Report on the 8th Workshop on Search-Oriented Conversational Artificial Intelligence (SCAI 2024) at CHIIR 2024

Alexander Frummet
Universität Regensburg
Germany
alexander.frummet@ur.de

Andrea Papenmeier
University of Twente
Netherlands
a.papenmeier@utwente.nl

Maik Fröbe
Friedrich-Schiller-Universität Jena
Germany
maik.froebe@uni-jena.de

Johannes Kiesel
Bauhaus-Universität Weimar
Germany
johannes.kiesel@uni-weimar.de

Vaibhav Adlakha, Norbert Braunschweiler, Mateusz Dubiel, Satanu Ghosh, Marcel Gohsen, Christin Kreutz, Milad Momeni, Markus Nilles, Sachin Pathiyan Cherumanal, Abbas Pirmoradi, Paul Thomas, Johanne R. Trippas, Ines Zelch, Oleg Zendel*

Abstract

Conversational Agents are increasingly integrated into our daily routines, assisting us with various tasks, from simple commands such as scheduling events to more complex conversational search interactions. Such conversational search systems are traditionally evaluated with word-overlap metrics such as F1 score and accuracy. The full-day workshop on Search-Oriented Conversational Artificial Intelligence (SCAI) at CHIIR 2024 explored the evaluation of conversational search systems from the user's perspective. This interactive workshop included multiple panel discussions and working groups focused on developing and discussing innovative, user-centered evaluation methods for these systems. This paper, co-authored by both organizers and participants of the workshop, presents a summary of the insights gathered from the panel discussions and working groups.

Date: 14 March 2024.

Website: https://scai.info/scai-2024/.

1 Introduction

The workshop series on Search-Oriented Conversational Artificial Intelligence (SCAI) is a discussion platform on Conversational AI for intelligent information access, bringing together researchers

*Affiliation not shown for all authors due to space limitations (see Appendix A for details).

Figure 1. Group photo of in-person and online participants.

and practitioners across natural language processing, information retrieval, machine learning and human-computer interaction fields. 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], independently organized (2021) [Vakulenko et al., 2022] and at SIGIR (2022) [Penha et al., 2022].

SCAI 2024 was co-located with CHIIR in Sheffield [Frummet et al., 2024]. The edition's theme, *Evaluation and Metrics*, focused on human-centered aspects of evaluating conversational systems beyond relevance, such as user engagement and fluency. The workshop adopted a hybrid format, with 25 participants attending on-site and 12 online (see Figure 1 for a snapshot of participants). The hybrid setup facilitated productive discussions and allowed for a broader audience.

This report provides an overview of the workshop program (Section 2) and highlights the key outcomes from the discussion rounds (Section 3).

2 Workshop Program

Table 1 shows the workshop program. First, a keynote talk introduced the topic of conversational search, followed by a panel discussion on general research directions and "big ideas" in SCAI. This first session ensured that all participants knew enough to actively participate in the working groups. The first session was followed by two more, each with a panel discussion that served as the basis for subsequent working group discussions (detailed in Section 3). The participants (online and on-site) formed the working groups themselves to match their own specific interests. At the end of the workshop, working group representatives gave a summary talk of their discussions.

2.1 Keynote

George Buchanan (RMIT University), Dana McKay (RMIT University).

At the start of the workshop, George Buchanan and Dana McKay gave a keynote[1] on conversational search to introduce the field and set the scene for the discussion rounds later. With their background in human-computer interaction, information interaction, and measuring human

[1]Slides available on the workshop web page, https://scai.info/scai-2024/

Table 1. The SCAI'24 workshop program.

Time	Event
09:30–09:45	Welcome by the Organizers
09:45–10:15	Keynote
10:15–11:00	Panel 1: Big Ideas in SCAI
11:30–12:30	Panel 2: Human-centered Metrics for SCAI
12:30–13:00	Working Groups
14:00–14:45	Panel 3: Challenges of Human-in-the-loop Evaluations in SCAI
14:45–15:30	Working Groups
16:00–17:00	Outcomes of Working Groups and Closing

behavior when interacting with technology, their keynote focused on quality aspects of human-machine conversations: What makes a good conversation (e.g., active listening, taking turns, staying on topic, paying attention to body language, showing empathy), and what makes a bad one (e.g., rambling, negativity, monopolizing the conversation, looking a the mobile phone). They asked today's large language models for quality aspects and found that the models also produce these. The keynote then posed the question on whether the models also adhere to these quality aspects when one interacts with them. Moreover, Dana and George asked the audience to look beyond the individual conversation, and also consider social and ethical implications for evaluating conversations: for example, we do not want a search bot that tells users they should harm themselves even if that would be beneficial for some "greater good."

