Rethinking the Evaluation of Dialogue Systems: Effects of User Feedback on Crowdworkers and LLMs

Clemencia Siro
University of Amsterdam
Amsterdam, The Netherlands
c.n.siro@uva.nl

Mohammad Aliannejadi
University of Amsterdam
Amsterdam, The Netherlands
m.aliannejadi@uva.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

ABSTRACT

In ad-hoc retrieval, evaluation relies heavily on user actions, including implicit feedback. In a conversational setting such signals are usually unavailable due to the nature of the interactions, and, instead, the evaluation often relies on crowdsourced evaluation labels. The role of user feedback in annotators' assessment of turns in a conversational perception has been little studied. We focus on how the evaluation of task oriented dialogue systems (TDSs), is affected by considering user feedback, explicit or implicit, as provided through the follow-up utterance of a turn being evaluated. We explore and compare two methodologies for assessing TDSs: one includes the user's follow-up utterance and one without. We use both crowdworkers and large language models (LLMs) as annotators to assess system responses across four aspects: relevance, usefulness, interestingness, and explanation quality. Our findings indicate that there is a distinct difference in ratings assigned by both annotator groups in the two setups, indicating that user feedback does influence system evaluation. Workers are more susceptible to user feedback on usefulness and interestingness compared to LLMs on interestingness and relevance. User feedback leads to a more personalized assessment of usefulness by workers, aligning closely with the user's explicit feedback. Additionally, in cases of ambiguous or complex user requests, user feedback improves agreement among crowdworkers. These findings emphasize the significance of user feedback in refining system evaluations and suggest the potential for automated feedback integration in future research. We publicly release the annotated data.[1]

CCS CONCEPTS

• **Information systems** → **Evaluation of retrieval results**; **Users and interactive retrieval**; Crowdsourcing; • **Computing methodologies** → Discourse, dialogue and pragmatics.

KEYWORDS

Evaluation, User feedback, Crowdworkers, Large language models

ACM Reference Format:
Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Rethinking the Evaluation of Dialogue Systems: Effects of User Feedback on Crowdworkers and LLMs. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3626772.3657712

[1] https://github.com/Clemenciah/LLMCrowdDialogueEval/tree/main/Data

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657712

1 INTRODUCTION

Evaluation of systems has been an integral part of the information retrieval (IR) research agenda for decades [e.g., 13]. Traditionally, IR evaluation has relied highly on user actions, including implicit feedback such as click-through rates. However, in a conversational setting such signals are not usually available due to the nature of the interactions. As a result, the evaluation of dialogue systems increasingly relies on human evaluation, leading to a growing interest in user-centric evaluation methods [45]. However, asking for explicit user feedback from a user can be intrusive and may negatively impact user experience [46]. Therefore, in recent years, the assessment of conversational systems has relied on crowdsourced evaluation, leveraging the collective wisdom of human annotators.

Turn-level assessments. When gathering evaluation feedback on individual turns in a conversational interaction, various design methods have been considered in the past. These include deciding on the type of judgment scale, as well as formulating annotation guidelines and methods for presenting dialogues under assessment at the turn level [36, 41]. Recent strategies for presenting turn-level utterances to annotators involve two main approaches: one displays both the user's initial request and the system's response for evaluation [40, 42], and the other shows only the system's response [35]. The choice between these approaches often depends on the specific evaluation metric in question. The first method, similar to the query–document pair setting in ad-hoc retrieval systems, operates under the premise that the user's initial request offers sufficient context for annotators to make well-informed evaluations.

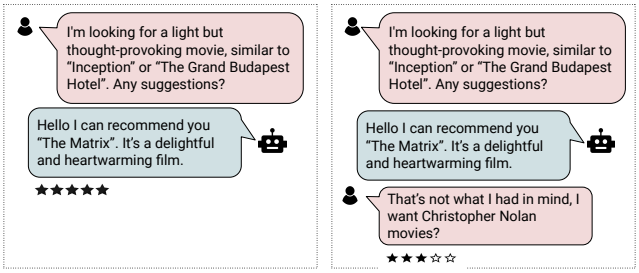


Figure 1: A dialogue showing an example of a complex user request with (right) and without (left) the user feedback. The star ratings show the assessment of external assessors judging the usefulness of the system utterance. As can be seen, based on the follow-up utterance the assessors lower their usefulness rating aligning with the user feedback.

Follow-up utterances. Users often do not articulate all of their intentions in a single request. Rather, they engage in an iterative dialogue, clarifying and refining their intentions through successive exchanges [1]. Their queries can be multifaceted, ambiguous, or overly generic, which complicates the process of evaluating individual turns. E.g., in Fig. 1, the user poses a multifaceted query, leaving substantial room for interpretation. First, the user is looking for a movie suitable for an evening watch, which typically suggests something not overly long or intense. Second, they desire a film that is "light but thought-provoking," implying a blend of easy-to-digest content with depth in storytelling. Last, by referencing specific movies like "Inception" and "The Grand Budapest Hotel," the user indicates a preference for a certain style or genre — perhaps celebratory narratives with unique storytelling or visually engaging films. Annotators tasked with evaluating the system's response to this query must consider these multiple layers.

When annotating turns in a conversational interaction, the complexity for annotators lies in assessing the system's response not just for its relevance to an overt request, e.g., for a movie recommendation, but also for its alignment with nuanced, implicit preferences indicated by the user. Systems may not always successfully address all aspects of such requests. When the system's response only partially meets the query's criteria, it becomes challenging for annotators to gauge which aspect of the request was most critical to the user. We believe that this is where a user's follow-up utterance be particularly informative. A user's subsequent response may provide valuable insights into what they valued most in their original request, giving annotators a clearer indication of the user's priorities. Thus, follow-up utterances may serve as crucial cues, helping annotators make a more informed assessment of the system's performance, particularly in how well it navigates and prioritizes the multifaceted aspects of a user's complex request.

Is a user's follow-up utterance crucial in ensuring evaluations align with actual user needs, especially since annotators, as external evaluators, may not fully grasp the user's perspective or context? We hypothesize that annotators who have access to a follow-up utterance produce more accurate and user-centric evaluations, improving the quality of evaluation labels in the process.

Research goals. We investigate the effect of a user's follow-up utterance on the annotation of turns in a task oriented dialogue system (TDS). We conduct experiments with two types of annotators: human and large language model (LLM)-based. Both types of annotators are asked to provide annotations of turn-level system responses along four dimensions: *relevance*, *usefulness*, *interestingness*, and *explanation quality*, on a 100 level scale, following [36]. We consider two contrastive setups for annotators to provide these annotations: (**Setup 1**) does not consider the user's follow-up utterance, and (**Setup 2**) does consider the user's follow-up utterance. In addition, we collect data on what sources of information human annotators rely on to arrive at their judgments. With this data, we aim to examine the bias introduced by these sources to the crowdsourced labels.

We use a subset of the recommendation dialogues (ReDial) [29] dataset to address the following research questions: (**RQ1**) How does a user's implicit feedback from the follow-up utterance influence the evaluation labels collected from both human annotators and LLMs? (**RQ2**) When is implicit user feedback significant in the evaluation of TDSs? (**RQ3**) What are the annotators' perceptions in terms of the sources of information they rely on to make their assessments, and what are the potential biases might that have on their performance?

