



Behavior Alignment: A New Perspective of Evaluating LLM-based Conversational Recommendation Systems

Dayu Yang
University of Delaware
Newark, DE, USA
dayu@udel.edu

Fumian Chen
University of Delaware
Newark, DE, USA
fmchen@udel.edu

Hui Fang
University of Delaware
Newark, DE, USA
hfang@udel.edu

ABSTRACT

Large Language Models (LLMs) have demonstrated great potential in Conversational Recommender Systems (CRS). However, the application of LLMs to CRS has exposed a notable discrepancy in behavior between LLM-based CRS and human recommenders: LLMs often appear inflexible and passive, frequently rushing to complete the recommendation task without sufficient inquiry. This behavior discrepancy can lead to decreased accuracy in recommendations and lower user satisfaction. Despite its importance, existing studies in CRS lack a study about how to measure such behavior discrepancy. To fill this gap, we propose *Behavior Alignment*, a new evaluation metric to measure how well the recommendation strategies made by a LLM-based CRS are consistent with human recommenders'. Our experiment results show that the new metric is better aligned with human preferences and can better differentiate *how systems perform* than existing evaluation metrics. As Behavior Alignment requires explicit and costly human annotations on the recommendation strategies, we also propose a classification-based method to implicitly measure the Behavior Alignment based on the responses. The evaluation results confirm the robustness of the method.

CCS CONCEPTS

• **Information systems** → Users and interactive retrieval; • **Human-centered computing** → HCI design and evaluation methods.

KEYWORDS

Recommender Systems, Conversational Systems, Evaluation Metric.

ACM Reference Format:

Dayu Yang, Fumian Chen, and Hui Fang. 2024. Behavior Alignment: A New Perspective of Evaluating LLM-based Conversational Recommendation Systems. 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, 5 pages. <https://doi.org/10.1145/3626772.3657924>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0431-4/24/07

<https://doi.org/10.1145/3626772.3657924>

1 INTRODUCTION

Recent advancements have highlighted the potential of large language models (LLMs) across a range of tasks. Their impressive recommendation performance [16, 8, 22] and language generation capability [20] have attracted the attention of researchers in the conversational recommender system (CRS) community [4, 13, 21, 7]. For example, Google Research team [4] recently proposed an LLM-based CRS that achieved huge success on YouTube. LLMs are inherently well-suited to the needs of CRS. Unlike traditional recommendation systems, which depend on user profiles and past activities, CRS prioritizes the identification of user preferences through real-time interactions [17, 21, 4], striving to mimic the interactions of human recommenders [6, 12]. These characteristics require strong capabilities in language understanding and generation [12, 25], areas where LLMs excel. Consequently, there is an increasing trend of adopting LLMs within the CRS domain.

However, existing studies also found a significant weakness of LLMs when implemented into conversational recommendation [21, 13]: the behavior of LLMs lacks proactiveness and adaptivity throughout dialogues, therefore leads to the shortage of information for understanding the users' preference [13].

To validate these findings, we conducted a comparative analysis involving two popular LLMs and a human recommender across 20 randomly selected datapoints from INSPIRED dataset [6], the only popular conversational recommendation dataset providing behavior labels. This comparison revealed a noticeable difference in the behavior of LLM-based CRS and human recommenders. Specifically, LLM-based CRS systems tend to be passive and inflexible, often rushing to make recommendations without conducting any inquiry. In contrast, human recommenders display much greater patience, dynamism, and adaptability. They show a wider range of complex information-seeking strategies contributing to recommendations, as illustrated in Figure 1. A more tangible illustration is that human recommenders averagely conduct 2.5 conversational turns before making their first recommendation, a number far exceeds those by GPT3.5[14] or Llama2[18], as shown in Table 1. The user information obtained from initial inquiries helps human recommenders make better recommendations and achieve higher success rates.

Recently, many studies on LLMs have focused on improving their behavior alignment with humans [14, 10, 1]. The alignment between the generated sentences from CRS and human recommenders is also desirable for LLM-based CRS. First, better behavior alignment can enhance the user experience by creating a more efficient and proactively interactive conversation. More importantly, the ability to use complex recommendation strategies can help the system receive more user preference information and better understand user profiles, leading to more accurate recommendations [6]. Despite

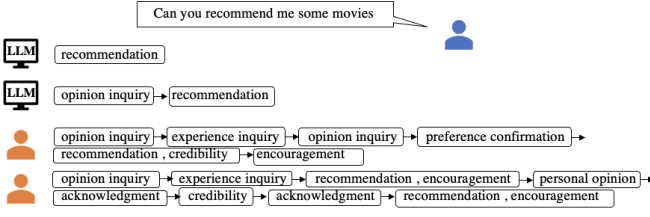


Figure 1: Comparing the strategies that LLM typically uses and human recommenders use for making conversational recommendations.

