Towards Human-centered Proactive Conversational Agents

Yang Deng
National University of Singapore
Singapore
ydeng@nus.edu.sg

Lizi Liao
Singapore Management University
Singapore
lzliao@smu.edu.sg

Zhonghua Zheng
Harbin Institute of Technology, Shenzhen
Shenzhen, China
polang1999@gmail.com

Grace Hui Yang
Georgetown University
Washington, D.C., United States
grace.yang@georgetown.edu

Tat-Seng Chua
National University of Singapore
Singapore
chuats@comp.nus.edu.sg

ABSTRACT

Recent research on proactive conversational agents (PCAs) mainly focuses on improving the system's capabilities in anticipating and planning action sequences to accomplish tasks and achieve goals before users articulate their requests. This perspectives paper highlights the importance of moving towards building human-centered PCAs that emphasize human needs and expectations, and that considers ethical and social implications of these agents, rather than solely focusing on technological capabilities. The distinction between a proactive and a reactive system lies in the proactive system's initiative-taking nature. Without thoughtful design, proactive systems risk being perceived as intrusive by human users. We address the issue by establishing a new taxonomy concerning three key dimensions of human-centered PCAs, namely **Intelligence**, **Adaptivity**, and **Civility**. We discuss potential research opportunities and challenges based on this new taxonomy upon the five stages of PCA system construction. This perspectives paper lays a foundation for the emerging area of conversational information retrieval research and paves the way towards advancing human-centered proactive conversational systems.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; • **Information systems** → *Users and interactive retrieval*; • **Human-centered computing** → *Natural language interfaces.*

KEYWORDS

Proactive Agent, Conversational Agent, Human-centered Design

ACM Reference Format:
Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards Human-centered Proactive Conversational Agents. 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, 12 pages. https://doi.org/10.1145/3626772.3657843

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

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657843

1 INTRODUCTION

With the advent of large language models (LLMs), the emergence and integration of conversational systems mark a significant leap forward in information retrieval (IR), which evolves many traditional interactive IR systems into conversational IR systems. For instance, Microsoft recently released a new version of Bing with its integration with ChatGPT [52] under the idea of conversational search. In the rapidly evolving field of conversational systems, proactive conversational agents (PCAs) [14, 30, 41] are emerging to revolutionize how systems interact with human users. In the literature [7, 15], the proactivity of a conversational system typically refers to the system's ability of being aware of the long-term conversational goal and capable of taking initiatives to lead the conversation towards the goal. Recent years have witnessed a number of advanced designs that address proactivity on a range of conversational systems. For instance, in conversational information seeking, PCAs are developed to further eliminate the uncertainty for more efficient and precise information seeking by initiating ambiguity clarification [2, 84] or eliciting the user preference [38, 90], instead of simply reacting to user queries. While in open-domain dialogue systems, different from passively echoing the user-initiated discussion topics or emotion requirements, various PCA designs arise to be capable of directing the conversations [21, 39]. The distinction between a proactive and a reactive system lies in the proactive system's initiative-taking nature, effectively increasing the number of initiators in an interactive system from one (the user) to two (both the user and the machine). Without thoughtful design, proactive systems risk being perceived as intrusive by human users. Consequently, the key to the widespread acceptance and effectiveness of PCAs lies in their design being fundamentally human-centered, rather than solely advancing technical efficiency and proficiency.

To this end, this perspectives paper discusses the intricate balance of technological advancement and human-centered design principles in the creation of proactive conversational agents. We envision human-centered PCAs to be a kind of PCA that *emphasizes human needs and expectations*, and *considers the ethical and social implications* of these agents, beyond technological capabilities. We propose to establish a new taxonomy concerning three key dimensions of human-centered PCAs, namely **Intelligence**, **Adaptivity**, and **Civility**, as shown in Figure 1. We first investigate the past work on proactive conversational agents based on the new taxonomy and then prospect a board research agenda for

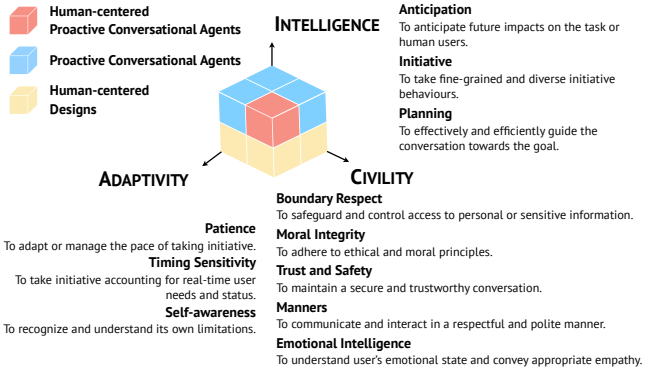

Figure 1: Three key dimensions of human-centered proactive conversational agents with representative abilities.

building human-centered PCAs. The goal of this paper is to act as a handbook for discussions on the human-centered designs in every stage of the construction of PCAs, including Task Formulation, Data Preparation, Model Learning, Evaluation, and System Deployment.

2 OVERVIEW

2.1 Key Dimensions

To develop human-centered proactive conversational agents, we have identified three key dimensions specific to these agents. We propose deriving both design principles and construction guidelines from these dimensions to inform the development of human-centered proactive conversational systems. These dimensions are **Intelligence**, **Adaptivity**, and **Civility** (shown in Figure 1).

- **Intelligence**: Intelligence in a proactive conversational agent is characterized by its capabilities to anticipate future development of the task and to perform strategic planning ahead of user requests, essential for achieving the conversation's goals proactively. This involves taking nuanced initiative and anticipating both the short-term and long-term impacts on the task or human users. PCAs with low-level intelligence may exhibit inaccurate, and unfocused initiative, like a well-intentioned but amateurish helper, eager to assist but lacking expertise or skills.
- **Adaptivity**: Adaptivity refers to the capability of PCAs to dynamically adjust and manage the timing and pacing of its actions and interventions in response to the user's real-time context and evolving needs. This requires the agent to be designed with patience in determining the initiative's pace, sensitivity to the impact of taking initiative while considering real-time user needs and status, and self-awareness of its capabilities and limitations, particularly in understanding when and how often to intervene in a manner that is most beneficial and relevant to the user.
- **Civility**: Civility in proactive conversational agents refers to the agent's capability to recognize and respect the physical, mental, and social boundaries set by the user, the conversational task, and general ethical standards. These agents should be adept at understanding both personal and task-related boundaries and respecting them while taking proactive initiatives. This covers a broad spectrum of personal and social norms, including maintaining privacy, ensuring trust and integrity, and avoiding interactions that are intrusive or disrespectful.

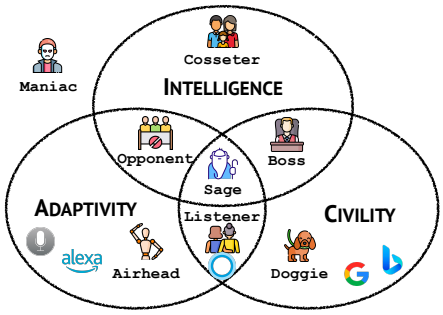

Figure 2: Different types of proactive conversational agents in terms of three key dimensions of human-centered PCAs.

