

An Evaluation Framework for Conversational Information Retrieval Using User Simulation*

Xiao Fu^{1,2}, Aldo Lipani² and Noriko Kando¹

¹National Institute of Informatics (NII), Tokyo 101-8430, Japan

²University College London (UCL), Gower Street, London, WC1E 6BT, UK

Abstract

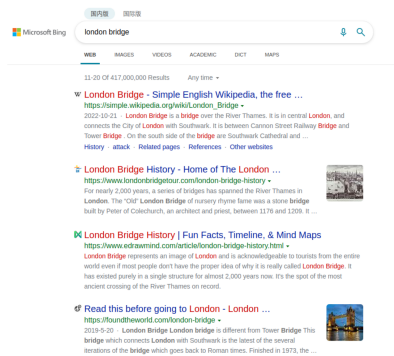
Recent advancements in the field of Conversational Information Retrieval (CIR) have increased the demand for more sophisticated modelling and evaluation approaches. This paper introduces a novel framework for user simulation in CIR, aimed at enhancing the modeling and evaluation of user interactions. Additionally, this study explores the potential integration of large language models (LLMs) within this domain. Furthermore, the paper anticipates future developments in CIR, particularly in the context of the widespread use of LLMs. The study emphasizes the necessity for robust evaluation paradigms that go beyond traditional methods to effectively measure the success of CIR systems.

Keywords

Conversation Information Retrieval, Evaluation, User Simulation, Large Language Models

1. Introduction

Recent advancements in Machine Learning (ML), Natural Language Processing (NLP), and the proliferation of smart devices have significantly enhanced conversational AI. This progress has led to a variety of commercial conversational services that enable natural spoken interactions, thereby increasing the demand for more human-centric approaches in *information retrieval (IR)* [1]. The objective of *Conversational Information Retrieval (CIR)* is to facilitate information seeking through multi-turn natural language dialogues between users and systems, a longstanding yet challenging area of research [2].



U: "I love candy, what's a good brand?"
S: "I don't know how good they are, but Dylan's Candy Bar has a chain of candy shops in various cities."
U: "Oh, they do? What kind of candy do they sell?"
S: "I don't know, really, but they also are a supplier of candy."
U: "Oh I see, what kind of candy do they offer?"
S: "It stocks over 7,000 candies from across the world."
U: "I see, do they have nougat? I love nougat!"
S: "I couldn't really say, but they're owned by the daughter of Ralph Lauren, Dylan Lauren."

Figure 1: Comparative Example: Web Search (Left) vs. CIR Dialogue (Right)

Figure 1 illustrates a comparison between a web search in a traditional IR system and a dialogue in CIR. Traditional IR systems typically focus solely on the user's current query, treating each query as an independent event. These systems do not account for the influence of previous queries on the current search.

UM-CIR 2024: The 1st Workshop on User Modelling in Conversational Information Retrieval, December 12, 2024, Tokyo, Japan

✉ xiao.fu.20@ucl.ac.uk (X. Fu); aldo.lipani@ucl.ac.uk (A. Lipani); Noriko.Kando@nii.ac.jp (N. Kando)

🆔 0000-0003-4676-8608 (X. Fu); 0000-0002-3643-6493 (A. Lipani); 0000-0002-2133-0215 (N. Kando)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



In contrast, CIR systems emphasize the importance of context in the retrieval process. Previous queries and responses influence the current response. The second part of Figure 1 shows a conversation between a user and a CIR system from the FaithDial dataset [3], where the interaction is more natural, and queries form sequences akin to conversations. For instance, the word *they* in the second turn of the conversation refers to the shops mentioned in the system’s previous response.

This shift in communication style introduces several differences between traditional IR and CIR. The conversation-like query format in CIR allows for richer, more interactive responses. Firstly, CIR systems can provide more detailed responses compared to traditional IR, which typically returns documents directly. CIR systems can refine information needs and the search space through user or system revelation. Secondly, users have more strategic options in their interactions. If unsatisfied with a response, users can ask follow-up questions to refine their search or request additional information. This necessitates the introduction of ad-hoc search methods (such as session-based and task-based search) in CIR, enabling the system to refine the search space based on the conversation, querying related sections instead of the entire database. Thirdly, context is crucial in CIR for understanding user queries, unlike traditional IR where queries are treated independently. For example, the meaning of pronouns can change depending on their order in a conversation.

Currently, CIR systems are predominantly used for simple tasks, as they are not sufficiently effective for complex and exploratory information-seeking conversations. Nonetheless, advancements in key components, particularly in ML, are driving a trend within the IR community towards more conversational methodologies [4, 5].

