Report on the 9th Workshop on Search-Oriented Conversational Artificial Intelligence (SCAI 2025) at IJCAI 2025

Svitlana Vakulenko
Vienna University of Economics and Business, Austria
svitlana.vakulenko@wu.ac.at

Philipp Christmann
CISPA Helmholtz Center for Information Security, Germany
philipp.christmann@cispa.de

Isabel Feustel
Ulm University, Germany
isabel.feustel@uni-ulm.de

Vaibhav Adlakha
McGill University, Mila, ServiceNow, Canada
vaibhav.adlakha@mila.quebec

Vahid Sadiri Javadi
University of Bonn, Germany
vahidsj@bit.uni-bonn.de

Patricia Schmidtova
Charles University, Czech Republic
schmidtova@ufal.mff.cuni.cz

Abstract

The goal of Search-oriented Conversational AI is to design systems that allow for a more convenient information access by means of a conversational user interface. Further development of Conversational Search systems requires closer integration and better information exchange between the diverse research communities. Over the past years, the Search-Oriented Conversational Artificial Intelligence (SCAI) workshop became an established venue that provides a discussion platform on Conversational AI for intelligent information access, bringing together researchers and practitioners across artificial intelligence, natural language processing, information retrieval, recommender systems, machine learning, dialogue systems and human-computer interaction subfields. This year, the full-day SCAI workshop at IJCAI 2025 once again brought together a group of researchers interested in informing the design of a new generation of systems for conversational information access. This paper, co-authored by both organizers and participants of the workshop, presents a summary of the insights gathered from the joint discussions that followed the invited talks.

Date: 18 August 2025.

Website: https://scai.info/scai-2025/.

1 Introduction

The workshop series on Search-Oriented Conversational Artificial Intelligence (SCAI) is one of the first and long-standing workshops dedicated to conversational search, with previous editions at ICTIR (2017) [Burtsev et al., 2017], EMNLP (2018) [Chuklin et al., 2018], WebConf (2019), IJCAI (2019), EMNLP (2020) [Dalton et al., 2020], an independently-organized online event in Gather.Town (2021) [Vakulenko et al., 2021, 2022], at SIGIR (2022) [Penha et al., 2022] and CHIIR (2024) [Frummet et al., 2024].

SCAI 2025 was co-located with IJCAI and took place in the Palais des Congres de Montreal, Canada. The workshop was organised around six invited talks. Five of the talks were presented by last-year PhD students, who took the opportunity to summarize their research, report the main findings and suggested directions for future work, which provided excellent ground for the follow-up discussions. They covered a broad range of topics starting from the application of Large Language Models (LLMs) in Information Retrieval and Question Answering to simulating dialogues, evaluation and transparency of the model's predictions. Professor Laurent Charlin from HEC Montreal was invited to deliver a keynote presentation sharing his in-depth research experience on recommender systems and the more recent extensions that the members of his research group are currently working on.

The workshop participants came from different research areas combining core Machine Learning (ML), Natural Language Processing (NLP), Information Retrieval (IR), KR (Knowledge Representation) and Human-Computer Interaction (HCI) expertise. There were also key industry researchers present in the audience, who actively engaged in the follow-up discussions. This report provides an overview of the workshop program (Section 2), summarizes each of the talks separately and highlights the key outcomes of our discussions.

2 Workshop Program

Table 1 shows the workshop program. The first invited talk introduced the topic of using LLMs for IR, followed by the second invited talk that focused on the subsequent task of Question Answering and the challenge of integrating heterogeneous information sources, such as tables and knowledge graphs in addition to textual corpora. This first session laid out the ground by discussing the core tasks in IR and the state-of-the-art approaches used for ranking and generation.

In the second session, the talks 3 and 4 introduced more advanced tasks: recommendation and conversational search that builds on top of the ranking and generation task by combining them into an interactive process. Traditional approaches for modeling user-item preferences were revisited and an extension to dialogue settings was further discussed.

Finally, the last session was dedicated to the challenges of evaluating generated texts and using grounded dialogue for explaining model's predictions. While there is a clear potential in using LLMs to generate fluent dialogues that help users to engage with the model and better understand its predictions, there is a clear gap in our ability to evaluate such dialogues on a large scale.

Table 1. The SCAI'25 program. All slides for the talks are available on the workshop website.

