Should We Tailor the Talk? Understanding the Impact of Conversational Styles on Preference Elicitation in Conversational Recommender Systems

Ivica Kostric
Dept. of Electrical Eng. and Comp. Science
University of Stavanger
Stavanger, Norway
ivica.kostric@uis.no

Krisztian Balog
Dept. of Electrical Eng. and Comp. Science
University of Stavanger & Google Research
Stavanger, Norway
krisztian.balog@uis.no

Ujwal Gadiraju
Web Information Systems, Software Technology Department
Delft University of Technology
Delft, Netherlands
u.k.gadiraju@tudelft.nl

Abstract

Conversational recommender systems (CRSs) provide users with an interactive means to express preferences and receive real-time personalized recommendations. The success of these systems is heavily influenced by the preference elicitation process. While existing research mainly focuses on what questions to ask during preference elicitation, there is a notable gap in understanding what role broader interaction patterns—including tone, pacing, and level of proactiveness—play in supporting users in completing a given task. This study investigates the impact of different conversational styles on preference elicitation, task performance, and user satisfaction with CRSs. We conducted a controlled experiment in the context of scientific literature recommendation, contrasting two distinct conversational styles—*high involvement* (fast-paced, direct, and proactive with frequent prompts) and *high considerateness* (polite and accommodating, prioritizing clarity and user comfort)—alongside a flexible experimental condition where users could switch between the two. Our results indicate that adapting conversational strategies based on user expertise and allowing flexibility between styles can enhance both user satisfaction and the effectiveness of recommendations in CRSs. Overall, our findings hold important implications for the design of future CRSs.

CCS Concepts

• **Information systems** → **Users and interactive retrieval**; **Recommender systems**; • **Human-centered computing** → **User studies**; *Natural language interfaces.*

Keywords

Conversational AI; Conversational recommender systems; Conversational styles

1 Introduction

Conversational recommender systems (CRSs) are interactive platforms that engage users in a dialogue, enabling the expression of preferences and feedback in real-time [17]. This dynamic interaction not only distinguishes CRSs from traditional recommender systems but also underlies their core advantage: the ability to provide personalized, context-aware recommendations that adapt to the user's current needs and preferences. Consequently, the success of CRSs largely depends on the effectiveness of preference elicitation, as it directly influences the relevance and accuracy of the recommendations [32].

Existing approaches in CRSs emphasize the significance of question-based user preference elicitation, focusing on what to ask users, and how to adapt the recommendations based on their responses [11]. The goal here is to understand and utilize user preferences to improve recommendation accuracy. However, there is still a limited understanding of how to ask preference elicitation questions within the broader context of a conversation. The interaction between users and CRSs plays a critical role in how effectively the system can support users in completing their tasks, yet there is a notable lack of research focusing on the design choices and conversational styles that lead to successful interactions [18, 30]. Recent studies have demonstrated that different conversational styles can significantly influence user engagement, task performance, and perceived cognitive load in online work environments [34, 35]. These findings suggest that conversational styles can indeed have a positive impact on interaction outcomes, underscoring the potential for similar benefits in CRSs.

The main objective of our work in this paper is to investigate how conversational styles influence the preference elicitation process within CRSs, and in turn, affect overall task performance and user satisfaction. We examine the impact of conversational styles through three specific research questions (RQs) within the context of a given CRS:

- **RQ1** What is the impact of various conversational styles on preference elicitation?
- **RQ2** What is the impact of various conversational styles on task performance?
- **RQ3** What is the impact of various conversational styles on user satisfaction?

To address these research questions, we conducted a controlled experiment within the context of scientific literature recommendations, where study subjects were tasked with finding relevant literature given a topic description. This user study involves an interactive agent designed with distinct conversational styles across three experimental conditions: (1) *high involvement*, where the system takes initiative, makes frequent suggestions, and utilizes fast interaction by use of buttons; (2) *high considerateness*, a relatively more passive interaction approach, supporting users in their decisions but not taking initiative; and (3) a *flexible setting* that grants users the agency to switch between styles at will. Each condition is analyzed to assess its impact on the preference elicitation stage, task performance, and user satisfaction in a CRS.

We recruited 30 master's and PhD students in computer science as participants for our study. They were randomly assigned a set of three tasks from one of two domains: *Computer Science* (expected high familiarity) and *Physics* (expected low familiarity). This division allowed us to examine how domain familiarity influences the effectiveness of different conversational styles in a CRS.

Our main findings indicate that participants with higher domain familiarity prefer the high involvement style for its efficiency and directness, enabling them to express their preferences quickly without what could be perceived as 'unnecessary explanations.' In contrast, participants with less familiarity favored the high considerateness style due to its explanatory nature, which helped them understand the choices and make informed decisions. Objective measures of task performance confirm the benefits of the high considerateness style for enhancing the quality of user selections, particularly among participants with lower domain familiarity. Moreover, allowing users to switch between conversational styles enables them to optimize their interaction based on personal preferences or situational needs, further improving task performance.

The main contributions of this work are as follows: (1) We design and conduct a comprehensive user study to empirically investigate how different conversational styles impact preference elicitation, task performance, and user satisfaction in CRSs (Section 3). (2) We develop a novel conversational scholarly assistant, specifically designed for this study, which are open-sourced to enable further research and development in the field of conversational recommender systems (Section 4). (3) We perform a quantitative analysis of the collected data, showing that different conversational styles are more effective depending on the level of user familiarity with the task domain (Section 5). The code and study materials developed within this paper are made publicly available at https://github.com/iai-group/umap2025-convstyles.

2 Related Work

We are inspired by, build on, and position our work and contributions in the following realms of existing literature: conversational recommender systems (Section 2.1), preference elicitation in CRSs (Section 2.2), and conversational styles (Section 2.3).

2.1 Conversational Recommender Systems

Early recommender systems, such as collaborative filtering [36], logistic regression [31], and gradient boosting decision trees [5], rely heavily on historical data to predict user preferences. While effective, these approaches struggle in cold-start scenarios where user data is sparse and fail to adapt to evolving preferences in real time. Additionally, they treat recommendation as a one-shot process, making them less suitable for high-involvement products (i.e., products bought rarely but chosen with more care and time) that require deliberation [17, 26].

