CRS Arena: Crowdsourced Benchmarking of Conversational Recommender Systems

Nolwenn Bernard
University of Stavanger
Stavanger, Norway
nolwenn.m.bernard@uis.no

Hideaki Joko
Radboud University
Nijmegen, The Netherlands
hideaki.joko@ru.nl

Faegheh Hasibi
Radboud University
Nijmegen, The Netherlands
faegheh.hasibi@ru.nl

Krisztian Balog
University of Stavanger
Stavanger, Norway
krisztian.balog@uis.no

Abstract

We introduce *CRS Arena*, a research platform for scalable benchmarking of Conversational Recommender Systems (CRS) based on human feedback. The platform displays pairwise battles between anonymous conversational recommender systems, where users interact with the systems one after the other before declaring either a winner or a draw. CRS Arena collects conversations and user feedback, providing a foundation for reliable evaluation and ranking of CRSs. We conduct experiments with CRS Arena on both open and closed crowdsourcing platforms, confirming that both setups produce highly correlated rankings of CRSs and conversations with similar characteristics. We release *CRSArena-Dial*, a dataset of 474 conversations and their corresponding user feedback, along with a preliminary ranking of the systems based on the Elo rating system. The platform is accessible at https://iai-group-crsarena.hf.space/.

CCS Concepts

• **Information systems → Recommender systems**.

Keywords

Conversational recommender systems; Benchmarking; Conversational dataset

ACM Reference Format:
Nolwenn Bernard, Hideaki Joko, Faegheh Hasibi, and Krisztian Balog. 2025. CRS Arena: Crowdsourced Benchmarking of Conversational Recommender Systems. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM '25), March 10–14, 2025, Hannover, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3701551.3704120

1 Introduction

Conversational recommender systems (CRSs) are attracting attention due to their ability to provide personalized recommendations. Indeed, unlike traditional recommender systems, CRSs are interactive, offering users the possibility to express their current preferences and provide direct feedback on recommended items through natural language [8]. Despite the potential of CRSs, current evaluation often relies on offline metrics or user studies that focus on some specific aspect of the system, which is assessed based on static dialogue corpora. For example, Manzoor and Jannach [13] focus only on the quality of response generation in pre-defined context, i.e., do not consider a sequence of generated responses. Therefore, we argue that current evaluations often overlook the interactive nature of the problem. This may be due to the fact that most CRSs are research prototypes and performing a fair comparison of them with real users is both time consuming and expensive. While user simulation can alleviate some of these issues [1], simulation-based results are only indicative and need to be validated with real users.

In this work, we introduce CRS Arena, a platform to benchmark existing CRSs in a crowdsourced environment that is inspired by Chatbot Arena [4]. In CRS Arena, pairs of CRSs face each other in side-by-side "battles" where users can interact with them and declare a winner or a tie. The platform makes the evaluation process and CRSs accessible to a wide range of users, including those without technical expertise, in a fun and engaging way. Based on the outcomes of battles, a ranking of CRSs can be established based on first-party user feedback. In addition to the pairwise preferences (i.e., battle outcomes), users also leave explicit feedback regarding their (dis)satisfaction with the individual CRSs.

We tested CRS Arena in two crowdsourced environments, open and closed. Open corresponds to public access, while closed means restricted access to a selected group of workers recruited using a crowdsourcing platform. Using nine CRSs, we collected a total of 474 conversations, annotated with human feedback regarding overall satisfaction and battle outcome. The resulting dataset, *CRSArena-Dial*, is unique in that it contains conversations between CRSs and real users as opposed to a Wizard-of-Oz setting that is typically employed in existing dialogue corpora (i.e., there is a human worker acting as the CRS) [10, 12]. The dataset is made available to the community for further research and can be analyzed to better understand the capabilities and limitations of CRSs. Additionally, we compute Elo rating for each CRS based on the results of the battles to rank them. We find that the Elo ranking disagrees with the recall-based ranking reported for these systems, and user feedback shows overall low user satisfaction, highlighting the importance

WSDM '25, March 10–14, 2025, Hannover, Germany

ACM ISBN 979-8-4007-1329-3/25/03
https://doi.org/10.1145/3701551.3704120

of considering actual users and the interactive nature of the task when evaluating CRSs.