Findings. Our findings indicate that both the crowdworkers and the LLM exhibit sensitivity to user cues from follow-up utterances. There is a significant difference in the mean ratings from both annotators except for relevance when follow-up utterance is included, indicating user feedback does influence system evaluation. Workers are more susceptible to user feedback in usefulness and interestingness, compared to LLMs in interestingness and relevance. Specifically, there is a clear distinction in relevance and usefulness ratings of crowdworkers **Setup 2** ratings unlike in **Setup 1** where these aspects are often conflated. This indicates that crowdworkers not only evaluate response usefulness based on topical relevance but also align with user needs and preferences expressed in follow-up utterances. This suggests that follow-up utterances enable a more personalized assessment of usefulness, aligning closely with the user's explicit feedback. In **Setup 2**, we observe an increase in annotator agreement. This is particularly evident in scenarios characterized by uncertainty in user requests. Complex user requests with multiple criteria or preferences posed challenges during evaluation, but follow-up utterances helped to clarify the user intent. Similarly, generic user requests, initially broad and challenging to address, became more focused on follow-up utterances, allowing human-/LLM-based annotators to tailor their assessment effectively.

These findings not only show the significance of user feedback in system evaluation but also provide a foundation for integrating user feedback in the automatic evaluation of conversational systems.

2 RELATED WORK

Recent studies emphasize evaluating TDSs through a user experience lens [14]. Traditionally, TDSs were assessed primarily for task completion. While task completion remains a fundamental criterion, there is a growing recognition that solely measuring task success may not provide a comprehensive understanding of system performance [40]. Consequently, there is a shift towards incorporating user experience metrics into the evaluation of TDSs [39]. Here, we provide a brief overview of the studies on TDS evaluation from the perspective of user feedback, evaluation bias, and LLMs.

2.1 User feedback

In web search, implicit user signals including click-through rates and dwell time on search results are available in vast amounts and these signals are leveraged to evaluate a system and continually improve the search results [27, 28, 30, 31, 46]. However, such signals are not accessible for conversational systems due to their interactive nature. Consequently, automatic evaluation of conversational systems would primarily rely on explicit user feedback in the form of ratings [8, 12], which could be intrusive and lead to poor user experience [7]. In this work, we propose using the user's next utterance as a proxy for both explicit and implicit user feedback. For instance, when a user expresses satisfaction or dissatisfaction in their next utterance following a system recommendation, it represents explicit feedback that should not be disregarded when assessing system

performance. Our study investigates how user feedback from the next utterance influences the evaluation labels provided by both human- and LLM-based annotators.

2.2 Bias in crowdsourcing evaluation labels

The use of crowdsourcing for evaluating TDSs in IR research, while offering scalability and diversity, brings inherent biases. Research, such as [4, 16, 38], highlights cognitive biases and load as significant factors influencing crowdworker judgments, which can skew the evaluations. For example, workers' preconceived notions or mental fatigue might lead to inconsistent results. Further, Hube et al. [24] and Han et al. [20] emphasize the impact of workers' relevance strategies and personal biases on their assessment of IR systems. To mitigate these biases, strategies such as task design adjustments and worker training are suggested [17]. For instance, presenting tasks neutrally and providing clear, unbiased instructions can help reduce bias, as discussed in [24]. Additionally, the choice and implementation of judgment scales, as explored in [6], play a crucial role in assessor bias. Different from previous studies, in this work we focus on assessing how the sources of information relied on by workers to make their judgment bias their evaluation labels.

2.3 LLMs as annotators

Recently, there has been a notable surge in the use of LLMs as annotators in various tasks [10]. These models show good performance comparable to human annotators and in some cases outperform them [18]. Additionally, they have proven to reduce the time and cost for annotation, making them a preferred choice compared to human annotators [44]. However, most research efforts have primarily focused on assessing how well LLMs' labels correlate to human labels [15, 23]. There has been a relative lack of investigation into whether LLMs are susceptible to the same influencing factors as crowdworkers.

Several studies have delved into understanding the factors that impact crowdworkers, encompassing aspects like task design, judgment scales, and protocols designed to enhance the quality of annotations [21, 26, 37]. These studies have contributed valuable insights into optimizing crowdworker performance. However, a similar examination of the impact of these factors on LLMs is notably absent from the research landscape. We contribute towards understanding how task design influences evaluation labels assigned by LLMs. We investigate the influence of user feedback from the user's follow-up utterance on the evaluation of TDSs by both crowdworkers and LLMs.

3 THE ANNOTATION TASK

Our objective is to understand the influence of user feedback from the user's follow-up utterance on the evaluation of TDSs. We conduct our study as an annotation effort with crowdworkers from amazon mechanical turk (AMT) [3]. Additionally, due to the increased use of LLMs as annotators [15, 23] we seek to understand how LLMs are affected by feedback from the user's next utterance. User feedback in this case can be either implicit or explicit. Explicit feedback refers to straightforward, direct responses from users, like specific comments on a certain dialogue aspect (e.g., "I don't like this movie"). Implicit feedback is more subtle, encompassing aspects like tone or contextual hints within the user's follow-up utterance (e.g., "Thanks for the suggestion. How about some action movies?"). We gather annotations for four fine-grained dialogue qualities: relevance, usefulness, interestingness, and explanation quality, across two experimental conditions.

3.1 Dialogue qualities

We experiment with four dialogue qualities in the domain of TDSs [5, 40] that have been investigated extensively, elaborated below.

Relevance. The relevance [2, 25, 32, 40] of a dialogue response is a crucial factor in assessing the effectiveness of a TDS. To evaluate relevance, we ask the workers to determine how well the system's response addresses the user's request. This aspect gauges the system's ability to understand and appropriately respond to user input.

Usefulness. Usefulness [33, 41, 43] of a dialogue response pertains to its practical value from the user's perspective. Apart from just being relevant, workers assess whether the system's response gives additional information to the user on the recommended item. In **Setup 2** the user's perspective is captured by asking workers to rely on the user's follow-up utterance to gauge the usefulness of the system response. E.g., if a user says, "I have already watched that," it suggests that the recommendation is not new or helpful to the user, even though it might be relevant. Usefulness helps to measure the system's overall utility in real-world scenarios.

Explanation quality. Understanding how well a TDS communicates its reasoning is important for user trust and comprehension. Explainability in IR systems has witnessed a notable surge in recent times [5, 19, 47, 48]. Following Guo et al. [19] we instruct workers to assess explanation quality, by evaluating the clarity and informativeness of the system's justifications or explanations accompanying its responses. This aspect provides insights into the system's transparency and user-friendly communication.

Interestingness. Beyond system functionality, the interestingness [39] of a system response adds a subjective layer to the evaluation. Workers are asked to evaluate whether the system's responses are engaging, captivating, or exhibit qualities that make the interaction more enjoyable for the user. This encapsulates the language used to make recommendations by the system. This aspect contributes to a holistic assessment of user experience.

3.2 Data

We use the ReDial dataset [29], a well-known collection of over 11,000 dialogues specifically focused on movie recommendations. We sample the dialogue turns for annotation by focusing on selecting user utterances that explicitly request movie recommendations or express movie preferences. Phrases like "I prefer," "recommend me," and "my favorite" were key indicators in this selection process. This approach ensured that our sampled data contained user utterances that were explicit and straightforward, facilitating a more accurate assessment of the dialogue system's responses.