2.2 Types of Proactive Conversational Agents

As illustrated in Figure 2, based on the proficiency level of the three dimensions, we can categorize the proactive conversational agents into eight general types:

- **`Sage`** (High Intelligence, High Adaptivity, High Civility) denotes the type of PCAs that meet the standards of three dimensions of human-centered PCAs. Its sophisticated, personalized, and respectful interactions, making it a valuable asset in diverse fields that require nuanced and human-centered AI assistance.
- **`Opponent`** (High Intelligence, High Adaptivity, Low Civility) denotes the type of PCAs that are designed to engage in thorough and persistent interactions, but possibly challenging or negotiating the user's views and decisions in an opposite position. In order to achieve their specific goals, these systems sometimes may intrude the user's personal or social boundaries. For example, negotiation systems sometimes involve strategies [24] that can be intrusive and potentially disrespectful to human users, such as attacking the opponent's stance.
- **`Boss`** (High Intelligence, Low Adaptivity, High Civility) denotes the type of PCAs that would likely offer efficient assistance and be considerate of the user's boundary and privacy, but their interactions might be more direct and to-the-point, prioritizing effectiveness and clarity over user needs and engagement, just like the authoritative boss at work. The analysis of proactive robotic assistants in HCI studies [33, 34] reveals that taking initiative at different timing leads to different impacts on the user's trust. In high-stakes, fast-paced settings like emergency response or critical business decisions, a `Boss`-type PCA excels by providing clear, direct guidance while respecting boundaries, ensuring swift and accurate task completion.
- **`Cosseter`** (High Intelligence, Low Adaptivity, Low Civility) denotes the type of PCAs that are overly involved and excessively monitoring or controlling in their interactions with users, similar to the overprotective behaviors associated with helicopter parenting. For instance, some conversational recommender systems may excessively acquire users' personal information by asking preference elicitation questions [18], which potentially lead to user discomfort or a sense of intrusion, as it may be perceived as invasive or aggressive in its attempt to reach their goals.
- **`Listener`** (Low Intelligence, High Adaptivity, High Civility) denotes the type of PCAs that are friendly, engaging, and empathetic in their interactions. These systems are often referred to social chatbots, such as Cortana, which would likely be designed

Table 1: Case studies of task formulation. ✓ and ✗ denote whether certain dimension has been considered or not.

Task Formulation	Intelligence	Adaptivity	Civility	PCA Type
Asking Clarifying Question	✗ (single-type strategy)	✗ (frequently take initiative)	✓	Doggie
Mixed-initiative Information Seeking	✓(multi-type strategy)	✓(depend on initiative needs)	✓	Sage
Empathetic Dialogue	✗ (single-type strategy)	✓	✓	Listener
Emotional Support Dialogue	✓(multi-type strategy)	✓	✓	Sage
Negotiation Dialogue	✓	✓	✗ (no restriction on strategies)	Opponent
Pro-social Negotiation Dialogue	✓	✓	✓ (constrained by social norms)	Sage
Target-guided Dialogue	✓	✗ (favour aggressiveness)	✗ (no restriction on targets)	Cosseter
Personalized Target-guided Dialogue	✓	✓ (considering user engagement)	✓ (constrained by user preference)	Sage

to entertain users by shifting topics [75] or comforting users. Generally speaking, in casual, low-stress environments such as home settings or social spaces, a Listener-type PCA thrives by offering companionship and emotional support, engaging users with empathy [58] and flexible conversation.

- **Airhead** (Low Intelligence, High Adaptivity, Low Civility) denotes the type of PCAs that are perceived as lacking depth, sophistication, or serious functionality, similar to the colloquial use of "airhead" to describe someone who is not very thoughtful or intelligent. For example, the proactivity in voice assistants (*e.g.*, early versions of Siri and Alexa) lies on agent-initiated interactions triggered by contextual and environmental events or user behaviours [49]. Their simplistic yet responsive nature makes them suitable for straightforward tasks or as novice-friendly, unintimidating interfaces for technology newcomers. While users raise concerns on privacy protection and intrusiveness [59, 86].
- **Doggie** (Low Intelligence, Low Adaptivity, High Civility) denotes the type of PCAs that are friendly, highly responsive, and possibly intuitive, like the typical characteristics associated with dogs. For example, recent years witnessed that many search engines, such as Google and Bing, are equipped with conversational features for proactive interactions, such as suggesting useful queries [60] or asking clarification questions [2, 84]. A user study in Zou et al. [96] shows that systems should ask clarification questions only when necessary, instead of frequently asking clarification questions or suggesting useful queries.
- **Maniac** (Low Intelligence, Low Adaptivity, Low Civility) denotes the type of PCAs characterized by their aggressive and irrational initiative behaviours, closely resembling the unpredictable nature of an uncontrollable maniac.

The current landscape of commercial conversational systems predominantly features a foundational level of proactive Intelligence, highlighting an exciting area for ongoing research to elevate their capabilities in this domain. Additionally, while significant strides have been made in developing PCAs, there is an emerging recognition of the importance of further exploring two other vital dimensions: Adaptivity and Civility. These aspects are essential for crafting PCAs that are truly centered around human needs and preferences, offering a well-rounded and user-friendly experience.

2.3 Five Stages for PCA System Construction

In terms of PCA system construction, we can briefly organize the process into five stages sequentially, namely Task Formulation, Data Preparation, Model Learning, Evaluation, and System Deployment. The proposed human-centered designs are supposed to be present in every stage during the construction of PCAs:

- **Task Formulation** is the initial stage where the objectives and scope of the PCA are defined, setting its development foundation.
- **Data Preparation** involves collecting, cleaning, and organizing the necessary data that is used for the PCA to learn.
- **Model Learning** is the phase where the PCA is trained using algorithms and the data to make desired decisions and responses.
- **Evaluation** is a critical stage where the agent's performance is assessed to ensure it meets the desired standards of interaction.
- **System Deployment** is the final stage where the developed PCA is integrated into the environment to interact with users.

In what follows, we first re-interpret the current studies of building PCAs from the new human-centered taxonomy upon the five stages for PCA construction, and then correspondingly prospect future research directions and challenges under each stage.

3 TASK FORMULATION

Existing task formulation of PCAs mainly prioritizes the dimension of Intelligence aimed at goal completion, while it is also crucial to ensure the integration of user emotions, preferences, and values, and adherence to ethical standards. The other two dimensions, Adaptivity and Civility, are key to fostering trust, satisfaction, and seamless interactions between humans and systems.

With the three key dimensions from human-centered perspectives, we re-interpret the task formulation of existing PCA literature, and discuss how new tasks can be derived from them. As shown in Table 1, we elaborate the discussions with several widely-studied and representative research task formulations in IR community.

- **Current: Asking Clarifying Question**. A proactive conversational information-seeking system [12, 85] might ask for clarification on an ambiguous query. However, frequent clarification requests can negatively impact user experience [95], resembling a Doggie-type PCA's approach.
- **Desired: Mixed-initiative Information Seeking**. Key system-initiated behaviors include asking clarifying questions [2], managing out-of-scope queries [74], and providing extra information [11]. Recent studies [74, 78] focus on diverse strategies to enhance the agent's Intelligence. Besides, system initiative prediction [47] (*e.g.*, clarification need prediction [1, 17]) is vital for improving Adaptivity in mixed-initiative information seeking.
- **Current: Empathetic Dialogues**. Traditional task formulation [58, 94] on emotion-aware dialogue systems has predominantly focused on crafting responses that echo the user's emotions or mirror their feelings, being a Lisenter-type PCA.
- **Desired: Emotional Support Dialogues**. In contrast, emotional support dialogue systems [42] are designed with the objective of

improving the user's emotional well-being from certain negative emotional states with interventions like in Cognitive Behavioral Therapy (CBT). The task is formulated to extend beyond merely demonstrating empathy; they should proactively take different emotional support strategies to engage with the user's concerns and deliver actionable advice or encouragement to aid in resolving the issues. An empirical analysis [22] shows that proactive behaviours at different phases of the conversation may lead to different impacts on the user's emotional state. Similarly, it is also a crucial subtask for emotional support dialogues to determine when to take the initiative, ensuring Adaptivity.