Time	Event
09:00	Welcome
09:10	Talk 1: Large Language Models in Retrieval – Mapping Potential and Pitfalls *Vaibhav Adlakha (McGill University/Mila - Quebec AI Institute)*
09:50	Talk 2: Question Answering over Structured Data and Unstructured Text *Philipp Christmann (Max Planck Institute for Informatics; now at CISPA)*
10:30	Coffee break
11:00	Talk 3: TEARS: Textual Representations for Scrutable Recommendations *Laurent Charlin (HEC Montreal)*
11:45	Talk 4: Subjective Debate as Conversational Search: Simulating Sales Negotiations Grounded in Preferences *Vahid Sadiri Javadi (University of Bonn)*
12:30	Lunch Break
14:00	Talk 5: Evaluation Challenges in the Era of Large Language Models *Patricia Schmidtova (Charles University)*
14:45	Talk 6: Enhancing model transparency: Effects of Domain Knowledge Integration for Conversational Explainable Artificial Intelligence (XAI) *Isabel Feustel (Ulm University)*

2.1 Large Language Models in Retrieval – Mapping Potential and Pitfalls

Vaibhav Adlakha (McGill University/Mila - Quebec AI Institute)

At the start of the workshop, Vaibhav comprehensively examined the influence of large language models (LLMs) on dense retrieval, with a focus on both their promise and their limitations. Drawing on several of his recent works, the keynote provided a holistic overview of how LLMs affect multiple components in the development and training of dense retrievers.

On the architecture side, Vaibhav discussed adapting LLMs for dense retrieval by modifying the casual autoregressive architecture of LLMs into a bidirectional architecture. LLM2Vec [BehnamGhader et al., 2024] introduces such a modification and proposes a parameter- and sample-efficient training objective called Masked Next Token Prediction (MNTP). Similar to masked language modeling, the goal is to predict the masked token in the input, however, the prediction is done from the token preceding the mask token, aligning the objective with how LLMs are trained with next-token prediction. The resulting model leads to state-of-the-art performance in both unsupervised and supervised settings, underscoring the upside re-purposing LLM's inductive biases for retrieval.

In contrast, improvements on the data front have been far more ambiguous. Vaibhav challenged the widespread assumption that augmenting training data with LLM-generated synthetic

data reliably improves generalization. Over the last couple of years, several methods for training generalizable, instruction-following retrievers have included LLM-generated synthetic data in their training mixture [Wang et al., 2024; Muennighoff et al., 2024]. On popular diverse multi-task benchmarks such as MTEB [Muennighoff et al., 2023], such methods have reported substantial improvements over models trained without synthetic data. The diversity of synthetic data is often accredited for this superior performance. However, this data is typically not publicly released, making these claims difficult to evaluate or reproduce. In Springer et al. [2025], the authors first make a public reproduction of the synthetic data, as described in Wang et al. [2024], using open-source models such as Llama-3.1-70B [Meta, 2024]. After verifying that the reproduced data mixture yields comparable performance gains, they conduct a fine-grained analysis of performance across different task types within MTEB. They found that substantial gains occur in a small number of datasets, creating an overall impression of uniform improvement that does not reflect actual distributional effects. In other words, the benefits of synthetic augmentation are sparse and unevenly distributed rather than broadly generalizable.

In the end, Vaibhav offered a forward-looking perspective on how these lessons should shape the next generation of retrieval systems. As information-seeking increasingly shifts toward LLM-mediated interfaces, he anticipates a broader pivot toward designing retrievers for LLM agents rather than solely for human users.

Together, these findings offer a nuanced perspective on integrating LLMs into retrieval systems: architectural innovation appears to offer clear, reproducible benefits, while data-centric claims – especially those relying on undisclosed synthetic sources – should be scrutinized more carefully before being treated as broadly reliable advances.

2.2 Question Answering over Structured Data and Unstructured Text

Philipp Christmann (Max Planck Institute for Informatics; now at CISPA)

In this talk, Philipp discussed the importance of integrating both structured data and unstructured text for answering factual questions. The first part of the talk introduced the need for heterogeneous sources for question answering (QA), exemplified by the question *"first Chinese NBA player?"*. For questions like this, a single type of an information source (e.g., a knowledge base, a table, or a document) is not sufficient. Instead, information from multiple source types should be integrated to produce a correct and complete answer. To motivate the need for research on QA in the era of LLMs, the next part discussed the major weaknesses of recent LLMs. To summarize, while LLMs demonstrate strong general capabilities, they often underperform task-specific QA approaches at substantially higher computational cost, answers cannot be traced, and more complex questions requiring operations such as aggregation or grouping (e.g., *"Universities with most graduates that later published at IJCAI?"*) are out of scope.