A *conversational recommender system* (CRS) engages users in multi-turn conversations to elicit detailed, real-time preferences using natural language [17]. Beyond capturing explicit preferences, CRS can process feedback and explain recommendations. Gupta et al. [13] found that a CRS was more effective in building user trust and satisfaction in an online housing recommendation system when compared to a conventional web interface. Gao et al. [11] identified five primary challenges in CRS research: effective preference elicitation, multi-turn strategy optimization, improving natural language understanding and generation, balancing exploration-exploitation trade-offs, and robust evaluation including user simulation. This work shifts the focus of question-based preference elicitation from *what* to ask to *how* to ask it. Specifically, we investigate how conversational style influences preference elicitation, task completion, and user satisfaction. We conduct this study in the scholarly domain, which serves diverse information needs, involves non-sensitive data, and engages users with domain expertise [2].

2.2 Preference Elicitation

Preference elicitation is a core challenge in CRSs, involving two key questions: *What to ask?* and *How to adjust the recommendations based on user responses?* [11]. There are two primary approaches: item-based and attribute-based elicitation. Item-based methods directly query users about specific items, updating recommendation lists based on feedback [7, 37, 49]. This shifts recommendation models from static historical data to more dynamic, interactive systems. Attribute-based methods focus on user preferences for item attributes and include critiquing-based techniques, where users refine recommendations by providing feedback on specific attributes [29, 47]. Reinforcement learning-based approaches further optimize attribute selection by adapting to user preferences over multiple interactions [27, 40]. Kostric et al. [24] argued that users with lower domain expertise may struggle to answer attribute-based and proposed usage-oriented questions as an alternative. Our work employs an attribute-based approach but focuses on *how* questions are asked rather than *what* is asked. We investigate how conversational style influences user responses and recommendation effectiveness, potentially enabling more personalized and efficient interactions. Prior work has also explored guiding user preference expression through nudging techniques, where systems directly suggest or emphasize specific areas for further inquiry [12]. Such techniques include repetition or speech modifications to steer users toward particular content. In contrast, our study does not aim to

guide users but instead observes the effects of conversational style on preference elicitation.

2.3 Conversational Styles

Conversational styles play a crucial role in shaping lexical communication. Early work by Lakoff [25] categorized them into four types based on relational dynamics between participants: (1) *clarity*, an ideal mode of discourse; (2) *distance*, a style that does not impose others; (3) *deference*, a style giving options; and (4) *camaraderie*, direct expression of desires. Based on this, Tannen [42, 43] analyzed social discourse and identified two broad styles: *high involvement*, characterized by active engagement and assertiveness, and *high considerateness*, which prioritizes politeness and listener comfort.

In a relevant study, Kim et al. [22] compared survey response data quality acquired from a web platform in comparison to a chatbot having either a *casual* or a *formal* conversational style. The chatbot using a *casual* conversational style attempted to establish a relationship with users using linguistic features that overlap with both of Tannen's *high-involvement* and *high-considerateness* conversational styles. The chatbot using *formal* style is akin to the *clarity style* (i.e., showing no involvement and the least relationship with the user) as summarized by Lakoff [25]. Additionally, Shamekhi et al. [38] found that users prefer agents mirroring their own style, a finding reinforced by studies of information-seeking conversations Thomas et al. [44] using the MISC dataset [45].

While prior CRS work has explored human-like dialogue generation [28], our study systematically examines the impact of conversational style on preference elicitation and user experience. We operationalize Tannen's *high-involvement* and *high-considerateness* styles to investigate their effects on CRS interactions.

Traditionally, response styles in dialogue systems were implemented via templates [35]. Recently, a new approach has emerged where large language models (LLMs) are used to simulate personas [1, 39, 48], where personas are loosely defined fictional characters with traits such as tone, backstory, and personality [41]. Our work differs by controlling not just tone but also interaction dynamics, including pacing and engagement level. Exploring the longitudinal impact of conversational styles, particularly on user trust and reliance [15, 16, 46], and the use of anthropomorphized avatars [20] on preference elicitation in CRSs are important avenues for future work.

3 Study Design

In this section, we detail the experimental design for our user study, where our main objective is to explore how conversational styles affect the preference elicitation process in CRSs and shape task performance and user satisfaction. Specifically, we contrast two conversational styles, which are detailed in Section 3.1: *high involvement*, a fast-paced, direct, and proactive style, where the system takes the initiative and frequently offers suggestions, and *high considerateness*, a polite and accommodating style that prioritizes clarity and alignment with users, allowing them to guide the interaction while the system provides support and additional information as needed. The task given to the participants is described in Section 3.2. The evaluation, where we focus on both objective and subjective

Table 1: A comparison between high-involvement and high-considerateness conversational styles, adopted from [35].

Criteria	High involvement	High considerateness
C1. Rate of speech	fast	slow
C2. Turn-taking	fast	slow
C3. Introduction of topics	w/o hesitation	w/ hesitation
C4. Use of syntax	simple	complex
C5. Directness of content	direct	indirect
C6. Utterance of questions	frequent	rare

measures, is described in Section 3.3. The experiments are carried out on a prototype CSA system, which is detailed in Section 4.

3.1 Conversational Styles: High Involvement and High Considerateness

In our study, we operationalize two distinct conversational styles for the CRS: *high involvement* (`HighInv`) and *high considerateness* (`HighCon`), drawing on the framework developed by Tannen [43]. These conversational styles, originally observed in human-human interactions, reflect varying social and cultural norms that shape communication patterns. They have been adapted to human-agent interactions, with specific criteria guiding conversational agent design to emulate these styles [35].

Table 1 summarizes the key differences between the `HighInv` and `HighCon` styles across six criteria (C1–C6). The criteria include *rate of speech* (C1), *turn-taking* (C2), *introduction of topics* (C3), *use of syntax* (C4), *directness of content* (C5), and *utterance of questions* (C6). In the `HighInv` style, the system communicates with a fast rate of speech and rapid turn-taking, introduces topics without hesitation, uses simple syntax, is direct in content delivery, and asks questions frequently. Conversely, the `HighCon` style features a slower rate of speech and turn-taking, introduces topics with hesitation, employs complex syntax, communicates indirectly, and rarely asks questions. Figure 1 shows excerpts from conversations contrasting the two styles where the user expresses the same information need.