- **Current: Negotiation Dialogues**. Negotiation dialogue [87] is a process of strategic interaction aimed at finding mutually acceptable solutions between parties, but, meanwhile, maximizing the profit from one side. This concept, deeply rooted in psychology, political science, and communication, has a wide range of applications in everyday life, including price bargaining, strategic gaming, and persuasion. To successfully negotiate, Intelligence and Adaptivity are key components considered in the current task formulation. This indicates that the prevalent formulation typically focuses on building `Opponent`-type PCAs for negotiation dialogues, neglecting the perspective of Civility. Strategy modeling is the primary aspect formulated in negotiation dialogue problems. Negotiation strategies encompass various tactics aimed at achieving goals, but some can be intrusive and potentially disrespectful. Real-world negotiation sometimes involves strategies [24] like contesting (attacking the opponent's stance), empowerment (emphasizing personal preferences to counter claims), and self-pity (evoking guilt). When used inappropriately, these tactics can cross the trust boundary of human users, harming the ethical conduct of the negotiation.
- **Desired: Pro-social Negotiation Dialogues**. To maintain boundary respect, the task formulation should involve constraints of avoiding strategies that might humiliate or provoke the other party and promoting polite and empathetic interactions [50, 61].
- **Current: Target-guided Dialogues**. This task involves conversational agents proactively leading the dialogue towards a specific target, *e.g.,* certain topics for chit-chats [63] or particular items for recommendation [44]. This approach has gained significant attention for its potential to enhance system effectiveness. However, current formulations often neglect Adaptivity and Civility, aligning with a `Cosseter`-type PCA. In terms of Adaptivity, LLM-based systems [16] show proficiency in goal-directed conversation, but abrupt topic shifts can reduce user satisfaction and engagement [39], resembling aggressive sales tactics. Regarding Civility, current task formulations do not impose restrictions on the choice of targets. If a target topic is harmful or toxic, the conversation may violate ethical boundaries. Similarly, if a target item is chosen solely by the seller without considering user preferences, it can erode users' trust, as they may feel the system prioritizes profits over their needs.
- **Desired: Personalized Target-guided Dialogues**. When considering Adaptivity in target-guided dialogues, a human-centered task formulation should encompass not only the efficiency of achieving the target but also the constraints related to user satisfaction and the smoothness of topic transitions. Besides, the target should align with the user's interests and needs. For example, in target-guided conversational recommendation [13, 44], the target can be first customized from a set of items for promotion based on user preferences, instead of being directly assigned.

The potential of promoting genuine value to the system side ensures that the development of PCAs increasingly garners attention from both academic and industrial sectors. However, the societal acceptance of PCAs hinges crucially on meeting standards of Adaptivity and Civility. Consequently, it is of great importance to establish a well-defined objective for the foundation of PCAs.

4 DATA PREPARATION

To gather conversational data, traditional methods often involve the recording and collection of raw dialogue samples, such as customer service logs, online forum threads, or video transcripts, capturing the data in its natural state. On the other hand, a significant portion of existing proactive dialogue datasets employs context-based data collection, either annotated by crowdworkers or generated by AI. Context-based data collection refers to the process of gathering dialogue data with underlying circumstances or background information that are pre-defined to direct the conversations.

4.1 Issues on Current Data Preparation Schemes

Apart from aiming at collecting data for the agent's Intelligence, we analyze the data preparation of existing proactive conversational datasets from the other two dimensions: Adaptivity and Civility.

4.1.1 Fabricated User Needs. From the perspective of Adaptivity, context-based data collection typically fabricates user needs for system-initiated behaviours to construct proactive conversation data. Upon training on the data with fabricated user needs for the agent's proactivity, it may result in inappropriate proactive behaviours regardless of real user needs, harming the adaptivity of PCAs. As listed in Table 2, we analyze the data preparation process of several widely-studied datasets for various PCA tasks. For example, in conversational information seeking datasets, some ambiguous queries do not naturally occur, while the ambiguity is introduced by deliberately truncating conversations [26] or omitting information [17]. Similar findings are drawn in a recent survey of asking clarification questions datasets [56]. In target-guided dialogue datasets [63, 79], some target topics are assigned without considering actual user needs while just specifying some random words with unclear meanings, like "blue" or "tired". In emotional support dialogue datasets, dialogues can be produced by asking annotators to role-play as patients dealing with a specific imagined emotional problem [42], instead of the real user needs for counseling [46, 55]. Similar patterns are also found in negotiation dialogue [27, 40] and target-guided dialogue datasets [44], where annotators are instructed to play a pre-defined role. Overall, those datasets constructed by context-based crowdworker annotations or rule-based reconstructions are more likely to contain proactive dialogues with fabricated user needs.

4.1.2 Ethical Concerns. Language toxicity has been an essential consideration in the perspective of Civility for human-centered PCAs. Following the toxicity assessment in previous studies [92, 93], we assess the toxicity of the utterances in the datasets listed in Table 2 by reporting the Toxicity and Severe Toxicity scores computed

Table 2: Analysis of user needs and toxicity in existing proactive conversation datasets. (The lower the better ↓.)

Problem	Dataset	Description of Data Preparation	User Needs	Toxicity ↓	Severe Toxicity ↓
Conversational Information Seeking	Qulac [3]	Created from the logs of search engine	Real	0.052	0.004
	Abg-CoQA [26]	Truncate conversations to induce ambiguity	Fabricated	0.095	0.003
	PACIFIC [17]	Manually rewrite queries to induce ambiguity	Fabricated	0.019	0.001
Target-guided Dialogue	TGC [63]	Rule-based keyword extractor to label targets	Fabricated	0.197	0.020
	TGConv [79]	Randomly specify an easy target and a hard target	Fabricated	0.202	0.012
	DuRecDial [44]	Crowdworker annotations based on given user profiles	Fabricated	0.118	0.007
Emotional Support Dialogue	HOPE [46]	Created from the transcriptions of counselling videos	Real	0.151	0.007
	MI [55]	Created from the transcriptions of counselling videos	Real	0.122	0.005
	ESConv [42]	Crowdworker annotations based on given scenarios	Fabricated	0.076	0.004
Negotiation Dialogue	CraigslistBargain [27]	Crowdworker annotations based on given bargaining targets	Fabricated	0.160	0.011
	AntiScam [40]	Crowdworker annotations based on given intents	Fabricated	0.080	0.005
	P4G [70]	Crowdworker annotations with a pre-task survey as user profiles	Real	0.048	0.002

by `Perspective API` [37]. In general, crowdworker-annotated datasets exhibit safer conversations (lower toxicity scores) than those datasets collected from real-world conversations. For example, among three emotional support dialogue datasets, the toxicity degree of the two datasets collected from the raw transcriptions of counseling videos (*i.e.*, HOPE [46] and MI [55]) are substantially higher than those of the crowdsourced dataset, ESConv [42].

4.1.3 Issues on Different Types of Data Preparation Schemes. Besides the above issues drawn from our analysis, we summarize the drawbacks of different types of data preparation schemes by further combining additional evidences from the literature as follows:

- **Real-world Collection**. (1) **Ethical Concerns**: Real-world data collection often involves sensitive personal information or potentially toxic content, raising significant privacy issues and necessitating toxicity assessment and detoxification. (2) **Quality Variability**: Real-world conversations can vary greatly in quality, relevance, and clarity.
- **Crowdworker Annotations**. (1) **Fabricated User Needs**: Crowdworkers are asked to perform conversations in a constrained setting, which could be different from how people with real user needs interact in natural conversations. (2) **Homogeneous Dialogue Patterns**: Crowdworkers generate dialogues specific to the provided context and are asked to generate dialogues of a certain style. Dialogues created in this way have a high degree of pattern overlap, either in lexical or logic [6].
- **Generative AI Annotations**. (1) **Lack of Human Intuition**: AI-generated dialogue annotations by following human instructions [36] might not capture the intricate contexts or subtleties of human conversation. For instance, while humans can grasp underlying emotions or subtext, AI might provide responses that feel shallow or off-mark. (2) **Propagation of Biases in Dialogues**: If trained on a biased dialogue dataset, AI might reproduce and magnify these biases in the annotated dialogues [23]. This can lead to datasets that inadvertently favor or disfavor certain topics, demographics, etc.

4.2 New Perspectives on Data Preparation

The dataset analysis in Section 4.1 reveals the limitations of existing data preparation schemes for proactive conversations. To mitigate these issues, we discuss two promising solutions as follows:

4.2.1 Reflecting Real Human Needs. A human-centered approach to data collection should emulate real-world scenarios to ensure the data genuinely represents actual human needs. A famous example is the data collection of Natural Questions [35], which consists of real anonymized, aggregated queries issued to the Google search engine. It revolutionizes the context-based QA data collection approaches [57] where annotators are asked to first read the passage containing the answer to generate the question. This is also valid under the context of proactive dialogue data collection. Collecting raw data from real-world conversations, like Qulac [3], HOPE [46], and MI [55], is the most straight-forward way to ensure real user needs. Similarly, Liu et al. [43] construct a user needs-centric E-Commerce conversational recommendation dataset (U-NEED) from real-world E-Commerce pre-sales dialogues, where the target item is described by the real human users, instead of random target assignment. However, it can be difficult to obtain the desired resources and the collected real-world dialogues suffer from quality and ethical issues. In the case where annotators are necessary to construct a conversation dataset, the real user needs can be reflected by asking the annotators to just be themselves and collecting their own background information, such as using pre-task survey [70].