Similar to Guo et al. [19], we observe a lack of in-depth explanations in the ReDial dialogues. Our initial analysis of the dataset indicates that longer system utterances often include attempts to explain movie recommendations, whereas shorter ones do not. As a result, we selected system utterances with more than 14 words (average length of system responses in the dataset) to better focus on responses that are more likely to include explanations. In total, we sampled 100 unique dialogue turns from the dataset, each representing a different conversation.

3.3 Annotation scale

Following the approach outlined in [36], our study adopts the S100 scale for evaluation purposes. This scale is employed through a sliding window mechanism, allowing annotators to provide detailed feedback on the dialogue systems. The sliding scale's interactive nature enables a more precise and flexible assessment compared to traditional binary or categorical scales. To enhance its usability and ensure intuitive responses, the default value on this slider is set to 0. This design choice is based on the rationale that a neutral starting point encourages annotators to consciously adjust the slider based on their judgment of the dialogue turn's effectiveness, rather than being biased by any preset values. To ensure consistency and accuracy in evaluations, we provide the annotators with several examples demonstrating how to effectively use the S100 scale. We adopt the same scale for annotation with LLM.

3.4 Preliminary experiments

Our research included preliminary experiments to refine the design and methodology. These experiments assessed the practicality of our setups, refined annotation guidelines, and identified data collection challenges. Two setups were tested:

Exp 1 *Single worker, two conditions*: Workers evaluated a dialogue turn under two conditions within a single human intelligence task (HIT). The only difference was the presence or absence of the user's follow-up utterance, which served as user feedback.

Exp 2 *Random assignment of conditions*: Workers were randomly assigned to one of the two conditions to incorporate diverse perspectives and reduce potential biases, gathering a range of rationales behind annotator evaluations.

Preliminary results. We used 13 dialogue turns, primarily focusing on comparing **Exp 1** and **Exp 2** to determine the most effective approach. The mean ratings obtained from both setups indicated a high degree of consistency in annotator assessments for relevance and usefulness, suggesting that both methods performed similarly in these aspects. However, an interesting observation emerged concerning annotation time and the diversity of justifications. In **Exp 1**, resulted in shorter annotation times but exhibited limited diversity in justifications. In contrast, **Exp 2**, yielded a more diverse set of justifications. Considering these findings, we decided to proceed with the **Exp 2** setup for our main experiments. This choice was motivated by the goal of obtaining a broader and more diverse range of annotations and justifications, a critical requirement for the comprehensive evaluation of dialogue systems in our study.

3.5 Experimental conditions

Following **Exp 2**, we designed two distinct experimental conditions to evaluate the effect of user feedback on the evaluation of TDSs with human annotators, as well as LLMs.

Setup 1 Following the conventional annotation method, this condition provides only the initial user query and the system's response to the annotators and LLMs, omitting the user's follow-up utterance. This setup focuses on evaluating the TDS based on a single interaction, reflecting the traditional approach in dialogue annotation.

Setup 2 This condition incorporates the user's follow-up utterance along with the initial query and the system's response. The aim is to allow annotators and LLMs to evaluate the TDS within the full context of the conversation, assessing the impact of subsequent user feedback on annotations.

3.6 Human annotators

For this task, we recruited master workers from AMT. We employed multiple HIT templates to conduct our study, aiming to investigate the impact of the user's next utterance on annotator ratings for various aspects. We collected annotation labels for relevance, usefulness, interestingness, and explanation quality in the two experimental conditions. Each aspect was annotated in a separate HIT. Importantly, we did not disclose the research angle to the annotators, framing it as an annotation effort.

During each HIT, we provided the annotators with instructions, definitions of the aspect to be assessed, and examples. We maintained consistent instructions across all aspects and setups, with variations limited to definitions and examples. In each HIT, annotators rated the aspect and provided justifications for their ratings, a practice known to reduce randomness and enhance assessment quality [34]. Additionally, we sought to understand the sources of information relied on by annotators to make their assessments, asking them to select sources they considered when making their assessments, including personal knowledge, external sources such as web searches, educated guesses, user's request, system response, user's feedback, and an "other" option for sources not covered in the provided options.

84 unique workers participated in the study (46 female and 38 male), with an average age of 30–45. Each worker received a reward of $0.4 per HIT, which was determined based on the minimum wage.

3.7 LLM as annotator

Since LLMs have shown a notable performance as annotators, we investigate whether LLMs are influenced to a comparable degree as human annotators with user feedback from the user's follow-up utterance. This investigation seeks to shed light on the potential similarities and differences in how LLMs and human annotators respond to such contextual input, ultimately contributing to an understanding of LLM behavior in annotation tasks.

We used ChatGPT (gpt-3.5-turbo API[2]) a subseries of GPT models [9] for the annotation task. Typically, crowdworkers are provided with annotation examples to improve the quality of the labels they assign, similarly, we provide the same to ChatGPT, thus using it in the few-shot setting. To ensure consistency, we replicate the experimental conditions used for human annotators across the four aspects. In our experiments, we employ two distinct experimental conditions, varying the prompts and data presentation to align with these setups.[3] For each aspect, we provided ChatGPT with the corresponding human annotation instructions. This approach allows us to comprehensively assess the performance of ChatGPT in generating evaluation labels in a manner that mirrors human annotation practices.

4 CROWDSOURCED JUDGMENTS

Before addressing our research questions (in Sec. 5–7), we examine the judgments collected through crowdsourcing.

Internal agreement. To assess the quality of labels collected from the crowdworkers, we compute pairwise Cohen's Kappa and report

[2]Temperature = 1, Top p = 1

[3]The prompts are available in our **Online Appendix** in Sec. A.

Table 1: ICC and pairwise Cohen's Kappa for all aspects across both setups.

	ICC		Kappa	
Aspects	Setup 1	Setup 2	Setup 1	Setup 2
Relevance	0.8358	0.7427	0.6978	0.5701
Usefulness	0.7553	0.7419	0.5892	0.5631
Interestingness	0.5456	0.4685	0.2825	0.2231
Explanation quality	0.5136	0.5351	0.2380	0.2812

the results in Tab. 1. The Kappa scores indicate varying levels of agreement in different evaluation setups and aspects:

Relevance shows a substantial agreement in **Setup 1**, compared to **Setup 2** (Kappa: 0.69 vs. 0.57). This indicates that the inclusion of the user's follow-up utterance during evaluation introduces complexity, impacting crowdworkers' judgments. Both setups in *usefulness* exhibit high agreement, indicating that the presence or absence of user follow-up utterances has minimal influence on crowdworkers' perceptions of the system's response utility. This suggests that, overall crowdworkers are consistent in the evaluation of usefulness. **Setup 1** shows a moderate agreement (Kappa: 0.28) in evaluating *interestingness* while **Setup 2** has a slightly lower agreement (Kappa: 0.22), reflecting the added complexity introduced by the user's follow-up utterance. In *explanation quality*, **Setup 2** exhibits higher agreement (Kappa: 0.28 vs. 0.23), possibly because it allows crowdworkers to assess explanations within a broader context that includes both the initial response and the user reaction in the follow-up.