To operationalize the criteria in our CRS, we group the criteria into two categories: *Pacing and Tone* (C1, C2, C4, C5) and *Content Introduction and Questioning* (C3, C6).

3.1.1 **Pacing and Tone**. Pacing refers to how quickly the conversation progresses, influenced by the *rate of speech* (C1) and *turn-taking* (C2). In the `HighInv` style, we simulate a fast rate of speech by displaying messages word by word with minimal delays between words (ranging from 5 to 50 milliseconds between each word), creating a sense of urgency. Shorter texts are displayed with slightly longer delays to simulate user typing speeds, preventing messages from appearing too abruptly. For longer texts, the delay is reduced to avoid making users wait too long. Turn-taking is rapid, with the system responding promptly after the user's input, maintaining a dynamic interaction flow. Additionally, responses in this style are separated into several shorter messages, which contributes further to the sense of urgency. In contrast, the `HighCon` style features a slower rate of speech, with messages displayed
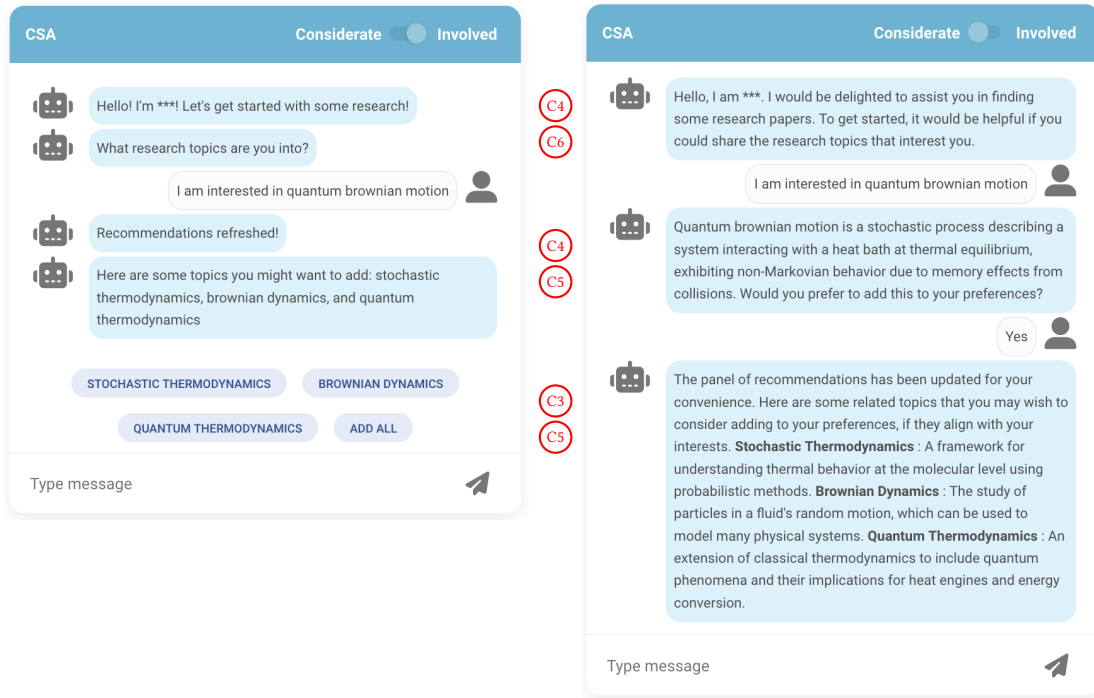
Figure 1: Contrasting example conversations between two conversational styles for the same user need. `HighInv` style on the left and `HighCon` style on the right.

character by character in a single long message and with longer delays (7 to 70 milliseconds between each character), fostering a more relaxed interaction. This is in stark contrast to the `HighInv` style, where messages are presented word by word. The delay difference is highly discernible: the word-based delay in the `HighInv` style contributes to a sense of urgency and momentum. In contrast, the character-based delay in the `HighCon` style slows the interaction down, encouraging users to process information carefully before responding. Turn-taking is deliberate, with pauses before the system responds, allowing users more time to process information. This delay is built-in to cater for the time taken to generate detailed explanations.

Tone encompasses the *use of syntax* (C4) and *directness of content* (C5), affecting the system's manner of communication. The `HighInv` style uses simple syntax and direct language, conveying information concisely. This approach creates an assertive and efficient tone. The `HighCon` style employs more complex syntax and indirect language, often using polite expressions and detailed explanations. This results in a considerate and supportive tone.

*3.1.2 **Content Introduction and Questioning**. Content introduction* (C3) pertains to how the system introduces new topics or suggestions. In the `HighInv` style, the system introduces topics without hesitation, frequently suggesting new keywords or preferences. It actively guides the conversation by offering options for the user to consider, using buttons to encourage users to act. In the `HighCon` style, the system introduces topics with hesitation, providing explanations before making suggestions. It waits for the

user's cues by asking for confirmation before adding a preference, allowing the user to steer the conversation.

Utterance of questions (C6) relates to how often the system asks questions. The `HighInv` style asks questions frequently, aiming to gather information rapidly and keep the conversation moving forward. These questions are often direct and require immediate responses. The `HighCon` style asks questions rarely, and when it does, they tend to be open-ended or for confirmation to make sure there is alignment between the user and the agent. This approach reduces pressure on the user and encourages reflective responses.