4.2.2 Human-AI Collaborative Data Collection. To compensate for the limitations in both crowdworker and generative AI annotations, Macina et al. [45] pair human teachers with an LLM that simulates students and their errors for tutoring dialogue data collection. By integrating the nuanced understanding of human teachers with the scalable and controllable generation capabilities of LLMs, this approach can produce tutoring dialogue data with more diverse patterns and more educationally valuable intuition. In order to diversify the target accomplishment process with different user personalities for augmenting target-guided dialogue data, Wang et al. [67] create the TopDial dataset, which employs generative AI to simulate a variety of users by using Big-5 personality traits and user profiles. While improving generative AI annotations, an effective approach is to incorporate human knowledge for guiding the generative AI. Zheng et al. [93] leverage expert knowledge to design various emotional support strategies and collect real-world counseling cases for creating the emotional support dialogue dataset, named ExTES. Additionally, the issue of ethical concerns in real-world data collection can also be alleviated by the AI re-evaluation for conforming to social rule-of-thumbs [32].

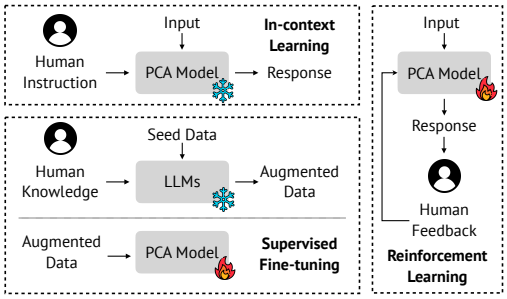

Figure 3: Three types of human alignment approaches.

5 MODEL LEARNING

Most existing studies propose various advanced methods for building an intelligent proactive conversational agents that equip sophisticated planning capabilities. Despite the improved Intelligence, these PCAs are not always adept at interpreting a wide range of real-world situations or may produce responses that deviate from human expectations. While recently many efforts have been made in aligning language models with human values and expectations, namely *Human Alignment* [71, 80], which is a valuable technique for integrating Adaptivity and Civility into the model learning of PCAs. As illustrated in Figure 3, we discuss three main types of human alignment approaches for building human-centered PCAs, including In-context Learning (ICL), Supervised Fine-tuning (SFT), and Reinforcement Learning (RL). Furthermore, we prospect the remaining challenges and future research agenda accordingly.

5.1 Prompting by Human Instructions

ICL has emerged as a highly efficient learning paradigm in the era of LLMs, since LLMs possess substantial knowledge and exceptional instruction-following capabilities.

- **Intelligence**. With the emerging capabilities of LLMs [72], plan generation by prompting LLMs is becoming the main-streamed paradigm for complex task solving. Motivated by this, recent studies design various prompting schemes for instructing LLMs to conduct self-thinking of strategy planning, including Chain-of-Thought (CoT) [16, 66] and multi-agent debate [25, 88].
  Challenges: Current prompt-based methods fail to do anticipation to optimize the long-term goal of the conversation.
- **Adaptivity**. There is few work investigating the prompt-based approaches for taking account the Adaptivity of PCAs. Notable observations are drawn from Deng et al. [16], where the human-designed proactive CoT (ProCoT) prompting scheme mitigates the aggressive topic shift of LLM-based PCAs in target-guided dialogues. Despite the lack of explicit designs for Adaptivity, the self-thinking instruction in ProCoT may enable the PCA to capture nuances in conversations, such as user satisfaction.
  Challenges: The underlying reasons for the enhanced Adaptivity remain unclear and the improvement is still far from satisfactory where LLMs with ProCoT prompting still tend to make more aggressive topic transitions than desired.
- **Civility**. By integrating the alignment goals directly into the prompts, ICL can guide and regulate the responses of PCAs, ensuring they align more closely with desired outcomes and guidelines. For example, Chen et al. [8] empirically show that PCAs with manually designed mixed-initiative strategy prompts become more honest and thoughtful (higher Civility) in emotional support and persuasion dialogues.
  Challenges: The manually designed prompts for each type of strategy lack transferability to unseen scenarios and are limited to specific abilities, such as Manners and Moral Integrity in Chen et al. [8], while neglecting other abilities of Civility.

5.2 Data Augmentation with Human Knowledge

The utilization of LLMs for data augmentation in conversational systems has gained substantial attention for offering promising avenues to improve response quality and dialogue performance.

- **Intelligence**. There are two typical paradigms: (1) Self-chat distillation methods [92, 93] directly distill the conversational intelligence from LLMs by prompting LLMs to complete the conversations with specific human instructions. (2) Role-play simulation methods [45, 67] employ LLMs to simulate a specific role in the conversation to communicate with humans or other role-playing agents for collecting conversation data.
  Challenges: Despite the remarkable quality of LLM-augmented dialogue data, this type of data inevitably inherits the limitation of LLMs in handling proactive dialogues, such as limited abilities to make strategic decisions and plans for long-term goals.
- **Adaptivity**. The practice of SFT using limited annotated datasets may result in a lack of generality in PCAs, particularly when encountering diverse user personalities or unforeseen scenarios. To address this, recent research [67, 93] has increasingly focused on incorporating human knowledge into LLMs to enhance the Adaptivity of these systems, allowing them to better handle a wider range of conversations and user needs.
  Challenges: The generality of the augmented data is still limited by the available human knowledge.
- **Civility**. To enable PCAs appropriately respond to unsafe or unethical user utterances, Kim et al. [32] augment morality-related dialogue datasets with social rule-of-thumb knowledge from human annotators. After being fine-tuned on the augmented data, PCAs can lead the conversation in a prosocial manner.
  Challenges: Due to the absence of negative feedback from humans, the fine-tuned model only knows what should do but may fail to prevent from what should not do for achieving Civility.

5.3 Learning from Human Feedback

Reinforcement learning from Human Feedback (RLHF) [54] is designed to align the language model with humans from human preference signals under the RL framework.

- **Intelligence**. Due to the high cost of human feedback on the intelligent completion for long-term goals, researchers [20, 83] simulate the goal-oriented human feedback by using generative AI for enhancing the planning and anticipation abilities of PCAs.
  Challenges: RLHF assumes that human advances in their abilities for aligning the model behaviours. However, considering a future PCA model may become much more intelligent than humans, humans will no longer be reliable to supervise the model in those complex tasks that we don't understand. This situation

Table 3: Experimental results on the ESConv dataset [42]. (The lower the better ↓. The higher the better ↑.)

Type	Method	Intelligence		Adaptivity			Civility				
		Succ. Rate ↑	Avg. Turn ↓	Smoothness ↑	Satisfaction ↑	ECE ↓	Toxicity ↓	Identity Attack ↓	Threat ↓	Insult ↓	Relaxation ↑
-	ChatGPT [52]	0.7692	5.10	0.3933	4.29	0.4631	0.0591	0.0019	0.0105	0.0261	0.3773
ICL	Ask-an-Expert [88]	0.8000	4.76	0.3346	4.16	0.3814	0.0633	0.0082	0.0089	0.0284	0.3958
	ProCoT [16]	0.7769	4.83	0.3704	4.26	**0.3199**	0.0586	**0.0061**	**0.0080**	0.0265	0.3525
SFT	AugESC [92]	0.7445	5.43	0.4181	3.80	0.3856	0.0605	0.0086	0.0184	0.0254	0.3482
	ExTES [93]	0.7954	4.67	0.4437	**4.35**	0.3321	**0.0526**	0.0071	0.0082	**0.0071**	**0.4110**
RL	RLHF [54]	0.8592	4.51	0.4398	3.92	0.4053	0.0629	0.0100	0.0245	0.0273	0.3851
	Aligned$_{\text{d-PM}}$ [68]	**0.8785**	**4.46**	**0.4525**	4.09	0.3816	0.0554	0.0065	0.0080	0.0275	0.4092

raises the needs for exploring Superalignment approaches [53] that ensure superintelligent PCAs reliably follow human intents.