We make similar observations with the intraclass correlation coefficient (ICC) values in the two setups. In general, workers exhibit a substantial to moderate agreement for all aspects. The low agreement in interestingness and explanation quality can be attributed to the nature of their subjectivity and the large scale of evaluation.

Crowdworker judgments. Fig. 2 shows the distributions of scores provided by crowdworkers for the four dialogue qualities over the two setups. In line with the findings of Roitero et al. [36], we see that 60% of the scores are multiples of 5 or 10, showing that crowdworkers tend to select such values as their judgments in an S100 range.

Relevance scores in both setups display a long-tailed distribution toward the extremes, as shown in Fig. 2a and 2b. The median rating in **Setup 1** is 71 with 65 for **Setup 2**, suggesting that including the user's follow-up utterance leads to lower relevance scores from the workers. The scores for both setups range between 0 to 100.

In **Setup 1**, there is a noticeable drop in *usefulness scores* within the 20–40 range, as seen in Fig. 2c. This setup also indicates similarity in the distribution of relevance and usefulness scores, suggesting that, without user feedback, workers tend to rate usefulness similarly to relevance due to limited information. In **Setup 2**, which includes user follow-up, there is a decrease in responses rated as not useful (0–20) and an increase in the 30–70 range. This indicates that some responses are considered useful even when not directly relevant to the user's initial query, possibly because users found the recommended movie intriguing or had not considered it before. The median scores are 70 and 56 respectively in Setups 1 and 2.

The *interestingness* aspect is highly subjective, therefore making it more prone to individual worker bias as observed with the moderate agreement between workers in Tab. 1. In **Setup 1** the

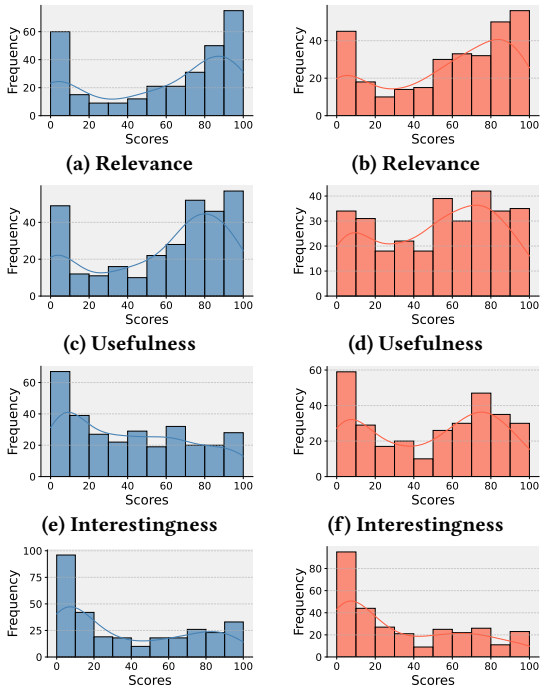


Figure 2: A comparison of individual worker scores distributions for Setup 1 (left column) and Setup 2 (right column).

scores are skewed towards the left, indicating most workers found the system responses less interesting (see Fig. 2f) with a median score of 37 compared to 50 for **Setup 2**.

Scores for *explanation quality* are skewed towards the left, with a median score of 25 (**Setup 1**) and 22 (**Setup 2**), as shown in Fig. 2g and 2h. This shows that even though most workers find that system recommendations relevant there is a lack of explanation on why the recommendations are made. These findings are in line with recent work conducted by Guo et al. [19], showing that most conversational recommender system (CRS) dialogues lack explanation in their recommendations.

5 EFFECT OF USER FEEDBACK

In this section, we answer **(RQ1)**: How does user feedback from the follow-up utterance influence the evaluation labels collected from both crowdworkers and LLMs?

Distributions. For the *crowdsourced labels*, each turn is annotated by three workers; their ratings are averaged to get the overall score per turn. Fig. 3 (a)–(d) show the density distributions of the scores: *Relevance and usefulness* are skewed towards the right (Fig. 3a and 3b) for both setups, showing that more turns are found to be more relevant and useful by the crowdworkers. For *usefulness*, the peak for **Setup 2** is towards the center compared to **Setup 1**, indicating a decrease in the number of turns that are highly useful as workers have access to the user's follow-up utterance, adding more context during the assessment. Cases where the system makes a relevant recommendation, are rated highly useful in **Setup 1**, but the user's feedback in **Setup 2** changes the worker's rating to lower values in certain cases. E.g., in cases where the user has already watched the movie or even though the movie satisfies their requirements

(e.g., genre and actor), they do not like other aspects. *Interestingness* has a more central peak with a wide range showing there was a lot of variability in the assessment (Fig. 3c). More turns are assessed as interesting in **Setup 2** compared to **Setup 1**, with fewer turns being scored as highly interesting in both setups (80–100). Similar observations pertain to *explanation quality* (Fig. 3d).

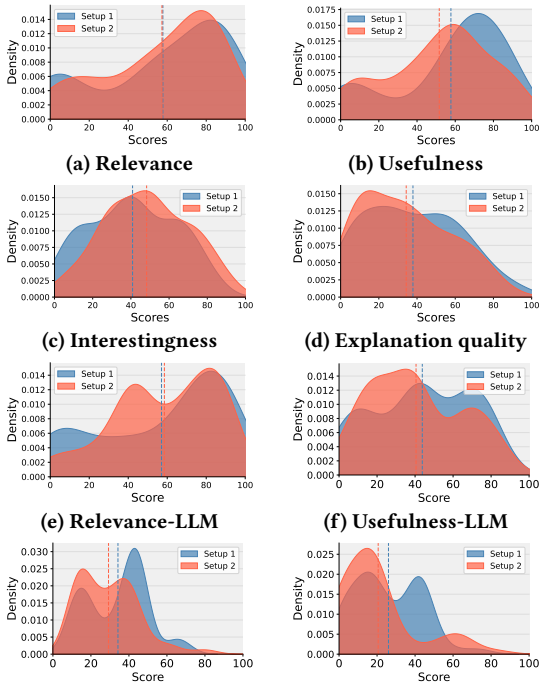
We also plot the distribution of scores from the LLM in Fig. 3 (e)–(f). The *relevance* kernel density estimation (KDE) plot exhibits a dual-peak distribution with a minor peak at lower values and a more pronounced peak at higher values, notably around 80 and above (Fig. 3e). In **Setup 2**, the KDE plot shows a distribution peaking in the mid to higher range of the score scale.

In contrast, *usefulness* (Fig. 3f) shows three peaks in **Setup 1** (not significant from each other), with **Setup 2** having a distinctive peak between scores of 10–40. The slight peak towards the high scores compared to **Setup 1** indicates that with the user's follow-up utterance, the LLM finds most turns not to be highly useful. We observe a different pattern for *interestingness* and *explanation quality* with scores skewed towards the left showing that the LLM rates most turns low on interestingness and explanation similar to observations made from the crowdworker scores.

Humans vs. LLM. The different peaks between the two setups across the four aspects indicate a significant divergence in how crowdworkers and LLMs perceive and rate the aspects in different setups. **Setup 1** is characterized by high peaks in high-range scores, compared to **Setup 2** which exhibits peaks in the moderate range except for relevance, which has both moderate and high peaks. This contrast suggests that user feedback from the follow-up utterance has a notable impact on both the crowdworkers and LLM assessments.

