

Factors Influencing the Perceived Meaningfulness of System Responses in Conversational Recommendation

Ahtsham Manzoor^{1,*}, Wanling Cai² and Dietmar Jannach¹

¹University of Klagenfurt, Universitätsstraße 65-67, Klagenfurt am Wörthersee, 9020, Austria

²Trinity College Dublin & Lero, College Green, Dublin 2, Ireland

Abstract

Conversational recommender systems (CRS) support users in finding satisfying items via multi-turn dialogs and have recently attracted increasing attention. Many research efforts have been made in producing quality responses including item recommendations, which can be jointly assessed in a user-centric manner using a subjective criterion, i.e., *meaningfulness*. However, human perceptions of meaningfulness are complex and can be nuanced, as individual users have their own preferences over conversations, and their perceptions can be affected by various factors, such as users' personal characteristics, system functionalities like informativeness of the generated responses, and dialog context. To better design CRS adapted to user needs and context, this work investigates the impact of three types of factors (i.e., user-related, system-related, and context-related) on users' perceived meaningfulness of system responses by analyzing a within-subject user study (N=90) data. Results indicate that users' *domain knowledge* and the *informativeness* of system responses positively influence users' perceived meaningfulness of system responses. A deep investigation reveals that users with previous chatbot experience tend to expect highly informative conversations. Moreover, older users seem to be less satisfied with lengthy dialogs.

Keywords

Conversational recommendation, personal characteristics, interaction context, evaluation

1. Introduction

Conversational recommender systems (CRS) assist users in finding items of interest and support their decision-making process while conversing with the system in natural language. Technically, regarding the development of a CRS, we observe two streams of works, i.e., retrieval-based and generation-based approaches to CRS [1]. However, unlike traditional recommender systems that mainly provide one-shot recommendations (e.g., a ranked list of items), both kinds of approaches facilitate “free-style” multi-turn conversations between a user and system. From the interaction standpoint, a user can express her preferences to the system (e.g., “*Can you suggest any good sci-fi movies?*”), and the system in response can proactively suggest recommendations or takes

IntrRS'23: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, September 18, 2023, Singapore (hybrid event).


*Corresponding author.

✉ ahtsham.manzoora@au.at (A. Manzoora); wanling.cai@tcd.ie (W. Cai); dietmar.jannach@au.at (D. Jannach)

ORCID 0000-0001-9418-7539 (A. Manzoora); 0000-0001-8506-3825 (W. Cai); 0000-0002-4698-8507 (D. Jannach)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

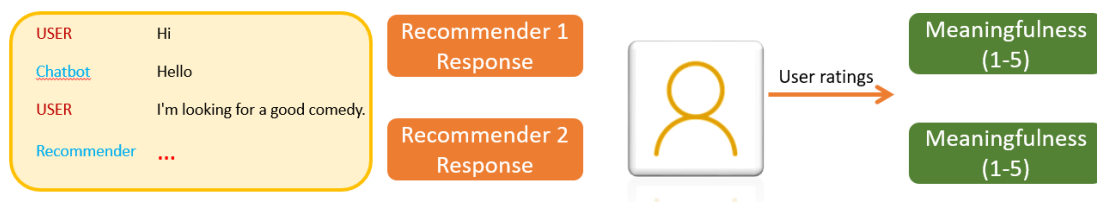


Figure 1: Procedure for Assessing the Meaningfulness of System Responses

the lead to elicit further user preferences, (e.g., “Are you looking for recent or old movies?”), thereby supporting the recommendation process via *producing* effective dialogs, see also [2].

Regarding the evaluation, a CRS can be evaluated using both computational (“*offline*”) experiments or studies involving humans, see e.g., a survey on the evaluation of CRS [3]. Given that various user-centric metrics to evaluate item recommendations are available [38], human evaluations of CRS predominantly centered around assessing *only* linguistic quality aspects like fluency, consistency, or naturalness [4, 5, 6], and if the quality of item recommendations is adequate in the given dialog context seems missing. For example, the system-response “have you seen *Black Panther (2018)*?” to a user utterance asking for *romantic* movie recommendations appears obscure [7, 8]. In addition, approaches that determine a relative ranking of different systems via human evaluators, as was done in [9], do not inform us if the highly ranked system would be useful in practice [10].

To address such limitations, the concept of *meaningfulness*, as an evaluation criterion of system responses, was introduced in [7]. In this approach, the authors specifically relied on a single subjective metric for assessing the overall quality of system responses, including recommendations, as shown in Fig 1. Specifically, *perceived meaningfulness* refers to assessing the recommendation quality, response appropriateness, and coherency in a given dialog context. An example of when a response should be meaningful can be when a user expresses interest in horror movies. In such cases, a response like “*The Conjuring (2013) is a really good one*” should be considered meaningful. More such cases and examples can be found in [11].

Recent works on CRS however have shown that human perceptions of quality may differ between groups of people and can impact user trust and satisfaction [12, 13, 14, 15, 16]. For example, users with a higher level of domain knowledge may benefit more from conversational recommendation interactions due to their ability to express their preferences better than novices [17]. Similarly, more aware users may expect shorter but more efficient dialogs. In addition, Araujo et al. [18] suggest that personal characteristics and context can be linked to different perceptions of automated decision-making. Likewise, individual perceptions of meaningfulness can be complicated and nuanced. For example, in a study on CRS responses regarding cases about seekers’ specific questions [19], participants prefer system responses that provide appropriate recommendations over responses that are entirely irrelevant. Yet, it remains unclear how different aspects corresponding to the user, system, and the context in which a dialog happens influence human perceptions of the meaningfulness of system responses.

Therefore, in this work, we identify and categorize different aspects into one of the three categories, i.e., user, system, or context, and investigate the following research questions (RQs).