Despite the progress, several critical questions remain unresolved, presenting challenges to the development of CIR systems. Zamani et al. [5] identifies four key directions in the CIR field with potential for significant advancements:

1. *Modelling and Producing Conversational Interactions*
2. *Result Presentation*
3. *Exploring Under-Explored Conversational Tasks*
4. *Measuring Interaction Success and Evaluation*

This paper primarily focuses on the first and last directions, which are both challenging and interconnected. The first direction involves modelling and producing conversational interactions, addressing the uncertainty of information needs between multiple agents through a mixed-initiative approach. Additionally, understanding long-term conversational interactions and addressing associated privacy and transparency concerns are critical topics in this direction.

The final direction pertains to the measurement and evaluation of CIR. Both academia and industry face limitations due to the absence of a robust definition of success in this field. As CIR continues to evolve, there is an urgent need for an evaluation paradigm that transcends the traditional Cranfield Paradigm, especially in frontier tasks such as personalized evaluation and transparency.

These two directions are highly interconnected. The definition of success relies on the proper modelling of conversational interactions, while precise measures support the modelling process. This paper introduces a new potential contribution to this field: introducing user simulation to CIR.

In this paper, we present a framework for the automated evaluation of CIR systems via user simulation. This paper also includes the recent advancements within this framework, while exploring the encountered opportunities and challenges. Section 3 details the two-stage user simulation prototype, which amalgamates psychological principles and ML to enhance explainability, and utilizes Large Language Models (LLMs) to sustain high performance. Section 4 focuses on the application of this prototype in assessing CIR systems. The methodology proposed aims to connect user simulation with well-established CIR conversation modelling approaches, such as effort and cost, and seeks to align the simulations closely with real user interactions through indirect assessment. Section 5 discusses the challenges and future perspectives in this field.

2. Related Works

Conversational search, a well-established field, continues to be a popular research topic due to its relevance for modern devices with small or no screens [1].

2.1. Evaluation

Despite significant progress, the evaluation of CIR remains relatively underdeveloped [6, 7, 8]. While CIR extends functionalities from traditional IR systems [9], many studies still rely on conventional metrics such as Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (nDCG), and Mean Reciprocal Rank (MRR) [7, 10, 11]. At the heart of these metrics is the concept of relevance, where the documents retrieved are fitting to the topics (keywords) sought by the user [12]. Additionally, metrics from other domains like ROUGE and BLEU are also utilized [6, 13, 14]. However, recent research indicates that without real user interaction, these metrics may not accurately reflect user satisfaction [15, 16].

User satisfaction, a highly abstract and subjective measure, pertains to the overall experience and interaction with a search system [10, 8]. It is defined as the fulfillment users achieve in pursuing their goals [17]. Extensive research has been conducted to understand this measure [10, 15, 18, 19, 20, 21, 22]. Studies like Yilmaz et al. [18] offer various metrics to reflect user satisfaction, considering factors such as effort. Some studies further break down user satisfaction into query-level satisfactions [19, 20], although others, such as Järvelin et al. [21], argue that summing query-level satisfaction misses the contextual journey where user queries are interconnected. Unlike traditional search systems, conversational search systems allow users to ask follow-up questions to refine their answers [22].

Despite widespread adoption, measuring user satisfaction remains an open question. Many studies rely on real-time user participation to gather feedback, providing fresh and realistic insights but requiring significant resources and participant incentives [10, 15]. Alternatively, satisfaction prediction proxies using deep learning models offer a computational approach, overcoming temporal and spatial constraints but demanding high-quality computational resources and datasets [8].

Emerging trends in user simulation offer promising solutions to these challenges [5, 2], though prior studies are not yet comprehensive. Gao et al. [23] noted that earlier simulators relied on randomly generated scenarios to approximate users' states of mind. While such models prove beneficial in specific domains like recommender systems [24], issues of explainability and scrutability remain unresolved. Our proposed framework addresses these issues by incorporating a user-profile-based personalized simulation approach, facilitating easier alignment with actual user behaviours and providing a scrutable means to control the simulation by modifying textual user profiles.

2.2. User Simulation

Azzopardi et al. [25] define simulation as the imitation of the operation of real-world phenomena. Simulations enable detailed experimental design and control tailored to specific research questions. These high-level controls allow for experiments with user simulations to be conducted with several advantages [25].

Firstly, *what-if* experiments can be performed by setting up different *scenarios* [26, 27]. Secondly, user simulators ensure the repeatability of experimental results. Additionally, user simulations can achieve these benefits at a low cost [28].

In the IR community, user simulation methods are primarily divided into cognitive and statistical approaches [29]. Cognitive approaches were among the first used in this field. Belkin [30] described users, information resources, and IR models, characterizing users by their objectives, problems, and knowledge. Subsequent studies expanded on this foundation [31, 32, 33].