	#Turns before Rec	Success Rate
LLM-based CRS		
GPT 3.5	1.158	15.8%
Llama 2	1.000	5.3%
Reference		
Human	2.500	57.1%

Table 1: Comparing the behavior of LLM-based CRSs and human’s behavior including: average length of conversational turns before making the first recommendation(#Turns before Rec), average success rate.

the importance of measuring behavior alignment, the evaluation metrics currently used in CRS have no ability to measure this.

To fill this gap, we propose a new evaluation metric: *Behavior Alignment*, a metric that *explicitly* evaluates how well LLM-based CRS’s recommendation strategies are aligned with humans. We also conduct experiments to demonstrate the effectiveness of the measure. In particular, Behavior Alignment has high agreement with the human preferences, and can differentiate the performance of different LLM-based CRS systems much better than the existing metrics. However, one limitation of Behavior Alignment is the requirement of having human annotations of recommendation strategies, which can be costly and time-consuming to obtain. To overcome this limitation, we propose a classification-based method to *implicitly* estimate the behavior alignment without the annotations of recommendation strategies. Experiment results confirm the robustness of the proposed implicit estimation of the Behavior Alignment measure across multiple CRS datasets.

2 RELATED WORK

Existing research practices primarily evaluate CRS’s performance in two aspects: recommendation accuracy, and the quality of sentence generation. In terms of recommendation accuracy, most CRS studies adopt metrics from traditional recommender systems, like hit rates, precision, and recall, from IR community [24, 5].

To evaluate the quality of generated sentences, automatic metrics borrowed from dialogue systems, such as Perplexity, BLEU, and DIST, are used [2, 26, 24, 19]. However, their alignment with *quality* and *user satisfaction* in the context of CRS is questioned [9, 23].

Under such questions, several manual evaluation metrics are proposed. Chen et al. [2] advocated for *Consistency*. Zhou et al. [25] proposed *fluency* and *informativeness* as substitutes for BLEU. However, with the introduction of LLM for building CRS [7, 11, 21, 4], fluency and informativeness have lost their relevance as LLM can generate highly fluent sentences and the informativeness can be

conveniently adjusted by prompting. In addition, manual evaluation’s inherent subjectivity and cost can pose a challenge. Consequently, recent works in LLM-based CRS [7, 11] opted to only evaluate CRS based on accuracy metrics, and overlooked the quality of generated sentences.

3 BEHAVIOR ALIGNMENT: THE METRIC

Users tend to find the conversations more natural and engaging with human recommenders. And those conversations often lead to better recommendations compared with automatic conversational recommendation systems [6]. This is largely because human recommenders usually deploy various sociable and preference elicitation strategies to grab more personal information of users [6]. In addition, we and existing studies [13, 21] consistently observed a substantial disparity between human’s strategies and those employed by LLM-based CRS where latter often appears naive and passive. Therefore, human recommenders’ behaviors could become valuable references for the LLM-based CRS.

3.1 Definition

Based on the assumption that human recommenders’ behaviors can perform as good references for CRS, given a context (e.g., user’s past interactions with a CRS), a CRS generate a response C . And a human recommender writes a response H for the same context. Assuming we know the recommendation strategy of C is R_C and the recommendation strategy of H is R_H , the behavior alignment score of the pair (i.e., $BA(C, H)$) can be computed as in Equation (1). In a test collection with N generated responses, the system’s Behavior Alignment score can be calculated using Equation (2).

$$BA(C, H) = \begin{cases} 1 & \text{if } R_C = R_H \\ 0 & \text{if } R_C \neq R_H \end{cases} \quad (1)$$

$$Behavior\ Alignment = \frac{\sum_{k=2..N} BA(C_k, H_k)}{N} \quad (2)$$

In Equation (2), the first interactive turn is not counted into Behavior Alignment since the start point of every conversation can be random in practice. After the first turn is finished, the behaviors of following turns should be conditioned on the existing context.