The remainder of the talk then outlined strategies for addressing these shortcomings and integrating heterogeneous sources into the QA pipeline for enhanced answer coverage. The first strategy, *verbalization*, transforms relevant information from all sources into textual form, which is then provided as context to the LLM [Oguz et al., 2022]. This yields competitive performance at low cost, though the answers are still not traceable. The second strategy combines small-scale language models with graph neural networks for identifying a small set of relevant ev-

idence [Christmann et al., 2023]. These ML models are trained on graphs that combine evidence from heterogeneous sources to predict the answer from a small subgraph, making it suitable for causal explanations of the derived answer. The third strategy utilizes LLMs for generating code that can then be executed over heterogeneous sources for answering complex questions [Christmann and Weikum, 2025]. Notably, the generated code calls operators for retrieval or information extraction, that are themselves empowered by small-scale language models. This strategy enables targeting complex questions at low cost, and yields traceable answers.

The overarching finding is that integrating heterogeneous sources in the QA pipeline yields consistent and substantial performance improvements compared to single-source approaches. Another major takeaway was that combining language models with structured representations, such as graphs or generated code, can significantly improve performance over approaches relying solely on LLMs. Finally, the talk stressed that the task of question answering is not solved: many open problems in QA remain out of scope for current LLMs, and likely also for the upcoming generations.

2.3 TEARS: Textual Representations for Scrutable Recommendation

Laurent Charlin (HEC Montreal & Mila – Quebec AI Institute)

In this talk, Prof. Charlin described the use of LLMs in modern recommender systems (RecSys). He started by briefly introducing the RecSys task and the idea behind modeling user preferences. Then he motivated the need for scrutable RecSys using LLMs that can summarize user preferences in natural language [Radlinski et al., 2022]. He then proceeded to describe their approach, called TEARS, and reported on the evaluation of its recommendation performance in comparison with standard non-scrutable models [Penaloza et al., 2025].

Instead of representing user's interests as a vector embedding, TEARS encodes them as natural text, providing transparency and allowing users to edit them. To do so, TEARS uses an LLM to generate user summaries based on user preferences. Afterwards these generated summaries are used in a hybrid approach that involves an optimal transport procedure to align the summaries' representation with the learned representation of a standard Variational Autoencoder (VAE) for collaborative filtering. This approach surpasses the performance of three popular VAE models while providing user-controllable recommendations. Three simulated user tasks demonstrate controllability of TEAR and help to evaluate the effectiveness of a user editing their summary.

2.4 Subjective Debate as Conversational Search: Simulating Sales Negotiations Grounded in Preferences

Vahid Sadiri Javadi (University of Bonn)

In this talk, Vahid introduced a conversational simulation framework featuring LLM agents, in which expert and sales agents engage in a subjective debate to assist a persona-driven user in making purchase decisions. The agent's functionality extends beyond dialogue into reflection, tool usage, and expert consultations. Framed as the task of conversational search, they proactively

elicit user preferences, retrieve review-grounded evidence, and compete through persuasive multi-turn interactions. The framework offers a scalable approach to oversee diverse subjective decision-making scenarios, where an adversarial dialogue helps to surface relevant product attributes and align recommendations with the user needs.

OpinionConv is the core contribution introduced in Javadi et al. [2023], which proposes a pipeline used to generate ∼200K opinionated multi-turn conversations grounded in real Amazon product reviews [Hou et al., 2024]. This research directly extends the task of conversational search towards subjectivity-aware reasoning by simulating how agents negotiate preferences and uncertainties. By incorporating negotiation tactics such as *"Deny-Disagreement"*, *"Deny-Switch Product"*, and *"Search-Agreement"*, the system simulates realistic sales interactions where agents express and counter subjective opinions about product features like battery life, and camera quality. Vahid concluded his talk by outlining future directions for applying reinforcement learning frameworks to subjective reasoning, extensions to new product domains and adversarial sales strategies.