In contrast, statistical approaches focus on analyzing user behaviours and satisfaction [29, 34, 35, 36, 37]. These approaches underpin early user simulators based on statistical models [38, 39]. Although these simulators heavily relied on corpora, they faced limitations such as the diversity of user intent [29].

Table 1

Performance of models trained with priming factors in predicting users' actions, while AUC refers to the area under P-R curves. Adopted from Fu and Lipani [46].

Action	Dataset	Precision	Recall	F1	AUC
	All	0.844	0.574	0.684	0.835
	Topi [47]	0.977	0.951	0.963	0.995
	TREC ¹	1.000	0.625	0.769	0.800
	Cran [28]	0.795	0.875	0.833	0.883
	QReCC [48]	0.953	0.997	0.974	0.990
	ORC [49]	0.965	0.979	0.972	0.997
Following up	ORC	0.656	0.779	0.712	0.698
Switching topic	Topi	0.569	0.162	0.252	0.476

Agenda-based user simulations are popular due to their realistic responses and straightforward dialogue strategies [40, 41, 42]. The latest trend involves employing deep learning models, including adversarial generative approaches [43], reinforcement learning [44], and inverse reinforcement learning to abstract knowledge from data [45].

Evaluating CIR with user simulations is becoming a key trend, enabling efficient and cost-effective evaluation at various levels of CIR [2]. Our proposed prototype integrates benefits from the aforementioned approaches. Initially, it simulates user behaviour guided by statistical signals derived from the context, enhancing explainability. Subsequently, the prototype employs deep learning models, utilizing textual user profiles to produce realistic and diverse responses within controlled parameters, thereby enabling further exploration of the target system. The details of the user simulation prototype are in the next section.

3. Modelling and Simulating Users

In this section, we introduce an two-stage prototype for constructing a robust user simulation for CIR.

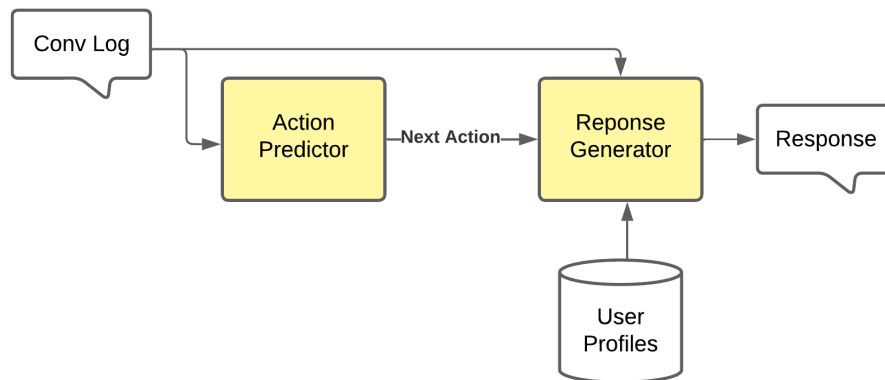


Figure 2: The basic structure of the user simulation.

The main structure of the prototype is divided into two parts to control simulated users: the Action Predictor and the Response Generator, as illustrated in Fig 2.

3.1. Modeling Actions

The Action Predictor controls the simulated users’ actions during the current conversation. This concept originates from the psychology community’s notion of priming, which refers to the unconscious influence of past experiences on current performance or behaviour [50, 51]. Many studies have established models to explain this mechanism [52, 50]. For instance, Tulving et al. [53] conducted an experiment where participants viewed a list of 96 words and later completed graphemic word fragments both one hour and seven days after studying the list. The results demonstrated a significant influence of the word list on subsequent tests.

As shown in Table 1, results from Fu and Lipani [46] demonstrate the potential of predicting users’ next actions. The benefits of using lexical and textual patterns to predict users’ next actions include low cost and ease of interpretation, which are valuable for evaluation.

In the current prototype, three user actions are modelled based on the dataset available. **Stopping** is defined as the action where users opt to end a conversation, typically indicating the conclusion of the exchange. This action is critical for evaluating effects such as the principle of least effort [54]. Sessions of conversation often include **Following up** on queries that build upon previous interactions, acknowledging missing contexts and references to earlier discussed topics [49]. As noted by Stede and Schlangen [55], an inquisitive user engaged in an ongoing dialogue may express interest in further related subjects as a response to the information provided. **Switching topics** is commonly seen in information-seeking dialogues, especially when using search systems for data acquisition [56].

As output from the Action Predictor, a general action as described above will be predicted, with the details elaborated in the Response Generator.