- **Adaptivity**. Most RLHF approaches generally use majority voting or averaging to combine inconsistent preferences into a unified one. However, this process represents only a narrow segment of individuals, failing to effectively reveal the full scope of human preferences universally. To remedy this, Wang et al. [68] propose to account for the distribution of disagreements among human preferences. Zhang et al. [89] designs role-playing user simulators with various personality to interact with PCAs for enhancing the diversity of human feedback.
  Challenges: The cost of collecting diverse human feedback is high and the alignment relies heavily on the quality of human preference feedback. Finding a balance between minimizing human cost while still maintaining high-quality alignment is a key challenge in the development of human-aligned Adaptivity.
- **Civility**. Besides the general safety alignment in most RLHF approaches, Xu et al. [77] further propose to employ an external memory to store established rules for alignment with diverse and customized human values, including legal and moral rules.
  Challenges: As for PCAs, the measurement of feedback quality is supposed to be multifaceted, involving the three key dimensions. Therefore, the widely-used binary feedback becomes indistinguishable in quality, while it is critical to investigate a more comprehensive form of human feedback collection for PCAs.

6 EVALUATION

An ideal human-centered proactive conversational agent should fulfill certain criteria across three key dimensions. However, existing evaluations of PCAs mainly focus on the Intelligence perspective, using metrics like Goal Completion and Response Quality. According to various goals, the evaluation of goal completion can be referred to the achievement of the target topic [39, 79], the clarification-based information seeking [62], the negotiation gain [31, 70], etc. Meanwhile, the quality of responses is typically judged by human annotators, involving factors like fluency, informativeness, etc. However, the other two aspects are often overlooked.

Due to the lack of existing metrics for assessing the representative abilities of Adaptivity and Civility, we propose a preliminary multidimensional evaluation framework by adopting some alternative evaluation protocols to evaluate these abilities for PCAs. Specifically, we take a widely-studied PCA task as a case study, namely Emotional Support Dialogues introduced in Section 3, and conduct the multidimensional evaluation on different human-centered model learning techniques discussed in Section 5.

6.1 Proposed Evaluation Framework

Besides Intelligence which has been extensively evaluated in the literature, including reference-based response quality metrics (*e.g.,* BLEU and ROUGE) and goal completion metrics (*e.g.,* Success Rate and Average Turn [20]), we first introduce some alternative evaluation protocols for the evaluation of Adaptivity and Civility.

6.1.1 Evaluation Protocols for Adaptivity. There are three representative abilities of Adaptivity:

- **Patience** refers to the ability to adaptively manage the pace of taking initiative, which is intrinsically linked to the smoothness of the conversation. Following some existing studies of PCAs [16, 79], we measure the smoothness by the contextual semantic similarity between the last utterance and the generated response.
- **Timing Sensitivity** refers to the ability to take initiative accounting for real-time user needs and status, which can be alternatively evaluated by the user satisfaction at each conversation turn. There are different approaches to measure user satisfaction, such as topic-based scoring [39], data-driven estimation [19, 81], and LLM-based prediction [28] (adopted).
- **Self-awareness** refers to the ability to recognize its own limitations. Inspired by uncertainty calibration studies [10, 64], we adopt Expected Calibration Error (ECE) to measures how well an agent's confidence match the observed accuracy. Specifically, we regard the success rate as the accuracy and the probability of taking initiative behaviour as the confidence of PCAs.

6.1.2 Evaluation Protocols for Civility. There are five representative abilities of Civility, including Boundary Respect, Moral Integrity, Trust and Safety, Manners, and Emotional Intelligence. Similar to the analysis in Section 4.1.2, we also adopt the `Perspective API` [37] for automatically scoring the first four abilities based on the corresponding attributes of Identity Attack, Toxicity, Threat, and Insult. As for the last ability, *i.e.*, Emotional Intelligence, we adopt Emotional Intensity Relaxation [22] for evaluation.

6.2 Empirical Analysis

We adopt a widely-studied emotional support dialogue dataset, ESConv [42], as the testbed for analysis. According to the three main types of human alignment approaches in Section 5, we adopt corresponding methods for evaluation, including two prompt-based methods (Ask-an-Expert [88] and ProCoT [16]), two supervised fine-tuning methods (AugESC [92] and ExTES [93]), and two RL-based methods (RLHF [54] and Aligned$_{\text{d-PM}}$ [68]). Almost all the existing studies only evaluate the PCA from the dimension of Intelligence,

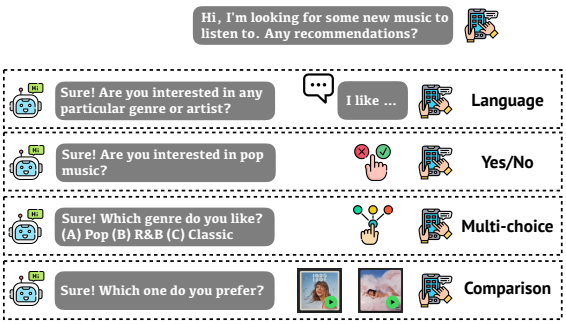

Figure 4: Different user interface designs for user preference elicitation in conversational recommender systems.

while in our analysis, we further include the aforementioned evaluation protocols for assessing the agents' abilities from the dimensions of Adaptivity and Civility. Experimental results are summarized in Table 3. There are two notable observations from the results:

- **Methods perform diversely in terms of metrics of three dimensions.** For example, although $\text{Aligned}_{\text{d-PM}}$ achieves the best performance in Intelligence, *i.e.*, it still under-performs in some metrics in the other two dimensions.
- **It attaches great importance in involving the three dimensions into model learning.** Compared with the standard RLHF, $\text{Aligned}_{\text{d-PM}}$ actually improves the performance in the dimension Adaptivity by taking into account the diversified user preferences. Similarly, ExTES performs better than AugESC in Adaptivity by using human knowledge to design various strategies and collect real-world counseling cases.

6.3 Prospects on Multidimensional Evaluation

As discussed in the empirical analysis, higher scores in terms of Intelligence-related metrics is not necessarily accompanied with better performance in Adaptivity and Civility. This helps us understand the limitations of current evaluation methods and sheds light on future opportunities.

6.3.1 Robust Evaluation Protocols. Due to the subjective nature and long-term impact for the evaluation of Adaptivity and Civility, human interactive judgments remain the most effective method for assessing these two human-centered dimensions. However, human judgements are challenging to scale and lack standardization for comparisons. For example, Huang et al. [29] integrates both system- and user-centric factors into the evaluation of conversational recommender systems. Our proposed taxonomy offers a solid foundation for developing suitable automatic evaluation metrics. While we explore various alternative metrics in Section 6.1, there is a clear need for more reliable and robust automatic metrics that seamlessly reflect the attributes of human-like conversational agents.

6.3.2 Customized Evaluation Framework. Different types of PCAs will need different evaluation gold standard, rather than solely evaluating on the Intelligence dimension or requiring high proficiency in all three dimensions. For example, for some `Listener`-type PCAs, like social chatbots, even Intelligence is not a necessary evaluation standard as humans may only need a listener who can just listen to their stories with limited initiative and planning capabilities. In certain emergency scenarios, the importance of respecting boundaries may diminish, as indicated by the user study in Zargham et al. [86] which reveals that concerns about privacy can be overshadowed by the physical safety. Therefore, it is crucial to customize the human-centered evaluation framework of PCAs for different social contexts or applied scenarios.

7 SYSTEM DEPLOYMENT

From the perspective of ubiquitous computing [73], the most profound human-centered proactive conversational agents are those that seamlessly integrate into daily life, becoming a natural part of it. To achieve this, the design of human-centered system deployment should focus more on human behaviour patterns rather than just extending technology functionalities.