RQ1: How do user-related, system-related, and context-related factors affect users' perceived meaningfulness of responses of a CRS?

RQ2: How do various factors interact to affect the perceived meaningfulness of responses in a CRS?

To address these questions, based on the literature on CRS, in particular works like [20] and [21], we carefully derived a set of user-related characteristics such as gender, age, domain knowledge, or prior chatbot experience. To study system-related factors, we rely on two CRS, i.e., KBRD [6] and CRB-CRS [19]. Finally, inspired by [19], we derived a contextual factor, i.e. dialog discourse length, which might have an impact on meaningfulness. Additional factors like dialog initiative strategy could be taken into account as well, but due to the non-availability of ReDial [9] data annotations, we consider that to be beyond the scope of the current work.

Overall, we examined user feedback data collected in an online study conducted via the MTurk crowdsourcing platform with 90 participants. Specifically, the participants were presented with a dialog situation curated from the ReDial dataset where two humans, one with a *user* role and the other as a *recommender*, discuss movie recommendations. The dialog situation always ends with a user (or: seeker) utterance followed by responses by two CRS that appeared in random order, as shown in Figure 1. Different system responses are evaluated in parallel. We note that this is in principle not *necessary* to understand the given RQs and may lead to implicit bias; but here the intuition was to collect user feedback through a smaller number of user interactions.

Our findings reveal that users with domain knowledge and informativeness of responses positively influence their perceptions of meaningfulness. Also, users with prior chatbot experience tend to have higher expectations and require more advanced features for satisfactory CRS performance. In addition, we found that the user's age affects the relationship between the dialog discourse length and perceived meaningfulness. CRS practitioners should aim to provide more concise responses for older users as relatively young users seem reasonably satisfied. To our best knowledge, this is the first work investigating the factors that influence the perceived meaningfulness of responses in the context of conversational recommendation. We believe our findings will contribute to the research on AI dialog systems and facilitate improved CRS design in terms of a more personalized and inclusive user experience.

2. RELATED WORK

Conversational Recommender Systems Conversational recommender systems (CRS) that use natural language offer free-style conversations with users and help them make online choices. We observe two main types of CRS: language generation-based and retrieval-based. Language generation-based systems are popular due to their ability to incorporate new contexts, but recent studies have found that they often generate identical responses [10, 7]. Retrieval-based methods on the other hand fetch appropriate responses from recorded dialog datasets, and adapt them to the ongoing dialog context. Such responses are usually informative, fluent, and semantically meaningful as they were originally made by humans. However, retrieval approaches struggle with unseen dialog contexts or open-ended questions, see also [22] for a comparison between generation and retrieval-based approaches in NLP. To study system-related aspects, we consider two recent open-source CRS (KBRD and CRB-CRS) as representatives of

both streams of work. These systems have shown significant performance compared to others and were published in high-quality research venues.

Technically, KBRD [6], a generation-based CRS, relies on a sequence-to-sequence encoder-decoder Transformer framework [23] while mapping user domain concepts to an additional knowledge graph, named DBpedia [24] to aid informativeness in the generated responses. The authors chose the Transformer framework over HRED [25] due to its better performance in various NLP tasks, such as Q&A or machine translation [26, 27, 28]. On the other hand, CRB-CRS is a contextual retrieval-based system that utilizes a dialog corpus to fetch, adapt and integrate item recommendations given the user utterance or dialog history as input. Note that both CRS are developed in the context of the *ReDial* dataset, see more details about the collection procedure of the dataset in [9], which provides further grounds to construct fair analyses in our study.

Evaluation A CRS can be evaluated based on various quality dimensions, including system effectiveness, efficiency, conversation quality, and subtask effectiveness, see also a survey on CRS evaluation in [3]. Quality measurements can be assessed through offline experiments or studies involving humans. Computational experiments typically evaluate system effectiveness using metrics like Recall or RMSE [29, 30, 31], and linguistic quality measures such as BLEU and Perplexity [32, 33, 4, 34]. However, the representativeness of such measures for human perceptions remains unresolved [8]. In the CRS domain, human studies mainly assess linguistic quality aspects using measures like informativeness, naturalness, persuasiveness, or engagingness [6, 5, 4, 29, 35]. However, studies show language generation-based CRS have limited capability to generate new sentences, therefore evaluating linguistic aspects primarily assesses the quality of human-produced utterances, see also [10, 7].

From the overview of predominant evaluation approaches to CRS, it is evident that various quality constructs as a proxy to the effectiveness of a CRS can be practically realized by human perceptions, which can be influenced by various factors such as context, beliefs, emotions, and personal differences [36, 37, 38]. It poses further challenges for researchers to explore factors that assert and predict users' needs accurately. In that regard, the *perceived meaningfulness* of conversations can represent a practical approach to assess the overall effectiveness of a CRS in a human-centered fashion to estimate how human perceptions vary along various user, system, and contextual dimensions.

User, System and Context-related Factors Among user-related aspects, inspired by previous works [39, 40, 12], we consider personal features such as gender, age, education, prior chatbot experience, and domain knowledge about movies that refer to enduring characteristics pertaining to individuals' cognition, emotions, and attitude [37]. For example, in [41], it is revealed that younger users prefer less human-like chatbots, while older users prefer more human-like characteristics. Therefore, we aim to identify in what ways and to what degree users' personal differences have an impact on *perceived meaningfulness*.