3.2. Personalized Responses

The Response Generator will generate realistic and diversified responses based on the conversation, user profiles and the predicted actions from the Action Predictor

LLMs such as GPTs and Llama, renowned for their sophisticated natural language processing capabilities, present unique opportunities to enhance Conversational Systems through mechanisms like pre-training, fine-tuning, and prompting [57]. The ability of LLMs to mimic diverse demographic characteristics offers a novel approach to simulating user behaviour and preferences [58]. High-quality user simulations, which closely mirror real user behaviour distributions, can significantly advance CRS development, currently dependent on real data for training, with its inherent constraints and disadvantages.

Ramos et al. [59] offers a valuable method for generating user profiles from the Amazon dataset, introducing a more compact style of personalization into the user simulator. Table 2 demonstrates the agreement between simulated and real users’ responses to the same items in the Amazon dataset. Since the simulated users are based on LLMs and prompts with user profiles, the results suggest the potential of generating personalized responses based on textual user profiles.

Table 2

Agreements (Cohen’s K_c , Randolph’s K_r and Krippendorff’s α) between simulated users and actual users in response. Cohen’s index considers the marginal distribution of categories, Randolph’s assumes a uniform distribution, and Krippendorff’s offers a broader approach to assessing agreement. Fair agreements for each metric (>0.2) are marked as bold. Three simulation methods are evaluated: Real UP, where users are simulated based on their profiles; Rand UP, involving users simulated with random profiles; and Rand Sc, where scores are randomly generated based on the dataset’s historical distribution.

Setting	K_c	K_r	α
Real UP	0.20	0.34	0.20
Rand UP	0.14	0.26	0.14
Rand Sc	-0.03	0.14	-0.03

In this simulation, responses are tailored based on user profiles, which consist of concise text that

summarizes the attributes of users in a few succinct sentences. This can include motivations for task-oriented CIR systems. Furthermore, the Response Generator is tasked with handling clarification questions posed by the CIR system.

At the conclusion of this phase, the user simulator is equipped to interact with CIR systems. The subsequent section proposes a linkage as the remaining component of this framework, specifically addressing the evaluation of the CIR system using this user simulator, given the absence of a direct indicator from the user simulation on the quality of the target CIR system.

4. Evaluation of CIR

This section discusses potential methods for applying evaluation tasks based on the user simulation prototype established in Section 3.

As introduced in Section 2, the target of evaluation originates from modelling users and conversations. According to Yilmaz et al. [18], user satisfaction can be reflected by the effort exerted. Effort also forms the foundation of modelling user actions, particularly stopping behaviours.

In the IR community, several studies have been conducted to depict stopping behaviours [60, 61, 62, 63]. These studies aim to quantify the feeling of having "enough." For instance, users may decide to stop a conversation when they feel frustrated or satisfied.

A previous study by Fu and Lipani [46] provided a reliable method for predicting stopping behaviours. The subsequent step is to explore the relationship between each stopping point and user satisfaction.

4.1. Evaluating User Simulation

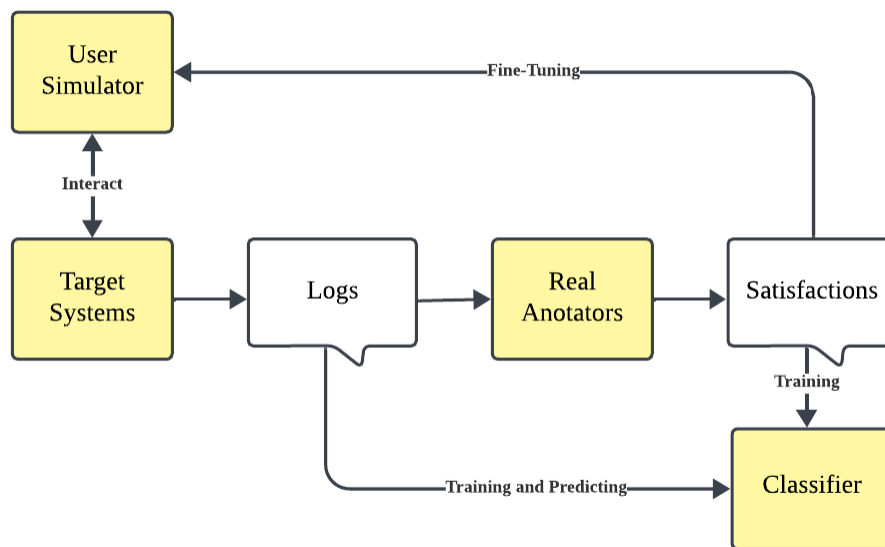


Figure 3: The basic flow of the evaluation with the user simulator.