3.2 Task Design

Figure 2 illustrates the user journey in our study, which consists of three experimental conditions with varying conversational styles of the CRS: `HighInv` (high involvement), `HighCon` (high considerateness), and `Agency` (providing users with the agency to select and manipulate the conversational style of the CRS between `HighInv` and `HighCon`). We employ a mixed experimental design, with conversational style as a within-subjects factor and the topic set as a between-subjects factor. Specifically, we recruit master's and Ph.D. students in computer science as participants to engage with each condition. We ask participants to use a scholarly conversational assistant to find research papers relevant to a given research topic.[1] To investigate the influence of prior knowledge, we divide participants into two groups. One group is presented with research topics from their primary domain (computer science). This allows us to

---

[1]In our study, a research topic is a specific subject or study area within a broader field.
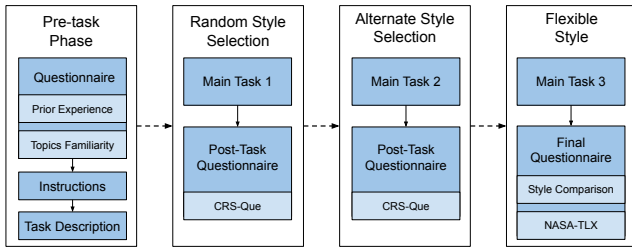
Figure 2: The user journey through the study phases: pre-task familiarization, main task interaction with the conversational system (counterbalanced with respect to which style is first encountered by participants), and post-task feedback. Participants completed three tasks in a randomized order with varying conversational styles, followed by feedback after each task.

assess how prior knowledge may affect their interaction with the CRS and task performance. In contrast, the second group engages with topics outside their primary domain of expertise (physics), ensuring their reliance on the CRS is not confounded by pre-existing knowledge. This approach allows us to isolate the CRS's performance from the participant's familiarity with the subject matter. To minimize order effects, the topic order is randomized for each participant. The topics used in our study were carefully chosen based on the expected familiarity of our participants and the availability of relevant academic papers. The computer science topics were fictional and designed by the authors to reflect the knowledge students are expected to have in our university's applied data science curriculum at the time of the study. This design choice ensured the tasks were engaging and appropriately challenging for our participants. In contrast, the physics topics were actual master's thesis topics provided to us by the Department of Physics at the University of Stavanger. To support the CRS functionality, we utilized papers from the arXiv[2] database. ArXiv was chosen for its extensive and up-to-date coverage of both Computer Science and Physics domains, making it an ideal source for scholarly literature recommendations. The study is split into three phases: the pre-task phase, where participants familiarize themselves with the system; the main task phase, where they interact with the CRS to complete their assignment; and the post-task phase, where they provide feedback on their experience.

In the pre-task phase, participants begin by completing a questionnaire designed to capture their familiarity with conversational assistants and assess their prior knowledge of the research topics included in the study. The questionnaire includes questions on participants' experience with conversational systems, such as how frequently they use these systems, their satisfaction with past interactions, and specific likes or dislikes. Additionally, they are asked to self-report their familiarity with two domains: Machine Learning and Neural Networks, as well as Theoretical Physics and Quantum Mechanics. This helps us gauge the participants' perceived domain expertise before they engage with the system. Following the questionnaire, participants are provided with detailed instructions

about the study, including screenshots of the system's interface that highlight key functionalities. No formal training or hands-on exploration of the system is provided, as the design prioritizes an intuitive, conversational interaction experience.

During the main task, each participant engages with all three experimental conditions (`HighInv`, `HighCon`, `Agency`) in a counterbalanced sequence. Simple randomization is employed for the first two conditions (`HighInv` and `HighCon`) to control for order and learning effects, minimizing potential biases. This setup ensures an even distribution, with half of the participants beginning with the `HighInv` condition and the other half starting with the `HighCon` condition. The `Agency` condition is consistently positioned as the final condition for all participants. This design choice was made to increase the ecological validity of users' agency in selecting their preferred conversational style, based on an informed understanding and expectation obtained after exposure to both styles. We recruited 30 participants, all master's or PhD students in the field of Computer Science, prioritizing academic expertise and domain knowledge as the primary inclusion criteria. General demographic information, such as age or gender, was not explicitly collected.

Participants in the study are asked to assume the role of a student, tasked with compiling a set of 5 research papers relevant to a given topic,[3] thereby simulating the process of gathering related work for a master thesis. The requirement to select 5 papers ensures sustained engagement with the conversational system and provides a clear objective. While interacting with the agent, participants bookmark relevant articles in real time to indicate their choices. This serves a dual purpose: it provides an end goal that participants work towards and allows us to gather data to assess their performance across the experimental conditions. While there is no algorithmic difference in paper recommendations between conditions, the interaction patterns and participant keyword choices influence the outcomes.

After completing each task, participants are asked to complete the CRS-QUE questionnaire [19], designed to evaluate user experience across four categories: Perceived Qualities, User Beliefs, User Attitudes, and Behavioral Intentions. CRS-QUE, which builds upon the ResQue [33] framework for recommender systems, captures both recommendation and conversational aspects of the system's performance. Following the final condition (`Agency`), where participants can freely switch between the two conversational styles, they answer a set of custom questions. These questions focus on their preferred conversational style, the specific aspects they found appealing, and whether they interacted more frequently with one style over the other. Additionally, participants are asked to reflect on how their interaction might change based on familiarity with the research topics. Finally, the NASA-TLX [14] questionnaire is leveraged at this stage to assess the participants' perceived cognitive load during their interactions with the system.

3.3 Evaluation

The evaluation of our user study is two-fold, encompassing both subjective user experience and objective performance metrics. For the subjective component, we gather feedback using post-task questionnaires designed to capture participants' perceptions of our CRS.

---

[2]https://arxiv.org/

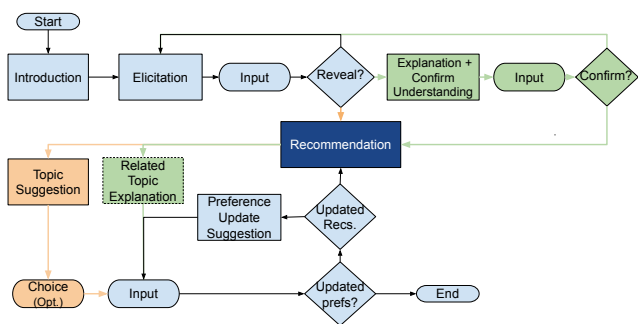[3]Participants in our study had no prior knowledge of the specific topics.

Figure 3: Dialogue flow. The green path shows the main flow for the HighInv, while the orange shows the path for the HighCon conversational style.

Figure 4: Architecture of the Conversational Scholarly Assistant (CSA) system used in this study.