For recommendation strategy R , existing CRS studies [6] have provided a comprehensive categorization that includes 13 mutually exclusive types. These types are: *acknowledgment*, *redibility*, *encouragement*, *experience inquiry*, *offer help*, *opinion inquiry*, *personal experience*, *personal opinion*, *preference confirmation*, *rephrase preference*, *self modeling*, *similarity*, and *transparency*.

4 BEHAVIOR ALIGNMENT: EFFECTIVENESS AS AN EVALUATION METRIC

Behavior Alignment can measure the behavior differences between generated sentences from CRS and human recommenders, but it remains unclear whether such a metric is more desirable than existing metrics. To answer this question, we design two sets of experiments to evaluate (1) *whether Behavior Alignment reflects user preferences*; and (2) *whether it can differentiate well-performed systems from those poorly performed ones*.

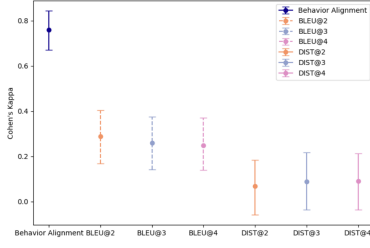


Figure 2: Agreement with human judgments over "Is the response from Falcon-based CRS better than Llama2-based CRS ?", measured by Cohen's Kappa.

4.1 Agreement with Human Preferences

The first set of experiments aims to assess whether the evaluation results of Behavior Alignment are closely consistent with human preferences. Specifically, responses from two CRS systems are evaluated by Behavior Alignment and humans. A strong agreement between them would reflect the effectiveness of a metric.

We randomly sampled 1,000 instances from INSPIRED dataset. The two CRS systems are built on two LLMs respectively: Falcon-7B [15] and Llama2-7B [18]. They were specifically chosen due to the distinct training data they utilize, which leads to varying response styles in reaction to dialog histories. To gauge human preferences, we enlisted two annotators who have experience in conversational recommendations. They assumed the role of CRS users and annotated preferences on the responses generated by the two LLM-based CRS systems (i.e., same, A is better than B, or B is better than A). In addition to preferences between the two systems, we ask an expert trained with linguistic backgrounds to annotate the recommendation strategies for each generated response, which is necessary for computing Behavior Alignment.

An ideal evaluation metric should yield ratings that highly correlate to human preferences [3]. We use Cohen's Kappa to measure such correlation. In addition to Cohen's Kappa, we employed the Bootstrap method to establish the 2.5% to 97.5% confidence interval.

To compare how various metrics reveal user preferences, we conduct a comparative analysis with two popular automatic metrics that existing CRS studies used to evaluate *generation quality* [2, 25, 24]: BLEU@K and DIST@K. The results are depicted in Figure 2.

We can see Behavior Alignment demonstrates a considerable level of agreement with human preference, as evidenced by a Cohen's Kappa of 0.74, a value typically interpreted as indicating "Substantial Agreement". On the contrary, DIST@K and BLEU@K exhibit a much lower level of agreement with human preferences. One potential reason for the poor alignment is they are word-level metrics that are directly adopted from Machine Translation and Summarization tasks, which do not fit the context and goal of the conversational recommendation task very well [3].

4.2 Differentiation of System Performance

A desirable evaluation metric should be able to differentiate various systems [3]. Also, the increase/decrease of rating should be consistent with the increase/decrease of human preference for different systems. However, it is hard to quantify human preference for each LLM-based CRS and it is almost impossible to find large numbers

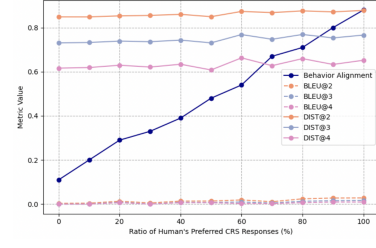


Figure 3: Metric score changes (y-axis) along with human preference changes (x-axis)

of LLMs covering different levels of human preferences to meet the necessity for conducting evaluations.

To address these challenges, we build synthetic LLM-based CRS systems with different levels of user preferences by mixing different ratios of "good generations" and "bad generations" based on the annotated sentences generated from two LLM-based CRS we implemented. Specifically, 100 generation pairs were randomly chosen from annotated data, with one system's response identified as superior to the other. We selected the responses preferred by humans to create a hypothetical "ideal" system, while the responses that were less preferred were selected to build a "worst-case" system. Following this, we crafted synthetic CRS systems of varying quality by blending 90%, 80%, ..., 10% of the "ideal" system with 10%, 20%, ..., 90% of the "worst-case" system. After the synthesis, we compute each metric including Behavior Alignment to see if they can effectively differentiate between different synthetic systems.