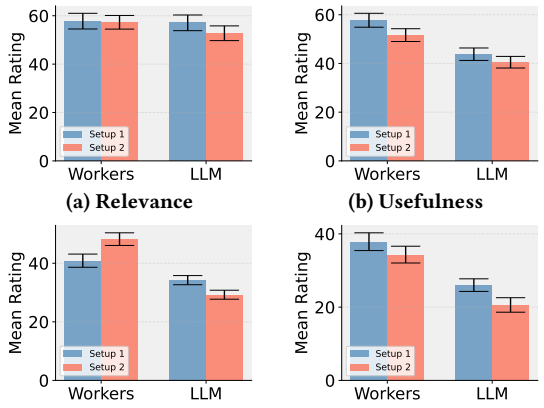
External agreement. Next, we compute the overall mean score for each setup for both the crowdworkers and the LLM with confidence intervals; see Fig. 4. There is no significant difference in *relevance scores* between the two annotator groups. This indicates that the presence of the user's follow-up utterance does not significantly affect the relevance assessment. Relevance primarily relies on the system's ability to provide a topically relevant recommendation to the user's request, and having only the user's initial request appears sufficient for this assessment. However, some differences in relevance scoring emerge when the follow-up utterance is available to certain workers. In many instances, workers influenced by the follow-up utterances tend to assign high scores to non-relevant system responses if the user accepts the recommendation, even if it deviates from their initial request. Conversely, they may assign low scores to relevant system responses if the user dislikes the recommendation based on specific attributes, despite its topical relevance. Similar patterns are observed in the scores assigned by LLM, with the LLM having a low mean score for **Setup 2**.

LLM-based annotations are consistently lower in terms of *usefulness scores* compared to crowdworkers', as indicated by lower mean scores in both setups (Fig. 4b). The mean scores are statistically significant for crowdworkers but not for the LLM, highlighting the substantial influence of the follow-up utterance on usefulness assessments. In **Setup 1**, crowdworkers assign high usefulness scores to system responses, closely aligning with relevance scores. This



(a) Relevance (b) Usefulness
(c) Interestingness (d) Explanation quality
(e) Relevance-LLM (f) Usefulness-LLM
(g) Interestingness-LLM (h) Explanation qual.-LLM

Figure 3: Kernel density estimation plots comparing aggregated crowdworker and LLM scores for both setups. The dotted lines represent the overall mean for each setup.



(a) Relevance (b) Usefulness
(c) Interestingness (d) Explanation quality

Figure 4: Mean rating for each aspect across the two setups, for both the crowdworkers and LLM.

suggests that annotators assess usefulness like relevance in the absence of the follow-up utterance.

Conversely, in **Setup 2**, there is a significant drop in the mean usefulness score. This reflects workers transitioning from assessing relevance to considering how well the system response addresses various facets of the user's needs, often revealed in the follow-up utterance. E.g., a user may request an action movie initially, but specific preferences may emerge in the subsequent utterance, such as actor or director preferences as users typically reveal their complete information needs through a back-and-forth exchange [1].

In contrast to other aspects, *interestingness* presents an intriguing observation. Workers assign lower scores in **Setup 1** compared

Table 2: Spearman's r correlation coefficient between the aspects and expert user satisfaction ratings for both the crowdworkers and LLM. * indicate non-significant values ($p < 0.05$).

	Crowdworkers		LLM	
Aspects	**Setup 1**	**Setup 2**	**Setup 1**	**Setup 2**
Relevance	0.56	0.51	0.63	0.52
Usefulness	0.55	0.66	0.45	0.41
Interestingness	0.31*	0.39	0.27*	0.21*
Explanation quality	0.44	0.42	0. 54	0.50

to **Setup 2** (Fig. 4c). Both groups of annotators assign lower scores for interestingness in both setups, and these differences in mean scores are statistically significant. Examination of annotators' justifications reveals that they hold strict criteria for rating system responses as interesting and had relatively high expectations for a response to be deemed interesting. This is reflected in the score distributions depicted in Fig. 3c, where a smaller proportion of turns receive a score of 100 for interestingness. The disparity between the setups is particularly notable in **Setup 1**, where only 2% of the turns received a score of 90 or higher, as opposed to 7% in **Setup 2**.

The mean score for explanation quality is not statistically significant for crowdworkers, although there is a noticeable drop of over 3 points from **Setup 1** to **Setup 2**. However, it is statistically significant for LLM annotators. It is worth noting that this aspect consistently yields low mean scores compared to the other aspects, ranging from 20 to 37, as indicated by the mean bars in Fig. 4d. Fewer turns receive a perfect score of 100 in the aggregated scores for **Setup 2**; see Fig. 3d. Both annotator groups agree in assigning lower scores for this aspect, highlighting the lack of recommendation explanations in the dataset. An analysis of annotator justifications reveals varying expectations regarding system explainability. While some workers expect the system to provide explanations related to aspects like the movie's cast or director, others focus on different facets. This subjectivity in worker bias and expectations regarding system explanations contributes to the variation in scores for explanation quality.

Humans vs. LLM. Overall, both groups agree on assessing relevance, but differences emerge when considering follow-up utterances, influencing relevance scores for LLM. LLMs consistently assign lower usefulness scores than human annotators, indicating the challenge of defining usefulness when follow-up utterances reveal complex user needs. Unlike humans, LLMs do not personalize the system's usefulness to the user. This highlights the importance of including follow-up utterances for more accurate evaluation labels that reflect the user's perspective. Both groups agree on the lack of explanations from the system.

Agreement with expert ratings. Here, we examine how well human and LLM labels align with expert ratings. We collect expert ratings on the user satisfaction aspect to investigate the correlation of the fine-grained aspects to overall user satisfaction following Siro et al. [40]. Using the same setup, we collect the expert ratings from two experts. Since our initial annotation was on an S100 scale we transformed the labels to S3 scale (1–3) for all aspects [22] and then calculated the Spearman's r between each aspect and the expert rating. We report our results in Tab. 2.

In **Setup 1**, relevance displays a moderate positive correlation of 0.56 (crowdworkers) and 0.63 (LLM) with expert ratings, indicating a similar alignment with expert satisfaction in the absence of follow-up utterances. Usefulness shows stronger correlations, with 0.55 for crowdworkers and 0.45 for LLM, suggesting that usefulness judgments are significantly influenced by the absence of follow-up utterances in this setup. Interestingness exhibits weaker correlations in both groups, suggesting potential challenges or subjectivity in assessing this aspect. Explanation quality demonstrates moderate correlations (0.47 for crowdworkers and 0.54 for LLM), indicating moderate alignment with expert satisfaction ratings.

In **Setup 2**, relevance maintains positive correlations, 0.51 (crowdworkers) and 0.52 (LLM). Usefulness shows notably stronger correlations, 0.66 for crowdworkers and 0.41 for LLM, indicating that crowdworkers assign scores closely aligned with the user feedback. Interestingness continues to exhibit weaker correlations (0.39 for crowdworkers and 0.21 for LLM). Explanation quality, while still aligned with expert ratings, has slightly weaker correlations (0.44 for crowdworkers and 0.50 for LLM) compared to **Setup 1**.