When considering system-related aspects, a CRS can be equipped with several attributes [20, 3], for example, enhancing transparency via explanations [42, 43]. Similarly, *informativeness* (amount of domain concepts [6, 29, 35, 34]) is critical for the consistency and relevance of

responses. Many proposals in CRS integrate additional domain knowledge to manifest domain-related concepts in the CRS output and thereby evaluate its effectiveness through offline and human evaluations [6, 29, 35, 34]. In this context, both KBRD and CRB-CRS incorporate additional knowledge and metadata into CRS. Therefore, we rely on the *item ratio* in responses, as was done in [29] to estimate the extent of informativeness. Moreover, *response length* (i.e., token count) can be an important system factor that can affect human perceptions in chatbot interactions. For example, studies have shown that response length can impact engagement, satisfaction, and perceived quality of responses [4, 35]. To this end, CRB-CRS implemented *Response length* as a system design parameter, heuristically curated based on ReDial dataset statistics, see also [19], while KBRD does not explicitly set a parameter for response length.

Finally, we explore context-related factors that relate to the specific context in which a dialog takes place, e.g., different stages of the dialog discourse. Dialog discourse length is an important consideration when assessing the usability of CRS, since certain CRS may excel during the initial phases of dialog, where chit-chat and preference elicitation interactions are expected [19]. Alternatively, other CRS may perform better during the middle stage, where we typically expect item recommendations, see also an overview of user intent and system actions in [14]. In addition, users may expect concise dialog, as lengthy conversations might lead to users being bored and dissatisfied [44].

3. Experiment Design

Like previous studies, e.g., in [5, 6, 9], we compared the quality of responses from both CRB-CRS and KBRD through an *online* user study. The details of the study procedure and participants are as follows.

Procedure A web interface is particularly developed for this study, where after the informed consent and data privacy statements, participants were presented with a dialog situation, starting from the first utterance and *always* ending by the user utterance, see also Figure 1. The dialog situations are automatically randomly curated from a set of 70 dialogs randomly selected from the *ReDial* [9] dataset. Dialog continuations (or responses) to the last user utterance from two different systems generated by KBRD and CRB-CRS were shown in random order. Study participants were tasked to rate the quality of responses *independently* in terms of *meaningfulness* of each response on an absolute 5-point scale ranging from “Entirely meaningless” to “Perfectly meaningful”. Note that when considering perceived meaningfulness, it entails evaluating the quality of recommendations, appropriateness of responses, and coherence within the context of a conversation.

Nonetheless, prior to the response rating task, participants were given instructions and examples of when a response should be meaningful. Precisely, we provided the following instructions to our study participants.

1. A response should be logical by the chatbot, for instance, the chatbot should make a recommendation when the user asks for one.
2. A response should be complete and grammatically correct.

3. *If a recommendation is made by the chatbot, it has to match the user’s stated preferences. If, for example, a user is looking for a funny movie, mostly funny movie recommendations are meaningful. Please make this assessment based on your knowledge, and expectations in the context of mentioned movies in dialog situations. You may always look up services like IMDb to check, e.g., about the movie genres and plots. Note that movie titles are shown in double quotes in both the dialog situations and corresponding responses.*
4. *If the chatbot response is not a movie recommendation, you are supposed to rate the meaningfulness of the response as a reply to the user’s last statement keeping in mind the context of the dialog situation.*

Despite the provided instructions, participants may still have varied perceptions of response meaningfulness. Nonetheless, we believe that allowing for subjective interpretation is more suitable for evaluating user perceptions in natural language interactions, rather than imposing strict evaluation metrics. During the experiment, each participant completed 10 trials, with one (hardcoded) randomly ordered dialog situation serving as an attention check. The CRS responses for this situation were also manually curated, with one required to select a particular rating from the given scale as a criterion to pass the attention check. After the rating task, participants were shown demographic questions to reveal their user-related aspects.

Participants We recruited 107 participants from Amazon Mechanical Turk with qualifications as fluent in English and having an interest in movies. After the experiment, we analyzed the data and found that 9 participants failed the attention check, and 8 participants provided inconsistent ratings. For example, in one of the cases, we found different ratings for similar responses by both systems. Other such cases can also be found in [19]. To avoid cherry-picking, we removed the entire data of such unreliable subjects, leaving us with valid data for 90 participants. In this way, we collected a total of 810 (9 per subject, after removing 1 attention-check trial) valid user ratings for each CRS algorithm. On average, participants took 9 minutes to complete the task, and they were paid 1.5 USD each. The study data is available online.

4. Methodology

In this section, we explain the procedure of curating data for each identified factor (variable) and the methodology for analyzing the research questions.

Dependent Variable. In our experiment, study participants rated their perceptions of meaningfulness as a proxy to the overall quality of responses from two CRS for a total of 810 dialog situations. So overall, we collected 1620 rating scores for both KBRD and CRB-CRS. The rating score for *perceived meaningfulness* was defined as a dependent variable in our analyses.

Independent Variables. We consider *six* different user-related aspects in our research, acquired from users’ self-reported responses. Four of them, i.e., *Gender*, *Age*, *Education*, and *English proficiency* describe the users’ demographic backgrounds. Additionally, inspired by works in [45, 12, 41, 46], two features, *Domain knowledge* and *Prior chatbot experience* were included in our analyses. To study system-related aspects, we consider three variables, *Informativeness*, *Response length*, and *Algorithmic performance differences* between KBRD [6] and CRB-CRS

[19]. Informativeness refers to the extent of domain items included in their system responses. Specifically, we compute the item ratio, as was done in [29], as a proxy for informativeness in the responses for each system. Moreover, *Response length* (i.e., token count of each response) might be an important factor, also introduced in [4, 35].

Regarding contextual aspects, we consider the *Dialog discourse stage*. To investigate this, we split the dialog situations in our study into three groups – Initial, Middle, and End – based on the number of seeker utterances. The dialogs in this study varied in length, ranging from one to nine seeker utterances. On average, each dialog had 2.5 seeker utterances. Dialogs in the Initial group consist of up to two seeker utterances, the Middle group had three to five utterances, and longer dialogs were classified into the End group.