7.1 Human-centered Designs of User Interface

While conversational agents often prioritize language as the main interface, this design can be imprecise for tasks needing specific user inputs. Language-based interactions can be challenging to control for desired outcomes and might be unsuitable for precision-required tasks. Additionally, these interactions can sometimes be inconvenient and inefficient for quick or straightforward tasks, where simpler interfaces might be more effective. Balancing language interfaces with other interaction modalities can enhance both user experience and task efficiency. In the realm of PCAs, it's essential to carefully design user interfaces that are minimally intrusive for system-initiated interactions. This consideration ensures a balance between functionality and user comfort.

Take conversational recommender systems as an example. The system-initiated interactions in conversational recommender systems typically refer to user preference elicitation, which aims to explicitly acquire user preference rather than solely inferring users' implicit preference from the conversation history. As shown in Figure 4, the most popular paradigm is called "System Ask, User Respond" [90], where the PCA uses **language** as the interface to ask eliciting questions for collecting users' preference descriptions. However, the language interface of PCAs faces several challenges in conversational recommender systems: 1) understanding user preferences from natural language itself is a challenging problem, and 2) users may feel uncomfortable to frequently provide their personal information to the technology company. In recent years, many other types of user preference elicitation interfaces have been studied, including asking **Yes/No** questions or **multiple-choice** questions [38, 91] for identifying users' preference towards specific item features, and presenting **comparisons** [76] for obtaining relative preference feedback. In this manner, human users can seamlessly and precisely convey their preference by tapping on certain icons on their devices to interact with PCAs.

7.2 New Emphasis on Trust and Reliance

When interacting with human-centered proactive conversational agents, users should feel comfortable using and depending on the system for proactively achieving their goals. Inspired by HCI studies [4, 65], human-centered PCAs should exhibit appropriate trust and reliance from human users. As illustrated in Figure 5, when provided with a system-initiated behaviour from a proactive conversational

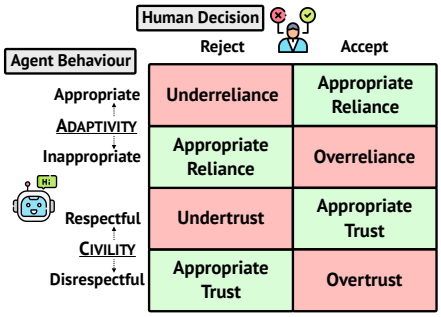

Figure 5: A simplified diagram outlining the different ways people rely on and trust proactive conversational agents.

agent, the human user has the choice to either accept or reject the interaction. The agent may exhibit appropriate or inappropriate behaviours in terms of their Adaptivity, and also respectful or disrespectful behaviours in terms of their Civility.

- **Appropriate Reliance**: the human accepts appropriate agent behaviours or corrects inappropriate agent behaviours.
- **Underreliance**: fails to accept appropriate agent behaviours.
- **Overreliance**: fails to correct inappropriate agent behaviours.
- **Appropriate Trust**: accepts respectful agent behaviours or rejects disrespectful agent behaviours.
- **Undertrust**: refuses to accept respectful agent behaviours.
- **Overtrust**: fails to reject disrespectful agent behaviours.

We advocate human-agent interactions with appropriate reliance and trust, since underreliance and undertrust can impede the goal achievement while overreliance and overtrust may pose risks to human users. As presented in Figure 6, we discuss several approaches from the HCI perspective to address this issue.

7.2.1 Explanability. Explanations help bridge the gap between complex AI decision-making processes and human understanding. Extensive HCI studies validate that presenting explanations can reduce users' overreliance [9, 65] and overtrust [51] on the prediction of AI systems. In the context of human-centered PCAs, we introduce four types of explanations that can be presented to human users. 1) Feature-based explanations show the contribution of different features to the decision making of PCAs. 2) Example-based explanations present either representative prototypes of the decision that the PCA makes for the given instance or examples that are similar to the given instance along with the PCA's decision. 3) Path-based explanations present the decision path made by the PCA between the initial state and the current state. 4) Attribution-based explanations attributes the output decision to specific parts or components of the input data or external knowledge.

7.2.2 Reliability. Another key finding in HCI studies [4, 69, 82] is the importance of "reliability disclosure" in shaping users' trust and reliance on system feedback. This concept involves explicitly informing users about the estimated reliability or uncertainty of the system's feedback. For example, human users were shown the system's confidence along with the decision, like "*The agent is 87% confident in its suggestion*". Meanwhile, how well the system can estimate its confidence is also a challenging problem, known as self-calibration. This involves a range of methods to enhance

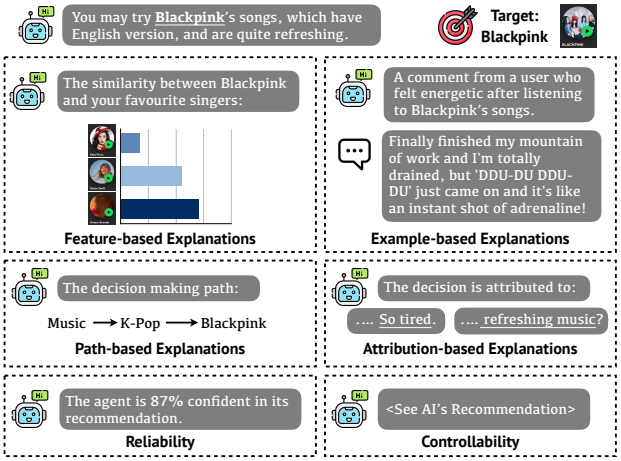

Figure 6: Example UI designs regarding appropriate trust and reliance for target-guided conversational recommendation.

the system's ability to assess and communicate its own confidence [48, 64]. When users are aware of an agent's confidence, as assessed by the agent itself, it significantly impacts their trust and reliance on the conversational agent.

7.2.3 Controllability. In conventional designs, proactive conversational agents often hold complete autonomy in deciding when to initiate interactions, leaving users with little choice but to engage with these system-initiated behaviors. However, some user studies in HCI literature [4, 5] suggests the importance of empowering users with control over these interactions. Allowing users the flexibility to determine the necessity of system-initiated behaviors can significantly address issues of underreliance and undertrust. For example, the system-initiated behaviours, such as asking clarification questions or providing suggestions, are on demand and not presented to the users by default, while users could choose to see this content by clicking on a button, like "*See AI's suggestion*".

8 CONCLUSIONS

This perspectives paper investigated proactive conversational agents from the human-centered perspective. We first proposed a new taxonomy concerning three key dimensions of human-centered PCAs, including Intelligence, Adaptivity, and Civility. According to this taxonomy, we re-interpreted existing literature on PCAs upon the five stages of PCA system construction (*i.e.*, Task Formulation, Data Preparation, Model Learning, Evaluation, and System Deployment). In the light of the limitations, we envisioned future research agenda and prospects for achieving human-centered PCAs. Meanwhile, PCAs are advancing towards the realm of superintelligent AI, where maintaining a human-centered system is crucial to ensure these superintelligent AIs continue to serve human's interests.

ACKNOWLEDGEMENT

This research was supported by U.S. National Science Foundation IIS-2336768 and NExT Research Center. This research was also supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (Proposal ID: 23-SIS-SMU-010).

REFERENCES

[1] Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and Evaluating Open-Domain Dialogue Corpora with Clarifying Questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*. 4473–4484.

[2] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *SIGIR 2019*. ACM, 475–484.

[3] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*. ACM, 475–484.

[4] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum. Comput. Interact.* 5, CSCW1 (2021), 188:1–188:21.

[5] Wanling Cai, Yucheng Jin, and Li Chen. 2022. Impacts of Personal Characteristics on User Trust in Conversational Recommender Systems. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*. ACM, 489:1–489:14.

[6] Iñigo Casanueva, Ivan Vulic, Georgios Spithourakis, and Pawel Budzianowski. 2022. NLU++: A Multi-Label, Slot-Rich, Generalisable Dataset for Natural Language Understanding in Task-Oriented Dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*. Association for Computational Linguistics, 1998–2013.