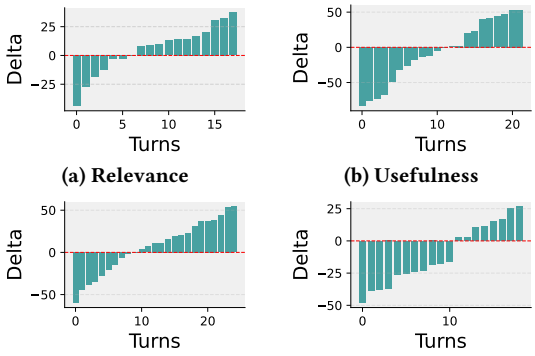
Humans vs. LLM. Correlations with expert ratings highlight that relevance and usefulness assessments generally have a stronger alignment with expert user satisfaction ratings across setups and annotator types. However, interestingness shows weaker correlations, indicating potential challenges in assessing this aspect consistently and objectively [39]. In general, we note that humans perform well in assessing user experience measures such as usefulness and interestingness while LLMs perform well in assessing utility measures such as relevance and explanation quality.

Further analysis of the correlation of the aspects in the two setups and task duration is available in our **Online Appendix**.

6 SIGNIFICANCE OF USER FEEDBACK

In this section, we examine the impact of user feedback from follow-up utterances on reducing annotator variability in their evaluation labels as part of our analysis to answer **(RQ2)**. To assess agreement, we calculate the standard deviation of workers' scores for each turn and categorize the data into two groups: **Group 1** (scores below the median standard deviation, indicating high agreement) and **Group 2** (scores above the median, indicating low agreement). We find that *interestingness* and *explanation quality* consistently exhibit higher agreement among annotators when the system response is uninteresting or lacks explanation. However, there is no clear agreement pattern among workers for *relevance* and *usefulness*.

We compare turns in Group 2 from **Setup 1** with Group 1 from **Setup 2**, where Group 2 initially exhibits high variability in evaluation labels, but shows increased agreement in **Setup 2** due to the presence of the user's follow-up utterance. Specifically, there are 18 turns for relevance, 22 for usefulness, 25 for interestingness, and 19 for explanation quality in this analysis (Group 2 initially consisted of 48 to 50 turns). Overall, at least 30% of the turns demonstrate improved worker agreement in **Setup 2**. To quantify score differences between Group 1 and Group 2, we calculate their delta and present the results in Fig. 5. We observe significant score differences for the same instances under different conditions. Relevance and interestingness have more turns rated highly in Group 1 (positive delta scores) (Fig. 5a & 5c) while usefulness and explanation quality have turns rated low in Group 1 (Fig. 5b & 5d).



(a) Relevance (b) Usefulness

(c) Interestingness (d) Explanation quality

Figure 5: Difference in scores assigned to dialogues turns for four aspects in Group 1 with low variability vs. dialogues in Group 2 with high variability between the worker scores from the mean rating.

Manual analysis. Our manual analysis primarily focused on the usefulness aspect due to its substantial impact, with the highest mean delta difference (35) compared to other aspects (interestingness: 26, explanation quality: 22, relevance: 15). We analyze 22 turns to identify instances where the user's follow-up utterances notably enhance worker agreement, shedding light on cases where the presence of the user's next utterance significantly improves consensus.

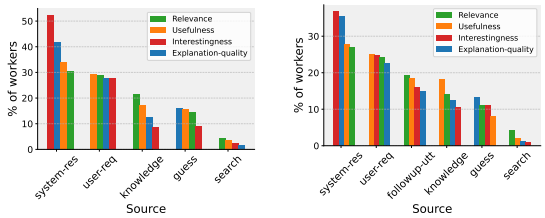
This analysis identifies specific scenarios where user feedback plays a pivotal role, such as addressing ambiguous requests by providing clarity, making generic requests more specific and actionable, simplifying complex requests, and compensating for annotators' lack of domain knowledge. In these scenarios, user feedback consistently improves the overall quality and consistency of the annotation process, highlighting its significance in enhancing system evaluations.[4]

Apart from resolving uncertainty in user requests, follow-up utterances are crucial when annotators encounter unfamiliar topics. An analysis of annotators' justifications reveals that when annotators lack prior knowledge, the user's knowledge about the recommended item, coupled with explicit feedback, bridges the knowledge gap, resulting in more precise evaluations of the system's performance, even when annotators lack subject matter expertise.

7 SOURCES AND BIAS

In this section, we answer **(RQ3)**. To understand the basis of workers' assessments and their choices when assigning evaluation labels, we conducted a survey where we asked workers to indicate the sources of information they relied on when making judgments. Both setups offered the same options for information sources, except for follow-up utterance, which was available only in **Setup 2**. The available sources included *personal knowledge, searched online, guessed, user request,* and *system response.* Additionally, we performed a manual analysis of the workers' justifications to evaluate any potential biases introduced by their chosen information sources.

[4]A detailed example of a complex user request can be found in our **Online Appendix** in Sec. B.1.



(a) **Setup 1** (b) **Setup 2**

Figure 6: Distribution of sources workers relied on to make their judgments for the four aspects in the two setups.

Sources. Fig. 6 shows the percentage of workers who relied on each information source during evaluation. The x-axis represents the information sources, while the y-axis indicates the percentage of workers relying on each source. Across both setups, it is evident that workers predominantly depend on information within the dialogue itself, specifically the user's request and the system's response, for their assessments. Interestingly, the evaluation of interestingness and explanation quality was primarily influenced by the system's response. While explanation quality considered both the user request and system response, we observe a notably higher reliance on the system's response compared to the user request. On the contrary for relevance and usefulness assessment, we observe that workers mostly rely on the user request to ensure the system meets the user's need. However, there is a marginal difference between system response and user request for these two aspects showing that the two are equally important during assessment. Without the user's follow-up utterance, some workers make an educated guess on the relevance and usefulness of the system response.

In **Setup 2** (Fig. 6b), we note a drop in the percentage of workers relying on the system response during evaluation, showing that the follow-up utterance introduces another dynamic to be considered by the workers during the assessment. Usefulness which measures how well the system response meets the user's needs has a high percentage of workers relying on the user's follow-up utterance. Unexpectedly we note a high number where workers relied on personal knowledge to gauge the usefulness of the system response, showing that they introduce personal bias in assessing this aspect.

A few workers utilize online sources, primarily when assessing relevance and usefulness, implying that workers without domain knowledge leverage online information for more accurate assessments. Interestingly, there is a decrease in the percentage of workers using online sources in **Setup 2**, specifically for evaluating usefulness. This suggests that the introduction of the user's follow-up utterance in **Setup 2** acts as an additional information source, assisting workers lacking domain knowledge in assessing the system response's usefulness.

Biases. Several studies highlight the influence of biases on crowdworkers' judgments [e.g., 16, 24, 38]. In our work, we specifically explore how the sources of information outlined in Sec. 7 introduce biases into crowdsourced labels. Fig. 6 illustrates a reliance on online sources for assessment, which, while potentially augmenting workers' domain knowledge, can introduce specific biases, such as popularity bias [11]. From the analysis of workers' justifications, we observe instances in assessing usefulness where some workers forego the user's feedback on the system's response and rely on online movie reviews. Justifications like "The movie seems to be liked

by many so it is useful," and "The movie is not rated highly" are observed, indicating that these external sources could bias workers, leading to ratings that may not accurately reflect the user feedback.