2.2 Panel 1: Big Ideas

Panelists: *George Buchanan (RMIT University), Ehsan Kamalloo (ServiceNow), Dana McKay (RMIT University), Johanne Trippas (RMIT University).* Moderator: *Andrea Papenmeier (University of Twente).*

The first panel aimed to discuss current themes in the field of SCAI and open up broad directions for research. The panel included the two keynote speakers, George Buchanan and Dana McKay, as well as Ehsan Kamalloo (*ServiceNow Research*) as industry researcher and Johanne Trippas (RMIT University). The panel extended the keynote talk to a discussion about user-centeredness in search systems and ideas for the future of search. Johanne suggested incorporating virtual reality and mixed reality into traditional search interfaces, inviting us to rethink the nature of our interactions with information. By recognizing that conversations are inherently interactive and multi-modal, this approach leverages the dynamic capabilities of virtual and mixed reality, enriching the user experience beyond textual interactions [Trippas et al., 2024]. It acknowledges that the scope of information-seeking is expanding with advancements in technologies like virtual reality and large language models (LLMs) and converts information seeking into an immersive experience where information is retrieved and interactively explored. Ehsan added that new search systems, especially those including an LLM, should be grounded in reliable knowledge. Dana pointed at current systems that do not engage in a holistic back-and-forth which is a characteristic of a human-human conversation. However, the panelists saw the aspiration to a high level of

human-likeness and the incorporation of contextual information rather critically, as it can have unwanted and uncontrollable side effects.

2.3 Panel 2: Human-centered Metrics

Panelists: *Vaibhav Adlakha (McGill University, AI Institute Mila in Quebec), Leif Azzopardi (University of Strathclyde), Mateusz Dubiel (University of Luxembourg), Satanu Ghosh (University of New Hampshire).* Moderator: *Johannes Kiesel (Bauhaus-Universität Weimar).*

The discussion centered on the multifaceted evaluation of conversational systems. Vaibhav assumed that factual accuracy and completeness will be a comparatively minor problem in the future. However, Mateusz questioned whether insights from studies actually transfer to the real world ("ecological validity"). Leif suggested moving beyond Q&A in conversations and looking more at questions for which there is no direct answer in the data. Satanu emphasized that identifying speech acts can shed light on what users actually want. Other evaluation challenges mentioned by the panelists include (1) shifting goals during a search task, (2) including the user context, (3) long-term/longitudinal evaluation, and (4) measuring the user's cost of actions (e.g., time spent). Panelists emphasized the need to evaluate various people-centered aspects of conversations, particularly concerning the user's trust in the system: reliability of citations, comprehensibility of responses, and perceived confidence of the system in its statements.

2.4 Panel 3: Challenges of Human-in-the-loop Evaluations

Panelists: *Norbert Braunschweiler (Toshiba Europe Limited), Christin Kreutz (TH Mittelhessen), Markus Nilles (Universität Trier).* Moderator: *Alexander Frummet (Universität Regensburg).*

The final panel discussion focused on the evaluation challenges of conversational systems, highlighting the importance of selecting appropriate human-centered metrics that ensure realism and user satisfaction. The consensus among panelists was that metrics should align with the specific task requirements. Norbert advocated for multi-faceted metrics tailored to the task's needs, suggesting that, e.g., accuracy should be prioritized for tasks demanding high precision, while efficiency should be key for tasks requiring quick completion. Christin and Markus emphasized choosing metrics that accurately reflect the assumed user behavior, characteristics, or skills relevant to the situation, underlining that understanding user information needs is crucial for satisfaction. Norbert added ease of use, responsiveness, and output reliability as important factors for satisfaction. On realism in human-centered evaluations, Norbert mentioned the impracticality of real-life scenario testing, proposing gamification and virtual environments to stimulate user engagement. Christin advocated for user simulations from logs and interviews, whereas Markus recommended long-term user observations to capture genuine behavior, noting that short-term studies might not accurately reflect normal user actions.

