking is performed using GPT-4.1 (Achiam et al., 2023), which is also used for final response generation. We relied on GPT-4.1, as the strongest non-reasoning model from OpenAI, achieving high performances on other conversational shared tasks (Aliannejadi et al., 2024).

Query rewriting, response generation and listwise reranking components are used in a zero-shot setting without task-specific fine-tuning. Similarly, retrieval models are only trained on MSMARCO,

[1] hzeng/Lion-SP-8B-llama3-marco-mntp

[2] Qwen/Qwen3-Reranker-8B

Task A - Retrieval	ClapNQ	FiQA	Govt	IBM Cloud
uva-1 (ours)	0.645	0.330	0.587	0.552
nb. turns	(84)	(59)	(106)	(87)

Table 2: nDCG@5 retrieval performance of our submission on the four domains of MTRAG.

but not finetuned for the specific domains of the task. Inverted index search is implemented with the `Seismic` library (Bruch et al., 2024), enabling efficient search over sparse representations. We apply representation pruning with query and document thresholds of (400, 600) to balance effectiveness and efficiency (although retrieval is not the efficiency bottleneck of the method).

For the retrieval track, systems are evaluated using nDCG (Järvelin and Kekäläinen, 2000), with nDCG@5 as the main metric. For response generation, the organizers reported three metrics: $\mathbf{RB_{alg}}$, the harmonic mean of BERT-Recall, BERT-K-Precision, and ROUGE-L (Adlakha et al., 2024), $\mathbf{RB_{llm}}$, an LLM-based evaluation metric (Kuo et al., 2025), and $\mathbf{RL_F}$ measuring faithfulness using the RAGAS framework (Es et al., 2024). They also computed the harmonic mean of these three together to rank participants' submissions ($\mathbf{H.Avg}$). The organizers also split turns into answerable, partially answerable and unanswerable, based on the passage assessment made on the collection.

Overall the dataset contains 332 turns evaluated for retrieval, resp. 84, 59, 106 and 87 turns on ClapNQ, FiQA, GovT and IBM-Cloud subsets.

System	IDK -Conditoned*				Not Conditioned			
	RB_agg	RB_llm	RL_F	**H. Avg***	RB_agg	RB_llm	RL_F	**H. Avg**
Task B - Generation w/ Oracle Retrieval								
uva-oracle	0.3683	0.6807	0.5981	**0.5123**	0.3590	0.8280	0.5899	0.5274
(Rank 23 out of 26)								
Task C - Generation w/ Predicted Retrieval								
uva-rag	0.3197	0.6538	0.6626	**0.4865**	0.3455	0.8332	0.8035	0.5619
(Rank 20 out of 29)								

Table 3: Response generation performance of our submission on Task B and C of the MT-RAG SemEval 2026 MTRAG Task 8 (***uva-oracle*** and ***uva-rag***). (*) denotes the official metric used to compare participants (IDK-Conditioned H.Avg).

5 Results

Task A - Retrieval. Table 1 reports the retrieval performance of our submission. To better isolate the contribution of each component, we analyze the retrieval and reranking stages separately.

The main performance gains stem from the retrieval stage and the first (pointwise) reranking step. This implies that query rewriting with GPT-4.1 already provides a strong foundation. Our retrieval alone also achieves an nDCG@5 of 0.4841, which surpasses the official baseline based on GPT-OSS-20B query rewriting combined with ELSER, which reports an nDCG@5 of 0.4795. While this comparison suggests the strength of our rewriting and retrieval setup, the recall@1000 of the baseline system is not reported, which limits deeper analysis of candidate coverage.

Using LION-SP-8B retrieval alone yields a strong nDCG@1 of 0.4910. The subsequent pointwise reranking step improves early precision by +2.1 nDCG@1 points. Finally, the listwise LLM reranking stage provides an additional +2.1 point gain at rank 1. Notably, this second reranking stage primarily benefits top-ranked results (nDCG@1), while yielding only marginal improvements at nDCG@5 and nDCG@10. Overall, our final nDCG@5 of 0.5475 places our system 2nd out of 38 submissions, with nDCG@5 serving as the official evaluation metric of the shared task.

Domain-Specific Results. We then report in Table 2 the performance of our submission across domains. Consistent with observations from the original MT-RAG benchmark, *FiQA* emerges as the most challenging collection among the four domains. Compared to the others, performance on FiQA is approximately half as high. Despite this relative drop, the results remain competitive. On average, the first relevant passage for FiQA appears at rank 3, whereas for the other domains it appears around rank 2. This indicates that, although ranking quality is lower for FiQA, relevant evidence is still retrieved early in the ranked list.