[7] Ana Paula Chaves and Marco Aurélio Gerosa. 2021. How Should My Chatbot Interact? A Survey on Social Characteristics in Human-Chatbot Interaction Design. *Int. J. Hum. Comput. Interact.* 37, 8 (2021), 729–758.

[8] Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. 2023. Controllable Mixed-Initiative Dialogue Generation through Prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 951–966.

[9] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *CoRR* abs/2301.07255 (2023).

[10] Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. 2023. A Close Look into the Calibration of Pre-trained Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 1343–1367.

[11] Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul A. Crook, and William Yang Wang. 2022. KETOD: Knowledge-Enriched Task-Oriented Dialogue. In *Findings of ACL: NAACL 2022*. 2581–2593.

[12] Jeffrey Dalton, Sophie Fischer, Paul Owoicho, Filip Radlinski, Federico Rossetto, Johanne R. Trippas, and Hamed Zamani. 2022. Conversational Information Seeking: Theory and Application. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*. ACM, 3455–3458.

[13] Huy Dao, Lizi Liao, Dung D. Le, and Yuxiang Nie. 2023. Reinforced Target-driven Conversational Promotion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 12583–12596.

[14] Yang Deng, Wenqiang Lei, Minlie Huang, and Tat-Seng Chua. 2023. Rethinking Conversational Agents in the Era of LLMs: Proactivity, Non-collaborativity, and Beyond. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2023, Beijing, China, November 26-28, 2023*. ACM, 298–301.

[15] Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023. A Survey on Proactive Dialogue Systems: Problems, Methods, and Prospects. In *IJCAI 2023*. 6583–6591.

[16] Yang Deng, Wenqiang Lei, Lizi Liao, and Tat-Seng Chua. 2023. Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration. *CoRR* abs/2305.13626 (2023).

[17] Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 2022. PACIFIC: Towards Proactive Conversational Question Answering over Tabular and Textual Data in Finance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*. 6970–6984.

[18] Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. Unified Conversational Recommendation Policy Learning via Graph-based Reinforcement Learning. In *SIGIR 2021*. ACM, 1431–1441.

[19] Yang Deng, Wenxuan Zhang, Wai Lam, Hong Cheng, and Helen Meng. 2022. User Satisfaction Estimation with Sequential Dialogue Act Modeling in Goal-oriented Conversational Systems. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*. ACM, 2998–3008.

[20] Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2023. Plug-and-Play Policy Planner for Large Language Model Powered Dialogue Agents. *CoRR* abs/2311.00262 (2023).

[21] Yang Deng, Wenxuan Zhang, Weiwen Xu, Wenqiang Lei, Tat-Seng Chua, and Wai Lam. 2023. A Unified Multi-task Learning Framework for Multi-goal Conversational Recommender Systems. *ACM Trans. Inf. Syst.* 41, 3 (2023), 77:1–77:25.

[22] Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced Mixed-initiative Dialogue System for Emotional Support Conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*. 4079–4095.

[23] Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*. Association for Computational Linguistics, 8173–8188.

[24] Ritam Dutt, Sayan Sinha, Rishabh Joshi, Surya Shekhar Chakraborty, Meredith Riggs, Xinru Yan, Haogang Bao, and Carolyn P. Rosé. 2021. ResPer: Computationally Modelling Resisting Strategies in Persuasive Conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*. Association for Computational Linguistics, 78–90.

[25] Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback. *CoRR* abs/2305.10142 (2023). https://doi.org/10.48550/arXiv.2305.10142

[26] Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. Abg-CoQA: Clarifying Ambiguity in Conversational Question Answering. In *AKBC 2021*.

[27] He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling Strategy and Generation in Negotiation Dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Association for Computational Linguistics, 2333–2343.

[28] Zhiyuan Hu, Yue Feng, Anh Tuan Luu, Bryan Hooi, and Aldo Lipani. 2023. Unlocking the Potential of User Feedback: Leveraging Large Language Model as User Simulators to Enhance Dialogue System. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*. ACM, 3953–3957.

[29] Chen Huang, Peixin Qin, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Tat-Seng Chua. 2024. Concept–An Evaluation Protocol on Conversation Recommender Systems with System-and User-centric Factors. *arXiv preprint arXiv:2404.03304* (2024).

[30] Gareth J. F. Jones, Procheta Sen, Debasis Ganguly, and Emine Yilmaz. 2022. Workshop on Proactive and Agent-Supported Information Retrieval (PASIR). In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. ACM, 5167–5168.

[31] Rishabh Joshi, Vidhisha Balachandran, Shikhar Vashishth, Alan W. Black, and Yulia Tsvetkov. 2021. DialoGraph: Incorporating Interpretable Strategy-Graph Networks into Negotiation Dialogues. In *ICLR 2021*.

[32] Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A Prosocial Backbone for Conversational Agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*. Association for Computational Linguistics, 4005–4029.

[33] Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. 2020. Effects of Proactive Dialogue Strategies on Human-Computer Trust. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2020, Genoa, Italy, July 12-18, 2020*. ACM, 107–116.

[34] Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. 2021. Modelling and Predicting Trust for Developing Proactive Dialogue Strategies in Mixed-Initiative Interaction. In *ICMI '21: International Conference on Multimodal Interaction, Montréal, QC, Canada, October 18-22, 2021*. ACM, 131–140.

[35] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a Benchmark for Question Answering Research. *Trans. Assoc. Comput. Linguistics* 7 (2019), 452–466.

[36] Tiziano Labruna, Sofia Brenna, Andrea Zaninello, and Bernardo Magnini. 2023. Unraveling ChatGPT: A Critical Analysis of AI-Generated Goal-Oriented Dialogues and Annotations. In *AIxIA 2023 - Advances in Artificial Intelligence - XXIInd International Conference of the Italian Association for Artificial Intelligence, AIxIA 2023, Rome, Italy, November 6-9, 2023, Proceedings (Lecture Notes in Computer Science, Vol. 14318)*. Springer, 151–171.

[37] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Prakash Gupta, Donald Metzler, and Lucy Vasserman. 2022. A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*. ACM, 3197–3207.

[38] 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 *WSDM 2020*. 304–312.

[39] Wenqiang Lei, Yao Zhang, Feifan Song, Hongru Liang, Jiaxin Mao, Jiancheng Lv, Zhenglu Yang, and Tat-Seng Chua. 2022. Interacting with Non-Cooperative User: A New Paradigm for Proactive Dialogue Policy. In *SIGIR 2022*. 212–222.

[40] Yu Li, Kun Qian, Weiyan Shi, and Zhou Yu. 2020. End-to-End Trainable Non-Collaborative Dialog System. In *AAAI 2020*. 8293–8302.

[41] 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, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*. ACM, 3452–3455.

[42] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards Emotional Support Dialog Systems. In *ACL/IJCNLP 2021*. 3469–3483.

[43] Yuanxing Liu, Weinan Zhang, Baohua Dong, Yan Fan, Hang Wang, Fan Feng, Yifan Chen, Ziyu Zhuang, Hengbin Cui, Yongbin Li, and Wanxiang Che. 2023. U-NEED: A Fine-grained Dataset for User Needs-Centric E-commerce Conversational Recommendation. In *SIGIR 2023*.

[44] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards Conversational Recommendation over Multi-Type Dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*. 1036–1049.

[45] Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 5602–5621.

[46] Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and Time-aware Joint Contextual Learning for Dialogue-act Classification in Counselling Conversations. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*. ACM, 735–745.

[47] Chuan Meng, Mohammad Aliannejadi, and Maarten de Rijke. 2023. System Initiative Prediction for Multi-turn Conversational Information Seeking. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*. ACM, 1807–1817.

[48] Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing Conversational Agents' Overconfidence through Linguistic Calibration. *Trans. Assoc. Comput. Linguistics* 10 (2022), 857–872.

[49] Ondrej Miksik, I. Munasinghe, J. Asensio-Cubero, S. Reddy Bethi, S.-T. Huang, S. Zylfo, X. Liu, T. Nica, A. Mitrocsak, S. Mezza, Rory Beard, Ruibo Shi, Raymond W. M. Ng, Pedro A. M. Mediano, Zafeirios Fountas, S.-H. Lee, J. Medvesek, H. Zhuang, Yvonne Rogers, and Pawel Swietojanski. 2020. Building Proactive Voice Assistants: When and How (not) to Interact. *CoRR* abs/2005.01322 (2020).