3 Outcome of the Working Groups

The on-site and online participants formed five thematic working groups in the two respective sessions of the workshop (see Table 1). Some working groups were part of both sessions (with changing participants), whereas others took place during only one session.

3.1 Multimodal Conversations

Participants: *Sachin Pathiyan Cherumanal (RMIT University), Johannes Kiesel (Bauhaus-Universität Weimar), Alieh Hajizadeh Saffar (Queensland University of Technology), Laurianne Sitbon (Queensland University of Technology), Paul Thomas (Microsoft).*

The working group looked at different aspects of online and offline evaluation of a multimodal conversational search system, focusing on input, output, and overall system performance. Depending on the input modalities, different channels may be required for evaluation. For instance, one could utilize the tone of voice from audio inputs, facial expressions such as smiles from video inputs, or even physiological and EEG signals [Ang et al., 2002; Ji et al., 2023, 2024; McDuff et al., 2021].[2] However, there is a need for research into which aspects of a search-oriented conversation such measurements correlate with (satisfaction, effectiveness, attention, task completion, happiness, ...), and care is needed to account for individual and cultural differences. This may call for further study of user models (e.g., represented by a logarithmic decay function, which is commonly used in certain offline IR metrics) and exploring its applicability to the offline evaluation of multimodal conversational search. While considering the need for developing metrics, it is also imperative to account for individual and cultural differences (e.g., Internet meme images presuppose knowledge of other elements). Some aspects such as comprehensibility can probably be adapted to different modalities (e.g. from readability to listenability).

Evaluation aspects specific to multimodal systems are (1) appropriateness of the system's own modality (are we using the best "embodiment", do we have the right appearance or tone of voice?), (2) appropriateness of output modality (were the best modalities chosen?) and (3) coordination of modalities (do the outputs across modalities duplicate or complement each other appropriately?). But even within a modality, the evaluation must distinguish different channels, e.g., voice and background music. Additionally, it is necessary to distinguish measurements for conversation quality and the individual aspects both at each turn of a conversation as well as across the whole session of the conversation. However, the importance of the aforementioned individual aspects may be task-specific. Moreover, the evaluation should also consider negative signals (e.g., indicating distraction or distress), which are potentially easier to detect than positive ones.

One particular danger that is especially (but not only) present when interacting with multimodal conversational systems is failing to account for the user's cognitive and social biases, which could potentially addressed through user simulation. Already when conducting user studies involving multimodal conversational systems one has to account for a user's cognitive and social biases and pre-existing beliefs. One possible approach to isolate bias and knowledge effects is to utilize user simulations in the evaluation process. To this end, the group discussed one particular idea: to use a (multimodal) language model as a user simulator as proposed in earlier works by

[2] Apple patented earbuds that allow to monitor brain activity (US patent number 20230225659).

Balog and Zhai [2023]. The group further discussed different ways a multimodal language model could be used as a user simulator that lacks certain knowledge (e.g., of events that happened after its training, or specifically removed knowledge), and then to instruct it to use the multimodal search engine to (re-)learn this knowledge. The simulated conversations can then be analyzed, although human input signals such as smiles or EEG would not be available unless the language models evolve in that direction.[3] Indeed, simulating the visual appearances of human searchers may not be that far away, as similar simulations are already being used to train medical staff.

3.2 Adapting to the User

Participants: *Marina Ernst (University of Koblenz), Marcel Gohsen (Bauhaus-Universität Weimar), Markus Nilles (Trier University), Oleg Zendel (RMIT University).*

The working group had a particular interest in the personalization of conversational search systems, distinguishing between short-term and long-term personalization. While short-term personalization should take into account, for example, the current emotional state, which may change from turn to turn, long-term personalization should establish a user model with slowly changing (or not changing at all) attributes such as occupation, interests or demographics. Both personalization scopes require a conversational agent to adapt to changes from turn to turn or across dialogs. To assess whether an agent adapts to changes (for better or worse), lexical or semantic features could be observed across a sequence of turns. The "adequacy" of the adaptation of the conversational system could be quantified by the satisfaction rated by humans. In any case, a conversational system would have to build an implicit or explicit model of the user in order to be able to react adequately. If an explicit user model has been created, the correctness can be verified directly. However, implicit user models demand proxy questions to verify if the model is accurate. For example, content recommendation effectiveness metrics could be proxy metrics to analyze the generated user model of a conversational agent.