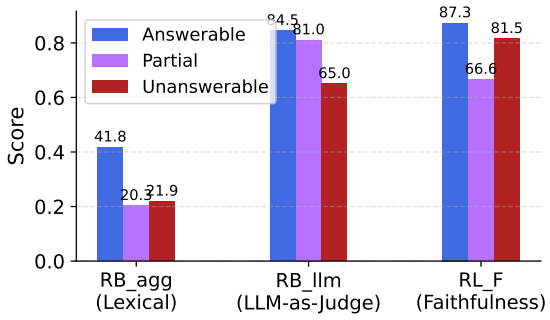

Figure 2: Response generation performance for different levels of answerability.

Task B & C - Response Generation. Table 3 presents the performance of our RAG submissions using two answer quality metrics (RB_agg and RB_llm), a faithfulness metric (RL_F), and their harmonic mean (H. Avg.). Results are presented separately for Task B and Task C. The official scores correspond to the evaluation conditioned on the IDK judge (left side of the table), which excludes underspecified turns.

Task C reflects a standard RAG setting: we use the retrieved passages from our Task A submission (uva-1) as context to generate answers within the conversational setting (uva-rag). In contrast, Task B serves as an oracle upper bound, where the generation model receives gold passages as context

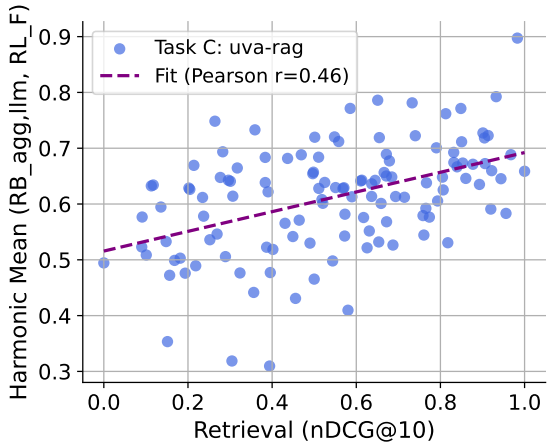

Figure 3: Retrieval and response generation correlation on partial and answerable turns. Pearson values for RB_agg, RB_llm and RL_F are 0.48, 0.41 and 0.08.

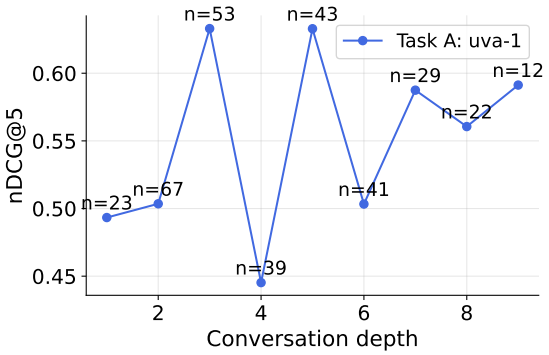

Figure 4: Retrieval Performance at varying depths.

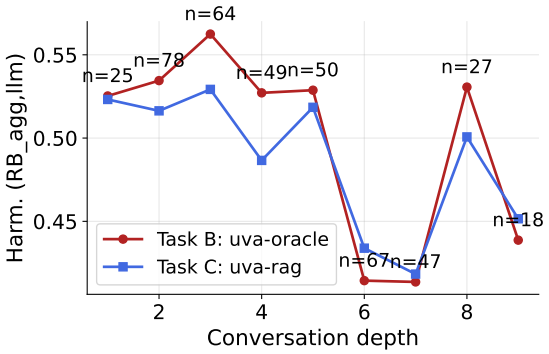

Figure 5: Response Generation at varying depths.

(uva-oracle). As expected, the oracle configuration achieves higher answer quality scores, indicating that access to gold evidence improves content relevance and completeness. However, our RAG pipeline achieves higher faithfulness (RL_F). We attribute this to the alignment between the reranking step and the final answer generation model, as both rely on the same LLM. In the oracle setting, although gold passages are provided, the generation model may selectively omit parts of the evidence or filter information it deems less relevant, which can reduce faithfulness.

Partial and Non-Answerable Turns. The organizers annotated a subset of turns as either non-answerable or partially answerable, requiring systems to explicitly identify underspecified cases. This results in two versions of each metric: the conditional variant (_idk_underspecified) and the original, unconditioned scores (the right section of Table 3 reports the latter). When not conditioning on the IDK judgment, performance increases substantially. In particular, the harmonic mean reaches 0.5619, representing an 8-point improvement over the conditioned evaluation. Under this setup, our RAG pipeline (Task C) outperforms the oracle configuration (Task B), which further emphasizes the strength of our retrieval component. Notably, we observe a faithfulness score of 0.8035, indicating strong alignment between generated answers and retrieved evidence.