Method. Given the collected data, the goal of this study is to examine the relationship between various independent variables and the dependent variable (i.e., user rating scores for the meaningfulness of CRS responses). Therefore, we used a Mixed Linear Effect Model regression (MLEM) to analyze the data using Python’s statsmodels library¹. Specifically, we used MLEM because (i) individual observations in our experiment are not independent, (ii) the ratings for perceived meaningfulness from the same subject are likely to be more similar to each other given similar dialog situations, and (iii) the model is appropriate for within-subject study designs where each participant is measured multiple times [47, 48]. Our model, therefore, includes a fixed effects term for the CRS method (i.e., KBRD and CRB-CRS) and a random intercept to account for any variability between different participants. We set α to 0.05 for the statistical significance threshold. We apply this model twice aiming to address two research questions.

Specifically, at first, MLEM is applied to identify relationships between the dependent variable and independent variables, i.e., addressing RQ1. Second, our goal was to investigate the interaction effects between the dependent variable and various interaction terms of the independent variables, targeting RQ2.

Before applying the model, we employed the Shapiro-Wilk test [49] for normality checks on the dialogs from which dialog situations were curated. The skewness turned out to be 0.01, which indicates a slightly right-skewed, probably due to the large sample size (810 rating trials), but an approximately symmetric distribution.

5. Analyses and Results

5.1. Descriptive Statistics

The descriptive statistics of the categorical and continuous variables in our study are shown in Table 1 and Table 2, respectively. Overall, we have an almost balanced gender distribution making our findings inclusive. All participants are fluent in English and have varied levels of domain understanding, suggesting they are intellectually and linguistically suitable for our study. Note however that before fitting the models for our further analyses: (i) we do not consider language fluency variable as all subjects are fluent in English, and (ii) discarded entire ratings from a subject stated “Prefer not to say” for its gender disclosure, thus leaving us with

¹<https://pypi.org/project/statsmodels/>

Table 1
Descriptive Statistics of our Study Data (Categorical Variables)

Feature	Scale	Total
Gender	Male	45 (50.00%)
	Female	44 (48.59%)
	Prefer not to say	1 (1.11%)
Age	18-25	6 (6.67%)
	25-30	26 (28.88%)
	30-35	20 (22.22%)
	35-45	21 (23.33%)
	45-70	17 (18.89%)
Education level	High school or less	13 (14.44%)
	Bachelor's	51 (56.57 %)
	Master's	22 (24.44 %)
	Doctorate	2 (2.22%)
	Other	2 (2.22%)
English fluency level	Beginner	0
	Intermediate	0
	Fluent	87 (96.67%)
	Advanced	3 (3.33%)
Movie watching frequency	Everyday	17 (18.89%)
	Several times a week	39 (43.33%)
	Once in a week	21 (23.33%)
	Once every few weeks	9 (10.00%)
	Less frequent	4 (4.44%)
Chatbot experience	Yes	33 (36.67%)
	No	57 (63.33%)
Dialog discourse stage	Initial	387 (35.43%)
	Middle	306 (37.78%)
	End	117 (14.44%)

Table 2
Descriptive Statistics of our Study Data (Continuous Variables)

	Min	Mean	Median	Mode	Max	STD	Variance
Perceived Meaningfulness (CRB-CRS)	1.0	3.78	4.00	5.00	5.00	1.22	1.48
Perceived Meaningfulness (KBRD)	1.0	3.62	4.00	5.00	5.00	1.32	1.75
Domain items (CRB-CRS)	0	0.84	1.00	1.00	4.00	0.77	0.59
Domain items (KBRD)	0	0.52	1.00	1.00	1.00	0.50	0.25
Response length (CRB-CRS)	2.00	6.83	5.00	4.00	47	5.30	28.13
Response length (KBRD)	1.0	6.88	6.00	4.00	34	4.91	24.08

801 rating trials for model fit. Based on the average scores, CRB-CRS with an average score of 3.78 (STD= 1.48) outperformed KBRD having a score of 3.62 (STD=1.75). Also, a deeper analysis of the *response length* shows in user requests asking for a movie plot or explanation, the system response gets longer, thus reflecting high variance.

Table 3

(Model 1: RQ1) Mixed Linear Effect Model (MLEM) regression results for predicting *perceived meaningfulness* from user, system, and context-related factors.

	Perceived Meaningfulness
Gender	-0.170* (0.068)
Age	0.027 (0.026)
Education	-0.059* (0.025)
Domain knowledge	0.062** (0.021)
Chatbot experience	0.094 (0.065)
Informativeness	0.336*** (0.054)
Response length	-0.014* (0.007)
Algorithmic differences	0.000** (0.003)
Dialog discourse stage	0.045 (0.046)
Groups	2
Group Size	801
Model Convergence	Yes
Scale	1.54

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

5.2. Effects of User, System and Context-related Factors (RQ1)

The goal of this analysis is to reveal what kind of different factors related to either user, system, or context influence human perceptions of meaningfulness. By following the above methodology, we report the results of our *Model 1* in Table 3, which indicates that users' *gender*, *education*, and *domain knowledge* have a significant influence on the outcome variable (i.e., perceived meaningfulness). Among system-related aspects, results indicate that the *informativeness* of system responses, i.e., the extent of mentioned domain entities in responses, with a high coefficient score, is a leading predictor of users' perceptions of meaningfulness (coeff: 0.336, $p < 0.001$). Individual algorithmic differences are also significant ($p < 0.01$) indicating that users perceived CRB-CRS responses as meaningful more often than KBRD, however, the mutual difference is minimal (coeff: 0.000). In addition, regarding a contextual feature, i.e., dialog discourse stage is *not* significant ($p > 0.05$) when keeping the remaining variables constant. On the other hand, response length was negatively correlated to meaningfulness with a coefficient of -0.014 ($p < 0.01$), meaning that users tended to rate shorter responses as more meaningful.