2.5 Evaluation Challenges in the Era of Large Language Models

Patricia Schmidtova (Charles University)

In this talk, Patricia discussed critical challenges in the evaluation of dialogue systems and generative tasks in general. She reported on the current state of both automatic and human evaluation, highlighting common pitfalls and their implications for assessing system performance, especially with respect to the model's hallucinations. While the majority of Patricia's work is not focused specifically on dialogue, it was one of the main tasks she considered in all of her papers that she discussed in this talk.

The first core pitfall pertinent to the evaluation of LLMs is the issue of data contamination, where closed-source models may be indirectly exposed to benchmark data, which casts doubt on the validity of many benchmark results [Balloccu et al., 2024]. The second common pitfall is the widespread use of the automatic metrics that often show poor correlation with human judgment and fail to capture important issues in generated texts, such as factuality and faithfulness [Schmidtova et al., 2024]. Conducting human evaluations presents challenges related to scalability and high costs. Also, ensuring unbiased high-quality feedback is very hard, especially when employing non-expert annotators [Schmidtova et al., 2025]. Finally, LLM as a judge may work well for certain tasks [Kasner et al., 2025], yet fail dramatically for other tasks in unseen domains [Schmidtová et al., 2025].

Ultimately, Patricia offered key recommendations that are crucial for the community to foster more rigorous and meaningful evaluation practices:

- implement careful quality control measures for human evaluations, such as using pilot studies and attention checks;
- ensure fair and objective comparison with baselines when reporting results;
- make evaluations reproducible by sharing prompts, code, and version details;
- share and openly discuss best evaluation practices, such as selection and combination of automated metrics.

2.6 Enhancing Model Transparency: Effects of Domain Knowledge Integration for Conversational Explainable Artificial Intelligence

Isabel Feustel (Ulm University)

As conversational systems become a central medium for information access, ensuring their transparency and user trust is critical, especially when these systems explain AI model decisions. In this talk, Isabel presented recent work on using grounded dialogues for Explainable Artificial Intelligence (XAI). She investigated how domain knowledge (DK) can be integrated into a dialogue-based explanation to help users understand model predictions more effectively [Feustel et al., 2024, 2025].

Building on principles from computational argumentation, Isabel presented an approach which structures DK as bipolar argumentation trees, linking model features and outputs to domain-relevant concepts and allowing the system to retrieve and present supporting or counter-acting arguments during the interaction.

Isabel discussed insights from user studies showing that explanations enriched with DK were perceived as more useful and consistent, leading to higher likeability of the dialogue and a reduction in cognitive effort required to understand the AI's decision. Moreover, the presence of DK increased participants' confidence in the system's usefulness.

Overall, the talk emphasized that integrating domain knowledge through computational argumentation enriches conversational XAI by providing more transparent, context-aware explanations, while also pointing out practical hurdles such as the effort required to model DK and the balance between explanatory depth and user-friendly interaction. These insights contribute to the broader discussion on designing interactive AI systems that support meaningful human-AI collaboration.

3 Discussion

Here, we summarize the most salient discussions among SCAI participants that emerged both during the official Q&A sessions following the talks and through informal social interactions during breaks and after the workshop.

3.1 How to Optimize Information Retrieval *for* LLMs?

Retrieval systems are traditionally optimized towards the use within traditional search engines or question answering systems. However, as users turn to conversational assistants and LLM agents, ad-hoc search is increasingly becoming a background process that enables grounded response generation by an LLM. This raises several important research questions:

- How should one design a retrieval system to be used by LLM agents, instead of humans?
- And how does one evaluate such a retrieval system designed for LLMs?

Current retrieval models are optimized towards high-precision results, as humans rarely inspect more than top-10 search results to satisfy their information need. However, retrieval-augmented

LLMs can ingest much more information. For long-context LLMs the input length can accommodate up to a 1K search snippets. Therefore, designing modern retrieval models require careful understanding of their interaction with an LLM. This includes a deep understanding of how LLMs process the retrieved results, since they may exhibit positional biases different from humans.

Similar challenges extend towards evaluation approaches. The most commonly-used IR metrics, such as NDCG@k or MRR@k, are precision-oriented with 5 or 10 as typical values usually chosen for the k cut-offs. Designing retrieval models that are to be used within LLM-powered systems will require re-thinking our basic assumptions, including evaluation protocols, that should be guided by a better understanding of LLM architectures and their properties.