3.3 Expectations of Human-likeness

Participants: *Vaibhav Adlakha (McGill University), Mateusz Dubiel (University of Luxembourg), Maik Fröbe (Friedrich-Schiller-Universität Jena), Satanu Ghosh (University of New Hampshire), Christin Kreutz (TH Mittelhessen), Abbas Pirmoradi (University of Regina).*

The working group began by discussing the characteristics of conversational agents that are related to the human likeness of conversational agents. The spectrum of human likeness ranges from automated dashboards that are obviously not human to some bots that appear to be completely human. The working group hypothesized that the degree to which a conversational agent appears to be human might impact which information needs users submit to the conversational agent. For example, users might prefer not to ask a question to humans they might feel somewhat embarrassed because the information need might appear stupid. On the contrary, users might feel more comfortable asking stupid questions in a conversational search system. LLMs are currently trained in human-to-human interactions. Therefore, LLMs might mimic human interaction. For

[3] After the workshop, OpenAI announced GPT-4o, with much improved capacities in that direction.

some tasks, it might, therefore, be more comfortable to interact with a system that can be clearly identified as non-humanoid. Interaction with a human-like system with a specific personality might be preferred for other tasks. In all cases, a bot should be transparent in acknowledging being a bot, similar to the California bot disclosure bill.[4]

Currently, many customer bots for banking or travel services are role-playing. Such systems can generate incorrect information[5] or be misused into performing actions for which an LLM in such web pages was not intended, e.g., rapping a song. One problem we see with current systems is their inability to admit to not knowing or having answers but fabricating a response nonetheless.

Existing evaluation measures which could be used for human-likeness (beyond the Turing test) if reference answers are available are BERTscore [Zhang et al., 2020], METEOR [Banerjee and Lavie, 2005], BLEU [Papineni et al., 2002] or ROUGE [Lin, 2004]. Still, these reference-answer based evaluations might not be able to cover all important aspects of conversational agents, e.g., aspects that can only be evaluated over multiple subsequent request response pairs, like patience, empathy, or transparently announcing that something is unknown or does not exist.

Finally, the working group brainstormed on potentially new measures for emotional aspects of conversations, as these metrics might enable conversational agents that understand not only content but also context and emotional dynamics, resulting in a more natural and satisfying user experience [Abbasian et al., 2024]. The brainstorming on potential new measures for this included "patience", "social sensitivity" (can I ask very stupid questions without feeling kind of embarrassed?), "domain/task suitability" (is the system helpful for the task?), "intention to adopt" (how likely are you to use this system?), and "social acceptance" (how likely would you be to use the system on a regular basis?).

3.4 Human-in-the-loop Evaluations

Participants: *Norbert Braunschweiler (Toshiba Europe Limited), Maik Fröbe (Friedrich-Schiller-Universität Jena), Christin Kreutz (TH Mittelhessen).*

The group started the discussion by noting that every system should aim for a unique selling point that should be the target of subsequent human-in-the-loop evaluations. Furthermore, evaluations are highly task-dependent, e.g., they substantially differ for tasks that require high accuracy versus tasks that require fast dynamic responses. Consequently, the human-in-the-loop setup must be adjusted to capture the essential aspects and metrics of the task. The working group subsequently brainstormed that an advertisement perspective could help shape human-in-the-loop evaluations, as advertisements must deliver the task-dependent unique selling point concisely and convincingly. This advertisement perspective emerged from the thought experiment in which one gets some search system and subsequently has to decide how to advertise this system in a highly competitive market. This perspective enables the development of a minimum viable experiment to evaluate a hypothesis (that, if successful, would come the advertisement if it confirms the unique selling point) while, at the same time, ensuring that the to-be-evaluated aspect is not yet addressed and/or evaluated (e.g., after a review on what existing systems try to advertise/sell).