The mixed effect model's *Scale* value of 1.54 implies that there is a certain extent of variability in user ratings supporting the validity of our model in light of the mean ratings and standard deviations of both KBRD and CRB-CRS. Additionally, our model converged on the data, indicating its reliability and the ability to successfully capture the underlying relationships between the independent and dependent variables, see also e.g., [50, 51].

5.3. Interaction Effects (RQ2)

Regarding RQ2, inspired by prior research [17, 52, 53], we examine how user-related factors (e.g., age, gender, and chatbot experience) interact with system-related (response length and informativeness) and context-related (dialog discourse stage) factors to shape users' perceptions

Table 4

(Model 2: RQ2) MLEM regression results for estimating the interaction effects between user and system, user and context, and system and context-related factors on *perceived meaningfulness*.

	Perceived Meaningfulness
<i>User and System</i>	
Age	-0.079 (0.073)
Gender	-0.091 (0.181)
Response length	-0.009 (0.027)
Informativeness	0.436** (0.168)
Domain knowledge	0.074 (0.046)
Chatbot experience	0.477** (0.175)
Age X Informativeness	-0.017 (0.042)
Age X Response length	0.000 (0.006)
Gender X Informativeness	0.139 (0.105)
Gender X Response length	0.015 (0.014)
Domain knowledge X Informativeness	0.009 (0.032)
Domain knowledge X Response length	-0.002 (0.005)
Chatbot experience X Informativeness	-0.482*** (0.104)
Chatbot experience X Response length	-0.016 (0.014)
<i>User and Context</i>	
Dialog discourse stage	-0.119 (0.146)
Age X Dialog discourse stage	0.103** (0.038)
Gender X Dialog discourse stage	-0.173 (0.096)
Chatbot experience X Dialog discourse stage	0.052 (0.094)
<i>System and Context</i>	
Informativeness X Dialog discourse stage	0.006 (0.072)
Response length X Dialog discourse stage	-0.002 (0.011)

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

and attitudes toward two CRS. In addition, to understand how two systems interact in a particular dialog context, we explore the effects of system-related and contextual aspects as well.

Table 4 presents the overall results of our MLEM model, which successfully converged on the data (with a Scale value of 1.52). Model 2 revealed significant interaction effects between users' previous chatbot experience and informativeness ($p < 0.001$), as well as between users' age and dialog discourse stage on perceived meaningfulness. We further explain the findings for significant (highlighted in bold) interaction terms below.

Interaction Effects between Chatbot Experience and Informativeness. Model 2 revealed a significant yet negatively correlated two-way interaction between users' prior chatbot experience and the extent of *informativeness* in responses from two CRS algorithms, impacting perceived meaningfulness. As shown in Figure 2, specifically, the increase in informativeness somehow linearly influences users' perceptions of meaningfulness for those *with* prior chatbot experience ($p < 0.001$). However, for users *without* prior chatbot experience, the results did not provide a definitive conclusion on such interaction effects, though it appears participants perceived higher meaningfulness when one-three items were mentioned in the responses.

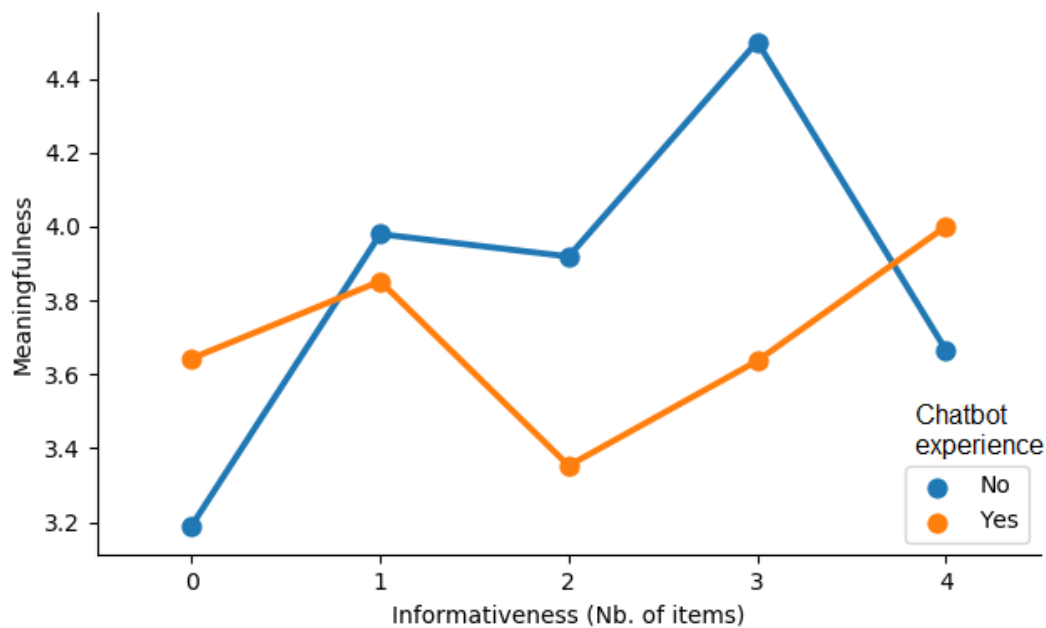


Figure 2: Interaction Effect of Chatbot Experience and Informativeness on Perceived Meaningfulness

Interaction Effects between Age and Dialog Discourse Stage. Figure 3 displays the interaction effects between age and dialog discourse stage on perceived meaningfulness. Shorter dialogs have a positive influence on perceived meaningfulness across all age groups, indicating the good performance of KBRD and CRB-CRS in generating responses for initial chitchat or preference elicitation-based user utterances. Middle-length dialogs show mixed effects based on age group, likely due to subjective assessments of the made item recommendations. Interestingly, longer dialogs have a positive correlation for perceived meaningfulness with younger participants, but a negative correlation with older participants (35 years onward). On the contrary, younger adults (especially 18-25 years old) perceived the system response at the end stage of dialog as quite meaningful. This indicates that old users expect concise but effective responses to appropriately serve the user’s specific queries too, like asking for explanations about made recommendations.