A notable percentage of workers rely on the user's request when assessing interestingness. However, interestingness assessments should primarily be based on the system's response. Therefore, we note several workers get biased by the relevance of the recommended item to the user request during their assessments. To mitigate this bias, it may be prudent to restrict access to the user request when evaluating aspects like interestingness, where the focus is on assessing the system independently of the user's input.

In comparison to other aspects, we notice that workers rely on their knowledge to evaluate the usefulness of the system response (see Fig. 6), introducing personal preference bias despite the explicit tie to the user's needs. Also, workers display a bias towards rating longer system responses highly for explanation quality.

8 DISCUSSION AND CONCLUSION

In this work, we addressed the question of how the inclusion of user feedback, both implicit and explicit, from the user's follow-up utterance, influences the evaluation labels from crowdworkers and LLMs. Our analysis revealed intriguing patterns across various aspects, providing valuable insights into the impact of user feedback on the quality of assessments.

Relevance. We considered two experimental setups, one without the user's follow-up utterance and one with. In both setups, crowdworkers and LLM-based annotators largely concur when evaluating relevance. However, subtle differences emerge with the inclusion of follow-up utterances, particularly for LLMs which tend to assign lower scores compared to humans. Though there is no significant mean difference in **Setup 1** for both annotator groups, we note that LLM scores show a higher correlation to expert user satisfaction ratings than humans. Crowdworkers relied more on the system response, user request, and their prior knowledge to gauge the relevance of the recommended item. With lots of candidate movies to recommend, crowdworkers may lack knowledge of some of these movies to assess their relevance, which results in making an educated guess. However, compared to humans LLMs are rich in internal knowledge on these movies, thus improving their relevance assessment.

Usefulness. The usefulness ratings show a distinctive contrast between the two annotator groups. Human annotators displayed strong correlations with user satisfaction in **Setup 2**, suggesting their ability to personalize the system's usefulness to individual users. In contrast, LLMs consistently assigned lower usefulness scores in both setups, highlighting the challenge of assessing usefulness when follow-up utterances reveal conflicting user needs from their initial request. This shows that when user feedback conflicts with a model's internal knowledge it leads to inconsistency in the ratings.

Interestingness. The aspect of interestingness presented unique challenges. Both crowdworkers and LLMs exhibited lower correlations with user satisfaction ratings, indicating that both annotator groups struggled to capture the user's subjective perception of interestingness. The presence of user feedback had a limited impact on improving assessments in this aspect. This is also observed with the low Kappa and ICC scores in Tab. 1. Nonetheless, utterances such as "that's interesting," "sounds good," and "you are funny" lead to an increase in annotator agreement and correlation with user satisfaction ratings (**Setup 2**), emphasizing the significance of user feedback in improving system evaluation.

Explanation quality. Both annotator groups concur on the absence of explanations provided by the system. This shared observation underscores a significant limitation in current CRS, as both human evaluators and LLMs noted the lack of explanatory content in system responses. The LLM shows less sensitivity to user feedback with high correlating scores to overall user satisfaction compared to humans. Humans' ratings are affected by their personal expectations of the system's explainability, which is not evident in the LLM scores. This shows that LLMs can maintain objectivity when assessing system performance.

In general, there is a distinct difference in ratings assigned by both annotator groups in the two setups, indicating that user feedback does influence system evaluation. Human workers are more susceptible to user feedback in usefulness and interestingness compared to LLMs in interestingness. User feedback leads to personalized usefulness assessment by workers and improves worker agreement when uncertainty arises in the user's request. The lack of adaptability to user feedback by the LLM in assessing usefulness, suggests that LLMs may require additional mechanisms such as prompt tuning to enhance user-centric evaluations by LLMs. Therefore, it is important to assign annotation tasks to LLMs based on the nature of the task, leveraging their strengths in objective assessments like relevance annotation while complementing with human assessors for tasks demanding subjective evaluation or sensitivity to user preferences and feedback. Combining human annotators and LLMs can lead to better system evaluations by leveraging the unique strengths of each type of annotator for specific evaluation tasks.

However, user feedback can sometimes lead to assessments that do not align with overall user satisfaction, resulting in lower correlation scores as observed for the relevance aspect in **Setup 2**. It is important to note that this study employed a single LLM for annotation, and results may vary with different LLMs. Additionally, potential biases in the crowdworker pool and the LLM's training data could influence the findings. For future work, we will conduct further research to validate our findings across diverse conversational systems.

Online Appendix. Further analyses and discussions are available at https://github.com/Clemenciah/LLMCrowdDialogueEval/tree/main/Appendix.

Data. We publicly release the annotated data at https://github.com/Clemenciah/LLMCrowdDialogueEval/tree/main/Data.

Acknowledgments. This research was supported by the Dreams Lab, a collaboration between Huawei Finland, the University of Amsterdam, and the Vrije Universiteit Amsterdam, by the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.-1389.20.183, and KICH3.LTP.20.006, and by the European Union's Horizon Europe program under grant agreement No 101070212.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 475–484. https://doi.org/10.1145/3331184.3331265

[2] Omar Alonso, Daniel E. Rose, and Benjamin Stewart. 2008. Crowdsourcing for Relevance Evaluation. *SIGIR Forum* 42, 2 (2008), 9–15. https://doi.org/10.1145/1480506.1480508

[3] Amazon Mechanical Turk. 2023. https://www.mturk.com.

[4] Leif Azzopardi. 2021. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. In *CHIIR '21: ACM SIGIR Conference on Human Information Interaction and Retrieval, Canberra, ACT, Australia, March 14-19, 2021*, Falk Scholer, Paul Thomas, David Elsweiler, Hideo Joho, Noriko Kando, and Catherine Smith (Eds.). ACM, 27–37. https://doi.org/10.1145/3406522.3446023

[5] Krisztian Balog and Filip Radlinski. 2020. Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 329–338. https://doi.org/10.1145/3397271.3401032

[6] David La Barbera, Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2020. Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12036)*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer, 207–214. https://doi.org/10.1007/978-3-030-45442-5_26

[7] Praveen Kumar Bodigutla, Lazaros Polymenakos, and Spyros Matsoukas. 2019. Multi-domain Conversation Quality Evaluation via User Satisfaction Estimation. *CoRR* abs/1911.08567 (2019). arXiv:1911.08567 http://arxiv.org/abs/1911.08567

[8] Praveen Kumar Bodigutla, Aditya Tiwari, Spyros Matsoukas, Josep Valls-Vargas, and Lazaros Polymenakos. 2020. Joint Turn and Dialogue level User Satisfaction Estimation on Mulit-Domain Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 3897–3909. https://doi.org/10.18653/V1/2020.FINDINGS-EMNLP.347

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[10] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A Survey on Evaluation of Large Language Models. *CoRR* abs/2307.03109 (2023). https://doi.org/10.48550/ARXIV.2307.03109 arXiv:2307.03109

[11] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Trans. Inf. Syst.* 41, 3 (2023), 67:1–67:39. https://doi.org/10.1145/3564284

[12] Jason Ingyu Choi, Ali Ahmadvand, and Eugene Agichtein. 2019. Offline and Online Satisfaction Prediction in Open-Domain Conversational Systems. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 1281–1290. https://doi.org/10.1145/3357384.3358047

[13] Cyril W. Cleverdon, Jack Mills, and Michael Keen. 1966. Factors determining the performance of indexing systems. In *ASLIB Cranfield research project.*

[14] Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2020. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review* 54 (2020), 755–810.