Figure 2 further breaks down performance across answerable, partially answerable, and non-answerable subsets (without IDK conditioning). While answer quality metrics (RB_agg and RB_llm) decrease for partially or non-answerable turns, the model remains relatively faithful even for unanswerable questions. This suggests that, although content completeness is affected, the system generally avoids hallucinating unsupported information. We did not explore in more detail why our submission achieved such a high faithfulness, this could be attributed to a conservative prompt for the RAG component. For more details, prompts of our submission are included in Appendix A.

Correlation Retrieval and Response Generation. Figure 3 illustrates the relationship between retrieval effectiveness and response generation quality. For this analysis, we consider only answerable and partially answerable turns, and report results without conditioning on the IDK judge. We observe a positive correlation, with a Pearson correlation between retrieval and generation metrics of 0.46, indicating that strong retrieval is generally associated with improved response quality. When computing the Pearson correlation between nDCG@10 and each generation metric individually, we find that the reference-based metrics (RB_agg and RB_llm)

exhibit positive correlation with retrieval effectiveness. In contrast, faithfulness (RL_F) shows no correlation with nDCG@5. This is expected, as faithfulness measures alignment with the provided evidence rather than agreement with a gold reference answer, and is therefore less directly dependent on retrieval ranking quality.

Performance by Conversation Depth. Finally, we analyze performance as a function of conversational depth across all tasks. For retrieval (Task A), Figure 4 shows no clear degradation as depth increases, suggesting that the query rewriting and retrieval components handle longer conversational histories consistently. In contrast, for response generation (Tasks B and C), Figure 5 reveals a performance decline with increasing depth. For generations, we report the harmonic mean of the reference-based metrics only (excluding faithfulness), as we assume faithfulness to be independent of conversational depth.

6 Conclusion

We present in this report an effective ranking pipeline for conversational search that combines query rewriting and full conversation context modeling with a learned sparse retrieval backbone. Our analysis further shows a positive correlation between retrieval effectiveness and response quality, confirming the importance of ranking for downstream generation. At the same time, faithfulness appears less directly tied to retrieval metrics, highlighting its distinct role in evaluating grounded responses. We also observe that our submitted system remained largely faithful even in challenging cases. Although performance on reference-based metrics decreases for partially or non-answerable queries, the system generally avoids introducing unsupported information, even if it does not always explicitly acknowledge missing knowledge.

7 Acknowledgments

This research was partly supported by the Swiss National Science Foundation (SNSF), under the project PACINO (Personality And Conversational INformatiOn Access), grant number 215742.

References

Zahra Abbasiantaeb, Simon Lupart, and Mohammad Aliannejadi. 2024. Generating multi-aspect queries for conversational search. *arXiv preprint arXiv:2403.19302*.

Zahra Abbasiantaeb, Simon Lupart, Leif Azzopardi, Jeffrey Dalton, and Mohammad Aliannejadi. 2025. Conversational gold: Evaluating personalized conversational search system using gold nuggets. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3455–3465, New York, NY, USA. Association for Computing Machinery.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

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*, SIGIR '24, page 819–829, New York, NY, USA. Association for Computing Machinery.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Sebastian Bruch, Franco Maria Nardini, Cosimo Rulli, and Rossano Venturini. 2024. Efficient inverted indexes for approximate retrieval over learned sparse representations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 152–162.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use chatgpt. Technical report, National Bureau of Economic Research.

Jeffrey Dalton, Sophie Fischer, Paul Owoicho, Filip Radlinski, Federico Rossetto, Johanne R Trippas, and Hamed Zamani. 2022. Conversational information seeking: Theory and application. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 3455–3458.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. Cast 2020: The conversational assistance track overview. In *Text Retrieval Conference*.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5918–5924, Hong Kong, China. Association for Computational Linguistics.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural ir models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2353–2359, New York, NY, USA. Association for Computing Machinery.

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*, SIGIR '21, page 2288–2292, New York, NY, USA. Association for Computing Machinery.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, Haofen Wang, and 1 others. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1):32.

Marcel Gohsen, Zahra Abbasiantaeb, Mohammad Aliannejadi, Krisztian Balog, Timo Breuer, Jeffrey Dalton, Maik Fröbe, Christin Katharina Kreutz, Andreas Kruff, Simon Lupart, and 1 others. 2025. User simulation in practice: Lessons learned from three shared tasks. In *SIGIR Forum*, volume 59.

Kalervo Järvelin and Jaana Kekäläinen. 2000. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 41–48, New York, NY, USA. ACM.