Specifically, we employed the CRS-Que [19] questionnaire, which asks participants to rate their satisfaction across several dimensions: perceived usefulness and ease of use (*Perceived Qualities*), trust in the CRS (*User Beliefs*), overall attitudes toward its use (*User Attitudes*), and future usage intent (*Behavioral Intention*). The responses are captured on a 7-point Likert scale, allowing us to quantify participants' overall satisfaction and engagement with the system. Following the final task, open-ended questions are included to allow participants to elaborate on their experiences, offering qualitative insights into how the different conversational styles affected their interaction with the CRS.

In assessing the objective measures in our study, we focus on the efficiency of the conversational assistant. We do this by analyzing the number of interaction turns and time required for participants to identify relevant articles, both of which are standard metrics for evaluating the efficiency of task-oriented dialogue systems [8]. These metrics help us identify which conversational style leads to quicker and more effective recommendations and how this correlates with participants' subjective experiences. To assess task performance, we evaluate the quality of the bookmarks participants collected during their interactions with the system. Relevance scoring for these bookmarked articles is obtained using a Large Language Model (LLM), building on the approach by Faggioli et al. [10]. Then, an expert in Computer Science verified the LLM-generated relevance scores, achieving substantial inter-rater agreement ($\kappa = 0.78$). Each article is assigned a relevance score of **0** (*irrelevant*), **1** (*somewhat relevant*), or **2** (*highly relevant*), allowing us to determine how effectively the CRS helped participants identify the most relevant papers for their tasks.

4 Conversational Scholarly Assistant System

In this section, we describe the experimental system, i.e., the Conversational Scholarly Assistant (CSA), that we developed and used to conduct our study. In Section 4.1, we describe the two conversational flows that users experienced, while in Section 4.2, the system design and architecture are described.

4.1 Conversational Flow

To address our research questions, we designed a system with two conversational flows characterized by distinct conversational styles.
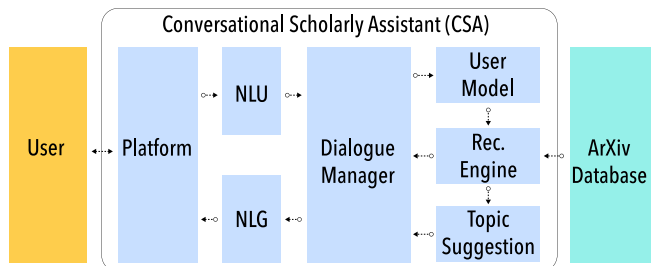
While the underlying task of eliciting preferences and presenting recommendations remains consistent across both styles, the interaction patterns differ to reflect the respective conversational styles. Figure 3 illustrates the core conversational flow, highlighting the differences between the HighInv and HighCon styles. There are two main points of divergence in the flow. First, after the initial elicitation and revealing of preferences, the HighInv style proceeds directly to presenting recommendations. In contrast, the HighCon style engages in an additional round of interaction, providing explanations of keywords the user mentioned and confirming understanding before offering recommendations. This ensures that the user's preferences are well-understood and aligned with the recommendations. The second distinction in the flow between the two conversational styles appears after recommendations are presented. In the HighInv style, the system actively suggests related topics and encourages users to make choices via buttons (though free-text input is always available). This proactive approach aligns with the HighInv style's emphasis on taking initiative and maintaining a dynamic interaction flow. Conversely, in the HighCon style, the system provides explanations of related topics without nudging the user to select them, allowing users to steer the conversation according to their preferences.

Throughout both conversational flows, syntax, and tone are carefully chosen to match the intended style. The HighInv style employs concise language and straightforward content, while the HighCon style opts for detailed explanations and a relatively more nuanced approach to information delivery (see Figure 1).

Although there are notable differences in the operationalization of the two conversational styles, it is important to note that the functionality of the system remains fundamentally the same across these two styles. The contrasting conversational styles merely offer flexibility in how information is requested or presented. For instance, although the HighInv style proactively makes suggestions without prompt, users can similarly request suggestions in the HighCon style. While the former may leverage buttons for quick responses, it also supports free text input, mirroring the latter's preference for open-ended dialogue. Furthermore, despite the frequent explanations characteristic of the HighCon style, users engaged in a HighInv style conversation can still request detailed explanations whenever necessary.

Table 2: Interaction metrics by conversational style. Note that for the number of turns and duration, lower is better. For elicited preferences and seen recommendations, higher is better. We report $\mu \pm \sigma$ (mean and standard deviation) and mixed-design ANOVA (style: within-subjects; domain: between-subjects) F-statistics with associated p-values.

Metric	HighCon	HighInv	Agency	Style Effect	Domain Effect	Interaction Effect
Num. Turns	16.79 ± 11.15	14.75 ± 8.46	$\mathbf{14.08 \pm 7.7}$	$F(1, 22) = 0.42, p = .519$	$F(1, 22) = 1.23, p = .277$	$\mathbf{F(1, 22) = 4.49, p = .045}$
Duration (min)	13.17 ± 6.19	12.77 ± 8.52	13.54 ± 9.84	$F(1, 22) = 0.38, p = .540$	$F(1, 22) = 0.05, p = .816$	$F(1, 22) = 0.00, p = .949$
Recs. Seen	20.58 ± 12.91	30.25 ± 20.46	21.92 ± 11.65	$F(1, 22) = 4.14, p = .054$	$\mathbf{F(1, 22) = 11.23, p = .002}$	$\mathbf{F(1, 22) = 14.49, p < .001}$
Prefs. Elicited	6.58 ± 4.72	6.58 ± 6.20	5.58 ± 3.75	$F(1, 17) = 0.80, p = .381$	$F(1, 17) = 0.00, p = .999$	$\mathbf{F(1, 17) = 9.19, p = .007}$

4.2 System Architecture and Modeling

Our system adopts a modular, component-based architecture [21, Chapter 15]. This widely adopted architecture [4, 23] comprises of three main modules: Natural Language Understanding (NLU), Dialogue Management (DM), and Natural Language Generation (NLG) (Figure 4). Additionally, CSA integrates User Modeling, Recommendation, and Topic Suggestion modules to enhance personalization and relevance in the user experience.