The final part of the evaluation focuses on assessing the user simulation. To effectively perform evaluation for CIR, the user simulation must exhibit not only a diversity of reasonable responses but also a high alignment with real users, particularly in reflecting satisfaction or frustration.

Following the study by Fu et al. [28], where direct and indirect assessments in CIR show substantial agreement, the alignment between user simulation and real users can be evaluated. Real users will review the conversations between the user simulation and the system to determine if the simulated user appears satisfied.

Figure 3 illustrates the operational flow within the evaluation component of the framework. In this component, a classifier, integrated with the user simulator, is utilized to predict user satisfaction. This classifier is aligned with annotations from real users. Feedback from these real annotators is employed to train both the user simulator and the classifier. The user simulator aims to accurately mimic real user behaviour through the Actions Predictor and the Response Generator. Similarly, the classifier is trained to align its judgments with those of real users regarding the same conversation.

During the development of this framework, numerous emerging trends were observed, particularly in LLMs. These observations are presented in the following section.

5. Towards Future

This represents a significant transformation since 2021, as the term *Large Language Models* (LLMs) has gained popularity. This shift has introduced both opportunities and challenges for contemporary research.

5.1. CIR and RAG

The integration of LLMs is not limited to the IR community; the ML community is also embracing IR techniques. A prominent development in this space is Retrieval-Augmented Generation (RAG), which enhances LLMs in domain-specific or knowledge-intensive tasks [64].

RAG involves multiple retrieval processes to enrich the context, going beyond traditional single retrieval methods. For instance, Self-RAG [65] refines the RAG framework by enabling LLMs to actively determine the optimal moments and content for retrieval, thus improving the efficiency and relevance of the sourced information. The key aspect here is not merely multiple retrievals, but the reliance on the judgment of LLMs, indicating that LLMs can further participate in the processing with minimal human intervention.

For the evaluation of RAG, integrating typical LLMs alone may not suffice. A realistic inquiry is how agents based on LLMs can assess the responses from RAG systems that incorporate retrieved documents. One feasible approach is using RAG to evaluate itself. Here, the crucial aspects include not only the quality of the conversation and the retrieval process but also *how effectively the documents are presented*.

Moreover, in the specialized domain of CIR, where the systems are relatively light, LLMs can still serve as experts. The following section provides an example.

5.2. Should We Ask LLMs First?

The TREC Interactive Knowledge Assistance Track (iKAT) builds upon the foundational work of the TREC Conversational Assistance Track (CASt) [66], with a key difference being the addition of personal context for each user in the dataset. The primary task remains similar to TREC CASt—retrieving and ranking documents from the corpus at each turn of the given conversations.

The best performance in iKAT 2023 introduced a novel approach [67]. In this approach, the LLM generates an initial answer to the user's query based on the context of the conversation and the user profile. This answer is derived through reasoning over the context and the user's profile, but it is not grounded in the documents within the collection. Subsequently, the LLM generates a set of five queries to achieve this answer.

The controversial aspect of this approach is using the LLM-generated answer as the target without initial retrieval, followed by employing an IR system to achieve it. This process assumes that LLMs' answers are sufficiently accurate. Alternatively, it suggests that the documents have likely been exposed to the LLMs, raising concerns of potential data leakage.

5.3. How Can We Go Beyond Our Knowledge Borders?

Training LLMs from scratch is a challenging task for most research groups due to high costs and the lack of storage and computational resources. The most common practice involves fine-tuning a public base version of LLMs and accessing them via APIs. Top LLMs are trained on a substantial portion of internet text documents, making it nearly impossible to prevent data leakage once any public data is used in a study.

Furthermore, the widespread use of LLMs will inevitably introduce LLM-generated text back into the internet, posing a significant challenge that has already raised considerable concerns within the community.

6. Conclusion

This paper proposes a framework for evaluating CIR systems automatically using a user simulator prototype. The framework comprises two main components:

1. A prototype of user simulation that leverages the advancements from both psychology and ML fields to conduct realistic and scrutable simulations targeted at CIR systems.
2. A component that employs sophisticated conversation modelling concepts from the IR community to provide reasonable feedback aimed at predicting user satisfaction alongside the user simulation prototype.

In addition to the framework, this study also presents emerging trends observed with the promising development of LLMs. The advent of systems such as RAG introduces both opportunities and challenges. It raises concerns that the current practices within the community utilizing LLMs may lead to increased data leakages.