6. Discussion and Findings

In this section, we discuss the implications and limitations of our research.

Findings and Implications Generally, we observe substantial progress in developing various techniques for conversational recommendation tasks. In this research, we attempt to understand the relationships between users’ perceptions of system quality estimated through meaningfulness and related factors. Specifically, from the literature, we investigated various factors (user-related, system-related, and context-related), and their influence on human perceptions of

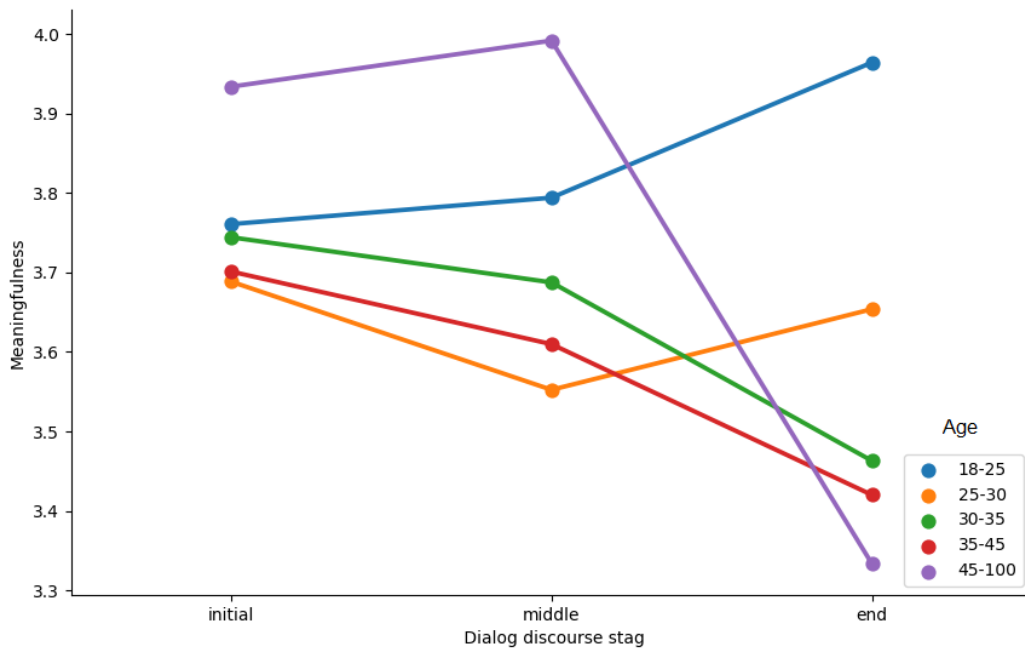


Figure 3: Interaction Effect of Age and Dialog Discourse Stage on Perceived Meaningfulness

meaningfulness. Based on a series of analyses, we highlight the findings of our research and their implications as follows.

Designing CRS for different user segments. The interaction effect between chatbot experience and informativeness suggests that users having prior experience with chatbots have higher expectations. Therefore, when designing (CRS) chatbots, it is important to consider the target user base and its familiarity with technology and provide more advanced features like access to item metadata, reviews proliferation, or explanations for item recommendations, to meet their expectations. This could also include incorporating more sophisticated natural language processing capabilities, personalized recommendations, or advanced system functionalities.

Tailoring dialog length based on user age group. The relationship between the dialog discourse stage and perceived meaningfulness being dependent on the user’s age indicates that different age groups prefer varying lengths of dialogs. Older users tend to prefer shorter dialogs, while younger users find longer dialogs more meaningful. To optimize user engagement and satisfaction, CRS designers should consider adapting the length and state of conversations to align with the preferences of different age segments. This may involve using concise responses for older users and providing explanatory, and interactive conversations for younger users.

Understanding gender differences in response expectations. Our finding, i.e., female users tend to have a higher need for informative responses, aligns with previous research

highlighting their higher expectations and critical evaluation of system responses, see also e.g., [54]. This suggests that CRS developers should pay attention to providing accurate and relevant information to meet their expectations. Incorporating robust knowledge bases, ensuring accurate information retrieval, and emphasizing clarity and completeness in responses can help address the needs of various gender groups to enhance their satisfaction with the system.

Limitations One potential limitation of our study is that we only examined two CRS as representatives of both language generation and retrieval methods. It remains an open question to what extent our findings are generalizable. Also, both analyzed CRS are developed using the same dataset from the movies domain, therefore limiting the generalizability of our findings to other domains and to other datasets e.g., [33, 55, 56]. However, we believe our results can be considered representative of the current state-of-the-art due to their superior performance over other baselines, and both systems were published in highly-ranked scientific venues. Second, the reliability of study participants is a potential threat to validity. However, we applied multiple quality-assurance measures, including participant selection qualifications like interest in movies and English fluency, attention checks, and manual inspection, which increases our confidence in the reliability of our results.