The NLU module interprets user inputs by identifying intents and extracting relevant entities. We utilize a fine-tuned JointBERT model [6] based on SciBERT [3] for improved scholarly-domain understanding. The model recognizes a set of predefined user intents and actions, enabling the handling of various user requests. The DM module maintains the dialogue state and directs the conversational flow based on Section 4.1. In the `HighInv` style, the DM proactively offers suggestions and encourages rapid interactions, while in the `HighCon` style, it provides explanations and seeks confirmations to ensure alignment with the user's intent. The User Modeling module tracks and updates user preferences as keyphrases reflecting areas of interest or disinterest. The system is stateful, maintaining a persistent record of all user preferences throughout the conversation unless the user chooses to remove them. This allows the system to personalize the interaction by dynamically updating user profiles, ensuring more relevant recommendations. The Recommendation module retrieves research papers from an arXiv-based external database and ranks them using BM25. It first selects the top 100 papers based on positive preference keyphrases, then refines rankings by subtracting BM25 scores for negative preference keyphrases. This ensures papers aligned with user disinterests are ranked lower. The final top 10 papers are presented as recommendations.

The Topic Suggestion module analyzes recommended articles to extract the three most common keyphrases, leveraging the Joint-BERT model from the NLU module for consistency across system components. This enables dynamic topic suggestions closely aligned with the user's current preferences. The NLG module generates system responses using a hybrid approach: predefined templates for standard interactions and LLM-based dynamic generation for complex tasks. For most intents—such as eliciting preferences, acknowledging updates, or suggesting topics—the system selects from multiple templates per style to ensure variety and coherence. These templates are designed to align with the conversational tone, concise and direct for the `HighInv` style, and more elaborate and polite for the `HighCon` style. For tasks requiring deeper explanation or justification, the system dynamically generates responses using

an LLM. Our implementation employs LLama 3.1 [9] to provide rich, context-aware explanations.

5 Results and Analysis

Next, we present the findings of our study examining how conversational styles influence the usage of a CRS in the context of identifying relevant scientific literature on a given topic. We address the research questions by analyzing the impact of conversational styles on preference elicitation, task performance, and user engagement. We conducted ANOVA analyses to assess differences across conditions, applying a standard significance threshold of p < 0.05. Given the exploratory nature of our study and a small sample size, our primary objective is to observe trends and patterns rather than establish statistically significant effects.

5.1 Effect on Preference Elicitation

To investigate how conversational styles affect preference elicitation (**RQ1**), we analyzed interaction metrics across the three conversational styles: `HighCon`, `HighInv`, and `Agency`. All 30 participants, master's and PhD students in computer science, completed tasks using each style within their randomly assigned domain—either *Computer Science* (high familiarity) or *Physics* (low familiarity). The average self-reported familiarity scores were significantly different ($p < 0.001$), 5.2 for Computer Science and 2.64 for Physics on a 7-point Likert scale.

Table 2 summarizes the interaction metrics. Although the average number of elicited preferences (6.58) is identical for `HighCon` and `HighInv` conditions and slightly higher than in the `Agency` condition, a significant style-by-domain interaction ($p = 0.007$) indicates that the effectiveness of these styles depends strongly on domain familiarity.

For recommendations seen, participants in `HighInv` viewed substantially more items (30.25) than in `HighCon` (20.58) or `Agency` (21.92). While the style effect approaches significance ($p = 0.054$), domain familiarity significantly influences the number of items participants choose to view ($p = 0.002$), with a highly significant style-by-domain interaction ($p < 0.001$). Additionally, participants in the `HighCon` condition spent more time interacting with the system (average duration of 13.17 minutes) and engaged in more conversational turns (average of 16.79 turns) compared to the `HighInv` condition (12.77 minutes and 14.75 turns). The number of turns shows a significant interaction effect ($p = 0.045$), indicating that the conversational style influenced interaction length differently across domains. The duration metric, however, did not yield significant effects.

Table 3: Average relevance scores of bookmarks ($\mu \pm \sigma$: mean and standard deviation) with mixed-design ANOVA results (style: within-subjects; domain: between-subjects).

Topic	HighCon	HighInv	Agency	Style Effect	Domain Effect	Interaction Effect
Computer Science	1.17 ± 0.36	**1.30 ± 0.57**	1.12 ± 0.80	—	—	—
Physics	1.15 ± 0.70	0.66 ± 0.51	1.25 ± 0.64	—	—	—
Overall	1.16 ± 0.56	0.95 ± 0.62	1.19 ± 0.70	$F(1, 22) = 2.68, p = .115$	$F(1, 22) = 3.40, p = .078$	**$F(1, 22) = 7.78, p = .010$**

Overall, these findings suggest that the `HighInv` style is better suited for exploration, yielding more recommendations overall. The `HighCon` style, while resulting in a slightly lower number of preferences and unique recommendations, can enhance the quality of the preference elicitation by supporting users in refining their preferences through explanations and more thoughtful interactions.

5.2 Effect on Task Performance

To assess the impact of conversational styles on task performance (**RQ2**), we evaluated the quality of the bookmarks participants made during each task using two measures: (1) the average relevance scores assigned by a language model (LLM) used as an auto-rater and (2) the distribution of bookmarks across relevance levels. Each bookmarked article received a score of 2 (highly relevant), 1 (somewhat relevant), or 0 (irrelevant), as described in the study design.

Table 3 presents the average relevance scores by conversational style and topic. In the Physics domain (low familiarity), the `Agency` style led to highest-quality bookmarks, with an average relevance score of 1.25 ± 0.64, compared to the `HighCon` (1.15 ± 0.70) and `HighInv` (0.66 ± 0.51) styles. This indicates that the additional explanations in the `HighCon` style aided participants in selecting more relevant articles when they were less familiar with the topic.

Conversely, in the Computer Science domain (high familiarity), the `HighInv` style resulted in the highest relevance scores (1.30 ± 0.57), slightly outperforming both `HighCon` (1.17 ± 0.36) and `Agency` (1.12 ± 0.80). Participants with high domain expertise likely benefited from the efficiency of the `HighInv` style, enabling them to quickly identify relevant keywords and articles without needing extensive explanations.