In summary, the main contributions of this work are:

- We develop CRS Arena, a platform where users can interact with CRSs and evaluate them in a realistic setting.
- We release a dataset of 474 conversations with various CRSs, along with first-party user feedback, collected using CRS Arena in both open and closed crowdsourced environments.
- We provide an initial analysis of the collected data, examining conversation characteristics, human feedback, and ranking of CRSs based on Elo rating.
- We demonstrate the robustness of CRS Arena across various crowdsourcing setups, showing its feasibility for fast and scalable human evaluation of CRSs.

CRS Arena is open source and accessible at https://iai-group-crsarena.hf.space/.[1]

2 Related Work

The evaluation of conversational recommender systems (CRSs) has been identified as an open challenge in the field [7]. Indeed, it is a complex task that requires the consideration of different aspects from the system and user perspectives, in addition to being highly interactive. Different evaluation methodologies have been proposed in the literature, including online experiments, user, and computational studies [8], with the last two being the most common. Toolkits and frameworks have been developed to facilitate the evaluation of CRSs, including CRSLab [20] and iEvaLM [18], which provide resources for computational studies, and INFACT [14] and CRS-Que [9], which facilitate user studies.

We make the following observations based on our examination of the literature. First, there is a lack of standardization in the evaluation of CRSs, which makes their comparison cumbersome. Second, directly interacting with CRSs is challenging due to the lack of user interface in existing toolkits and frameworks, e.g., CRSLab has an interface for only one out of five CRSs. Third, evaluation often focuses on specific aspects of the system, such as recommendation quality or the fluency of the responses, rather than the overall user experience. Finally, user studies often use grading rubrics resulting in pointwise comparisons; due to the subjective nature of the evaluation it can be difficult to ensure the consistency of the grading across many evaluators [11].

To mitigate some of these points, we propose CRS Arena, a platform to collect conversations and human feedback on CRSs in a realistic setting and to rank them based on pairwise comparisons. Generally, relative (pairwise) comparisons are easier for people and yield greater agreement among assessors than pointwise comparisons [2]. Our platform is inspired by the Chatbot Arena [4] that facilitates the benchmarking of large language models (LLMs) based on pairwise comparisons. The platform has led to the release of valuable resources, e.g., a dataset of realistic prompts and a leaderboard, for the community. We believe that CRS Arena can have a similar impact on the benchmarking of CRSs. The main differences between Chatbot Arena and CRS Arena are: (1) the focus, i.e., LLMs vs. CRSs, (2) the independence of the systems, i.e., in Chatbot Arena the systems receive the same prompts while in CRS Arena the conversations are independent, and (3) CRS Arena collects users' explicit feedback on task success, i.e., frustration or satisfaction, after each conversation.

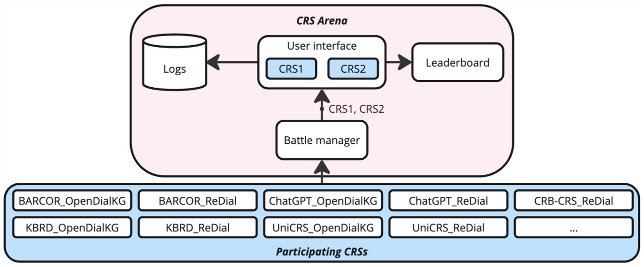


Figure 1: Overview of the main components of CRS Arena.

3 CRS Arena

This work presents CRS Arena, a platform for benchmarking conversational recommender systems (CRSs) in a realistic setting. The main components of the platform are shown in Fig. 1.