[15] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a Good Data Annotator?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 11173–11195. https://doi.org/10.18653/V1/2023.ACL-LONG.626

[16] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek (Eds.). ACM, 162–170. https://doi.org/10.1145/3159652.3159654

[17] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity is a Worthwhile Quality: On the Role of Task Clarity in Microtask Crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT 2017, Prague, Czech Republic, July 4-7, 2017*, Peter Dolog, Peter Vojtás, Francesco Bonchi, and Denis Helic (Eds.). ACM, 5–14. https://doi.org/10.1145/3078714.3078715

[18] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *CoRR* abs/2303.15056 (2023). https://doi.org/10.48550/ARXIV.2303.15056 arXiv:2303.15056

[19] Shuyu Guo, Shuo Zhang, Weiwei Sun, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2023. Towards Explainable Conversational Recommender Systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 2786–2795. https://doi.org/10.1145/3539618.3591884

[20] Lei Han, Eddy Maddalena, Alessandro Checco, Cristina Sarasua, Ujwal Gadiraju, Kevin Roitero, and Gianluca Demartini. 2020. Crowd Worker Strategies in Relevance Judgment Tasks. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang (Eds.). ACM, 241–249. https://doi.org/10.1145/3336191.3371857

[21] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. All Those Wasted Hours: On Task Abandonment in Crowdsourcing. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, J. Shane Culpepper, Alistair Moffat, Paul N. Bennett, and Kristina Lerman (Eds.). ACM, 321–329. https://doi.org/10.1145/3289600.3291035

[22] Lei Han, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2019. On Transforming Relevance Scales. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 39–48. https://doi.org/10.1145/3357384.3357988

[23] Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2023. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. *CoRR* abs/2303.16854 (2023). https://doi.org/10.48550/ARXIV.2303.16854 arXiv:2303.16854

[24] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos (Eds.). ACM, 407. https://doi.org/10.1145/3290605.3300637

[25] Jiepu Jiang, Daqing He, and James Allan. 2017. Comparing In Situ and Multidimensional Relevance Judgments. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 405–414. https://doi.org/10.1145/3077136.3080840

[26] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Inf. Retr.* 16, 2 (2013), 138–178. https://doi.org/10.1007/s10791-012-9205-0

[27] Jin Young Kim, Jaime Teevan, and Nick Craswell. 2016. Explicit In Situ User Feedback for Web Search Results. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, 829–832. https://doi.org/10.1145/2911451.2914754

[28] Youngho Kim, Ahmed Hassan Awadallah, Ryen W. White, and Imed Zitouni. 2014. Modeling dwell time to predict click-level satisfaction. In *Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY, USA, February 24-28, 2014*, Ben Carterette, Fernando Diaz, Carlos Castillo, and Donald Metzler (Eds.). ACM, 193–202. https://doi.org/10.1145/2556195.2556220

[29] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 9748–9758. https://proceedings.neurips.cc/paper/2018/hash/800de15c79c8d840f4e78d3af937d4d4-Abstract.html

[30] Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. "Satisfaction with Failure" or "Unsatisfied Success": Investigating the

Relationship between Search Success and User Satisfaction. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 1533–1542. https://doi.org/10.1145/3178876.3186065

[31] Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. 2015. Different Users, Different Opinions: Predicting Search Satisfaction with Mouse Movement Information. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, Ricardo Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier A. Ribeiro-Neto (Eds.). ACM, 493–502. https://doi.org/10.1145/2766462.2767721

[32] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. 2017. On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation. *ACM Trans. Inf. Syst.* 35, 3 (2017), 19:1–19:32. https://doi.org/10.1145/3002172

[33] Jiaxin Mao, Yiqun Liu, Noriko Kando, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. Investigating Result Usefulness in Mobile Search. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 10772)*, Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury (Eds.). Springer, 223–236. https://doi.org/10.1007/978-3-319-76941-7_17

[34] Tyler McDonnell, Matthew Lease, Mücahid Kutlu, and Tamer Elsayed. 2016. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2016, 30 October - 3 November, 2016, Austin, Texas, USA*, Arpita Ghosh and Matthew Lease (Eds.). AAAI Press, 139–148. https://doi.org/10.1609/HCOMP.V4I1.13287

[35] Shikib Mehri and Maxine Eskénazi. 2020. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 681–707. https://doi.org/10.18653/V1/2020.ACL-MAIN.64

[36] Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. 2018. On Fine-Grained Relevance Scales. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 675–684. https://doi.org/10.1145/3209978.3210052

[37] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can The Crowd Identify Misinformation Objectively?: The Effects of Judgment Scale and Assessor's Background. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 439–448. https://doi.org/10.1145/3397271.3401112

[38] Sashank Santhanam, Alireza Karduni, and Samira Shaikh. 2020. Studying the Effects of Cognitive Biases in Evaluation of Conversational Agents. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik (Eds.). ACM, 1–13. https://doi.org/10.1145/3313831.3376318

[39] Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2022. Understanding User Satisfaction with Task-oriented Dialogue Systems. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 2018–2023. https://doi.org/10.1145/3477495.3531798

[40] Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2023. Understanding and Predicting User Satisfaction with Conversational Recommender Systems. *ACM Trans. Inf. Syst.* 42, 2 (sep 2023), Article 55. https://doi.org/10.1145/3624989

[41] Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Context Does Matter: Implications for Crowdsourced Evaluation Labels in Task-Oriented Dialogue Systems. *arXiv preprint arXiv:2404.09980* (2024).

[42] Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating User Satisfaction for the Evaluation of Task-oriented Dialogue Systems. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2499–2506. https://doi.org/10.1145/3404835.3463241

[43] Pertti Vakkari, Michael Völske, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Modeling the usefulness of search results as measured by information use. *Inf. Process. Manag.* 56, 3 (2019), 879–894. https://doi.org/10.1016/J.IPM.2019.02.001

[44] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want To Reduce Labeling Cost? GPT-3 Can Help. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 4195–4205. https://doi.org/10.18653/V1/2021.FINDINGS-EMNLP.354

[45] Shuo Zhang and Krisztian Balog. 2020. Evaluating Conversational Recommender Systems via User Simulation. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (Eds.). ACM, 1512–1520. https://doi.org/10.1145/3394486.3403202

[46] Weinan Zhang, Lingzhi Li, Dongyan Cao, and Ting Liu. 2018. Exploring Implicit Feedback for Open Domain Conversation Generation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 547–554. https://doi.org/10.1609/AAAI.V32I1.11253

[47] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. *Found. Trends Inf. Retr.* 14, 1 (2020), 1–101. https://doi.org/10.1561/1500000066

[48] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast, QLD, Australia - July 06–11, 2014*, Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin (Eds.). ACM, 83–92. https://doi.org/10.1145/2600428.2609579