Overall, the style effect and domain familiarity alone were not significant. However, there was a statistically significant style-by-domain interaction ($p = 0.010$), confirming that the effectiveness of the conversational styles on bookmark quality strongly depends on the domain familiarity of users. The superior overall performance of the `Agency` condition (1.19 ± 0.70), despite fewer elicited preferences (see Table 2), also suggests that allowing users flexibility in choosing conversational styles enhances task outcomes by catering to their situational needs or preferred interaction style.

5.3 Effect on User Engagement

To explore the effect of conversational styles on user engagement (**RQ3**), we analyze the satisfaction scores from the post-task questionnaires (CRS-QUE). Participants rated their satisfaction on a 7-point Likert scale (1 = very unsatisfied, 7 = very satisfied).

Table 4 shows the user satisfaction scores by conversational style and domain. Overall satisfaction scores are very similar between the `HighCon` 4.34 ± 1.35 and `HighInv` (4.30 ± 1.20) conditions. However,

Table 4: User satisfaction scores by conversational style and topic (*Mean $\pm$ StandardDeviation*)

Measure	HighCon	HighInv	Average
Computer Science	4.01 ± 1.52	4.43 ± 1.03	4.22 ± 1.29
Perceived Qualities	4.28 ± 1.77	**4.68 ± 1.35**	4.48 ± 1.58
User Beliefs	3.72 ± 1.93	**4.01 ± 1.71**	3.87 ± 1.82
User Attitudes	3.92 ± 2.19	**4.65 ± 1.62**	4.29 ± 1.94
Behavioral Intention	3.85 ± 2.08	**4.38 ± 1.56**	4.12 ± 1.82
Physics	**4.63 ± 1.16**	4.21 ± 1.34	4.41 ± 1.26
Perceived Qualities	**4.77 ± 1.61**	4.35 ± 1.69	4.55 ± 1.67
User Beliefs	**4.40 ± 1.67**	4.08 ± 1.65	4.23 ± 1.66
User Attitudes	**5.00 ± 1.55**	4.26 ± 1.68	4.61 ± 1.65
Behavioral Intention	**4.20 ± 1.93**	3.88 ± 1.93	4.03 ± 1.91
Overall Satisfaction	**4.34 ± 1.35**	4.30 ± 1.20	4.32 ± 1.26

there are notable differences when considering domain familiarity. In the Computer Science domain, participants reported a higher satisfaction score in the `HighInv` condition (4.43 ± 1.03) compared to the `HighCon` condition (4.01 ± 1.52). Specifically, scores for all four sub-measures (*Perceived Qualities*, *User Beliefs*, *User Attitudes*, and *Behavioral Intention*) were higher in the `HighInv` condition, suggesting users with higher topic familiarity preferred the direct and efficient interaction offered by the `HighInv` style.

Conversely, in the **Physics** domain, participants reported a higher satisfaction score in the `HighCon` condition (4.63 ± 1.16) compared to the `HighInv` condition (4.21 ± 1.34). The `HighCon` style scored consistently higher across all four sub-measures, indicating that users benefited from the additional explanatory support provided by this style when interacting with less familiar content.

These findings indicate that the alignment between conversational style and user topic familiarity influences user satisfaction with the CRS. This supports our previous observation that users with higher familiarity found the `HighInv` style more satisfying due to its efficiency, while those with lower familiarity preferred the `HighCon` style for its supportive explanations.

6 Conclusions and Future Work

We investigated how conversational styles—high involvement and high considerateness—affect preference elicitation, task performance, and user satisfaction in a Conversational Recommender System (CRS). Through a controlled user study involving 30 master's and PhD students, we found that conversational style influences user interaction and outcomes. The involved style elicited more

user preferences and presented more recommendations, particularly benefiting users with high domain familiarity. In contrast, the considerate style led to higher-quality bookmarks in low-familiarity domains, suggesting that a slower pace and detailed explanations help users select more relevant articles. Allowing users to switch between styles resulted in the highest task performance, indicating the potential for dynamically adapting interactions to suit individual user preferences and situational demands.

However, in interpreting these results, it is important to consider certain limitations. The small sample size and focus on a single domain may limit the generalizability of our findings. Furthermore, the fixed ordering of the Agency condition may have introduced order effects, as participants could have been influenced by prior exposure to the other styles or experienced fatigue by the final task. Expanding to larger participant pools and application domains will validate and extend these findings beyond the discussed limitations.

This study raises ethical considerations around user experience, privacy, and undue influence. Conversational styles not only shape how users engage with the system but also how they articulate and refine their preferences, which can affect decision-making. While some styles may be more effective for certain users, system-driven choices should not unintentionally nudge users toward specific preferences or limit their ability to explore alternatives. Additionally, as our study focuses on non-sensitive academic recommendations without collecting personal data, privacy risks are minimized.

In future work, we aim to explore automatic user expertise detection to enable CRS systems to dynamically tailor interactions. This could further enhance the flexibility and effectiveness of CRSs in real-time recommendation scenarios.

Acknowledgments

References

[1] Fatima Ali Amer Jid Almahri, David Bell, and Mahir Arzoky. 2019. Personas Design for Conversational Systems in Education. *Informatics* 6, 4 (2019), 46.

[2] Krisztian Balog, Lucie Flekova, Matthias Hagen, Rosie Jones, Martin Potthast, Filip Radlinski, Mark Sanderson, Svitlana Vakulenko, and Hamed Zamani. 2020. Common Conversational Community Prototype: Scholarly Conversational Assistant. arXiv:2001.06910 [cs]

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (IJCNLP '19)*. 3615–3620.

[4] Nolwenn Bernard, Ivica Kostric, and Krisztian Balog. 2024. IAI MovieBot 2.0: An Enhanced Research Platform with Trainable Neural Components and Transparent User Modeling. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24)*. 1042–1045.

[5] Roi Blanco, Berkant Barla Cambazoglu, Peter Mika, and Nicolas Torzec. 2013. Entity Recommendations in Web Search. In *The Semantic Web – ISWC 2013 (ISWC '13)*. 33–48.

[6] Qian Chen, Zhu Zhuo, and Wen Wang. 2019. BERT for Joint Intent Classification and Slot Filling. arXiv:1902.10909

[7] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD '16)*. 815–824.