In addition, it is important to note that our study involved N=90 subjects, which might be considered relatively limited for analyzing user aspects. However, we believe that these subjects represent a subset of CRS users in practice. Thus, enabling CRS researchers to explore further directions with more comprehensive studies. Overall, our research sheds light on the diverse dimensions of user expectations and system functionalities supported by a substantial number of 1602 users' ratings. This understanding may further enable us to effectively design CRS that enhances system quality and ultimately lead to improved user satisfaction. In future research, we intend to explore additional factors like dialog initiative strategy and users' personal traits in a high-powered study, potentially influencing human perceptions of quality in conversational recommender systems.

Conclusion

Conversational recommender systems (CRS) that interact with users in natural language assist users in finding relevant items via multi-turn dialogs. To design interactive CRS that fulfills individual needs and expectations, it is vital to understand what factors influence users' perceptions of system quality. To this end, based on our user study specifically designed to obtain users' fine-grained feedback on various parts of the dialog, we investigate which user-related, system-related, and contextual factors are the main predictors of human quality perceptions of CRS. We further delved into investigating the combined effects of such factors to reveal how various terms interact to affect the perceived quality of CRS. We believe our findings will be helpful in tailoring various dimensions of user expectations and system functionalities, thus opening new research directions to understand how we *design* and *evaluate* today's CRS.

References

- [1] A. Manzoor, D. Jannach, Generation-based vs. retrieval-based conversational recommendation: A user-centric comparison, in: *RecSys '21*, 2021, pp. 515–520.
- [2] Y. Zhang, X. Chen, Q. Ai, L. Yang, W. B. Croft, Towards conversational search and recommendation: System ask, user respond, in: *CIKM '18*, 2018, pp. 177–186.
- [3] D. Jannach, Evaluating conversational recommender systems, *Artificial Intelligence Review* forthcoming (2022).
- [4] S. A. Hayati, D. Kang, Q. Zhu, W. Shi, Z. Yu, Inspired: Toward sociable recommendation dialog systems, in: *EMNLP '20*, 2020, pp. 8142–8152.
- [5] K. Zhou, W. X. Zhao, S. Bian, Y. Zhou, J.-R. Wen, J. Yu, Improving conversational recommender systems via knowledge graph based semantic fusion, in: *KDD '20*, 2020, pp. 1006–1014.
- [6] Q. Chen, J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang, J. Tang, Towards knowledge-based recommender dialog system, in: *EMNLP-IJCNLP '19*, 2019, pp. 1803–1813.
- [7] A. Manzoor, D. Jannach, Conversational recommendation based on end-to-end learning: How far are we?, *Computers in Human Behavior Reports* 4 (2021) 100139.
- [8] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, J. Pineau, How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation, in: *EMNLP '16*, 2016, pp. 2122–2132.
- [9] R. Li, S. Ebrahimi Kahou, H. Schulz, V. Michalski, L. Charlin, C. Pal, Towards deep conversational recommendations, *NIPS '18* 31 (2018).
- [10] D. Jannach, A. Manzoor, End-to-end learning for conversational recommendation: A long way to go?, in: *IntRS@ RecSys*, 2020, pp. 72–76.
- [11] A. Manzoor, D. Jannach, INFACT: An online human evaluation framework for conversational recommendation, in: *KARS@ RecSys*, 2022, pp. 72–76.
- [12] B. P. Knijnenburg, N. J. M. Reijmer, M. C. Willemsen, Each to his own: how different users call for different interaction methods in recommender systems, in: *RecSys '11*, 2011.
- [13] S. Berkovsky, R. Taib, D. Conway, How to recommend? user trust factors in movie recommender systems, in: *IUI '17*, 2017, pp. 287–300.
- [14] W. Cai, L. Chen, Predicting user intents and satisfaction with dialogue-based conversational recommendations, in: *UMAP '20*, 2020, pp. 33–42.
- [15] I. Benbasat, W. Wang, Trust in and adoption of online recommendation agents, *JAIS* 6 (2005) 4.
- [16] L. Chen, P. Pu, Trust building in recommender agents, in: *WPRS-IUI '05*, 2005, pp. 135–145.
- [17] W. Cai, Y. Jin, L. Chen, Impacts of personal characteristics on user trust in conversational recommender systems, in: *CHI '22*, 2022, pp. 1–14.
- [18] T. Araujo, N. Helberger, S. Kruikemeier, C. H. De Vreese, In ai we trust? perceptions about automated decision-making by artificial intelligence, *AI & Society* 35 (2020) 611–623.
- [19] A. Manzoor, D. Jannach, Towards retrieval-based conversational recommendation, *Information Systems* 109 (2022) 102083.
- [20] Y. Jin, L. Chen, W. Cai, P. Pu, Key qualities of conversational recommender systems: From users' perspective, in: *HAI '21*, 2021, pp. 93–102.
- [21] D. Jannach, C. Bauer, Escaping the McNamara Fallacy: Towards more impactful recom-