3.1 Conversational Recommender Systems

Currently, there are nine CRSs available in the arena; see ***Participating CRSs*** in Fig. 1. These include *KBRD* [3], *BARCOR* [17], *UniCRS* [19], *ChatGPT* [18], and *CRB-CRS* [15], which are trained and/or leverage external knowledge from either OpenDialKG [16] or ReDial [12]. These CRSs are implemented in an extended version of the iEvaLM framework [18] that facilitates the integration of new CRSs. They only need to inherit from the CRS base class and implement the method responsible for generating responses to incoming user utterances. Consequently, new participating systems can join the arena with minimal effort.

3.2 Battle Manager

The main responsibility of the battle manager is to pair two CRSs for a pairwise comparison, i.e., "battle." The current matchmaking algorithm selects the two CRSs with the fewest recorded conversations, resolving ties by random selection, to ensure that all CRSs can be compared to each other as uniformly as possible. In the future, the algorithm may be improved to consider other factors, such as the performance of the CRSs in previous battles.

3.3 User Interface

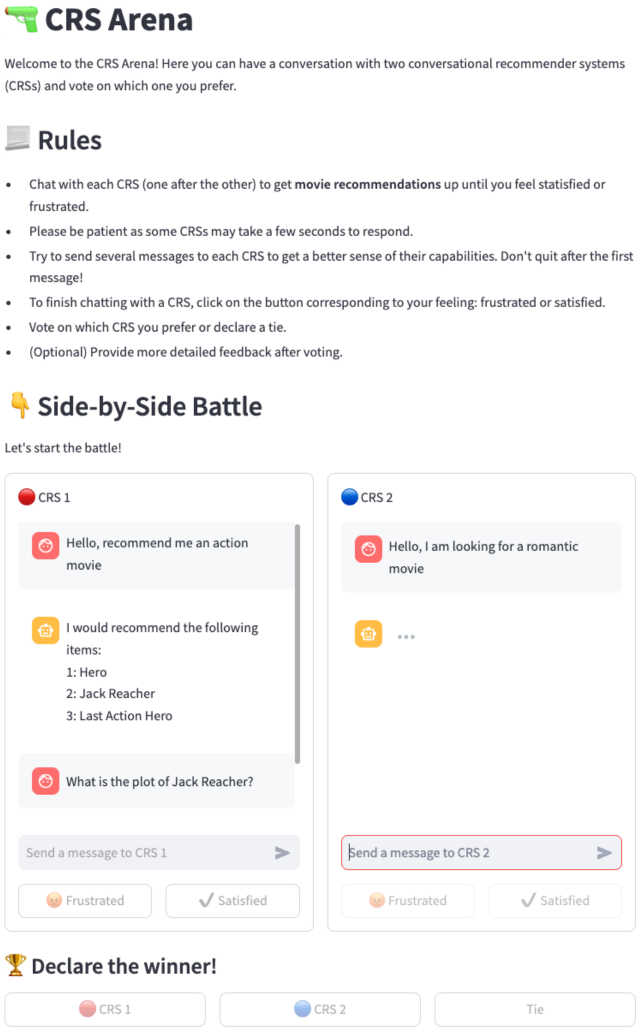
The user interface of CRS Arena is designed to be engaging and accessible to users with diverse backgrounds; see Fig. 2. It is divided into five sections, as follows (from top to bottom):

(1) *Introduction* provides a succinct description of the platform's purpose.

(2) *Rules* concisely outlines how sessions are conducted: users are expected to get movie recommendations by conversing with two anonymous CRSs sequentially, providing explicit feedback (frustration or satisfaction) for each. Once both conversations are completed, the user selects a winner or declares a draw.

(3) *Side-by-side battles* is the central part of the frontend. It contains two chat interfaces associated with the two CRSs. A chat interface is composed of: (1) a chat window, with the conversation history and an input field to type messages and (2) two buttons to express satisfaction or frustration before ending the conversation. Note that once a conversation is ended, the user cannot go back to it.