As displayed in Figure 3, the x-axis signifies the proportion of samples taken from the "ideal" system, which can be interpreted as a "human preference score". The results show that there is a consistent increase in Behavior Alignment which ranges from 0.11 to 0.88 as we shift from a "worst-case" system to an "ideal" system. This rise is directly proportional to each incremental inclusion of human-preferred samples. However, the other two metrics, BLEU@K and DIST@K, exhibited minimal variability despite the enhancement in human-preferred samples.

5 ESTIMATION OF IMPLICIT BEHAVIOR ALIGNMENT

Although the proposed Behavior Alignment is effective, it requires real-time annotations of behavior types, which is not readily available. To address this issue, we propose a classification-based method to estimate the implicit behavior alignment. Instead of directly predicting the behavior categories and then comparing whether the behaviors of two responses are the same, the method focuses on a binary problem: predicting whether the two responses belong to the same category, which would not require the explicit behavior labels anymore¹.

5.1 Methodology

Given a human-generated response H and a CRS-generated response C , we propose estimating the alignment through a binary classifier, which takes H and C as input and predicts whether they

¹The source code and the experiments can be found at <https://github.com/dayuyang1999/Behavior-Alignment>

Behavior Type	Accuracy	1st Misclassification	2nd Misclassification
personal experience	0.60	credibility	similarity
rephrase preference	0.45	preference confirmation	personal opinion
self modeling	0.31	personal experience	similarity
similarity	0.53	acknowledgment	self modeling
transparency	0.65	opinion inquiry	offer help

Table 2: Top 1st and 2nd misclassification for those Behavior classes having accuracy lower than 0.7

	Accuracy on hold-out test set	
	Original	Mixed-hard
Fold0	0.960	0.976
Fold1	0.958	0.978
Fold2	0.958	0.977
Fold3	0.940	0.976
Fold4	0.967	0.973
Averaged	0.957	0.976

Table 3: Cross-validation results

correspond to the same category or not. The binary classifier is built based on BERT² and fine-tuned on 100,000 human response pairs selected from INSPIRED dataset. How to select the pairs? We explore two strategies to construct the training/testing data.

The first training strategy is straightforward. We created 50,000 negative sentence pairs with different behavior types and another 50,000 positive pairs with the same behavior type. For each sentence pair, they are concatenated by a special token [SEP], which allows BERT to identify the segment of two sentences. This strategy is referred to as "Original".

The second strategy is to intentionally include more difficult examples in the training data to improve the robustness of the classifier. To achieve this, we first train a multi-class classifier to predict the exact behavior type of each sentence. The prediction results are compared with the ground truth labels in order to compute the misclassification rate. Specifically, we define "hard negatives" as negative pairs that a sentence having a prediction accuracy of less than 0.7 and another sentence is from its most misclassified category as shown in Table 2. Consequently, we created 10,000 hard negative samples across five categories. We then randomly replaced 10,000 negative examples in the original training data with hard negatives to build the "mixed-hard" training set.

5.2 Evaluation

The robustness of a binary classifier is important, as it directly influences the efficacy and applicability of the proposed metric. To assess the robustness, we employ three methodologies: (1) implementing cross-validation to ensure consistently high accuracy; (2) determining its Cohen's Kappa in relation to human annotations to verify an invariably strong agreement; and (3) evaluating its performance on out-of-distribution data to ascertain minimal performance degradation.

First, we employ cross-validation on the binary sentence pairs. As observed in Table 3, the accuracy remains consistently high regardless of which strategy is used for constructing training data.

Second, we utilized Cohen's Kappa to quantify the agreement between the classifier's predictions and human annotations. As illustrated in Table 4, our results depict that Cohen's Kappa exceeds

	Accuracy	Cohen's kappa
Original	0.957	0.913
Mixed-hard	0.976	0.952

Table 4: The agreement between predictions and human annotations on testing samples

	Accuracy	Cohen's kappa
Original	0.782	0.563
Mixed-hard	0.932	0.865

Table 5: The agreement between predictions and human annotations on out-of-distribution instances from ReDial dataset

0.9 on the testing set, denoting an almost impeccable alignment between the classifier's predictions and human annotations.