3.2 Trade-off between Performance and Transparency

As conversational AI systems become widely adopted in critical ecosystems, such as hospitals, it is crucial to make the underlying decision processes transparent. However, such transparency can also lead to notable performance degradations, as was noted in different talks throughout our workshop. An interesting example is the usage of "soft" retrieval queries, which are latent representations of information needs [Zhang et al., 2025]. The intuition here is that LLMs can more effectively represent the information they seek within latent space than in symbolic space. While such an approach can improve the end-to-end performance of conversational systems, this comes at the cost of reduced transparency.

3.3 Integration of Tools

Integrating tools into the LLM inference, such as the ones for making precise computations, can greatly enhance both the applicability and reliability of LLMs. However, the prevalent paradigm of enumerating all tools in the initial system-instructions prompt simply does not scale. Providing LLMs with a large set of tools can dramatically increase the context length, and thus the inference costs. Selecting an appropriate set of tools is specific to each query, which motivates the need for developing specialized retrieval techniques. These allow for ranking the available tools with respect to the conversation history such that they can be invoked when needed, and their results are finally integrated into an LLM response.

4 Conclusion

The workshop on Search-Oriented Conversational Artificial Intelligence (SCAI) at IJCAI 2025 facilitated discussions among participants from diverse backgrounds. The three informative sessions provided in-depth overviews of the number of relevant topics, such as retrieval, question answering, applications, evaluation, and transparency. In particular, we drew inspiration from the current state of the art in knowledge representation, natural language generation, and recommender systems, and made their relationships to conversational search more explicit. Moreover, our discussions yielded valuable insights with important implications for the design and evaluation of conversational search systems. Participants repeatedly emphasized the need for adapting information retrieval approaches to better integrate with LLMs, and discussed the challenges of aligning performance with transparency in such systems.

Acknowledgments

Svitlana Vakulenko was funded by the Vienna Science and Technology Fund (WWTF) Grant ID 10.47379/VRG24013.

References

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.5/.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=IW1PR7vEBf.

Mikhail Burtsev, Aleksandr Chuklin, Julia Kiseleva, and Alexey Borisov. Search-Oriented Conversational AI (SCAI). In Jaap Kamps, Evangelos Kanoulas, Maarten de Rijke, Hui Fang, and Emine Yilmaz, editors, *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2017, Amsterdam, The Netherlands, October 1-4, 2017*, pages 333–334. ACM, 2017. URL https://doi.org/10.1145/3121050.3121111.

Philipp Christmann and Gerhard Weikum. Recursive question understanding for complex question answering over heterogeneous personal data. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18269–18288, Vienna, Austria, July 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.findings-acl.939/.

Philipp Christmann, Rishiraj Saha Roy, and Gerhard Weikum. Explainable conversational question answering over heterogeneous sources via iterative graph neural networks. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 643–653, 2023. URL https://dl.acm.org/hidedoi/10.1145/3539618.3591682.

Aleksandr Chuklin, Jeff Dalton, Julia Kiseleva, Alexey Borisov, and Mikhail Burtsev, editors. *Proceedings of the 2nd International Workshop on Search-Oriented Conversational AI, SCAI@EMNLP 2018, Brussels, Belgium, October 31, 2018*, 2018. Association for Computational Linguistics. URL https://aclanthology.org/volumes/W18-57/.

Jeff Dalton, Aleksandr Chuklin, Julia Kiseleva, and Mikhail Burtsev, editors. *Proceedings of the 5th International Workshop on Search-Oriented Conversational AI (SCAI)*, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.scai-1.0.

Isabel Feustel, Niklas Rach, Wolfgang Minker, and Stefan Ultes. Enhancing model transparency: A dialogue system approach to XAI with domain knowledge. In Tatsuya Kawahara, Vera Demberg, Stefan Ultes, Koji Inoue, Shikib Mehri, David Howcroft, and Kazunori Komatani, editors, *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 248–258, Kyoto, Japan, September 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.sigdial-1.22/.

Isabel Feustel, Carolin Schindler, Niklas Rach, Wolfgang Minker, and Stefan Ultes. Towards a deeper understanding: Effects of domain knowledge integration for conversational xai. In *Proceedings of the 3rd International Workshop on Argumentation for eXplainable AI*, pages 63–77, 2025. URL https://ceur-ws.org/Vol-4066/paper_7.pdf.