[4] https://digitaldemocracy.calmatters.org/bills/ca_201720180sb1001

[5] https://www.forbes.com/sites/marisagarcia/2024/02/19/what-air-canada-lost-in-remarkable-lying-ai-chatbot-case/

Another important aspect of human-in-the-loop evaluations is measuring the effort required to switch to a new system. In reality, users already work with existing systems and might have adjusted and specialized their workflows. Consequently, the satisfaction of users must not only cover the satisfaction of the new system alone, but must also cover the satisfaction during the transition/familiarization with a new system given experience with the previous system (and every switch causes mental load). A practical approach could be that the new system initially mimics the previous system and evolves into the new system in small steps.

Finally, the group discussed how the research community can better document and share study designs, which is important because human-in-the-loop evaluations are very costly. Interesting approaches could be to discuss if pre-registrations could be implemented in our community (as in the field of psychology, but with the caveat that industry researchers might not be able to reveal all details). Furthermore, the discussion highlighted the importance of publishing the questionnaires as supplementary material and promoting a culture that is open to failures (e.g., with the ACL Workshop on Insights from Negative Results in NLP[6] as a positive example). There was a consensus that ethical considerations should be considered early on in the study design, as not everything that we can measure should be measured.

3.5 Trust in Conversational Systems

Participants: *Alexander Frummet (Universität Regensburg), Milad Momeni (University of Regina), Andrea Papenmeier (University of Twente), Alisa Rieger (TU Delft), Ines Zelch (Leipzig University).*

This group discussed how we can increase trust in IR systems. For example, the participants discussed the role of transparency. While this is often mentioned as a trust building facet, the implementation of this concept may not always be beneficial and can have undesirable effects. It could potentially lead to information overload or result in users becoming passive in their engagement with information. Furthermore, choice architectures that misuse trust or overwhelm users with transparency, such as ubiquitous cookie warnings, can have unintended consequences. To address these challenges, systems could consider directly asking users about their concerns, although users may be hesitant to trust such inquiries. Additionally, presenting metadata and integrating it into conversational interactions, perhaps through color schemes, could enhance trust. Some participants mentioned that the framing of information by both users and systems plays a significant role in shaping trust perceptions. For example, the experience that the formulation of a question can completely change the answer does not contribute to the user's trust in a conversational system. Transparency efforts, such as disclosing the sources of information in conversational search results and framing answers in a trustworthy manner, can also build trust. However, trust is influenced by various factors, including personal biases, the accuracy of prompts, and task-specific requirements for transparency. Systems should be designed to encourage users to seek information elsewhere if needed, while also ensuring appropriate levels of trust to facilitate efficient information retrieval without constant skepticism. Collaborative online evaluation methods could further enhance credibility and reliability. Yet, the complexity of trust extends beyond individual systems, requiring consideration of consistency across systems and tasks. Ul-

[6] https://insights-workshop.github.io/

timately, fostering appropriate levels of trust is essential for enabling users to effectively engage with systems and accomplish their objectives.

Given the 'black box' nature of conversational AI systems, where users cannot observe the provenance of information, addressing this issue is critical for building trust. This also ties closely with the principles of explainable AI, which advocates for systems that are transparent about their decision-making processes. Discussing the intersection of transparency, explainability, and trust could provide valuable insights for enhancing user trust in conversational systems. Does it warrant further discussion as a significant aspect of trust in these systems?

4 Conclusion

The workshop on Search-Oriented Conversational Artificial Intelligence (SCAI) at CHIIR 2024 facilitated discussions among participants from diverse backgrounds. One keynote and three panels provided an overview of the topic. Thematic working groups explored various aspects of conversational search system evaluation and human-centered metrics. The five groups focused on (1) multimodal conversations (both multimodal in input and in output), discussing how measurements can be correlated with user satisfaction, effectiveness, and attention, as well as multimodal user simulation; (2) adapting to the user (both short-term and long-term personalization), discussing the role of user models in assessing system adaptation and satisfaction; (3) expectations of human-likeness, discussing the spectrum of human-likeness and issues regarding transparency and trustworthiness of conversational agents, especially when fabricating information instead of admitting to not know; (4) human-in-the-loop evaluations, discussing focused evaluations of the selling points of systems, system switching, and methods for documenting and sharing study designs effectively; and (5) trust in conversational systems, discussing choice architectures, ethical considerations and transparency as well as user engagement. Overall, the workshop provided valuable insights into evaluating conversational search systems from a user-centric perspective. The discussions and outcomes outlined in this paper pave the way for future research in this direction.