[1] The Safari browser is currently not supported.



CRS Arena

Welcome to the CRS Arena! Here you can have a conversation with two conversational recommender systems (CRSs) and vote on which one you prefer.

Rules

- Chat with each CRS (one after the other) to get **movie recommendations** up until you feel statisfied or frustrated.
- Please be patient as some CRSs may take a few seconds to respond.
- Try to send several messages to each CRS to get a better sense of their capabilities. Don't quit after the first message!
- To finish chatting with a CRS, click on the button corresponding to your feeling: frustrated or satisfied.
- Vote on which CRS you prefer or declare a tie.
- (Optional) Provide more detailed feedback after voting.

Side-by-Side Battle

Let's start the battle!

CRS 1

Hello, recommend me an action movie

I would recommend the following items:
1: Hero
2: Jack Reacher
3: Last Action Hero

What is the plot of Jack Reacher?

CRS 2

Hello, I am looking for a romantic movie

...

Declare the winner!

Figure 2: Screenshot of the CRS Arena.

(4) *Vote* presents three buttons to choose a winner or declare a draw. After voting, the user has the option to provide additional feedback.

(5) *Terms of service and contact* (not visible in Fig. 2) explains that CRS Arena is a research platform that collects data and provides a point of contact for support.

3.4 Implementation

CRS Arena is implemented as a web application using Streamlit,[2] a Python library for building interactive applications. At the beginning of each session, a user is assigned an identifier ensuring that data remains anonymous (unless the user voluntarily discloses personal information during the conversations).

CRS Arena is publicly deployed on the HuggingFace Hub.[3] We observe that some models have high latency when interacting with users, which may affect their experience. It is likely due to high traffic and hardware limitations as the platform currently runs on 2 CPU cores and 16 GB of RAM.

4 Data and Analysis

CRS Arena has been publicly released in September 2024. After a period of approximately 10 days, we collected 254 conversations and 81 votes from users. Note that we do not apply any filtering, e.g., based on quality, to preserve the raw and uncurated nature of the data. Additionally, we conducted a study in a closed environment, with access restricted to selected crowd workers, to assess the robustness of CRS Arena across different user groups. Specifically, we recruited workers from English-speaking countries on Prolific, with a 100% approval rate and $\geq$ 1000 previous submissions, to ensure data quality. The data was collected within approximately 7 hours, with each battle taking around 9 minutes at a cost of roughly £1.33. Considering that some workers might behave to maximize their financial gain [5], which could lead to a limited number of interactions, we explicitly instructed them to interact with each CRS at least 5 times. Using this setup, we collected 220 conversations and 104 votes. Overall, the number of conversations collected from the different CRSs is fairly uniform.[4]

The evaluation of CRSs based on users' ratings is presented in Table 1. We observe a generally low level of user satisfaction in both environments, with ChatGPT_OpenDialKG and ChatGPT_ReDial achieving the highest average satisfaction rates of 52.1% and 43.9%, respectively. Notably, even the best-performing CRSs can only satisfy users in approximately half of the cases. Detailed user feedback indicates a few reasons for frustration, such as the lack of understanding of users' requests leading to irrelevant responses and recommendations, repetitive answers in some cases, lack of certain information, and latency issues.

Table 1 also presents a preliminary ranking of CRSs based on the Elo rating [6], computed from the 185 pairwise judgments collected. For Elo computation, initial rating is set to 1000 and K-factor is 16. We note that this preliminary ranking is consistent with the feeling of satisfaction expressed by users in the collected conversations, as indicated by a strong Spearman correlation of $\rho = 0.917$. The R@10 column shows the recall of each CRS reported in Wang et al. [18]. The Spearman's coefficient between R@10 and Elo ratings is $\rho = -0.238$, showing a negative ranking correlation. This shows that performing well on the recommendation aspect does not align well with user satisfaction, thus, highlighting the importance of evaluating CRSs in a holistic manner.