Tzu-Lin Kuo, FengTing Liao, Mu-Wei Hsieh, Fu-Chieh Chang, Po-Chun Hsu, and Da-Shan Shiu. 2025. Radbench: Evaluating large language models' capabilities in retrieval augmented dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 868–902.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*.

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

Simon Lupart, Mohammad Aliannejadi, and Evangelos Kanoulas. 2025b. DiSCo: LLM Knowledge Distillation for Efficient Sparse Retrieval in Conversational Search. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 9–19, New York, NY, USA. Association for Computing Machinery.

Simon Lupart, Daniël van Dijk, Eric Langezaal, Ian van Dort, and Mohammad Aliannejadi. 2025c. 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*, SIGIR-AP 2025, page 353–363, New York, NY, USA. Association for Computing Machinery.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

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*, pages 1211–1225, Singapore. Association for Computational Linguistics.

Fengran Mo, Yifan Gao, Sha Li, Hansi Zeng, Xin Liu, Zhaoxuan Tan, Xian Li, Jianshu Chen, Dakuo Wang, and Meng Jiang. 2026. Agentic conversational search with contextualized reasoning via reinforcement learning. *arXiv preprint arXiv:2601.13115*.

Fengran Mo, Yifan Gao, Chuan Meng, Xin Liu, Zhuofeng Wu, Kelong Mao, Zhengyang Wang, Pei Chen, Zheng Li, Xian Li, Bing Yin, and Meng Jiang. 2025. UniConv: Unifying retrieval and response generation for large language models in conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6936–6949, Vienna, Austria. Association for Computational Linguistics.

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*.

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*, page 101–116, Berlin, Heidelberg. Springer-Verlag.

Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, page 117–126, New York, NY, USA. Association for Computing Machinery.

Sara Rosenthal, Yannis Katsis, Vraj Shah, Lihong He, Lucian Popa, and Marina Danilevsky. 2026a. Mtrag-un: A benchmark for open challenges in multi-turn rag conversations. *Preprint*, arXiv:2602.23184.

Sara Rosenthal, Vraj Shah, Yannis Katsis, and Marina Danilevsky. 2026b. Semeval-2026 task 8: Mtrageval: Evaluating multi-turn rag conversations. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California. Association for Computational Linguistics.

Sara Rosenthal, Avirup Sil, Radu Florian, and Salim Roukos. 2025. Clapnq: C ohesive l ong-form a nswers from p assages in natural questions for rag systems. *Transactions of the Association for Computational Linguistics*, 13:53–72.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 14918–14937.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 355–363.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Hamed Zamani, Mostafa Dehghani, W. Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 497–506, New York, NY, USA. Association for Computing Machinery.

Hansi Zeng, Julian Killingback, and Hamed Zamani. 2025. Scaling sparse and dense retrieval in decoder-only llms. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 2679–2684, New York, NY, USA. Association for Computing Machinery.

A Prompts

We provide in this section the detailed prompts we used for our submission, including LLM query rewriting, listwise LLM reranking and response generation. All prompts were used on GPT-4.1.

```
You are a query rewriting model for
conversational information retrieval.
Rewrite ONLY the final user turn into a
single, standalone search query. Use earlier
turns ONLY to resolve references, ellipsis,
and ambiguity.

Guidelines:
- Maximize lexical clarity and keyword recall.
- Prefer explicit noun phrases over pronouns.
- Preserve important domain terms from all
turns.
- Do NOT answer the question.
- Do NOT add new facts or assumptions.
- Do NOT explain or comment.
- Output exactly one query.

Output ONLY the rewritten query text.
Conversation: {conversation_text}
Rewrite the last user message into a
standalone query.
Last user message: {last_user_utterance}
```

Table 4: Query Rewriting prompt to obtain a single standalone query from a full context conversation.

```
You are an expert relevance judge for
conversational search.
Conversation: {conversation_text}

Task:
Rank the following passages by how useful they
are for answering the user's current
information need. The full conversation
provides context, but prioritize the latest
user turn.

Guidelines:
- Prefer passages that directly answer the
current user need.
- Use earlier turns only to resolve references
or constraints.
- Penalize passages relevant only to earlier
turns.
- Do NOT generate an answer.

Passages: {top20_retrieved_text}
Return exactly 10 passage labels ranked from
best to worst, using only labels (D1..D20).
Example: D1 > D20 > D7 > ... > D14
```

Table 5: LLM Listwise reranking prompt.

```
You are a conversational question answering
assistant. Use the conversation history to
understand context and intent. Answer the LAST
user question using the information in the
documents. Answer in maximum 256 tokens.

Conversation: {conversation_text}
Retrieved Documents: {top5_retrieved_text}
```

Table 6: Response Generation with RAG prompt.