- mender systems research, *AI Magazine* 41 (2020) 79–95.
- [22] L. Yang, J. Hu, M. Qiu, C. Qu, J. Gao, W. B. Croft, X. Liu, Y. Shen, J. Liu, A hybrid retrieval-generation neural conversation model, in: *CIKM '19*, 2019, pp. 1341–1350.
 - [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *NIPS '17* 30 (2017).
 - [24] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, et al., Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia, *Semantic Web* 6 (2015) 167–195.
 - [25] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. Grue Simonsen, J.-Y. Nie, A hierarchical recurrent encoder-decoder for generative context-aware query suggestion, in: *CIKM '15*, 2015, pp. 553–562.
 - [26] M. Ott, S. Edunov, D. Grangier, M. Auli, Scaling neural machine translation, in: *MT '18*, 2018, pp. 1–9.
 - [27] Q. Chen, J. Lin, Y. Zhang, H. Yang, J. Zhou, J. Tang, Towards knowledge-based personalized product description generation in e-commerce, in: *SIGKDD '19*, 2019, pp. 3040–3050.
 - [28] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, C. D. Manning, HotpotQA: A dataset for diverse, explainable multi-hop question answering (2018). [arXiv:1809.09600](https://arxiv.org/abs/1809.09600).
 - [29] J. Zhou, B. Wang, R. He, Y. Hou, CRFR: Improving conversational recommender systems via flexible fragments reasoning on knowledge graphs, in: *EMNLP*, 2021, pp. 4324–4334.
 - [30] Y. Lu, J. Bao, Y. Song, Z. Ma, S. Cui, Y. Wu, X. He, Revcore: Review-augmented conversational recommendation, in: *ACL-IJCNLP '21*, 2021, pp. 1161–1173.
 - [31] K. Chen, S. Sun, Knowledge-based conversational recommender systems enhanced by dialogue policy learning, in: *ICKG '21*, 2021, pp. 10–18.
 - [32] S. Zhang, M.-C. Wang, K. Balog, Analyzing and simulating user utterance reformulation in conversational recommender systems, in: *SIGIR '22*, 2022, pp. 133–143.
 - [33] K. Zhou, Y. Zhou, W. X. Zhao, X. Wang, J.-R. Wen, Towards topic-guided conversational recommender system, in: *ICCL '20*, 2020, pp. 4128–4139.
 - [34] Y. He, L. Liao, Z. Zhang, T.-S. Chua, Towards enriching responses with crowd-sourced knowledge for task-oriented dialogue, in: *MuCAI '21*, 2021, pp. 3–11.
 - [35] T. Zhang, Y. Liu, P. Zhong, C. Zhang, H. Wang, C. Miao, Kecrs: Towards knowledge-enriched conversational recommendation system (2021). [arXiv:2105.08261](https://arxiv.org/abs/2105.08261).
 - [36] D. Jannach, H. Abdollahpouri, A survey on multi-objective recommender systems, *Frontiers in Big Data* 6 (2023).
 - [37] A. Beheshti, S. Yakhchi, S. Mousaeirad, S. M. Ghafari, S. R. Goluguri, M. A. Edrisi, Towards cognitive recommender systems, *Algorithms* 13 (2020) 176.
 - [38] P. Pu, L. Chen, R. Hu, A user-centric evaluation framework for recommender systems, in: *Proceedings of the fifth ACM conference on Recommender systems*, 2011, pp. 157–164.
 - [39] R. Wang, F. M. Harper, H. Zhu, Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences, in: *CHI '20*, 2020, pp. 1–14.
 - [40] N. Wang, L. Chen, User bias in beyond-accuracy measurement of recommendation algorithms, in: *RecSys '21*, 2021, pp. 133–142.
 - [41] A. Følstad, P. B. Brandtzaeg, Users' experiences with chatbots: findings from a question-

- naire study, *Quality and User Experience* 5 (2020) 3.
- [42] N. Sonboli, J. J. Smith, F. Cabral Berenfus, R. Burke, C. Fiesler, Fairness and transparency in recommendation: The users' perspective, in: *UMAP '21*, 2021, pp. 274–279.
 - [43] D. Elsweiler, C. Trattner, M. Harvey, Exploiting food choice biases for healthier recipe recommendation, in: *SIGIR '17*, 2017, pp. 575–584.
 - [44] W. Lei, X. He, M. de Rijke, T.-S. Chua, Conversational recommendation: Formulation, methods, and evaluation, in: *SIGIR '20*, 2020, pp. 2425–2428.
 - [45] M. Nourani, J. King, E. Ragan, The role of domain expertise in user trust and the impact of first impressions with intelligent systems, in: *AAAI '20*, volume 8, 2020, pp. 112–121.
 - [46] P. B. Brandtzaeg, A. Følstad, Chatbots: changing user needs and motivations, *interactions* 25 (2018) 38–43.
 - [47] H. Quan, W. J. Shih, Assessing reproducibility by the within-subject coefficient of variation with random effects models, *Biometrics* (1996) 1195–1203.
 - [48] D. J. Barr, R. Levy, C. Scheepers, H. J. Tily, Random effects structure for confirmatory hypothesis testing: Keep it maximal, *Journal of memory and language* 68 (2013) 255–278.
 - [49] S. Shapiro, M. Wilk, A goodness-of-fit test for normality: The shapiro-wilk test, *Journal of the American Statistical Association* 60 (1965) pp. 619–626.
 - [50] A. Gelman, J. Hill, *Data analysis using regression and multilevel/hierarchical models*, Cambridge University Press, 2006.
 - [51] H. F. Senter, *Applied linear statistical models*. michael h. kutner, christopher j. nachtsheim, john neter, and william li, *Journal of the American Statistical Association* 103 (2008) 880–880.
 - [52] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, C. Newell, Explaining the user experience of recommender systems, *UMUAI* 22 (2012) 441–504.
 - [53] C. M. Myers, A. Furqan, J. Zhu, The impact of user characteristics and preferences on performance with an unfamiliar voice user interface, in: *CHI '19*, 2019, pp. 1–9.
 - [54] P. K. Mo, S. H. Malik, N. S. Coulson, Gender differences in computer-mediated communication: A systematic literature review of online health-related support groups, *Patient Education and Counseling* 75 (2009) 16–24.
 - [55] A. Manzoor, D. Jannach, *Inspired2: An improved dataset for sociable conversational recommendation* (2022).
 - [56] Z. Liu, H. Wang, Z. Niu, H. Wu, W. Che, T. Liu, Towards conversational recommendation over multi-type dialogs, in: *ACL '20*, 2020, pp. 1036–1049.