Lastly, to verify the generalizability of the binary classifier, we selected another widely-used CRS dataset that the model has not seen during training, ReDial [12] for testing. Compared with INSPIRED dataset, ReDial dataset emphasizes more on making recommendations during chi-chat, rendering its linguistic style more colloquial. The out-of-distribution samples that have differences on using of language may introduce challenges to the classifier. Given the absence of behavior annotation in the ReDial dataset, we first instructed two annotators to annotate ReDial in a manner similar to INSPIRED, resulting in 42006 sentence pairs with ground-truth binary labels. As Table 5 showcases, the binary classifier trained on "Original" sentence pairs witnessed a marked decline in performance. Conversely, the classifier trained on sentence pairs augmented with hard negatives demonstrated superior generalizability, achieving over 93% accuracy and a Cohen's kappa of 0.86 with humans. This generally indicates a near-perfect agreement between the automatic annotator and human recommenders. The findings emphasize the importance of hard negatives, particularly for a dataset-agnostic evaluation metric.

6 CONCLUSION

Evaluating the generated sentences is crucial for CRS as they significantly influence user responses. This, in turn, impacts both information exposure and user profiling. In this paper, we introduce a novel evaluation metric designed to assess the quality of responses generated by LLM-based CRSs. Our experiments demonstrate this metric's effectiveness, particularly in its alignment with human preferences and its capacity to distinguish LLM-based CRSs by their ability to implement complex recommendation strategies.

7 ACKNOWLEDGEMENTS

This research is supported by the graduate fellowship from the Institute for Financial Services Analytics at the University of Delaware. We would also like to express our sincere gratitude to the reviewers for their insightful comments and suggestions.

²<https://huggingface.co/bert-large-uncased>

REFERENCES

- [1] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Tallrec: an effective and efficient tuning framework to align large language model with recommendation. *arXiv preprint arXiv:2305.00447*.
- [2] Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. *arXiv preprint arXiv:1908.05391*.
- [3] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54, 755–810.
- [4] Luke Friedman et al. 2023. Leveraging large language models in conversational recommender systems. *arXiv preprint arXiv:2305.07961*.
- [5] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: a survey. *AI open*, 2, 100–126.
- [6] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. Inspired: toward sociable recommendation dialog systems. *arXiv preprint arXiv:2009.14306*.
- [7] Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. Large language models as zero-shot conversational recommenders. *arXiv preprint arXiv:2308.10053*.
- [8] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2023. Large language models are zero-shot rankers for recommender systems. *arXiv preprint arXiv:2305.08845*.
- [9] Dietmar Jannach. 2023. Evaluating conversational recommender systems: a landscape of research. *Artificial Intelligence Review*, 56, 3, 2365–2400.
- [10] Andreas Köpf et al. 2023. Openassistant conversations—democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.
- [11] Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. 2023. Large language models for generative recommendation: a survey and visionary discussions. *arXiv preprint arXiv:2309.01157*.
- [12] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. *Advances in neural information processing systems*, 31.
- [13] Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3452–3455.
- [14] Long Ouyang et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- [15] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.
- [16] Damien Sileo, Wout Vossen, and Robbe Raymaekers. 2022. Zero-shot recommendation as language modeling. In *European Conference on Information Retrieval*. Springer, 223–230.
- [17] Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *The 41st international acm sigir conference on research & development in information retrieval*, 235–244.
- [18] Hugo Touvron et al. 2023. Llama 2: open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- [19] Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1929–1937.
- [20] Jason Wei et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- [21] Gangyi Zhang. 2023. User-centric conversational recommendation: adapting the need of user with large language models. In *Proceedings of the 17th ACM Conference on Recommender Systems*, 1349–1354.
- [22] Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023. Recommendation as instruction following: a large language model empowered recommendation approach. *arXiv preprint arXiv:2305.07001*.
- [23] Shuo Zhang and Krisztian Balog. 2020. Evaluating conversational recommender systems via user simulation. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining*, 1512–1520.
- [24] Kun Zhou, Xiaolei Wang, Yuanhang Zhou, Chenzhan Shang, Yuan Cheng, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2021. Crslab: an open-source toolkit for building conversational recommender system. *arXiv preprint arXiv:2101.00939*.
- [25] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving conversational recommender systems via knowledge graph based semantic fusion. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 1006–1014.
- [26] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards topic-guided conversational recommender system. *arXiv preprint arXiv:2010.04125*.