There is a strong correlation when comparing the open and closed crowdsourcing environments. Indeed, for Elo ratings, Pearson's and Spearman's correlation coefficients are $r = 0.763$ and $\rho = 0.700$, respectively, while for satisfaction, $r = 0.843$ and $\rho = 0.726$. Table 2 compares high-level characteristics of conversations from both environments. Response diversity is measured using Distinct-2, following Joko et al. [10], with the same hyperparameters. Both environments show similar statistics, with the average number of utterances per dialogue being slightly longer in the closed one, as crowd workers were instructed to have at least 5 interactions. Overall, the results suggest that user feedback and conversation characteristics are similar in both open and closed setups, indicating the robustness of CRS Arena across different user groups. This is particularly important because using a closed

[2] https://streamlit.io/
[3] https://huggingface.co/

[4] A detailed analysis including the conversation distribution is available at: https://github.com/iai-group/crsarena-dial/blob/main/DataAnalysis.md.

Table 1: Preliminary ranking of systems and level of satisfaction across different environments. R@10 shows the recall of each CRS as reported in Wang et al. [18]. The rank position is shown in parentheses for Elo rating and R@10.

CRS	Open crowdsourcing		Closed crowdsourcing		Closed & open crowdsourcing		R@10 (rank)
	Elo (rank)	% sat.	Elo (rank)	% sat.	Elo (rank)	% sat.	
BARCOR_OpenDialKG	968 (7)	17.2	1008 (4)	11.5	988 (5)	14.5	0.453 (3)
BARCOR_ReDial	1052 (2)	30.4	1044 (3)	29.2	1077 (3)	29.8	0.170 (7)
CRB-CRS_ReDial	930 (9)	5.4	964 (6)	11.5	930 (8)	7.9	—
ChatGPT_OpenDialKG	**1056 (1)**	50	1066 (2)	**54.2**	**1102 (1)**	**52.3**	**0.539** (1)
ChatGPT_ReDial	1036 (3)	**58.6**	**1085 (1)**	29.2	**1102 (1)**	45.3	0.174 (6)
KBRD_OpenDialKG	966 (8)	2.6	942 (9)	0.0	920 (9)	1.7	0.423 (4)
KBRD_ReDial	1027 (4)	8.6	985 (5)	7.4	994 (4)	8.1	0.169 (8)
UniCRS_OpenDialKG	974 (6)	9.5	953 (7)	0.0	937 (7)	4.8	0.513 (2)
UniCRS_ReDial	991 (5)	18.2	952 (8)	3.7	950 (6)	10.2	0.215 (5)

Table 2: Comparison between conversations collected in open and closed crowdsourcing environments.

Statistics	Open	Closed
#Utterances per dialogue	8.13	11.15
#Words per utterance	11.01	11.59
Diversity (Distinct-2)	0.527	0.538

crowdsourcing setup enables fast and scalable human evaluation of CRSs when a new system is released.

We release the resulting dataset, CRSArena-Dial,[5] which represents a unique resource. Indeed, unlike existing dialogue corpora, it contains conversations with multiple CRSs and real users, in addition to pairwise comparisons between the systems. We acknowledge that the dataset is noisy due to the minimally regulated environment, where only high-level guidelines were provided. However, it remains highly representative of how regular users naturally express preferences and seek recommendations. We thus believe that the dataset is a valuable resource for evaluating CRSs and studying authentic user interactions with such systems.

5 Conclusion

We introduced CRS Arena for benchmarking conversational recommender systems in a realistic setting. The platform anonymously collects conversations between CRSs and real users, along with feedback and pairwise CRS preferences. We released CRSArena-Dial, a dataset comprising 474 conversations and their corresponding user feedback, collected in open and closed crowdsourced environments. Analysis of the data reveals an overall low level of satisfaction, highlighting that further research is needed to improve the quality of CRSs. We also presented a preliminary ranking of nine CRSs based on Elo ratings, providing an early snapshot of a leaderboard.