Acknowledgments

Marcel Gohsen is supported by the Thüringer Ministerium für Wirtschaft, Wissenschaft und Digitale Gesellschaft (TMWWDG) under grant agreement 5575/10-5 (MetaReal). Ines Zelch is supported by the European Union's Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU).

A Authors and Affiliations

Workshop organizers

- Alexander Frummet, Universität Regensburg, Germany, alexander.frummet@ur.de
- Andrea Papenmeier, University of Twente, Netherlands, a.papenmeier@utwente.nl
- Maik Fröbe, Friedrich-Schiller-Universität Jena, Germany, maik.froebe@uni-jena.de
- Johannes Kiesel, Bauhaus-Universität Weimar, Germany, johannes.kiesel@uni-weimar.de

Other authors

- Vaibhav Adlakha, Mila-Quebec AI Institute, Canada, vaibhav.adlakha@mila.quebec
- Norbert Braunschweiler, Toshiba Europe Limited, UK, norbert.braunschweiler@toshiba.eu
- Mateusz Dubiel, University of Luxembourg, Luxembourg, mateusz.dubiel@uni.lu
- Satanu Ghosh, University of New Hampshire, US, satanu.ghosh@unh.edu
- Marcel Gohsen, Bauhaus-Universität Weimar, Germany, marcel.gohsen@uni-weimar.de
- Christin Kreutz, TH Mittelhessen, Germany, ckreutz@acm.org
- Milad Momeni, University of Regina, Canada, miladmomeni@uregina.ca
- Markus Nilles, Trier University, Germany, nillesm@uni-trier.de
- Sachin Pathiyan Cherumanal, RMIT University, Australia, s3874326@student.rmit.edu.au
- Abbas Pirmoradi, University of Regina, Canada, abbaspirmorady@uregina.ca
- Paul Thomas, Microsoft, Australia, pathom@microsoft.com
- Johanne R. Trippas, RMIT University, Australia, j.trippas@rmit.edu.au
- Ines Zelch, Leipzig University, Germany, ines.zelch@uni-leipzig.de
- Oleg Zendel, RMIT University, Australia, oleg.zendel@student.rmit.edu.au

References

Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, Olivier Gevaert, Li-Jia Li, Ramesh Jain, and Amir M. Rahmani. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative ai. *npj Digital Medicine*, 7(1): 82, Mar 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01074-z. URL https://doi.org/10.1038/s41746-024-01074-z.

Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of the International Conference on Spoken Language Processing*, 2002.

Krisztian Balog and ChengXiang Zhai. User simulation for evaluating information access systems. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP '23, page 302–305, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400704086. doi: 10.1145/3624918.3629549. URL https://doi.org/10.1145/3624918.3629549.

Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss, editors, *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics, 2005. URL https://aclanthology.org/W05-0909/.

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. doi: 10.1145/3121050.3121111. URL https://doi.org/10.1145/3121050.3121111.

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. ISBN 978-1-948087-75-9. 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.

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. doi: 10.1145/3627508.3638310. URL https://doi.org/10.1145/3627508.3638310.

Kaixin Ji, Damiano Spina, Danula Hettiachchi, Flora Dilys Salim, and Falk Scholer. Examining the impact of uncontrolled variables on physiological signals in user studies for information processing activities. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1971–1975, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591981. URL https://doi.org/10.1145/3539618.3591981.

Kaixin Ji, Danula Hettiachchi, Flora D. Salim, Falk Scholer, and Damiano Spina. Characterizing information seeking processes with multiple physiological signals. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, New York, NY, USA, 2024. ACM. doi: 10.1145/3626772.3657793.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Daniel McDuff, Paul Thomas, Kael Rowan, Nick Craswell, and Mary Czerwinski. Do affective cues validate behavioural metrics for search? In *SIGIR 2021*. ACM, July 2021. URL https://www.microsoft.com/en-us/research/publication/do-affective-cues-validate-behavioural-metrics-for-search/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational

Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

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. doi: 10.1145/3477495.3531700. URL https://doi.org/10.1145/3477495.3531700.

Johanne R. Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. What do users really ask large language models? an initial log analysis of google bard interactions in the wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*, SIGIR '24, New York, NY, USA, 2024. ACM. doi: 10.1145/3626772.3657914.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.