References

- [1] J. Gao, C. Xiong, P. Bennett, Recent advances in conversational information retrieval, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 2421–2424. URL: <https://doi.org/10.1145/3397271.3401418>. doi:10.1145/3397271.3401418.
- [2] J. Gao, C. Xiong, P. Bennett, N. Craswell, Neural approaches to conversational information retrieval, volume 44, Springer Nature, 2023.
- [3] N. Dziri, E. Kamaloo, S. Milton, O. Zaiane, M. Yu, E. Ponti, S. Reddy, Faithdial: A faithful benchmark for information-seeking dialogue, arXiv preprint, arXiv:2204.10757 (2022). URL: <https://arxiv.org/abs/2204.10757>.
- [4] G. Penha, C. Hauff, Challenges in the evaluation of conversational search systems., in: Converse@KDD, 2020.
- [5] H. Zamani, J. R. Trippas, J. Dalton, F. Radlinski, et al., Conversational information seeking, Foundations and Trends® in Information Retrieval 17 (2023) 244–456.
- [6] A. Lipani, B. Carterette, E. Yilmaz, How am i doing?: Evaluating conversational search systems offline, ACM Trans. Inf. Syst. 39 (2021). doi:10.1145/3451160.
- [7] J. Jiang, A. Hassan Awadallah, R. Jones, U. Ozertem, I. Zitouni, R. Gurunath Kulkarni, O. Z. Khan, Automatic online evaluation of intelligent assistants, in: Proceedings of the 24th International Conference on World Wide Web, WWW '15, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2015. doi:10.1145/2736277.2741669.
- [8] J. I. Choi, A. Ahmadvand, E. Agichtein, Offline and online satisfaction prediction in open-domain conversational systems, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19, Association for Computing Machinery, New York, NY, USA, 2019. doi:10.1145/3357384.3358047.

- [9] A. Anand, L. Cavedon, H. Joho, M. Sanderson, B. Stein, Conversational Search (Dagstuhl Seminar 19461), Dagstuhl Reports 9 (2020). doi:10.4230/DagRep.9.11.34.
- [10] J. Kiseleva, K. Williams, A. Hassan Awadallah, A. C. Crook, I. Zitouni, T. Anastasakos, Predicting user satisfaction with intelligent assistants, in: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16, Association for Computing Machinery, New York, NY, USA, 2016. doi:10.1145/2911451.2911521.
- [11] J. Dalton, C. Xiong, J. Callan, Trec cast 2019: The conversational assistance track overview, 2020.
- [12] T. Saracevic, Relevance reconsidered, in: Proceedings of the second conference on conceptions of library and information science (CoLIS 2), 1996, pp. 201–218.
- [13] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: A method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, USA, 2002. doi:10.3115/1073083.1073135.
- [14] E. Reiter, A structured review of the validity of BLEU, Computational Linguistics 44 (2018). doi:10.1162/coli_a_00322.
- [15] J. Kiseleva, K. Williams, J. Jiang, A. Hassan Awadallah, A. C. Crook, I. Zitouni, T. Anastasakos, Understanding user satisfaction with intelligent assistants, in: Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval, CHIIR '16, Association for Computing Machinery, New York, NY, USA, 2016. doi:10.1145/2854946.2854961.
- [16] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, J. Pineau, How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016. doi:10.18653/v1/D16-1230.
- [17] D. Kelly, Methods for evaluating interactive information retrieval systems with users, Foundations and Trends® in Information Retrieval 3 (2009). doi:10.1561/15000000012.
- [18] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, P. Bailey, Relevance and effort: An analysis of document utility, in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14, Association for Computing Machinery, New York, NY, USA, 2014, pp. 91–100. doi:10.1145/2661829.2661953.
- [19] J. Kiseleva, E. Crestan, R. Brigo, R. Dittel, Modelling and detecting changes in user satisfaction, in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14, Association for Computing Machinery, New York, NY, USA, 2014, pp. 1449–1458. doi:10.1145/2661829.2661960.
- [20] J. Kiseleva, J. Kamps, V. Nikulin, N. Makarov, Behavioral dynamics from the serp's perspective: What are failed serps and how to fix them?, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 1561–1570. doi:10.1145/2806416.2806483.
- [21] K. Järvelin, S. L. Price, L. M. L. Delcambre, M. L. Nielsen, Discounted cumulated gain based evaluation of multiple-query ir sessions, in: C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, R. W. White (Eds.), Advances in Information Retrieval, Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 4–15.
- [22] A. Al-Maskari, M. Sanderson, P. Clough, The relationship between ir effectiveness measures and user satisfaction, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, Association for Computing Machinery, New York, NY, USA, 2007, pp. 773–774. doi:10.1145/1277741.1277902.
- [23] J. Gao, M. Galley, L. Li, Neural approaches to conversational ai, in: The 41st international ACM SIGIR conference on research & development in information retrieval, 2018, pp. 1371–1374.
- [24] S. Zhang, K. Balog, Evaluating conversational recommender systems via user simulation, in: Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining, 2020, pp. 1512–1520.
- [25] L. Azzopardi, K. Järvelin, J. Kamps, M. D. Smucker, Report on the sigir 2010 workshop on the