[50] Kshitij Mishra, Azlaan Mustafa Samad, Palak Totala, and Asif Ekbal. 2022. PEPDS: A Polite and Empathetic Persuasive Dialogue System for Charity Donation. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*. International Committee on Computational Linguistics, 424–440.

[51] Sina Mohseni, Fan Yang, Shiva K. Pentyala, Mengnan Du, Yi Liu, Nic Lupfer, Xia Hu, Shuiwang Ji, and Eric D. Ragan. 2021. Machine Learning Explanations to Prevent Overtrust in Fake News Detection. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*. AAAI Press, 421–431.

[52] OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. *OpenAI blog* (2022). https://openai.com/blog/chatgpt/

[53] OpenAI. 2023. Introducing Superalignment. *OpenAI blog* (2023). https://openai.com/blog/introducing-superalignment

[54] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

[55] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence C. An. 2016. Building a Motivational Interviewing Dataset. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA*. The Association for Computational Linguistics, 42–51.

[56] Hossein A. Rahmani, Xi Wang, Yue Feng, Qiang Zhang, Emine Yilmaz, and Aldo Lipani. 2023. A Survey on Asking Clarification Questions Datasets in Conversational Systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*. Association for Computational Linguistics, 2698–2716.

[57] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *EMNLP 2016*. 2383–2392.

[58] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *ACL 2019*. Association for Computational Linguistics, 5370–5381.

[59] Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, and Rainer Malaka. 2021. May I Interrupt? Diverging Opinions on Proactive Smart Speakers. In *CUI 2021*. ACM, 34:1–34:10.

[60] Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul N. Bennett. 2020. Leading Conversational Search by Suggesting Useful Questions. In *WWW 2020*. ACM / IW3C2, 1160–1170.

[61] Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Empathetic Persuasion: Reinforcing Empathy and Persuasiveness in Dialogue Systems. In *Findings of ACL: NAACL 2022*. 844–856.

[62] Ivan Sekulic, Mohammad Aliannejadi, and Fabio Crestani. 2022. Evaluating Mixed-initiative Conversational Search Systems via User Simulation. In *WSDM 2022*.

[63] Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric P. Xing, and Zhiting Hu. 2019. Target-Guided Open-Domain Conversation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*. 5624–5634.

[64] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 5433–5442.

[65] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum. Comput. Interact.* 7, CSCW1 (2023), 1–38.

[66] Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 12047–12064.

[67] Jian Wang, Yi Cheng, Dongding Lin, Chak Tou Leong, and Wenjie Li. 2023. Target-oriented Proactive Dialogue Systems with Personalization: Problem Formulation and Dataset Curation. *CoRR* abs/2310.07397 (2023).

[68] Jiashuo Wang, Haozhao Wang, Shichao Sun, and Wenjie Li. 2023. Aligning Language Models with Human Preferences via a Bayesian Approach. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[69] Lu Wang, Greg A. Jamieson, and Justin G. Hollands. 2009. Trust and Reliance on an Automated Combat Identification System. *Hum. Factors* 51, 3 (2009), 281–291.

[70] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In *ACL 2019*. 5635–5649.

[71] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning Large Language Models with Human: A Survey. *CoRR* abs/2307.12966 (2023).

[72] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *Trans. Mach. Learn. Res.* 2022 (2022).

[73] Mark Weiser. 1991. The Computer for the 21 st Century. *Scientific american* 265, 3 (1991), 94–105.

[74] Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. InSCIt: Information-Seeking Conversations with Mixed-Initiative Interactions. *Transactions of the Association for Computational Linguistics* 11 (05 2023), 453–468.

[75] Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann A. Copestake. 2021. TIAGE: A Benchmark for Topic-Shift Aware Dialog Modeling. In *Findings of the ACL: EMNLP 2021*. Association for Computational Linguistics, 1684–1690.

[76] Zhihui Xie, Tong Yu, Canzhe Zhao, and Shuai Li. 2021. Comparison-based Conversational Recommender System with Relative Bandit Feedback. In *SIGIR 2021*. 1400–1409.

[77] Chunpu Xu, Steffi Chern, Ethan Chern, Ge Zhang, Zekun Wang, Ruibo Liu, Jing Li, Jie Fu, and Pengfei Liu. 2023. Align on the Fly: Adapting Chatbot Behavior to Established Norms. *CoRR* abs/2312.15907 (2023).

[78] Sitong Yan, Shengli Song, Jingyang Li, Shiqi Meng, and Guangneng Hu. 2023. TITAN : Task-oriented Dialogues with Mixed-Initiative Interactions. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023*. 5251–5259.

[79] Zhitong Yang, Bo Wang, Jinfeng Zhou, Yue Tan, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2022. TopKG: Target-oriented Dialog via Global

Planning on Knowledge Graph. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022*. International Committee on Computational Linguistics, 745–755.

[80] Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From Instructions to Intrinsic Human Values - A Survey of Alignment Goals for Big Models. *CoRR* abs/2308.12014 (2023).

[81] Fanghua Ye, Zhiyuan Hu, and Emine Yilmaz. 2023. Modeling User Satisfaction Dynamics in Dialogue via Hawkes Process. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*. Association for Computational Linguistics, 8875–8889.

[82] Ming Yin, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019*. ACM, 279.

[83] Xiao Yu, Maximillian Chen, and Zhou Yu. 2023. Prompt-Based Monte-Carlo Tree Search for Goal-oriented Dialogue Policy Planning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 7101–7125.

[84] Hamed Zamani, Susan T. Dumais, Nick Craswell, Paul N. Bennett, and Gord Lueck. 2020. Generating Clarifying Questions for Information Retrieval. In *WWW 2020*. ACM / IW3C2, 418–428.

[85] Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2023. Conversational Information Seeking. *Found. Trends Inf. Retr.* 17, 3-4 (2023), 244–456.

[86] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Voelkel, Johannes Schöning, Rainer Malaka, and Yvonne Rogers. 2022. Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma. In *CUI 2022*. ACM, 3:1–3:14.

[87] Haolan Zhan, Yufei Wang, Tao Feng, Yuncheng Hua, Suraj Sharma, Zhuang Li, Lizhen Qu, and Gholamreza Haffari. 2022. Let's Negotiate! A Survey of Negotiation Dialogue Systems. *CoRR* abs/2212.09072 (2022).

[88] Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023. Ask an Expert: Leveraging Language Models to Improve Strategic Reasoning in Goal-Oriented Dialogue Models. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 6665–6694.

[89] Tong Zhang, Chen Huang, Yang Deng, Hongru Liang, Jia Liu, Zujie Wen, Wenqiang Lei, and Tat-Seng Chua. 2024. Strength Lies in Differences! Towards Effective Non-collaborative Dialogues via Tailored Strategy Planning. *CoRR* abs/2403.06769 (2024).

[90] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*. ACM, 177–186.

[91] Yiming Zhang, Lingfei Wu, Qi Shen, Yitong Pang, Zhihua Wei, Fangli Xu, Bo Long, and Jian Pei. 2022. Multiple Choice Questions based Multi-Interest Policy Learning for Conversational Recommendation. In *WWW 2022*. 2153–2162.

[92] Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 1552–1568.

[93] Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building Emotional Support Chatbots in the Era of LLMs. *CoRR* abs/2308.11584 (2023).

[94] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In *AAAI 2018*. 730–739.

[95] Jie Zou, Mohammad Aliannejadi, Evangelos Kanoulas, Maria Soledad Pera, and Yiqun Liu. 2023. Users Meet Clarifying Questions: Toward a Better Understanding of User Interactions for Search Clarification. *ACM Trans. Inf. Syst.* 41, 1 (2023), 16:1–16:25.

[96] Jie Zou, Aixin Sun, Cheng Long, Mohammad Aliannejadi, and Evangelos Kanoulas. 2023. Asking Clarifying Questions: To benefit or to disturb users in Web search? *Inf. Process. Manag.* 60, 2 (2023), 103176.