Alexander Frummet, Andrea Papenmeier, Maik Fröbe, and Johannes Kiesel. The Eighth Workshop on Search-Oriented Conversational Artificial Intelligence (SCAI'24). In Paul D. Clough, Morgan Harvey, and Frank Hopfgartner, editors, *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval, CHIIR 2024, Sheffield, United Kingdom, March 10-14, 2024*, pages 433–435. ACM, 2024. URL https://doi.org/10.1145/3627508.3638310.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging Language and Items for Retrieval and Recommendation. *arXiv preprint arXiv:2403.03952*, 2024. URL https://arxiv.org/abs/2403.03952.

Vahid Sadiri Javadi, Martin Potthast, and Lucie Flek. OpinionConv: Conversational product search with grounded opinions. 2023. URL https://aclanthology.org/2023.sigdial-1.6/.

Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtová, Ivan Kartáč, Kristýna Onderková, Ondřej Plátek, Dimitra Gkatzia, Saad Mahamood, Ondřej Dušek, and Simone Balloccu. Large language models as span annotators, 2025. URL https://arxiv.org/abs/2504.08697.

Llama Team @ Meta. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.148/.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative Representational Instruction Tuning. In *The Thirteenth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BC4lIvfSzv.

Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. UniK-QA: Unified Representations of Structured and Unstructured Knowledge for Open-Domain Question Answering. In

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1535–1546, 2022. URL https://aclanthology.org/2022.findings-naacl.115/.

Emiliano Penaloza, Olivier Gouvert, Haolun Wu, and Laurent Charlin. Tears: Text representations for scrutable recommendations. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 4949–4968, New York, NY, USA, 2025. Association for Computing Machinery. URL https://doi.org/10.1145/3696410.3714948.

Gustavo Penha, Svitlana Vakulenko, Ondrej Dusek, Leigh Clark, Vaishali Pal, and Vaibhav Adlakha. The Seventh Workshop on Search-Oriented Conversational Artificial Intelligence (SCAI'22). In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3466–3469. ACM, 2022. URL https://doi.org/10.1145/3477495.3531700.

Filip Radlinski, Krisztian Balog, Fernando Diaz, Lucas Dixon, and Ben Wedin. On natural language user profiles for transparent and scrutable recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 2863–2874. ACM, 2022. URL https://dl.acm.org/doi/10.1145/3477495.3531873.

Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. Automatic metrics in natural language generation: A survey of current evaluation practices. In Saad Mahamood, Nguyen Le Minh, and Daphne Ippolito, editors, *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583, Tokyo, Japan, September 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.inlg-main.44/.

Patricia Schmidtova, Eduardo Calò, Simone Balloccu, Dimitra Gkatzia, Rudali Huidrom, Mateusz Lango, Fahime Same, Vilém Zouhar, Saad Mahamood, and Ondrej Dusek. Do my eyes deceive me? a survey of human evaluations of hallucinations in NLG. In Lucie Flek, Shashi Narayan, Lê Hong Phuong, and Jiahuan Pei, editors, *Proceedings of the 18th International Natural Language Generation Conference*, pages 60–79, Hanoi, Vietnam, October 2025. Association for Computational Linguistics. URL https://preview.aclanthology.org/ingest-inlg/2025.inlg-main.4/.

Patrícia Schmidtová, Ondrej Dusek, and Saad Mahamood. Real-world summarization: When evaluation reaches its limits. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25014–25026, Suzhou, China, November 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.findings-emnlp.1363/.

Jacob Mitchell Springer, Vaibhav Adlakha, Siva Reddy, Aditi Raghunathan, and Marius Mosbach. Understanding the influence of synthetic data for text embedders. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22551–22567, Vienna, Austria,

July 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.findings-acl.1160/.

Svitlana Vakulenko, Ondrej Dusek, and Leigh Clark. Report on the 6th workshop on search-oriented conversational AI (SCAI 2021). *SIGIR Forum*, 55(2), 2021.

Svitlana Vakulenko, Johannes Kiesel, and Maik Fröbe. SCAI-QReCC Shared Task on Conversational Question Answering. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 4913–4922. European Language Resources Association, 2022. URL https://aclanthology.org/2022.lrec-1.525.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.642/.

Wenzheng Zhang, Xi Victoria Lin, Karl Stratos, Wen-tau Yih, and Mingda Chen. ImpRAG: Retrieval-augmented generation with implicit queries. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8888–8900, Suzhou, China, November 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.findings-emnlp.472/.