- simulation of interaction, *SIGIR Forum* 44 (2011) 35–47. URL: <https://doi.org/10.1145/1924475.1924484>. doi:10.1145/1924475.1924484.
- [26] M. I. Kellner, R. J. Madachy, D. M. Raffo, Software process simulation modeling: why? what? how?, *Journal of Systems and Software* 46 (1999) 91–105.
- [27] K. Balog, D. Maxwell, P. Thomas, S. Zhang, Report on the 1st simulation for information retrieval workshop (sim4ir 2021) at sigir 2021, *SIGIR Forum* 55 (2022). URL: <https://doi.org/10.1145/3527546.3527559>. doi:10.1145/3527546.3527559.
- [28] X. Fu, E. Yilmaz, A. Lipani, Evaluating the cranfield paradigm for conversational search systems, in: *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, 2022, pp. 275–280.
- [29] P. Erbacher, L. Soulier, L. Denoyer, State of the art of user simulation approaches for conversational information retrieval, *arXiv preprint arXiv:2201.03435* (2022).
- [30] N. Belkin, Cognitive models and information transfer, *Social Science Information Studies* 4 (1984) 111–129. URL: <https://www.sciencedirect.com/science/article/pii/014362368490070X>. doi:[https://doi.org/10.1016/0143-6236\(84\)90070-X](https://doi.org/10.1016/0143-6236(84)90070-X), special Issue Seminar on the Psychological Aspects of Information Searching.
- [31] C. C. Kuhlthau, Inside the search process: Information seeking from the user’s perspective, *J. Am. Soc. Inf. Sci.* 42 (1991) 361–371.
- [32] P. Ingwersen, K. Järvelin, The Turn: Integration of Information Seeking and Retrieval in Context, 2005. doi:10.1007/1-4020-3851-8.
- [33] D. Ellis, A behavioral approach to information retrieval system design, *J. Doc.* 45 (1989) 171–212. URL: <https://doi.org/10.1108/eb026843>. doi:10.1108/eb026843.
- [34] N. Craswell, O. Zoeter, M. Taylor, B. Ramsey, An experimental comparison of click position-bias models, in: *Proceedings of the international conference on Web search and web data mining, WSDM ’08*, ACM, New York, NY, USA, 2008, pp. 87–94. URL: <http://doi.acm.org/10.1145/1341531.1341545>. doi:10.1145/1341531.1341545.
- [35] O. Chapelle, Y. Zhang, A dynamic bayesian network click model for web search ranking, in: *In WWW*, 2009.
- [36] G. E. Dupret, B. Piwowarski, A user browsing model to predict search engine click data from past observations., in: *SIGIR ’08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, 2008, pp. 331–338. URL: <http://portal.acm.org/citation.cfm?id=1390334.1390392>. doi:10.1145/1390334.1390392.
- [37] A. Chuklin, P. Serdyukov, M. Rijke, Modeling clicks beyond the first result page, 2013. doi:10.1145/2505515.2507859.
- [38] W. Eckert, E. Levin, R. Pieraccini, User modeling for spoken dialogue system evaluation, 1997 *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings* (1997) 80–87.
- [39] K. Scheffler, S. J. Young, Probabilistic simulation of human-machine dialogues, 2000 *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings* (Cat. No.00CH37100) 2 (2000) II1217–II1220 vol.2.
- [40] K. Komatani, S. Ueno, T. Kawahara, H. Okuno, User modeling in spoken dialogue systems to generate flexible guidance, *User Modeling and User-Adapted Interaction* 15 (2005) 169–183. doi:10.1007/s11257-004-5659-0.
- [41] J. Schatzmann, B. Thomson, K. Weillhammer, H. Ye, S. Young, Agenda-based user simulation for bootstrapping a pomdp dialogue system, 2007, pp. 149–152. doi:10.3115/1614108.1614146.
- [42] X. Li, Z. C. Lipton, B. Dhingra, L. Li, J. Gao, Y.-N. Chen, A user simulator for task-completion dialogues., *CoRR abs/1612.05688* (2016). URL: <http://dblp.uni-trier.de/db/journals/corr/corr1612.html#LiLDLGC16>.
- [43] X. Zhao, L. Xia, Z. Ding, D. Yin, J. Tang, Toward simulating environments in reinforcement learning based recommendations, *ArXiv abs/1906.11462* (2019).
- [44] X. Chen, S. Li, H. Li, S. Jiang, Y. Qi, L. Song, Generative adversarial user model for reinforcement learning based recommendation system, in: K. Chaudhuri, R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine*