[8] 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, 1 (2021), 755–810.

[9] Dubey et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783

[10] Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*. 39–50.

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

[12] Marcel Gohsen, Johannes Kiesel, Mariam Korashi, Jan Ehlers, and Benno Stein. 2023. Guiding Oral Conversations: How to Nudge Users Towards Asking Questions?. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval (CHIIR '23)*. 34–42.

[13] Akshit Gupta, Debadeep Basu, Ramya Ghantasala, Sihang Qiu, and Ujwal Gadiraju. 2022. To trust or not to trust: How a conversational interface affects trust in a decision support system. In *Proceedings of the ACM web conference 2022*. 3531–3540.

[14] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Human mental workload*. Advances in psychology, Vol. 52. 139–183.

[15] Gaole He, Nilay Aishwarya, and Ujwal Gadiraju. 2025. Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 907–924.

[16] Gaole He, Gianluca Demartini, and Ujwal Gadiraju. 2025. Plan-Then-Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents As A Daily Assistant. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

[17] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2022. A Survey on Conversational Recommender Systems. *Comput. Surveys* 54, 5 (2022), 1–36.

[18] Yucheng Jin, Li Chen, Wanling Cai, and Pearl Pu. 2021. Key Qualities of Conversational Recommender Systems: From Users' Perspective. In *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI '21)*. 93–102.

[19] Yucheng Jin, Li Chen, Wanling Cai, and Xianglin Zhao. 2024. CRS-Que: A User-centric Evaluation Framework for Conversational Recommender Systems. *ACM Transactions on Recommender Systems* 2, 1 (2024), 2:1–2:34.

[20] Ji-Youn Jung, Sihang Qiu, Alessandro Bozzon, and Ujwal Gadiraju. 2022. Great chain of agents: The role of metaphorical representation of agents in conversational crowdsourcing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–22.

[21] Daniel Jurafsky and James H. Martin. 2025. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models* (3rd ed.).

[22] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. 86:1–86:12.

[23] Ivica Kostric, Krisztian Balog, Tølløv Alexander Aresvik, Nolwenn Bernard, Eyvinn Thu Dørheim, Pholit Hantula, Sander Havn-Sørensen, Rune Henriksen, Hengameh Hosseini, Ekaterina Khlybova, Weronika Lajewska, Sindre Ekrheim Mosand, and Narmin Orujova. 2022. DAGFiNN: A Conversational Conference Assistant. In *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys '22)*.

[24] Ivica Kostric, Krisztian Balog, and Filip Radlinski. 2024. Generating Usage-related Questions for Preference Elicitation in Conversational Recommender Systems. *ACM Transactions on Recommender Systems* 2, 2 (2024), 1–24.

[25] Robin Tolmach Lakoff. 1979. Stylistic strategies within a grammar of style. *Annals of the New York Academy of Sciences* 327, 1 (1979), 53–78.

[26] Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. 2019. MeLU: Meta–Learned user preference estimator for Cold–Start recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (KDD '19)*. 1073–1082.

[27] Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020. Estimation-Action-Reflection: Towards Deep Interaction Between Conversational and Recommender Systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM '20)*. 304–312.

[28] Shuokai Li, Yongchun Zhu, Ruobing Xie, Zhenwei Tang, Zhao Zhang, Fuzhen Zhuang, Qing He, and Hui Xiong. 2022. Customized Conversational Recommender Systems. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 740–756.

[29] Kai Luo, Hojin Yang, Ga Wu, and Scott Sanner. 2020. Deep critiquing for VAE–Based recommender systems. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval (SIGIR '20)*. 1269–1278.

[30] Fedelucio Narducci, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. 2018. Improving the user experience with a conversational recommender system. In *AI*IA 2018 – advances in artificial intelligence (AI*IA '18)*. 528–538.

[31] J. A. Nelder and R. W. M. Wedderburn. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)* 135, 3 (1972), 370–384.

[32] Dhanya Pramod and Prafulla Bafna. 2022. Conversational recommender systems techniques, tools, acceptance, and adoption: A state of the art review. *Expert Systems with Applications* 203 (2022).

[33] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. 157–164.

[34] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Estimating conversational styles in conversational microtask crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

[35] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Improving Worker Engagement Through Conversational Microtask Crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. 1–12.

[36] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web (WWW '01)*. 285–295.

[37] Anna Sepliarskaia, Julia Kiseleva, Filip Radlinski, and Maarten de Rijke. 2018. Preference elicitation as an optimization problem. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)*. 172–180.

[38] Ameneh Shamekhi, Mary Czerwinski, Gloria Mark, Margeigh Novotny, and Gregory A Bennett. 2016. An exploratory study toward the preferred conversational style for compatible virtual agents. In *International Conference on Intelligent Virtual Agents*. Springer, 40–50.

[39] Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building Better AI Agents: A Provocation on the Utilisation of Persona in LLM-based Conversational Agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI '24)*. 1–6.

[40] Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. 235–244.

[41] Richard Sutcliffe. 2023. A Survey of Personality, Persona, and Profile in Conversational Agents and Chatbots. arXiv:2401.00609 [cs]

[42] Deborah Tannen. 1987. Conversational style. *Psycholinguistic models of production* (1987), 251–267.

[43] Deborah Tannen. 2005. *Conversational style: Analyzing talk among friends*. Oxford University Press.

[44] Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and alignment in information-seeking conversation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. ACM, 42–51.

[45] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A data set of information-seeking conversations. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, Vol. 5.

[46] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second chance for a first impression? Trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on user modeling, adaptation and personalization*. 77–87.

[47] Ga Wu, Kai Luo, Scott Sanner, and Harold Soh. 2019. Deep Language–Based critiquing for recommender systems. In *Proceedings of the 13th ACM conference on recommender systems (RecSys '19)*. 137–145.

[48] Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards Persona-Based Empathetic Conversational Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (EMNLP '20)*. 6556–6566.

[49] Lixin Zou, Long Xia, Yulong Gu, Xiangyu Zhao, Weidong Liu, Jimmy Xiangji Huang, and Dawei Yin. 2020. Neural Interactive Collaborative Filtering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 749–758.