As a resource built for the community, we rely on active contributions—both in adding new CRSs and participating as users. This, in turn, would allow for creation and release of a larger and more diverse version of CRSArena-Dial and enable a more comprehensive evaluation of CRSs. As the community grows, we also plan to scale up the platform to be able to accommodate more users.

Acknowledgments

We thank all the contributors of iEvaLM. This research was partially supported by the Norwegian Research Center for AI Innovation, NorwAI (309834, Research Council of Norway) and the Dutch Research Council (NWO) under the LESSEN project (NWA.1389.20.183).

References

[1] Krisztian Balog and ChengXiang Zhai. 2024. User Simulation for Evaluating Information Access Systems. *Found. Trends Inf. Retr.* 18, 1-2 (2024), 1–261.

[2] Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. 2008. Here or there: preference judgments for relevance. In *Proc. of ECIR '08.* 16–27.

[3] Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards Knowledge-Based Recommender Dialog System. In *Proc. of EMNLP-IJCNLP '19.* 1803–1813.

[4] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *Proc. of ICML '24.*

[5] Carsten Eickhoff and Arjen P. de Vries. 2011. How Crowdsourcable is Your Task?. In *Proc. of CSDM '11.* 11–14.

[6] Arpad E Elo. 1967. The proposed uscf rating system, its development, theory, and applications. *Chess life* 22, 8 (1967), 242–247.

[7] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open* 2 (2021), 100–126.

[8] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *ACM Comput. Surv.* 54, 5 (2021), 1–36.

[9] Yucheng Jin, Li Chen, Wanling Cai, and Xianglin Zhao. 2024. CRS-Que: A User-centric Evaluation Framework for Conversational Recommender Systems. *ACM Trans. Recomm. Syst.* 2, 1 (2024).

[10] Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P. de Vries, Jeff Dalton, and Faegheh Hasibi. 2024. Doing Personal LAPS: LLM-Augmented Dialogue Construction for Personalized Multi-Session Conversational Search. In *Proc. of SIGIR '24.* 796–806.

[11] Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of Search and Evaluation Strategies in Neural Dialogue Modeling. In *Proc. of INLG '19'.* 76–87.

[12] Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Proc. of NIPS '18.* 9748–9758.

[13] Ahtsham Manzoor and Dietmar Jannach. 2021. Generation-based vs. Retrieval-based Conversational Recommendation: A User-Centric Comparison. In *Proc. of RecSys '21.* 515–520.

[14] Ahtsham Manzoor and Dietmar Jannach. 2022. INFACT: An Online Human Evaluation Framework for Conversational Recommendation. In *Proc. of KARS '22.* 6–11.

[15] Ahtsham Manzoor and Dietmar Jannach. 2022. Towards retrieval-based conversational recommendation. *Inf. Sys.* 109 (2022), 102083.

[16] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs. In *Proc. of ACL '19'.* 845–854.

[17] Ting-Chun Wang, Shang-Yu Su, and Yun-Nung Chen. 2022. BARCOR: Towards A Unified Framework for Conversational Recommendation Systems. arXiv:2203.14257 [cs.CL]

[18] Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models. In *Proc. of EMNLP '23.* 10052–10065.

[19] Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Towards Unified Conversational Recommender Systems via Knowledge-Enhanced Prompt Learning. In *Proc. of KDD '22.* 1929–1937.

[20] Kun Zhou, Xiaolei Wang, Yuanhang Zhou, Chenzhan Shang, Yuan Cheng, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2021. CRSLab: An Open-Source Toolkit for Building Conversational Recommender System. In *Proc. of ACL-IJCNLP '21.* 185–193.

[5] https://github.com/iai-group/crsarena-dial