- Learning Research*, PMLR, 2019, pp. 1052–1061. URL: <https://proceedings.mlr.press/v97/chen19f.html>.
- [45] S. Chandramohan, M. Geist, F. Lefèvre, O. Pietquin, User simulation in dialogue systems using inverse reinforcement learning, in: INTERSPEECH, 2011.
- [46] X. Fu, A. Lipani, Priming and actions: An analysis in conversational search systems, SIGIR, 2023.
- [47] V. Adlakha, S. Dhuliawala, K. Suleman, H. de Vries, S. Reddy, Topi-OCQA: Open-domain conversational question answering with topic switching, volume 10, 2022, pp. 468–483. URL: https://doi.org/10.1162/tacl_a_00471. doi:10.1162/tacl_a_00471. arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00471/2008126/tacl_a_00471.pdf.
- [48] R. Anantha, S. Vakulenko, Z. Tu, S. Longpre, S. Pulman, S. Chappidi, Open-domain question answering goes conversational via question rewriting, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2021).
- [49] C. Qu, L. Yang, C. Chen, M. Qiu, W. B. Croft, M. Iyyer, Open-Retrieval Conversational Question Answering, in: SIGIR, 2020.
- [50] D. L. Schacter, R. L. Buckner, Priming and the brain, *Neuron* 20 (1998) 185–195.
- [51] J. A. Bargh, T. L. Chartrand, The mind in the middle: A practical guide to priming and automaticity research. (2014).
- [52] C. Janiszewski, R. S. Wyer Jr, Content and process priming: A review, *Journal of consumer psychology* 24 (2014) 96–118.
- [53] E. Tulving, D. L. Schacter, H. A. Stark, Priming effects in word-fragment completion are independent of recognition memory., *Journal of experimental psychology: learning, memory, and cognition* 8 (1982) 336.
- [54] G. K. Zipf, Human behavior and the principle of least effort: An introduction to human ecology, Ravenio Books, 2016.
- [55] M. Stede, D. Schlangen, Information-seeking chat : Dialogue management by topic structure, 2004.
- [56] A. Spink, H. Özmutlu, S. Özmutlu, Multitasking information seeking and searching processes, *JASIST* 53 (2002) 639–652. doi:10.1002/asi.10124.
- [57] W. Fan, Z. Zhao, J. Li, Y. Liu, X. Mei, Y. Wang, J. Tang, Q. Li, Recommender systems in the era of large language models (llms), arXiv preprint arXiv:2307.02046 (2023).
- [58] G. V. Aher, R. I. Arriaga, A. T. Kalai, Using large language models to simulate multiple humans and replicate human subject studies, in: International Conference on Machine Learning, PMLR, 2023, pp. 337–371.
- [59] J. Ramos, H. A. Rahmani, X. Wang, X. Fu, A. Lipani, Natural language user profiles for transparent and scrutable recommendations, 2024. arXiv:2402.05810.
- [60] D. Maxwell, L. Azzopardi, K. Järvelin, H. Keskustalo, Searching and stopping: An analysis of stopping rules and strategies, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 313–322. URL: <https://doi.org/10.1145/2806416.2806476>. doi:10.1145/2806416.2806476.
- [61] W. S. Cooper, On selecting a measure of retrieval effectiveness part ii. implementation of the philosophy, *Journal of the American Society for information Science* 24 (1973) 413–424.
- [62] D. H. Kraft, T. Lee, Stopping rules and their effect on expected search length, *Information Processing & Management* 15 (1979) 47–58.
- [63] K. R. Nickles, Judgment-based and reasoning-based stopping rules in decision-making under uncertainty, University of Minnesota, 1995.
- [64] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 (2023).
- [65] A. Asai, Z. Wu, Y. Wang, A. Sil, H. Hajishirzi, Self-rag: Learning to retrieve, generate, and critique through self-reflection, arXiv preprint arXiv:2310.11511 (2023).

- [66] M. Aliannejadi, Z. Abbasiantaeb, S. Chatterjee, J. Dalton, L. Azzopardi, Trec ikat 2023: The interactive knowledge assistance track overview, arXiv preprint arXiv:2401.01330 (2024).
- [67] Z. Abbasiantaeb, C. Meng, D. Rau, A. Krasakis, H. A. Rahmani, M. Aliannejadi, Llm-based retrieval and generation pipelines for trec interactive knowledge assistance track (ikat) 2023, TREC, 2023.